# Econometrics

Michael Creel

May 30, 2023

Universitat Autònoma de Barcelona

# Contents

**1 About this document** **25**

1.1 Prerequisites . . . . . 25

1.2 Contents . . . . . 28

1.3 License . . . . . 31

1.4 Obtaining and using the materials . . . . . 31

**2 Introduction to Julia** **35**

2.1 Why Julia? . . . . . 36

2.2 Why not Julia? . . . . . 37

2.3 Resources . . . . . 39
2.4 Installation of Julia and packages . . . . . 40
2.5 Using a system image . . . . . 43
2.6 Running Julia and the work flow . . . . . 44
2.7 Loading/saving data . . . . . 46
2.8 Exploratory analysis and plotting . . . . . 48

**3 Introduction: Economic and econometric models 49**

**4 Ordinary Least Squares 61**
4.1 The Linear Model . . . . . 61
4.2 Estimation by least squares . . . . . 64
4.3 Geometric interpretation of least squares estimation . . . . . 69
4.4 Influential observations and outliers . . . . . 76
4.5 Goodness of fit . . . . . 80
4.6 The classical linear regression model . . . . . 85

4.7 Small sample statistical properties of the least squares estimator . 88
4.8 Example: The Nerlove model . . . . . 102
4.9 Exercises . . . . . 110

**5 Asymptotic properties of the least squares estimator 112**
5.1 Consistency . . . . . 114
5.2 Asymptotic normality . . . . . 116
5.3 Asymptotic efficiency . . . . . 119
5.4 Exercises . . . . . 121

**6 Restrictions and hypothesis tests 124**
6.1 Exact linear restrictions . . . . . 124
6.2 Testing . . . . . 135
6.3 The asymptotic equivalence of the LR, Wald and score tests . . . . 149
6.4 Interpretation of test statistics . . . . . 156
6.5 Confidence intervals . . . . . 157

6.6 Bootstrapping . . . . . 160
6.7 Wald test for nonlinear restrictions: the delta method . . . . . 164
6.8 Example: the Nerlove data . . . . . 170
6.9 Exercises . . . . . 179

**7 Stochastic regressors 182**
7.1 Case 1 . . . . . 185
7.2 Case 2 . . . . . 188
7.3 Case 3 . . . . . 190
7.4 When are the assumptions reasonable? . . . . . 192
7.5 Exercises . . . . . 195

**8 Data problems 196**
8.1 Collinearity . . . . . 197
8.2 Measurement error . . . . . 226
8.3 Missing observations . . . . . 235

8.4 Missing regressors . . . . . 244
8.5 Exercises . . . . . 245

**9 Functional form and nonnested tests 247**
9.1 Flexible functional forms . . . . . 250
9.2 Testing nonnested hypotheses . . . . . 269

**10 Generalized least squares 276**
10.1 Effects of non-spherical disturbances on the OLS estimator . . . . 278
10.2 The GLS estimator . . . . . 284
10.3 Feasible GLS . . . . . 290
10.4 Heteroscedasticity . . . . . 293
10.5 Autocorrelation . . . . . 312
10.6 Exercises . . . . . 350

**11 Endogeneity and simultaneity 360**
11.1 Simultaneous equations . . . . . 362

11.2 Reduced form . . . . . 370
11.3 Estimation of the reduced form equations . . . . . 375
11.4 Bias and inconsistency of OLS estimation of a structural equation . 380
11.5 Note about the rest of this chapter . . . . . 384
11.6 Identification by exclusion restrictions . . . . . 384
11.7 2SLS . . . . . 401
11.8 Testing the overidentifying restrictions . . . . . 407
11.9 System methods of estimation . . . . . 417
11.10 Example: Klein's Model 1 . . . . . 429

**12 Numeric optimization methods 439**

12.1 Search . . . . . 443
12.2 Derivative-based methods . . . . . 446
12.3 Global methods . . . . . 469
12.4 Examples . . . . . 476
12.5 Practical Summary . . . . . 490

12.6 Exercises . . . . . 491

**13 Asymptotic properties of extremum estimators 495**

13.1 Extremum estimators . . . . . 496

13.2 Existence . . . . . 509

13.3 Consistency . . . . . 510

13.4 Example: Consistency of Least Squares . . . . . 521

13.5 More on the limiting objective function: correctly and incorrectly specified models . . . . . 527

13.6 Example: Inconsistency of Misspecified Least Squares . . . . . 532

13.7 Example: Linearization of a nonlinear model . . . . . 535

13.8 Asymptotic Normality . . . . . 543

13.9 Example: Classical linear model . . . . . 548

13.10 Practical Summary . . . . . 554

13.11 Exercises . . . . . 555

**14 Application: a simple DSGE model** **557**

14.1 The model . . . . . 560

14.2 Solution of the model and generation of data . . . . . 579

**15 Maximum likelihood estimation** **582**

15.1 The likelihood function . . . . . 584

15.2 Consistency of MLE . . . . . 605

15.3 The score function . . . . . 610

15.4 Asymptotic normality of MLE . . . . . 619

15.5 The information matrix equality . . . . . 634

15.6 The Cramér-Rao lower bound . . . . . 646

15.7 Likelihood ratio-type tests . . . . . 653

15.8 Examples . . . . . 659

15.9 ML estimation of the DSGE model . . . . . 695

15.10 Practical Summary . . . . . 701

15.11 Exercises . . . . . 702

**16 Generalized method of moments** **712**

16.1 Moment conditions . . . . . 713

16.2 Definition of GMM estimator . . . . . 724

16.3 Consistency . . . . . 727

16.4 Asymptotic normality . . . . . 732

16.5 Choosing the weighting matrix . . . . . 742

16.6 Estimation of the variance-covariance matrix . . . . . 750

16.7 Two step and continuously updated GMM estimators . . . . . 758

16.8 Estimation using conditional moments . . . . . 762

16.9 Generalized instrumental variables estimator for linear models . . . 769

16.10 The Hansen-Sargan (or J) test . . . . . 791

16.11 Other estimators interpreted as GMM estimators . . . . . 800

16.12 The Hausman Test . . . . . 804

16.13 More moments are better (?) . . . . . 819

16.14 Examples . . . . . 822

16.15 Practical Summary . . . . . 852
16.16 Exercises . . . . . 853

**17 Models for time series data** **871**
17.1 ARMA models . . . . . 881
17.2 VAR models . . . . . 904
17.3 ARCH, GARCH and Stochastic volatility . . . . . 915
17.4 Diffusion models . . . . . 934
17.5 State space models . . . . . 941
17.6 Nonstationarity and cointegration . . . . . 947
17.7 Exercises . . . . . 949

**18 Bayesian methods** **953**
18.1 Definitions . . . . . 955
18.2 Philosophy, etc. . . . . 959
18.3 Example . . . . . 963

18.4 Theory . . . . . 965
18.5 Computational methods . . . . . 969
18.6 Examples . . . . . 981
18.7 Full sample Bayesian estimation of the DSGE model . . . . . 983
18.8 Bayesian GMM for the DSGE model . . . . . 988
18.9 Exercises . . . . . 994

**19 Introduction to panel data** **995**
19.1 Generalities . . . . . 997
19.2 Static models and correlations between variables . . . . . 1008
19.3 Estimation of the simple linear panel model . . . . . 1013
19.4 Dynamic panel data . . . . . 1025
19.5 Examples . . . . . 1032
19.6 Practical summary . . . . . 1037
19.7 Exercises . . . . . 1038

**20 Nonparametric inference** **1041**

20.1 Estimation of regression functions . . . . . 1057

20.2 Density function estimation . . . . . 1085

20.3 Examples . . . . . 1096

20.4 Exercises . . . . . 1112

**21 Quantile regression** **1113**

21.1 Quantiles of the linear regression model . . . . . 1116

21.2 Fully nonparametric conditional quantiles . . . . . 1120

21.3 Quantile regression as a semi-parametric estimator . . . . . 1122

21.4 Returns to schooling: quantile regression, quantile IV regression, and Bayesian GMM via MCMC . . . . . 1128

**22 Simulation-based methods for estimation and inference** **1139**

22.1 Motivation . . . . . 1144

22.2 Simulated maximum likelihood (SML) . . . . . 1153

22.3 Method of simulated moments (MSM) . . . . . 1157
22.4 Example: stochastic volatility . . . . . 1170
22.5 Simulated Neural Moments estimation of the DSGE model . . . . 1181
22.6 Exercises . . . . . 1182
22.7 Practical summary . . . . . 1183

**23 Notation and Review** **1184**
23.1 Notation for differentiation of vectors and matrices . . . . . 1185
23.2 Convergence modes . . . . . 1187
23.3 Rates of convergence and asymptotic equality . . . . . 1194
23.4 Slutsky Theorem and Continuous Mapping Theorem . . . . . 1198

**24 The attic** **1200**
24.1 Efficient method of moments (EMM) . . . . . 1201
24.2 Parallel programming for econometrics . . . . . 1212
24.3 Quasi-ML . . . . . 1231

24.4 Nonlinear simultaneous equations . . . . . 1240
24.5 Example: The MEPS data . . . . . 1242
24.6 Hurdle models . . . . . 1265
24.7 Finite mixture models . . . . . 1275
24.8 Nonlinear least squares (NLS) . . . . . 1282
24.9 The Fourier functional form . . . . . 1306

**Bibliography** **1325**

# List of Figures


1.1 Julia . . . . . 30
1.2 LyX . . . . . 32

3.1 Price and Quantity, colored by income (blue is low, violet is high) . 54

4.1 Typical data, Classical Model . . . . . 65
4.2 Example OLS Fit . . . . . 70
4.3 The fit in observation space . . . . . 72
4.4 Detection of influential observations . . . . . 79
4.5 Uncentered $R^2$ . . . . . 82

4.6 Unbiasedness of OLS under classical assumptions: replications of $\hat{\beta}$ minus true $\beta$ . . . . . 90
4.7 Biasedness of OLS when an assumption fails: : replications of $\hat{\beta}$ minus true $\beta$ . . . . . 92
4.8 Gauss-Markov Result: The OLS estimator . . . . . 98
4.9 Gauss-Markov Resul: The split sample estimator . . . . . 99

6.1 Joint and Individual Confidence Regions . . . . . 159
6.2 RTS as a function of firm size . . . . . 178

8.1 $s(\beta)$ when there is no collinearity . . . . . 209
8.2 $s(\beta)$ when there is collinearity . . . . . 210
8.3 Collinearity: Monte Carlo results . . . . . 215
8.4 OLS and Ridge regression . . . . . 225
8.5 $\hat{\rho} - \rho$ with and without measurement error . . . . . 235
8.6 Sample selection bias . . . . . 241

10.1 Rejection frequency of 10% t-test, H0 is true. . . . . 282
10.2 Motivation for GLS correction when there is HET . . . . . 305
10.3 Residuals from time trend for CO2 data . . . . . 316
10.4 Autocorrelation induced by misspecification . . . . . 319
10.5 Efficiency of OLS and FGLS, AR1 errors . . . . . 335
10.6 Durbin-Watson critical values . . . . . 347
10.7 Dynamic model with MA(1) errors . . . . . 351

11.1 Exogeneity and Endogeneity (adapted from Cameron and Trivedi) 361

12.1 Search method . . . . . 444
12.2 Grid search, one dimension . . . . . 445
12.3 Increasing directions of search . . . . . 450
12.4 Newton iteration . . . . . 455
12.5 Trace of SA path to minimize sum of squared errors . . . . . 475
12.6 Multiple local maxima: Mountains with low fog . . . . . 488

13.1 OLS objective function contours . . . . . 502
13.2 Why uniform convergence of $s_n(\theta)$ is needed . . . . . 517
13.3 Consistency of OLS . . . . . 526
13.4 Linear Approximation . . . . . 540
13.5 Effects of $I_\infty$ and $J_\infty$ . . . . . 547

14.1 The DSGE data . . . . . 580

15.1 Alternative variance computations for the OBDV Poisson model . 645

16.1 Asymptotic Normality of GMM estimator, $\chi^2$ example . . . . . 742
16.2 GIV estimation results for $\hat{\rho}-\rho$, dynamic model with measurement error . . . . . 787
16.3 OLS and IV ($n = 1000$,cov(X,$\epsilon$) $= 0.5$,cov(W,$\epsilon$) $= 0.5$) . . . . . 806
16.4 Incorrect rank and the Hausman test . . . . . 814

17.1 NYSE weekly close price, 100 $\times$log differences . . . . . 918
17.2 SV model, typical data and density . . . . . 932

17.3 Returns from jump-diffusion model . . . . . 937
17.4 Spot volatility, jump-diffusion model . . . . . 938

18.1 Bayesian estimation, exponential likelihood, lognormal prior . . . . 965
18.2 Chernozhukov and Hong, Theorem 2 . . . . . 966
18.3 Metropolis-Hastings MCMC, exponential likelihood, lognormal prior 982
18.4 MCMC results for simple DSGE example model (two different runs using different observed variables), plus the GMM results for comparison . . . . . 986
18.5 CGHK model, posteriors . . . . . 987

20.1 True and simple approximating functions . . . . . 1047
20.2 True and approximating elasticities . . . . . 1049
20.3 True function and more flexible approximation . . . . . 1053
20.4 True elasticity and more flexible approximation . . . . . 1054
20.5 A simple neural net . . . . . 1077
20.6 Negative binomial raw moments . . . . . 1093

20.7 Kernel regression fits, OBDV health care usage versus AGE and INCOME . . . . . 1099
20.8 Dollar-Euro . . . . . 1100
20.9 Dollar-Yen . . . . . 1101
20.10 Kernel regression fitted conditional second moments, Yen/Dollar and Euro/Dollar . . . . . 1104

21.1 Inverse CDF for N(0,1) . . . . . 1118
21.2 Quantiles of classical linear regression model . . . . . 1119
21.3 Quantile regression results . . . . . 1127
21.4 QR results for the Card data, $\tau$ sequence . . . . . 1131
21.5 Two chains for $\beta_0(\tau = 0.5)$, independent and correlated proposals . 1135
21.6 IV-QR results . . . . . 1138

22.1 SV model, typical data and density . . . . . 1171
22.2 MSM for SV model . . . . . 1174

22.3 MCMC estimation using simulated moments and limited information quasi-likelihood . . . . . 1175
22.4 CI coverage, SV model, MSM and Bayesian MSM . . . . . 1177
22.5 90, 95 and 99% CI coverage, SV model, using simulated neural moments . . . . . 1180

24.1 Speedups from parallelization . . . . . 1223
24.2 Life expectancy of mongooses, Weibull model . . . . . 1227
24.3 Life expectancy of mongooses, mixed Weibull model . . . . . 1230

# List of Tables

14.1 Variables . . . . . 561
14.2 Parameters . . . . . 562
14.3 Deterministic steady state . . . . . 574
14.4 True parameters and support of uniform priors. . . . . 578

15.1 Marginal Variances, Sample and Estimated (Poisson) . . . . . 675
15.2 Marginal Variances, Sample and Estimated (NB-II) . . . . . 684
15.3 Average Information Criteria, OBDV . . . . . 692

24.1 Actual and Poisson fitted frequencies . . . . . 1266

24.2 Actual and Hurdle Poisson fitted frequencies . . . . . 1274

# Chapter 1

# About this document

## 1.1 Prerequisites

These notes have been prepared under the assumption that the reader understands basic statistics, linear algebra, and mathematical optimization. There are many sources for this material, for example, the appendices to *Introductory Econometrics: A Modern Approach* by Jeffrey Wooldridge. It is the student's responsibility

to get up to speed on this material, it will not be covered in class.

This document integrates lecture notes for a one year graduate level course with computer programs that illustrate and apply the methods that are studied. The immediate availability of executable (and modifiable) example programs when using the PDF version of the document is a distinguishing feature of these notes. If printed, the document is a somewhat terse approximation to a textbook. These notes are not intended to be a perfect substitute for a printed textbook. If you are a student of mine, please note that last sentence carefully. There are many good textbooks available. Students taking my courses should read the appropriate sections from at least one of the following books (or other textbooks with similar level and content)

- Cameron and Trivedi (2005), *Microeconometrics - Methods and Applications.* This is the book I recommend to use, if you don't have some reason to choose a different one.

- Davidson, R. and J.G. MacKinnon, *Econometric Theory and Methods*

- Gallant, A.R., *An Introduction to Econometric Theory*
- Hamilton, J.D., *Time Series Analysis*

Some more advanced books:

- Davidson, R. and J.G. MacKinnon (1993) *Estimation and Inference in Econometrics*, Oxford Univ. Press.
- Gallant, *Nonlinear Statistical Models*.

Undergraduate level texts, if you need to catch up with some concepts

- Wooldridge (2003), *Introductory Econometrics: A Modern Approach* (undergraduate level, for supplementary use only. Be sure to see the appendices, which give good coverage of foundations).
- Stock and Watson, *Introduction to Econometrics*. This is the book used at the UAB for undergraduate courses in econometrics.

## 1.2 Contents

With respect to contents, the emphasis is on estimation and inference within the world of stationary data. The notes have been used to teach first year masters and pre-doctoral students, in two 30-40 hour courses. The first part covers linear regression, and the second part goes on to cover ML and GMM estimation of potentially nonlinear models. There are some topical chapters after this core material that give introductions to more specialized methods. Student with interest in quantitative methods go on to study this material more deeply in elective courses, which is why the presentation here of the later chapters is more broad than deep.

The integrated examples and the support files (available online at the github repository) are an important part of these notes. Julia (julialang.org) has been used for most of the example programs, which are scattered though the document. The examples and code use the current stable version of Julia, version 1.x. This choice is motivated by several factors. Julia runs on all of the popular operating systems,

it is free, and it is fast, thanks to just-in-time compilation. It is a relatively new language, but is stable, with performance improving with each point release. The fundamental tools (manipulation of matrices, statistical functions, minimization, *etc.*) exist and are implemented in a way that make extending them fairly easy, plus new packages for more advanced applications are appearing constantly. Figure 1.1 shows Julia running one of the examples from this document. There are also some examples which use Gretl, the Gnu Regression, Econometrics, and Time-Series Library. This is an easy to use program, available in a number of languages, and it comes with a lot of data ready to use. It runs on the major operating systems. Sometimes, simple is better.

The main document was prepared using LyX (www.lyx.org). LyX is a free[1] "what you see is what you mean" word processor, basically working as a graphical frontend to LaTeX. It (with help from other applications) can export your work in LaTeX, HTML, PDF and several other forms. It will run on Linux, Windows, and

[1]"Free" is used in the sense of "freedom", but LyX is also free of charge (free as in "free beer").

Figure 1.1: Julia

MacOS systems. Figure 1.2 shows LyX editing this document. . The LyX source for the document is available on the web page.

## 1.3 License

All materials are copyrighted by Michael Creel with the date that appears above, under the MIT license. See the file License.md

## 1.4 Obtaining and using the materials

A video which explains the steps of this section is here. I recommend that you look at the video one time before installing.

The materials are available from a github repository. To run the examples embedded in the document, you need to

- install the Julia language. See the download link on that page.

Figure 1.2: LyX

- and add files of the github repository as a Julia package.

1. download the code:

   (a) download a zip of the repo, and uncompress it in a convenient directory, or

   (b) git clone the repository to the desired location

2. Go to that directory and start Julia `using` `julia --proj`

3. In Julia, the first time you use the files, do `using Pkg; Pkg.instantiate()` This will take some time, as Econometrics relies on a number of other packages.

4. then do `using Econometrics` in Julia to use the package. The first time you do this, it will take a **long** time, maybe 15 minutes or so. *Don't worry*, this is normal. All of the packages that were downloaded are being compiled for the first time. We will be able to make this go *much, much faster* when

we want to use the code, by following the steps explained in Section 2.5, below.

5. To run examples, cd into the relevant subdirectory of Econometrics/Examples, and then just include the script you would like to run.

6. Once this is done, you can use the code at any time by repeating steps 2 and 4.

7. I recommend setting your operating system to open .jl files with your favorite editor.

Please see the web page for links to videos that explain the installation and usage process, and for how to speed things up.

# Chapter 2

# Introduction to Julia

This document uses the Julia programming language for most of the examples. This chapter gives a very bare bones introduction to Julia. There are much better introductory materials from other sources, some of which are noted below.

## 2.1 Why Julia?

- free: free in terms of $$$, and also, source code is free, so you can know exactly what it does, and you can modify it and contribute to it
- multi-platform: runs on all the popular operating systems. For teaching econometrics, this is nice, because all students have equal access to the materials.
- fast: speed of well-written code is close to C or Fortan. Code is relatively easy to write and to read, similar to Python, Matlab or other matrix scripting languages
- reproducible: it is not difficult to control software versions exactly, so that your results can be reproduced exactly.
- the above 4 considerations are essentially necessary for a language for modern science, which requires accessibility, verifiability, and performance

## 2.2 Why not Julia?

- Julia code is compiled before it's run. This means that first calls to functions take a bit of time, as while are compiled. The second call will be much faster. So, interactive use may frustrate a bit, at least until you learn to work around this particularity.
  - one can use a system image to solve this problem. See https://www.youtube.com/watch?v=a-_ZNTBeLCw for a video that shows how to do this for the code that accompanies these notes.
  - this is getting better all the time, as support for pre-compilation improves
  - can be dealt with quite easily by warming up functions with toy usages, which you might include in your startup file.
  - Also, keep your Julia session running, and the things you use often will already be compiled from previous uses. With Linux, you can have

Julia running in a `byobu` or `screen` session, which you can re-connect to whenever you need it, which is amazingly convenient.

- the speed is only needed if your work is computationally demanding. Dividing epsilon by 2 is not very important when epsilon is small. For getting fast results for linear models, you may prefer a more complete and specialized package, e.g., Stata, etc.

- there is a lot of existing code for languages like Matlab, Python, Fortran, etc. You may prefer to use that code, depending on your research.

## 2.3 Resources

- Julia language: http://julialang.org/
- tutorials and resources:
  - Recommended! https://github.com/PaulSoderlind/JuliaTutorial
  - Recommended! https://julia.quantecon.org/intro.html

## 2.4 Installation of Julia and packages

- install Julia stable version from `https://julialang.org/downloads/`.
- Getting started: https://docs.julialang.org/en/v1/manual/getting-started/
- Package manager documentation: https://docs.julialang.org/en/v1/stdlib/Pkg/
- The first thing you will need to do to make full use of this document is to install the examples and the support code. The main commands for packages:
  - from the REPL, press ] to enter package mode.
  - `]?` : help for package mode.
  - `] add` : Add a package. e.g., to add a popular plotting package, do `] add Plots`
  - Recommended packages (amongst many others):

    * Plots: well-established plotting package. Makie.jl is a newer package, also widely used.
    * CSV: this is a more powerful package for working with CSV data. The DelimitedFiles module of basic Julia can also work in many cases.
    * DataFrames: organization of data with names and manipulations
    * Distributions: essential for statistics beyond the basics.
    * Optim: a well-established optimization package. NLopt.jl is a little less easy to use, but is well-tested. Optimization.jl is a newer package that provides a common interface to many optimization packages, including Optim.jl and NLopt.jl.
    * Revise: make development of code easier
    * OhMyREPL: colors and other features for the Julia prompt.
  - Recommended: put `using Revise; using OhMyREPL` in your `~/.julia/conf` file so that they are automatically used when you start Julia. If you

installed the Econometric package (see Section 1.4), you can add `using Econometrics`, too.

## 2.5 Using a system image

A system image is a file that contains precompiled images of functions that one might want to access as quickly as possible. The "time to first plot" (or first X) issue is the most common complaint about Julia of new users. Using a precompiled system image solves this problem. In the Econometrics code, the file MakeSysImage.jlMakeSysImage.jl will create a system image that contains the Econometrics functions, and the packages upon which they depend. Please see this video for instructions on how to make and use an image. Remember that you will need to create a new system image if you update the packages. See `https://www.youtube.com/watch?v=a-_ZNTBeLCw` for a video that shows how to do this for the code that accompanies these notes.

## 2.6 Running Julia and the work flow

There are several ways to use Julia, here is a basic description of some of them:

### REPL and text editor

The REPL (”read-eval-print loop” ), or in more plain parlance, the Julia command prompt, is my main way of working for research. Simple and easy to replicate. On Linux, just open a terminal and type ”julia”. You can run your code in one window, and edit it in another, using your favorite text editor. There are syntax highlighting schemes for many of the popular editors. This is what I use for research.

### Julia for VSCode

If you want something more modern and integrated looking than the REPL and a text editor, check out Julia for VSCode . This combines the editing, command,

and plot windows all on one interface, and may be more what you're used to if coming from Matlab or RStudio, for example. This is what I use when teaching. A video showing how to use VSCode with the Econometrics packageis here. I recommend that my students have a look at the video, to better follow what I'm doing in class.

## Notebook interfaces

This is a nice way to interactively explore relatively simple code.

- IJulia and Jupyter notebooks: `https://github.com/JuliaLang/IJulia.jl`
- Pluto notebooks: `https://github.com/fonsp/Pluto.jl`
- Neptune notebooks. A fork of Pluto that does not automatically update all cells when any changes.

## 2.7 Loading/saving data

The examples that follow in later chapters provide some examples. Relevant commands are:

- using DelimitedFiles; ?readdlm; ?writedlm
- using CSV;
  - ?CSV.read for reading CSV files into with variable names in the first row into a dataframe
  - ?CSV.write
- For more information, see https://github.com/PaulSoderlind/JuliaTutorial/blob/maste

Example data sets to practice on:

CSV with names: `https://github.com/mcreel/Econometrics/blob/master/Examples/Data/card.csv`, and

plain text, space delimited: https://github.com/mcreel/Econometrics/blob/master/Examples/Data/nerlove.data

## 2.8 Exploratory analysis and plotting

Here are a couple of examples of data preparation and basic analysis.

- Using DataFrames and StatPlots for exploratory analysis using the Card returns to education data set: BasicDataAnalysis.jl
- Data preparation and exploration using Oxford-Man realized library data on financial time series: Oxford-Man realized library data: SP500.jl

# Chapter 3

# Introduction: Economic and econometric models

Here's some data: observations on 3 economic variables.

***Draw a data block.***

Let's do some exploratory analysis using Gretl:

- histograms
- correlations
- x-y scatterplots

So, what can we say? Correlations? Yes. Causality? Who knows?

- *What are these variables*? So far, we don't know, so we have no mental model to sort out which variables might be causing others.

- We are missing a theoretical model!
- A theoretical model is a key ingredient to assign causal relationships (which we might subsequently try to test). Without a model (or the ability to do experiments) we can't distinguish correlation from causality.
- It turns out that the variables we're looking at are QUANTITY (q), PRICE (p), and INCOME (m), and the data were generated using this script

Economic theory tells us that the quantity of a good that consumers will purchase (the demand function) is something like:

$$q = d(p, m, z)$$

- $q$ is the quantity demanded
- $p$ is the price of the good
- $m$ is income

- $z$ is a vector of other variables that may affect demand

The supply of the good to the market is the aggregation of the firms' supply functions. The market supply function is something like

$$q = s(p, z)$$

Suppose we have a sample consisting of a number of observations on $q$ $p$ and $m$ at different time periods $t = 1, 2, ..., n$. Supply and demand in each period is

$$q_t = d(p_t, m_t, z_t)$$
$$q_t = s(p_t, z_t)$$

***Draw a theory block.*** **(draw some graphs showing roles of $m$ and $z$)**

This is the basic economic model of supply and demand: $q$ and $p$ are determined in the market equilibrium, given by the intersection of the two curves.

- These two variables are determined jointly by the model, and are the *endogenous variables*. Income ($m$) is not determined by this model, its value is determined independently of $q$ and $p$ by some other process.

- $m$ is an *exogenous variable*. So, $m$ causes $q$, though the demand function. Because $q$ and $p$ are jointly determined, $m$ also causes $p$.

- $p$ and $q$ do not cause $m$, according to this theoretical model. $q$ and $p$ have a joint causal relationship.

- Economic theory can help us to determine the causality relationships between correlated variables. According to theory, income does not affect the supply equation, so when income changes, supply stays the same. You can see in Figure 3.1 that when income increases, the upward movement of demand is tracing out the slope of the supply equation.

Figure 3.1: Price and Quantity, colored by income (blue is low, violet is high)

The model is essentially a theoretical construct up to now:

- We don't know the forms of the functions $s$ and $d$ (unless we studied the script that generated the data)
- Some components of $z_t$ may not be observable. For example, people don't

eat the same lunch every day, and you can't tell what they will order just by looking at them. There are unobservable components to supply and demand, and we can model them as random variables. Suppose we can break $z_t$ into two unobservable components $\varepsilon_{t1}$ and $\epsilon_{t2}$.

- Theory can make some predictions, too. For example, theory tells us that demand functions are homogeneous of degree zero in prices and income. Also, the compensated demand functions have a negative slope with respect to price. But theory gives us *qualitative information,* signs of effects and so forth, but not the actual values in a given economy, not the magnitudes. So, theory by itself has some limitations, just as data by itself has limitations.

An econometric model attempts to **quantify** the relationship more precisely. A step toward an estimable econometric model is to suppose that the model may be written as

$$q_t = \alpha_1 + \alpha_2 p_t + \alpha_3 m_t + \varepsilon_{t1}$$
$$q_t = \beta_1 + \beta_2 p_t + \varepsilon_{t2}$$

1. The functions $s$ and $d$ have been specified to be linear functions
2. The parameters ($\alpha_1$, $\beta_2$, etc.) are constant over time.
3. There is a single unobservable component in each equation, and it is additive.
4. $E(\epsilon_{tj}) = 0,\ j = 1, 2$
5. $E(m_t \epsilon_{tj}) = 0,\ j = 1, 2$
6. $E(p_t \epsilon_{t2}) = 0$

***Draw an assumptions block.*** With assumptions 1-3, we can always write the last two equations, as the errors simply make up the difference between the true demand and supply functions and the assumed forms. But in order for the $\beta$ coefficients to exist in a sense that has *economic meaning*, and in order to be able to use sample data to make reliable inferences about their values, we need to make assumptions like those in 4-6. These are assertions that the errors have mean zero and are uncorrelated with income, and that the second error is uncorrelated with price. Such assertions <u>may or may not be</u> reasonable. Later we will see how such assumptions may be used and/or tested.

*We can now use econometric methods to learn about the parameters.* ***Draw an econometric model block.***

- with assumptions 4 and 6, we could estimate the supply equation by OLS (do it with GRETL: and compare estimate with true parameter value)

- with assumptions 4 and 5, we could estimate the supply equation by instrumental variables (do it).

- All of the above assumptions 1-6 above have **no theoretical basis**, in that the theory of supply and demand doesn't imply these conditions.

- The validity of any econometric results we obtain using an econometric model will be contingent on these additional restrictions being at least approximately correct. If you study the script that generated the data, you will see that assumptions 1-5 are in fact valid for the data set, but assumption 6 is not. This data is an example of a *simultaneous equations* system, and we will see more about that in Chapter 11.

- Because validity of results depends on validity of assumptions, *specification testing* will be needed, to check that the model seems to be reasonable.
- Only when we are convinced that the model is at least approximately correct should we use it for economic analysis, or for possibly rejecting theories

**Exercise 1.** Given that we know the variable names of the above data, estimate the supply equation by two stage least squares, if you know how to. Compare the coefficient estimates with the values that generated the data. Note that the estimates are not bad, and get very close to the true values if you increase the sample size. This is because the model is correctly specified when we use assumptions 1-5, and the 2SLS estimator is consistent in this case.

# Chapter 4

# Ordinary Least Squares

## 4.1 The Linear Model

Consider approximating a variable $y$ using the variables $x_1, x_2, ..., x_k$. We can consider a model that is a linear approximation:

**Linearity**: the model is a linear function of the parameter vector $\beta^0$ :

$$y = \beta_1^0 x_1 + \beta_2^0 x_2 + ... + \beta_k^0 x_k + \epsilon$$

or, using vector notation:

$$y = \mathbf{x}'\beta^0 + \epsilon$$

The dependent variable $y$ is a scalar random variable, $\mathbf{x} = (\ x_1\ \ x_2\ \cdots\ x_k)'$ is a $k$-vector of explanatory variables, and $\beta^0 = (\ \beta_1^0\ \ \beta_2^0\ \cdots\ \beta_k^0)'$ . The superscript "0" in $\beta^0$ means this is the "true value" of the unknown parameter. It will be defined more precisely later, and usually suppressed when it's not necessary for clarity.

Suppose that we want to use data to try to determine the best linear approximation to $y$ using the variables $\mathbf{x}$. The data $\{(y_t, \mathbf{x}_t)\}$ , $t = 1, 2, ..., n$ are obtained

by some form of sampling[1]. An individual observation is

$$y_t = \mathbf{x}_t'\beta + \varepsilon_t$$

The $n$ observations can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{4.1}$$

where $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}'$ is $n \times 1$ and $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix}'$.

Linear models are more general than they might first appear, since one can employ nonlinear transformations of the variables:

$$\varphi_0(z) = \begin{bmatrix} \varphi_1(w) & \varphi_2(w) & \cdots & \varphi_p(w) \end{bmatrix} \beta + \varepsilon$$

where the $\phi_i()$ are known functions. Defining $y = \varphi_0(z)$, $x_1 = \varphi_1(w)$, *etc.* leads

[1]For example, cross-sectional data may be obtained by random sampling. Time series data accumulate historically.

to a model in the form of equation 4.4. For example, the Cobb-Douglas model

$$z = Aw_2^{\beta_2} w_3^{\beta_3} \exp(\varepsilon)$$

can be transformed logarithmically to obtain

$$\ln z = \ln A + \beta_2 \ln w_2 + \beta_3 \ln w_3 + \varepsilon.$$

If we define $y = \ln z$, $\beta_1 = \ln A$, *etc.*, we can put the model in the form needed. The approximation is linear in the parameters, but not necessarily linear in the variables.

## 4.2 Estimation by least squares

Figure 4.1, obtained by running TypicalData.jl shows some data that follows the linear model $y_t = \beta_1 + \beta_2 x_{t2} + \epsilon_t$. The blue line is the "true" regression line $\beta_1 + \beta_2 x_{t2}$, and the orange dots are the data points $(x_{t2}, y_t)$, where $\epsilon_t$ is a random

Figure 4.1: Typical data, Classical Model

error that has mean zero and is independent of $x_{t2}$. Exactly how the blue line is defined will become clear later. In practice, we only have the data, and we don't know where the blue line lies. We need to gain information about the straight line that best fits the data points.

The *ordinary least squares* (OLS) estimator is defined as the value that minimizes the sum of the squared errors:

$$\hat{\beta} \;=\; \arg\min s(\beta)$$

where

$$\begin{aligned} s(\beta) \;&=\; \sum_{t=1}^{n} \left(y_t - \mathbf{x}_t'\beta\right)^2 \\ &=\; (\mathbf{y} - \mathbf{X}\beta)'\,(\mathbf{y} - \mathbf{X}\beta) \\ &=\; \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &=\; \|\, \mathbf{y} - \mathbf{X}\beta \,\|^2 \end{aligned} \tag{4.2}$$

This last expression makes it clear how the OLS estimator is defined: it minimizes the Euclidean distance between $y$ and $X\beta$. The fitted OLS coefficients are those that give the best linear approximation to $y$ using $\mathbf{x}$ as basis functions, where

"best" means minimum Euclidean distance. One could think of other estimators based upon other metrics. For example, the *minimum absolute distance* (MAD) minimizes $\Sigma_{t=1}^{n} |y_t - \mathbf{x}_t'\beta|$. Later, we will see that which estimator is best in terms of their statistical properties, rather than in terms of the metrics that define them, depends upon the properties of $\epsilon$, about which we have as yet made no assumptions.

- To minimize the criterion $s(\beta)$, find the derivative with respect to $\beta$:

$$D_\beta s(\beta) \;=\; -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta$$

Then setting it to zeros gives

$$D_\beta s(\hat{\beta}) = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \equiv 0$$

so

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- To verify that this is a minimum, check the second order sufficient condition:

$$D^2_\beta s(\hat{\beta}) = 2\mathbf{X}'\mathbf{X}$$

  Since $\rho(\mathbf{X}) = K$, this matrix is positive definite, since it's a quadratic form in a p.d. matrix (identity matrix of order $n$), so $\hat{\beta}$ is in fact a minimizer.

- The *fitted values* are the vector $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$.

- The *residuals* are the vector $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$

- Note that

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \varepsilon \\ &= \mathbf{X}\hat{\beta} + \hat{\varepsilon} \end{aligned}$$

- Also, the first order conditions can be written as

$$\begin{aligned}
\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\beta} &= 0 \\
\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) &= 0 \\
\mathbf{X}'\hat{\varepsilon} &= 0
\end{aligned}$$

  which is to say, the OLS residuals are orthogonal to $\mathbf{X}$. Let's look at this more carefully.

## 4.3 Geometric interpretation of least squares estimation

### In $X, Y$ Space

Figure 4.2 shows a typical fit to data, along with the true regression line. Note that the true line and the estimated line are different. This figure was created

Figure 4.2: Example OLS Fit

by running the Julia program OlsFit.jl . You can experiment with changing the parameter values to see how this affects the fit, and to see how the fitted line will sometimes be close to the true line, and sometimes rather far away.

## In Observation Space

If we want to plot in observation space, we'll need to use only two or three observations, or we'll encounter some limitations of the blackboard. If we try to use 3, we'll encounter the limits of my artistic ability, so let's use two. With only two observations, we can't have $K > 1$.

- We can decompose $y$ into two components: the orthogonal projection onto the $K-$dimensional space spanned by $X$, $X\hat{\beta}$, and the component that is the orthogonal projection onto the $n - K$ subpace that is orthogonal to the span of $X$, $\hat{\varepsilon}$.

- Since $\hat{\beta}$ is chosen to make $\hat{\varepsilon}$ as short as possible, $\hat{\varepsilon}$ will be orthogonal to the space spanned by $X$. Since $X$ is in this space, $X'\hat{\varepsilon} = 0$. Note that the f.o.c. that define the least squares estimator imply that this is so.

Figure 4.3: The fit in observation space

# Projection Matrices

$X\hat{\beta}$ is the projection of $y$ onto the span of $X$, or

$$X\hat{\beta} = X\left(X'X\right)^{-1}X'y$$

Therefore, the matrix that projects $y$ onto the span of $X$ is

$$P_X = X(X'X)^{-1}X'$$

since

$$X\hat{\beta} = P_X y.$$

$\hat{\varepsilon}$ is the projection of $y$ onto the $N - K$ dimensional space that is orthogonal to the span of $X$. We have that

$$\begin{aligned}
\hat{\varepsilon} &= y - X\hat{\beta} \\
&= y - X(X'X)^{-1}X'y \\
&= \left[I_n - X(X'X)^{-1}X'\right] y.
\end{aligned}$$

So the matrix that projects $y$ onto the space orthogonal to the span of $X$ is

$$\begin{aligned}
M_X &= I_n - X(X'X)^{-1}X' \\
&= I_n - P_X.
\end{aligned}$$

We have

$$\hat{\varepsilon} = M_X y.$$

Therefore

$$\begin{aligned} y &= P_X y + M_X y \\ &= X\hat{\beta} + \hat{\varepsilon}. \end{aligned}$$

These two projection matrices decompose the $n$ dimensional vector $y$ into two orthogonal components - the portion that lies in the $K$ dimensional space defined by $X$, and the portion that lies in the orthogonal $n - K$ dimensional space.

- Note that both $P_X$ and $M_X$ are *symmetric* and *idempotent*.
  - A symmetric matrix $A$ is one such that $A = A'$.
  - An idempotent matrix $A$ is one such that $A = AA$.
  - The only nonsingular idempotent matrix is the identity matrix.

## 4.4 Influential observations and outliers

The OLS estimator of the $i^{th}$ element of the vector $\beta_0$ is simply

$$\begin{aligned}\hat{\beta}_i &= \left[(X'X)^{-1}X'\right]_{i\cdot} y \\ &= c_i' y\end{aligned}$$

This is how we define a linear estimator - it's a linear function of the dependent variable. Since it's a linear combination of the observations on the dependent variable, where the weights are determined by the observations on the regressors, some observations may have more influence than others.

To investigate this, let $e_t$ be an $n$ vector of zeros with a 1 in the t$^{th}$ position, *i.e.,* it's the $t$th column of the matrix $I_n$. Define

$$\begin{aligned}h_t &= (P_X)_{tt} \\ &= e_t' P_X e_t\end{aligned}$$

so $h_t$ is the t$^{th}$ element on the main diagonal of $P_X$. Note that

$$h_t \;=\; \| \; P_X e_t \; \|^2$$

so

$$h_t \leq \| \; e_t \; \|^2 = 1$$

So $0 < h_t < 1$. Also,

$$Tr P_X = K \Rightarrow \overline{h} = K/n.$$

So the average of the $h_t$ is $K/n$. The value $h_t$ is referred to as the *leverage* of the observation. If the leverage is much higher than average, the observation has the potential to affect the OLS fit importantly. However, an observation may also be influential due to the value of $y_t$, rather than the weight it is multiplied by, which only depends on the $x_t$'s.

To account for this, consider estimation of $\beta$ without using the $t^{th}$ observation (designate this estimator as $\hat{\beta}^{(t)}$). One can show (see Davidson and MacKinnon,

pp. 32-5 for proof) that

$$\hat{\beta}^{(t)} = \hat{\beta} - \left(\frac{1}{1-h_t}\right)(X'X)^{-1}X_t'\hat{\varepsilon}_t$$

so the change in the $t^{th}$ observations fitted value is

$$\mathbf{x}_t'\hat{\beta} - \mathbf{x}_t'\hat{\beta}^{(t)} = \left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$$

While an observation may be influential if it doesn't affect its own fitted value, it certainly *is* influential if it does. A fast means of identifying influential observations is to plot $\left(\frac{h_t}{1-h_t}\right)\hat{\varepsilon}_t$ (which I will refer to as the *own influence* of the observation) as a function of $t$. Figure 4.4 gives an example plot of data, fit, leverage and influence. The Julia program is InfluentialObservation.jl. If you re-run the program you will see that the leverage of the last observation (an outlying value of x) is always high, and the influence is sometimes high.

After influential observations are detected, one needs to determine *why* they

Figure 4.4: Detection of influential observations

are influential. Possible causes include:

- data entry error, which can easily be corrected once detected. Data entry errors *are very common.*
- special economic factors that affect some observations. These would need to be identified and incorporated in the model. This is the idea behind *structural change*: the parameters may not be constant across all observations.
- pure randomness may have caused us to sample a low-probability observation.

There exist *robust* estimation methods that downweight outliers.

## 4.5 Goodness of fit

The fitted model is

$$y = X\hat{\beta} + \hat{\varepsilon}$$

Take the inner product:

$$y'y = \hat{\beta}'X'X\hat{\beta} + 2\hat{\beta}'X'\hat{\varepsilon} + \hat{\varepsilon}'\hat{\varepsilon}$$

But the middle term of the RHS is zero since $X'\hat{\varepsilon} = 0$, so

$$y'y = \hat{\beta}'X'X\hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon} \tag{4.3}$$

The *uncentered* $R_u^2$ is defined as

$$\begin{aligned} R_u^2 &= 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} \\ &= \frac{\hat{\beta}'X'X\hat{\beta}}{y'y} \\ &= \frac{\| P_X y \|^2}{\| y \|^2} \\ &= \cos^2(\phi), \end{aligned}$$

where $\phi$ is the angle between $y$ and the span of $X$ .

- The uncentered $R^2$ changes if we add a constant to $y$, since this changes $\phi$ (see Figure 4.5, the yellow vector is a constant, since it's on the 45 degree line in observation space). Another, more common definition measures the

Figure 4.5: Uncentered $R^2$

contribution of the variables, other than the constant term, to explaining the variation in $y$. Thus it measures the ability of the model to explain the variation of $y$ about its unconditional sample mean.

Let $\iota = (1, 1, ..., 1)'$, a $n$ -vector. So

$$\begin{aligned} M_\iota &= I_n - \iota(\iota'\iota)^{-1}\iota' \\ &= I_n - \iota\iota'/n \end{aligned}$$

$M_\iota y$ just returns the vector of deviations from the mean. In terms of deviations from the mean, equation 4.3 becomes

$$y'M_\iota y = \hat{\beta}'X'M_\iota X\hat{\beta} + \hat{\varepsilon}'M_\iota\hat{\varepsilon}$$

The *centered* $R_c^2$ is defined as

$$R_c^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'M_\iota y} = 1 - \frac{ESS}{TSS}$$

where $ESS = \hat{\varepsilon}'\hat{\varepsilon}$ and $TSS = y'M_\iota y = \Sigma_{t=1}^{n}(y_t - \bar{y})^2$.

Supposing that $X$ contains a column of ones (*i.e.*, there is a constant term),

$$X'\hat{\varepsilon} = 0 \Rightarrow \sum_t \hat{\varepsilon}_t = 0$$

so $M_\iota \hat{\varepsilon} = \hat{\varepsilon}$. In this case

$$y'M_\iota y = \hat{\beta}'X'M_\iota X\hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon}$$

So

$$R_c^2 = \frac{RSS}{TSS}$$

where $RSS = \hat{\beta}'X'M_\iota X\hat{\beta}$

- Supposing that a column of ones is in the space spanned by $X$ ($P_X\iota = \iota$), then one can show that $0 \leq R_c^2 \leq 1$.

## 4.6 The classical linear regression model

Up to this point the model is empty of content beyond the definition of a best linear approximation to $y$ and some geometrical properties. There is no economic content to the model, and the regression parameters have no economic interpretation. For example, what is the partial derivative of $y$ with respect to $x_j$? The linear approximation is

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

The partial derivative is

$$\frac{\partial y}{\partial x_j} = \beta_j + \frac{\partial \epsilon}{\partial x_j}$$

Up to now, there's no guarantee that $\frac{\partial \epsilon}{\partial x_j}$=0. For the $\beta$ to have an economic meaning, we need to make additional assumptions. The assumptions that are appropriate to make depend on the data under consideration. We'll start with the classical linear regression model, which incorporates some assumptions that are clearly not realistic for economic data. This is to be able to explain some concepts

with a minimum of confusion and notational clutter. Later we'll adapt the results to what we can get with more realistic assumptions.

**Linearity**: the model is a linear function of the parameter vector $\beta^0$ :

$$y = \beta_1^0 x_1 + \beta_2^0 x_2 + ... + \beta_k^0 x_k + \epsilon \tag{4.4}$$

or, using vector notation:

$$y = \mathbf{x}'\beta^0 + \epsilon$$

**Nonstochastic linearly independent regressors**: $\mathbf{X}$ is a fixed matrix of constants, it has rank $K$ equal to its number of columns, and

$$\lim \frac{1}{n}\mathbf{X}'\mathbf{X} = Q_X \tag{4.5}$$

where $Q_X$ is a finite positive definite matrix. This is needed to be able to identify the individual effects of the explanatory variables.

**Independently and identically distributed errors**:

$$\epsilon \sim IID(0, \sigma^2 I_n) \tag{4.6}$$

$\varepsilon$ is jointly distributed IID. This implies the following two properties:

**Homoscedastic errors**:

$$V(\varepsilon_t) = \sigma_0^2, \forall t \tag{4.7}$$

**Nonautocorrelated errors:**

$$\mathcal{E}(\varepsilon_t \epsilon_s) = 0, \forall t \neq s \tag{4.8}$$

Optionally, we will sometimes assume that the errors are normally distributed.

**Normally distributed errors:**

$$\epsilon \sim N(0, \sigma^2 I_n) \tag{4.9}$$

## 4.7 Small sample statistical properties of the least squares estimator

Up to now, we have only examined numeric properties of the OLS estimator, that always hold. Now we will examine statistical properties. The statistical properties depend upon the assumptions we make.

### Unbiasedness

We have $\hat{\beta} = (X'X)^{-1}X'y$. By linearity,

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon\end{aligned}$$

By 4.5 and 4.6

$$\begin{aligned} E(X'X)^{-1}X'\varepsilon &= E(X'X)^{-1}X'\varepsilon \\ &= (X'X)^{-1}X'E\varepsilon \\ &= 0 \end{aligned}$$

so the OLS estimator is unbiased under the assumptions of the classical model.

Figure 4.6 shows the results of a small Monte Carlo experiment where the OLS estimator was calculated for 10000 samples from the classical model with $y = 1 + 2x + \varepsilon$, where $n = 20$, $\sigma_\varepsilon^2 = 9$, and $x$ is fixed across samples. We can see that the $\beta_2$ appears to be estimated without bias. The program that generates the plot is Unbiased.jl , if you would like to experiment with this.

With time series data, the OLS estimator will often be biased. Figure 4.7 shows the results of a small Monte Carlo experiment where the OLS estimator was calculated for 1000 samples from the AR(1) model with $y_t = 0 + 0.9y_{t-1} + \varepsilon_t$, where $n = 20$ and $\sigma_\varepsilon^2 = 1$. In this case, assumption 4.5 does not hold: the

Figure 4.6: Unbiasedness of OLS under classical assumptions: replications of $\hat{\beta}$ minus true $\beta$

regressors are stochastic. We can see that the bias in the estimation of $\beta_2$ is about -0.2.

The program that generates the plot is Biased.jl , if you would like to experiment with this.

## Normality

With the linearity assumption, we have $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$. This is a linear function of $\varepsilon$. Adding the assumption of normality (4.9, which implies strong exogeneity), then

$$\hat{\beta} \sim N\left(\beta, (X'X)^{-1}\sigma_0^2\right)$$

since a linear function of a normal random vector is also normally distributed. In Figure 4.6 you can see that the estimator appears to be normally distributed. It in fact is normally distributed, since the DGP (see the Octave program) has normal

Figure 4.7: Biasedness of OLS when an assumption fails: : replications of $\hat{\beta}$ minus true $\beta$

errors. Even when the data may be taken to be IID, the assumption of normality is often questionable or simply untenable. For example, if the dependent variable is the number of automobile trips per week, it is a count variable with a discrete distribution, and is thus not normally distributed. Many variables in economics can take on only nonnegative values, which, strictly speaking, rules out normality.[2]

[2]Normality may be a good model nonetheless, as long as the probability of a negative value occurring is negligible under the model. This depends upon the mean being large enough in relation to the variance.

## The variance of the OLS estimator and the Gauss-Markov theorem

Now let's make all the classical assumptions except the assumption of normality. We have $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$ and we know that $E(\hat{\beta}) = \beta$. So

$$\begin{aligned} Var(\hat{\beta}) &= E\left\{\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right\} \\ &= E\left\{(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right\} \\ &= (X'X)^{-1}\sigma_0^2 \end{aligned}$$

The OLS estimator is a *linear estimator*, which means that it is a linear function of the dependent variable, $y$.

$$\begin{aligned} \hat{\beta} &= \left[(X'X)^{-1}X'\right]y \\ &= Cy \end{aligned}$$

where $C$ is a function of the explanatory variables only, not the dependent variable.

It is also *unbiased* under the present assumptions, as we proved above. One could consider other weights $W$ that are a function of $X$ that define some other linear estimator. We'll still insist upon unbiasedness. Consider $\tilde{\beta} = Wy$, where $W = W(X)$ is some $k \times n$ matrix function of $X$. Note that since $W$ is a function of $X$, it is nonstochastic, too. If the estimator is unbiased, then we must have $WX = I_K$:

$$\begin{aligned}
\mathcal{E}(Wy) &= \mathcal{E}(WX\beta_0 + W\varepsilon) \\
&= WX\beta_0 \\
&= \beta_0 \\
&\Rightarrow \\
WX &= I_K
\end{aligned}$$

The variance of $\tilde{\beta}$ is

$$V(\tilde{\beta}) = WW'\sigma_0^2.$$

Define

$$D = W - (X'X)^{-1}X'$$

so

$$W = D + (X'X)^{-1}X'$$

Since $WX = I_K$, $DX = 0$, so

$$\begin{aligned} V(\tilde{\beta}) &= \left(D + (X'X)^{-1}X'\right)\left(D + (X'X)^{-1}X'\right)' \sigma_0^2 \\ &= \left(DD' + (X'X)^{-1}\right)\sigma_0^2 \end{aligned}$$

So

$$V(\tilde{\beta}) \geq V(\hat{\beta})$$

The inequality is a shorthand means of expressing, more formally, that $V(\tilde{\beta}) - V(\hat{\beta})$ is a positive semi-definite matrix. This is a proof of the Gauss-Markov Theorem. The OLS estimator is the "best linear unbiased estimator" (BLUE).

- It is worth emphasizing again that we have not used the normality assumption in any way to prove the Gauss-Markov theorem, so it is valid if the errors are not normally distributed, as long as the other assumptions hold.

To illustrate the Gauss-Markov result, consider the estimator that results from splitting the sample into $p$ equally-sized parts, estimating using each part of the data separately by OLS, then averaging the $p$ resulting estimators. You should be able to show that this estimator is unbiased, but inefficient with respect to the OLS estimator. The program Efficiency.jl illustrates this using a small Monte Carlo experiment, which compares the OLS estimator and a 3-way split sample estimator. The data generating process follows the classical model, with $n = 21$. The true parameter value is $\beta = 2$. In Figures 4.8 and 4.9 we can see that the OLS estimator is more efficient, since the tails of its histogram are more narrow.

We have that $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \left(X'X\right)^{-1}\sigma_0^2$, but we still need to estimate the variance of $\epsilon$, $\sigma_0^2$, in order to have an idea of the precision of the

Figure 4.8: Gauss-Markov Result: The OLS estimator

Figure 4.9: Gauss-Markov Resul: The split sample estimator

estimates of $\beta$. A commonly used estimator of $\sigma_0^2$ is

$$\widehat{\sigma_0^2} = \frac{1}{n-K}\hat{\varepsilon}'\hat{\varepsilon}$$

This estimator is unbiased:

$$
\begin{aligned}
\widehat{\sigma_0^2} &= \frac{1}{n-K}\hat{\varepsilon}'\hat{\varepsilon} \\
&= \frac{1}{n-K}\varepsilon' M \varepsilon \\
\mathcal{E}(\widehat{\sigma_0^2}) &= \frac{1}{n-K}E(Tr\varepsilon' M \varepsilon) \\
&= \frac{1}{n-K}E(TrM\varepsilon\varepsilon') \\
&= \frac{1}{n-K}TrE(M\varepsilon\varepsilon') \\
&= \frac{1}{n-K}\sigma_0^2 TrM \\
&= \frac{1}{n-K}\sigma_0^2\,(n-k) \\
&= \sigma_0^2
\end{aligned}
$$

where we use the fact that $Tr(AB) = Tr(BA)$ when both products are conformable. Thus, this estimator is also unbiased under these assumptions.

## 4.8 Example: The Nerlove model

### Theoretical background

For a firm that takes input prices $w$ and the output level $q$ as given, the cost minimization problem is to choose the quantities of inputs $x$ to solve the problem

$$\min_x w'x$$

subject to the restriction

$$f(x) = q.$$

The solution is the vector of factor demands $x(w, q)$. The *cost function* is obtained by substituting the factor demands into the criterion function:

$$Cw, q) = w'x(w, q).$$

- **Monotonicity** Increasing factor prices cannot decrease cost, so

$$\frac{\partial C(w,q)}{\partial w} \geq 0$$

  Remember that these derivatives give the conditional factor demands (Shephard's Lemma).

- **Homogeneity** The cost function is homogeneous of degree 1 in input prices: $C(tw,q) = tC(w,q)$ where $t$ is a scalar constant. This is because the factor demands are homogeneous of degree zero in factor prices - they only depend upon relative prices.

- **Returns to scale** The *returns to scale* parameter $\gamma$ is defined as the inverse of the elasticity of cost with respect to output:

$$\gamma = \left(\frac{\partial C(w,q)}{\partial q}\frac{q}{C(w,q)}\right)^{-1}$$

*Constant returns to scale* is the case where increasing production $q$ implies that cost increases in the proportion 1:1. If this is the case, then $\gamma = 1$.

## Cobb-Douglas functional form

The Cobb-Douglas functional form is linear in the logarithms of the regressors and the dependent variable. For a cost function, if there are $g$ factors, the Cobb-Douglas cost function has the form

$$C = Aw_1^{\beta_1}...w_g^{\beta_g}q^{\beta_q}e^{\varepsilon}$$

What is the elasticity of $C$ with respect to $w_j$?

$$\begin{aligned} e_{w_j}^{C} &= \left(\frac{\partial C}{\partial_{W_J}}\right)\left(\frac{w_j}{C}\right) \\ &= \beta_j A w_1^{\beta_1}.w_j^{\beta_j - 1}..w_g^{\beta_g}q^{\beta_q}e^{\varepsilon}\frac{w_j}{Aw_1^{\beta_1}...w_g^{\beta_g}q^{\beta_q}e^{\varepsilon}} \\ &= \beta_j \end{aligned}$$

This is one of the reasons the Cobb-Douglas form is popular - the coefficients are easy to interpret, since they are the elasticities of the dependent variable with respect to the explanatory variable. Not that in this case,

$$\begin{aligned} e^C_{w_j} &= \left(\frac{\partial C}{\partial_{W_J}}\right)\left(\frac{w_j}{C}\right) \\ &= x_j(w,q)\frac{w_j}{C} \\ &\equiv s_j(w,q) \end{aligned}$$

the *cost share* of the $j^{th}$ input. So with a Cobb-Douglas cost function, $\beta_j = s_j(w,q)$. The cost shares are constants.

Note that after a logarithmic transformation we obtain

$$\ln C = \alpha + \beta_1 \ln w_1 + ... + \beta_g \ln w_g + \beta_q \ln q + \epsilon$$

where $\alpha = \ln A$ . So we see that the transformed model is linear in the logs of the data.

One can verify that the property of HOD1 implies that

$$\sum_{i=1}^{g} \beta_i = 1$$

In other words, the cost shares add up to 1.

The hypothesis that the technology exhibits CRTS implies that

$$\gamma = \frac{1}{\beta_q} = 1$$

so $\beta_q = 1$. Likewise, monotonicity implies that the coefficients $\beta_i \geq 0, i = 1, ..., g$.

## The Nerlove data and OLS

The file nerlove.data contains data on 145 electric utility companies' cost of production, output and input prices. The data are for the U.S., and were collected by M. Nerlove. The observations are by row, and the columns are **COMPANY, COST** ($C$)**, OUTPUT** ($Q$), **PRICE OF LABOR** ($P_L$), **PRICE OF**

**FUEL** ($P_F$) and **PRICE OF CAPITAL** ($P_K$). Note that the data are sorted by output level (the third column).

We will estimate the Cobb-Douglas model

$$\ln C = \beta_1 + \beta_Q \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon \tag{4.10}$$

by OLS, using the Julia script Nerlove.jl , which uses ols.jl .

The results are

```
julia> include("Nerlove.jl")

OLS estimation, 145 observations
R²: 0.925955   σ²: 0.153943
```

| parameter | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -3.52650 | 1.77437 | -1.98747 | 0.04882 |
| output | 0.72039 | 0.01747 | 41.24448 | 0.00000 |
| labor | 0.43634 | 0.29105 | 1.49921 | 0.13607 |
| fuel | 0.42652 | 0.10037 | 4.24948 | 0.00004 |
| capital | -0.21989 | 0.33943 | -0.64782 | 0.51816 |

- Do the theoretical restrictions hold?

- Does the model fit well?
- What do you think about RTS?

We will most often use Julia programs that more or less directly implement the theory we learn in examples in this document. This is because following such transparent programming statements is a useful way of learning how theory is put into practice. Nevertheless, you may be interested in a more "user-friendly" environment for doing econometrics, especially after you have mastered the theory. Julia itself offers packages such as DataFrames.jl and GLM.jl which will allow you to avoid some of the nuts and bolts of econometric modeling. For example, see NerloveDF.jl for estimating the Nerlove model using these packages. If you run that, you will see that the estimated standard errors differ from what Nerlove.jl reports, we will get to the reason for that later. For a "canned" package, apart from what Julia offers, I heartily recommend Gretl, the GNU Regression, Econometrics, and Time-Series Library. Gretl is free software. This is an easy to use program, available in English, French, and Spanish, and it comes with a lot of data ready

to use. It even has an option to save output as LaTeX fragments, so that I can just include the results into this document, no muss, no fuss. Here is the Nerlove data in the form of a GRETL data set: nerlove.gdt . Here the results of the Nerlove model from GRETL:

Fortunately, Gretl and my OLS program agree upon the results. I recommend using GRETL to repeat the examples that are done using Julia.

The previous properties hold for finite sample sizes. Before considering the asymptotic properties of the OLS estimator it is useful to review the MLE estimator, since under the assumption of normal errors the two estimators coincide.

## 4.9 Exercises

1. Prove that the split sample estimator used to generate figure 4.9 is unbiased.

2. Calculate the OLS estimates of the Nerlove model using Julia and GRETL, and provide printouts of the results. Interpret the results.

3. Do an analysis of whether or not there are influential observations for OLS estimation of the Nerlove model. Discuss.

4. Using GRETL, examine the residuals after OLS estimation and tell me whether or not you believe that the assumption of independent identically distributed normal errors is warranted. No need to do formal tests, just look at the plots. Print out any that you think are relevant, and interpret them.

5. For a random vector $X \sim N(\mu_x, \Sigma)$, what is the distribution of $AX + b$, where $A$ and $b$ are conformable matrices of constants?

6. Using Julia, write a little program that verifies that $Tr(AB) = Tr(BA)$ for $A$ and $B$ 4x4 matrices of random numbers. Note: there is a Julia function `trace()`.

7. For the model with a constant and a single regressor, $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$, which satisfies the classical assumptions, prove that the variance of the OLS estimator declines to zero as the sample size increases.

# Chapter 5

# Asymptotic properties of the least squares estimator

The OLS estimator under the classical assumptions is BLUE[1], for all sample sizes. Now let's see what happens when the sample size tends to infinity.

[1] BLUE $\equiv$ best linear unbiased estimator if I haven't defined it before

## 5.1 Consistency

$$
\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + \varepsilon) \\
&= \beta_0 + (X'X)^{-1}X'\varepsilon \\
&= \beta_0 + \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}
\end{aligned}
$$

Consider the last two terms. By assumption $\lim_{n\to\infty}\left(\frac{X'X}{n}\right) = Q_X \Rightarrow \lim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1} = Q_X^{-1}$, since the inverse of a nonsingular matrix is a continuous function of the elements of the matrix. Considering $\frac{X'\varepsilon}{n}$,

$$
\frac{X'\varepsilon}{n} = \frac{1}{n}\sum_{t=1}^{n} x_t\varepsilon_t
$$

Each $x_t\varepsilon_t$ has expectation zero, so

$$E\left(\frac{X'\varepsilon}{n}\right) = 0$$

The variance of each term is

$$V\left(x_t\epsilon_t\right) \;=\; x_t x_t'\sigma^2.$$

As long as these are finite, and given a technical condition[2], the Kolmogorov SLLN applies, so

$$\frac{1}{n}\sum_{t=1}^{n} x_t\varepsilon_t \overset{a.s.}{\to} 0.$$

This implies that

$$\hat{\beta} \overset{a.s.}{\to} \beta_0.$$

[2]For application of LLN's and CLT's, of which there are very many to choose from, I'm going to avoid the technicalities. Basically, as long as terms that make up an average have finite variances and are not too strongly dependent, one will be able to find a LLN or CLT to apply. Which one it is doesn't matter, we only need the result. When working with particular models, it will be more relevant to consider which particular theorems will apply.

This is the property of *strong consistency:* the estimator converges in almost surely to the true value.

- The consistency proof does not use the normality assumption.
- Remember that almost sure convergence implies convergence in probability.

## 5.2 Asymptotic normality

We've seen that the OLS estimator is normally distributed *under the assumption of normal errors.* If the error distribution is unknown, we of course don't know the distribution of the estimator. However, we can get asymptotic results. *Assuming the distribution of* $\varepsilon$ is unknown, but the the other classical assumptions hold:

$$
\begin{aligned}
\hat{\beta} &= \beta_0 + (X'X)^{-1}X'\varepsilon \\
\hat{\beta} - \beta_0 &= (X'X)^{-1}X'\varepsilon \\
\sqrt{n}\left(\hat{\beta} - \beta_0\right) &= \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{\sqrt{n}}
\end{aligned}
$$

- Now as before, $\left(\frac{X'X}{n}\right)^{-1} \to Q_X^{-1}$.
- Considering $\frac{X'\varepsilon}{\sqrt{n}}$, the limit of the variance is

$$
\begin{aligned}
\lim_{n\to\infty} V\left(\frac{X'\varepsilon}{\sqrt{n}}\right) &= \lim_{n\to\infty} E\left(\frac{X'\epsilon\epsilon'X}{n}\right) \\
&= \sigma_0^2 Q_X
\end{aligned}
$$

The mean is of course zero. To get asymptotic normality, we need to apply a CLT. The best known CLTs are for averages of IID terms, but CLTs exist for averages of dependent, non-identically distributed terms, too. The basic

requirement is that variances of the terms in the average must not explode, and the terms in the average can not be too highly dependent. Without getting into the technical details, which are appropriate to discuss when working with some particular type of data, we assume one holds, so

$$\frac{X'\varepsilon}{\sqrt{n}} \xrightarrow{d} N\left(0, \sigma_0^2 Q_X\right)$$

Therefore,

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \sigma_0^2 Q_X^{-1}\right) \tag{5.1}$$

- In summary, the OLS estimator is normally distributed in small and large samples if $\varepsilon$ is normally distributed. If $\varepsilon$ is not normally distributed, $\hat{\beta}$ is asymptotically normally distributed when a CLT can be applied.

## 5.3 Asymptotic efficiency

The least squares objective function is

$$s(\beta) \;=\; \sum_{t=1}^{n} \left(y_t - x_t'\beta\right)^2$$

Supposing that $\varepsilon$ is normally distributed, the model is

$$y = X\beta_0 + \varepsilon,$$

$$\begin{aligned} \varepsilon \;&\sim\; N(0, \sigma_0^2 I_n), \text{ so} \\ f(\varepsilon) \;&=\; \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right) \end{aligned}$$

The joint density for $y$ can be constructed using a change of variables. We have $\varepsilon = y - X\beta$, so $\frac{\partial \varepsilon}{\partial y'} = I_n$ and $|\frac{\partial \varepsilon}{\partial y'}| = 1$, so

$$f(y) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - x_t'\beta)^2}{2\sigma^2}\right).$$

Taking logs,

$$\ln L(\beta, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \sum_{t=1}^{n} \frac{(y_t - x_t'\beta)^2}{2\sigma^2}.$$

Maximizing this function with respect to $\beta$ and $\sigma$ gives what is known as the maximum likelihood (ML) estimator. It turns out that ML estimators are asymptotically efficient, a concept that will be explained in detail later. It's clear that the first order conditions for the MLE of $\beta_0$ are the same as the first order conditions that define the OLS estimator (up to multiplication by a constant), so the OLS estimator of $\beta$ is also the ML estimator. *The estimators are the same, under the present assumptions.* Therefore, their properties are the same. *In partic-*

*ular, under the classical assumptions with normality, the OLS estimator $\hat{\beta}$ is asymptotically efficient.* Note that one needs to make an assumption about the distribution of the errors to compute the ML estimator. If the errors had a distribution other than the normal, then the OLS estimator and the ML estimator would not coincide.

As we'll see later, it will be possible to use (iterated) linear estimation methods and still achieve asymptotic efficiency even if the assumption that $Var(\varepsilon) \neq \sigma^2 I_n$, as long as $\varepsilon$ is still normally distributed. This is **not** the case if $\varepsilon$ is nonnormal. In general with nonnormal errors it will be necessary to use nonlinear estimation methods to achieve asymptotically efficient estimation.

## 5.4 Exercises

1. Write an Octave program that generates a histogram for $R$ Monte Carlo replications of $\sqrt{n}\left(\hat{\beta}_j - \beta_j\right)$, where $\hat{\beta}$ is the OLS estimator and $\beta_j$ is one

of the $k$ slope parameters. $R$ should be a large number, at least 1000. The model used to generate data should follow the classical assumptions, except that the errors should not be normally distributed (try $U(-a, a)$, $t(p)$, $\chi^2(p) - p$, etc). Generate histograms for $n \in \{20, 50, 100, 1000\}$. Do you observe evidence of asymptotic normality? Comment.

2. Consider the following regression model:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

where $E(x_{1i}) = 0, E(x_{2i}) = 0, Var(x_{1i}) = \sigma_1^2, Var(x_{2i}) = \sigma_2^2$ and $Cov(x_{1i}, x_{2i}) = \sigma_{12}$. The model satisfies the basic OLS assumptions. However, a loose econometrician estimates the following model:

$$y_i = \beta_1 x_{1i} + v_i$$

where $\tilde{\beta}_1$ is the estimator of the parameter $\beta_1$.

a) Compute $plim(\tilde{\beta}_1)$.

b) Compute the asymptotic bias of $\tilde{\beta}_1$ (i.e. $plim(\tilde{\beta}_1) - \beta_1$).

c) Under which conditions is $\tilde{\beta}_1$ a consistent estimate of $\beta_1$?

# Chapter 6

# Restrictions and hypothesis tests

## 6.1 Exact linear restrictions

In many cases, economic theory suggests restrictions on the parameters of a model. For example, a demand function is supposed to be homogeneous of degree zero in

prices and income. If we have a Cobb-Douglas (log-linear) model,

$$\ln q = \beta_0 + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m + \varepsilon,$$

then we need that

$$k^0 \ln q = \beta_0 + \beta_1 \ln kp_1 + \beta_2 \ln kp_2 + \beta_3 \ln km + \varepsilon,$$

so

$$\begin{aligned} \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m &= \beta_1 \ln kp_1 + \beta_2 \ln kp_2 + \beta_3 \ln km \\ &= (\ln k)(\beta_1 + \beta_2 + \beta_3) + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m. \end{aligned}$$

The only way to guarantee this for arbitrary $k$ is to set

$$\beta_1 + \beta_2 + \beta_3 = 0,$$

which is a *parameter restriction.* In particular, this is a linear equality restriction, which is probably the most commonly encountered case.

## Imposition

The general formulation of linear equality restrictions is the model

$$\begin{aligned} y &= X\beta + \varepsilon \\ R\beta &= r \end{aligned}$$

where $R$ is a $Q \times K$ matrix, $Q < K$ and $r$ is a $Q \times 1$ vector of constants.

- We assume $R$ is of rank $Q$, so that there are no redundant restrictions.
- We also assume that $\exists \beta$ that satisfies the restrictions: they aren't infeasible.

Let's consider how to estimate $\beta$ subject to the restrictions $R\beta = r$. The most obvious approach is to set up the Lagrangean

$$\min_{\beta,\lambda} s(\beta,\lambda) = \frac{1}{n}(y - X\beta)'(y - X\beta) + 2\lambda'(R\beta - r).$$

The Lagrange multipliers are scaled by 2, which makes things less messy. The fonc are

$$\begin{aligned} D_\beta s(\hat{\beta},\hat{\lambda}) &= -2X'y + 2X'X\hat{\beta}_R + 2R'\hat{\lambda} \equiv 0 \\ D_\lambda s(\hat{\beta},\hat{\lambda}) &= R\hat{\beta}_R - r \equiv 0, \end{aligned}$$

which can be written as

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'y \\ r \end{bmatrix}.$$

We get

$$\begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ r \end{bmatrix}.$$

Maybe you're curious about how to invert a partitioned matrix? I can help you with that:

Note that

$$\begin{aligned} \begin{bmatrix} (X'X)^{-1} & 0 \\ -R\,(X'X)^{-1} & I_Q \end{bmatrix} \begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} &\equiv AB \\ &= \begin{bmatrix} I_K & (X'X)^{-1}\,R' \\ 0 & -R\,(X'X)^{-1}\,R' \end{bmatrix} \\ &\equiv \begin{bmatrix} I_K & (X'X)^{-1}\,R' \\ 0 & -P \end{bmatrix} \\ &\equiv C, \end{aligned}$$

and

$$\begin{aligned}\begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix}\begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -P \end{bmatrix} &\equiv DC \\ &= I_{K+Q},\end{aligned}$$

so

$$\begin{aligned} DAB &= I_{K+Q} \\ DA &= B^{-1} \\ B^{-1} &= \begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix}\begin{bmatrix} (X'X)^{-1} & 0 \\ -R\,(X'X)^{-1} & I_Q \end{bmatrix} \\ &= \begin{bmatrix} (X'X)^{-1} - (X'X)^{-1}R'P^{-1}R\,(X'X)^{-1} & (X'X)^{-1}R'P^{-1} \\ P^{-1}R\,(X'X)^{-1} & -P^{-1} \end{bmatrix},\end{aligned}$$

If you weren't curious about that, please start paying attention again. Also, note

that we have made the definition $P = R\left(X'X\right)^{-1} R')$

$$
\begin{aligned}
\begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} &= \begin{bmatrix} \left(X'X\right)^{-1} - \left(X'X\right)^{-1} R' P^{-1} R \left(X'X\right)^{-1} & \left(X'X\right)^{-1} R' P^{-1} \\ P^{-1} R \left(X'X\right)^{-1} & -P^{-1} \end{bmatrix} \begin{bmatrix} X'y \\ r \end{bmatrix} \\
&= \begin{bmatrix} \hat{\beta} - \left(X'X\right)^{-1} R' P^{-1} \left(R\hat{\beta} - r\right) \\ P^{-1} \left(R\hat{\beta} - r\right) \end{bmatrix} \\
&= \begin{bmatrix} \left(I_K - \left(X'X\right)^{-1} R' P^{-1} R\right) \\ P^{-1} R \end{bmatrix} \hat{\beta} + \begin{bmatrix} \left(X'X\right)^{-1} R' P^{-1} r \\ -P^{-1} r \end{bmatrix}
\end{aligned}
$$

The fact that $\hat{\beta}_R$ and $\hat{\lambda}$ are linear functions of $\hat{\beta}$ makes it easy to determine their distributions, since the distribution of $\hat{\beta}$ is already known. Recall that for $x$ a random vector, and for $A$ and $b$ a matrix and vector of constants, respectively, $Var\left(Ax + b\right) = AVar(x)A'$.

Though this is the obvious way to go about finding the restricted estimator, an easier way, if the number of restrictions is small, is to impose them by substitution.

Write

$$
\begin{aligned}
y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\
\begin{bmatrix} R_1 & R_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} &= r
\end{aligned}
$$

where $R_1$ is $Q \times Q$ nonsingular. Supposing the $Q$ restrictions are linearly independent, one can always make $R_1$ nonsingular by reorganizing the columns of $X$. Then

$$\beta_1 = R_1^{-1}r - R_1^{-1}R_2\beta_2.$$

Substitute this into the model

$$
\begin{aligned}
y &= X_1R_1^{-1}r - X_1R_1^{-1}R_2\beta_2 + X_2\beta_2 + \varepsilon \\
y - X_1R_1^{-1}r &= \left[X_2 - X_1R_1^{-1}R_2\right]\beta_2 + \varepsilon
\end{aligned}
$$

or with the appropriate definitions,

$$y_R = X_R \beta_2 + \varepsilon.$$

This model satisfies the classical assumptions, *supposing the restriction is true.* One can estimate by OLS. The variance of $\hat{\beta}_2$ is as before

$$V(\hat{\beta}_2) = (X_R' X_R)^{-1} \sigma_0^2$$

and the estimator is

$$\hat{V}(\hat{\beta}_2) = (X_R' X_R)^{-1} \hat{\sigma}^2$$

where one estimates $\sigma_0^2$ in the normal way, using the restricted model, *i.e.,*

$$\widehat{\sigma_0^2} = \frac{\left(y_R - X_R \widehat{\beta_2}\right)' \left(y_R - X_R \widehat{\beta_2}\right)}{n - (K - Q)}$$

To recover $\hat{\beta}_1$, use the restriction. To find the variance of $\hat{\beta}_1$, use the fact that it

is a linear function of $\hat{\beta}_2$, so

$$\begin{aligned} V(\hat{\beta}_1) &= R_1^{-1} R_2 V(\hat{\beta}_2) R_2' \left(R_1^{-1}\right)' \\ &= R_1^{-1} R_2 \left(X_2' X_2\right)^{-1} R_2' \left(R_1^{-1}\right)' \sigma_0^2 \end{aligned}$$

## Properties of the restricted estimator

We have that

$$\begin{aligned} \hat{\beta}_R &= \hat{\beta} - (X'X)^{-1} R' P^{-1} \left(R\hat{\beta} - r\right) \\ &= \hat{\beta} + (X'X)^{-1} R' P^{-1} r - (X'X)^{-1} R' P^{-1} R (X'X)^{-1} X' y \\ &= \beta + (X'X)^{-1} X'\varepsilon + (X'X)^{-1} R' P^{-1} \left[r - R\beta\right] - (X'X)^{-1} R' P^{-1} R (X'X)^{-1} X'\varepsilon \\ \hat{\beta}_R - \beta &= (X'X)^{-1} X'\varepsilon \\ &+ (X'X)^{-1} R' P^{-1} \left[r - R\beta\right] \\ &- (X'X)^{-1} R' P^{-1} R (X'X)^{-1} X'\varepsilon \end{aligned}$$

Mean squared error is

$$MSE(\hat{\beta}_R) = \mathcal{E}(\hat{\beta}_R - \beta)(\hat{\beta}_R - \beta)'$$

Noting that the crosses between the second term and the other terms expect to zero, and that the cross of the first and third has a cancellation with the square of the third, we obtain

$$\begin{aligned} MSE(\hat{\beta}_R) &= (X'X)^{-1}\sigma^2 \\ &+ (X'X)^{-1}R'P^{-1}\left[r - R\beta\right]\left[r - R\beta\right]' P^{-1}R(X'X)^{-1} \\ &- (X'X)^{-1}R'P^{-1}R(X'X)^{-1}\sigma^2 \end{aligned}$$

So, the first term is the OLS covariance. The second term is PSD, and the third term is NSD.

- If the restriction is true, the second term is 0, so we are better off. *True restrictions improve efficiency of estimation.*

- If the restriction is false, we may be better or worse off, in terms of MSE, depending on the magnitudes of $r - R\beta$ and $\sigma^2$.

## 6.2 Testing

In many cases, one wishes to test economic theories. If theory suggests parameter restrictions, as in the above homogeneity example, one can test theory by testing parameter restrictions. A number of tests are available. The first two (t and F) have a known small sample distributions, when the errors are normally distributed. The third and fourth (Wald and score) do not require normality of the errors, but their distributions are known only approximately, so that they are not exactly valid with finite samples.

### t-test

Suppose one has the model

$$y = X\beta + \varepsilon$$

and one wishes to test the *single restriction* $H_0 : R\beta = r$ vs. $H_A : R\beta \neq r$ . Under $H_0$, with normality of the errors,

$$R\hat{\beta} - r \sim N\left(0, R(X'X)^{-1}R'\sigma_0^2\right)$$

so

$$\frac{R\hat{\beta} - r}{\sqrt{R(X'X)^{-1}R'\sigma_0^2}} = \frac{R\hat{\beta} - r}{\sigma_0\sqrt{R(X'X)^{-1}R'}} \sim N(0,1).$$

The problem is that $\sigma_0^2$ is unknown. One could use the consistent estimator $\widehat{\sigma_0^2}$ in place of $\sigma_0^2$, but the test would only be valid asymptotically in this case.

**Proposition 2.** $\frac{N(0,1)}{\sqrt{\frac{\chi^2(q)}{q}}} \sim t(q)$

*as long as the $N(0,1)$ and the $\chi^2(q)$ are independent.*

We need a few results on the $\chi^2$ distribution.

**Proposition 3.** *If $x \sim N(\mu, I_n)$ is a vector of $n$ independent r.v.'s., then $x'x \sim \chi^2(n, \lambda)$ where $\lambda = \Sigma_i \mu_i^2 = \mu'\mu$ is the* noncentrality parameter.

When a $\chi^2$ r.v. has the noncentrality parameter equal to zero, it is referred to as a central $\chi^2$ r.v., and it's distribution is written as $\chi^2(n)$, suppressing the noncentrality parameter.

**Proposition 4.** *If the $n$ dimensional random vector $x \sim N(0, V)$, then $x'V^{-1}x \sim \chi^2(n)$.*

We'll prove this one as an indication of how the following unproven propositions could be proved.

Proof: Factor $V^{-1}$ as $P'P$ (this is the Cholesky factorization, where $P$ is defined to be upper triangular). Then consider $y = Px$. We have

$$y \sim N(0, PVP')$$

but

$$\begin{aligned} VP'P &= I_n \\ PVP'P &= P \end{aligned}$$

so $PVP' = I_n$ and thus $y \sim N(0, I_n)$. Thus $y'y \sim \chi^2(n)$ but

$$y'y = x'P'Px = xV^{-1}x$$

and we get the result we wanted.

A more general proposition which implies this result is

**Proposition 5.** *If the $n$ dimensional random vector $x \sim N(0, V)$, then $x'Bx \sim \chi^2(\rho(B))$ if and only if $BV$ is idempotent.*

An immediate consequence is

**Proposition 6.** *If the random vector (of dimension $n$) $x \sim N(0, I)$, and $B$ is idempotent with rank $r$, then $x'Bx \sim \chi^2(r)$.*

Consider the random variable

$$\begin{aligned}\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma_0^2} &= \frac{\varepsilon' M_X \varepsilon}{\sigma_0^2} \\ &= \left(\frac{\varepsilon}{\sigma_0}\right)' M_X \left(\frac{\varepsilon}{\sigma_0}\right) \\ &\sim \chi^2(n-K)\end{aligned}$$

**Proposition 7.** *If the random vector (of dimension n) $x \sim N(0, I)$, then $Ax$ and $x'Bx$ are independent if $AB = 0$.*

Now consider (remember that we have only one restriction in this case)

$$\frac{\frac{R\hat{\beta}-r}{\sigma_0\sqrt{R(X'X)^{-1}R'}}}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{(n-K)\sigma_0^2}}} = \frac{R\hat{\beta}-r}{\widehat{\sigma_0}\sqrt{R(X'X)^{-1}R'}}$$

This will have the $t(n-K)$ distribution if $\hat{\beta}$ and $\hat{\varepsilon}'\hat{\varepsilon}$ are independent. But $\hat{\beta} =$

$\beta + (X'X)^{-1}X'\varepsilon$ and

$$(X'X)^{-1}X'M_X = 0,$$

so

$$\frac{R\hat{\beta} - r}{\widehat{\sigma_0}\sqrt{R(X'X)^{-1}R'}} = \frac{R\hat{\beta} - r}{\hat{\sigma}_{R\hat{\beta}}} \sim t(n-K)$$

In particular, for the commonly encountered *test of significance* of an individual coefficient, for which $H_0 : \beta_i = 0$ vs. $H_0 : \beta_i \neq 0$ , the test statistic is

$$\frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}i}} \sim t(n-K)$$

- **Note**: the $t-$ test is strictly valid only if the errors are actually normally distributed. If one has nonnormal errors, one could use the above asymptotic result to justify taking critical values from the $N(0,1)$ distribution, since $t(n-K) \xrightarrow{d} N(0,1)$ as $n \to \infty$. In practice, a conservative procedure is to take critical values from the $t$ distribution if nonnormality is suspected. This will reject $H_0$ less often since the $t$ distribution is fatter-tailed than is

the normal.

## $F$ test

The $F$ test allows testing multiple restrictions jointly.

**Proposition 8.** *If $x \sim \chi^2(r)$ and $y \sim \chi^2(s)$, then $\frac{x/r}{y/s} \sim F(r,s)$, provided that $x$ and $y$ are independent.*

**Proposition 9.** *If the random vector (of dimension $n$) $x \sim N(0, I)$, then $x'Ax$*

*and $x'Bx$ are independent if $AB = 0$.*

Using these results, and previous results on the $\chi^2$ distribution, it is simple to show that the following statistic has the $F$ distribution:

$$F = \frac{\left(R\hat{\beta} - r\right)' \left(R\left(X'X\right)^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right)}{q\hat{\sigma}^2} \sim F(q, n-K).$$

A numerically equivalent expression is

$$\frac{\left(ESS_R - ESS_U\right)/q}{ESS_U/(n-K)} \sim F(q, n-K).$$

- **Note:** The $F$ test is strictly valid only if the errors are truly normally distributed. The following tests will be appropriate when one cannot assume normally distributed errors.

## Wald-type tests

The $t$ and $F$ tests require normality of the errors. The Wald test does not, but it is an asymptotic test - it is only approximately valid in finite samples.

The Wald principle is based on the idea that if a restriction is true, the unrestricted model should "approximately" satisfy the restriction. Given that the least

squares estimator is asymptotically normally distributed:

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N\left(0, \sigma_0^2 Q_X^{-1}\right)$$

then under $H_0 : R\beta_0 = r$, we have

$$\sqrt{n}\left(R\hat{\beta}-r\right) \xrightarrow{d} N\left(0, \sigma_0^2 R Q_X^{-1} R'\right)$$

so by Proposition [4]

$$n\left(R\hat{\beta}-r\right)'\left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1}\left(R\hat{\beta}-r\right) \xrightarrow{d} \chi^2(q)$$

Note that $Q_X^{-1}$ or $\sigma_0^2$ are not observable. The test statistic we use substitutes the consistent estimators. Use $(X'X/n)^{-1}$ as the consistent estimator of $Q_X^{-1}$. With this, there is a cancellation of $n's$, and the statistic to use is

$$\left(R\hat{\beta}-r\right)'\left(\widehat{\sigma_0^2} R (X'X)^{-1} R'\right)^{-1}\left(R\hat{\beta}-r\right) \xrightarrow{d} \chi^2(q)$$

- The Wald test is a simple way to test restrictions without having to estimate the restricted model.

- Note that this formula is similar to one of the formulae provided for the $F$ test.

## Score-type tests (Rao tests, Lagrange multiplier tests)

The score test is another asymptotically valid test that does not require normality of the errors.

In some cases, an unrestricted model may be nonlinear in the parameters, but the model is linear in the parameters under the null hypothesis. For example, the model

$$y = (X\beta)^{\gamma} + \varepsilon$$

is nonlinear in $\beta$ and $\gamma$, but is linear in $\beta$ under $H_0 : \gamma = 1$. Estimation of nonlinear models is a bit more complicated, so one might prefer to have a test based upon

the restricted, linear model. The score test is useful in this situation.

- Score-type tests are based upon the general principle that the gradient vector of the unrestricted model, evaluated at the restricted estimate, should be asymptotically normally distributed with mean zero, if the restrictions are true. The original development was for ML estimation, but the principle is valid for a wide variety of estimation methods.

We have seen that

$$\begin{aligned}\hat{\lambda} &= \left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right)\\ &= P^{-1}\left(R\hat{\beta}-r\right)\end{aligned}$$

so

$$\sqrt{n}\hat{P}\lambda = \sqrt{n}\left(R\hat{\beta}-r\right)$$

Given that

$$\sqrt{n}\left(R\hat{\beta}-r\right) \xrightarrow{d} N\left(0, \sigma_0^2 R Q_X^{-1} R'\right)$$

under the null hypothesis, we obtain

$$\sqrt{n}\hat{P}\lambda \xrightarrow{d} N\left(0, \sigma_0^2 R Q_X^{-1} R'\right)$$

So

$$\left(\sqrt{n}\hat{P}\lambda\right)' \left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1} \left(\sqrt{n}\hat{P}\lambda\right) \xrightarrow{d} \chi^2(q)$$

Noting that $\lim nP = RQ_X^{-1}R'$, we obtain,

$$\hat{\lambda}' \left(\frac{R(X'X)^{-1}R'}{\sigma_0^2}\right) \hat{\lambda} \xrightarrow{d} \chi^2(q)$$

since the powers of $n$ cancel. To get a usable test statistic substitute a consistent estimator of $\sigma_0^2$.

- This makes it clear why the test is sometimes referred to as a Lagrange multiplier test. It may seem that one needs the actual Lagrange multipliers to calculate this. If we impose the restrictions by substitution, these are not

available. Note that the test can be written as

$$\frac{\left(R'\hat{\lambda}\right)'(X'X)^{-1}R'\hat{\lambda}}{\sigma_0^2} \xrightarrow{d} \chi^2(q)$$

However, we can use the fonc for the restricted estimator:

$$-X'y + X'X\hat{\beta}_R + R'\hat{\lambda}$$

to get that

$$\begin{aligned} R'\hat{\lambda} &= X'(y - X\hat{\beta}_R) \\ &= X'\hat{\varepsilon}_R \end{aligned}$$

Substituting this into the above, we get

$$\frac{\hat{\varepsilon}_R' X(X'X)^{-1}X'\hat{\varepsilon}_R}{\sigma_0^2} \xrightarrow{d} \chi^2(q)$$

but this is simply

$$\hat{\varepsilon}_R' \frac{P_X}{\sigma_0^2} \hat{\varepsilon}_R \xrightarrow{d} \chi^2(q).$$

To see why the test is also known as a score test, note that the fonc for restricted least squares

$$-X'y + X'X\hat{\beta}_R + R'\hat{\lambda}$$

give us

$$R'\hat{\lambda} = X'y - X'X\hat{\beta}_R$$

and the rhs is simply the gradient (score) of the unrestricted model, evaluated at the restricted estimator. The scores evaluated at the unrestricted estimate are identically zero. The logic behind the score test is that the scores evaluated at the restricted estimate should be approximately zero, if the restriction is true. The test is also known as a Rao test, since P. Rao first proposed it in 1948.

## 6.3 The asymptotic equivalence of the LR, Wald and score tests

Note: the discussion of the LR test has been moved forward in these notes. I no longer teach the material in this section, but I'm leaving it here for reference.

We have seen that the three tests all converge to $\chi^2$ random variables. In fact, they all converge to the *same* $\chi^2$ rv, under the null hypothesis. We'll show that the Wald and LR tests are asymptotically equivalent. We have seen that the Wald test is asymptotically equivalent to

$$W \stackrel{a}{=} n\left(R\hat{\beta} - r\right)' \left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1} \left(R\hat{\beta} - r\right) \xrightarrow{d} \chi^2(q) \tag{6.1}$$

Using

$$\hat{\beta} - \beta_0 = (X'X)^{-1} X'\varepsilon$$

and

$$R\hat{\beta} - r = R(\hat{\beta} - \beta_0)$$

we get

$$\begin{aligned}\sqrt{n}R(\hat{\beta}-\beta_0) &= \sqrt{n}R(X'X)^{-1}X'\varepsilon \\ &= R\left(\frac{X'X}{n}\right)^{-1} n^{-1/2}X'\varepsilon\end{aligned}$$

Substitute this into [6.1] to get

$$\begin{aligned}W &\stackrel{a}{=} n^{-1}\varepsilon' X Q_X^{-1} R' \left(\sigma_0^2 R Q_X^{-1} R'\right)^{-1} R Q_X^{-1} X'\varepsilon \\ &\stackrel{a}{=} \varepsilon' X (X'X)^{-1} R' \left(\sigma_0^2 R (X'X)^{-1} R'\right)^{-1} R (X'X)^{-1} X'\varepsilon \\ &\stackrel{a}{=} \frac{\varepsilon' A (A'A)^{-1} A'\varepsilon}{\sigma_0^2} \\ &\stackrel{a}{=} \frac{\varepsilon' P_R \varepsilon}{\sigma_0^2}\end{aligned}$$

where $P_R$ is the projection matrix formed by the matrix $X(X'X)^{-1}R'$.

- Note that this matrix is idempotent and has $q$ columns, so the projection matrix has rank $q$.

Now consider the likelihood ratio statistic

$$LR \stackrel{a}{=} n^{1/2} g(\theta_0)' \mathcal{I}(\theta_0)^{-1} R' \left( R\mathcal{I}(\theta_0)^{-1} R' \right)^{-1} R\mathcal{I}(\theta_0)^{-1} n^{1/2} g(\theta_0) \tag{6.2}$$

Under normality, we have seen that the likelihood function is

$$\ln L(\beta, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2} \frac{(y - X\beta)'(y - X\beta)}{\sigma^2}.$$

Using this,

$$\begin{aligned} g(\beta_0) &\equiv D_\beta \frac{1}{n} \ln L(\beta, \sigma) \\ &= \frac{X'(y - X\beta_0)}{n\sigma^2} \\ &= \frac{X'\varepsilon}{n\sigma^2} \end{aligned}$$

Also, by the information matrix equality:

$$\begin{aligned}
\mathcal{I}(\theta_0) &= -H_\infty(\theta_0) \\
&= \lim -D_{\beta'} g(\beta_0) \\
&= \lim -D_{\beta'} \frac{X'(y - X\beta_0)}{n\sigma^2} \\
&= \lim \frac{X'X}{n\sigma^2} \\
&= \frac{Q_X}{\sigma^2}
\end{aligned}$$

so

$$\mathcal{I}(\theta_0)^{-1} = \sigma^2 Q_X^{-1}$$

Substituting these last expressions into [6.2], we get

$$\begin{aligned} LR &\stackrel{a}{=} \varepsilon' X'(X'X)^{-1}R' \left(\sigma_0^2 R(X'X)^{-1}R'\right)^{-1} R(X'X)^{-1}X'\varepsilon \\ &\stackrel{a}{=} \frac{\varepsilon' P_R \varepsilon}{\sigma_0^2} \\ &\stackrel{a}{=} W \end{aligned}$$

This completes the proof that the Wald and LR tests are asymptotically equivalent. Similarly, one can show that, *under the null hypothesis*,

$$qF \stackrel{a}{=} W \stackrel{a}{=} LM \stackrel{a}{=} LR$$

- The proof for the statistics except for $LR$ does not depend upon normality of the errors, as can be verified by examining the expressions for the statistics.
- The $LR$ statistic *is* based upon distributional assumptions, since one can't write the likelihood function without them.
- However, due to the close relationship between the statistics $qF$ and $LR$,

supposing normality, the $qF$ statistic can be thought of as a *pseudo-LR statistic,* in that it's like a LR statistic in that it uses the value of the objective functions of the restricted and unrestricted models, but it doesn't require distributional assumptions.

- The presentation of the score and Wald tests has been done in the context of the linear model. This is readily generalizable to nonlinear models and/or other estimation methods.

Though the four statistics *are* asymptotically equivalent, they are numerically different in small samples. The numeric values of the tests also depend upon how $\sigma^2$ is estimated, and we've already seen than there are several ways to do this. For example all of the following are consistent for $\sigma^2$ under $H_0$

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$
$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$
$$\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n-k+q}$$
$$\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n}$$

and in general the denominator call be replaced with any quantity $a$ such that $\lim a/n = 1$.

It can be shown, for linear regression models subject to linear restrictions, and if $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$ is used to calculate the Wald test and $\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n}$ is used for the score test, that

$$W > LR > LM.$$

For this reason, the Wald test will always reject if the LR test rejects, and in turn the LR test rejects if the LM test rejects. This is a bit problematic: there

is the possibility that by careful choice of the statistic used, one can manipulate reported results to favor or disfavor a hypothesis. A conservative/honest approach would be to report all three test statistics when they are available. In the case of linear models with normal errors the $F$ test is to be preferred, since asymptotic approximations are not an issue.

The small sample behavior of the tests can be quite different. The true size (probability of rejection of the null when the null is true) of the Wald test is often dramatically higher than the nominal size associated with the asymptotic distribution. Likewise, the true size of the score test is often smaller than the nominal size.

## 6.4 Interpretation of test statistics

Now that we have a menu of test statistics, we need to know how to use them.

## 6.5 Confidence intervals

Confidence intervals for single coefficients are generated in the normal manner. Given the $t$ statistic

$$t(\beta) = \frac{\hat{\beta} - \beta}{\widehat{\sigma_{\hat{\beta}}}}$$

a $100\,(1-\alpha)\,\%$ confidence interval for $\beta_0$ is defined by the bounds of the set of $\beta$ such that $t(\beta)$ does not reject $H_0 : \beta_0 = \beta$, using a $\alpha$ significance level:

$$C(\alpha) = \{\beta : -c_{\alpha/2} < \frac{\hat{\beta} - \beta}{\widehat{\sigma_{\hat{\beta}}}} < c_{\alpha/2}\}$$

The set of such $\beta$ is the interval

$$\hat{\beta} \pm \widehat{\sigma_{\hat{\beta}}} c_{\alpha/2}$$

A confidence ellipse for two coefficients jointly would be, analogously, the set of $\{\beta_1, \beta_2\}$ such that the $F$ (or some other test statistic) doesn't reject at the

specified critical value. This generates an ellipse, if the estimators are correlated.

- The region is an ellipse, since the CI for an individual coefficient defines a (infinitely long) rectangle with total prob. mass $1 - \alpha$, since the other coefficient is marginalized (e.g., can take on any value). Since the ellipse is bounded in both dimensions but also contains mass $1 - \alpha$, it must extend beyond the bounds of the individual CI.

- From the picture we can see that:

  - Rejection of hypotheses individually does not imply that the joint test will reject.

  - Joint rejection does not imply individual tests will reject.

Figure 6.1: Joint and Individual Confidence Regions

## 6.6 Bootstrapping

When we rely on asymptotic theory to use the normal distribution-based tests and confidence intervals, we're often at serious risk of making important errors. If the sample size is small and errors are highly nonnormal, the small sample distribution of $\sqrt{n}\left(\hat{\beta}-\beta_0\right)$ may be very different from its large sample distribution. Also, the distributions of test statistics may not resemble their limiting distributions at all. A means of trying to gain information on the small sample distribution of test statistics and estimators is the *bootstrap.* We'll consider a simple example, just to get the main idea.

Suppose that

$$\begin{aligned} y &= X\beta_0+\varepsilon \\ \varepsilon &\sim IID(0,\sigma_0^2) \\ X &\text{ is nonstochastic} \end{aligned}$$

Given that the distribution of $\varepsilon$ is unknown, the distribution of $\hat{\beta}$ will be unknown in small samples. However, since we have random sampling, we could generate *artificial data.* The steps are:

1. Draw $n$ observations from $\hat{\varepsilon}$ **with replacement**. Call this vector $\tilde{\varepsilon}^j$ (it's a $n \times 1$).

2. Then generate the data by $\tilde{y}^j = X\hat{\beta} + \tilde{\varepsilon}^j$

3. Now take this and estimate

$$\tilde{\beta}^j = (X'X)^{-1}X'\tilde{y}^j.$$

4. Save $\tilde{\beta}^j$

5. Repeat steps 1-4, until we have a large number, $J$, of $\tilde{\beta}^j$.

With this, we can use the replications to calculate the *empirical distribution of* $\tilde{\beta}_j$. One way to form a 100(1-$\alpha$)% confidence interval for $\beta_0$ would be to order the

$\tilde{\beta}^j$ from smallest to largest, and drop the first and last $J\alpha/2$ of the replications, and use the remaining endpoints as the limits of the CI. Note that this will not give the shortest CI if the empirical distribution is skewed.

- Suppose one was interested in the distribution of some function of $\hat{\beta}$, for example a test statistic. Simple: just calculate the transformation for each $j$, and work with the empirical distribution of the transformation.

- If the assumption of iid errors is too strong (for example if there is heteroscedasticity or autocorrelation, see below) one can work with a bootstrap defined by sampling from $(y, x)$ with replacement.

- How to choose $J$: $J$ should be large enough that the results don't change with repetition of the entire bootstrap. This is easy to check. If you find the results change a lot, increase $J$ and try again.

- The bootstrap is based fundamentally on the idea that the empirical distribution of the sample data converges to the actual sampling distribution as

$n$ becomes large, so statistics based on sampling from the empirical distribution should converge in distribution to statistics based on sampling from the actual sampling distribution.

- In finite samples, this doesn't hold. At a minimum, the bootstrap is a good way to check if asymptotic theory results offer a decent approximation to the small sample distribution.

- Bootstrapping can be used to test hypotheses. Basically, use the bootstrap to get an approximation to the empirical distribution of the test statistic under the alternative hypothesis, and use this to get critical values. Compare the test statistic calculated using the real data, under the null, to the bootstrap critical values. There are many variations on this theme, which we won't go into here.

## 6.7 Wald test for nonlinear restrictions: the delta method

Testing nonlinear restrictions of a linear model is not much more difficult, at least when the model is linear. Since estimation subject to nonlinear restrictions requires nonlinear estimation methods, which are beyond the score of this course, we'll just consider the Wald test for nonlinear restrictions on a linear model.

Consider the $q$ nonlinear restrictions

$$r(\beta_0) = 0.$$

where $r(\cdot)$ is a $q$-vector valued function. Write the derivative of the restriction evaluated at $\beta$ as

$$D_{\beta'} r(\beta)\big|_{\beta} = R(\beta)$$

We suppose that the restrictions are not redundant in a neighborhood of $\beta_0$, so

that

$$\rho(R(\beta)) = q$$

in a neighborhood of $\beta_0$. Take a first order Taylor's series expansion of $r(\hat{\beta})$ about $\beta_0$:

$$r(\hat{\beta}) = r(\beta_0) + R(\beta^*)(\hat{\beta} - \beta_0)$$

where $\beta^*$ is a convex combination of $\hat{\beta}$ and $\beta_0$. Under the null hypothesis we have

$$r(\hat{\beta}) = R(\beta^*)(\hat{\beta} - \beta_0)$$

Due to consistency of $\hat{\beta}$ we can replace $\beta^*$ by $\beta_0$, asymptotically, so

$$\sqrt{n}r(\hat{\beta}) \stackrel{a}{=} \sqrt{n}R(\beta_0)(\hat{\beta} - \beta_0)$$

We've already seen the distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$. Using this we get

$$\sqrt{n}r(\hat{\beta}) \xrightarrow{d} N\left(0, R(\beta_0)Q_X^{-1}R(\beta_0)'\sigma_0^2\right).$$

Considering the quadratic form

$$\frac{nr(\hat{\beta})'\left(R(\beta_0)Q_X^{-1}R(\beta_0)'\right)^{-1}r(\hat{\beta})}{\sigma_0^2} \xrightarrow{d} \chi^2(q)$$

under the null hypothesis. Substituting consistent estimators for $\beta_{0,}Q_X$ and $\sigma_0^2$, the resulting statistic is

$$\frac{r(\hat{\beta})'\left(R(\hat{\beta})(X'X)^{-1}R(\hat{\beta})'\right)^{-1}r(\hat{\beta})}{\widehat{\sigma^2}} \xrightarrow{d} \chi^2(q)$$

under the null hypothesis.

- This is known in the literature as the *delta method*, or as *Klein's approximation*.

- Since this is a Wald test, it will tend to over-reject in finite samples. The score and LR tests are also possibilities, but they require estimation methods for nonlinear models, which aren't in the scope of this course.

Note that this also gives a convenient way to estimate nonlinear functions and associated asymptotic confidence intervals. If the nonlinear function $r(\beta_0)$ is not hypothesized to be zero, we just have

$$\sqrt{n}\left(r(\hat{\beta}) - r(\beta_0)\right) \xrightarrow{d} N\left(0, R(\beta_0)Q_X^{-1}R(\beta_0)'\sigma_0^2\right)$$

so an approximation to the distribution of the function of the estimator is

$$r(\hat{\beta}) \approx N(r(\beta_0), R(\beta_0)(X'X)^{-1}R(\beta_0)'\sigma_0^2)$$

For example, the vector of elasticities of a function $f(x)$ is

$$\eta(x) = \frac{\partial f(x)}{\partial x} \odot \frac{x}{f(x)}$$

where $\odot$ means element-by-element multiplication. Suppose we estimate a linear function

$$y = x'\beta + \varepsilon.$$

The elasticities of $y$ w.r.t. $x$ are

$$\eta(x) = \frac{\beta}{x'\beta} \odot x$$

(note that this is the entire vector of elasticities). The estimated elasticities are

$$\widehat{\eta}(x) = \frac{\hat{\beta}}{x'\hat{\beta}} \odot x$$

To calculate the estimated standard errors of all five elasticities, use

$$\begin{aligned} R(\beta) &= \frac{\partial \eta(x)}{\partial \beta'} \\ &= \frac{\begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & x_k \end{bmatrix} x'\beta - \begin{bmatrix} \beta_1 x_1^2 & 0 & \cdots & 0 \\ 0 & \beta_2 x_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \beta_k x_k^2 \end{bmatrix}}{(x'\beta)^2}. \end{aligned}$$

To get a consistent estimator just substitute in $\hat{\beta}$. Note that the elasticity and the standard error are functions of $x$. The program ExampleDeltaMethod.jl shows how this can be done.

In many cases, nonlinear restrictions can also involve the data, not just the parameters. For example, consider a model of expenditure shares. Let $x(p, m)$ be a demand function, where $p$ is prices and $m$ is income. An expenditure share system for $G$ goods is

$$s_i(p, m) = \frac{p_i x_i(p, m)}{m}, i = 1, 2, ..., G.$$

Now demand must be positive, and we assume that expenditures sum to income, so we have the restrictions

$$\begin{aligned} 0 \leq s_i(p, m) &\leq 1, \ \forall i \\ \sum_{i=1}^{G} s_i(p, m) &= 1 \end{aligned}$$

Suppose we postulate a linear model for the expenditure shares:

$$s_i(p, m) = \beta_1^i + p'\beta_p^i + m\beta_m^i + \varepsilon^i$$

It is fairly easy to write restrictions such that the shares sum to one, but the restriction that the shares lie in the $[0, 1]$ interval depends on both parameters and the values of $p$ and $m$. It is impossible to impose the restriction that $0 \leq s_i(p, m) \leq 1$ for all possible $p$ and $m$. In such cases, one might consider whether or not a linear model is a reasonable specification.

## 6.8 Example: the Nerlove data

Remember that we in a previous example (section 4.8) that the OLS results for the Nerlove model are

**********

```
OLS estimation, 145 observations
RÃ²:  0.925955    ÏÂ²:  0.153943
White's covariance estimator

              coef        se         t         p
constant  -3.52650   1.68871  -2.08828   0.03858
  output   0.72039   0.03203  22.49083   0.00000
   labor   0.43634   0.24136   1.80782   0.07278
    fuel   0.42652   0.07417   5.75054   0.00000
 capital  -0.21989   0.31818  -0.69108   0.49066
*****************************************************
```

Note that $s_K = \beta_K < 0$, and that $\beta_L + \beta_F + \beta_K \neq 1$.

Remember that if we have constant returns to scale, then $\beta_Q = 1$, and if there is homogeneity of degree 1 then $\beta_L + \beta_F + \beta_K = 1$. We can test these hypotheses either separately or jointly. NerloveRestrictions.jl imposes and tests CRTS and then HOD1. From it we obtain the results that follow:

Imposing and testing CRTS

Restricted LS estimation, 145 observations
$R^2$: 0.790420  $\sigma^2$: 0.438861
White's covariance estimator

| parameter | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -7.53038 | 2.91950 | -2.57934 | 0.01094 |
| output | 1.00000 | 0.00000 | Inf | 0.00000 |
| labor | 0.01955 | 0.37573 | 0.05202 | 0.95859 |
| fuel | 0.71501 | 0.15923 | 4.49040 | 0.00001 |
| capital | 0.07580 | 0.57629 | 0.13154 | 0.89554 |

| | Value | p-value |
|---|---|---|
| qF | 256.26200 | 0.00000 |
| Wald | 265.41421 | 0.00000 |
| LR | 150.86281 | 0.00000 |
| Score | 93.77127 | 0.00000 |

Imposing and testing HOD1

Restricted LS estimation, 145 observations
$R^2$: 0.925652  $\sigma^2$: 0.155686
White's covariance estimator

| parameter | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -4.69079 | 0.80354 | -5.83766 | 0.00000 |
| output | 0.72069 | 0.03201 | 22.51556 | 0.00000 |
| labor | 0.59291 | 0.16672 | 3.55631 | 0.00051 |
| fuel | 0.41447 | 0.07186 | 5.76815 | 0.00000 |
| capital | -0.00738 | 0.15363 | -0.04805 | 0.96175 |

| | Value | p-value |
|---|---|---|
| qF | 0.57366 | 0.44881 |
| Wald | 0.59415 | 0.44082 |
| LR | 0.59293 | 0.44129 |
| Score | 0.59172 | 0.44175 |

Notice that the input price coefficients in fact sum to 1 when HOD1 is imposed.

HOD1 is not rejected at usual significance levels (*e.g.*, $\alpha = 0.10$). Also, $R^2$ does not drop much when the restriction is imposed, compared to the unrestricted results. For CRTS, you should note that $\beta_Q = 1$, so the restriction is satisfied. Also note that the hypothesis that $\beta_Q = 1$ is rejected by the test statistics at all reasonable significance levels. Note that $R^2$ drops quite a bit when imposing CRTS. If you look at the unrestricted estimation results, you can see that a t-test for $\beta_Q = 1$ also rejects, and that a confidence interval for $\beta_Q$ does not overlap 1.

From the point of view of neoclassical economic theory, these results are not anomalous: HOD1 is an implication of the theory, but CRTS is not.

**Exercise 10.** Modify the NerloveRestrictions.jl program to impose and test the restrictions jointly.

**The Chow test** Since CRTS is rejected, let's examine the possibilities more carefully. Recall that the data is sorted by output (the third column). Define 5 subsamples of firms, with the first group being the 29 firms with the lowest

output levels, then the next 29 firms, etc. The five subsamples can be indexed by $j = 1, 2, ..., 5$, where $j = 1$ for $t = 1, 2, ...29$, $j = 2$ for $t = 30, 31, ...58$, etc. Define *indicator variables* $D_1, D_2, ..., D_5$ where

$$D_1 = \begin{cases} 1 & t \in \{1, 2, ...29\} \\ 0 & t \notin \{1, 2, ...29\} \end{cases}$$

$$D_2 = \begin{cases} 1 & t \in \{30, 31, ...58\} \\ 0 & t \notin \{30, 31, ...58\} \end{cases}$$

$$\vdots$$

$$D_5 = \begin{cases} 1 & t \in \{117, 118, ..., 145\} \\ 0 & t \notin \{117, 118, ..., 145\} \end{cases}$$

Define the model

$$\ln C_t = \sum_{j=1}^{5} \alpha_1 D_j + \sum_{j=1}^{5} \gamma_j D_j \ln Q_t + \sum_{j=1}^{5} \beta_{Lj} D_j \ln P_{Lt} + \sum_{j=1}^{5} \beta_{Fj} D_j \ln P_{Ft} + \sum_{j=1}^{5} \beta_{Kj} D_j \ln P_{Kt} + \epsilon_t \tag{6.3}$$

Note that the first column of nerlove.data indicates this way of breaking up the sample, and provides and easy way of defining the indicator variables. The new model may be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \\ y_5 \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & & 0 \\ 0 & X_2 & & & \\ \vdots & & X_3 & & \\ & & & X_4 & 0 \\ 0 & & & & X_5 \end{bmatrix} \begin{bmatrix} \beta^1 \\ \beta^2 \\ \\ \\ \beta^5 \end{bmatrix} + \begin{bmatrix} \epsilon^1 \\ \epsilon^2 \\ \vdots \\ \\ \epsilon^5 \end{bmatrix} \tag{6.4}$$

where $y_1$ is $29\times1$, $X_1$ is $29\times5$, $\beta^j$ is the $5 \times 1$ vector of coefficients for the $j^{th}$ subsample (e.g., $\beta^1 = (\alpha_1, \gamma_1, \beta_{L1}, \beta_{F1}, \beta_{K1})'$), and $\epsilon^j$ is the $29 \times 1$ vector of errors

for the $j^{th}$ subsample.

The Julia program Restrictions/ChowTest.jl estimates the above model. It also tests the hypothesis that the five subsamples share the same parameter vector, or in other words, that there is coefficient stability across the five subsamples. The null to test is that the parameter vectors for the separate groups are all the same, that is,

$$\beta^1 = \beta^2 = \beta^3 = \beta^4 = \beta^5$$

This type of test, that parameters are constant across different sets of data, is sometimes referred to as a *Chow test.*

- There are 20 restrictions. If that's not clear to you, look at the Julia program.
- The restrictions are rejected at all conventional significance levels.

Since the restrictions are rejected, we should probably use the unrestricted model for analysis. What is the pattern of RTS as a function of the output group (small

Figure 6.2: RTS as a function of firm size

to large)? Figure 6.2 plots RTS. We can see that there is increasing RTS for small firms, but that RTS is approximately constant for large firms.

## 6.9 Exercises

1. Using the Chow test on the Nerlove model, we reject that there is coefficient stability across the 5 groups. But perhaps we could restrict the input price coefficients to be the same but let the constant and output coefficients vary by group size. This new model is

$$\ln C = \sum_{j=1}^{5} \alpha_j D_j + \sum_{j=1}^{5} \gamma_j D_j \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon \quad (6.5)$$

   (a) estimate this model by OLS, giving $R^2$, estimated standard errors for coefficients, t-statistics for tests of significance, and the associated p-values. Interpret the results in detail.

   (b) Test the restrictions implied by this model (relative to the model that lets all coefficients vary across groups) using the F, qF, Wald, score and likelihood ratio tests. Comment on the results.

   (c) Estimate this model but imposing the HOD1 restriction, *using an OLS*

estimation program. Give estimated standard errors for all coefficients.

(d) Plot the estimated RTS parameters as a function of firm size. Compare the plot to that given in the notes for the unrestricted model. Comment on the results.

2. For the model of the above question, compute 95% confidence intervals for RTS for each of the 5 groups of firms, using the delta method to compute standard errors. Comment on the results.

3. Perform a Monte Carlo study that generates data from the model

$$y = -2 + 1x_2 + 1x_3 + \epsilon$$

where the sample size is 30, $x_2$ and $x_3$ are independently uniformly distributed on $[0, 1]$ and $\epsilon \sim IIN(0, 1)$

(a) Compare the means and standard errors of the estimated coefficients

using OLS and restricted OLS, imposing the restriction that $\beta_2+\beta_3 = 2$.

(b) Compare the means and standard errors of the estimated coefficients using OLS and restricted OLS, imposing the restriction that $\beta_2+\beta_3 = 1$.

(c) Discuss the results.

# Chapter 7

# Stochastic regressors

Up to now we have treated the regressors as fixed, which is clearly unrealistic. Now we will assume they are random. There are several ways to think of the problem. First, if we are interested in an analysis *conditional* on the explanatory variables, then it is irrelevant if they are stochastic or not, since conditional on the values of they regressors take on, they are nonstochastic, which is the case already considered.

- In cross-sectional analysis it is usually reasonable to make the analysis conditional on the regressors.

- In dynamic models, where $y_t$ may depend on $y_{t-1}$, a conditional analysis is not sufficiently general, since we may want to predict into the future many periods out, so we need to consider the behavior of $\hat{\beta}$ and the relevant test statistics unconditional on $X$.

The model we'll deal will involve a combination of the following assumptions

**Assumption 11.** ***Linearity****: the model is a linear function of the parameter vector* $\beta_0$ :

$$y_t = x_t'\beta_0 + \varepsilon_t,$$

*or in matrix form,*

$$y = X\beta_0 + \varepsilon,$$

*where* $y$ *is* $n \times 1$, $X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}'$, *where* $x_t$ *is* $K \times 1$, *and* $\beta_0$ *and* $\varepsilon$ *are conformable.*

**Assumption 12.** ***Stochastic, linearly independent regressors***

*$X$ has rank $K$ with probability 1*

*$X$ is stochastic*

$\lim_{n\to\infty} \Pr\left(\frac{1}{n}X'X = Q_X\right) = 1$, *where $Q_X$ is a finite positive definite matrix.*

**Assumption 13.** ***Central limit theorem***

$n^{-1/2}X'\varepsilon \xrightarrow{d} N(0, Q_X\sigma_0^2)$

**Assumption 14.** ***Normality (Optional)***: $\varepsilon|X \sim N(0, \sigma^2 I_n)$: *$\epsilon$ is normally distributed*

**Assumption 15.** ***Strongly exogenous regressors***. *The regressors* $\mathbf{X}$ *are strongly exogenous if*

$$\mathcal{E}(\varepsilon_t|\mathbf{X}) \;=\; 0, \forall t \tag{7.1}$$

**Assumption 16.** ***Weakly exogenous regressors***: *The regressors are weakly exogenous if*

$$\mathcal{E}(\varepsilon_t|\mathbf{x}_t) \;=\; 0, \forall t$$

In both cases, $\mathbf{x}_t'\beta$ is the conditional mean of $y_t$ given $\mathbf{x}_t$: $E(y_t|\mathbf{x}_t) = \mathbf{x}_t'\beta$

## 7.1 Case 1

*Normality of* $\varepsilon$, strongly exogenous regressors

In this case,

$$\hat{\beta} = \beta_0 + (X'X)^{-1}X'\varepsilon$$

$$\begin{aligned}\mathcal{E}(\hat{\beta}|X) &= \beta_0 + (X'X)^{-1}X'\mathcal{E}(\varepsilon|X) \\ &= \beta_0\end{aligned}$$

and since this holds for all $X$, $E(\hat{\beta}) = \beta$, unconditional on $X$. Likewise,

$$\hat{\beta}|X \sim N\left(\beta, (X'X)^{-1}\sigma_0^2\right)$$

- If the density of $X$ is $d\mu(X)$, the marginal density of $\hat{\beta}$ is obtained by multiplying the conditional density by $d\mu(X)$ and integrating over $X$. Doing this leads to a nonnormal density for $\hat{\beta}$, in small samples.

- However, conditional on $X$, the usual test statistics have the $t$, $F$ and $\chi^2$ distributions. *Importantly,* these distributions don't depend on $X$, so when

marginalizing to obtain the unconditional distribution, nothing changes. The tests are valid in small samples.

- Summary: When $X$ is stochastic but strongly exogenous and $\varepsilon$ is normally distributed:

  1. $\hat{\beta}$ is unbiased
  2. $\hat{\beta}$ is nonnormally distributed
  3. The usual test statistics have the same distribution as with nonstochastic $X$.
  4. The Gauss-Markov theorem still holds, since it holds conditionally on $X$, and this is true for all $X$.
  5. Asymptotic properties are treated in the next section.

## 7.2 Case 2

*$\varepsilon$ nonnormally distributed, strongly exogenous regressors*

The unbiasedness of $\hat{\beta}$ carries through as before. However, the argument regarding test statistics doesn't hold, due to nonnormality of $\varepsilon$. Still, we have

$$\begin{aligned}\hat{\beta} &= \beta_0 + (X'X)^{-1}X'\varepsilon \\ &= \beta_0 + \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}\end{aligned}$$

Now

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{p} Q_X^{-1}$$

by assumption, and

$$\frac{X'\varepsilon}{n} = \frac{n^{-1/2}X'\varepsilon}{\sqrt{n}} \xrightarrow{p} 0$$

since the numerator converges to a $N(0, Q_X\sigma^2)$ r.v. and the denominator still goes to infinity. We have unbiasedness and the variance disappearing, so, *the estimator*

*is consistent*:

$$\hat{\beta} \xrightarrow{p} \beta_0.$$

Considering the asymptotic distribution

$$\begin{aligned}\sqrt{n}\left(\hat{\beta} - \beta_0\right) &= \sqrt{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n} \\ &= \left(\frac{X'X}{n}\right)^{-1} n^{-1/2}X'\varepsilon\end{aligned}$$

so

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, Q_X^{-1}\sigma_0^2)$$

directly following the assumptions. *Asymptotic normality of the estimator still holds.* Since the asymptotic results on all test statistics only require this, all the previous asymptotic results on test statistics are also valid in this case.

- Summary: Under strongly exogenous regressors, with $\varepsilon$ normal or nonnormal, $\hat{\beta}$ has the properties:

1. Unbiasedness
2. Consistency
3. Gauss-Markov theorem holds, since it holds in the previous case and doesn't depend on normality.
4. Asymptotic normality
5. Tests are asymptotically valid
6. Tests are not valid in small samples if the error is normally distributed

## 7.3 Case 3

*Weakly exogenous regressors*

An important class of models are *dynamic models*, where lagged dependent variables have an impact on the current value. A simple version of these models

that captures the important points is

$$\begin{aligned} y_t &= z_t'\alpha + \sum_{s=1}^{p} \gamma_s y_{t-s} + \varepsilon_t \\ &= x_t'\beta + \varepsilon_t \end{aligned}$$

where now $x_t$ contains lagged dependent variables. Clearly, even with $E(\epsilon_t|\mathbf{x}_t) = 0$, $X$ and $\varepsilon$ are not uncorrelated, so one can't show unbiasedness. For example,

$$\mathcal{E}(\varepsilon_{t-1}x_t) \neq 0$$

since $x_t$ contains $y_{t-1}$ (which is a function of $\varepsilon_{t-1}$) as an element.

- This fact implies that all of the small sample properties such as unbiasedness, Gauss-Markov theorem, and small sample validity of test statistics *do not hold* in this case. Recall Figure 4.7. This is a case of weakly exogenous regressors, and we see that the OLS estimator is biased in this case.

- Nevertheless, under the above assumptions, all asymptotic properties con-

tinue to hold, using the same arguments as before.

## 7.4 When are the assumptions reasonable?

The two assumptions we've added are

1. $\lim_{n\to\infty} \Pr\left(\frac{1}{n}X'X = Q_X\right) = 1$, a $Q_X$ finite positive definite matrix.
2. $n^{-1/2}X'\varepsilon \xrightarrow{d} N(0, Q_X\sigma_0^2)$

The most complicated case is that of dynamic models, since the other cases can be treated as nested in this case. There exist a number of central limit theorems for dependent processes, many of which are fairly technical. We won't enter into details (see Hamilton, Chapter 7 if you're interested). A main requirement for use of standard asymptotics for a dependent sequence

$$\{s_t\} = \{\frac{1}{n}\sum_{t=1}^{n} z_t\}$$

to converge in probability to a finite limit is that $z_t$ be *stationary*, in some sense.

- Strong stationarity requires that the joint distribution of the set

$$\{z_t, z_{t+s}, z_{t-q}, ...\}$$

  not depend on $t$.

- Covariance (weak) stationarity requires that the first and second moments of this set not depend on $t$.

- An example of a sequence that doesn't satisfy this is an AR(1) process with a unit root (a *random walk)*:

$$\begin{aligned} x_t &= x_{t-1} + \varepsilon_t \\ \varepsilon_t &\sim IIN(0, \sigma^2) \end{aligned}$$

  One can show that the variance of $x_t$ depends upon $t$ in this case, so it's not weakly stationary.

- The series $\sin t + \epsilon_t$ has a first moment that depends upon $t$, so it's not weakly stationary either.

Stationarity prevents the process from trending off to plus or minus infinity, and prevents cyclical behavior which would allow correlations between far removed $z_t$ znd $z_s$ to be high. *Draw a picture here.*

- In summary, the assumptions are reasonable when the stochastic conditioning variables have variances that are finite, and are not too strongly dependent. The AR(1) model with unit root is an example of a case where the dependence is too strong for standard asymptotics to apply.

- The study of nonstationary processes is an important part of econometrics, but it isn't in the scope of this course.

## 7.5 Exercises

1. Show that for two random variables $A$ and $B$, if $E(A|B) = 0$, then $E\left(Af(B)\right) = 0$. How is this used in the proof of the Gauss-Markov theorem?

2. Is it possible for an AR(1) model for time series data, *e.g.*, $y_t = 0 + 0.9y_{t-1} + \varepsilon_t$ satisfy weak exogeneity? Strong exogeneity? Discuss.

# Chapter 8

# Data problems

In this section we'll consider problems associated with the regressor matrix: collinearity, missing observations and measurement error.

# 8.1 Collinearity

## Motivation: Data on Mortality and Related Factors

The data set mortality.data contains annual data from 1947 - 1980 on death rates in the U.S., along with data on factors like smoking and consumption of alcohol. The data description is:

> DATA4-7: Death rates in the U.S. due to coronary heart disease and their determinants. Data compiled by Jennifer Whisenand

- chd = death rate per 100,000 population (Range 321.2 - 375.4)
- cal = Per capita consumption of calcium per day in grams (Range 0.9 - 1.06)
- unemp = Percent of civilian labor force unemployed in 1,000 of persons 16 years and older (Range 2.9 - 8.5)
- cig = Per capita consumption of cigarettes in pounds of tobacco by persons

18 years and older–approx. 339 cigarettes per pound of tobacco (Range 6.75 - 10.46)

- edfat = Per capita intake of edible fats and oil in pounds–includes lard, margarine and butter (Range 42 - 56.5)
- meat = Per capita intake of meat in pounds–includes beef, veal, pork, lamb and mutton (Range 138 - 194.8)
- spirits = Per capita consumption of distilled spirits in taxed gallons for individuals 18 and older (Range 1 - 2.9)
- beer = Per capita consumption of malted liquor in taxed gallons for individuals 18 and older (Range 15.04 - 34.9)
- wine = Per capita consumption of wine measured in taxed gallons for individuals 18 and older (Range 0.77 - 2.65)

Consider estimation results for several models:

$$\widehat{\text{chd}} = \underset{(58.939)}{334.914} + \underset{(5.156)}{5.41216}\,\text{cig} + \underset{(7.373)}{36.8783}\,\text{spirits} - \underset{(1.2513)}{5.10365}\,\text{beer}$$
$$+ \underset{(12.735)}{13.9764}\,\text{wine}$$
$$T = 34 \quad \bar{R}^2 = 0.5528 \quad F(4, 29) = 11.2 \quad \hat{\sigma} = 9.9945$$

(standard errors in parentheses)

$$\widehat{\text{chd}} = \underset{(56.624)}{353.581} + \underset{(4.7523)}{3.17560}\,\text{cig} + \underset{(7.275)}{38.3481}\,\text{spirits} - \underset{(1.0102)}{4.28816}\,\text{beer}$$
$$T = 34 \quad \bar{R}^2 = 0.5498 \quad F(3, 30) = 14.433 \quad \hat{\sigma} = 10.028$$

(standard errors in parentheses)

$$\widehat{\text{chd}} = \underset{(67.21)}{243.310} + \underset{(6.1508)}{10.7535}\,\text{cig} + \underset{(8.0359)}{22.8012}\,\text{spirits} - \underset{(12.638)}{16.8689}\,\text{wine}$$

$$T = 34 \quad \bar{R}^2 = 0.3198 \quad F(3, 30) = 6.1709 \quad \hat{\sigma} = 12.327$$

(standard errors in parentheses)

$$\widehat{\text{chd}} = \underset{(49.119)}{181.219} + \underset{(4.4371)}{16.5146}\,\text{cig} + \underset{(6.2079)}{15.8672}\,\text{spirits}$$

$$T = 34 \quad \bar{R}^2 = 0.3026 \quad F(2, 31) = 8.1598 \quad \hat{\sigma} = 12.481$$

(standard errors in parentheses)

Note how the signs of the coefficients change depending on the model, and that the magnitudes of the parameter estimates vary a lot, too. The parameter estimates are highly sensitive to the particular model we estimate. Why? We'll see that the problem is that the data exhibit *collinearity*.

# Collinearity: definition

Collinearity is the existence of linear relationships amongst the regressors. We can always write

$$\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_K \mathbf{x}_K + v = 0$$

where $\mathbf{x}_i$ is the $i^{th}$ column of the regressor matrix $X$, and $v$ is an $n \times 1$ vector. In the case that there exists collinearity, the variation in $v$ is relatively small, so that there is an approximately exact linear relation between the regressors.

- "relative" and "approximate" are imprecise, so it's difficult to define when collinearilty exists.

In the extreme, if there are exact linear relationships (every element of $v$ equal) then $\rho(X) < K$, so $\rho(X'X) < K$, so $X'X$ is not invertible and the OLS estimator

is not uniquely defined. For example, if the model is

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t \\ x_{2t} &= \alpha_1 + \alpha_2 x_{3t} \end{aligned}$$

then we can write

$$\begin{aligned} y_t &= \beta_1 + \beta_2 \left(\alpha_1 + \alpha_2 x_{3t}\right) + \beta_3 x_{3t} + \varepsilon_t \\ &= \beta_1 + \beta_2 \alpha_1 + \beta_2 \alpha_2 x_{3t} + \beta_3 x_{3t} + \varepsilon_t \\ &= \left(\beta_1 + \beta_2 \alpha_1\right) + \left(\beta_2 \alpha_2 + \beta_3\right) x_{3t} \\ &= \gamma_1 + \gamma_2 x_{3t} + \varepsilon_t \end{aligned}$$

- The $\gamma's$ can be consistently estimated, but since the $\gamma'$s define two equations in three $\beta's$, the $\beta's$ can't be consistently estimated (there are multiple values of $\beta$ that solve the first order conditions). The $\beta's$ are *unidentified* in the case of perfect collinearity.

- Perfect collinearity is unusual, except in the case of an error in construction of the regressor matrix, such as including the same regressor twice.

Another case where perfect collinearity may be encountered is with models with indicator variables, if one is not careful. Consider a model of rental price ($y_i$) of an apartment. This could depend factors such as size, quality etc., collected in $x_i$, as well as on the location of the apartment. Let $B_i = 1$ if the $i^{th}$ apartment is in Barcelona, $B_i = 0$ otherwise. Similarly, define $G_i$, $T_i$ and $L_i$ for Girona, Tarragona and Lleida. One could use a model such as

$$y_i = \beta_1 + \beta_2 B_i + \beta_3 G_i + \beta_4 T_i + \beta_5 L_i + x_i' \gamma + \varepsilon_i$$

In this model, $B_i + G_i + T_i + L_i = 1$, $\forall i$, so there is an exact relationship between these variables and the column of ones corresponding to the constant. One must either drop the constant, or one of the qualitative variables.

# A brief aside on indicator variables

**indicator variable**: A indicator variable is a binary-valued variable that indicates whether or not some condition is true. It is customary to assign the value 1 if the condition is true, and 0 if the condition is false.

indicator variables are used essentially like any other regressor. Use $d$ to indicate that a variable is a indicator, so that variables like $d_t$ and $d_{t2}$ are understood to be indicator variables. Variables like $x_t$ and $x_{t3}$ are ordinary continuous regressors. You know how to interpret the following models:

$$y_t = \beta_1 + \beta_2 d_t + \epsilon_t$$

$$y_t = \beta_1 d_t + \beta_2 (1 - d_t) + \epsilon_t$$

$$y_t = \beta_1 + \beta_2 d_t + \beta_3 x_t + \epsilon_t$$

**Interaction terms:** an interaction term is the product of two variables, so that the effect of one variable on the dependent variable depends on the value of the other. The following model has an interaction term. Note that $\frac{\partial E(y|x)}{\partial x} = \beta_3 + \beta_4 d_t$. The slope depends on the value of $d_t$.

$$y_t = \beta_1 + \beta_2 d_t + \beta_3 x_t + \beta_4 d_t x_t + \epsilon_t$$

**Multiple indicator variables:** we can use more than one indicator variable in a model. We will study models of the form

$$y_t = \beta_1 + \beta_2 d_{t1} + \beta_3 d_{t2} + \beta_4 x_t + \epsilon_t$$

$$y_t = \beta_1 + \beta_2 d_{t1} + \beta_3 d_{t2} + \beta_4 d_{t1} d_{t2} + \beta_5 x_t + \epsilon_t$$

**Incorrect usage:** You should understand why the following models are not correct usages of indicator variables:

1. overparameterization:

$$y_t = \beta_1 + \beta_2 d_t + \beta_3(1 - d_t) + \epsilon_t$$

2. multiple values assigned to multiple categories. Suppose that we a condition that defines 4 possible categories, and we create a variable $d = 1$ if the observation is in the first category, $d = 2$ if in the second, etc. (This is not strictly speaking a indicator variable, according to our definition). Why is the following model not a good one?

$$y_t = \beta_1 + \beta_2 d + \epsilon$$

What is the correct way to deal with this situation?

**Multiple parameterizations.** To formulate a model that conditions on a given set of categorical information, there are multiple ways to use indicator vari-

ables. For example, the two models

$$y_t = \beta_1 d_t + \beta_2(1 - d_t) + \beta_3 x_t + \beta_4 d_t x_t + \epsilon_t$$

and

$$y_t = \alpha_1 + \alpha_2 d_t + \alpha_3 x_t d_t + \alpha_4 x_t(1 - d_t) + \epsilon_t$$

are equivalent. You should know what are the 4 equations that relate the $\beta_j$ parameters to the $\alpha_j$ parameters, $j = 1, 2, 3, 4$. You should know how to interpret the parameters of both models.

## Back to collinearity

The more common case, if one doesn't make mistakes such as these, is the existence of inexact linear relationships, *i.e.*, correlations between the regressors that are less than one in absolute value, but not zero. The basic problem is that when two (or

more) variables move together, it is difficult to determine their separate influences.

**Example 17.** Two children are in a room, along with a broken lamp. Both say "I didn't do it!". This is like two regressors giving exactly the same information. How can we tell who broke the lamp?

Lack of knowledge about the separate influences of variables is reflected in imprecise estimates, *i.e.*, estimates with high variances. *With economic data, collinearity is commonly encountered, and is often a severe problem.*

When there is collinearity, the minimizing point of the objective function that defines the OLS estimator ($s(\beta)$, the sum of squared errors) is relatively poorly defined. This is seen in Figures 8.1 and 8.2.

To see the effect of collinearity on variances, partition the regressor matrix as

$$X = \begin{bmatrix} \mathbf{x} & W \end{bmatrix}$$

where $\mathbf{x}$ is the first column of $X$ (note: we can interchange the columns of $X$ isf we like, so there's no loss of generality in considering the first column). Now, the

Figure 8.1: $s(\beta)$ when there is no collinearity

Figure 8.2: $s(\beta)$ when there is collinearity

variance of $\hat{\beta}$, under the classical assumptions, is

$$V(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

Using the partition,

$$X'X = \begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'W \\ W'\mathbf{x} & W'W \end{bmatrix}$$

and following a rule for partitioned inversion,

$$\begin{aligned} (X'X)^{-1}_{1,1} &= \left(\mathbf{x}'\mathbf{x} - \mathbf{x}'W(W'W)^{-1}W'\mathbf{x}\right)^{-1} \\ &= \left(\mathbf{x}'\left(I_n - W(W'W)'^{1}W'\right)\mathbf{x}\right)^{-1} \\ &= \left(ESS_{\mathbf{x}|W}\right)^{-1} \end{aligned}$$

where by $ESS_{\mathbf{x}|W}$ we mean the error sum of squares obtained from the regression

$$\mathbf{x} = W\lambda + v.$$

Since

$$R^2 = 1 - ESS/TSS,$$

we have

$$ESS = TSS(1 - R^2)$$

so the variance of the coefficient corresponding to $\mathbf{x}$ is

$$V(\hat{\beta}_{\mathbf{x}}) = \frac{\sigma^2}{TSS_{\mathbf{x}}(1 - R^2_{\mathbf{x}|W})} \tag{8.1}$$

We see three factors influence the variance of this coefficient. It will be high if

1. $\sigma^2$ is large

2. There is little variation in $\mathbf{x}$. *Draw a picture here.*

3. There is a strong linear relationship between $x$ and the other regressors, so that $W$ can explain the movement in $\mathbf{x}$ well. In this case, $R^2_{\mathbf{x}|W}$ will be close to 1. As $R^2_{\mathbf{x}|W} \to 1, V(\hat{\beta}_{\mathbf{x}}) \to \infty$.

The last of these cases is collinearity.

Intuitively, when there are strong linear relations between the regressors, it is difficult to determine the separate influence of the regressors on the dependent variable. This can be seen by comparing the OLS objective function in the case of no correlation between regressors with the objective function with correlation between the regressors. See the figures nocollin.ps (no correlation) and collin.ps (correlation), available on the web site.

**Example 18.** The Julia script DataProblems/collinearity.jl performs a Monte Carlo study with correlated regressors. The model is $y = 1 + x_2 + x_3 + \epsilon$, where the correlation between $x_2$ and $x_3$can be set. Three estimators are used: OLS, OLS dropping $x_3$ (a false restriction), and restricted LS using $\beta_2 = \beta_3$ (a true restriction). The output when the correlation between the two regressors is 0.9 is

```
correlation between x2 and x3: 0.9
descriptive statistics for 1000 OLS replications
                 mean       std      skew      kurt       min       max
```

```
         1      0.997      0.188      0.073      0.176      0.190      1.606
         2      1.017      0.436     -0.057      0.394     -0.638      2.599
         3      0.995      0.437      0.031      0.379     -0.635      2.341
descriptive statistics for 1000 OLS replications, dropping x3
                 mean        std       skew       kurt        min        max
         1      0.999      0.199      0.088      0.196      0.123      1.598
         2      1.912      0.208     -0.169      0.490      1.099      2.549
descriptive statistics for 1000 Restricted OLS replications, b2=b3
                 mean        std       skew       kurt        min        max
         1      0.999      0.186      0.083      0.189      0.188      1.626
         2      1.007      0.098     -0.004      0.336      0.661      1.319
         3      1.007      0.098     -0.004      0.336      0.661      1.319
```

Figure 8.3 shows histograms for the estimated $\beta_2$, for each of the three estimators.

- Check the biases and variances.

Figure 8.3: Collinearity: Monte Carlo results

- repeat the experiment with a lower value of rho, and note how the standard errors of the OLS estimator change.

## Detection of collinearity

The best way is simply to regress each explanatory variable in turn on the remaining regressors. If any of these auxiliary regressions has a high $R^2$, there is a problem of collinearity. Furthermore, this procedure identifies which parameters are affected.

- Sometimes, we're only interested in certain parameters. Collinearity isn't a problem if it doesn't affect what we're interested in estimating.

An alternative is to examine the matrix of correlations between the regressors. High correlations are sufficient but not necessary for severe collinearity.

Also indicative of collinearity is that the model fits well (high $R^2$), but none of the variables is significantly different from zero (e.g., their separate influences aren't well determined).

In summary, the artificial regressions are the best approach if one wants to be careful.

**Example 19.** Nerlove data and collinearity. The simple Nerlove model is

$$\ln C = \beta_1 + \beta_2 \ln Q + \beta_3 \ln P_L + \beta_4 \ln P_F + \beta_5 \ln P_K + \epsilon$$

When this model is estimated by OLS, some coefficients are not significant (see subsection 4.8). Maybe this is due to collinearity? The Julia script DataProblems/NerloveCollinearity.jl checks the regressors for collinearity. If you run this, you will see that collinearity is not a problem with this data. Why is the coefficient of $\ln P_K$ not significantly different from zero?

# Dealing with collinearity

## More information

Collinearity is a problem of an uninformative sample. The first question is: is all the available information being used? Is more data available? Are there coefficient restrictions that have been neglected? *Picture illustrating how a restriction can*

*solve problem of perfect collinearity.*

## Stochastic restrictions and ridge regression

Note: here's a nice introduction to ridge regression: https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db.

Supposing that there is no more data or neglected restrictions, one possibility is to change perspectives, to Bayesian econometrics. One can express prior beliefs regarding the coefficients using stochastic restrictions. A stochastic linear restriction would be something of the form

$$R\beta = r + v$$

where $R$ and $r$ are as in the case of exact linear restrictions, but $v$ is a random

vector. For example, the model could be

$$
\begin{aligned}
y &= X\beta + \varepsilon \\
R\beta &= r + v \\
\begin{pmatrix} \varepsilon \\ v \end{pmatrix} &\sim N\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 I_n & 0_{n\times q} \\ 0_{q\times n} & \sigma_v^2 I_q \end{pmatrix}
\end{aligned}
$$

This sort of model isn't in line with the classical interpretation of parameters as constants: according to this interpretation the left hand side of $R\beta = r + v$ is constant but the right is random. This model does fit the Bayesian perspective: we combine information coming from the model and the data, summarized in

$$
\begin{aligned}
y &= X\beta + \varepsilon \\
\varepsilon &\sim N(0, \sigma_\varepsilon^2 I_n)
\end{aligned}
$$

with prior beliefs regarding the distribution of the parameter, summarized in

$$R\beta \sim N(r, \sigma_v^2 I_q)$$

Since the sample is random it is reasonable to suppose that $\mathcal{E}(\varepsilon v') = 0$, which is the last piece of information in the specification. How can you estimate using this model? The solution is to treat the restrictions as artificial data. Write

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ v \end{bmatrix}$$

This model is heteroscedastic, since $\sigma_\varepsilon^2 \neq \sigma_v^2$. Define the *prior precision* $k = \sigma_\varepsilon / \sigma_v$. This expresses the degree of belief in the restriction relative to the variability of the data. Supposing that we specify $k$, then the model

$$\begin{bmatrix} y \\ kr \end{bmatrix} = \begin{bmatrix} X \\ kR \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ kv \end{bmatrix}$$

is homoscedastic and can be estimated by OLS. Note that this estimator is biased.

It is consistent, however, given that $k$ is a fixed constant, even if the restriction is false (this is in contrast to the case of false exact restrictions). To see this, note that there are $Q$ restrictions, where $Q$ is the number of rows of $R$. As $n \to \infty$, these $Q$ artificial observations have no weight in the objective function, so the estimator has the same limiting objective function as the OLS estimator, and is therefore consistent.

To motivate the use of stochastic restrictions, consider the expectation of the squared length of $\hat{\beta}$:

$$\begin{aligned}
\mathcal{E}(\hat{\beta}'\hat{\beta}) &= \mathcal{E}\left\{\left(\beta + (X'X)^{-1} X'\varepsilon\right)' \left(\beta + (X'X)^{-1} X'\varepsilon\right)\right\} \\
&= \beta'\beta + \mathcal{E}\left(\varepsilon' X (X'X)^{-1} (X'X)^{-1} X'\varepsilon\right) \\
&= \beta'\beta + Tr\,(X'X)^{-1}\,\sigma^2 \\
&= \beta'\beta + \sigma^2 \sum_{i=1}^{K} \lambda_i (\text{the trace is the sum of eigenvalues}) \\
&> \beta'\beta + \lambda_{\max\left(X'X^{-1}\right)} \sigma^2 (\text{the eigenvalues are all positive, since} X'X \text{ is p.d.}
\end{aligned}$$

so

$$\mathcal{E}(\hat{\beta}'\hat{\beta}) > \beta'\beta + \frac{\sigma^2}{\lambda_{\min(X'X)}}$$

where $\lambda_{\min(X'X)}$ is the minimum eigenvalue of $X'X$ (which is the inverse of the maximum eigenvalue of $(X'X)^{-1}$). As collinearity becomes worse and worse, $X'X$ becomes more nearly singular, so $\lambda_{\min(X'X)}$ tends to zero (recall that the determinant is the product of the eigenvalues) and $\mathcal{E}(\hat{\beta}'\hat{\beta})$ tends to infinite. On the other hand, $\beta'\beta$ is finite.

Now considering the restriction $I_K\beta = 0 + v$. With this restriction the model becomes

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ kI_K \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ kv \end{bmatrix}$$

and the estimator is

$$\begin{aligned}\hat{\beta}_{ridge} &= \left(\begin{bmatrix} X' & kI_K \end{bmatrix}\begin{bmatrix} X \\ kI_K \end{bmatrix}\right)^{-1}\begin{bmatrix} X' & I_K \end{bmatrix}\begin{bmatrix} y \\ 0 \end{bmatrix} \\ &= \left(X'X + k^2 I_K\right)^{-1} X'y\end{aligned}$$

This is the ordinary *ridge regression* estimator. The ridge regression estimator can be seen to add $k^2 I_K$, which is nonsingular, to $X'X$, which is more and more nearly singular as collinearity becomes worse and worse. As $k \to \infty$, the restrictions tend to $\beta = 0$, that is, the coefficients are shrunken toward zero. Also, the estimator tends to

$$\hat{\beta}_{ridge} = \left(X'X + k^2 I_K\right)^{-1} X'y \to \left(k^2 I_K\right)^{-1} X'y = \frac{X'y}{k^2} \to 0$$

so $\hat{\beta}'_{ridge}\hat{\beta}_{ridge} \to 0$. This is clearly a false restriction in the limit, if our original model is at all sensible.

There should be some amount of shrinkage that is in fact a true restriction.

The problem is to determine the $k$ such that the restriction is correct. The interest in ridge regression centers on the fact that it can be shown that there exists a $k$ such that $MSE(\hat{\beta}_{ridge}) < \hat{\beta}_{OLS}$. The problem is that this $k$ depends on $\beta$ and $\sigma^2$, which are unknown.

The ridge trace method plots $\hat{\beta}'_{ridge}\hat{\beta}_{ridge}$ as a function of $k$, and chooses the value of $k$ that "artistically" seems appropriate (e.g., where the effect of increasing $k$ dies off). *Draw picture here.* This means of choosing $k$ is obviously subjective. This is not a problem from the Bayesian perspective: the choice of $k$ reflects prior beliefs about the length of $\beta$.

In summary, the ridge estimator offers some hope, but it is impossible to guarantee that it will outperform the OLS estimator. Collinearity is a fact of life in econometrics, and there is no clear solution to the problem.

The Julia script DataProblems/RidgeRegression.jl does a Monte Carlo study that shows that ridge regression can help to deal with collinearity. This script generates the following figures, which show the Monte Carlo sampling frequency

Figure 8.4: OLS and Ridge regression

of the OLS and ridge estimators, after subtracting the true parameter values. You can see that the ridge estimator has much lower RMSE: both histograms are centered close to zero, but the ridge histogram is much tighter.

## 8.2 Measurement error

Measurement error is exactly what it says, either the dependent variable or the regressors are measured with error. Thinking about the way economic data are reported, measurement error is probably quite prevalent. For example, estimates of growth of GDP, inflation, etc. are commonly revised several times. Why should the last revision necessarily be correct?

### Error of measurement of the dependent variable

Measurement errors in the dependent variable and the regressors have important differences. First consider error in measurement of the dependent variable. The

data generating process is presumed to be

$$\begin{aligned} y^* &= X\beta + \varepsilon \\ y &= y^* + v \\ v_t &\sim iid(0, \sigma_v^2) \end{aligned}$$

where $y^* = y + v$ is the unobservable true dependent variable, and $y$ is what is observed. We assume that $\varepsilon$ and $v$ are independent and that $y^* = X\beta + \varepsilon$ satisfies the classical assumptions. Given this, we have

$$y + v = X\beta + \varepsilon$$

so

$$\begin{aligned} y &= X\beta + \varepsilon - v \\ &= X\beta + \omega \\ \omega_t &\sim iid(0, \sigma_\varepsilon^2 + \sigma_v^2) \end{aligned}$$

- As long as $v$ is uncorrelated with $X$, this model satisfies the classical assumptions and can be estimated by OLS. This type of measurement error isn't a problem, then, except in that the increased variability of the error term causes an increase in the variance of the OLS estimator (see equation 8.1).

## Error of measurement of the regressors

The situation isn't so good in this case. The DGP is

$$\begin{aligned} y_t &= x_t^{*\prime}\beta + \varepsilon_t \\ x_t &= x_t^* + v_t \\ v_t &\sim iid(0, \Sigma_v) \end{aligned}$$

where $\Sigma_v$ is a $K \times K$ matrix. Now $X^*$ contains the true, unobserved regressors, and $X$ is what is observed. Again assume that $v$ is independent of $\varepsilon$, and that the

model $y = X^*\beta + \varepsilon$ satisfies the classical assumptions. Now we have

$$\begin{aligned} y_t &= (x_t - v_t)'\beta + \varepsilon_t \\ &= x_t'\beta - v_t'\beta + \varepsilon_t \\ &= x_t'\beta + \omega_t \end{aligned}$$

The problem is that now there is a correlation between $x_t$ and $\omega_t$, since

$$\begin{aligned} \mathcal{E}(x_t\omega_t) &= \mathcal{E}\left((x_t^* + v_t)\left(-v_t'\beta + \varepsilon_t\right)\right) \\ &= -\Sigma_v\beta \end{aligned}$$

where

$$\Sigma_v = \mathcal{E}\left(v_t v_t'\right).$$

Because of this correlation, the OLS estimator is biased and inconsistent, just as in the case of autocorrelated errors with lagged dependent variables. In matrix

notation, write the estimated model as

$$y = X\beta + \omega$$

We have that

$$\hat{\beta} = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'y}{n}\right)$$

and

$$\begin{aligned} plim\left(\frac{X'X}{n}\right)^{-1} &= plim\frac{\left(X^{*\prime} + V'\right)\left(X^{*} + V\right)}{n} \\ &= \left(Q_{X^*} + \Sigma_v\right)^{-1} \end{aligned}$$

since $X^*$ and $V$ are independent, and

$$\begin{aligned} plim\frac{V'V}{n} &= \lim \mathcal{E}\frac{1}{n}\sum_{t=1}^{n} v_t v_t' \\ &= \Sigma_v \end{aligned}$$

Likewise,

$$\begin{aligned} plim\left(\frac{X'y}{n}\right) &= plim\frac{\left(X^{*\prime}+V'\right)\left(X^{*}\beta+\varepsilon\right)}{n} \\ &= Q_{X^{*}}\beta \end{aligned}$$

so

$$plim\hat{\beta} = \left(Q_{X^{*}}+\Sigma_{v}\right)^{-1}Q_{X^{*}}\beta$$

So we see that the least squares estimator is inconsistent when the regressors are measured with error.

- A potential solution to this problem is the instrumental variables (IV) estimator, which we'll discuss shortly.

**Example 20.** Measurement error in a dynamic model. Consider the model

$$\begin{aligned} y_t^* &= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t \\ y_t &= y_t^* + \upsilon_t \end{aligned}$$

where $\epsilon_t$ and $\upsilon_t$ are independent Gaussian white noise errors. Suppose that $y_t^*$ is not observed, and instead we observe $y_t$. What are the properties of the OLS regression on the equation

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \eta_t$$

? The error is

$$\begin{aligned}
\eta_t &= y_t - \alpha - \rho y_{t-1} - \beta x_t \\
&= y_t^* + \upsilon_t - \alpha - \rho y_{t-1}^* - \rho \upsilon_{t-1} - \beta x_t \\
&= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t + \upsilon_t - \alpha - \rho y_{t-1}^* - \rho \upsilon_{t-1} - \beta x_t \\
&= \epsilon_t + \upsilon_t - \rho \upsilon_{t-1}
\end{aligned}$$

So the error term is autocorrelated. Note that

$$y_{t-1} = \alpha + \rho y_{t-2} + \beta x_{t-1} + \eta_{t-1}$$

and

$$\eta_{t-1} = \epsilon_{t-1} + \upsilon_{t-1} - \rho\upsilon_{t-2},$$

so the error $\eta_t$ and the regressor $y_{t-1}$ are correlated, because they share the common term $\upsilon_{t-1}$. This means that the equation

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \eta_t$$

does not satisfy weak exogeneity, and the OLS estimator will be biased and inconsistent.

The Julia script DataProblems/MeasurementError.jl does a Monte Carlo study. The sample size is $n = 100$. Figure 8.5 gives the results. The first panel shows a histogram for 1000 replications of $\hat{\rho} - \rho$, when $\sigma_\nu = 1$, so that there is significant measurement error. The second panel repeats this with $\sigma_\nu = 0$, so that there is not measurement error. Note that there is much more bias with measurement error. There is also bias without measurement error. This is due to the same reason that we saw bias in Figure 4.7: one of the classical assumptions (nonstochastic

regressors) that guarantees unbiasedness of OLS does not hold for this model. Without measurement error, the OLS estimator *is* consistent. By re-running the script with larger $n$, you can verify that the bias disappears when $\sigma_\nu = 0$, but not when $\sigma_\nu > 0$.

Figure 8.5: $\hat{\rho} - \rho$ with and without measurement error

(a) with measurement error: $\sigma_\nu = 1$

(b) without measurement error: $\sigma_\nu = 0$

## 8.3 Missing observations

Missing observations occur quite frequently: time series data may not be gathered in a certain year, or respondents to a survey may not answer all questions. We'll consider two cases: missing observations on the dependent variable and missing observations on the regressors.

## Missing observations on the dependent variable

In this case, we have

$$y = X\beta + \varepsilon$$

or

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

where $y_2$ is not observed. Otherwise, we assume the classical assumptions hold.

- A clear alternative is to simply estimate using the compete observations

  $$y_1 = X_1\beta + \varepsilon_1$$

  Since these observations satisfy the classical assumptions, one could estimate by OLS.

- The question remains whether or not one could somehow replace the unobserved $y_2$ by a predictor, and improve over OLS in some sense. Let $\hat{y}_2$ be

the predictor of $y_2$. Now

$$\begin{aligned}\hat{\beta} &= \left\{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}' \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right\}^{-1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ \hat{y}_2 \end{bmatrix} \\ &= \left[X_1'X_1 + X_2'X_2\right]^{-1} \left[X_1'y_1 + X_2'\hat{y}_2\right]\end{aligned}$$

Recall that the OLS fonc are

$$X'X\hat{\beta} = X'y$$

so if we regressed using only the first (complete) observations, we would have

$$X_1'X_1\hat{\beta}_1 = X_1'y_1.$$

Likewise, an OLS regression using only the second (filled in) observations would give

$$X_2'X_2\hat{\beta}_2 = X_2'\hat{y}_2.$$

Substituting these into the equation for the overall combined estimator gives

$$\begin{aligned}
\hat{\beta} &= [X_1'X_1 + X_2'X_2]^{-1}\left[X_1'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2\right] \\
&= [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1\hat{\beta}_1 + [X_1'X_1 + X_2'X_2]^{-1} X_2'X_2\hat{\beta}_2 \\
&\equiv A\hat{\beta}_1 + (I_K - A)\hat{\beta}_2
\end{aligned}$$

where

$$A \equiv [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1$$

and we use

$$\begin{aligned}
[X_1'X_1 + X_2'X_2]^{-1} X_2'X_2 &= [X_1'X_1 + X_2'X_2]^{-1} [(X_1'X_1 + X_2'X_2) - X_1'X_1] \\
&= I_K - [X_1'X_1 + X_2'X_2]^{-1} X_1'X_1 \\
&= I_K - A.
\end{aligned}$$

Now,

$$\mathcal{E}(\hat{\beta}) = A\beta + (I_K - A)\mathcal{E}\left(\hat{\beta}_2\right)$$

and this will be unbiased only if $\mathcal{E}\left(\hat{\beta}_2\right) = \beta$.

- The conclusion is that the filled in observations alone would need to define an unbiased estimator. This will be the case only if

$$\hat{y}_2 = X_2\beta + \hat{\varepsilon}_2$$

  where $\hat{\varepsilon}_2$ has mean zero. Clearly, it is difficult to satisfy this condition without knowledge of $\beta$.

- Note that putting $\hat{y}_2 = \bar{y}_1$ does not satisfy the condition and therefore leads to a biased estimator.

**Exercise 21.** Formally prove this last statement.

## The sample selection problem

In the above discussion we assumed that the missing observations are random. The sample selection problem is a case where the missing observations are not

random. Consider the model

$$y_t^* = x_t'\beta + \varepsilon_t$$

which is assumed to satisfy the classical assumptions. However, $y_t^*$ is not always observed. What is observed is $y_t$ defined as

$$y_t = y_t^* \text{ if } y_t^* \geq 0$$

Or, in other words, $y_t^*$ is missing when it is less than zero.

The difference in this case is that the missing values are not random: they are correlated with the $x_t$. Consider the case

$$y^* = x + \varepsilon$$

with $V(\varepsilon) = 25$, but using only the observations for which $y^* > 0$ to estimate. Figure 8.6 illustrates the bias. The Julia program is sampsel.jl

There are means of dealing with sample selection bias, but we will not go into

Figure 8.6: Sample selection bias

it here. One should at least be aware that nonrandom selection of the sample will normally lead to bias and inconsistency if the problem is not taken into account.

## Missing observations on the regressors

Again the model is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

but we assume now that each row of $X_2$ has an unobserved component(s). Again, one could just estimate using the complete observations, but it may seem frustrating to have to drop observations simply because of a single missing variable. In general, if the unobserved $X_2$ is replaced by some prediction, $X_2^*$, then we are in the case of errors of observation. As before, this means that the OLS estimator is biased when $X_2^*$ is used instead of $X_2$. Consistency is salvaged, however, as long as the number of missing observations doesn't increase with $n$.

- Including observations that have missing values replaced by *ad hoc* values

can be interpreted as introducing false stochastic restrictions. In general, this introduces bias. It is difficult to determine whether MSE increases or decreases. Monte Carlo studies suggest that it is dangerous to simply substitute the mean, for example.

- In the case that there is only one regressor other than the constant, substitution of $\bar{x}$ for the missing $x_t$ *does not lead to bias*. This is a special case that doesn't hold for $K > 2$.

**Exercise 22.** Prove this last statement.

- In summary, if one is strongly concerned with bias, it is best to drop observations that have missing components. There is potential for reduction of MSE through filling in missing elements with intelligent guesses, but this could also increase MSE.

## 8.4 Missing regressors

Suppose that the model $y = X\beta + W\gamma + \epsilon$ satisfies the classical assumptions, so OLS would be a consistent estimator. However, let's suppose that the regressors $W$ are not available in the sample. What are the properties of the OLS estimator of the model $y = X\beta + \omega$? We can think of this as a case of imposing false restrictions: $\gamma = 0$ when in fact $\gamma \neq 0$. We know that the restricted least squares estimator is biased and inconsistent, in general, when we impose false restrictions. Another way of thinking of this is to look to see if $X$ and $\omega$ are correlated. We have

$$\begin{aligned} E(X_t\omega_t) &= E\left(X_t\left(W_t'\gamma + \epsilon_t\right)\right) \\ &= E(X_tW_t'\gamma) + E(X_t\epsilon_t) \\ &= E(X_tW_t'\gamma) \end{aligned}$$

where the last line follows because $E(X_t\epsilon_t) = 0$ by assumption. So, there will be correlation between the error and the regressors if there is collinearity between the included regressors $X_t$ and the missing regressors $W_t$. If there is not, the OLS estimator will be consistent. Because the normal thing is to have collinearity between regressors, we expect that missing regressors will lead to bias and inconsistency of the OLS estimator.

## 8.5 Exercises

1. Consider the simple Nerlove model

$$\ln C = \beta_1 + \beta_2 \ln Q + \beta_3 \ln P_L + \beta_4 \ln P_F + \beta_5 \ln P_K + \epsilon$$

When this model is estimated by OLS, some coefficients are not significant. We have seen that collinearity is not an important problem. Why is $\beta_5$ not significantly different from zero? Give an economic explanation.

2. For the model $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$,

   (a) verify that the level sets of the OLS criterion function (defined in equation 4.2) are straight lines when there is perfect collinearity

   (b) For this model with perfect collinearity, the OLS estimator does not exist. Depict what this statement means using a drawing.

   (c) Show how a restriction $R_1\beta_1 + R_2\beta_2 = r$ causes the restricted least squares estimator to exist, using a drawing.

# Chapter 9

# Functional form and nonnested tests

Though theory often suggests which conditioning variables should be included, and suggests the signs of certain derivatives, it is usually silent regarding the functional form of the relationship between the dependent variable and the regressors. For

example, considering a cost function, one could have a Cobb-Douglas model

$$c = Aw_1^{\beta_1} w_2^{\beta_2} q^{\beta_q} e^{\varepsilon}$$

This model, after taking logarithms, gives

$$\ln c = \beta_0 + \beta_1 \ln w_1 + \beta_2 \ln w_2 + \beta_q \ln q + \varepsilon$$

where $\beta_0 = \ln A$. Theory suggests that $A > 0, \beta_1 > 0, \beta_2 > 0, \beta_3 > 0$. This model isn't compatible with a fixed cost of production since $c = 0$ when $q = 0$. Homogeneity of degree one in input prices suggests that $\beta_1 + \beta_2 = 1$, while constant returns to scale implies $\beta_q = 1$.

While this model may be reasonable in some cases, an alternative

$$\sqrt{c} = \beta_0 + \beta_1 \sqrt{w_1} + \beta_2 \sqrt{w_2} + \beta_q \sqrt{q} + \varepsilon$$

may be just as plausible. Note that $\sqrt{x}$ and $\ln(x)$ look quite alike, for certain

values of the regressors, and up to a linear transformation, so it may be difficult to choose between these models.

The basic point is that many functional forms are compatible with the linear-in-parameters model, since this model can incorporate a wide variety of nonlinear transformations of the dependent variable and the regressors. For example, suppose that $g(\cdot)$ is a real valued function and that $x(\cdot)$ is a $K-$ vector-valued function. The following model is linear in the parameters but nonlinear in the variables:

$$\begin{aligned} x_t &= x(z_t) \\ y_t &= x_t'\beta + \varepsilon_t \end{aligned}$$

There may be $P$ fundamental conditioning variables $z_t$, but there may be $K$ regressors, where $K$ may be smaller than, equal to or larger than $P$. For example, $x_t$ could include squares and cross products of the conditioning variables in $z_t$.

## 9.1 Flexible functional forms

Given that the functional form of the relationship between the dependent variable and the regressors is in general unknown, one might wonder if there exist parametric models that can closely approximate a wide variety of functional relationships. A "Diewert-Flexible" functional form is defined as one such that the function, the vector of first derivatives and the matrix of second derivatives can take on an arbitrary value *at a single data point.* Flexibility in this sense clearly requires that there be at least

$$K = 1 + P + \left(P^2 - P\right)/2 + P$$

free parameters: one for each independent effect that we wish to model.

Suppose that the model is

$$y = g(x) + \varepsilon$$

A second-order Taylor's series expansion (with remainder term) of the function

$g(x)$ about the point $x = 0$ is

$$g(x) = g(0) + x' D_x g(0) + \frac{x' D_x^2 g(0) x}{2} + R$$

Use the approximation, which simply drops the remainder term, as an approximation to $g(x)$ :

$$g(x) \simeq g_K(x) = g(0) + x' D_x g(0) + \frac{x' D_x^2 g(0) x}{2}$$

As $x \to 0$, the approximation becomes more and more exact, in the sense that $g_K(x) \to g(x)$, $D_x g_K(x) \to D_x g(x)$ and $D_x^2 g_K(x) \to D_x^2 g(x)$. For $x = 0$, the approximation is exact, up to the second order. The idea behind many flexible functional forms is to note that $g(0)$, $D_x g(0)$ and $D_x^2 g(0)$ are all constants. If we treat them as parameters, the approximation will have exactly enough free parameters to approximate the function $g(x)$, which is of unknown form, exactly,

up to second order, at the point $x = 0$. The model is

$$g_K(x) = \alpha + x'\beta + 1/2x'\Gamma x$$

so the regression model to fit is

$$y = \alpha + x'\beta + 1/2x'\Gamma x + \varepsilon$$

- While the regression model has enough free parameters to be Diewert-flexible, the question remains: is $plim\hat{\alpha} = g(0)$? Is $plim\hat{\beta} = D_x g(0)$? Is $plim\hat{\Gamma} = D_x^2 g(0)$?

- The answer is no, in general. The reason is that if we treat the true values of the parameters as these derivatives, then $\varepsilon$ is forced to play the part of the remainder term, which is a function of $x$, so that $x$ and $\varepsilon$ are correlated in this case. As before, the estimator is biased in this case.

- A simpler example would be to consider a first-order T.S. approximation to

a quadratic function. *Draw picture.*

- The conclusion is that "flexible functional forms" aren't really flexible in a useful statistical sense, in that neither the function itself nor its derivatives are consistently estimated, unless the function belongs to the parametric family of the specified functional form. In order to lead to consistent inferences, the regression model must be correctly specified.

## The translog form

In spite of the fact that FFF's aren't really flexible for the purposes of econometric estimation and inference, they are useful, and they are certainly subject to less bias due to misspecification of the functional form than are many popular forms, such as the Cobb-Douglas or the simple linear in the variables model. The translog model is probably the most widely used FFF. This model is as above, except that the variables are subjected to a logarithmic tranformation. Also, the expansion

point is usually taken to be the sample mean of the data, after the logarithmic transformation. The model is defined by

$$\begin{aligned} y &= \ln(c) \\ x &= \ln\left(\frac{z}{\bar{z}}\right) \\ &= \ln(z) - \ln(\bar{z}) \\ y &= \alpha + x'\beta + 1/2x'\Gamma x + \varepsilon \end{aligned}$$

In this presentation, the $t$ subscript that distinguishes observations is suppressed for simplicity. Note that

$$\begin{aligned} \frac{\partial y}{\partial x} &= \beta + \Gamma x \\ &= \frac{\partial \ln(c)}{\partial \ln(z)} \text{(the other part of } x \text{ is constant)} \\ &= \frac{\partial c}{\partial z}\frac{z}{c} \end{aligned}$$

which is the elasticity of $c$ with respect to $z$. This is a convenient feature of the translog model. Note that at the means of the conditioning variables, $\bar{z}$, $x = 0$, so

$$\left.\frac{\partial y}{\partial x}\right|_{z=\bar{z}} = \beta$$

so the $\beta$ are the first-order elasticities, at the means of the data.

To illustrate, consider that $y$ is cost of production:

$$y = c(w, q)$$

where $w$ is a vector of input prices and $q$ is output. We could add other variables by extending $q$ in the obvious manner, but this is supressed for simplicity. By Shephard's lemma, the conditional factor demands are

$$x = \frac{\partial c(w, q)}{\partial w}$$

and the cost shares of the factors are therefore

$$s = \frac{wx}{c} = \frac{\partial c(w,q)}{\partial w}\frac{w}{c}$$

which is simply the vector of elasticities of cost with respect to input prices. If the cost function is modeled using a translog function, we have

$$\begin{aligned}
\ln(c) &= \alpha + x'\beta + z'\delta + 1/2\begin{bmatrix} x' & z \end{bmatrix}\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{12}' & \Gamma_{22} \end{bmatrix}\begin{bmatrix} x \\ z \end{bmatrix} \\
&= \alpha + x'\beta + z'\delta + 1/2x'\Gamma_{11}x + x'\Gamma_{12}z + 1/2z^2\gamma_{22}
\end{aligned}$$

where $x = \ln(w/\bar{w})$ (element-by-element division) and $z = \ln(q/\bar{q})$, and

$$\begin{aligned}
\Gamma_{11} &= \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{bmatrix} \\
\Gamma_{12} &= \begin{bmatrix} \gamma_{13} \\ \gamma_{23} \end{bmatrix} \\
\Gamma_{22} &= \gamma_{33}.
\end{aligned}$$

Note that symmetry of the second derivatives has been imposed.

Then the share equations are just

$$s = \beta + \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}$$

Therefore, the share equations and the cost equation have parameters in common. By pooling the equations together and imposing the (true) restriction that the parameters of the equations be the same, we can gain efficiency.

To illustrate in more detail, consider the case of two inputs, so

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

In this case the translog model of the logarithmic cost function is

$$\ln c = \alpha + \beta_1 x_1 + \beta_2 x_2 + \delta z + \frac{\gamma_{11}}{2} x_1^2 + \frac{\gamma_{22}}{2} x_2^2 + \frac{\gamma_{33}}{2} z^2 + \gamma_{12} x_1 x_2 + \gamma_{13} x_1 z + \gamma_{23} x_2 z$$

The two cost shares of the inputs are the derivatives of $\ln c$ with respect to $x_1$ and $x_2$:

$$\begin{aligned} s_1 &= \beta_1 + \gamma_{11} x_1 + \gamma_{12} x_2 + \gamma_{13} z \\ s_2 &= \beta_2 + \gamma_{12} x_1 + \gamma_{22} x_2 + \gamma_{13} z \end{aligned}$$

Note that the share equations and the cost equation have parameters in common. One can do a pooled estimation of the three equations at once, imposing that the parameters are the same. In this way we're using more observations and there-

fore more information, which will lead to imporved efficiency. Note that this does assume that the cost equation is correctly specified (*i.e.*, not an approximation), since otherwise the derivatives would not be the true derivatives of the log cost function, and would then be misspecified for the shares. To pool the equations, write the model in matrix form (adding in error terms)

$$\begin{bmatrix} \ln c \\ s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 & z & \frac{x_1^2}{2} & \frac{x_2^2}{2} & \frac{z^2}{2} & x_1x_2 & x_1z & x_2z \\ 0 & 1 & 0 & 0 & x_1 & 0 & 0 & x_2 & z & 0 \\ 0 & 0 & 1 & 0 & 0 & x_2 & 0 & x_1 & 0 & z \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \delta \\ \gamma_{11} \\ \gamma_{22} \\ \gamma_{33} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{23} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

This is *one* observation on the three equations. With the appropriate notation, a single observation can be written as

$$y_t = X_t\theta + \varepsilon_t$$

The overall model would stack $n$ observations on the three equations for a total of $3n$ observations:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Next we need to consider the errors. For observation $t$ the errors can be placed in a vector

$$\varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}$$

First consider the covariance matrix of this vector: the shares are certainly correlated since they must sum to one. (In fact, with 2 shares the variances are equal and the covariance is -1 times the variance. General notation is used to allow easy extension to the case of more than 2 inputs). Also, it's likely that the shares and the cost equation have different variances. Supposing that the model is covariance stationary, the variance of $\varepsilon_t$ won't depend upon $t$:

$$Var\varepsilon_t = \Sigma_0 = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \cdot & \sigma_{22} & \sigma_{23} \\ \cdot & \cdot & \sigma_{33} \end{bmatrix}$$

Note that this matrix is singular, since the shares sum to 1. Assuming that there is no autocorrelation, the overall covariance matrix has the *seemingly unrelated*

*regressions* (SUR) structure.

$$
\begin{aligned}
Var \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} &= \Sigma \\
&= \begin{bmatrix} \Sigma_0 & 0 & \cdots & 0 \\ 0 & \Sigma_0 & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & \cdots & 0 & \Sigma_0 \end{bmatrix} \\
&= I_n \otimes \Sigma_0
\end{aligned}
$$

where the symbol $\otimes$ indicates the *Kronecker product*. The Kronecker product of

two matrices $A$ and $B$ is

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & \ddots & & \vdots \\ \vdots & & & \\ a_{pq}B & \cdots & & a_{pq}B \end{bmatrix}.$$

## FGLS estimation of a translog model

So, this model has heteroscedasticity and autocorrelation, so OLS won't be efficient. The next question is: how do we estimate efficiently using FGLS? FGLS is based upon inverting the estimated error covariance $\hat{\Sigma}$. So we need to estimate $\Sigma$.

An asymptotically efficient procedure is (supposing normality of the errors)

1. Estimate each equation by OLS

2. Estimate $\Sigma_0$ using

$$\hat{\Sigma}_0 = \frac{1}{n} \sum_{t=1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_t'$$

3. Next we need to account for the singularity of $\Sigma_0$. It can be shown that $\hat{\Sigma}_0$ will be singular when the shares sum to one, so FGLS won't work. The solution is to drop one of the share equations, for example the second. The model becomes

$$\begin{bmatrix} \ln c \\ s_1 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 & z & \frac{x_1^2}{2} & \frac{x_2^2}{2} & \frac{z^2}{2} & x_1x_2 & x_1z & x_2z \\ 0 & 1 & 0 & 0 & x_1 & 0 & 0 & x_2 & z & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \delta \\ \gamma_{11} \\ \gamma_{22} \\ \gamma_{33} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{23} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

or in matrix notation for the observation:

$$y_t^* = X_t^* \theta + \varepsilon_t^*$$

and in stacked notation for all observations we have the $2n$ observations:

$$\begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{bmatrix} = \begin{bmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_n^* \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1^* \\ \varepsilon_2^* \\ \vdots \\ \varepsilon_n^* \end{bmatrix}$$

or, finally in matrix notation for all observations:

$$y^* = X^* \theta + \varepsilon^*$$

Considering the error covariance, we can define

$$\begin{aligned}\Sigma_0^* &= Var\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix}\\ \Sigma^* &= I_n \otimes \Sigma_0^*\end{aligned}$$

Define $\hat{\Sigma}_0^*$ as the leading $2 \times 2$ block of $\hat{\Sigma}_0$ , and form

$$\hat{\Sigma}^* = I_n \otimes \hat{\Sigma}_0^*.$$

This is a consistent estimator, following the consistency of OLS and applying a LLN.

4. Next compute the Cholesky factorization

$$\hat{P}_0 = Chol\left(\hat{\Sigma}_0^*\right)^{-1}$$

(I am assuming this is defined as an upper triangular matrix, which is con-

sistent with the way Octave does it) and the Cholesky factorization of the overall covariance matrix of the 2 equation model, which can be calculated as

$$\hat{P} = Chol\hat{\Sigma}^* = I_n \otimes \hat{P}_0$$

5. Finally the FGLS estimator can be calculated by applying OLS to the transformed model

$$\hat{P}'y^* = \hat{P}'X^*\theta + \hat{\hat{P}}'\varepsilon^*$$

or by directly using the GLS formula

$$\hat{\theta}_{FGLS} = \left(X^{*\prime}\left(\hat{\Sigma}_0^*\right)^{-1}X^*\right)^{-1}X^{*\prime}\left(\hat{\Sigma}_0^*\right)^{-1}y^*$$

It is equivalent to transform each observation individually:

$$\hat{P}_0'y_y^* = \hat{P}_0'X_t^*\theta + \hat{P}_0'\varepsilon^*$$

and then apply OLS. This is probably the simplest approach.

A few last comments.

1. We have assumed no autocorrelation across time. This is clearly restrictive. It is relatively simple to relax this, but we won't go into it here.

2. Also, we have only imposed symmetry of the second derivatives. Another restriction that the model should satisfy is that the estimated shares should sum to 1. This can be accomplished by imposing

$$\begin{aligned} \beta_1 + \beta_2 &= 1 \\ \sum_{i=1}^{3} \gamma_{ij} &= 0, \; j = 1, 2, 3. \end{aligned}$$

   These are linear parameter restrictions, so they are easy to impose and will improve efficiency if they are true.

3. The estimation procedure outlined above can be *iterated.* That is, estimate

$\hat{\theta}_{FGLS}$ as above, then re-estimate $\Sigma_0^*$ using errors calculated as

$$\hat{\varepsilon} = y - X\hat{\theta}_{FGLS}$$

These might be expected to lead to a better estimate than the estimator based on $\hat{\theta}_{OLS}$, since FGLS is asymptotically more efficient. Then re-estimate $\theta$ using the new estimated error covariance. It can be shown that if this is repeated until the estimates don't change (*i.e.*, iterated to convergence) then the resulting estimator is the MLE. At any rate, the asymptotic properties of the iterated and uniterated estimators are the same, since both are based upon a consistent estimator of the error covariance.

## 9.2 Testing nonnested hypotheses

Given that the choice of functional form isn't perfectly clear, in that many possibilities exist, how can one choose between forms? When one form is a parametric

restriction of another, the previously studied tests such as Wald, LR, score or $qF$ are all possibilities. For example, the Cobb-Douglas model is a parametric restriction of the translog: The translog is

$$y_t = \alpha + x_t'\beta + 1/2x_t'\Gamma x_t + \varepsilon$$

where the variables are in logarithms, while the Cobb-Douglas is

$$y_t = \alpha + x_t'\beta + \varepsilon$$

so a test of the Cobb-Douglas versus the translog is simply a test that $\Gamma = 0$.

The situation is more complicated when we want to test *non-nested hypotheses*. If the two functional forms are linear in the parameters, and use the same

transformation of the dependent variable, then they may be written as

$$\begin{aligned} M_1 : y &= X\beta + \varepsilon \\ \varepsilon_t &\sim iid(0, \sigma_\varepsilon^2) \\ M_2 : y &= Z\gamma + \eta \\ \eta &\sim iid(0, \sigma_\eta^2) \end{aligned}$$

We wish to test hypotheses of the form: $H_0 : M_i$ *is correctly specified* versus $H_A : M_i$ *is misspecified*, for $i = 1, 2$.

- One could account for non-iid errors, but we'll suppress this for simplicity.
- There are a number of ways to proceed. We'll consider the $J$ test, proposed by Davidson and MacKinnon, *Econometrica* (1981). The idea is to artificially nest the two models, e.g.,

$$y = (1 - \alpha)X\beta + \alpha(Z\gamma) + \omega$$

If the first model is correctly specified, then the true value of $\alpha$ is zero. On the other hand, if the second model is correctly specified then $\alpha = 1$.

- The problem is that this model is not identified in general. For example, if the models share some regressors, as in

$$
\begin{aligned}
M_1 : y_t &= \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t \\
M_2 : y_t &= \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{4t} + \eta_t
\end{aligned}
$$

then the composite model is

$$
y_t = (1-\alpha)\beta_1 + (1-\alpha)\beta_2 x_{2t} + (1-\alpha)\beta_3 x_{3t} + \alpha\gamma_1 + \alpha\gamma_2 x_{2t} + \alpha\gamma_3 x_{4t} + \omega_t
$$

Combining terms we get

$$
\begin{aligned}
y_t &= ((1-\alpha)\beta_1 + \alpha\gamma_1) + ((1-\alpha)\beta_2 + \alpha\gamma_2)\, x_{2t} + (1-\alpha)\beta_3 x_{3t} + \alpha\gamma_3 x_{4t} + \omega_t \\
&= \delta_1 + \delta_2 x_{2t} + \delta_3 x_{3t} + \delta_4 x_{4t} + \omega_t
\end{aligned}
$$

The four $\delta's$ are consistently estimable, but $\alpha$ is not, since we have four equations in 7 unknowns, so one can't test the hypothesis that $\alpha = 0$.

The idea of the $J$ test is to substitute $\hat{\gamma}$ in place of $\gamma$. This is a consistent estimator supposing that the second model is correctly specified. It will tend to a finite probability limit even if the second model is misspecified. Then estimate the model

$$\begin{aligned} y &= (1-\alpha)X\beta + \alpha(Z\hat{\gamma}) + \omega \\ &= X\theta + \alpha\hat{y} + \omega \end{aligned}$$

where $\hat{y} = Z(Z'Z)^{-1}Z'y = P_Z y$. In this model, $\alpha$ is consistently estimable, and one can show that, under the hypothesis that the first model is correct, $\alpha \xrightarrow{p} 0$ and that the ordinary $t$ -statistic for $\alpha = 0$ is asymptotically normal:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}} \overset{a}{\sim} N(0,1)$$

- If the second model is correctly specified, then $t \xrightarrow{p} \infty$, since $\hat{\alpha}$ tends in

probability to 1, while it's estimated standard error tends to zero. Thus the test will always reject the false null model, asymptotically, since the statistic will eventually exceed any critical value with probability one.

- We can reverse the roles of the models, testing the second against the first.

- It may be the case that *neither* model is correctly specified. In this case, the test will still reject the null hypothesis, asymptotically, if we use critical values from the $N(0,1)$ distribution, since as long as $\hat{\alpha}$ tends to something different from zero, $|t| \xrightarrow{p} \infty$. Of course, when we switch the roles of the models the other will also be rejected asymptotically.

- In summary, there are 4 possible outcomes when we test two models, each against the other. Both may be rejected, neither may be rejected, or one of the two may be rejected.

- There are other tests available for non-nested models. The $J-$ test is simple to apply when both models are linear in the parameters. The $P$-test is

similar, but easier to apply when $M_1$ is nonlinear.

- The above presentation assumes that the same transformation of the dependent variable is used by both models. MacKinnon, White and Davidson, *Journal of Econometrics*, (1983) shows how to deal with the case of different transformations.

- Monte-Carlo evidence shows that these tests often over-reject a correctly specified model. Can use bootstrap critical values to get better-performing tests.

# Chapter 10

# Generalized least squares

Recall the assumptions of the classical linear regression model, in Section 4.6. One of the assumptions we've made up to now is that

$$\varepsilon_t \sim IID(0, \sigma^2)$$

or occasionally

$$\varepsilon_t \sim IIN(0, \sigma^2).$$

Now we'll investigate the consequences of non-identically and/or dependently distributed errors. We'll assume fixed regressors for now, to keep the presentation simple, and later we'll look at the consequences of relaxing this admittedly unrealistic assumption. The model is

$$\begin{aligned} y &= X\beta + \varepsilon \\ \mathcal{E}(\varepsilon) &= 0 \\ V(\varepsilon) &= \Sigma \end{aligned}$$

where $\Sigma$ is a general symmetric positive definite matrix (we'll write $\beta$ in place of $\beta_0$ to simplify the typing of these notes).

- The case where $\Sigma$ is a diagonal matrix gives uncorrelated, non-identically distributed errors. This is known as *heteroscedasticity*: $\exists i, j \; s.t. \; V(\epsilon_i) \neq V(\epsilon_j)$

- The case where $\Sigma$ has the same number on the main diagonal but nonzero

elements off the main diagonal gives identically (assuming higher moments are also the same) dependently distributed errors. This is known as *auto-correlation*: $\exists i \neq j \; s.t. \; E(\epsilon_i \epsilon_j) \neq 0$)

- The general case combines heteroscedasticity and autocorrelation. This is known as "non-spherical" disturbances, though why this term is used, I have no idea. Perhaps it's because under the classical assumptions, a joint confidence region for $\varepsilon$ would be an $n-$ dimensional hypersphere.

## 10.1 Effects of non-spherical disturbances on the OLS estimator

The least square estimator is

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ &= \beta + (X'X)^{-1}X'\varepsilon \end{aligned}$$

- We have unbiasedness, as before.

- The variance of $\hat{\beta}$ is

$$\begin{aligned} \mathcal{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] &= \mathcal{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} \end{aligned} \tag{10.1}$$

  Due to this, any test statistic that is based upon an estimator of $\sigma^2$ is invalid, since there *isn't* any $\sigma^2$, it doesn't exist as a feature of the true d.g.p. In particular, the formulas for the $t$, $F$, $\chi^2$ based tests given above do not lead to statistics with these distributions.

- $\hat{\beta}$ is still consistent, following exactly the same argument given before.

- If $\varepsilon$ is normally distributed, then

$$\hat{\beta} \sim N\left(\beta, (X'X)^{-1}X'\Sigma X(X'X)^{-1}\right)$$

The problem is that $\Sigma$ is unknown in general, so this distribution won't be useful for testing hypotheses.

- Without normality, and with stochastic $X$ (e.g., weakly exogenous regressors) we still have

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta\right) &= \sqrt{n}(X'X)^{-1}X'\varepsilon \\ &= \left(\frac{X'X}{n}\right)^{-1} n^{-1/2}X'\varepsilon\end{aligned}$$

Define the limiting variance of $n^{-1/2}X'\varepsilon$ (supposing a CLT applies) as

$$\lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) = \Omega, \text{ a.s.}$$

so we obtain $\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right)$. Note that the true asymptotic distribution of the OLS has changed with respect to the results under the classical assumptions. If we neglect to take this into account, the Wald

and score tests will not be asymptotically valid. So we need to figure out *how* to take it into account.

To see the invalidity of test procedures that are correct under the classical assumptions, when we have non-spherical errors, consider the Julia script GLS/EffectsOLS.jl. This script does a Monte Carlo study, generating data that are either heteroscedastic or homoscedastic, and then computes the empirical rejection frequency of a nominally 10% t-test. When the data are heteroscedastic, we obtain something like what we see in Figure 10.1. This sort of heteroscedasticity causes us to reject a true null hypothesis regarding the slope parameter much too often. You can experiment with the script to look at the effects of other sorts of HET, and to vary the sample size.

Figure 10.1: Rejection frequency of 10% t-test, H0 is true.

**Summary**: OLS with heteroscedasticity and/or autocorrelation is:

- unbiased with fixed or strongly exogenous regressors
- biased with weakly exogenous regressors
- has a different variance than before, so the previous test statistics aren't valid
- is consistent
- is asymptotically normally distributed, but with a different limiting covariance matrix. Previous test statistics aren't valid in this case for this reason.
- is inefficient, as is shown below.

## 10.2 The GLS estimator

Suppose $\Sigma$ were known. Then one could form the Cholesky decomposition

$$P'P = \Sigma^{-1}$$

Here, $P$ is an upper triangular matrix. We have

$$P'P\Sigma = I_n$$

so

$$P'P\Sigma P' = P',$$

which implies that

$$P\Sigma P' = I_n$$

Let's take some time to play with the Cholesky decomposition. Try out the GLS/cholesky.jl Julia script to see that the above claims are true, and also to see

how one can generate data from a $N(0, V)$ distribution.

Consider the model

$$Py = PX\beta + P\varepsilon,$$

or, making the obvious definitions,

$$y^* = X^*\beta + \varepsilon^*.$$

This variance of $\varepsilon^* = P\varepsilon$ is

$$\begin{aligned} \mathcal{E}(P\varepsilon\varepsilon' P') &= P\Sigma P' \\ &= I_n \end{aligned}$$

Therefore, the model

$$\begin{aligned} y^* &= X^*\beta + \varepsilon^* \\ \mathcal{E}(\varepsilon^*) &= 0 \\ V(\varepsilon^*) &= I_n \end{aligned}$$

satisfies the classical assumptions. The GLS estimator is simply OLS applied to the transformed model:

$$\begin{aligned} \hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\ &= (X'P'PX)^{-1}X'P'Py \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \end{aligned}$$

The GLS estimator is unbiased in the same circumstances under which the

OLS estimator is unbiased. For example, assuming $X$ is nonstochastic

$$\begin{aligned}
\mathcal{E}(\hat{\beta}_{GLS}) &= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y\right\} \\
&= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(X\beta+\varepsilon\right\} \\
&= \beta.
\end{aligned}$$

To get the variance of the estimator, we have

$$\begin{aligned}
\hat{\beta}_{GLS} &= (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*} \\
&= (X^{*\prime}X^{*})^{-1}X^{*\prime}\left(X^{*}\beta+\varepsilon^{*}\right) \\
&= \beta+(X^{*\prime}X^{*})^{-1}X^{*\prime}\varepsilon^{*}
\end{aligned}$$

so

$$
\begin{aligned}
\mathcal{E}\left\{\left(\hat{\beta}_{GLS}-\beta\right)\left(\hat{\beta}_{GLS}-\beta\right)'\right\} &= \mathcal{E}\left\{(X^{*\prime}X^{*})^{-1}X^{*\prime}\varepsilon^{*}\varepsilon^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1}\right\} \\
&= (X^{*\prime}X^{*})^{-1}X^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1} \\
&= (X^{*\prime}X^{*})^{-1} \\
&= (X'\Sigma^{-1}X)^{-1}
\end{aligned}
$$

Either of these last formulas can be used.

- All the previous results regarding the desirable properties of the least squares estimator hold, when dealing with the transformed model, since the transformed model satisfies the classical assumptions..

- Tests are valid, using the previous formulas, as long as we substitute $X^{*}$ in place of $X$. Furthermore, any test that involves $\sigma^2$ can set it to 1. This is preferable to re-deriving the appropriate formulas.

- The GLS estimator is more efficient than the OLS estimator. This is a

consequence of the Gauss-Markov theorem, since the GLS estimator is based on a model that satisfies the classical assumptions but the OLS estimator is not. To see this directly, note that

$$\begin{aligned} Var(\hat{\beta}) - Var(\hat{\beta}_{GLS}) &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} - (X'\Sigma^{-1}X)^{-1} \\ &= A\Sigma A' \end{aligned}$$

where $A = \left[(X'X)^{-1}X' - (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\right]$. This may not seem obvious, but it is true, as you can verify for yourself. Then noting that $A\Sigma A'$ is a quadratic form in a positive definite matrix, we conclude that $A\Sigma A'$ is positive semi-definite, and that GLS is efficient relative to OLS.

- As one can verify by calculating first order conditions, the GLS estimator is the solution to the minimization problem

$$\hat{\beta}_{GLS} = \arg\min(y - X\beta)'\Sigma^{-1}(y - X\beta)$$

so the *metric* $\Sigma^{-1}$ is used to weight the residuals.

## 10.3 Feasible GLS

The problem is that $\Sigma$ ordinarily isn't known, so this estimator isn't available.

- Consider the dimension of $\Sigma$ : it's an $n \times n$ matrix with $(n^2 - n)/2 + n = (n^2 + n)/2$ unique elements (remember - it is symmetric, because it's a covariance matrix).

- The number of parameters to estimate is larger than $n$ and increases faster than $n$. There's no way to devise an estimator that satisfies a LLN without adding restrictions.

- The *feasible GLS estimator* is based upon making sufficient assumptions regarding the form of $\Sigma$ so that a consistent estimator can be devised.

Suppose that we *parameterize* $\Sigma$ as a function of $X$ and $\theta$, where $\theta$ may include $\beta$ as well as other parameters, so that

$$\Sigma = \Sigma(X, \theta)$$

where $\theta$ is of fixed dimension. If we can consistently estimate $\theta$, we can consistently estimate $\Sigma$, as long as the elements of $\Sigma(X, \theta)$ are continuous functions of $\theta$ (by the Slutsky theorem). In this case,

$$\widehat{\Sigma} = \Sigma(X, \hat{\theta}) \xrightarrow{p} \Sigma(X, \theta)$$

If we replace $\Sigma$ in the formulas for the GLS estimator with $\widehat{\Sigma}$, we obtain the FGLS estimator. **The FGLS estimator shares the same asymptotic properties as GLS. These are**

1. Consistency

2. Asymptotic normality

3. Asymptotic efficiency *if* the errors are normally distributed. (Cramér-Rao).

4. Test procedures are asymptotically valid.

**In practice, the usual way to proceed is**

1. Define a consistent estimator of $\theta$. This is a case-by-case proposition, depending on the parameterization $\Sigma(\theta)$. We'll see examples below.

2. Form $\widehat{\Sigma} = \Sigma(X, \hat{\theta})$

3. Calculate the Cholesky factorization $\widehat{P} = Chol(\hat{\Sigma}^{-1})$.

4. Transform the model using

$$\hat{P}y = \hat{P}X\beta + \hat{P}\varepsilon$$

5. Estimate using OLS on the transformed model.

## 10.4 Heteroscedasticity

Heteroscedasticity is the case where

$$\mathcal{E}(\varepsilon\varepsilon') = \Sigma$$

is a diagonal matrix, so that the errors are uncorrelated, but have different variances. Heteroscedasticity is usually thought of as associated with cross sectional data, though there is absolutely no reason why time series data cannot also be heteroscedastic. Actually, the popular ARCH (autoregressive conditionally heteroscedastic) models explicitly assume that a time series is conditionally heteroscedastic.

Consider a supply function

$$q_i = \beta_1 + \beta_p P_i + \beta_s S_i + \varepsilon_i$$

where $P_i$ is price and $S_i$ is some measure of size of the $i^{th}$ firm. One might suppose

that unobservable factors (e.g., talent of managers, degree of coordination between production units, *etc.*) account for the error term $\varepsilon_i$. If there is more variability in these factors for large firms than for small firms, then $\varepsilon_i$ may have a higher variance when $S_i$ is high than when it is low.

Another example, individual demand.

$$q_i = \beta_1 + \beta_p P_i + \beta_m M_i + \varepsilon_i$$

where $P$ is price and $M$ is income. In this case, $\varepsilon_i$ can reflect variations in preferences. There are more possibilities for expression of preferences when one is rich, so it is possible that the variance of $\varepsilon_i$ could be higher when $M$ is high.

*Add example of group means.*

# OLS with heteroscedastic consistent varcov estimation

Eicker (1967) and White (1980) showed how to modify test statistics to account for heteroscedasticity of unknown form. The OLS estimator has asymptotic distribution

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right)$$

as we've already seen. Recall that we defined

$$\lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) = \Omega$$

This matrix has dimension $K \times K$ and can be consistently estimated, even if we can't estimate $\Sigma$ consistently. The consistent estimator, under heteroscedasticity but no autocorrelation is

$$\widehat{\Omega} = \frac{1}{n}\sum_{t=1}^{n} x_t x_t' \hat{\varepsilon}_t^2$$

One can then modify the previous test statistics to obtain tests that are valid when there is heteroscedasticity of unknown form. For example, the Wald test for

$H_0 : R\beta - r = 0$ would be

$$n\left(R\hat{\beta} - r\right)'\left(R\left(\frac{X'X}{n}\right)^{-1}\hat{\Omega}\left(\frac{X'X}{n}\right)^{-1}R'\right)^{-1}\left(R\hat{\beta} - r\right) \overset{a}{\sim} \chi^2(q)$$

# Detection

There exist many tests for the presence of heteroscedasticity. We'll discuss three methods.

**Goldfeld-Quandt** The sample is divided in to three parts, with $n_1, n_2$ and $n_3$ observations, where $n_1 + n_2 + n_3 = n$. The model is estimated using the first and third parts of the sample, separately, so that $\hat{\beta}^1$ and $\hat{\beta}^3$ will be independent. Then we have

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^1}{\sigma^2} = \frac{\varepsilon^{1\prime}M^1\varepsilon^1}{\sigma^2} \xrightarrow{d} \chi^2(n_1 - K)$$

and

$$\frac{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^3}{\sigma^2} = \frac{\varepsilon^{3\prime} M^3 \varepsilon^3}{\sigma^2} \xrightarrow{d} \chi^2(n_3 - K)$$

so

$$\frac{\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^1/(n_1 - K)}{\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^3/(n_3 - K)} \xrightarrow{d} F(n_1 - K, n_3 - K).$$

The distributional result is exact if the errors are normally distributed. This test is a two-tailed test. Alternatively, and probably more conventionally, if one has prior ideas about the possible magnitudes of the variances of the observations, one could order the observations accordingly, from largest to smallest. In this case, one would use a conventional one-tailed F-test. *Draw picture.*

- Ordering the observations is an important step if the test is to have any power.

- The motive for dropping the middle observations is to increase the difference between the average variance in the subsamples, supposing that there exists heteroscedasticity. This can increase the power of the test. On the other

hand, dropping too many observations will substantially increase the variance of the statistics $\hat{\varepsilon}^{1\prime}\hat{\varepsilon}^{1}$ and $\hat{\varepsilon}^{3\prime}\hat{\varepsilon}^{3}$. A rule of thumb, based on Monte Carlo experiments is to drop around 25% of the observations.

- If one doesn't have any ideas about the form of the het. the test will probably have low power since a sensible data ordering isn't available.

**White's test** When one has little idea if there exists heteroscedasticity, and no idea of its potential form, the White test is a possibility. The idea is that if there is homoscedasticity, then

$$\mathcal{E}(\varepsilon_t^2|x_t) = \sigma^2, \forall t$$

so that $x_t$ or functions of $x_t$ shouldn't help to explain $\mathcal{E}(\varepsilon_t^2)$. The test works as follows:

1. Since $\varepsilon_t$ isn't available, use the consistent estimator $\hat{\varepsilon}_t$ instead.

2. Regress

$$\hat{\varepsilon}_t^2 = \sigma^2 + z_t'\gamma + v_t$$

where $z_t$ is a $P$-vector. $z_t$ may include some or all of the variables in $x_t$, as well as other variables. White's original suggestion was to use $x_t$, plus the set of all unique squares and cross products of variables in $x_t$.

3. Test the hypothesis that $\gamma = 0$. The $qF$ statistic in this case is

$$qF = \frac{P\left(ESS_R - ESS_U\right)/P}{ESS_U/\left(n - P - 1\right)}$$

Note that $ESS_R = TSS_U$, so dividing both numerator and denominator by this we get

$$qF = \left(n - P - 1\right)\frac{R^2}{1 - R^2}$$

Note that this is the $R^2$ of the artificial regression used to test for heteroscedasticity, not the $R^2$ of the original model.

An asymptotically equivalent statistic, under the null of no heteroscedasticity (so that $R^2$ should tend to zero), is

$$nR^2 \overset{a}{\sim} \chi^2(P).$$

This doesn't require normality of the errors, though it does assume that the fourth moment of $\varepsilon_t$ is constant, under the null. **Question**: why is this necessary?

- The White test has the disadvantage that it may not be very powerful unless the $z_t$ vector is chosen well, and this is hard to do without knowledge of the form of heteroscedasticity.

- It also has the problem that specification errors other than heteroscedasticity may lead to rejection.

- Note: the null hypothesis of this test may be interpreted as $\theta = 0$ for the variance model $V(\varepsilon_t^2) = h(\alpha + z_t'\theta)$, where $h(\cdot)$ is an arbitrary function

of unknown form. The test is more general than is may appear from the regression that is used.

**Plotting the residuals** A very simple method is to simply plot the residuals (or their squares). *Draw pictures here.* Like the Goldfeld-Quandt test, this will be more informative if the observations are ordered according to the suspected form of the heteroscedasticity.

## Correction

Correcting for heteroscedasticity requires that a parametric form for $\Sigma(\theta)$ be supplied, and that a means for estimating $\theta$ consistently be determined. The estimation method will be specific to the assumed form $\Sigma(\theta)$. In recent years, there has been a trend toward simply estimating by OLS, and using robust standard errors. This may be somewhat unfortunate, as the weighted least squares estimator (GLS when there is only HET) is still consistent even if the specification of $\Sigma(\theta)$ is in-

correct, and it may be a good deal more efficient than OLS. Also, robust standard errors don't always work so well.

**Example 23.** The GRETL script GLS/Heteroscedasticity.inp illustrates these points.

Perhaps a middle ground is to attempt to use GLS when severe HET is detected, but to continue to use robust standard errors, to account for misspecifications in the modeling of $\Sigma(\theta)$.

We'll consider two examples. Before this, let's consider the general nature of GLS when there is heteroscedasticity. When we have HET but no AUT, $\Sigma$ is a diagonal matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ \vdots & \sigma_2^2 & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}$$

Likewise, $\Sigma^{-1}$ is diagonal

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ \vdots & \frac{1}{\sigma_2^2} & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_n^2} \end{bmatrix}$$

and so is the Cholesky decomposition $P = chol(\Sigma^{-1})$

$$P = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ \vdots & \frac{1}{\sigma_2} & & \vdots \\ & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_n} \end{bmatrix}$$

We need to transform the model, just as before, in the general case:

$$Py = PX\beta + P\varepsilon,$$

or, making the obvious definitions,

$$y^* = X^*\beta + \varepsilon^*.$$

Note that multiplying by $P$ just divides the data for each observation $(y_i, x_i)$ by the corresponding standard error of the error term, $\sigma_i$. That is, $y_i^* = y_i/\sigma_i$ and $x_i^* = x_i/\sigma_i$ (note that $x_i$ is a $K$-vector: we divided each element, including the 1 corresponding to the constant).

This makes sense. Consider Figure 10.2, which shows a true regression line with heteroscedastic errors. Which sample is more informative about the location of the line? The ones with observations with smaller variances. So, the GLS solution is equivalent to OLS on the transformed data. By the transformed data is the original data, weighted by the inverse of the standard error of the observation's error term. When the standard error is small, the weight is high, and vice versa. The GLS correction for the case of HET is also known as weighted least squares, for this reason.

Figure 10.2: Motivation for GLS correction when there is HET

## Multiplicative heteroscedasticity

Suppose the model is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \sigma_t^2 &= \mathcal{E}(\varepsilon_t^2) = (z_t'\gamma)^\delta \end{aligned}$$

but the other classical assumptions hold. In this case

$$\varepsilon_t^2 = (z_t'\gamma)^\delta + v_t$$

and $v_t$ has mean zero. Nonlinear least squares could be used to estimate $\gamma$ and $\delta$ consistently, were $\varepsilon_t$ observable. The solution is to substitute the squared OLS residuals $\hat{\varepsilon}_t^2$ in place of $\varepsilon_t^2$, since it is consistent by the Slutsky theorem. Once we have $\hat{\gamma}$ and $\hat{\delta}$, we can estimate $\sigma_t^2$ consistently using

$$\hat{\sigma}_t^2 = (z_t'\hat{\gamma})^{\hat{\delta}} \xrightarrow{p} \sigma_t^2 \ .$$

In the second step, we transform the model by dividing by the standard deviation:

$$\frac{y_t}{\hat{\sigma}_t} = \frac{x_t'\beta}{\hat{\sigma}_t} + \frac{\varepsilon_t}{\hat{\sigma}_t}$$

or

$$y_t^* = x_t^{*\prime}\beta + \varepsilon_t^*.$$

Asymptotically, this model satisfies the classical assumptions.

- This model is a bit complex in that NLS is required to estimate the model of the variance. A simpler version would be

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \sigma_t^2 &= \mathcal{E}(\varepsilon_t^2) = \sigma^2 z_t^\delta \end{aligned}$$

  where $z_t$ is a single variable. There are still two parameters to be estimated, and the model of the variance is still nonlinear in the parameters. However, the *search method* can be used in this case to reduce the estimation problem

to repeated applications of OLS.

- First, we define an interval of reasonable values for $\delta$, e.g., $\delta \in [0, 3]$.
- Partition this interval into $M$ equally spaced values, e.g., $\{0, .1, .2, ..., 2.9, 3\}$.
- For each of these values, calculate the variable $z_t^{\delta_m}$.
- The regression

$$\hat{\varepsilon}_t^2 = \sigma^2 z_t^{\delta_m} + v_t$$

  is linear in the parameters, conditional on $\delta_m$, so one can estimate $\sigma^2$ by OLS.
- Save the pairs $(\sigma_m^2, \delta_m)$, and the corresponding $ESS_m$. Choose the pair with the minimum $ESS_m$ as the estimate.
- Next, divide the model by the estimated standard deviations.
- Can refine. *Draw picture.*

- Works well when the parameter to be searched over is low dimensional, as in this case.

## Groupwise heteroscedasticity

A common case is where we have repeated observations on each of a number of economic agents: e.g., 10 years of macroeconomic data on each of a set of countries or regions, or daily observations of transactions of 200 banks. This sort of data is a *pooled cross-section time-series model.* It may be reasonable to presume that the variance is constant over time within the cross-sectional units, but that it differs across them (e.g., firms or countries of different sizes...). The model is

$$\begin{aligned} y_{it} &= x_{it}'\beta + \varepsilon_{it} \\ \mathcal{E}(\varepsilon_{it}^2) &= \sigma_i^2, \forall t \end{aligned}$$

where $i = 1, 2, ..., G$ are the agents, and $t = 1, 2, ..., n$ are the observations on each agent.

- The other classical assumptions are presumed to hold.
- In this case, the variance $\sigma_i^2$ is specific to each agent, but constant over the $n$ observations for that agent.
- In this model, we assume that $\mathcal{E}(\varepsilon_{it}\varepsilon_{is}) = 0$. This is a strong assumption that we'll relax later.

To correct for heteroscedasticity, just estimate each $\sigma_i^2$ using the natural estimator:

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{t=1}^{n} \hat{\varepsilon}_{it}^2$$

- Note that we use $1/n$ here since it's possible that there are more than $n$ regressors, so $n - K$ could be negative. Asymptotically the difference is unimportant.
- With each of these, transform the model as usual:

$$\frac{y_{it}}{\hat{\sigma}_i} = \frac{x_{it}'\beta}{\hat{\sigma}_i} + \frac{\varepsilon_{it}}{\hat{\sigma}_i}$$

Do this for each cross-sectional group. This transformed model satisfies the classical assumptions, asymptotically.

## Example: the Nerlove model

- Here's the data in Gretl format: nerlove.gdt
- estimate the basic Nerlove model by OLS, using Gretl, and plot the residuals: evidence of HET and AUT
- include square of $\ln Q$, now there's no AUT, but still HET. Conclusion: apparent AUT may be evidence of misspecification, rather than true autocorrelation.
- estimate using HET correction, and compare standard error estimates.

## 10.5 Autocorrelation

Autocorrelation, which is the serial correlation of the error term, is a problem that is usually associated with time series data, but also can affect cross-sectional data. For example, a shock to oil prices will simultaneously affect all countries, so one could expect contemporaneous correlation of macroeconomic variables across countries.

### Example

Consider the Keeling-Whorf data on atmospheric CO2 concentrations an Mauna Loa, Hawaii (see http://en.wikipedia.org/wiki/Keeling_Curve and http://cdiac.ornl.gov/ftp/ndp001/maunaloa.txt).

From the file maunaloa.txt: "THE DATA FILE PRESENTED IN THIS SUB-DIRECTORY CONTAINS MONTHLY AND ANNUAL ATMOSPHERIC CO2 CONCENTRATIONS DERIVED FROM THE SCRIPPS INSTITUTION OF

OCEANOGRAPHY'S (SIO's) CONTINUOUS MONITORING PROGRAM AT MAUNA LOA OBSERVATORY, HAWAII. THIS RECORD CONSTITUTES THE LONGEST CONTINUOUS RECORD OF ATMOSPHERIC CO2 CONCENTRATIONS AVAILABLE IN THE WORLD. MONTHLY AND ANNUAL AVERAGE MOLE FRACTIONS OF CO2 IN WATER-VAPOR-FREE AIR ARE GIVEN FROM MARCH 1958 THROUGH DECEMBER 2003, EXCEPT FOR A FEW INTERRUPTIONS."

The data is available in Octave format at CO2.data .

If we fit, using this script , the model $CO2_t = \beta_1 + \beta_2 t + \epsilon_t$, we get the results

```
octave:8> CO2Example
warning: load: file found in load path

*****************************************************
OLS estimation results
Observations 468
R-squared 0.979239
Sigma-squared 5.696791

Results (Het. consistent var-cov estimator)

     estimate      st.err.      t-stat.      p-value
1     316.918        0.227     1394.406        0.000
```

```
2        0.121        0.001      141.521        0.000

*****************************************************
```

It seems pretty clear that CO2 concentrations have been going up in the last 50 years, surprise, surprise. Let's look at a residual plot for the last 3 years of the data, see Figure 10.3. Note that there is a very predictable pattern. This is pretty strong evidence that the errors of the model are not independent of one another, which means there seems to be autocorrelation.

- this data is clearly nonstationary. The very large t-statistics that you get from OLS are suspicious, no?
- What is the limit of $X'X/n$ when there is a time trend in the regressor matrix?

Figure 10.3: Residuals from time trend for CO2 data

## Causes

Autocorrelation is the existence of correlation across the error term:

$$\mathcal{E}(\varepsilon_t \varepsilon_s) \neq 0, t \neq s.$$

Why might this occur? Plausible explanations include

1. Lags in adjustment to shocks. In a model such as

$$y_t = x_t' \beta + \varepsilon_t,$$

   one could interpret $x_t' \beta$ as the equilibrium value. Suppose $x_t$ is constant over a number of observations. One can interpret $\varepsilon_t$ as a shock that moves the system away from equilibrium. If the time needed to return to equilibrium is long with respect to the observation frequency, one could expect $\varepsilon_{t+1}$ to be positive, conditional on $\varepsilon_t$ positive, which induces a correlation.

2. Unobserved factors that are correlated over time. The error term is often

assumed to correspond to unobservable factors. If these factors are correlated over time, there will be autocorrelation.

3. Misspecification of the model. Suppose that the DGP is

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \varepsilon_t$$

but we estimate

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

The effects are illustrated in Figure 10.4.

4. Neglecting to include dynamics in a model. Lags of the dependent variable may be relevant regressors, and if they are omitted, their effects go into the error term, which will be autocorrelated.

Figure 10.4: Autocorrelation induced by misspecification

## Effects on the OLS estimator

The variance of the OLS estimator is the same as in the case of heteroscedasticity - the standard formula does not apply. The correct formula is given in equation 10.1. Next we discuss two GLS corrections for OLS. This sort of solution may lead to inconsistent estimation of betas in some cases, and it has definitely gone completely out of style. The standard procedure is to include enough lags of the dependent variable so that detectable AUT disappears, and then to use robust covariance estimation to take care of residual effects (see section 10.5). For reference, a couple of examples of the old-fashioned GLS corrections follow, but I will not discuss this in class.

## AR(1)

There are many types of autocorrelation. We'll consider two examples. The first is the most commonly encountered case: autoregressive order 1 (AR(1) errors. The

model is

$$
\begin{aligned}
y_t &= x_t'\beta + \varepsilon_t \\
\varepsilon_t &= \rho\varepsilon_{t-1} + u_t \\
u_t &\sim iid(0, \sigma_u^2) \\
\mathcal{E}(\varepsilon_t u_s) &= 0, t < s
\end{aligned}
$$

We assume that the model satisfies the other classical assumptions.

- We need a stationarity assumption: $|\rho| < 1$. Otherwise the variance of $\varepsilon_t$ explodes as $t$ increases, so standard asymptotics will not apply.

- By recursive substitution we obtain

$$\begin{aligned}
\varepsilon_t &= \rho\varepsilon_{t-1} + u_t \\
&= \rho\left(\rho\varepsilon_{t-2} + u_{t-1}\right) + u_t \\
&= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t \\
&= \rho^2\left(\rho\varepsilon_{t-3} + u_{t-2}\right) + \rho u_{t-1} + u_t
\end{aligned}$$

In the limit the lagged $\varepsilon$ drops out, since $\rho^m \to 0$ as $m \to \infty$, so we obtain

$$\varepsilon_t = \sum_{m=0}^{\infty} \rho^m u_{t-m}$$

With this, the variance of $\varepsilon_t$ is found as

$$\begin{aligned}
\mathcal{E}(\varepsilon_t^2) &= \sigma_u^2 \sum_{m=0}^{\infty} \rho^{2m} \\
&= \frac{\sigma_u^2}{1-\rho^2}
\end{aligned}$$

- If we had directly assumed that $\varepsilon_t$ were covariance stationary, we could obtain this using

$$\begin{aligned} V(\varepsilon_t) &= \rho^2 \mathcal{E}(\varepsilon_{t-1}^2) + 2\rho \mathcal{E}(\varepsilon_{t-1} u_t) + \mathcal{E}(u_t^2) \\ &= \rho^2 V(\varepsilon_t) + \sigma_u^2, \end{aligned}$$

so

$$V(\varepsilon_t) = \frac{\sigma_u^2}{1 - \rho^2}$$

- The variance is the $0^{th}$ order autocovariance: $\gamma_0 = V(\varepsilon_t)$

- Note that the variance does not depend on $t$

Likewise, the first order autocovariance $\gamma_1$ is

$$\begin{aligned} Cov(\varepsilon_t, \varepsilon_{t-1}) &= \gamma_s = \mathcal{E}((\rho\varepsilon_{t-1} + u_t)\,\varepsilon_{t-1}) \\ &= \rho V(\varepsilon_t) \\ &= \frac{\rho\sigma_u^2}{1-\rho^2} \end{aligned}$$

- Using the same method, we find that for $s < t$

$$Cov(\varepsilon_t, \varepsilon_{t-s}) = \gamma_s = \frac{\rho^s\sigma_u^2}{1-\rho^2}$$

- The autocovariances don't depend on $t$: the process $\{\varepsilon_t\}$ is *covariance stationary*

The *correlation (*in general, for r.v.'s $x$ and $y$) is defined as

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\text{se}(x)\text{se}(y)}$$

but in this case, the two standard errors are the same, so the $s$-order autocorrelation $\rho_s$ is

$$\rho_s = \rho^s$$

- All this means that the overall matrix $\Sigma$ has the form

$$\Sigma = \underbrace{\frac{\sigma_u^2}{1-\rho^2}}_{\text{this is the variance}} \underbrace{\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & & \ddots & & \vdots \\ & & & \ddots & \rho \\ \rho^{n-1} & \cdots & & & 1 \end{bmatrix}}_{\text{this is the correlation matrix}}$$

  So we have homoscedasticity, but elements off the main diagonal are not zero. All of this depends only on two parameters, $\rho$ and $\sigma_u^2$. If we can estimate these consistently, we can apply FGLS.

It turns out that it's easy to estimate these consistently. The steps are

1. Estimate the model $y_t = x_t'\beta + \varepsilon_t$ by OLS.

2. Take the residuals, and estimate the model

$$\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t^*$$

   Since $\hat{\varepsilon}_t \xrightarrow{p} \varepsilon_t$, this regression is asymptotically equivalent to the regression

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

   which satisfies the classical assumptions. Therefore, $\hat{\rho}$ obtained by applying OLS to $\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t^*$ is consistent. Also, since $u_t^* \xrightarrow{p} u_t$, the estimator

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{t=2}^{n}(\hat{u}_t^*)^2 \xrightarrow{p} \sigma_u^2$$

3. With the consistent estimators $\hat{\sigma}_u^2$ and $\hat{\rho}$, form $\hat{\Sigma} = \Sigma(\hat{\sigma}_u^2, \hat{\rho})$ using the previous structure of $\Sigma$, and estimate by FGLS. Actually, one can omit the

factor $\hat{\sigma}_u^2/(1-\rho^2)$, since it cancels out in the formula

$$\hat{\beta}_{FGLS} = \left(X'\hat{\Sigma}^{-1}X\right)^{-1}(X'\hat{\Sigma}^{-1}y).$$

- One can iterate the process, by taking the first FGLS estimator of $\beta$, re-estimating $\rho$ and $\sigma_u^2$, etc. If one iterates to convergences it's equivalent to MLE (supposing normal errors).

- An asymptotically equivalent approach is to simply estimate the transformed model

$$y_t - \hat{\rho}y_{t-1} = (x_t - \hat{\rho}x_{t-1})'\beta + u_t^*$$

using $n-1$ observations (since $y_0$ and $x_0$ aren't available). This is the method of Cochrane and Orcutt. Dropping the first observation is asymptotically irrelevant, but *it can be very important in small samples.* One

can recuperate the first observation by putting

$$
\begin{aligned}
y_1^* &= y_1 \sqrt{1-\hat{\rho}^2} \\
x_1^* &= x_1 \sqrt{1-\hat{\rho}^2}
\end{aligned}
$$

This somewhat odd-looking result is related to the Cholesky factorization of $\Sigma^{-1}$. See Davidson and MacKinnon, pg. 348-49 for more discussion. Note that the variance of $y_1^*$ is $\sigma_u^2$, asymptotically, so we see that the transformed model will be homoscedastic (and nonautocorrelated, since the $u's$ are uncorrelated with the $y's$, in different time periods.

# MA(1)

The linear regression model with moving average order 1 errors is

$$\begin{aligned} y_t &= x_t'\beta + \varepsilon_t \\ \varepsilon_t &= u_t + \phi u_{t-1} \\ u_t &\sim iid(0, \sigma_u^2) \\ \mathcal{E}(\varepsilon_t u_s) &= 0, t < s \end{aligned}$$

In this case,

$$\begin{aligned} V(\varepsilon_t) = \gamma_0 &= \mathcal{E}\left[(u_t + \phi u_{t-1})^2\right] \\ &= \sigma_u^2 + \phi^2 \sigma_u^2 \\ &= \sigma_u^2 (1 + \phi^2) \end{aligned}$$

Similarly

$$\begin{aligned}\gamma_1 &= \mathcal{E}\left[\left(u_t + \phi u_{t-1}\right)\left(u_{t-1} + \phi u_{t-2}\right)\right] \\ &= \phi\sigma_u^2\end{aligned}$$

and

$$\begin{aligned}\gamma_2 &= \left[\left(u_t + \phi u_{t-1}\right)\left(u_{t-2} + \phi u_{t-3}\right)\right] \\ &= 0\end{aligned}$$

so in this case

$$\Sigma = \sigma_u^2 \begin{bmatrix} 1+\phi^2 & \phi & 0 & \cdots & 0 \\ \phi & 1+\phi^2 & \phi & & \\ 0 & \phi & \ddots & & \vdots \\ \vdots & & & \ddots & \phi \\ 0 & \cdots & & \phi & 1+\phi^2 \end{bmatrix}$$

Note that the first order autocorrelation is

$$\begin{aligned}\rho_1 &= \tfrac{\phi\sigma_u^2}{\sigma_u^2(1+\phi^2)} = \frac{\gamma_1}{\gamma_0} \\ &= \frac{\phi}{(1+\phi^2)}\end{aligned}$$

- This achieves a maximum at $\phi = 1$ and a minimum at $\phi = -1$, and the maximal and minimal autocorrelations are 1/2 and -1/2. Therefore, series that are more strongly autocorrelated can't be MA(1) processes.

Again the covariance matrix has a simple structure that depends on only two parameters. The problem in this case is that one can't estimate $\phi$ using OLS on

$$\hat{\varepsilon}_t = u_t + \phi u_{t-1}$$

because the $u_t$ are unobservable and they can't be estimated consistently. However, there is a simple way to estimate the parameters.

- Since the model is homoscedastic, we can estimate

$$V(\varepsilon_t) = \sigma_\varepsilon^2 = \sigma_u^2(1+\phi^2)$$

using the typical estimator:

$$\widehat{\sigma_\varepsilon^2} = \sigma_u^2\widehat{(1+\phi^2)} = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2$$

- By the Slutsky theorem, we can interpret this as defining an (unidentified) estimator of both $\sigma_u^2$ and $\phi$, e.g., use this as

$$\widehat{\sigma_u^2}(1+\widehat{\phi^2}) = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t^2$$

However, this isn't sufficient to define consistent estimators of the parameters, since it's unidentified - two unknowns, one equation.

- To solve this problem, estimate the covariance of $\varepsilon_t$ and $\varepsilon_{t-1}$ using

$$\widehat{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \widehat{\phi\sigma_u^2} = \frac{1}{n}\sum_{t=2}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}$$

  This is a consistent estimator, following a LLN (and given that the epsilon hats are consistent for the epsilons). As above, this can be interpreted as defining an unidentified estimator of the two parameters:

$$\hat{\phi}\widehat{\sigma_u^2} = \frac{1}{n}\sum_{t=2}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}$$

- Now solve these two equations to obtain identified (and therefore consistent) estimators of both $\phi$ and $\sigma_u^2$. Define the consistent estimator

$$\hat{\Sigma} = \Sigma(\hat{\phi}, \widehat{\sigma_u^2})$$

  following the form we've seen above, and transform the model using the Cholesky decomposition. The transformed model satisfies the classical as-

sumptions asymptotically.

- Note: there is no guarantee that $\Sigma$ estimated using the above method will be positive definite, which may pose a problem. Another method would be to use ML estimation, if one is willing to make distributional assumptions regarding the white noise errors.

## Monte Carlo example: AR1

(too lazy to convert the code to Julia)

Let's look at a Monte Carlo study that compares OLS and GLS when we have AR1 errors. The model is

$$y_t = 1 + x_t + \epsilon_t$$
$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

with $\rho = 0.9$. The sample size is $n = 30$, and 1000 Monte Carlo replications are

Figure 10.5: Efficiency of OLS and FGLS, AR1 errors

done. The Octave script is GLS/AR1Errors.jl. Figure 10.5 shows histograms of the estimated coefficient of $x$ (the true value is 1) . We can see that the GLS histogram is much more concentrated about 1, which is indicative of the efficiency of GLS relative to OLS.

# Asymptotically valid inferences with autocorrelation of unknown form

See Hamilton Ch. 10, pp. 261-2 and 280-84.

When the form of autocorrelation is unknown, one may decide to use the OLS estimator, without correction. We've seen that this estimator has the limiting distribution

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right)$$

where, as before, $\Omega$ is

$$\Omega = \lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon' X}{n}\right)$$

We need a consistent estimate of $\Omega$. Define $m_t = x_t\varepsilon_t$ (recall that $x_t$ is defined as

a $K \times 1$ vector). Note that

$$\begin{aligned} X'\varepsilon &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \sum_{t=1}^{n} x_t \varepsilon_t \\ &= \sum_{t=1}^{n} m_t \end{aligned}$$

so that

$$\Omega = \lim_{n \to \infty} \frac{1}{n} \mathcal{E} \left[ \left( \sum_{t=1}^{n} m_t \right) \left( \sum_{t=1}^{n} m_t' \right) \right]$$

We assume that $m_t$ is covariance stationary (so that the covariance between $m_t$ and $m_{t-s}$ does not depend on $t$).

Define the $v - th$ autocovariance of $m_t$ as

$$\Gamma_v = \mathcal{E}(m_t m_{t-v}').$$

Note that $\mathcal{E}(m_t m'_{t+v}) = \Gamma'_v$. *(show this with an example).* In general, we expect that:

- $m_t$ will be autocorrelated, since $\varepsilon_t$ is potentially autocorrelated:

$$\Gamma_v = \mathcal{E}(m_t m'_{t-v}) \neq 0$$

  Note that this autocovariance does not depend on $t$, due to covariance stationarity.

- contemporaneously correlated ( $\mathcal{E}(m_{it} m_{jt}) \neq 0$ ), since the regressors in $x_t$ will in general be correlated (more on this later).

- and heteroscedastic ($\mathcal{E}(m_{it}^2) = \sigma_i^2$ , which depends upon $i$ ), again since the regressors will have different variances.

While one could estimate $\Omega$ parametrically, we in general have little information upon which to base a parametric specification. Recent research has focused on consistent nonparametric estimators of $\Omega$.

Now define

$$\Omega_n = \mathcal{E}\frac{1}{n}\left[\left(\sum_{t=1}^{n} m_t\right)\left(\sum_{t=1}^{n} m_t'\right)\right]$$

We have (*show that the following is true, by expanding sum and shifting rows to left)*

$$\Omega_n = \Gamma_0 + \frac{n-1}{n}\left(\Gamma_1 + \Gamma_1'\right) + \frac{n-2}{n}\left(\Gamma_2 + \Gamma_2'\right)\cdots + \frac{1}{n}\left(\Gamma_{n-1} + \Gamma_{n-1}'\right)$$

The natural, consistent estimator of $\Gamma_v$ is

$$\widehat{\Gamma_v} = \frac{1}{n}\sum_{t=v+1}^{n} \hat{m}_t \hat{m}_{t-v}'.$$

where

$$\hat{m}_t = x_t \hat{\varepsilon}_t$$

(note: one could put $1/(n-v)$ instead of $1/n$ here). So, a natural, but inconsistent,

estimator of $\Omega_n$ would be

$$\begin{aligned}\hat{\Omega}_n &= \widehat{\Gamma_0} + \frac{n-1}{n}\left(\widehat{\Gamma_1} + \widehat{\Gamma'_1}\right) + \frac{n-2}{n}\left(\widehat{\Gamma_2} + \widehat{\Gamma'_2}\right) + \cdots + \frac{1}{n}\left(\widehat{\Gamma_{n-1}} + \widehat{\Gamma'_{n-1}}\right) \\ &= \widehat{\Gamma_0} + \sum_{v=1}^{n-1} \frac{n-v}{n}\left(\widehat{\Gamma_v} + \widehat{\Gamma'_v}\right).\end{aligned}$$

This estimator is inconsistent in general, since the number of parameters to estimate is more than the number of observations, and increases more rapidly than $n$, so information does not build up as $n \to \infty$.

On the other hand, supposing that $\Gamma_v$ tends to zero sufficiently rapidly as $v$ tends to $\infty$, a modified estimator

$$\hat{\Omega}_n = \widehat{\Gamma_0} + \sum_{v=1}^{q(n)} \left(\widehat{\Gamma_v} + \widehat{\Gamma'_v}\right),$$

where $q(n) \xrightarrow{p} \infty$ as $n \to \infty$ will be consistent, provided $q(n)$ grows sufficiently slowly.

- The assumption that autocorrelations die off is reasonable in many cases.

For example, the AR(1) model with $|\rho| < 1$ has autocorrelations that die off.

- The term $\frac{n-v}{n}$ can be dropped because it tends to one for $v < q(n)$, given that $q(n)$ increases slowly relative to $n$.

- A disadvantage of this estimator is that is may not be positive definite. This could cause one to calculate a negative $\chi^2$ statistic, for example!

- Newey and West proposed and estimator (*Econometrica*, 1987) that solves the problem of possible nonpositive definiteness of the above estimator. Their estimator is

$$\hat{\Omega}_n = \widehat{\Gamma}_0 + \sum_{v=1}^{q(n)} \left[1 - \frac{v}{q+1}\right] \left(\widehat{\Gamma}_v + \widehat{\Gamma}_v'\right).$$

  This estimator is p.d. by construction. The condition for consistency is that $n^{-1/4}q(n) \to 0$. Note that this is a very slow rate of growth for $q$. This estimator is nonparametric - we've placed no parametric restrictions on the form of $\Omega$. It is an example of a *kernel* estimator.

Finally, since $\Omega_n$ has $\Omega$ as its limit, $\hat{\Omega}_n \xrightarrow{p} \Omega$. We can now use $\hat{\Omega}_n$ and $\widehat{Q_X} = \frac{1}{n}X'X$ to consistently estimate the limiting distribution of the OLS estimator under heteroscedasticity and autocorrelation of unknown form. With this, asymptotically valid tests are constructed in the usual way.

# Testing for autocorrelation

## Breusch-Godfrey test

This test uses an auxiliary regression, as does the White test for heteroscedasticity. The regression is

$$\hat{\varepsilon}_t = x_t'\delta + \gamma_1\hat{\varepsilon}_{t-1} + \gamma_2\hat{\varepsilon}_{t-2} + \cdots + \gamma_P\hat{\varepsilon}_{t-P} + v_t$$

and the test statistic is the $nR^2$ statistic, just as in the White test. There are $P$ restrictions, so the test statistic is asymptotically distributed as a $\chi^2(P)$.

- The intuition is that the lagged errors shouldn't contribute to explaining the

current error if there is no autocorrelation.

- $x_t$ is included as a regressor to account for the fact that the $\hat{\varepsilon}_t$ are not independent even if the $\varepsilon_t$ are. This is a technicality that we won't go into here.

- This test is valid even if the regressors are stochastic and contain lagged dependent variables, so it is considerably more useful than the DW test for typical time series data.

- The alternative is not that the model is an AR(P), following the argument above. The alternative is simply that some or all of the first $P$ autocorrelations are different from zero. This is compatible with many specific forms of autocorrelation.

**Durbin-Watson test**

The Durbin-Watson test is not strictly valid in most situations where we would like to use it. Nevertheless, it is encountered often enough so that one should know

something about it (perhaps: I'm no longer teaching this, but I'll leave it in the notes for reference). The Durbin-Watson test statistic is

$$\begin{aligned} DW &= \frac{\Sigma_{t=2}^{n}\left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\Sigma_{t=1}^{n}\hat{\varepsilon}_t^2} \\ &= \frac{\Sigma_{t=2}^{n}\left(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2\right)}{\Sigma_{t=1}^{n}\hat{\varepsilon}_t^2} \end{aligned}$$

- The null hypothesis is that the first order autocorrelation of the errors is zero: $H_0 : \rho_1 = 0$. The alternative is of course $H_A : \rho_1 \neq 0$. Note that the alternative is not that the errors are AR(1), since many general patterns of autocorrelation will have the first order autocorrelation different than zero. For this reason the test is useful for detecting autocorrelation in general. For the same reason, one shouldn't just assume that an AR(1) model is appropriate when the DW test rejects the null.

- Under the null, the middle term tends to zero, and the other two tend to one, so $DW \xrightarrow{p} 2$.

- Supposing that we had an AR(1) error process with $\rho = 1$. In this case the middle term tends to $-2$, so $DW \xrightarrow{p} 0$

- Supposing that we had an AR(1) error process with $\rho = -1$. In this case the middle term tends to 2, so $DW \xrightarrow{p} 4$

- These are the extremes: $DW$ always lies between 0 and 4.

- The distribution of the test statistic depends on the matrix of regressors, $X$, so tables can't give exact critical values. The give upper and lower bounds, which correspond to the extremes that are possible. See Figure 10.6. There are means of determining exact critical values conditional on $X$.

- Note that DW can be used to test for nonlinearity (add discussion).

- The DW test is based upon the assumption that the matrix $X$ is fixed in repeated samples. This is often unreasonable in the context of economic time series, which is precisely the context where the test would have application.

It is possible to relate the DW test to other test statistics which are valid without strict exogeneity.

## Lagged dependent variables and autocorrelation

We've seen that the OLS estimator is consistent under autocorrelation, as long as $plim\frac{X'\varepsilon}{n} = 0$. This will be the case when $\mathcal{E}(X'\varepsilon) = 0$, following a LLN. An important exception is the case where $X$ contains lagged $y's$ and the errors are autocorrelated.

**Example 24.** Dynamic model with MA1 errors. Consider the model

$$\begin{aligned} y_t &= \alpha + \rho y_{t-1} + \beta x_t + \epsilon_t \\ \epsilon_t &= \upsilon_t + \phi \upsilon_{t-1} \end{aligned}$$

Figure 10.6: Durbin-Watson critical values

We can easily see that a regressor is not weakly exogenous:

$$\begin{aligned}\mathcal{E}(y_{t-1}\varepsilon_t) &= \mathcal{E}\left\{(\alpha + \rho y_{t-2} + \beta x_{t-1} + \upsilon_{t-1} + \phi\upsilon_{t-2})(\upsilon_t + \phi\upsilon_{t-1})\right\} \\ &\neq 0\end{aligned}$$

since one of the terms is $\mathcal{E}(\phi\upsilon_{t-1}^2)$ which is clearly nonzero. In this case $\mathcal{E}(\mathbf{x}_t\varepsilon_t) \neq 0$, and therefore $plim\frac{X'\varepsilon}{n} \neq 0$. Since

$$plim\hat{\beta} = \beta + plim\frac{X'\varepsilon}{n}$$

the OLS estimator is inconsistent in this case. One needs to estimate by instrumental variables (IV), which we'll get to later

The Octave (reminder to self: translate this?) script GLS/DynamicMA.m does a Monte Carlo study. The sample size is $n = 100$. The true coefficients are $\alpha = 1$ $\rho = 0.9$ and $\beta = 1$. The MA parameter is $\phi = -0.95$. Figure 10.7 gives the results. You can see that the constant and the autoregressive parameter have a

lot of bias. By re-running the script with $\phi = 0$, you will see that much of the bias disappears (not all - why?).

## 10.6 Exercises

1. Consider the following model with $n$ observations and only one explanatory variable:

$$y_i = \beta x_i + u_i \ (1)$$

which satisfies the basic OLS assumptions except for the homoskedasticity assumption. More precisely, the variance of the error term for the first $n/2$ observations is $Var(u_i) = 2$, and for the remaining observations $Var(u_i) = \sigma^2$, where $\sigma^2$ is unknown.

a) How can you get an asymptotically efficient estimator of $\beta$? Provide an expression for the estimator.

b) How can you test the hypothesis that the explanatory variable ($x_i$) is relevant in explaining $y_i$ in the model in equation (1)? Provide the statistic to be used in the test as well as its distribution.

2. Comparing the variances of the OLS and GLS estimators, I claimed that the

Figure 10.7: Dynamic model with MA(1) errors

(a) $\hat{\alpha} - \alpha$

(b) $\hat{\rho} - \rho$

(c) $\hat{\beta} - \beta$

following holds:

$$Var(\hat{\beta}) - Var(\hat{\beta}_{GLS}) \;=\; A\Sigma A^{'}$$

Verify that this is true.

3. Show that the GLS estimator can be defined as

$$\hat{\beta}_{GLS} = \arg\min(y - X\beta)'\Sigma^{-1}(y - X\beta)$$

4. The limiting distribution of the OLS estimator with heteroscedasticity of unknown form is

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, Q_X^{-1}\Omega Q_X^{-1}\right),$$

where

$$\lim_{n\to\infty} \mathcal{E}\left(\frac{X'\varepsilon\varepsilon'X}{n}\right) = \Omega$$

Explain why

$$\widehat{\Omega} = \frac{1}{n} \sum_{t=1}^{n} x_t x_t' \hat{\varepsilon}_t^2$$

is a consistent estimator of this matrix.

5. Define the $v-th$ autocovariance of a covariance stationary process $m_t$, where $E(m_t) = 0$ as

$$\Gamma_v = \mathcal{E}(m_t m_{t-v}').$$

Show that $\mathcal{E}(m_t m_{t+v}') = \Gamma_v'$.

6. Perhaps we can be a little more parsimonious with the Nerlove data (nerlove.data ), rather than using so many parameters to account for non-constant returns to scale, and to account for heteroscedasticity. Consider the original model

$$\ln C = \beta + \beta_Q \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon$$

   (a) Estimate by OLS, plot the residuals, and test for autocorrelation and heteroscedasticity. Explain your findings.

(b) Consider the model

$$\ln C = \beta + \beta_Q \ln Q + \gamma_Q (\ln Q)^2 + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon$$

i. Explain how this model can account for non-constant returns to scale.

ii. estimate this model, and test for autocorrelation and heteroscedasticity. You should find that there is HET, but no strong evidence of AUT. Why is this the case?

iii. Do a GLS correction where it is assumed that $V(\epsilon_i) = \frac{\sigma^2}{(\ln Q_i)^2}$. In GRETL, there is a weighted least squares option that you can use. Why does this assumed form of HET make sense?

iv. plot the weighted residuals versus output. Is there evidence of HET, or has the correction eliminated the problem?

v. plot the fitted values for returns to scale, for all of the firms.

7. The hall.csv or hall.gdt dataset contains monthly observation on 3 variables: the consumption ratio $c_t/c_{t-1}$; the gross return of an equally weighted index of assets $ewr_t$; and the gross return of the same index, but weighted by value, $vwr_t$. The idea is that a representative consumer may finance consumption by investing in assets. Present wealth is used for two things: consumption and investment. The return on investment defines wealth in the next period, and the process repeats. For the moment, explore the properties of the variables.

   (a) Are the variances constant over time?

   (b) Do the variables appear to be autocorrelated? Hint: regress a variable on its own lags.

   (c) Do the variable seem to be normally distributed?

   (d) Look at the properties of the growth rates of the variables: repeat a-c for growth rates. The growth rate of a variable $x_t$ is given by $\log(x_t/x_{t-1})$.

8. Consider the model

$$
\begin{aligned}
y_t &= C + A_1 y_{t-1} + \epsilon_t \\
E(\epsilon_t \epsilon_t') &= \Sigma \\
E(\epsilon_t \epsilon_s') &= 0, t \neq s
\end{aligned}
$$

where $y_t$ and $\epsilon_t$ are $G \times 1$ vectors, $C$ is a $G \times 1$ of constants, and $A_1$is a $G \times G$ matrix of parameters. The matrix $\Sigma$ is a $G \times G$ covariance matrix. Assume that we have $n$ observations. This is a *vector autoregressive* model, of order 1 - commonly referred to as a VAR(1) model.

(a) Show how the model can be written in the form $Y = X\beta + \nu$, where $Y$ is a $Gn \times 1$ vector, $\beta$ is a $(G + G^2) \times 1$ parameter vector, and the other items are conformable. What is the structure of $X$? What is the structure of the covariance matrix of $\nu$?

(b) This model has HET and AUT. Verify this statement.

(c) Set $G = 2$,$C = (0\,0)'A = \begin{bmatrix} 0.8 & -0.1 \\ 0.2 & 0.5 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. Simulate data from this model, then estimate the model using OLS and feasible GLS. You should find that the two estimators are identical, which might seem surprising, given that there is HET and AUT.

(d) (optional, and advanced). Prove analytically that the OLS and GLS estimators are identical. Hint: this model is of the form of *seemingly unrelated regressions*.

9. Consider the model

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t$$

where $\epsilon_t$ is a $N(0,1)$ white noise error. This is an autogressive model of order 2 (AR2) model. Suppose that data is generated from the AR2 model, but the econometrician mistakenly decides to estimate an AR1 model ($y_t =$

$\alpha + \rho_1 y_{t-1} + \epsilon_t)$.

(a) simulate data from the AR2 model, setting $\rho_1 = 0.5$ and $\rho_2 = 0.4$, using a sample size of $n = 30$.

(b) Estimate the AR1 model by OLS, using the simulated data

(c) test the hypothesis that $\rho_1 = 0.5$

(d) test for autocorrelation using the test of your choice

(e) repeat the above steps 10000 times.

   i. What percentage of the time does a t-test reject the hypothesis that $\rho_1 = 0.5$?

   ii. What percentage of the time is the hypothesis of no autocorrelation rejected?

(f) discuss your findings. Include a residual plot for a representative sample.

10. Modify the script given in Subsection 10.5 so that the first observation is dropped, rather than given special treatment. This corresponds to using the Cochrane-Orcutt method, whereas the script as provided implements the Prais-Winsten method. Check if there is an efficiency loss when the first observation is dropped.

# Chapter 11

# Endogeneity and simultaneity

Several times we've encountered cases where correlation between regressors and the error term lead to biasedness and inconsistency of the OLS estimator. Cases include autocorrelation with lagged dependent variables (Example 24), measurement error in the regressors (Example 20) and missing regressors (Section 8.4).

Figure 11.1: Exogeneity and Endogeneity (adapted from Cameron and Trivedi)

Another important case we have not seen yet is that of simultaneous equations. The cause is different, but the effect is the same: bias and inconsistency when OLS is applied to a single equation. The basic idea is presented in Figure 11.1. A simple regression will estimate the overall effect of x on y. If we're interested in the direct effect, $\beta$, then we have a problem when the overall effect and the direct effect differ.

## 11.1 Simultaneous equations

Up until now our model is

$$y = X\beta + \varepsilon$$

where we assume weak exogeneity of the regressors, so that $E(x_t\epsilon_t) = 0$. With weak exogeneity, the OLS estimator has desirable large sample properties (consistency, asymptotic normality).

Simultaneous equations is a different prospect. An example of a simultaneous equation system is a simple supply-demand system:

$$\begin{aligned}
\text{Demand: } q_t &= \alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t} \qquad (11.1)\\
\text{Supply: } q_t &= \beta_1 + \beta_2 p_t + \varepsilon_{2t}\\
\mathcal{E}\left(\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} & \varepsilon_{2t} \end{bmatrix}\right) &= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \cdot & \sigma_{22} \end{bmatrix}\\
&\equiv \Sigma, \forall t
\end{aligned}$$

The presumption is that $q_t$ and $p_t$ are jointly determined at the same time by the intersection of these equations. We'll assume that $y_t$ is determined by some unrelated process.

It's easy to see that we have correlation between regressors and errors. Solving for $p_t$ :

$$\begin{aligned}
\alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t} &= \beta_1 + \beta_2 p_t + \varepsilon_{2t} \\
\beta_2 p_t - \alpha_2 p_t &= \alpha_1 - \beta_1 + \alpha_3 y_t + \varepsilon_{1t} - \varepsilon_{2t} \\
p_t &= \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}
\end{aligned}$$

Now consider whether $p_t$ is uncorrelated with $\varepsilon_{1t}$ :

$$\begin{aligned}
\mathcal{E}(p_t \varepsilon_{1t}) &= \mathcal{E}\left\{\left(\frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\varepsilon_{1t}\right\} \\
&= \frac{\sigma_{11} - \sigma_{12}}{\beta_2 - \alpha_2}
\end{aligned}$$

Because of this correlation, weak exogeneity does not hold, and OLS estimation of the demand equation will be biased and inconsistent. The same applies to the supply equation, for the same reason.

A GRETL script which generates data according to this simple supply-demand

system is here: Simeq/simeq.inp. It does a Monte Carlo study which verifies that the OLS estimator is inconsistent, and that the IV estimator is consistent.

In this model, $q_t$ and $p_t$ are the *endogenous* variables (endogs), that are determined within the system. $y_t$ is an *exogenous* variable (exogs). These concepts are a bit tricky, and we'll return to it in a minute. First, some notation. Suppose we group together current endogs in the vector $Y_t$. If there are $G$ endogs, $Y_t$ is $G \times 1$. Group current and lagged exogs, as well as lagged endogs in the vector $X_t$ , which is $K \times 1$. Stack the errors of the $G$ equations into the error vector $E_t$. The model, with additional assumptions, can be written as

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t' \\ E_t &\sim N(0, \Sigma), \forall t \\ \mathcal{E}(E_t E_s') &= 0, t \neq s \end{aligned} \tag{11.2}$$

There are $G$ equations here, and the parameters that enter into each equation are contained in the *columns* of the matrices $\Gamma$ and $B$.

We can stack all $n$ observations and write the model as

$$\begin{aligned} Y\Gamma &= XB + E \\ \mathcal{E}(X'E) &= 0_{(K\times G)} \\ \text{vec}(E) &\sim N(0, \Psi) \end{aligned}$$

where

$$Y = \begin{bmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_n' \end{bmatrix}, X = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}, E = \begin{bmatrix} E_1' \\ E_2' \\ \vdots \\ E_n' \end{bmatrix}$$

$Y$ is $n \times G$, $X$ is $n \times K$, and $E$ is $n \times G$. Recall that the vec() operator just stacks the columns of a matrix on top of one another into a vector.

- This system is *complete*, in that there are as many equations as endogs.
- There is a normality assumption. This isn't necessary for obtaining con-

sistent estimates, but, if it holds, it allows us to consider the relationship between least squares and ML estimators.

- Since there is no autocorrelation of the $E_t$ 's, and since the columns of $E$ are individually homoscedastic, then

$$\begin{aligned}\Psi &= \begin{bmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \cdots & \sigma_{1G}I_n \\ & \sigma_{22}I_n & & \vdots \\ & & \ddots & \vdots \\ \cdot & & & \sigma_{GG}I_n \end{bmatrix} \\ &= I_n \otimes \Sigma\end{aligned}$$

- $X$ may contain lagged endogenous and exogenous variables. These variables are *predetermined.*

- We need to define what is meant by "endogenous" and "exogenous" when classifying the current period variables. Remember the definition of weak ex-

ogeneity Assumption 16, the regressors are weakly exogenous if $E(E_t|X_t) = 0$. Endogenous regressors are those for which this assumption does not hold. As long as there is no autocorrelation, lagged endogenous variables are weakly exogenous.

## 11.2 Reduced form

Recall that the model is

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t' \\ V(E_t) &= \Sigma \end{aligned}$$

This is the model in *structural form.*

**Definition 25.** [Structural form] An equation is in structural form when more than one current period endogenous variable is included.

The solution for the current period endogs is easy to find. It is

$$\begin{aligned} Y_t' &= X_t'B\Gamma^{-1} + E_t'\Gamma^{-1} \\ &= X_t'\Pi + V_t' \end{aligned}$$

Now only one current period endog appears in each equation. This is the *reduced form.*

**Definition 26.** [Reduced form] An equation is in reduced form if only one current period endog is included.

An example is our supply/demand system. The reduced form for quantity is obtained by solving the supply equation for price and substituting into demand:

$$
\begin{aligned}
q_t &= \alpha_1 + \alpha_2 \left( \frac{q_t - \beta_1 - \varepsilon_{2t}}{\beta_2} \right) + \alpha_3 y_t + \varepsilon_{1t} \\
\beta_2 q_t - \alpha_2 q_t &= \beta_2 \alpha_1 - \alpha_2 \left( \beta_1 + \varepsilon_{2t} \right) + \beta_2 \alpha_3 y_t + \beta_2 \varepsilon_{1t} \\
q_t &= \frac{\beta_2 \alpha_1 - \alpha_2 \beta_1}{\beta_2 - \alpha_2} + \frac{\beta_2 \alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\beta_2 \varepsilon_{1t} - \alpha_2 \varepsilon_{2t}}{\beta_2 - \alpha_2} \\
&= \pi_{11} + \pi_{21} y_t + V_{1t}
\end{aligned}
$$

Similarly, the rf for price, as we've seen above, is

$$\begin{aligned} p_t &= \frac{\alpha_1 - \beta_1}{\beta_2 - \alpha_2} + \frac{\alpha_3 y_t}{\beta_2 - \alpha_2} + \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2} \\ &= \pi_{12} + \pi_{22} y_t + V_{2t} \end{aligned}$$

The interesting thing about the rf is that the equations individually satisfy the classical assumptions, since $y_t$ is uncorrelated with $\varepsilon_{1t}$ and $\varepsilon_{2t}$ by assumption, and therefore $\mathcal{E}(y_t V_{it}) = 0$, i=1,2, $\forall t$. The errors of the rf are

$$\begin{bmatrix} V_{1t} \\ V_{2t} \end{bmatrix} = \begin{bmatrix} \frac{\beta_2 \varepsilon_{1t} - \alpha_2 \varepsilon_{2t}}{\beta_2 - \alpha_2} \\ \frac{\varepsilon_{1t} - \varepsilon_{2t}}{\beta_2 - \alpha_2} \end{bmatrix}$$

The variance of $V_{1t}$ is

$$\begin{aligned} V(V_{1t}) &= \mathcal{E}\left[\left(\frac{\beta_2 \varepsilon_{1t} - \alpha_2 \varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\left(\frac{\beta_2 \varepsilon_{1t} - \alpha_2 \varepsilon_{2t}}{\beta_2 - \alpha_2}\right)\right] \\ &= \frac{\beta_2^2 \sigma_{11} - 2\beta_2 \alpha_2 \sigma_{12} + \alpha_2 \sigma_{22}}{(\beta_2 - \alpha_2)^2} \end{aligned}$$

- This is constant over time, so the first rf equation is homoscedastic.
- Likewise, since the $\varepsilon_t$ are independent over time, so are the $V_t$.

The variance of the second rf error is

$$\begin{aligned} V(V_{2t}) &= \mathcal{E}\left[\left(\frac{\varepsilon_{1t}-\varepsilon_{2t}}{\beta_2-\alpha_2}\right)\left(\frac{\varepsilon_{1t}-\varepsilon_{2t}}{\beta_2-\alpha_2}\right)\right] \\ &= \frac{\sigma_{11}-2\sigma_{12}+\sigma_{22}}{(\beta_2-\alpha_2)^2} \end{aligned}$$

and the contemporaneous covariance of the errors across equations is

$$\begin{aligned} \mathcal{E}(V_{1t}V_{2t}) &= \mathcal{E}\left[\left(\frac{\beta_2\varepsilon_{1t}-\alpha_2\varepsilon_{2t}}{\beta_2-\alpha_2}\right)\left(\frac{\varepsilon_{1t}-\varepsilon_{2t}}{\beta_2-\alpha_2}\right)\right] \\ &= \frac{\beta_2\sigma_{11}-(\beta_2+\alpha_2)\,\sigma_{12}+\sigma_{22}}{(\beta_2-\alpha_2)^2} \end{aligned}$$

- In summary the rf equations individually satisfy the classical assumptions, under the assumptions we've made, but they are contemporaneously correlated.

The general form of the rf is

$$\begin{aligned} Y_t' &= X_t' B \Gamma^{-1} + E_t' \Gamma^{-1} \\ &= X_t' \Pi + V_t' \end{aligned}$$

so we have that

$$V_t = \left(\Gamma^{-1}\right)' E_t \sim N\left(0, \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}\right), \forall t$$

and that the $V_t$ are timewise independent (note that this wouldn't be the case if the $E_t$ were autocorrelated).

From the reduced form, we can easily see that the endogenous variables are correlated with the structural errors:

$$\begin{aligned} E(E_t Y_t') &= E\left(E_t\left(X_t' B \Gamma^{-1} + E_t' \Gamma^{-1}\right)\right) \\ &= E\left(E_t X_t' B \Gamma^{-1} + E_t E_t' \Gamma^{-1}\right) \\ &= \Sigma \Gamma^{-1} \end{aligned} \tag{11.3}$$

## 11.3 Estimation of the reduced form equations

From above, the RF equations are

$$\begin{aligned} Y_t' &= X_t' B \Gamma^{-1} + E_t' \Gamma^{-1} \\ &= X_t' \Pi + V_t' \end{aligned}$$

and

$$V_t \sim N\left(0, \Xi\right), \forall t$$

where we define $\Xi \equiv \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1}$. The rf parameter estimator $\hat{\Pi}$, is simply OLS applied to this model, equation by equation::

$$\hat{\Pi} = \left(X'X\right)^{-1} X'Y$$

which is simply

$$\hat{\Pi} = (X'X)^{-1}X' \begin{bmatrix} y_1 & y_2 & \cdots & y_G \end{bmatrix}$$

that is, OLS equation by equation using *all* the exogs in the estimation of each column of $\Pi$.

It may seem odd that we use OLS on the reduced form, since the rf equations are correlated, because $\Xi \equiv (\Gamma^{-1})' \Sigma \Gamma^{-1}$ is a full matrix. Why don't we do GLS to improve efficiency of estimation of the RF parameters?

OLS equation by equation to get the rf is equivalent to

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} = \begin{bmatrix} X & 0 & \cdots & 0 \\ 0 & X & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_G \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_G \end{bmatrix}$$

where $y_i$ is the $n \times 1$ vector of observations of the $i^{th}$ endog, $X$ is the entire $n \times K$ matrix of exogs, $\pi_i$ is the $i^{th}$ column of $\Pi$, and $v_i$ is the $i^{th}$ column of $V$. Use the

notation

$$y = \mathbf{X}\pi + v$$

to indicate the pooled model. Following this notation, the error covariance matrix is

$$V(v) = \Xi \otimes I_n$$

- This is a special case of a type of model known as a set of *seemingly unrelated equations (SUR)* since the parameter vector of each equation is different. The important feature of this special case is that *the regressors are the same in each equation.* The equations are contemporanously correlated, because of the non-zero off diagonal elements in $\Xi$.

- Note that each equation of the system individually satisfies the classical assumptions.

- Normally when doing SUR, one simply does GLS on the whole system $y = \mathbf{X}\pi + v$, where $V(v) = \Xi \otimes I_n$, which is in general more efficient than OLS

on each equation.

- However, when the regressors are the same in all equations, as is true in the present case of estimation of the RF parameters, SUR $\equiv$OLS. To show this note that in this case $\mathbf{X} = I_n \otimes X$. Using the rules

  1. $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$
  2. $(A \otimes B)' = (A' \otimes B')$ and

3. $(A \otimes B)(C \otimes D) = (AC \otimes BD)$, we get

$$
\begin{aligned}
\hat{\pi}_{SUR} &= \left((I_n \otimes X)' (\Xi \otimes I_n)^{-1} (I_n \otimes X)\right)^{-1} (I_n \otimes X)' (\Xi \otimes I_n)^{-1} y \\
&= \left((\Xi^{-1} \otimes X') (I_n \otimes X)\right)^{-1} (\Xi^{-1} \otimes X') y \\
&= \left(\Xi \otimes (X'X)^{-1}\right) (\Xi^{-1} \otimes X') y \\
&= \left[I_G \otimes (X'X)^{-1} X'\right] y \\
&= \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_G \end{bmatrix}
\end{aligned}
$$

- Note that this provides the answer to the exercise 8d in the chapter on GLS.
- So the unrestricted rf coefficients can be estimated efficiently (assuming normality) by OLS, even if the equations are correlated.
- We have ignored any potential zeros in the matrix $\Pi$, which if they exist

could potentially increase the efficiency of estimation of the rf.

- Another example where SUR≡OLS is in estimation of vector autoregressions which is discussed in Section 17.2.

## 11.4 Bias and inconsistency of OLS estimation of a structural equation

Considering the first equation (this is without loss of generality, since we can always reorder the equations) we can partition the $Y$ matrix as

$$Y = \begin{bmatrix} y & Y_1 & Y_2 \end{bmatrix}$$

- $y$ is the first column
- $Y_1$ are the other endogenous variables that enter the first equation
- $Y_2$ are endogs that are excluded from this equation

Similarly, partition $X$ as

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

- $X_1$ are the included exogs, and $X_2$ are the excluded exogs.

Finally, partition the error matrix as

$$E = \begin{bmatrix} \varepsilon & E_{12} \end{bmatrix}$$

Assume that $\Gamma$ has ones on the main diagonal. These are normalization restrictions that simply scale the remaining coefficients on each equation, and which scale the variances of the error terms.

Given this scaling and our partitioning, the coefficient matrices can be written

as

$$\Gamma = \begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \end{bmatrix}$$

$$B = \begin{bmatrix} \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

With this, the first equation can be written as

$$\begin{aligned} y &= Y_1\gamma_1 + X_1\beta_1 + \varepsilon \\ &= Z\delta + \varepsilon \end{aligned} \tag{11.4}$$

The problem, as we've seen, is that the columns of $Z$ corresponding to $Y_1$ are correlated with $\varepsilon$, because these are endogenous variables, and as we saw in equation 11.3, the endogenous variables are correlated with the structural errors, so they don't satisfy weak exogeneity. So, $E(Z'\epsilon) \neq 0$. What are the properties of the

OLS estimator in this situation?

$$\begin{aligned}\hat{\delta} &= (Z'Z)^{-1} Z'y \\ &= (Z'Z)^{-1} Z' (Z\delta^0 + \varepsilon) \\ &= \delta^0 + (Z'Z)^{-1} Z'\epsilon\end{aligned}$$

It's clear that the OLS estimator is biased in general. Also,

$$\hat{\delta} - \delta^0 = \left(\frac{Z'Z}{n}\right)^{-1} \frac{Z'\epsilon}{n}$$

Say that $\lim \frac{Z'\epsilon}{n} = A$,a.s., and $\lim \frac{Z'Z}{n} = Q_Z$, $a.s.$ Then

$$\lim \left(\hat{\delta} - \delta^0\right) = Q_Z^{-1} A \neq 0, \; a.s.$$

So the OLS estimator of a structural equation is inconsistent. In general, correlation between regressors and errors leads to this problem, whether due to measurement error, simultaneity, or omitted regressors.

A GRETL script which generates data according to a simple supply-demand system is here: Simeq/simeq.inp. It does a Monte Carlo study which verifies that the OLS estimator is inconsistent, and that the IV estimator is consistent.

## 11.5 Note about the rest of this chapter

In class, I will not teach the material in the rest of this chapter at this time, but instead we will go on to GMM. The material that follows is easier to understand in the context of GMM, where we get a nice unified theory.

## 11.6 Identification by exclusion restrictions

The material in the rest of this chapter is no longer used in classes, but I'm leaving it in the notes for reference.

The identification problem in simultaneous equations is in fact of the same

nature as the identification problem in any estimation setting: does the limiting objective function have the proper curvature so that there is a unique global minimum or maximum at the true parameter value? In the context of IV estimation, this is the case if the limiting covariance of the IV estimator is positive definite and $plim \frac{1}{n} W'\varepsilon = 0$. This matrix is

$$V_\infty(\hat{\beta}_{IV}) = (Q_{XW} Q_{WW}^{-1} Q'_{XW})^{-1} \sigma^2$$

- The necessary and sufficient condition for identification is simply that this matrix be positive definite, and that the instruments be (asymptotically) uncorrelated with $\varepsilon$.

- For this matrix to be positive definite, we need that the conditions noted above hold: $Q_{WW}$ must be positive definite and $Q_{XW}$ must be of full rank ( $K$ ).

- These identification conditions are not that intuitive nor is it very obvious how to check them.

## Necessary conditions

If we use IV estimation for a single equation of the system, the equation can be written as

$$y = Z\delta + \varepsilon$$

where

$$Z = \begin{bmatrix} Y_1 & X_1 \end{bmatrix}$$

**Notation:**

- Let $K$ be the total numer of weakly exogenous variables.
- Let $K^* = cols(X_1)$ be the number of included exogs, and let $K^{**} = K - K^*$ be the number of excluded exogs (in this equation).
- Let $G^* = cols(Y_1) + 1$ be the total number of included endogs, and let $G^{**} = G - G^*$ be the number of excluded endogs.

Using this notation, consider the selection of instruments.

- Now the $X_1$ are weakly exogenous and can serve as their own instruments.

- It turns out that $X$ exhausts the set of possible instruments, in that if the variables in $X$ don't lead to an identified model then no other instruments will identify the model either. Assuming this is true (we'll prove it in a moment), then a necessary condition for identification is that $cols(X_2) \geq cols(Y_1)$ since if not then at least one instrument must be used twice, so $W$ will not have full column rank:

$$\rho(W) < K^* + G^* - 1 \Rightarrow \rho(Q_{ZW}) < K^* + G^* - 1$$

This is the *order condition* for identification in a set of simultaneous equations. When the only identifying information is exclusion restrictions on the variables that enter an equation, then the number of excluded exogs must be greater than or equal to the number of included endogs, minus 1 (the

normalized lhs endog), e.g.,

$$K^{**} \geq G^{*} - 1$$

- To show that this is in fact a necessary condition consider some arbitrary set of instruments $W$. A necessary condition for identification is that

$$\rho\left(plim\frac{1}{n}W'Z\right) = K^{*} + G^{*} - 1$$

where

$$Z = \left[\, Y_1 \;\; X_1 \,\right]$$

Recall that we've partitioned the model

$$Y\Gamma = XB + E$$

as

$$Y = \left[\, y \;\; Y_1 \;\; Y_2 \,\right]$$

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

Given the reduced form

$$Y = X\Pi + V$$

we can write the reduced form using the same partition

$$\begin{bmatrix} y & Y_1 & Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \pi_{11} & \Pi_{12} & \Pi_{13} \\ \pi_{21} & \Pi_{22} & \Pi_{23} \end{bmatrix} + \begin{bmatrix} v & V_1 & V_2 \end{bmatrix}$$

so we have

$$Y_1 = X_1\Pi_{12} + X_2\Pi_{22} + V_1$$

so

$$\frac{1}{n}W'Z = \frac{1}{n}W' \begin{bmatrix} X_1\Pi_{12} + X_2\Pi_{22} + V_1 & X_1 \end{bmatrix}$$

Because the $W$ 's are uncorrelated with the $V_1$ 's, by assumption, the cross between

$W$ and $V_1$ converges in probability to zero, so

$$plim\frac{1}{n}W'Z = plim\frac{1}{n}W'\left[\ X_1\Pi_{12} + X_2\Pi_{22}\ \ X_1\ \right]$$

Since the far rhs term is formed only of linear combinations of columns of $X$, the rank of this matrix can never be greater than $K$, regardless of the choice of instruments. If $Z$ has more than $K$ columns, then it is not of full column rank. When $Z$ has more than $K$ columns we have

$$G^* - 1 + K^* > K$$

or noting that $K^{**} = K - K^*$,

$$G^* - 1 > K^{**}$$

In this case, the limiting matrix is not of full column rank, and the identification condition fails.

## Sufficient conditions

Identification essentially requires that the structural parameters be recoverable from the data. This won't be the case, in general, unless the structural model is subject to some restrictions. We've already identified necessary conditions. Turning to sufficient conditions (again, we're only considering identification through zero restricitions on the parameters, for the moment).

The model is

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t \\ V(E_t) &= \Sigma \end{aligned}$$

This leads to the reduced form

$$
\begin{aligned}
Y_t' &= X_t' B \Gamma^{-1} + E_t \Gamma^{-1} \\
&= X_t' \Pi + V_t \\
V(V_t) &= \left(\Gamma^{-1}\right)' \Sigma \Gamma^{-1} \\
&= \Omega
\end{aligned}
$$

The reduced form parameters are consistently estimable, but none of them are known *a priori,* and there are no restrictions on their values. The problem is that more than one structural form has the same reduced form, so knowledge of the reduced form parameters alone isn't enough to determine the structural parameters. To see this, consider the model

$$
\begin{aligned}
Y_t' \Gamma F &= X_t' B F + E_t F \\
V(E_t F) &= F' \Sigma F
\end{aligned}
$$

where $F$ is some arbirary nonsingular $G \times G$ matrix. The rf of this new model is

$$\begin{aligned} Y_t' &= X_t' B F \left(\Gamma F\right)^{-1} + E_t F \left(\Gamma F\right)^{-1} \\ &= X_t' B F F^{-1} \Gamma^{-1} + E_t F F^{-1} \Gamma^{-1} \\ &= X_t' B \Gamma^{-1} + E_t \Gamma^{-1} \\ &= X_t' \Pi + V_t \end{aligned}$$

Likewise, the covariance of the rf of the transformed model is

$$\begin{aligned} V(E_t F \left(\Gamma F\right)^{-1}) &= V(E_t \Gamma^{-1}) \\ &= \Omega \end{aligned}$$

Since the two structural forms lead to the same rf, and the rf is all that is directly estimable, the models are said to be *observationally equivalent.* What we need for identification are restrictions on $\Gamma$ and $B$ such that the only admissible $F$ is an identity matrix (if all of the equations are to be identified). Take the coefficient

matrices as partitioned before:

$$\begin{bmatrix} \Gamma \\ B \end{bmatrix} = \begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \\ \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

The coefficients of the first equation of the transformed model are simply these coefficients multiplied by the first column of $F$. This gives

$$\begin{bmatrix} \Gamma \\ B \end{bmatrix} \begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = \begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \\ \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} f_{11} \\ F_2 \end{bmatrix}$$

For identification of the first equation we need that there be enough restrictions so that the only admissible

$$\begin{bmatrix} f_{11} \\ F_2 \end{bmatrix}$$

be the leading column of an identity matrix, so that

$$\begin{bmatrix} 1 & \Gamma_{12} \\ -\gamma_1 & \Gamma_{22} \\ 0 & \Gamma_{32} \\ \beta_1 & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -\gamma_1 \\ 0 \\ \beta_1 \\ 0 \end{bmatrix}$$

Note that the third and fifth rows are

$$\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix} F_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Supposing that the leading matrix is of full column rank, e.g.,

$$\rho\left(\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix}\right) = cols\left(\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix}\right) = G - 1$$

then the only way this can hold, without additional restrictions on the model's parameters, is if $F_2$ is a vector of zeros. Given that $F_2$ is a vector of zeros, then the first equation

$$\begin{bmatrix} 1 & \Gamma_{12} \end{bmatrix} \begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = 1 \Rightarrow f_{11} = 1$$

Therefore, as long as

$$\rho\left(\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix}\right) = G - 1$$

then

$$\begin{bmatrix} f_{11} \\ F_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0_{G-1} \end{bmatrix}$$

The first equation is identified in this case, so the condition is sufficient for identification. It is also necessary, since the condition implies that this submatrix must have at least $G - 1$ rows. Since this matrix has

$$G^{**} + K^{**} = G - G^{*} + K^{**}$$

rows, we obtain

$$G - G^{*} + K^{**} \geq G - 1$$

or

$$K^{**} \geq G^{*} - 1$$

which is the previously derived necessary condition.

The above result is fairly intuitive (draw picture here). The necessary condition ensures that there are enough variables not in the equation of interest to potentially move the other equations, so as to trace out the equation of interest. The sufficient condition ensures that those other equations in fact do move around as the variables

change their values. Some points:

- When an equation has $K^{**} = G^{*} - 1$, is is *exactly identified*, in that omission of an identifiying restriction is not possible without loosing consistency.

- When $K^{**} > G^{*} - 1$, the equation is *overidentified*, since one could drop a restriction and still retain consistency. Overidentifying restrictions are therefore testable. When an equation is overidentified we have more instruments than are strictly necessary for consistent estimation. Since estimation by IV with more instruments is more efficient asymptotically, one should employ overidentifying restrictions if one is confident that they're true.

- We can repeat this partition for each equation in the system, to see which equations are identified and which aren't.

- These results are valid assuming that the only identifying information comes from knowing which variables appear in which equations, e.g., by exclusion

restrictions, and through the use of a normalization. There are other sorts of identifying information that can be used. These include

1. Cross equation restrictions
2. Additional restrictions on parameters within equations (as in the Klein model discussed below)
3. Restrictions on the covariance matrix of the errors
4. Nonlinearities in variables

- When these sorts of information are available, the above conditions aren't necessary for identification, though they are of course still sufficient.

To give an example of how other information can be used, consider the model

$$Y\Gamma = XB + E$$

where $\Gamma$ is an upper triangular matrix with 1's on the main diagonal. This is a *triangular system* of equations. In this case, the first equation is

$$y_1 = XB_{\cdot 1} + E_{\cdot 1}$$

Since only exogs appear on the rhs, this equation is identified.

The second equation is

$$y_2 = -\gamma_{21} y_1 + XB_{\cdot 2} + E_{\cdot 2}$$

This equation has $K^{**} = 0$ excluded exogs, and $G^* = 2$ included endogs, so it fails the order (necessary) condition for identification.

- However, suppose that we have the restriction $\Sigma_{21} = 0$, so that the first and second structural errors are uncorrelated. In this case

$$\mathcal{E}(y_{1t}\varepsilon_{2t}) = \mathcal{E}\left\{(X_t' B_{\cdot 1} + \varepsilon_{1t})\varepsilon_{2t}\right\} = 0$$

so there's no problem of simultaneity. If the entire $\Sigma$ matrix is diagonal, then following the same logic, all of the equations are identified. This is known as a *fully recursive* model.

## 11.7 2SLS

When we have no information regarding cross-equation restrictions or the structure of the error covariance matrix, one can estimate the parameters of a single equation of the system without regard to the other equations.

- This isn't always efficient, as we'll see, but it has the advantage that misspecifications in other equations will not affect the consistency of the estimator of the parameters of the equation of interest.

- Also, estimation of the equation won't be affected by identification problems in other equations.

The 2SLS estimator is very simple: it is the GIV estimator, using all of the weakly exogenous variables as instruments. In the first stage, each column of $Y_1$ is regressed on *all* the weakly exogenous variables in the system, e.g., the entire $X$ matrix. The fitted values are

$$\begin{aligned}\hat{Y}_1 &= X(X'X)^{-1}X'Y_1 \\ &= P_X Y_1 \\ &= X\hat{\Pi}_1\end{aligned}$$

Since these fitted values are the projection of $Y_1$ on the space spanned by $X$, and since any vector in this space is uncorrelated with $\varepsilon$ by assumption, $\hat{Y}_1$ is uncorrelated with $\varepsilon$. Since $\hat{Y}_1$ is simply the reduced-form prediction, it is correlated with $Y_1$, The only other requirement is that the instruments be linearly independent. This should be the case when the order condition is satisfied, since there are more columns in $X_2$ than in $Y_1$ in this case.

The second stage substitutes $\hat{Y}_1$ in place of $Y_1$, and estimates by OLS. This

original model is

$$\begin{aligned} y &= Y_1\gamma_1 + X_1\beta_1 + \varepsilon \\ &= Z\delta + \varepsilon \end{aligned}$$

and the second stage model is

$$y = \hat{Y}_1\gamma_1 + X_1\beta_1 + \varepsilon.$$

Since $X_1$ is in the space spanned by $X$, $P_X X_1 = X_1$, so we can write the second stage model as

$$\begin{aligned} y &= P_X Y_1\gamma_1 + P_X X_1\beta_1 + \varepsilon \\ &\equiv P_X Z\delta + \varepsilon \end{aligned}$$

The OLS estimator applied to this model is

$$\hat{\delta} = (Z'P_X Z)^{-1} Z'P_X y$$

which is exactly what we get if we estimate using IV, with the reduced form predictions of the endogs used as instruments. Note that if we define

$$\begin{aligned} \hat{Z} &= P_X Z \\ &= \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix} \end{aligned}$$

so that $\hat{Z}$ are the instruments for $Z$, then we can write

$$\hat{\delta} = (\hat{Z}'Z)^{-1} \hat{Z}'y$$

- Important note: OLS on the transformed model can be used to calculate the 2SLS estimate of $\delta$, since we see that it's equivalent to IV using a particular set of instruments. However *the OLS covariance formula is not valid.* We need to apply the IV covariance formula already seen above.

Actually, there is also a simplification of the general IV variance formula. Define

$$\begin{aligned}\hat{Z} &= P_X Z \\ &= \begin{bmatrix} \hat{Y} & X \end{bmatrix}\end{aligned}$$

The IV covariance estimator would ordinarily be

$$\hat{V}(\hat{\delta}) = \left(\hat{Z}'Z\right)^{-1}\left(\hat{Z}'\hat{Z}\right)\left(Z'\hat{Z}\right)^{-1}\hat{\sigma}^2_{IV}$$

However, looking at the first term in parentheses

$$\hat{Z}'Z = \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix}' \begin{bmatrix} Y_1 & X_1 \end{bmatrix} = \begin{bmatrix} Y_1'(P_X)Y_1 & Y_1'(P_X)X_1 \\ X_1'Y_1 & X_1'X_1 \end{bmatrix}$$

but since $P_X$ is idempotent and since $P_X X = X$, we can write

$$\begin{aligned}
\begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix}' \begin{bmatrix} Y_1 & X_1 \end{bmatrix} &= \begin{bmatrix} Y_1' P_X P_X Y_1 & Y_1' P_X X_1 \\ X_1' P_X Y_1 & X_1' X_1 \end{bmatrix} \\
&= \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix}' \begin{bmatrix} \hat{Y}_1 & X_1 \end{bmatrix} \\
&= \hat{Z}' \hat{Z}
\end{aligned}$$

Therefore, the first and second terms in the variance formula cancel, so the 2SLS varcov estimator simplifies to

$$\hat{V}(\hat{\delta}) = \left(Z' \hat{Z}\right)^{-1} \hat{\sigma}_{IV}^2$$

which, following some algebra similar to the above, can also be written as

$$\hat{V}(\hat{\delta}) = \left(\hat{Z}' \hat{Z}\right)^{-1} \hat{\sigma}_{IV}^2 \tag{11.5}$$

Finally, recall that though this is presented in terms of the first equation, it is

general, since any equation can be placed first.

**Properties of 2SLS:**

1. Consistent
2. Asymptotically normal
3. Biased when the mean esists (the existence of moments is a technical issue we won't go into here).
4. Asymptotically inefficient, except in special circumstances (more on this later).

## 11.8 Testing the overidentifying restrictions

The selection of which variables are endogs and which are exogs *is part of the specification of the model*. As such, there is room for error here: one might erroneously classify a variable as exog when it is in fact correlated with the error

term. A general test for the specification on the model can be formulated as follows:

The IV estimator can be calculated by applying OLS to the transformed model, so the IV objective function at the minimized value is

$$s(\hat{\beta}_{IV}) = \left(y - X\hat{\beta}_{IV}\right)' P_W \left(y - X\hat{\beta}_{IV}\right),$$

but

$$\begin{aligned}
\hat{\varepsilon}_{IV} &= y - X\hat{\beta}_{IV} \\
&= y - X(X'P_W X)^{-1} X' P_W y \\
&= \left(I - X(X'P_W X)^{-1} X' P_W\right) y \\
&= \left(I - X(X'P_W X)^{-1} X' P_W\right)(X\beta + \varepsilon) \\
&= A\,(X\beta + \varepsilon)
\end{aligned}$$

where

$$A \equiv I - X(X'P_W X)^{-1} X' P_W$$

so

$$s(\hat{\beta}_{IV}) = (\varepsilon' + \beta' X')\, A' P_W A\, (X\beta + \varepsilon)$$

Moreover, $A' P_W A$ is idempotent, as can be verified by multiplication:

$$\begin{aligned} A' P_W A &= \left(I - P_W X(X'P_W X)^{-1} X'\right) P_W \left(I - X(X'P_W X)^{-1} X' P_W\right) \\ &= \left(P_W - P_W X(X'P_W X)^{-1} X' P_W\right)\left(P_W - P_W X(X'P_W X)^{-1} X' P_W\right) \\ &= \left(I - P_W X(X'P_W X)^{-1} X'\right) P_W. \end{aligned}$$

Furthermore, $A$ is orthogonal to $X$

$$\begin{aligned} AX &= \left(I - X(X'P_W X)^{-1} X' P_W\right) X \\ &= X - X \\ &= 0 \end{aligned}$$

so

$$s(\hat{\beta}_{IV}) = \varepsilon' A' P_W A \varepsilon$$

Supposing the $\varepsilon$ are normally distributed, with variance $\sigma^2$, then the random variable

$$\frac{s(\hat{\beta}_{IV})}{\sigma^2} = \frac{\varepsilon' A' P_W A \varepsilon}{\sigma^2}$$

is a quadratic form of a $N(0,1)$ random variable with an idempotent matrix in the middle, so

$$\frac{s(\hat{\beta}_{IV})}{\sigma^2} \sim \chi^2(\rho(A' P_W A))$$

This isn't available, since we need to estimate $\sigma^2$. Substituting a consistent estimator,

$$\frac{s(\hat{\beta}_{IV})}{\widehat{\sigma^2}} \overset{a}{\sim} \chi^2(\rho(A' P_W A))$$

- Even if the $\varepsilon$ aren't normally distributed, the asymptotic result still holds. The last thing we need to determine is the rank of the idempotent matrix.

We have

$$A'P_W A = \left(P_W - P_W X(X'P_W X)^{-1} X'P_W\right)$$

so

$$\begin{aligned}
\rho(A'P_W A) &= Tr\left(P_W - P_W X(X'P_W X)^{-1} X'P_W\right) \\
&= TrP_W - TrX'P_W P_W X(X'P_W X)^{-1} \\
&= TrW(W'W)^{-1}W' - K_X \\
&= TrW'W(W'W)^{-1} - K_X \\
&= K_W - K_X
\end{aligned}$$

where $K_W$ is the number of columns of $W$ and $K_X$ is the number of columns of $X$. The degrees of freedom of the test is simply the number of overidentifying restrictions: the number of instruments we have beyond the number that is strictly necessary for consistent estimation.

- This test is an overall specification test: the joint null hypothesis is that

the model is correctly specified *and* that the $W$ form valid instruments (e.g., that the variables classified as exogs really are uncorrelated with $\varepsilon$. Rejection can mean that either the model $y = Z\delta + \varepsilon$ is misspecified, or that there is correlation between $X$ and $\varepsilon$.

- This is a particular case of the GMM criterion test, which is covered in the second half of the course. See Section 16.10.

- Note that since

$$\hat{\varepsilon}_{IV} = A\varepsilon$$

and

$$s(\hat{\beta}_{IV}) = \varepsilon' A' P_W A\varepsilon$$

we can write

$$\begin{aligned}
\frac{s(\hat{\beta}_{IV})}{\widehat{\sigma^2}} &= \frac{(\hat{\varepsilon}'W(W'W)^{-1}W')\,(W(W'W)^{-1}W'\hat{\varepsilon})}{\hat{\varepsilon}'\hat{\varepsilon}/n} \\
&= n(RSS_{\hat{\varepsilon}_{IV}|W}/TSS_{\hat{\varepsilon}_{IV}}) \\
&= nR_u^2
\end{aligned}$$

where $R_u^2$ is the uncentered $R^2$ from a regression of the $IV$ residuals on all of the instruments $W$. This is a convenient way to calculate the test statistic.

On an aside, consider IV estimation of a just-identified model, using the standard notation

$$y = X\beta + \varepsilon$$

and $W$ is the matrix of instruments. If we have exact identification then $cols(W) =$

$cols(X)$, so $W^{'}X$ is a square matrix. The transformed model is

$$P_W y = P_W X\beta + P_W \varepsilon$$

and the fonc are

$$X' P_W (y - X\hat{\beta}_{IV}) = 0$$

The IV estimator is

$$\hat{\beta}_{IV} = (X' P_W X)^{-1} X' P_W y$$

Considering the inverse here

$$\begin{aligned}
(X' P_W X)^{-1} &= \left(X' W (W'W)^{-1} W' X\right)^{-1} \\
&= (W'X)^{-1} \left(X' W (W'W)^{-1}\right)^{-1} \\
&= (W'X)^{-1} (W'W) \left(X'W\right)^{-1}
\end{aligned}$$

Now multiplying this by $X'P_W y$, we obtain

$$\begin{aligned}
\hat{\beta}_{IV} &= (W'X)^{-1}(W'W)\,(X'W)^{-1}\,X'P_W y \\
&= (W'X)^{-1}(W'W)\,(X'W)^{-1}\,X'W(W'W)^{-1}W'y \\
&= (W'X)^{-1}W'y
\end{aligned}$$

The objective function for the generalized IV estimator is

$$\begin{aligned}
s(\hat{\beta}_{IV}) &= \left(y - X\hat{\beta}_{IV}\right)' P_W \left(y - X\hat{\beta}_{IV}\right) \\
&= y'P_W\left(y - X\hat{\beta}_{IV}\right) - \hat{\beta}'_{IV}X'P_W\left(y - X\hat{\beta}_{IV}\right) \\
&= y'P_W\left(y - X\hat{\beta}_{IV}\right) - \hat{\beta}'_{IV}X'P_W y + \hat{\beta}'_{IV}X'P_W X\hat{\beta}_{IV} \\
&= y'P_W\left(y - X\hat{\beta}_{IV}\right) - \hat{\beta}'_{IV}\left(X'P_W y + X'P_W X\hat{\beta}_{IV}\right) \\
&= y'P_W\left(y - X\hat{\beta}_{IV}\right)
\end{aligned}$$

by the fonc for generalized IV. However, when we're in the just indentified case,

this is

$$
\begin{aligned}
s(\hat{\beta}_{IV}) &= y' P_W \left(y - X(W'X)^{-1}W'y\right) \\
&= y' P_W \left(I - X(W'X)^{-1}W'\right) y \\
&= y' \left(W(W'W)^{-1}W' - W(W'W)^{-1}W'X(W'X)^{-1}W'\right) y \\
&= 0
\end{aligned}
$$

*The value of the objective function of the IV estimator is zero in the just identified case.* This makes sense, since we've already shown that the objective function after dividing by $\sigma^2$ is asymptotically $\chi^2$ with degrees of freedom equal to the number of overidentifying restrictions. In the present case, there are no overidentifying restrictions, so we have a $\chi^2(0)$ rv, which has mean 0 and variance 0, e.g., it's simply 0. This means we're not able to test the identifying restrictions in the case of exact identification.

## 11.9 System methods of estimation

2SLS is a single equation method of estimation, as noted above. The advantage of a single equation method is that it's unaffected by the other equations of the system, so they don't need to be specified (except for defining what are the exogs, so 2SLS can use the complete set of instruments). The disadvantage of 2SLS is that it's inefficient, in general.

- Recall that overidentification improves efficiency of estimation, since an overidentified equation can use more instruments than are necessary for consistent estimation.

- Secondly, the assumption is that

$$
\begin{aligned}
Y\Gamma &= XB + E \\
\mathcal{E}(X'E) &= 0_{(K\times G)} \\
vec(E) &\sim N(0, \Psi)
\end{aligned}
$$

- Since there is no autocorrelation of the $E_t$ 's, and since the columns of $E$ are individually homoscedastic, then

$$\begin{aligned}\Psi &= \begin{bmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \cdots & \sigma_{1G}I_n \\ & \sigma_{22}I_n & & \vdots \\ & & \ddots & \vdots \\ \cdot & & & \sigma_{GG}I_n \end{bmatrix} \\ &= \Sigma \otimes I_n\end{aligned}$$

  This means that the structural equations are heteroscedastic and correlated with one another

- In general, ignoring this will lead to inefficient estimation, following the section on GLS. When equations are correlated with one another estimation should account for the correlation in order to obtain efficiency.

- Also, since the equations are correlated, information about one equation

is implicitly information about all equations. Therefore, overidentification restrictions in any equation improve efficiency for *all* equations, even the just identified equations.

- Single equation methods can't use these types of information, and are therefore inefficient (in general).

## 3SLS

Note: It is easier and more practical to treat the 3SLS estimator as a generalized method of moments estimator (see Chapter 16). I no longer teach the following section, but it is retained for its possible historical interest. Another alternative is to use FIML (Subsection 11.9), if you are willing to make distributional assumptions on the errors. This is computationally feasible with modern computers.

Following our above notation, each structural equation can be written as

$$\begin{aligned} y_i &= Y_i\gamma_1 + X_i\beta_1 + \varepsilon_i \\ &= Z_i\delta_i + \varepsilon_i \end{aligned}$$

Grouping the $G$ equations together we get

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & Z_G \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_G \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_G \end{bmatrix}$$

or

$$y = Z\delta + \varepsilon$$

where we already have that

$$\begin{aligned}\mathcal{E}(\varepsilon\varepsilon') &= \Psi \\ &= \Sigma \otimes I_n\end{aligned}$$

The 3SLS estimator is just 2SLS combined with a GLS correction that takes advantage of the structure of $\Psi$. Define $\hat{Z}$ as

$$\begin{aligned}\hat{Z} &= \begin{bmatrix} X(X'X)^{-1}X'Z_1 & 0 & \cdots & 0 \\ 0 & X(X'X)^{-1}X'Z_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X(X'X)^{-1}X'Z_G \end{bmatrix} \\ &= \begin{bmatrix} \hat{Y}_1 & X_1 & 0 & & \cdots & 0 \\ 0 & & \hat{Y}_2 & X_2 & & \vdots \\ \vdots & & & & \ddots & 0 \\ 0 & & \cdots & & 0 & \hat{Y}_G & X_G \end{bmatrix}\end{aligned}$$

These instruments are simply the *unrestricted* rf predicitions of the endogs, combined with the exogs. The distinction is that if the model is overidentified, then

$$\Pi = B\Gamma^{-1}$$

may be subject to some zero restrictions, depending on the restrictions on $\Gamma$ and $B$, and $\hat{\Pi}$ does not impose these restrictions. Also, note that $\hat{\Pi}$ is calculated using OLS equation by equation, as was discussed in Section 11.3.

The 2SLS estimator would be

$$\hat{\delta} = (\hat{Z}'Z)^{-1}\hat{Z}'y$$

as can be verified by simple multiplication, and noting that the inverse of a block-diagonal matrix is just the matrix with the inverses of the blocks on the main diagonal. This IV estimator still ignores the covariance information. The natural extension is to add the GLS transformation, putting the inverse of the error

covariance into the formula, which gives the 3SLS estimator

$$\begin{aligned}\hat{\delta}_{3SLS} &= \left(\hat{Z}'\left(\Sigma \otimes I_n\right)^{-1} Z\right)^{-1} \hat{Z}'\left(\Sigma \otimes I_n\right)^{-1} y \\ &= \left(\hat{Z}'\left(\Sigma^{-1} \otimes I_n\right) Z\right)^{-1} \hat{Z}'\left(\Sigma^{-1} \otimes I_n\right) y\end{aligned}$$

This estimator requires knowledge of $\Sigma$. The solution is to define a feasible estimator using a consistent estimator of $\Sigma$. The obvious solution is to use an estimator based on the 2SLS residuals:

$$\hat{\varepsilon}_i = y_i - Z_i \hat{\delta}_{i,2SLS}$$

**(IMPORTANT NOTE**: this is calculated using $Z_i$, not $\hat{Z}_i$). Then the element $i, j$ of $\Sigma$ is estimated by

$$\hat{\sigma}_{ij} = \frac{\hat{\varepsilon}_i' \hat{\varepsilon}_j}{n}$$

Substitute $\hat{\Sigma}$ into the formula above to get the feasible 3SLS estimator.

Analogously to what we did in the case of 2SLS, the asymptotic distribution

of the 3SLS estimator can be shown to be

$$\sqrt{n}\left(\hat{\delta}_{3SLS}-\delta\right) \overset{a}{\sim} N\left(0, \lim_{n\to\infty} \mathcal{E}\left\{\left(\frac{\hat{Z}'\left(\Sigma\otimes I_n\right)^{-1}\hat{Z}}{n}\right)^{-1}\right\}\right)$$

A formula for estimating the variance of the 3SLS estimator in finite samples (cancelling out the powers of $n$) is

$$\hat{V}\left(\hat{\delta}_{3SLS}\right) = \left(\hat{Z}'\left(\hat{\Sigma}^{-1}\otimes I_n\right)\hat{Z}\right)^{-1}$$

- This is analogous to the 2SLS formula in equation (11.5), combined with the GLS correction.

- In the case that all equations are just identified, 3SLS is numerically equivalent to 2SLS. Proving this is easiest if we use a GMM interpretation of 2SLS and 3SLS. GMM is presented in the next econometrics course. For now, take it on faith.

# FIML

Full information maximum likelihood is an alternative estimation method. FIML will be asymptotically efficient, since ML estimators based on a given information set are asymptotically efficient w.r.t. all other estimators that use the same information set, and in the case of the full-information ML estimator we use the entire information set. The 2SLS and 3SLS estimators don't require distributional assumptions, while FIML of course does. Our model is, recall

$$\begin{aligned} Y_t'\Gamma &= X_t'B + E_t' \\ E_t &\sim N(0,\Sigma), \forall t \\ \mathcal{E}(E_t E_s') &= 0, t \neq s \end{aligned}$$

The joint normality of $E_t$ means that the density for $E_t$ is the multivariate normal, which is

$$(2\pi)^{-g/2}\left(\det \Sigma^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}E_t'\Sigma^{-1}E_t\right)$$

The transformation from $E_t$ to $Y_t$ requires the Jacobian

$$|\det \frac{dE_t}{dY_t'}| = |\det \Gamma|$$

so the density for $Y_t$ is

$$(2\pi)^{-G/2}|\det \Gamma| \left(\det \Sigma^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\left(Y_t'\Gamma - X_t'B\right)\Sigma^{-1}\left(Y_t'\Gamma - X_t'B\right)'\right)$$

Given the assumption of independence over time, the joint log-likelihood function is

$$\ln L(B,\Gamma,\Sigma) = -\frac{nG}{2}\ln(2\pi) + n\ln(|\det \Gamma|) - \frac{n}{2}\ln\det\Sigma^{-1} - \frac{1}{2}\sum_{t=1}^{n}\left(Y_t'\Gamma - X_t'B\right)\Sigma^{-1}\left(Y_t'\Gamma - X$$

- This is a nonlinear in the parameters objective function. Maximixation of this can be done using iterative numeric methods. We'll see how to do this in the next section.

- It turns out that the asymptotic distribution of 3SLS and FIML are the

same, *assuming normality of the errors.*

- One can calculate the FIML estimator by iterating the 3SLS estimator, thus avoiding the use of a nonlinear optimizer. The steps are

  1. Calculate $\hat{\Gamma}_{3SLS}$ and $\hat{B}_{3SLS}$ as normal.
  2. Calculate $\hat{\Pi} = \hat{B}_{3SLS}\hat{\Gamma}_{3SLS}^{-1}$. This is new, we didn't estimate $\Pi$ in this way before. This estimator may have some zeros in it. When Greene says iterated 3SLS doesn't lead to FIML, he means this for a procedure that doesn't update $\hat{\Pi}$, but only updates $\hat{\Sigma}$ and $\hat{B}$ and $\hat{\Gamma}$. If you update $\hat{\Pi}$ you *do* converge to FIML.
  3. Calculate the instruments $\hat{Y} = X\hat{\Pi}$ and calculate $\hat{\Sigma}$ using $\hat{\Gamma}$ and $\hat{B}$ to get the estimated errors, applying the usual estimator.
  4. Apply 3SLS using these new instruments and the estimate of $\Sigma$.
  5. Repeat steps 2-4 until there is no change in the parameters.

- FIML is fully efficient, since it's an ML estimator that uses all information. This implies that 3SLS is fully efficient *when the errors are normally distributed.* Also, if each equation is just identified and the errors are normal, then 2SLS will be fully efficient, since in this case 2SLS≡3SLS.

- When the errors aren't normally distributed, the likelihood function is of course different than what's written above.

## 11.10 Example: Klein's Model 1

To give a practical example, consider the following (old-fashioned, but illustrative) macro model (this is the widely known Klein's Model 1)

$$
\begin{aligned}
\text{Consumption: } C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \varepsilon_{1t} \\
\text{Investment: } I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} \\
\text{Private Wages: } W_t^p &= \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} \\
\text{Output: } X_t &= C_t + I_t + G_t \\
\text{Profits: } P_t &= X_t - T_t - W_t^p \\
\text{Capital Stock: } K_t &= K_{t-1} + I_t
\end{aligned}
$$

$$
\begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix} \sim IID \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{pmatrix} \right)
$$

The other variables are the government wage bill, $W_t^g$, taxes, $T_t$, government nonwage spending, $G_t$,and a time trend, $A_t$. The endogenous variables are the lhs variables,

$$Y_t' = \left[\, C_t \;\; I_t \;\; W_t^p \;\; X_t \;\; P_t \;\; K_t \,\right]$$

and the predetermined variables are all others:

$$X_t' = \left[\, 1 \;\; W_t^g \;\; G_t \;\; T_t \;\; A_t \;\; P_{t-1} \;\; K_{t-1} \;\; X_{t-1} \,\right].$$

The model assumes that the errors of the equations are contemporaneously correlated, but nonautocorrelated. The model written as $Y\Gamma = XB + E$ gives

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 \\ -\alpha_3 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & -\gamma_1 & 1 & -1 & 0 \\ -\alpha_1 & -\beta_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} \alpha_0 & \beta_0 & \gamma_0 & 0 & 0 & 0 \\ \alpha_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & \gamma_3 & 0 & 0 & 0 \\ \alpha_2 & \beta_2 & 0 & 0 & 0 & 0 \\ 0 & \beta_3 & 0 & 0 & 0 & 1 \\ 0 & 0 & \gamma_2 & 0 & 0 & 0 \end{bmatrix}$$

To check this identification of the consumption equation, we need to extract $\Gamma_{32}$ and $B_{22}$, the submatrices of coefficients of endogs and exogs that *don't* appear in this equation. These are the rows that have zeros in the first column, and we need

to drop the first column. We get

$$\begin{bmatrix} \Gamma_{32} \\ B_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 & 0 & -1 \\ 0 & -\gamma_1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & \gamma_3 & 0 & 0 & 0 \\ \beta_3 & 0 & 0 & 0 & 1 \\ 0 & \gamma_2 & 0 & 0 & 0 \end{bmatrix}$$

We need to find a set of 5 rows of this matrix gives a full-rank 5×5 matrix. For example, selecting rows 3,4,5,6, and 7 we obtain the matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & \gamma_3 & 0 & 0 & 0 \\ \beta_3 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This matrix is of full rank, so the sufficient condition for identification is met. Counting included endogs, $G^* = 3$, and counting excluded exogs, $K^{**} = 5$, so

$$\begin{aligned} K^{**} - L &= G^* - 1 \\ 5 - L &= 3 - 1 \\ L &= 3 \end{aligned}$$

- The equation is over-identified by three restrictions, according to the count-

ing rules, which are correct when the only identifying information are the exclusion restrictions. However, there is additional information in this case. Both $W_t^p$ and $W_t^g$ enter the consumption equation, and their coefficients are restricted to be the same. For this reason the consumption equation is in fact overidentified by four restrictions.

The Octave program Simeq/Klein2SLS.m performs 2SLS estimation for the 3 equations of Klein's model 1, assuming nonautocorrelated errors, so that lagged endogenous variables can be used as instruments. The results are:

```
CONSUMPTION EQUATION
*********************************************
  OLS estimation, 21 observations
  R^2: 0.981008   Sig^2: 1.051732
  White's covariance estimator

                 coef        se         t         p
```

```
         1    16.237     1.618    10.035     0.000
         2     0.193     0.061     3.172     0.006
         3     0.090     0.066     1.362     0.191
         4     0.796     0.051    15.527     0.000
*********************************************
*****************************************************
Klein model 1 GMM example, plain covariance
GMM Estimation Results
BFGS convergence: Normal convergence
Observations: 21
Hansen-Sargan statistic: 6.74212
Hansen-Sargan p-value: 0.15016

            estimate   st. err    t-stat   p-value
  Constant    14.310     1.127    12.702     0.000
```

```
   Profits     0.085     0.099     0.858     0.403
 Profits-1     0.158     0.087     1.808     0.088
     Wages     0.859     0.037    23.441     0.000

*****************************************************
*****************************************************
Klein model 1 GMM example, NW covariance
GMM Estimation Results
BFGS convergence: No convergence
Observations: 21
Hansen-Sargan statistic: 6.49607
Hansen-Sargan p-value: 0.16504

            estimate   st. err    t-stat   p-value
  Constant    14.402     0.869    16.574     0.000
```

```
   Profits      0.084      0.133      0.631      0.537
 Profits-1      0.149      0.103      1.449      0.166
     Wages      0.859      0.039     22.182      0.000

*****************************************************************
```

The above results are not valid (specifically, they are inconsistent) if the errors are autocorrelated, since lagged endogenous variables will not be valid instruments in that case. You might consider eliminating the lagged endogenous variables as instruments, and re-estimating by 2SLS, to obtain consistent parameter estimates in this more complex case. Standard errors will still be estimated inconsistently, unless use a Newey-West type covariance estimator. Food for thought...

Here's a Gretl script to estimate Klein's model 1: http://gretl.sourceforge.net/gretl-help/scripts/klein.inp.

# Chapter 12

# Numeric optimization methods

**Readings:** Cameron and Trivedi (2005), Ch. 10; Hamilton, ch. 5, section 7 (pp. 133-139)*; Gourieroux and Monfort, Vol. 1, ch. 13, pp. 443-60*; Goffe, et. al. (1994).

The next chapter introduces extremum estimators, which are minimizers or maximizers of objective functions. If we're going to be applying extremum estimators, we'll need to know how to find an extremum. This section gives a very brief introduction to what is a large literature on numeric optimization methods. We'll consider a few well-known techniques, and one of the global optimization methods that may allow one to solve difficult problems.

The main objectives are

- to become familiar with the issues, which should lead you to take a cautious attitude
- to learn how to use gradient-based local minimizers such as `fminunc` and `fmincon` at the practical level.

The general problem we consider is how to find the maximizing element $\hat{\theta}$ (a $K$ -vector) of a function $s(\theta)$. This function may not be continuous, and it may not be differentiable. Even if it is twice continuously differentiable, it may not be globally concave, so local maxima, minima and saddlepoints may all exist. Supposing $s(\theta)$ were a quadratic function of $\theta$, *i.e.*,

$$s(\theta) = a + b'\theta + \frac{1}{2}\theta' C\theta,$$

the first order conditions would be linear:

$$D_\theta s(\theta) = b + C\theta$$

so the maximizing (minimizing) element would be $\hat{\theta} = -C^{-1}b$. This is the sort of problem we have with linear models estimated by OLS. It's also the case for feasible GLS, since conditional on the estimate of the varcov matrix, we have a quadratic objective function in the remaining parameters.

More general problems will not have linear f.o.c., and we will not be able to

solve for the maximizer analytically. This is when we need a numeric optimization method.

## 12.1 Search

The idea is to create a grid over the parameter space and evaluate the function at each point on the grid. Select the best point. Then refine the grid in the neighborhood of the best point, and continue until the accuracy is "good enough". See Figure 12.1. One has to be careful that the grid is fine enough in relationship to the irregularity of the function to ensure that sharp peaks are not missed entirely.

To check $q$ values in each dimension of a $K$ dimensional parameter space, we need to check $q^K$ points. For example, if $q = 100$ and $K = 10$, there would be $100^{10}$ points to check. If 1000 points can be checked in a second, it would take $3.171 \times 10^9$ years to perform the calculations, which is approximately 2/3 the age of the earth. The search method is a very reasonable choice if $K$ is small, but it quickly becomes infeasible if $K$ is moderate or large.

The Julia function GridExample.jl allows you to play around with a simple one dimensional grid search, selecting the number of evenly spaced values to try. Try running GridExample(5) and GridExample(10). The result of GridExample(10)

Figure 12.1: Search method

Figure 12.2: Grid search, one dimension

is in Figure 12.2. In this example, we're in the neighborhood of the minimizer, but still not too close to the minimizer. However, we're close enough so refinement will lead us to converge to the global minimizer.

## 12.2 Derivative-based methods

*In the following, the superscript $k$ is used as the index of the iterations of a given method. It is not an exponent, and it is not the dimension of the parameter vector.*

We assume that the objective function is at least one time differentiable. Otherwise, these methods are not applicable, obviously. Derivative-based methods are defined by

1. the method for choosing the initial value, $\theta^1$
2. the iteration method for choosing $\theta^{k+1}$, given that we're at $\theta^k$ at iteration $k$ (based upon derivatives)
3. the stopping criterion.

The iteration method can be broken into two problems: choosing the stepsize $a^k$ (a scalar) and choosing the direction of movement, $d^k$, which is of the same dimension

of $\theta$, so that

$$\theta^{(k+1)} = \theta^{(k)} + a^k d^k.$$

*A locally increasing direction of search* $d$ is a direction such that

$$\frac{\partial s(\theta + ad)}{\partial a} > 0.$$

That is, if we go in direction $d$, we will improve on the objective function, at least if we don't go too far.

- As long as the gradient at $\theta^k$ is not zero, there exist increasing directions, and they can all be represented as $Q^k g(\theta^k)$ where $Q^k$ is a symmetric pd matrix and $g(\theta) = D_\theta s(\theta)$ is the gradient at $\theta$. To see this, take a Taylor's series expansion around $a^0 = 0$

$$\begin{aligned} s(\theta + ad) &= s(\theta + 0d) + (a - 0)\, g(\theta + 0d)'d + o(1) \\ &= s(\theta) + ag(\theta)'d + o(1) \end{aligned}$$

  For small enough $a$ the $o(1)$ term can be ignored. If $d$ is to be an increasing direction, we need $g(\theta)'d > 0$, because this term is proportional to the

directional derivative in direction $d$. Defining $d = Qg(\theta)$, where $Q$ is positive definite, we guarantee that

$$g(\theta)'d = g(\theta)'Qg(\theta) > 0$$

unless $g(\theta) = 0$. Every increasing direction can be represented in this way (p.d. matrices are those such that the angle between $g$ and $Qg(\theta)$ is less than 90 degrees). See Figure 12.3.

Figure 12.3: Increasing directions of search

- With this, the iteration rule becomes

$$\theta^{(k+1)} = \theta^{(k)} + a^k Q^k g(\theta^k)$$

and we keep going until the gradient becomes zero, so that there is no increasing direction. The problem is now *how to choose $a$ and $Q$.*

- **Conditional on** $Q$, choosing $a$ is fairly straightforward. A simple line (1 dimensional grid) search is an attractive possibility, since $a$ is a scalar. But there are other methods that may be better (bisection, golden, etc.). Optimization packages will generally do this well for you, so you usually don't need to worry too much about it.

- The remaining problem is how to choose $Q$.

- Note also that this gives no guarantees to find a global maximum.

## Steepest ascent

Steepest ascent (descent if we're minimizing) just sets $Q$ to an identity matrix, since the gradient provides the direction of maximum rate of increase of the objective function.

- Advantages: fast, per iteration - doesn't require anything more than first derivatives.

- Disadvantages: May not be fast after all, as we may need many iterations: see the Rosenbrock, or "banana" function: http://en.wikipedia.org/wiki/Rosenbrock_function. (note for lectures: click on the image in the web page to enlarge it)

# Newton's method

Newton's method uses information about the slope and curvature of the objective function to determine which direction and how far to move from an initial point. Supposing we're trying to maximize $s_n(\theta)$. Take a second order Taylor's series approximation of $s_n(\theta)$ about $\theta^k$ (an initial guess).

$$s_n(\theta) \approx s_n(\theta^k) + g(\theta^k)' \left(\theta - \theta^k\right) + 1/2 \left(\theta - \theta^k\right)' H(\theta^k) \left(\theta - \theta^k\right)$$

($g$ is the gradient vector and $H$ is the Hessian matrix). To attempt to maximize $s_n(\theta)$, we can maximize the portion of the right-hand side that depends on $\theta$, *i.e.*, we can maximize

$$\tilde{s}(\theta) = g(\theta^k)'\theta + 1/2 \left(\theta - \theta^k\right)' H(\theta^k) \left(\theta - \theta^k\right)$$

with respect to $\theta$. This is a much easier problem, since it is a quadratic function in $\theta$, so it has linear first order conditions. These are

$$D_\theta \tilde{s}(\theta) = g(\theta^k) + H(\theta^k)\left(\theta - \theta^k\right)$$

So the solution for the next round estimate is

$$\theta^{k+1} = \theta^k - H(\theta^k)^{-1} g(\theta^k)$$

So, the $Q^k$ from above is set to $-H^{-1}(\theta^k)$ when we use Newton's method. See http://en.wikipedia.org/wiki/Newton%27s_method_in_optimization for more information. This is illustrated in Figure 12.4.

However, it's good to include a stepsize, since the approximation to $s_n(\theta)$ may be bad far away from the maximizer $\hat{\theta}$, so the actual iteration formula is

$$\theta^{k+1} = \theta^k - a^k H(\theta^k)^{-1} g(\theta^k)$$

Figure 12.4: Newton iteration

- A potential problem is that the Hessian may not be negative definite when we're far from the maximizing point. So $-H(\theta^k)^{-1}$ may not be positive definite, and $-H(\theta^k)^{-1}g(\theta^k)$ may not define an increasing direction of search. This can happen when the objective function has flat regions, in which case the Hessian matrix is very ill-conditioned (e.g., is nearly singular), or when we're in the vicinity of a local minimum, $H(\theta^k)$ is positive definite, and our direction is a *decreasing* direction of search. Matrix inverses by computers are subject to large errors when the matrix is ill-conditioned. Also, we certainly don't want to go in the direction of a minimum when we're maximizing. To solve this problem, *Quasi-Newton* methods simply add a positive definite component to $H(\theta)$ to ensure that the resulting matrix is positive definite, *e.g.*, $Q = -H(\theta) + b\mathbf{I}$, where $b$ is chosen large enough so that $Q$ is well-conditioned and positive definite. This has the benefit that improvement in the objective function is guaranteed. See http://en.wikipedia.org/wiki/Quasi-Newton_method.

- Another variation of quasi-Newton methods is to approximate the Hessian by using successive gradient evaluations. This avoids actual calculation of the Hessian, which is an order of magnitude (in the dimension of the parameter vector) more costly than calculation of the gradient. They can be done to ensure that the approximation is p.d. DFP and BFGS are two well-known examples.

**Example 27.** BFGS minimization: cut and paste the following code into julia to see some BFGS minimization of the Rosenbrock function

```
using Optim
rosenbrock(x) = (1.0 - x[1])^2 + 100.0 * (x[2] - x[1]^2)^2
result = optimize(rosenbrock, zeros(2), BFGS())
```

Try replacing `BFGS()` with `GradientDescent()` and see what happens.

Also, run the RosenbrockTrace.jl code to see how the path to find the optimizer changes between steepest descent and Newton's method.

**Stopping criteria**

The last thing we need is to decide when to stop. A digital computer is subject to limited machine precision and round-off errors. For these reasons, it is unreasonable to hope that a program can **exactly** find the point that maximizes a function. We need to define acceptable tolerances. Some stopping criteria are:

- Negligible change in parameters:

$$|\theta_j^k - \theta_j^{k-1}| < \varepsilon_1, \forall j$$

- Negligible relative change:

$$|\frac{\theta_j^k - \theta_j^{k-1}}{\theta_j^{k-1}}| < \varepsilon_2, \forall j$$

- Negligible change of function:

$$|s(\theta^k) - s(\theta^{k-1})| < \varepsilon_3$$

- Gradient negligibly different from zero:

$$|g_j(\theta^k)| < \varepsilon_4, \forall j$$

- Or, even better, check all of these. Observe that the BFGS snippet from above checks a number of criteria.

- Also, it's good to check that the last round (real, not approximate) Hessian has the appropriate definiteness. This can be done by computing the Hessian at the final estimates, and then computing its eigenvalues.

**Starting values**

The Newton-Raphson and related algorithms work well if the objective function is concave (when maximizing), but not so well if there are convex regions and local minima or multiple local maxima. The algorithm may converge to a local minimum or to a local maximum that is not optimal. The algorithm may also have difficulties converging at all.

- The usual way to "ensure" that a global maximum has been found is to use many different starting values, and choose the solution that returns the highest objective function value. **THIS IS IMPORTANT in practice.** More on this later.

- an alternative is to use a global optimization algorithm, e.g., simulated annealing or others, which may or may not be gradient based. This may be slow, but is more likely to give you the correct answer, if your problem is difficult.

**Calculating derivatives**

Gradient-based methods obviously require first and possibly second derivatives. It is often difficult to calculate derivatives (especially the Hessian) analytically if the function $s_n(\cdot)$ is complicated. Fortunately, there are some good options:

- computing derivatives by hand: this works well, if you're able to do it, and you don't make mistakes. I don't recommend it for complex problems, as you are likely to make errors and/or program your results incorrectly.

- symbolic differentiation: Mathematica and many other similar packages can give analytic solutions, in many cases, and they can output the solution in code form, to be pasted in to your programs.

- automatic differentiation: this is now a common way to compute derivatives, and should probably be your default way to go.

- numeric derivatives based on finite differences were historically widely used. They are less accurate than analytic derivatives, and are usually more costly

to evaluate. Both factors usually cause optimization programs to be less successful when numeric derivatives are used. However, numeric derivatives provide a reasonably reliable fall-back option if other methods don't work, for some reason. They can be used to check derivatives computed by hand.

**Example 28.** Computing some derivatives. Here's an example of computing the exact gradient using automatic differentiation, and the approximate gradient using finite differences. You will see that the finite difference version has some error. If you replace "gradient" with "hessian", you can get the Hessian matrix.

```
using ForwardDiff, Calculus
rosenbrock(x) = (1.0 - x[1])^2 + 100.0 * (x[2] - x[1]^2)^2
ForwardDiff.gradient(rosenbrock, 2.0*ones(2))
Calculus.gradient(rosenbrock, 2.0*ones(2))
```

## Scaling your data

- Gradient based optimization methods are much more likely to be successful and give accurate results if the data are scaled so that the elements of the gradient are of the same order of magnitude. Example: if the model is $y_t = h(\alpha x_t + \beta z_t) + \varepsilon_t$, and estimation is by NLS. Let $g()$ be the derivative of $h()$.

$$
\begin{aligned}
s_n(\theta) =& \frac{1}{n} \sum_t \left(y_t - h(\alpha x_t + \beta z_t)\right)^2 \\
D_\alpha s_n(\cdot) =& \frac{1}{n} \sum_t 2\left(y_t - h(\alpha x_t + \beta z_t)\right) g(\alpha x_t + \beta z_t) x_t \\
D_\beta s_n(\cdot) =& \frac{1}{n} \sum_t 2\left(y_t - h(\alpha x_t + \beta z_t)\right) g(\alpha x_t + \beta z_t) z_t
\end{aligned}
$$

  - suppose that $D_\alpha s_n(\cdot) = 1000$ and $D_\beta s_n(\cdot) = 0.001$. One could define $\alpha^* = 1000\alpha$; $x_t^* = x_t/1000$;$\beta^* = \beta/1000$; $z_t^* = 1000 z_t$.

- then

$$D_{\alpha^*} s_n(\cdot) = \frac{1}{n} \sum_t 2\left(y_t - h(\alpha^* x_t^* + \beta^* z_t^*)\right) g(\alpha^* x_t^* + \beta^* z_t^*) x_t^*$$

  Everything is the same as before, because the 1000s cancel out, except there is an $x_t^*$ at the end, instead of $x_t$, which causes the derivative to be 1 now, where it was 1000 before.

- the same is true for the other derivative, it will be 1.
- this scaling causes the derivatives to be of the same order.

In general, estimation programs always work better if data is scaled in this way, regardless of what type of derivatives you are using (analytic, automatic, finite difference) since roundoff errors are less likely to become important. *This is important in practice.* It is easy to loose precision in calculation if you don't take care. In the future, if you start to do empirical work and get results that seem meaningless or crazy, try to remember this point. This is

really a very common stumbling block, and many people have wasted a lot of time due to forgetting about it.

**Example 29.** Here are some Hessian computations, using the same Rosenbrock function, but with $x_1$ scaled differently from the previous example .

```
using ForwardDiff, Calculus
rosenbrock(x) = (1.0 - 1000x[1])^2 + 100.0 * (x[2] - (1000x[1])^2)^2
ForwardDiff.hessian(rosenbrock, 2.0*ones(2))
Calculus.hessian(rosenbrock, 2.0*ones(2))
```

Look at the (2,2) element: the finite difference version is off by 6 orders of magnitude! The off diagonal elements are not too good, either. If this were part of a larger program, the results that build on this would be meaningless garbage!

## 12.3 Global methods

Global optimization methods seek to find the global optimum, over a defined parameter space. They can (hopefully) find an optimum in the presence of nonconcavities, discontinuities and multiple local minima/maxima. There are methods backup up by theory that gives conditions under which there is a guarantee of finding the global optimum (perhaps using an infeasible amount of time), and there are so-called heuristic methods, which may perform well, but which may offer no guarantees of working. A well known and popular heuristic method is the Nelder-Mead (simplex) method.

Simulated annealing (SA) is an example of an algorithm which has theoretical support. There are many others, so we'll just focus on this one to get the basic idea. Essentially, the SA algorithm

- randomly selects evaluation points within pre-established limits
- accepts all points that yield an increase in the objective function, but also

accepts some points that decrease the objective function. This allows the algorithm to escape from local minima.

- As more and more points are tried, periodically the algorithm focuses on the best point so far, and reduces the range over which random points are generated. Also, the probability that a negative move is accepted reduces.

- The algorithm relies on many evaluations, as in the search method, but focuses in on promising areas, which reduces function evaluations with respect to the search method. It does not require derivatives to be evaluated.

- The theoretical support for SA requires that the region being search contracts very slowly, so slowly that it becomes impractical for most problems. In practice, this is usually done more quickly than what theory asks for, so the method as used is actually heuristic. One can try to check that it is working by running it more than once, to see if the same solution is found.

- Run `samin()` (with the `Econometrics` package installed) to see an example,

and see the source code for samin.jl, or run `edit(samin,())` to get an idea of how the algorithm works.

**Example 30.** Paste this code into Julia to see an example of simulated annealing, using the version in the package Optim.jl. Try experimenting by passing rt=0.9 or rt=0.25 as arguments, and see what happens. NOTE: as of 22 March, 2022, the convergence message of SAMIN in Optim is missleading. Don't worry about a `* Status: failure` message, if you see one, as long as you see `==> Normal convergence <==`.

```
using Optim
junk=2.
# shows use of obj. fun. as a closure
function sse(x) # very simple quadratic objective function
  objvalue = junk + sum(x.*x)
end
  k = 5
  x = rand(k,1)
  lb = -ones(k,1)
  ub = -lb
xopt= (Optim.optimize(sse, lb, ub, x, SAMIN(verbosity=2),Optim.Options(
```

```
    iterations=10^6))).minimizer
```

An additional example is OLSviaSA.jl , which uses SA (the version the code accompanying these notes, which is functionally equivalent to the version in Optim.jl, but with a different interface) to compute the OLS estimator, and plots the trace of the improvements as they are found. You can see how SA narrows in on the solution, in Figure 12.5

Figure 12.5: Trace of SA path to minimize sum of squared errors

## 12.4 Examples

### The Nerlove model via numeric minimization

The Nerlove model we studied previously (equation 4.10), when estimated by OLS, gives results that do not satisfy homogeneity of degree one, and one of the estimated cost shares is negative. Theory says those things should not happen. Let's impose the restrictions (note, this is not a case of simple linear restrictions, because the restriction that shares are in the $[0, 1]$ interval is not a linear equality restriction). Here's the code: EstimateRestrictedNerlove.jl which shows how to use unconstrained and constrained minimization to estimate the simple Nerlove model with and without parameter restrictions. You should use the objective function values to compute the $qF$ test.

# Maximum likelihood estimation using count data: The MEPS data and the Poisson model

To show optimization methods in practice, using real economic data, this section presents maximum likelihood estimation results for a particular model using real data. The focus at present is simply on numeric optimization. Later, after studying maximum likelihood estimation, this section can be read again.

Demand for health care is usually thought of a a derived demand: health care is an input to a home production function that produces health, and health is an argument of the utility function. Grossman (1972), for example, models health as a capital stock that is subject to depreciation (e.g., the effects of ageing). Health care visits restore the stock. Under the home production framework, individuals decide when to make health care visits to maintain their health stock, or to deal with negative shocks to the stock in the form of accidents or illnesses. As such, individual demand will be a function of the parameters of the individuals' utility functions.

- The MEPS health data file , `meps1996.data,` contains 4564 observations on six measures of health care usage. The data is from the 1996 Medical Expenditure Panel Survey (MEPS). There are now more than 20 years of data! You can get more information at `http://www.meps.ahrq.gov/`.

- The six measures of use are are office-based visits (OBDV), outpatient visits (OPV), inpatient visits (IPV), emergency room visits (ERV), dental visits (VDV), and number of prescription drugs taken (PRESCR). These form columns 1 - 6 of `meps1996.data`.

The conditioning variables are public insurance (PUBLIC), private insurance (PRIV), sex (SEX), age (AGE), years of education (EDUC), and income (INCOME). These form columns 7 - 12 of the file, in the order given here. PRIV and PUBLIC are 0/1 binary variables, where a 1 indicates that the person has access to public or private insurance coverage. SEX is also 0/1, where 1 indicates that the person is female. This data will be used in several examples in what follows.

- Perhaps there may be common factors behind health care visits and the decision to buy private insurance. If there are common unobserved factors (perhaps, health), then private insurace would be an endogenous variable. We will ignore that at present, but come back to it later.

The program ExploreMEPS.jl shows how the data may be read in, and gives some descriptive information about variables, which follows:

```
MEPS data, 1996, complete data set statistics
4564 observations
            mean    st. dev.        min         max
OBDV       3.279       6.171      0.000     133.000
OPV        0.260       1.962      0.000      78.000
IPV        0.194       0.637      0.000      17.000
ERV        0.086       0.389      0.000       5.000
DV         1.054       1.875      0.000      32.000
RX         8.384      18.852      0.000     316.000
PUB        0.141       0.334      0.000       1.000
PRIV       0.674       0.449      0.000       1.000
SEX        0.517       0.500      0.000       1.000
AGE       39.354      12.198     18.000      65.000
EDUC      12.652       2.896      0.000      17.000
INC    42803.630   34108.362      0.000  250463.330
```

All of the measures of use are count data, which means that they take on the values $0, 1, 2, \ldots$. It might be reasonable to try to use this information by specifying the density as a count data density. One of the simplest count data densities is the Poisson density, which is

$$f_Y(y) = \frac{\exp(-\lambda)\lambda^y}{y!}, \; y = 0, 1, 2, \ldots.$$

For this density, $E(Y) = V(Y) = \lambda$. The Poisson average log-likelihood function (assuming the cross sectional observations are independent of one another, so the joint density is the product of the marginals) is

$$s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(-\lambda_i + y_i \ln \lambda_i - \ln y_i!\right)$$

We will parameterize the model as

$$\begin{aligned} \lambda_i &= \exp(\mathbf{x}_i'\beta) \\ \mathbf{x}_i &= [1 \; PUBLIC \; PRIV \; SEX \; AGE \; EDUC \; INC]' \end{aligned} \tag{12.1}$$

This ensures that the mean is positive, as is required for the Poisson model, and now the mean (and the variance) depend upon explanatory variables. Note that for this parameterization

$$\frac{\partial \lambda}{\partial x_j} = \lambda \beta_j,$$

so the elasticity of the conditional mean of $y$ with respect to the $j^{th}$ conditioning variable is

$$\frac{\partial \lambda}{\partial x_j} \frac{x_j}{\lambda} = \beta_j x_j.$$

Thus, the interpretation of the parameters is the same as for a semi-log linear regression model.

The program EstimatePoisson.jl estimates a Poisson model using the full data set. The results of the estimation, using OBDV as the dependent variable are here:

```
julia> include("EstimatePoisson.jl");
*****************************************************
Poisson model, OBDV,  MEPS 1996 full data set
MLE Estimation Results
BFGS convergence: Normal convergence
Average Log-L: -3.67109
Observations: 4564

            estimate   st. err    t-stat   p-value
  constant    -0.791     0.149    -5.290     0.000
 pub. ins.     0.848     0.076    11.092     0.000
priv. ins.     0.294     0.071     4.136     0.000
       sex     0.487     0.055     8.796     0.000
```

```
        age     0.024     0.002    11.470     0.000
        edu     0.029     0.010     3.060     0.002
        inc    -0.001     0.001    -0.978     0.328

Information Criteria
                Crit.    Crit/n
      CAIC 33575.688     7.357
       BIC 33568.688     7.355
       AIC 33523.706     7.345
*****************************************************************
```

**Poor scaling of the data**

When the data is scaled so that the magnitudes of the first and second derivatives are of different orders, problems can easily result, because numerical accuracy can be lost, so gradient-based directions of movement in iterations can become very poor. **UPDATE:** the MLE code now uses automatic differentiation when possible (as it is here). This is less sensitive to scaling problems than is code that uses finite difference derivatives, so we will not observe a difference in the results in this example (except for in the coefficient of income). However, it is still desirable to scale data, to avoid potential round off errors and loss of precision. I~~f we comment the appropriate line in EstimatePoisson.jl, the data will not be scaled, and the estimation program will have difficulty converging (note that the likelihood value is lower with poor scaling).~~

## Multiple optima

Multiple optima (one global, others local) can complicate life, since we have limited means of determining if there is a higher maximum than the one we're at. Think of climbing a mountain in an unknown range, in a very foggy place. A nice picture is Figure 12.6, panel (a), but try to imagine the scene if the clouds were 2000m thicker. A mathematical representation of a similar problem is in the second panel. You can go up until there's nowhere else to go up, but since you're in the fog you don't know if the true summit is across the gap that's at your feet. Do you claim victory (convergence) and go home, or do you trudge down the gap and explore the other side? (example inspired by H.W. Tilman *Snow on the Equator).*

The best way to avoid stopping at a local maximum is to use many starting values, for example on a grid, or randomly generated. Or perhaps one might have priors about possible values for the parameters (*e.g.,* from previous studies of similar data).

Let's try to find the true minimizer of minus 1 times the foggy mountain

Figure 12.6: Multiple local maxima: Mountains with low fog

(a)

(b)

(c)

```
julia> include("FoggyMountain.jl")

The result with poor start values
objective function value: -0.02854
local minimizer: [9.949837851595285, 9.949838970954335]


_________________________________________________________
SAMIN results
==> Normal convergence <==
total number of objective function evaluations: 2350

     Obj. value:     -0.0503702592

       parameter      search width
         0.14843           0.00000
         0.14843           0.00001
_________________________________________________________
the estimate using BFGS with multiple start values: [0.14842659241202327, 0.14842657289310665]
the best objective function value: -0.05037025917950971

julia>
```

function (since the algorithms are set up to minimize). From the picture, you can see it's close to $(0.1, 0.1)$, but let's pretend there is fog, and that we don't know that. The program FoggyMountain.jl shows that poor start values can lead to problems. It uses SA, which finds the true global minimum, and it shows that BFGS using a battery of random start values can also find the global minimum help. The output of one run is in panel (c) of the figure: In that run, the single BFGS run with bad start values converged to a point far from the true minimizer, while simulated annealing found the true minimizer. BFGS using multiple start values also gets the correct solution, and if you check, you'll find that it's faster than SA. The moral of the story is to be cautious and don't publish your results too quickly.

## 12.5 Practical Summary

The practical summary for the Chapter is here. A video that works through the summary is here.

## 12.6 Exercises

1. Numerically minimize the function $\sin(x) + 0.01\,(x-a)^2$, setting $a = 0$, using the software of your choice. Plot the function over the interval $(-2\pi, 2\pi)$. Does the software find the global minimum? Does this depend on the starting value you use? Outline a strategy that would allow you to find the minimum reliably, when $a$ can take on any given value in the interval $(-\pi, \pi)$.

2. Numerically compute the OLS estimator of the Nerlove model

$$\ln C = \beta + \beta_Q \ln Q + \beta_L \ln P_L + \beta_F \ln P_F + \beta_K \ln P_K + \epsilon$$

by using the fminunc function in the `Econometrics` package to minimize the sum of squared residuals. See Subsection 12.4 for a really good hint. The data is at the link nerlove.data . Verify that the results coincide with those

given in subsection 4.8, or with what you get from GRETL, i.e.:

$$\widehat{\text{l_cost}} = \underset{(1.7744)}{-3.52650} + \underset{(0.017466)}{0.720394}\,\text{l_output} + \underset{(0.29105)}{0.436341}\,\text{l_labor} + \underset{(0.10037)}{0.426517}\,\text{l_fuel}$$
$$- \underset{(0.33943)}{0.219888}\,\text{l_capital}$$
$$T = 145 \quad \bar{R}^2 = 0.9238 \quad F(4, 140) = 437.69 \quad \hat{\sigma} = 0.39236$$

(standard errors in parentheses)

The important part of this problem is to learn how to minimize a function that depends on both parameters and data. Try to write your function so that it is re-usable, with a different data set.

3. Take the code from the previous problem, and modify it to estimate the model

$$\ln W = \beta_0 + \beta_{EDUC} EDUC + \beta_X EXP + \beta_{EXP^2} \frac{EXP^2}{100} + \beta_{BLACK} BLACK + \beta_{SMSA} SM$$

using the Card data, which was presented in Section 2.8.

4. Suppose we have an $AR(1)$ model $y_t = \rho y_{t-1} + u_t$. Suppose that $y_t$ is stationary, and that the error $u_t$ is white noise. Explain how one could compute an estimator of $\rho$ using the grid search method. You must define your criterion function and explain how to implement the grid search.

5. Consider the data generating process $y_t = \beta_1 + \beta_2 x_t + u_t$, where $x_t \sim U(0, 5)$ and $\epsilon_t \sim N(0, 1)$. Suppose we have $n$ observations on $(y_t, x_t)$. Suppose that the data satisfies the assumptions of the classical linear regression model. Suppose that we know that $\beta_1 = 1$ and that $0 < \beta_2 < 5$. Generate $n = 100$ observations from this model, and compute an estimator of $\beta_2$ *using the grid search method.* Your answer should include the program that you wrote, and the estimate of $\beta_2$.

6. In Julia (with the `Econometrics` package installed), type `fminunc()`, to learn a bit more about the `fminunc` function for unconstrained minimization, and to see a simple example. This is a convenience function, for Matlab users, and uses the Julia Optim.jl package in the background. If you get into Julia,

it's better to use Optim.jl directly.

7. In Julia (with the `Econometrics` package installed), type `fmincon()`, to learn a bit more about the `fmincon` function for unconstrained minimization, and to see a simple example. This is a convenience function, for Matlab users, and uses the Julia NLopt.jl package in the background. If you get into Julia, it's better to use NLopt.jl directly.

8. In Julia (with the `Econometrics` package installed), type `samin()`, to learn a bit more about the `samin` function for minimization by simulated annealing, and to see a simple example.

# Chapter 13

# Asymptotic properties of extremum estimators

**Readings**: Cameron and Trivedi (2005), Ch. 5; Hayashi (2000), Ch. 7; Gourieroux and Monfort (1995), Vol. 2, Ch. 24; Amemiya, Ch. 4 section 4.1; Davidson and MacKinnon, pp. 591-96; Gallant, Ch. 3; Newey and McFadden (1994).

## 13.1 Extremum estimators

We'll begin with study of *extremum estimators* in general.

- Let $\mathbf{Z}_n = \{z_1, z_2, ..., z_n\}$ be the available data, based on a sample of size $n$.
- Suppose there are $p$ variables, so each $z_i$, $i = 1, 2, ..., n$, is a $p$-vector.
  - so, we have $n$ observations on $p$ variables.
  - $n$ could be a number of economic agents in a cross sectional sample, or the number of time periods in a time series sample.
  - it's probably most natural to organize the data in a $n \times p$ array, like what you would have in a spreadsheet. That is not important here, but it may help to fix the ideas.
- Our paradigm is that data are generated as a draw from a joint density $f_{Z_n}(z)$. This density may not be known, but it exists in principle. It is multivariate, with high dimension.

- The draw from the density may be thought of as the outcome of a random experiment that is characterized by the probability space $\{\Omega, \mathcal{F}, P\}$. When the experiment is performed, $\omega \in \Omega$ is the result, and $\mathbf{Z}_n(\omega) = \{Z_1(\omega), Z_2(\omega), ..., Z_n(\omega)\} = \{z_1, z_2, ..., z_n\}$ is the realized data.

- The probability space is rich enough to allow us to consider events defined in terms of an infinite sequence of data $\mathbf{Z} = \{z_1, z_2, ..., \}$.

**Definition 31.** [Extremum estimator] An extremum estimator $\hat{\theta}$ is the optimizing element of an objective function $s_n(\mathbf{Z}_n, \theta)$ over a compact set $\overline{\Theta}$.

Because the data $\mathbf{Z}_n(\omega)$ depends on $\omega$, we can emphasize this by writing $s_n(\omega, \theta)$ or $s_n(\mathbf{Z}_n(\omega), \theta)$. I'll be loose with notation and interchange when convenient, to emphasize what we're focusing on, at the moment. These points of view are:

- A data set, once $\omega$ has been drawn. This is just the collection of numbers in your spreadsheet. At this point, $\mathbf{Z}_n(\omega)$ is no longer random, and we may as well write simply $\mathbf{Z}_n$. When working with $s_n(\mathbf{Z}_n, \theta)$, which is not at this point a random function, we will be focused on how to maximize or minimize the function, to get the estimate, and how to compute estimated variances, *etc.*

- For the purposes of theory, we need to be able to analyze how well our estimators perform in repeated sampling, so we want to treat the sample as

random. This is when we need to think about $\mathbf{Z}_n(\omega)$, and how $s_n(\mathbf{Z}_n(\omega), \theta)$ behaves with different samples.

**Example 32.** OLS. Let the d.g.p. be $y_t = \mathbf{x}_t'\theta^0 + \varepsilon_t$, $t = 1, 2, ..., n$, $\theta^0 \in \Theta$. Stacking observations vertically, $\mathbf{y}_n = \mathbf{X}_n\theta^0 + \varepsilon_n$, where $\mathbf{X}_n = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{pmatrix}'$. Let $\mathbf{Z}_n = [\mathbf{y}_n \, \mathbf{X}_n]$. The least squares estimator is defined as

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\mathbf{Z}_n, \theta)$$

where

$$s_n(\mathbf{Z}_n, \theta) = 1/n \sum_{t=1}^{n} \left(y_t - \boldsymbol{x}_t'\theta\right)^2$$

As you already know, $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, as this is a case where we can solve the f.o.c. analytically.

The contours of the OLS objective function are plotted in Figure 13.1, based on the Julia script OLScontours.jl . This illustrates the idea that an extremum estimator minimizes or maximizes a function, and that the estimate and parameters that we are trying to estimate are distinct points. As the sample gets large, the two points will get close together, as we will see, if the estimator is *consistent*.

Figure 13.1: OLS objective function contours

**Example 33.** Maximum likelihood. Suppose that the continuous random variables $Y_t \sim IIN(\mu^0, \sigma_0^2)$, $t = 1, 2, ..., n$. Suppose we have a sample $\{y_1, y_2, ..., y_n\}$. The density at the realization $y_t$ is

$$f_Y(y_t; \theta) = (2\pi)^{-1/2} (1/\sigma) \exp\left(-\frac{1}{2}\left(\frac{y_t - \mu}{\sigma}\right)^2\right).$$

where $\theta = (\mu, \sigma)$. The maximum likelihood estimator maximizes the joint density of the sample. Because the data are i.i.d., the joint density of the sample $\{y_1, y_2, ..., y_n\}$ is the product of the densities of each observation, and the ML estimator is

$$\hat{\theta} \equiv \arg\max_{\Theta} \ \mathcal{L}_n(\theta) = \prod_{t=1}^{n} f_Y(y_t; \theta)$$

Because the natural logarithm is strictly increasing on $(0, \infty)$, maximization of the average logarithmic likelihood function is achieved at the same $\hat{\theta}$ as for the likelihood function. So, the ML estimator $\hat{\theta} = \arg\max_{\Theta} s_n(\theta)$ where

$$\begin{aligned} s_n(\theta) &= (1/n) \ln \mathcal{L}_n(\theta) = (1/n) \sum_{t=1}^{n} \ln f_Y(y_t; \theta) \\ &= -\ln \sqrt{2\pi} - \log\sigma - (1/n) \sum_{t=1}^{n} \left( \frac{y_t - \mu}{\sigma} \right)^2 \end{aligned}$$

Solution of the f.o.c. leads to the familiar result that $\hat{\mu} = \bar{\mathbf{y}}$. We'll come back to this in more detail later.

**Example 34.** Bayesian estimator

(reminder to self in lectures: that squiggle is a "zeta") Bayesian point estimators such as the posterior mode, median or mean can be expressed as extremum estimators. For example, the posterior mean $E(\theta|Z_n)$ is the minimizer (with respect to $\zeta$) of the function

$$s_n(\zeta) = \int_\Theta (\theta - \zeta)^2 f(Z_n; \theta)\pi(\theta)/f(Z_n)d\theta$$

where $f(Z_n; \theta)$ is the likelihood function, $\pi(\theta)$ is a prior density, and $f(Z_n)$ is the marginal likelihood of the data. These concepts are explained later, for now the point is that Bayesian point estimators can be thought of as extremum estimators, and the theory for extremum estimators will apply.

- Note that the objective function $s_n(\mathbf{Z}_n, \theta)$ is a random function, because it depends on $\mathbf{Z}_n(\omega) = \{Z_1(\omega), Z_2(\omega), ..., Z_n(\omega)\} = \{z_1, z_2, ..., z_n\}$.

- We need to consider what happens as different outcomes $\omega \in \Omega$ occur. These different outcomes lead to different data being generated, and the different data causes the objective function to change.

- Note, however, that for a fixed $\omega \in \Omega$, the data $\mathbf{Z}_n(\omega) = \{Z_1(\omega), Z_2(\omega), ..., Z_n(\omega)\} = \{z_1, z_2, ..., z_n\}$ are a fixed realization, and the objective function $s_n(\mathbf{Z}_n, \theta)$ becomes a non-random function of $\theta$.

- When actually *computing* an extremum estimator, we condition on the observed data, and treat it as fixed. Then we compute estimators either by solving the f.o.c., as in the case of OLS, or if that is not possible, by employing algorithms for optimization of nonstochastic functions. How to do this is the topic of Chapter 12.

- When *analyzing the properties* of an extremum estimator, we need to in-

vestigate what happens throughout $\Omega$: we do not focus only on the $\omega$ that generated the observed data. This is because we would like to find estimators that work well on average, for any data set that can result from $\omega \in \Omega$. This is the topic of the remainder of the present Chapter.

- We'll often write the objective function suppressing the dependence on $\mathbf{Z}_n$, as $s_n(\omega, \theta)$ or simply $s_n(\theta)$, depending on context. The first of these emphasizes the fact that the objective function is random, and the second is more compact. However, the data is still in there, and because the data is randomly sampled, the objective function is random, too.

- For lectures, jump to Ch. 12 now, and then come back here.

## 13.2 Existence

If $s_n(\theta)$ is continuous in $\theta$ and $\overline{\Theta}$ is compact, then a maximizer exists, by the Weierstrass maximum theorem (Debreu, 1959). In some cases of interest, $s_n(\theta)$ may not be continuous. Nevertheless, it may still converge to a continuous function, in which case existence will not be a problem, at least asymptotically. Henceforth in this course, we assume that $s_n(\theta)$ is continuous, unless it is stated otherwise.

## 13.3 Consistency

The following theorem is patterned on a proof in Gallant (1987b). It is interesting to compare the following proof with Amemiya's Theorem 4.1.1, which is done in terms of convergence in probability.

**Theorem 35.** [Consistency of e.e.] *Suppose that $\hat{\theta}_n$ is obtained by maximizing $s_n(\theta)$ over $\overline{\Theta}$.*

*Assume*

*(a) Compactness: The parameter space $\Theta$ is an open bounded subset of Euclidean space $\mathbb{R}^K$. So, the closure of $\Theta$, $\overline{\Theta}$, is compact.*

*(b) Uniform Convergence: There is a nonstochastic function $s_\infty(\theta)$ that is continuous in $\theta$ on $\overline{\Theta}$ such that*

$$\lim_{n\to\infty} \sup_{\theta\in\overline{\Theta}} |s_n(\omega,\theta) - s_\infty(\theta)| = 0, \ a.s.$$

*(c) Identification: $s_\infty(\cdot)$ has a unique global maximum at $\theta^0 \in \Theta$, i.e.,*

$s_\infty(\theta^0) > s_\infty(\theta)$, $\forall \theta \neq \theta^0, \theta \in \overline{\Theta}$

*Then* $\hat{\theta}_n \xrightarrow{a.s.} \theta^0$.

**Proof:**

- Select a $\omega \in \Omega$ and hold it fixed. Then $\{s_n(\omega, \theta)\}$ is a fixed sequence of functions. Suppose that $\omega$ is such that $s_n(\omega, \theta)$ converges to $s_\infty(\theta)$.

- The sequence $\{\hat{\theta}_n\}$ lies in the compact set $\overline{\Theta}$, by assumption (a) and the fact that maximization is over $\overline{\Theta}$. Since every sequence from a compact set has at least one limit point (Bolzano-Weierstrass), say that $\hat{\theta}$ is a limit point of $\{\hat{\theta}_n\}$. As such, there is a subsequence $\{\hat{\theta}_{n_m}\}$ ($\{n_m\}$ is simply a sequence of increasing integers) with $\lim_{m\to\infty} \hat{\theta}_{n_m} = \hat{\theta}$.

- By uniform convergence and continuity of the limiting objective function,

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_{n_m}) = s_\infty(\hat{\theta}).$$

To see this, first of all, select an element $\hat{\theta}_t$ from the sequence $\left\{\hat{\theta}_{n_m}\right\}$. Then uniform convergence (assn. b) implies

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_t) = s_\infty(\hat{\theta}_t)$$

Continuity of $s_\infty(\cdot)$ implies that

$$\lim_{t\to\infty} s_\infty(\hat{\theta}_t) = s_\infty(\hat{\theta})$$

since the limit as $t \to \infty$ of $\left\{\hat{\theta}_t\right\}$ is $\hat{\theta}$. So the above claim is true.

- Next, by maximization

$$s_{n_m}(\hat{\theta}_{n_m}) \geq s_{n_m}(\theta^0)$$

which holds in the limit, so

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_{n_m}) \geq \lim_{m\to\infty} s_{n_m}(\theta^0).$$

- However, for the left hand side, the previous slide showed that

$$\lim_{m\to\infty} s_{n_m}(\hat{\theta}_{n_m}) = s_\infty(\hat{\theta}),$$

For the right hand side, a similar argument gives

$$\lim_{m\to\infty} s_{n_m}(\theta^0) = s_\infty(\theta^0)$$

by uniform convergence, so the above inequality can be written as

$$s_\infty(\hat{\theta}) \geq s_\infty(\theta^0).$$

- But by assumption (c), there is a unique global maximum of $s_\infty(\theta)$ at $\theta^0$, so we must have $s_\infty(\hat{\theta}) = s_\infty(\theta^0)$, and, therefore, $\hat{\theta} = \theta^0$.

- Finally, so far we have held $\omega$ fixed, but now we need to consider all $\omega \in \Omega$. All of the above limits hold almost surely, by assumption (b). Therefore $\{\hat{\theta}_n\}$ has only one limit point, $\theta^0$, except on a set $C \subset \Omega$ with $P(C) = 0$.

*Discussion of the proof:*

- We assume that $\hat{\theta}_n$ is in fact a global maximum of $s_n(\theta)$. It is not required to be unique for $n$ finite, though the identification assumption requires that the limiting objective function have a unique maximizing argument. The previous section on numeric optimization methods showed that actually finding the global maximum of $s_n(\theta)$ may be a non-trivial problem.

- See Amemiya's Example 4.1.4 for a case where discontinuity leads to breakdown of consistency.

- uniform convergence is needed, so that the maximum of $s_n(\theta)$ is eventually close to $\theta_0$. See Figure 13.2. If the objective function is not converging at $\theta_*$, there's no guarantee that $s_n(\theta_*)$ will be lower than $s_\infty(\theta_0)$ as $n$ gets large.

- The assumption that $\theta^0$ is in the interior of $\overline{\Theta}$ (part of the identification assumption) has not been used to prove consistency, so we could directly assume that $\theta^0$ is simply an element of a compact set $\overline{\Theta}$. The reason that we

Figure 13.2: Why uniform convergence of $s_n(\theta)$ is needed

assume it's in the interior here is that this is necessary for subsequent proof of asymptotic normality, and I'd like to maintain a minimal set of simple assumptions, for clarity. Parameters on the boundary of the parameter set cause theoretical difficulties that we will not deal with in this course. Just note that conventional hypothesis testing methods do not apply in this case.

- Note that $s_n(\theta)$ is not required to be continuous, though $s_\infty(\theta)$ is.

## Sufficient conditions for assumption (b)

We need a uniform strong law of large numbers in order to verify assumption (2) of Theorem 35. To verify the uniform convergence assumption, it is often feasible to employ the following set of stronger assumptions:

- the parameter space is compact, which is given by assumption (b)
- the objective function $s_n(\theta)$ is continuous and bounded with probability one on the entire parameter space
- a standard SLLN can be shown to apply to some point $\theta$ in the parameter space. That is, we can show that $s_n(\theta) \xrightarrow{a.s.} s_\infty(\theta)$ for some $\theta$. Note that in most cases, the objective function will be an average of terms, such as

$$s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} s_t(\theta)$$

  As long as the $s_t(\theta)$ are not too strongly dependent, and have finite variances, we can usually find a SLLN that will apply.

With these assumptions, it can be shown that assumption (b) holds.

These are reasonable conditions in many cases, and henceforth when dealing with specific estimators we'll simply assume that assumption (b) holds.

## 13.4 Example: Consistency of Least Squares

Thus example shows how the above theorem can be used to show that the OLS estimator under the classical assumptions is consistent. Of course, this is not the easiest way to show that. The purpose is to show that the theorem gives us a result that we already know to be true. This may help you to believe it in the cases where we do not have external confirmation.

We suppose that data is generated by random sampling of $(Y, X)$, where $y_t = \beta_0 x_t + \varepsilon_t$. $(X, \varepsilon)$ has the distribution function $F_Z = \mu_x \mu_\varepsilon$ ($x$ and $\varepsilon$ are independent) with support $\mathcal{Z} = \mathcal{X} \times \mathcal{E}$. Suppose that the variances $\sigma_X^2$ and $\sigma_\varepsilon^2$ are finite. The sample objective function for a sample size $n$ is

$$\begin{aligned}
s_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n}(y_t - \beta x_t)^2 = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 x_t + \varepsilon_t - \beta x_t)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((\beta_0 - \beta)x_t + \varepsilon_t)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}((\beta_0 - \beta)x_t)^2 + \frac{2}{n}\sum_{t=1}^{n}(\beta_0 - \beta)x_t\varepsilon_t + \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2
\end{aligned}$$

- Considering the last term, by the SLLN,

$$\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 \overset{a.s.}{\to} \int_{\mathcal{X}}\int_{\mathcal{E}}\varepsilon^2 d\mu_{\mathcal{X}} d\mu_{\mathcal{E}} = \sigma_\varepsilon^2.$$

- Considering the second term, since $E(\varepsilon) = 0$ and $X$ and $\varepsilon$ are independent, the SLLN implies that it converges to zero.

- Finally, for the first term, for a given $\beta$, we assume that a SLLN applies so that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\left((\beta_0-\beta)x_t\right)^2 &\overset{a.s.}{\to} \int_{\mathcal{X}}\left((\beta_0-\beta)x\right)^2 d\mu_{\mathcal{X}} \\ &= \left(\beta^0-\beta\right)^2\int_{\mathcal{X}} x^2 d\mu_{\mathcal{X}} \\ &= \left(\beta^0-\beta\right)^2 E\left(X^2\right)\end{aligned} \tag{13.1}$$

Finally, the objective function is clearly continuous, and the parameter space is

assumed to be compact, so the convergence is also uniform. Thus,

$$s_\infty(\beta) = \left(\beta^0 - \beta\right)^2 E\left(X^2\right) + \sigma_\varepsilon^2$$

A minimizer of this is clearly $\beta = \beta^0$.

**Exercise 36.** Show that in order for the above solution to be unique it is necessary that $E(X^2) \neq 0$. Interpret this condition.

This example shows that Theorem 35 can be used to prove strong consistency of the OLS estimator. There are easier ways to show this, of course - this is only an example of application of the theorem.

For a more concrete example, Figure 13.3 shows computations that illustrate that the OLS estimator is consistent, when the true relationship is $y = 1 + 1x + \epsilon$, $\epsilon$ satisfies the classical assumptions, and $x$ is distributed uniform$(0, 1)$. The computations show that the true parameter values satisfy the first order conditions for minimization of the first term of the limiting objective function, eqn. 13.1, above. You should know how to do this by hand, but I use software here, just to illustrate that it can do this sort of thing.

Figure 13.3: Consistency of OLS

## 13.5 More on the limiting objective function: correctly and incorrectly specified models

The limiting objective function in assumption (b) is $s_\infty(\theta)$. What is the nature of this function and where does it come from?

- Remember our paradigm - data is presumed to be generated as a draw from $f_{Z_n}(z)$, and the objective function is $s_n(Z_n, \theta)$.
- Usually, $s_n(Z_n, \theta)$ is an average of terms. (*e.g.,* sum of squared errors, or the log likelihood function of Example 33)
- The limiting objective function is found by applying a strong (weak) law of large numbers to $s_n(Z_n, \theta)$.
- A strong (weak) LLN says that an average of terms converges almost surely

(in probability) to the limit of the expectation of the average.

Supposing one holds,

$$s_\infty(\theta) = \lim_{n\to\infty} \mathcal{E} s_n(Z_n, \theta) = \lim_{n\to\infty} \int_{\mathcal{Z}_n} s_n(z, \theta) f_{Z_n}(z) dz$$

Now suppose that the density $f_{Z_n}(z)$ that characterizes the DGP is parametric: $f_{Z_n}(z;\rho)$, $\rho \in \varrho$, and the data is generated by $\rho^0 \in \varrho$. Now we have two parameters to worry about, $\theta$ and $\rho$. We are probably interested in learning about the true DGP, which means that $\rho^0$ is the item of interest. When the DGP is parametric, the limiting objective function is

$$s_\infty(\theta) = \lim_{n\to\infty} \mathcal{E} s_n(Z_n,\theta) = \lim_{n\to\infty} \int_{\mathcal{Z}_n} s_n(z,\theta) f_{Z_n}(z;\rho^0) dz \tag{13.2}$$

and we can write the limiting objective function as $s_\infty(\theta,\rho^0)$ to emphasize the dependence on the parameter of the DGP. From the theorem, we know that $\hat{\theta}_n \xrightarrow{a.s.} \theta^0$ *What is the relationship between $\theta^0$ and $\rho^0$? Does the econometric estimator tell us something about the true unknown parameter?*

- $\rho$ and $\theta$ may have different dimensions. Often, the statistical model (with parameter $\theta$) only partially describes the DGP. For example, the case of OLS with errors of unknown distribution. In some cases, the dimension of $\theta$ may be greater than that of $\rho$. For example, fitting a polynomial to an unknown

nonlinear function.

- case 1: If knowledge of $\theta^0$ is sufficient for knowledge of $\rho^0$, then we have a correctly and fully specified model. $\theta^0$ is referred to as the *true parameter value*. Example 13.4 illustrates this case.

- case 2: If knowledge of $\theta^0$ is sufficient for knowledge of some but not all elements of $\rho^0$, we have a correctly specified *semiparametric* model. $\theta^0$ is referred to as the *true parameter value*, understanding that not all parameters of the DGP are estimated. An example would be OLS with heteroscedasticity of unknown form: we can learn about the parameters of the conditional mean, but not about the conditional variances.

- case 3: If knowledge of $\theta^0$ is not sufficient for knowledge of any elements of $\rho^0$, or if it causes us to draw false conclusions regarding at least some of the elements of $\rho^0$, our model is *misspecified*. $\theta^0$ is referred to as the *pseudo-true parameter value*. The next section provides an example.

## 13.6 Example: Inconsistency of Misspecified Least Squares

You already know that the OLS estimator is inconsistent when relevant variables are omitted. Let's verify this result in the context of extremum estimators. We suppose that data is generated by random sampling of $(Y, X)$, where

$$y_t = \beta_0 x_t + \varepsilon_t.$$

$(X, \varepsilon)$ has the distribution function $F_Z = \mu_x \mu_\varepsilon$ ($x$ and $\varepsilon$ are independent) with support $Z = \mathcal{X} \times \mathcal{E}$. Suppose that the variances $\sigma_X^2$ and $\sigma_\varepsilon^2$ are finite. However, the econometrician is unaware of the true DGP, and instead proposes the misspecified model

$$y_t = \gamma_0 w_t + \eta_t.$$

Suppose that $E(W\epsilon) = 0$ and that $E(WX) \neq 0$.

The sample objective function for a sample size $n$ is

$$\begin{aligned} s_n(\gamma) &= 1/n \sum_{t=1}^{n} (y_t - \gamma w_t)^2 = 1/n \sum_{i=1}^{n} (\beta_0 x_t + \varepsilon_t - \gamma w_t)^2 \\ &= 1/n \sum_{t=1}^{n} (\beta_0 x_t)^2 + 1/n \sum_{t=1}^{n} (\gamma w_t)^2 + 1/n \sum_{t=1}^{n} \varepsilon_t^2 \\ &\quad +2/n \sum_{t=1}^{n} \beta_0 x_t \varepsilon_t - 2/n \sum_{t=1}^{n} \beta_0 \gamma x_t w_t - 2/n \sum_{t=1}^{n} \varepsilon_t x_t w_t, \end{aligned}$$

which one can verify if armed with patience. Using arguments similar to the correctly specified case, above,

$$s_\infty(\gamma) = \gamma^2 E\left(W^2\right) - 2\beta_0 \gamma E(WX) + C$$

where $C$ holds the red terms, that do not depend on $\gamma$, and the terms that are not given a color converge to 0. So, finding the minimizer with respect to $\gamma$, we get $\gamma_0 = \frac{\beta_0 E(WX)}{E(W^2)}$, which is the true parameter of the DGP, multiplied by the pseudo-true value of a regression of $X$ on $W$. The OLS estimator *is not consistent* for the true parameter, $\beta_0$

## Summary

The theorem for consistency is really quite intuitive. It says that, with probability one, an extremum estimator converges to the value that maximizes the limit of the expectation of the objective function. Because the objective function may or may not make sense, depending on how good or poor is the econometric model, we may or may not be estimating parameters of the DGP.

## 13.7 Example: Linearization of a nonlinear model

See White (1980b) and Gourieroux and Monfort, section 8.3.4.

Suppose we have a nonlinear model

$$y_i = h(x_i, \theta^0) + \varepsilon_i$$

where

$$\varepsilon_i \sim iid(0, \sigma^2)$$

The *nonlinear least squares* estimator solves

$$\hat{\theta}_n = \arg\min \frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i, \theta))^2$$

We'll study this more later, but for now it is clear that the foc for minimization will require solving a set of nonlinear equations. A common approach to the problem seeks to avoid this difficulty by *linearizing* the model. A first order Taylor's series expansion about the point $x_0$ with remainder gives

$$y_i = h(x^0, \theta^0) + (x_i - x_0)' \frac{\partial h(x_0, \theta^0)}{\partial x} + \nu_i$$

where $\nu_i$ encompasses both $\varepsilon_i$ and the Taylor's series remainder.

- Note that $\nu_i$ is no longer a classical error - its mean is not zero. We should expect problems.

- Define

$$\begin{aligned} \alpha^* &= h(x_0, \theta^0) - x_0' \frac{\partial h(x^0, \theta^0)}{\partial x} \\ \beta^* &= \frac{\partial h(x_0, \theta^0)}{\partial x} \end{aligned}$$

as the intercept and slope of the Taylor's series tangent line.

- Given this, one might try to estimate $\alpha^*$ and $\beta^*$ by applying OLS to

$$y_i = \alpha + \beta x_i + \nu_i$$

- Question, will $\hat{\alpha}$ and $\hat{\beta}$ be consistent for $\alpha^*$ and $\beta^*$?
- The answer is no, as one can see by interpreting $\hat{\alpha}$ and $\hat{\beta}$ as extremum estimators. Let $\gamma = (\alpha, \beta')'$.

$$\hat{\gamma} = \arg\min s_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

The objective function converges to its expectation

$$s_n(\gamma) \overset{u.a.s.}{\to} s_\infty(\gamma) = \mathcal{E}_X \mathcal{E}_{Y|X} (y - \alpha - \beta x)^2$$

and $\hat{\gamma}$ converges $a.s.$ to the $\gamma^0$ that minimizes $s_\infty(\gamma)$:

$$\gamma^0 = \arg\min \mathcal{E}_X \mathcal{E}_{Y|X} (y - \alpha - \beta x)^2$$

Noting that

$$\begin{aligned} \mathcal{E}_X \mathcal{E}_{Y|X} (y - \alpha - x'\beta)^2 &= \mathcal{E}_X \mathcal{E}_{Y|X} \left(h(x, \theta^0) + \varepsilon - \alpha - \beta x\right)^2 \\ &= \sigma^2 + \mathcal{E}_X \left(h(x, \theta^0) - \alpha - \beta x\right)^2 \end{aligned}$$

since cross products involving $\varepsilon$ drop out. $\alpha^0$ and $\beta^0$ correspond to the hyperplane that is closest to the true regression function $h(x, \theta^0)$ according to the mean squared error criterion. This depends on both the shape of $h(\cdot)$ and the density function of the conditioning variables. See Figure 13.4.

Figure 13.4: Linear Approximation

- It is clear that the tangent line does not minimize MSE, since, for example, if $h(x, \theta^0)$ is concave, all errors between the tangent line and the true function are negative.

- Note that the true underlying parameter $\theta^0$ is not estimated consistently, either (it may be of a different dimension than the dimension of the parameter of the approximating model, which is 2 in this example).

- Second order and higher-order approximations suffer from exactly the same problem, though to a less severe degree, of course. For this reason, translog, Generalized Leontiev and other "flexible functional forms" based upon second-order approximations in general suffer from bias and inconsistency. The bias may not be too important for analysis of conditional means, but it can be very important for analyzing first and second derivatives. In production and consumer analysis, first and second derivatives (*e.g.,* elasticities of substitution) are often of interest, so in this case, one should be cautious of unthinking application of models that impose stong restrictions on second derivatives.

- This sort of linearization about a long run equilibrium is a common practice in working with dynamic macroeconomic models. It is justified for the purposes of theoretical analysis of a model *given* the model's parameters, but it will lead to *bias and inconsistency* if it is done before estimation of the parameters of the model using data. The section on simulation-based

methods offers a means of obtaining consistent estimators of the parameters of dynamic macro models that are too complex for standard methods of analysis.

## 13.8 Asymptotic Normality

A consistent estimator is oftentimes not very useful unless we know how fast it is likely to be converging to the true value, and the probability that it is far away from the true value. Establishment of asymptotic normality with a known scaling factor solves these two problems. The following theorem is similar to Amemiya's Theorem 4.1.3 (pg. 111).

**Theorem 37.** [Asymptotic normality of e.e.] *In addition to the assumptions of Theorem 35, assume*

*(a)* $\mathcal{J}_n(\theta) \equiv D^2_\theta s_n(\theta)$ *exists and is continuous in an open, convex neighborhood of* $\theta^0$.

*(b)* $\mathcal{J}_n(\theta_n) \xrightarrow{a.s.} \mathcal{J}_\infty(\theta^0)$, *a finite negative definite matrix, for any sequence of* $\theta_n$ *that converges almost surely to* $\theta^0$.

*(c)* $\sqrt{n} D_\theta s_n(\theta^0) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta^0)\right]$, *where* $\mathcal{I}_\infty(\theta^0) = \lim_{n\to\infty} Var \sqrt{n} D_\theta s_n(\theta^0)$

*Then* $\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$

**Proof:** By Taylor expansion:

$$D_\theta s_n(\hat{\theta}_n) = D_\theta s_n(\theta^0) + D_\theta^2 s_n(\theta^*)\left(\hat{\theta} - \theta^0\right)$$

where $\theta^* = \lambda\hat{\theta} + (1-\lambda)\theta^0$, $0 \leq \lambda \leq 1$.

- Note that $\hat{\theta}$ will be in the neighborhood where $D_\theta^2 s_n(\theta)$ exists with probability one as $n$ becomes large, by consistency.

- Now the l.h.s. of this equation is zero, at least asymptotically, since $\hat{\theta}_n$ is a maximizer and the f.o.c. must hold exactly since the limiting objective function is strictly concave in a neighborhood of $\theta^0$ (assns. a and b)

- Also, since $\theta^*$ is between $\hat{\theta}_n$ and $\theta^0$, and since $\hat{\theta}_n \overset{a.s.}{\to} \theta^0$ , assumption (b) gives

$$D_\theta^2 s_n(\theta^*) \overset{a.s.}{\to} \mathcal{J}_\infty(\theta^0)$$

So

$$0 = D_\theta s_n(\theta^0) + \left[\mathcal{J}_\infty(\theta^0) + o_s(1)\right]\left(\hat{\theta} - \theta^0\right),$$

and

$$\sqrt{n} D_\theta s_n(\theta^0) = -\left[\mathcal{J}_\infty(\theta^0) + o_s(1)\right]\sqrt{n}\left(\hat{\theta} - \theta^0\right).$$

Now $\sqrt{n} D_\theta s_n(\theta^0) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta^0)\right]$ by assumption c, so

$$-\left[\mathcal{J}_\infty(\theta^0) + o_s(1)\right]\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta^0)\right]$$

Also, $\left[\mathcal{J}_\infty(\theta^0) + o_s(1)\right] \xrightarrow{a.s.} \mathcal{J}_\infty(\theta^0)$, so

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$$

by the Slutsky Theorem (quick version, formal version: see Gallant (1997), Theorem 4.6).

Figure 13.5 shows the effects of the two components, the variability of the gradient, and the slope of the gradient.

Figure 13.5: Effects of $I_\infty$ and $J_\infty$

## 13.9 Example: Classical linear model

Let's use the results to get the asymptotic distribution of the OLS estimator applied to the classical model, to verify that we obtain the results seen before. The OLS criterion is

$$
\begin{aligned}
s_n(\beta) &= \frac{1}{n}(y - X\beta)'(y - X\beta) \\
&= \frac{1}{n}\left(X\beta^0 + \epsilon - X\beta\right)'\left(X\beta^0 + \epsilon - X\beta\right) \\
&= \frac{1}{n}\left(X(\beta^0 - \beta) + \epsilon\right)'\left(X(\beta^0 - \beta) + \epsilon\right) \\
&= \frac{1}{n}\left[\left(\beta^0 - \beta\right)' X'X\left(\beta^0 - \beta\right) + 2\epsilon' X(\beta_0 - \beta) + \epsilon'\epsilon\right]
\end{aligned}
$$

The first derivative is

$$
D_\beta s_n(\beta) = \frac{1}{n}\left[-2X'X\left(\beta^0 - \beta\right) - 2X'\epsilon\right]
$$

so, evaluating at $\beta^0$,

$$D_\beta s_n(\beta^0) = -2\frac{X'\epsilon}{n}$$

- note that this is an average of terms, each of which has expectation zero: $-2\frac{X'\epsilon}{n} = -2\frac{1}{n}\,\Sigma_t\, x_t\epsilon_t$
- thus, a LLN tells us this converges almost surely to 0.
- to keep this from happening, we can multiply by something that is converging to infinity. It turns out that $\sqrt{n}$ is the right choice, because then the asymptotic distribution will be stable, and a CLT will apply.

Now, let's get the form of $\mathcal{I}_\infty$ of Assumption (c): Considering $\sqrt{n}D_\beta s_n(\beta^0)$, it has expectation 0, so the variance is the expectation of the outer product (there's no

need to subtract the mean):

$$
\begin{aligned}
Var\sqrt{n}D_\beta s_n(\beta^0) &= E\left[\left(-\sqrt{n}2\frac{X'\epsilon}{n}\right)\left(-\sqrt{n}2\frac{X'\epsilon}{n}\right)'\right] \\
&= E4\frac{X'\epsilon\epsilon'X}{n} \\
&= 4\sigma_\epsilon^2 E\left(\frac{X'X}{n}\right)
\end{aligned}
$$

(because regressors are independent of errors). Therefore

$$
\begin{aligned}
\mathcal{I}_\infty(\beta^0) &= \lim_{n\to\infty} Var\sqrt{n}D_\beta s_n(\beta^0) \\
&= 4\sigma_\epsilon^2 Q_X
\end{aligned}
$$

where $Q_X = \lim E\left(\frac{X'X}{n}\right)$, is a finite p.d. matrix (by the classical assumption of no perfect collinearity).

The second derivative is

$$\mathcal{J}_n(\beta) = D^2_\beta s_n(\beta^0) = \frac{1}{n}\left[2X'X\right].$$

A SLLN tells us that this converges almost surely to the limit of its expectation:

$$\mathcal{J}_\infty(\beta^0) = 2Q_X$$

There's no parameter in that last expression, so uniformity is not an issue.

The asymptotic normality theorem (37) tells us that

$$\sqrt{n}\left(\hat{\beta}-\beta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\beta^0)^{-1}\mathcal{I}_\infty(\beta^0)\mathcal{J}_\infty(\beta^0)^{-1}\right]$$

which is, given the above,

$$\sqrt{n}\left(\hat{\beta}-\beta^0\right) \xrightarrow{d} N\left[0, \left(\frac{Q_X^{-1}}{2}\right)\left(4\sigma_\epsilon^2 Q_X\right)\left(\frac{Q_X^{-1}}{2}\right)\right]$$

or

$$\sqrt{n}\left(\hat{\beta}-\beta^0\right) \xrightarrow{d} N\left[0, Q_X^{-1}\sigma_\epsilon^2\right].$$

This is the same thing we saw in equation 5.1, of course. So, the theory seems to work :-)

## 13.10 Practical Summary

The practical summary for the Chapter is here.

## 13.11 Exercises

1. Suppose that $x_i \sim$ uniform(0,1), and $y_i = 1 - x_i^2 + \varepsilon_i$, where $\varepsilon_i$ is iid(0,$\sigma^2$). Suppose we estimate the misspecified model $y_i = \alpha + \beta x_i + \eta_i$ by OLS.

   (a) Find, analytically, the numeric values of $\alpha^0$ and $\beta^0$ that are the probability limits of $\hat{\alpha}$ and $\hat{\beta}$. Hint: the correct answers are 7/6 and -1. To get some help with this exercise, you can use a computer algebra program, as was done in Figure 13.3. A small modification of that code would solve this problem. In particular, paste the following into a new Xmaxima session:

```
integrate((1-x^2-a-b*x)^2,x,0,1);
diff(%,a);
diff(%th(2),b);
solve([%o2=0,%o3=0],[a,b]);
```

   (b) Verify your results by generating data that follows the above model,

and calculating the OLS estimator. When the sample size is very large, the estimator should be very close to the analytical results you obtained in the previous question.

2. Use the asymptotic normality theorem to find the asymptotic distribution of the ML estimator of $\beta^0$ for the model $y = x\beta^0 + \varepsilon$, where $\varepsilon \sim N(0,1)$ and is independent of $x$. This means finding $\frac{\partial^2}{\partial\beta\partial\beta'} s_n(\beta)$, $\mathcal{J}(\beta^0)$, $\left.\frac{\partial s_n(\beta)}{\partial\beta}\right|$, and $\mathcal{I}(\beta^0)$. The expressions may involve the unspecified density of $x$.

# Chapter 14

# Application: a simple DSGE model

This short chapter presents a simple DSGE model, which will be used to illustrate several estimators: ML, GMM, VARs, and Bayesian methods. DSGE models are quite widely used by central banks, etc., but they are not without their critics: see Paul Romer's WP "The Trouble with Macroeconomics", for example. I like the

DSGE model as an example because it allows illustrating a variety of econometric techniques and methods. Any other structural nonlinear model could serve the same purpose.

- To build an econometric model and to know how to interpret the results, it is very useful to have an economically meaningful model in mind, at least vaguely.
  - selecting variables, lags, moment conditions, instruments, etc.
  - here, we will have an explicit model as a reference point. Often, the reference is not so clearly defined.
- We will investigate structural estimation methods, which attempt to estimate the actual parameters of the data generating process, and reduced form methods, which characterize the data, but which do not recover the parameters of the data generating process.

- Knowing the true DGP will allow us to measure sensibly how well the different methods work. This chapter describes the dgp for a fairly simple structural model.

## 14.1 The model

**The model is as follows:**

- The consumer chooses consumption, hours of work, and investment to maximize expected discounted utility.
- Using capital and labor provided by the consumer, a competitive firm produces an output to maximize profits, and pays the consumer according to the marginal productivity of the inputs.
- The price of the consumption good is normalized to one.

Table 14.1: Variables

| | |
|---|---|
| $y$ | output |
| $c$ | consumption |
| $k$ | capital |
| $i$ | investment |
| $n$ | hours |
| $w$ | return to labor |
| $r$ | return to capital |
| $\eta$ | preference shock |
| $z$ | technology shock |

## Variables

There are 9 endogenous variables, listed in Table 14.1.

Table 14.2: Parameters

| | |
|---|---|
| $\alpha$ | production |
| $\beta$ | discount |
| $\delta$ | depreciation |
| $\gamma$ | risk aversion |
| $\psi$ | MRS |
| $\rho_z$ | persistence technology shock |
| $\sigma_z$ | variability technology shock |
| $\rho_\eta$ | persistence preference shock |
| $\sigma_\eta$ | variability preference shock |

## Parameters

There are 9 parameters, listed in Table 14.2

**Consumer's problem**

At the beginning of period $t$, the household owns a given amount of capital, $k_t$, and chooses $c_t$, $i_t$ and $n_t$ to maximize expected discounted utility

$$E_t \sum_{s=0}^{\infty} \beta^s \left( \frac{c_{t+s}^{1-\gamma}}{1-\gamma} + (1 - n_{t+s})\eta_t \psi \right)$$

- subject to the budget constraint $c_t + i_t = r_t k_t + w_t n_t$
- available time $0 \leq n_t \leq 1$
- and the accumulation of capital $k_{t+1} = i_t + (1-\delta)k_t$ : investment and depreciation
- There is a shock, $\eta_t$, that affects the desirability of leisure relative to consumption
  - The shock evolves according to $\ln \eta_t = \rho_\eta \ln \eta_{t-1} + \sigma_\eta \epsilon_t$.

- sometimes, people want to work more, and sometimes, they want to take it easy. There is some persistence in this mood.

**Firm's problem**

The competitive firm maximizes profits $y_t - w_t n_t - r_t k_t$ from production of the good $y_t$, taking $w_t$ and $r_t$ as given, using the constant returns to scale technology (remember that the price of $y$ is normalized to 1).

$$y_t = k_t^{\alpha} n_t^{1-\alpha} z_t \tag{14.1}$$

- Technology shocks $z_t$ also follow an AR(1) process in logarithms: $\ln z_t = \rho_z \ln z_{t-1} + \sigma_z u_t$.

**Comments:**

- The innovations to the preference and technology shocks, $\epsilon_t$ and $u_t$, are both i.i.d. standard normal random variables, and are independent of one another.
- The good $y_t$ can be allocated by the consumer to consumption or investment: $y_t = c_t + i_t$.
- The consumer provides capital and labor to the firm, and is paid at the rates $r_t$ and $w_t$, respectively.
- The representative agent chooses actions in period $t$ using rational expectations, with full information about all variables indexed $t-1$ and earlier.
- The variables available for estimation are $y, c, n, w$, and $r$. We will see that $k$ can be recovered from these.
- the model is nonlinear in the parameters, equations depend on multiple endogenous variables, and 4 of the endogenous variables are not observed by

the econometrician (the two shocks, capital and investment).

**First order conditions**

Four definitions are

$$
\begin{aligned}
\text{marginal utility of consumption: } MUC_t &:= c_t^{-\gamma} \\
\text{marginal utility of leisure: } MUL_t &:= \psi \eta_t \\
\text{marginal rate of substitution: } MRS_t &:= MUC_t / MUL_t \\
\text{marginal product of labor: } MPL_t &:= (1-\alpha)\, z_t k_t^{\alpha} n_t^{-\alpha} \\
\text{marginal product of capital: } MPK_t &:= \alpha z_t k_t^{\alpha-1} n_t^{1-\alpha}
\end{aligned}
$$

With these definitions, the 9 equations that characterize the solution for the 9 endogenous variables are:

$$
\begin{aligned}
MUC_t &= E\left(\beta \cdot MUC_{t+1}\left[1 + r_{t+1} - \delta\right]\right) && (14.2)\\
1/MRS_t &= w_t && (14.3)\\
w_t &= MPL_t && (14.4)\\
r_t &= MPK_t && (14.5)\\
\ln \eta_t &= \rho_\eta \ln \eta_{t-1} + \sigma_\eta \epsilon_t && (14.6)\\
\ln z_t &= \rho_z \ln z_{t-1} + \sigma_z u_t && (14.7)\\
y_t &= z_t k_t^{\alpha} n_t^{(1-\alpha)} && (14.8)\\
i_t &= y_t - c_t && (14.9)\\
k_{t+1} &= i_t + (1-\delta) k_t && (14.10)
\end{aligned}
$$

where the first two are from utility maximization, the second two are from profit maximization, and the remaining 5 are directly from the model.

### The steady state

We use a third-order perturbation solution, which combines good accuracy with moderate computational demands Aruoba et al. (2006). This is done using SolveDSGE.jl, a registered Julia package which can be added using `]add SolveDSGE`. A first step for solving the model is to find the deterministic steady state. SolveDSGE.jl can compute the steady state itself, given a reasonably good starting value, but it is faster to use an analytic solution, if one can be computed. In this case, it can be done.

- the deterministic steady state is the equilibrium value of each variable that obtains when all shocks are set to 0.

Let variables without the $t$ subscript indicate the deterministic steady state level of the variable. The deterministic steady state values of the two shocks $\eta$ and $z$ are both 1.

- For example, take $z$. The assumption is that $\ln z_t = \rho_z \ln z_{t-1} + \sigma_z u_t$. If we set $u_t = 0$, to make the equation deterministic, we get $\ln z_t = \rho_z \ln z_{t-1}$.

Drop the subscript to reflect the equilibrium condition: $\ln z = \rho_z \ln z$. This only holds when $\ln z = 0$, for abritrary values of $\rho_z$. So, the steady state value of $z = \exp(0) = 1$.

Using equations 14.3 and 14.4, dropping $t$ subscripts, and setting the two shocks to 1, we obtain

$$\begin{aligned} \frac{\psi}{c^{-\gamma}} &= (1-\alpha)\, k^{\alpha} n^{-\alpha} \\ n &= \left(\frac{1-\alpha}{\psi}\right)^{1/\alpha} k c^{-\gamma/\alpha} \end{aligned} \tag{14.11}$$

Define

$$\theta := \left(\frac{1-\alpha}{\psi}\right)^{1/\alpha} \tag{14.12}$$

so

$$n = \theta k c^{-\gamma/\alpha} \tag{14.13}$$

From the Euler equation (equation 14.2)

$$\begin{aligned} c &= \beta c\left(1+\alpha k^{\alpha-1}n^{1-\alpha}\right) \\ n^{1-\alpha} &= k^{1-\alpha}\left[\frac{1}{\alpha}\left(\frac{1}{\beta}-1+\delta\right)\right] \\ n &= k\left[\frac{1}{\alpha}\left(\frac{1}{\beta}-1+\delta\right)\right]^{\frac{1}{1-\alpha}} \end{aligned}$$

Define

$$\varphi := \left[\frac{1}{\alpha}\left(\frac{1}{\beta}-1+\delta\right)\right]^{\frac{1}{1-\alpha}} \tag{14.14}$$

so

$$n = \varphi k \tag{14.15}$$

Now set equations 14.13 and 14.15 equal, and solve for steady state level of con-

sumption:

$$\begin{aligned}
\theta k c^{-\gamma/\alpha} &= \varphi k \\
c^{-\gamma/\alpha} &= \frac{\varphi}{\theta} \\
c &= \left(\frac{\theta}{\varphi}\right)^{\frac{\alpha}{\gamma}}
\end{aligned}$$

From equation 14.10, steady state investment satisfies

$$i = \delta k$$

and combining this with equations 14.8 and 14.9, we obtain (using equation 14.15)

$$c + \delta k = k^{\alpha}(\varphi k)^{1-\alpha}$$

which solves as

$$k = \frac{c}{\varphi^{1-\alpha} - \delta}$$

Table 14.3: Deterministic steady state

| Variable | Description | Steady state |
|---|---|---|
| $y$ | output | $k^{\alpha}n^{1-\alpha}$ |
| $c$ | consumption | $c = \left(\frac{\theta}{\varphi}\right)^{\frac{\alpha}{\gamma}}$ |
| $k$ | capital | $k = \frac{c}{\varphi^{1-\alpha}-\delta}$ |
| $i$ | investment | $y - c$ |
| $n$ | hours | $n = \varphi k$ |
| $w$ | return to labor | $(1-\alpha)\, k^{\alpha}n^{-\alpha}$ |
| $r$ | return to capital | $\alpha k^{\alpha-1}n^{1-\alpha}$ |
| $\eta$ | preference shock | 1 |
| $z$ | technology shock | 1 |

To summarize, the steady state values for the 9 endogenous variables are given in Table 14.3, and given the 9 parameters of the model, these can be solved for in the order $c$, $k$, $n$, $y$, $i$, $w, r$.

## True parameter values and priors

- For this model, given the observable variables, we can recover $\alpha = 1 - \frac{wn}{y}$, by substituting the production function into the MPL or MPK.

- Once we have $\alpha$, we can solve for $k$, by substituting the production function into the $r = MPK$ equation.

- with $k$, we can recover $\delta$, from the law of motion of capital.

- So, we will take $\alpha$ and $\delta$ as known parameters, as they can be recovered exactly from the observable data (assuming there is no measurement error, which is the case we are in). Henceforth, we set $\alpha = 0.33$ and $\delta = 0.025$, which are commonly used in the literature.

Estimation by GMM does not require specifying a prior distribution for the parameters, but Bayesian methods do, so here, I discuss the true parameter values used for the simulations, as well as priors. Economic intuition can guide choice of priors for most parameters. However, we may be less confident specifying a prior for the marginal rate of substitution, $\psi$.

- Instead, following Ruge-Murcia (2012), we may place a prior on steady state hours, $n$.
- Given that hours $n_t$ must satisfy the constraint $0 \leq n_t \leq 1$, and we know that normally around 8 hours per day is dedicated to work, it is relatively straightforward to place a prior on $n$.
- If steady state hours, $n$ is given, say as a draw from it's prior, then this, along with the parameters, excepting $\psi$, allows us to solve for the steady state values of the other variables and for $\psi$, as follows:
  - Given $n$, we can compute $k = n/\varphi$, then $i = \delta k$, then $y = k^{\alpha} n^{1-\alpha}$,

then $c = y - i$

- finally, using equation 14.11, we can compute the $\psi$ that is consistent with the given steady state $n$ and the other parameters of the model as

$$\psi \;=\; c^{-\gamma}\,(1-\alpha)\,k^{\alpha}n^{-\alpha}$$

We use independent uniform priors for all parameters except $\psi$, and a uniform prior for steady state hours, $n$. The true values of the parameters and the supports of the uniform priors are given in Table 14.4. We believe that most economists will find these priors to be quite loose, and the parameter values to be reasonable.

Table 14.4: True parameters and support of uniform priors.

| Parameter | Lower bound | Upper bound | True value |
|---|---|---|---|
| $\beta$ | 0.95 | 0.995 | 0.990 |
| $\gamma$ | 0 | 5 | 2.000 |
| $\rho_z$ | 0 | 0.995 | 0.900 |
| $\sigma_z$ | 0 | 0.1 | 0.020 |
| $\rho_\eta$ | 0 | 0.995 | 0.700 |
| $\sigma_\eta$ | 0 | 0.1 | 0.01 |
| $\bar{n}$ | 6/24 | 9/24 | 1/3 |

## 14.2 Solution of the model and generation of data

Given a draw of the parameters other than $\psi$, $\psi$ is computed as above, and we can form the full parameter vector, $\theta^s$, where $s$ indexes the simulations. The model is solved using SolveDSGE.jl, using a third order perturbation about the steady state. Once the model is solved, a simulation of length 360 is done, initialized at the steady state. We drop the first 200 observations, retaining the last 160 observations, which mimic 40 years of quarterly data. The observable variables are $y$, $c$, $n$, $w$, and $r$. The selection of observable variables is in line with much empirical work (e.g., Smets and Wouters (2007), Guerrón-Quintana (2010)).

The model file that SolveDSGE.jl uses is CK.txt . It's remarkably close to the way the equations were presented above, I find. With this file, we can generate samples at the chosen parameter values. The script that generates the data file dsgedata.txt is GenData.jl . Here's a plot of the 160 observations of the five

Figure 14.1: The DSGE data

observed variables.

- recall that the true parameter values that generate this data were described above.

- Note that consumption is much more smooth than output
- the interest rate is pretty much constant, hours worked is a little more responsive
- wages move around quite a bit in response to shocks
- this data file will be treated as the true sample file in the examples that follow.

# Chapter 15

# Maximum likelihood estimation

Cameron and Trivedi (2005), Ch. 5

The maximum likelihood estimator is important because it uses all of the information in a fully specified statistical model. Its use of all of the information causes it to have a number of attractive properties, foremost of which is *asymptotic*

*efficiency*. For this reason, the ML estimator can serve as a benchmark against which other estimators may be measured. The ML estimator requires that the statistical model be fully specified, which essentially means that there is enough information to draw data from the DGP, given the parameter. This is a fairly strong requirement, and for this reason we need to be concerned about the possible misspecification of the statistical model. If this is the case, the ML estimator will not have the nice properties that it has under correct specification.

## 15.1 The likelihood function

Suppose we have a sample of size $n$ of the random vectors $y$ and $z$. Suppose the joint density of $Y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix}$ and $Z = \begin{pmatrix} z_1 & \dots & z_n \end{pmatrix}$ is characterized by a parameter vector $\psi_0$ :

$$f_{YZ}(Y, Z, \psi_0).$$

This is the joint density of the sample (note: it's the same as what we see in eq. 13.2, except we're partitioning the data into two groups). The *likelihood function* is just this density evaluated at other values $\psi$

$$L(Y, Z, \psi) = f(Y, Z, \psi), \psi \in \Psi,$$

where $\Psi$ is a *parameter space.*

The *maximum likelihood estimator* of $\psi_0$ is the value of $\psi$ that maximizes the likelihood function.

So, when we're doing ML, we assume we're in the Case 1 of the discussion

around eq. 13.2. Maybe we could call this estimator the maximum assumptions estimator, as we assume everything about the data generating process is known, except for the particular values of the parameter vector.

**Example 38.** Count data. Suppose we have a sample $Y = \{y_1, ..., y_n\}$ where the data are counts: the number of times some event occurs in a given interval of time, e.g., number of visits to the doctor in a year. The simplest count data density is the Poisson:

$$f_Y(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

If the observations are i.i.d. distributed according to this density, then the joint density of the sample is

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\Sigma y_i}}{\Pi_i y_i!}$$

A little calculus and algebra shows us that the value that maximizes this is $\hat{\lambda} = \bar{y}$.

**Exercise 39.** Prove this last statement

- In Example 38 we can compute the ML estimator without much trouble, and we have asymptotic theory that will allow us to test hypotheses about $\lambda$, because the estimator is the sample mean, and the LLN and CLT will apply.

- however, now suppose that each observation has its own $\lambda_i = \exp(x_i'\beta)$, which depends on conditioning variables and a new parameter vector. We can now write the likelihood function in terms of $\beta$ (as was done in Section 12.4)

  - The problem is, we can't find an analytic solution for the ML estimator of $\beta$.
  - Even if we could, the $\hat{\beta}$ which solves the f.o.c. is a nonlinear function of the data, rather than a simple average. How could we test hypotheses? The t and F tests developed for the classical linear model do not apply.
  - To solve these two problems, we need the methods from Ch. 11 and

the theory from Ch. 12.

## Exogenous variables

Engle et al. (1983) is a good reference for this part. The likelihood function can be factored as

$$f_{YZ}(Y, Z, \psi) = f_{Y|Z}(Y|Z, \theta) f_Z(Z, \rho)$$

where $\theta$ are whatever elements of $\psi$ that happen to enter in the conditional density, and $\rho$ are the elements that enter into the marginal density.

Note that if $\theta$ and $\rho$ share no elements, then the maximizer of the conditional likelihood function $f_{Y|Z}(Y|Z, \theta)$ with respect to $\theta$ is the same as the maximizer of the overall likelihood function $f_{YZ}(Y, Z, \psi) = f_{Y|Z}(Y|Z, \theta) f_Z(Z, \rho)$, for the elements of $\psi$ that correspond to $\theta$.

- In this case, the variables $Z$ are said to be *exogenous* for estimation of $\theta$, and we may more conveniently work with the conditional likelihood function $f_{Y|Z}(Y|Z, \theta)$ for the purposes of estimating $\theta_0$.

- With exogeneity of $Z$, the maximum likelihood estimator of $\theta_0$ will be

$\arg\max f_{Y|Z}(Y|Z, \theta)$.

- We'll suppose this framework holds in what follows. If it didn't, for some variables in $Z$, then just move those variables from $Z$ to $Y$, until it does hold.

# A convenient factorization of the likelihood function

- If the $n$ observations are independent, the likelihood function can be written as

$$L(Y|Z,\theta) = \prod_{t=1}^{n} f(y_t|z_t,\theta)$$

- If this is not possible, we can always factor the likelihood into *contributions of observations,* by using the fact that a joint density can be factored into the product of a marginal and conditional.

- Then

$$\underbrace{f(y_1,y_2,\ldots y_{n-1},y_n|Z,\theta)}_{\text{joint}} = \underbrace{f(y_n|y_1,y_2,\ldots y_{n-1},Z,\theta)}_{\text{conditional}} \underbrace{f(y_1,y_2,\ldots y_{n-1}|Z,\theta)}_{\text{marginal}}$$

- do the same thing for $y_{n-1}$ in the last term, and keep iterating.

- Then, in the end, we have

$$
\begin{aligned}
L(Y|Z,\theta) = & f(y_n|y_{1,}y_2,\ldots y_{n-1},Z,\theta) \\
& \cdot f(y_{n-1}|y_{1,}y_2,\ldots y_{n-2},Z,\theta) \\
& \cdots \\
& \cdot f(y_2|y_1,Z,\theta) \\
& \cdot f(y_1|Z,\theta)
\end{aligned}
$$

To simplify notation, define

$$x_t \;=\; \{y_1, y_2, ..., y_{t-1}, Z\}$$

- Usually, for time series data, conditional densities depend only on current period exogenous variables, as the effects of lagged exogenous variables are transmitted though the realizations of the lagged endogenous variables, and economic models normally don't involve leads of exogenous variables. If

this is the case, $x_1 = z_1$, $x_2 = \{y_1, z_2\}$, *etc.* - it contains exogenous and predetermined endogenous variables.

- it is also often the case that the data is Markovian, which means that only a limited number of lags of $y$ affect the current value. When this is the case, if $m$ is the maximum lag that still matters, then

$$x_t = \{y_{t-m}, y_{t-m+1}, ..., y_{t-1}, Z\}$$

  when $t > m$. (Treatment of the observations where $t \leq m$ is a bit complicated - these observations are often dropped, to keep things simple. Here, we will not worry about this problem.)

- Regardless of the specific contents of $x_t$, the likelihood function can now be written as

$$L(Y|Z;\theta) = \prod_{t=1}^{n} f(y_t|x_t, \theta)$$

## The log likelihood function

The criterion function can be defined as the average log-likelihood function:

$$s_n(\theta) = \frac{1}{n} \ln L(Y|Z;\theta) = \frac{1}{n} \sum_{t=1}^{n} \ln f(y_t|x_t;\theta) \tag{15.1}$$

The maximum likelihood estimator may thus be defined equivalently as

$$\hat{\theta} = \arg\max s_n(\theta),$$

where the set maximized over is defined below. Since $\ln(\cdot)$ is a monotonic increasing function, $\ln L$ and $L$ maximize at the same value of $\theta$. Dividing by $n$ has no effect on $\hat{\theta}$.

- *Question: why do we do it, then?* There are both theoretical and practical reasons:
  - to get a LLN to apply: LNNs apply to averages of terms, not products

- to avoid loss of precision on a digital computer: the likelihood function in product form will tend rapidly to zero when each term is between 0 and 1. For a discrete r.v., this will be the case, and it's usually the case for continuous R.V.s, too, unless they're highly concentrated. When this happens, then, as the sample size gets larger, the objective function gets smaller, and, before long, will dip below machine precision.

**Example 40.** Example: Bernoulli trial

Suppose that we are flipping a coin that may be biased, so that the probability of a heads may not be 0.5. Maybe we're interested in estimating the probability of a heads. Let $Y = 1(heads)$ be a binary variable that indicates whether or not a heads is observed. The outcome of a toss is a Bernoulli random variable:

$$\begin{aligned} f_Y(y, p_0) &= p_0^y (1 - p_0)^{1-y}, y \in \{0, 1\} \\ &= 0, y \notin \{0, 1\} \end{aligned}$$

So a representative term that enters the likelihood function is

$$f_Y(y, p) = p^y (1 - p)^{1-y}$$

and

$$\ln f_Y(y, p) = y \ln p + (1 - y) \ln (1 - p)$$

For this example, the average log-likelihood function is

$$s_n(p) = \frac{1}{n} \sum_{t=1}^{n} y_t \ln p + (1 - y_t) \ln (1 - p) \,.$$

* To explore the behavior of the likelihood, the log-likelihood, and the average log-likelihood, see the code PlainLF.jl

The derivative of a representative term is

$$\begin{aligned}\frac{\partial \ln f_Y(y,p)}{\partial p} &= \frac{y}{p} - \frac{(1-y)}{(1-p)} \\ &= \frac{y-p}{p\,(1-p)}\end{aligned}$$

so, averaging this over a sample of size $n$, the gradient is

$$\frac{\partial s_n(p)}{\partial p} = \frac{1}{n}\sum_{t=1}^{n} \frac{y_t - p}{p\,(1-p)}.$$

Setting to zero and solving gives

$$\hat{p} = \bar{y} \tag{15.2}$$

So it's easy to calculate the MLE of $p_0$ in this case.

- We also know that the sample mean converges almost surely to the true mean, from basic statistics (LLN).
- The mean is $E(Y) = \Sigma_{y=0}^{y=1} y p_0^y (1 - p_0)^{1-y} = p_0$.
- Thus, the MLE, which is the sample mean, is a consistent estimator of the parameter, because the population mean is the parameter..
- For future reference, note that $Var(Y) = E(Y^2) - [E(Y)]^2 = p_0 - p_0^2$.

We see that the ML estimator is consistent, for this model. However, let's verify that the consistency theorem for extremum estimators gives us the same result:

- A LLN tells us that, for a given $p$, the objective function converges to the limit of its expectation:

$$s_n(p) = \frac{1}{n} \sum_{t=1}^{n} y_t \ln p + (1 - y_t) \ln (1 - p) \rightarrow^{a.s.} p_0 \ln p + (1 - p_0) \ln(1 - p).$$

- The parameter space must be compact. We know that $p_0$ lies between 0 and 1, so this helps set the parameter space.

- the objective function is obviously continuous in the parameter

- we need the objective function to be bounded, for the simple sufficient conditions for the consistency theorem to hold. So, we need to assume that $p$ can't go to 0 or to 1. This means that the parameter space must be a compact subset of $(0, 1)$.

- With these conditions, the a.s. convergence is also uniform.

- The consistency theorem for extremum estimators tells us that the ML estimator converges to the value that maximizes the limiting objective function. Because

$$s_\infty(p) = p_0 \ln p + (1 - p_0) \ln(1 - p),$$

  we can easily check that the unique maximizer is $p_0$, by computing first and second derivatives.

- So, the three assumptions of the consistency theorem hold, and thus the ML estimator is consistent for the true probability.

**Example 41.** *Likelihood function and MLE of classical linear regression model.* Let's suppose that a dependent variable is normally distributed: $y \sim N(\mu_0, \sigma_0^2)$, so

$$f_y(y; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y-\mu_0)^2}{2\sigma_0^2}\right)$$

Suppose that the mean, $\mu_0$, depends on some regressors, $x$. The simplest way to do this is to assume that $\mu_0 = x'\beta_0$. With this, the density, conditional on $x$ is

$$f_y(y|x; \beta_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y-x'\beta_0)^2}{2\sigma_0^2}\right)$$

This is an example of *parameterization* of a density, making some parameters depend on additional variables and new parameters. With an i.i.d. sample of size $n$, the overall conditional density is the product of the conditional density of each observation:

$$f_y(y_1, y_2, ..., y_n|x_1, x_2, ..., x; \beta_0, \sigma_0^2) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y_t - x_t'\beta_0)^2}{2\sigma_0^2}\right)$$

Taking logarithms, and evaluating at some point in the parameter space, we get the log-likelihood function:

$$\ln L(Y|X;\beta,\sigma^2) = -n\ln\sqrt{2\pi} - n\ln\sigma - \sum_{t=1}^{n}\frac{(y_t - x_t'\beta)^2}{2\sigma^2}$$

- Observe that the first order conditions for $\beta$ are the same as for the OLS estimator. For the $\beta$'s, the OLS and ML estimators are identical.

- We know that the OLS estimator is consistent without making distributional assumptions regarding the errors. As long as the assumptions for consistency of OLS hold (fundamentally, weak exogeneity), then the "ML" estimator will be consistent for $\beta$ as well, even if the normality assumption is not correct. This would be an example of *quasi-maximum likelihood* estimation: "ML" estimation of a misspecified model. Sometimes the QML estimator is consistent, sometimes it's not.

- A Julia example that shows how to compute the maximum likelihood estima-

tor for data that follows the CLRM with normality is in NormalExample.jl . Examine the code, and figure out how the likelihood function is defined.

## 15.2 Consistency of MLE

- The MLE is an extremum estimator, given basic assumptions it is consistent for the value that maximizes the limiting objective function, following Theorem 35.

- The question is: what is the value that maximizes $s_\infty(\theta)$ when the criterion function is the average log-likelihood?

- For two cases (Bernoulli trial and ML of the linear model with normality) we have seen that the ML estimator converges to the true parameter of the d.g.p.

- Is this a general result?

Remember that $s_n(\theta) = \frac{1}{n} \ln L(Y|Z, \theta)$, and $L(Y|Z, \theta_0)$ is the true density of the sample data. For any $\theta \neq \theta_0$

$$\mathcal{E}\left(\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right) \leq \ln\left(\mathcal{E}\left(\frac{L(\theta)}{L(\theta_0)}\right)\right)$$

by Jensen's inequality ( $\ln(\cdot)$ is a concave function).

Now, the expectation on the RHS is

$$\mathcal{E}\left(\frac{L(\theta)}{L(\theta_0)}\right) = \int \frac{L(\theta)}{L(\theta_0)} L(\theta_0) dy = 1,$$

since $L(\theta_0)$ *is* the density function of the observations, and since the integral of any density is 1.

Therefore, since $\ln(1) = 0$,

$$\mathcal{E}\left(\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right) \leq 0,$$

or (both sides have implicitly been multiplied by $1/n$)

$$\mathcal{E}\left(s_n\left(\theta\right)\right) - \mathcal{E}\left(s_n\left(\theta_0\right)\right) \leq 0$$

or

$$\mathcal{E}\left(s_n\left(\theta\right)\right) \leq \mathcal{E}\left(s_n\left(\theta_0\right)\right)$$

Taking limits of each side:

$$s_\infty(\theta) \leq s_\infty(\theta_0)$$

except on a set of zero probability (by assumption b of Theorem 35). So the true parameter value is the maximizer of the limiting objective function (we are in Case 1 of the three cases discussed above - a fully correctly specified model).

If the identification assumption holds, then there is a unique maximizer, so the inequality is strict if $\theta \neq \theta_0$:

$$s_\infty(\theta) < s_\infty(\theta_0), \forall \theta \neq \theta_0, \text{a.s.}$$

Therefore, $\theta_0$ is the unique maximizer of $s_\infty(\theta)$, and thus, Theorem 35 tells us that

$$\lim_{n \to \infty} \hat{\theta} = \theta_0, \text{ a.s.}$$

So, the ML estimator is consistent for the true parameter value. In practice, we

will need to check identification for the specific model under consideration.

**Exercise 42.** Verify by experiment the consistency of the ML estimator of the CLRM with normality by increasing $n$ in the example NormalExample.jl .

## 15.3 The score function

**Assumption:** (Differentiability) Assume that $s_n(\theta)$ is twice continuously differentiable in a neighborhood $N(\theta_0)$ of $\theta_0$, at least when $n$ is large enough.

- with this, and with the result on consistency from above, assumptions (a) and (b) of theorem 37 hold, and $\mathcal{J}_n(\hat{\theta}) \overset{a.s.}{\to} \mathcal{J}_\infty(\theta^0)$

To maximize the log-likelihood function, take derivatives:

$$\begin{aligned} g_n(Y|Z,\theta) &\equiv D_\theta s_n(\theta) \\ &= \frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f(y_t|x_t,\theta) \\ &\equiv \frac{1}{n}\sum_{t=1}^{n} g_t(\theta). \end{aligned} \tag{15.3}$$

This is the *score vector* (with dim $K \times 1$). Note that the score function has $Y$ as an argument, which implies that it is a random function. $Y$ (and any exogeneous

variables) will often be suppressed for clarity, but one should not forget that they are still there.

Note that

$$s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \ln f(y_t|x_t, \theta)$$

If we define the contributions of each observation to the objective function as

$$s^{(t)} = \ln f(y_t|x_t, \theta)$$

then the objective function is the average of these terms. We can arrange the contributions to the objective function into the $n$-vector

$$s^{(1:n)} = \begin{bmatrix} s^{(1)} \\ s^{(2)} \\ \vdots \\ s^{(n)} \end{bmatrix}$$

Note that the Jacobian of this vector is an $n \times K$ matrix with the score contribu-

tions $g_t(\theta) = \frac{\partial}{\partial\theta} s^{(t)}$ (transposed) in its rows:

$$\frac{\partial s^{(1:n)}}{\partial \theta'} = \begin{bmatrix} g_1'(\theta) \\ g_2'(\theta) \\ \vdots \\ g_n'(\theta) \end{bmatrix}$$

The averages of the columns of this matrix are the elements of the overall score vector.

The ML estimator $\hat{\theta}$ sets the derivatives to zero:

$$g_n(\hat{\theta}) = \frac{1}{n} \sum_{t=1}^{n} g_t(\hat{\theta}) \equiv 0.$$

We will show that $E_\theta \left[ g_t(\theta) | x_t \right] = 0$, $\forall t$. *This is the expectation taken with respect to the density* $f(y_t|x_t, \theta)$, not necessarily $f\left(y_t|x_t, \theta_0\right)$.

$$\begin{aligned} E_\theta \left[ g_t(\theta) | x_t \right] &= \int [D_\theta \ln f(y_t|x_t, \theta)] f(y_t|x, \theta) dy_t \\ &= \int \frac{1}{f(y_t|x_t, \theta)} \left[ D_\theta f(y_t|x_t, \theta) \right] f(y_t|x_t, \theta) dy_t \\ &= \int D_\theta f(y_t|x_t, \theta) dy_t. \end{aligned}$$

Given some regularity conditions on boundedness of $D_\theta f$, we can switch the order of integration and differentiation, by the dominated convergence theorem (Gallant

(1997), Theorem 2.8). This gives

$$\begin{aligned} E_\theta\left[g_t(\theta)|x_t\right] &= D_\theta \int f(y_t|x_t,\theta)dy_t \\ &= D_\theta 1 \\ &= 0 \end{aligned} \tag{15.4}$$

where we use the fact that the integral of the density is 1.

- So $E_\theta(g_t(\theta)|x_t) = 0$ : *the conditional expectation of the score vector is zero*. Because this is true for all $x_t$, the unconditional expectation is also zero.

- This hold for all $t$, so it implies that $\mathcal{E}_\theta g_n(Y|Z,\theta) = 0$.

**Proposition 43.** *Law of iterated expectations. Suppose that a random variable $Y$ has zero expectation conditional on the random variable $X$*

$$\mathcal{E}_{Y|X}Y = \int Y f(Y|X) dY = 0$$

*Then the unconditional expectation of the product of $Y$ and a function $g(X)$ of $X$ is also zero.*

To see this, the unconditional expectation is

$$\mathcal{E}\left(Yg(X)\right) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} Yg(X)f(Y,X)dY\right) dX.$$

Factor the joint density in to conditional and marginal:

$$\mathcal{E}\left(Yg(X)\right) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} Y g(X) f(Y|X) f(X) dY\right) dX.$$

Because $f(X)$ and $g(X)$ don't depend on $Y$, they can be pulled out of the integral

$$\mathcal{E}\left(Yg(X)\right) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} Y f(Y|X) dY\right) g(X) f(X) dX.$$

But the term in parentheses on the rhs is zero by assumption, so

$$\mathcal{E}\left(Yg(X)\right) = 0$$

as claimed.

- This result implies that $E_\theta(g_t(y_t|x_t,\theta)g_{t-s}(y_{t-s}|x_{t-s},\theta)') = 0_{k\times k}$: the score contributions are uncorrelated with one another.

- This comes from the fact that all random variables in $g_{t-s}(y_{t-s}|x_{t-s},\theta)$ are in the information set $x_t$, and applying iterated expectations.

- If one computes score contributions, and observes that they appear to be correlated, this casts doubt on the correct specification of the model. There are formal specification tests based on this result (White (1982)).

## 15.4 Asymptotic normality of MLE

Recall that we assume that the log-likelihood function $s_n(\theta)$ is twice continuously differentiable. Take a first order Taylor's series expansion of $g(Y, \hat{\theta})$ about the true value $\theta_0$ :

$$0 \equiv g(\hat{\theta}) = g(\theta_0) + (D_{\theta'} g(\theta^*)) \left(\hat{\theta} - \theta_0\right)$$

or with appropriate definitions

$$\mathcal{J}(\theta^*) \left(\hat{\theta} - \theta_0\right) = -g(\theta_0),$$

where $\theta^* = \lambda\hat{\theta} + (1 - \lambda)\theta_0, 0 < \lambda < 1$. Assume $\mathcal{J}(\theta^*)$ is invertible (we'll justify this in a minute). So

$$\sqrt{n} \left(\hat{\theta} - \theta_0\right) = -\mathcal{J}(\theta^*)^{-1} \sqrt{n} g(\theta_0) \tag{15.5}$$

Now consider $\mathcal{J}(\theta^*)$, the matrix of second derivatives of the average log likeli-

hood function. This is

$$\begin{aligned}\mathcal{J}(\theta^*) &= D_{\theta'} g(\theta^*) \\ &= D_\theta^2 s_n(\theta^*) \\ &= \frac{1}{n} \sum_{t=1}^{n} D_\theta^2 \ln f_t(\theta^*)\end{aligned}$$

where the notation

$$D_\theta^2 s_n(\theta^*) \equiv \left. \frac{\partial^2 s_n(\theta)}{\partial\theta\partial\theta'} \right|_{\theta=\theta^*}.$$

- Given that this is an average of terms, it should usually be the case that this satisfies a strong law of large numbers (SLLN).

- *Regularity conditions* are a set of assumptions that guarantee that this will happen. There are different sets of assumptions that can be used to justify appeal to different SLLN's. For example, the $D_\theta^2 \ln f_t(\theta^*)$ must not be too strongly dependent over time, and their variances must not become infinite. We don't assume any particular set here, since the appropriate assumptions

will depend upon the particularities of a given model. However, we assume that a SLLN applies.

Also, since we know that $\hat{\theta}$ is consistent, and since $\theta^* = \lambda\hat{\theta} + (1-\lambda)\theta_0$, we have that $\theta^* \overset{a.s.}{\to} \theta_0$. Also, by the above differentiability assumption, $\mathcal{J}(\theta)$ is continuous in $\theta$. Given this, $\mathcal{J}(\theta^*)$ converges to the limit of its expectation:

$$\mathcal{J}(\theta^*) \overset{a.s.}{\to} \lim_{n\to\infty} \mathcal{E}\left(D_\theta^2 s_n(\theta_0)\right) = \mathcal{J}_\infty(\theta_0) < \infty$$

*This matrix converges to a finite limit.*

Re-arranging orders of limits and differentiation, which is legitimate given certain regularity conditions related to the boundedness of the log-likelihood function, we get

$$\begin{aligned}\mathcal{J}_\infty(\theta_0) &= D^2_\theta \lim_{n\to\infty} \mathcal{E}\left(s_n(\theta_0)\right) \\ &= D^2_\theta s_\infty(\theta_0)\end{aligned}$$

From assumption (c) of the consistency proof,

$$s_\infty(\theta) < s_\infty(\theta_0)$$

*i.e.,* $\theta_0$ maximizes the limiting objective function. Since there is a unique maximizer, and by the assumption that $s_n(\theta)$ is twice continuously differentiable (which holds in the limit), then $\mathcal{J}_\infty(\theta_0)$ must be negative definite, and therefore of full rank. When $n$ is large enough so that we are close enough to $\theta_0$,the inversion of

$\mathcal{J}(\theta^*)$ is justified, so we get

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\mathcal{J}(\theta^*)^{-1}\sqrt{n}g(\theta_0). \tag{15.6}$$

Now consider $\sqrt{n}g(\theta_0)$. For assumption (c) of Theorem 37 to apply, this quantity must follow a Central Limit Theorem. We have

$$\begin{aligned}
\sqrt{n}g_n(\theta_0) &= \sqrt{n}D_\theta s_n(\theta) \\
&= \sqrt{n}\frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f_t(y_t|x_t,\theta_0) \\
&= \sqrt{n}\frac{1}{n}\sum_{t=1}^{n} g_t(\theta_0)
\end{aligned}$$

- We've already seen that $\mathcal{E}_\theta\left[g_t(\theta)\right] = 0$, for all $\theta$, and, thus, for $\theta_0$, too. Thus, the last line, without scaling by $\sqrt{n}$, would converge almost surely to zero, by a LLN (assuming the variances of the $g_t$ are bounded). To get a stable limiting distribution, we need to multiply by something that is tending to infinity, at the proper rate. It turns out that $\sqrt{n}$ is the quantity that fits the bill (see a proof of a CLT to understand why).

- Also, the elements of the sum are uncorrelated with one another (iterated expectations, above)

- As long as the $g_t$ have finite variances, a CLT will apply. Checking this requires knowing what specific model we're working with, so for this general treatment, we will assume that the model satisfies this condition.

Assuming that a CLT applies:

$$\sqrt{n} g_n(\theta_0) \xrightarrow{d} N\left[0, \mathcal{I}_\infty(\theta_0)\right] \tag{15.7}$$

where

$$\begin{aligned}
\mathcal{I}_\infty(\theta_0) &= \lim_{n\to\infty} \mathcal{E}_{\theta_0}\left(n\left[g_n(\theta_0)\right]\left[g_n(\theta_0)\right]'\right) \\
&= \lim_{n\to\infty} V_{\theta_0}\left(\sqrt{n} g_n(\theta_0)\right)
\end{aligned}$$

- $\mathcal{I}_\infty(\theta_0)$ is known as the *information matrix*. It is the asymptotic variance of the score vector

- From the previous expression $\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\mathcal{J}(\theta^*)^{-1}\sqrt{n} g(\theta_0)$, and noting that $\mathcal{J}(\theta^*) \xrightarrow{a.s.} \mathcal{J}_\infty(\theta_0)$, we see that, following the Slutsky theorem

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to^d N\left[0, \mathcal{J}_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)\mathcal{J}_\infty(\theta_0)^{-1}\right].$$

*The MLE estimator is asymptotically normally distributed.*

- The form of this expression is the same as what we saw for extremum estimators in general. This is no surprise, as the MLE is an extremum estimator.

## Example, Coin flipping, again. Finding $\mathcal{J}_\infty(p_0)$ and $\mathcal{I}_\infty(p_0)$.

In section 40 we saw that the MLE for the parameter of a Bernoulli trial, with i.i.d. data, is the sample mean: $\hat{p} = \bar{y}$ (equation 15.2). Now let's find the limiting variance of $\sqrt{n}\,(\hat{p} - p_0)$.

We can do this in a simple way:

$$
\begin{aligned}
\lim Var\sqrt{n}\left(\hat{p}-p_0\right) &= \lim nVar\left(\hat{p}-p_0\right) \\
&= \lim nVar\left(\hat{p}\right) \\
&= \lim nVar\left(\bar{y}\right) \\
&= \lim nVar\left(\frac{\Sigma\, y_t}{n}\right) \\
&= \lim \frac{1}{n}\Sigma\, Var(y_t) \text{ (by independence of obs.)} \\
&= \lim \frac{1}{n} nVar(y) \text{ (by identically distributed obs.)} \\
&= Var(y) \\
&= p_0\left(1-p_0\right) \qquad (15.8)
\end{aligned}
$$

While that is simple, let's verify this using the methods we're studing in this chapter, because this simple method is not possible to use in general.

**Finding $\mathcal{J}_\infty(p_0)$** The log-likelihood function is

$$s_n(p) \;=\; \frac{1}{n}\sum_{t=1}^{n}\left\{y_t \ln p + (1-y_t)\ln(1-p)\right\}$$

so

$$\mathcal{E}_{p_0} s_n(p) = p^0 \ln p + \left(1-p^0\right)\ln(1-p)$$

by the fact that the observations are i.i.d. Note that this does not depend on the sample size, $n$, thus,

$$s_\infty(p) = p^0 \ln p + \left(1-p^0\right)\ln(1-p)\,.$$

A bit of calculation shows that

$$\mathcal{J}_\infty(p^0) = D^2_\theta s_\infty(p)\big|_{p=p^0} = \frac{-1}{p^0\left(1-p^0\right)}.$$

(note: we used dominated convergence).

**Finding $\mathcal{I}_\infty(p_0)$** As we have seen before, the score vector is

$$g_n(p) = \frac{\partial s_n(p)}{\partial p} = \frac{1}{n}\sum_{t=1}^{n}\frac{y_t - p}{p\,(1-p)}.$$

and we wish to compute the covariance of the score vector, after weighting by root-$n$: $\mathcal{I}_\infty(p_0) = \lim_{n\to\infty}\mathcal{E}_{p_0}\left(n\left[g_n(p_0)\right]\left[g_n(p_0)\right]'\right)$. Note that, playing with the $n$s

$$\begin{aligned}\mathcal{E}_{p_0}\left(n\left[g_n(p_0)\right]\left[g_n(p_0)\right]'\right) &= n\mathcal{E}_{p_0}\left\{\left(\frac{1}{n}\sum_{t=1}^{n}\frac{y_t - p_0}{p_0\,(1-p_0)}\right)^2\right\},\\ &= \frac{1}{n}\mathcal{E}_{p_0}\sum_{t=1}^{n}\left(\frac{y_t - p_0}{p_0\,(1-p_0)}\right)^2\end{aligned}$$

Next, the terms in the sum are i.i.d., so the expection of the sum is $n$ times the expectation of a representatiive term, so

$$\begin{aligned}
\frac{1}{n}\mathcal{E}_{p_0}\sum_{t=1}^{n}\left(\frac{y_t-p_0}{p_0\left(1-p_0\right)}\right)^2 &= \frac{1}{n}n\mathcal{E}_{p_0}\left(\frac{y-p_0}{p_0\left(1-p_0\right)}\right)^2 \\
&= \mathcal{E}_{p_0}\left(\frac{y-p_0}{p_0\left(1-p_0\right)}\right)^2 \\
&= \frac{1}{p_0^2\left(1-p_0\right)^2}\mathcal{E}_{p_0}\left(y^2-2yp_0+p_0^2\right) \\
&= \frac{1}{p_0^2\left(1-p_0\right)^2}(p_0-2p_0^2+p_0^2) \\
&= \frac{1}{p_0(1-p_0)}
\end{aligned}$$

which does not depend on $n$, so its limit is the same thing, and, therefore,

$$\mathcal{I}_\infty(p_0)=\frac{1}{p_0(1-p_0)}.$$

So, combining things, we know that the theory says that

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \rightarrow^d N\left[0, \mathcal{J}_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)\mathcal{J}_\infty(\theta_0)^{-1}\right].$$

In the present case,

$$\begin{aligned}\mathcal{J}_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)\mathcal{J}_\infty(\theta_0)^{-1} &= p^0\left(1-p^0\right)\frac{1}{p_0(1-p_0)}p^0\left(1-p^0\right)\\ &= p_0 - p_0^2,\end{aligned}$$

which is the same limiting variance that we computed directly, above, in eq. 15.8.

## 15.5 The information matrix equality

We will show that $\mathcal{J}_\infty(\theta) = -\mathcal{I}_\infty(\theta)$. The example we just looked at exhibits this property. Now, we will see that it's a general property of ML estimators.

Let $f_t(\theta)$ be short for $f(y_t|x_t, \theta)$. This is a density function, which integrates to 1, so:

$$\begin{aligned} 1 &= \int f_t(\theta)dy, \text{ so} \\ 0 &= \int D_\theta f_t(\theta)dy \\ &= \int \left(D_\theta \ln f_t(\theta)\right) f_t(\theta)dy \end{aligned}$$

Now differentiate again, using the product rule, because $\theta$ appears in two terms

that multiply one another:

$$
\begin{aligned}
0 &= \int \left[D_\theta^2 \ln f_t(\theta)\right] f_t(\theta) dy + \int \left[D_\theta \ln f_t(\theta)\right] \left\{D_{\theta'} f_t(\theta)\right\} dy \\
&= \mathcal{E}_\theta \left[D_\theta^2 \ln f_t(\theta)\right] + \int \left[D_\theta \ln f_t(\theta)\right] \left\{\left[D_{\theta'} \ln f_t(\theta)\right] f_t(\theta)\right\} dy \\
&= \mathcal{E}_\theta \left[D_\theta^2 \ln f_t(\theta)\right] + \mathcal{E}_\theta \left[D_\theta \ln f_t(\theta)\right] \left[D_{\theta'} \ln f_t(\theta)\right] \\
&= \mathcal{E}_\theta \left[\mathcal{J}_t(\theta)\right] + \mathcal{E}_\theta \left[g_t(\theta)\right] \left[g_t(\theta)\right]' \qquad (15.9)
\end{aligned}
$$

Now sum over $n$ and multiply by $\frac{1}{n}$

$$\mathcal{E}_\theta \frac{1}{n} \sum_{t=1}^{n} \left[\mathcal{J}_t(\theta)\right] = \mathcal{E}_\theta \left[\mathcal{J}_n(\theta)\right] = -\mathcal{E}_\theta \left[\frac{1}{n} \sum_{t=1}^{n} \left[g_t(\theta)\right] \left[g_t(\theta)\right]'\right] \tag{15.10}$$

- The scores $g_t$ and $g_s$ are uncorrelated for $t \neq s$, as we saw above (43). This allows us to write:

$$\mathcal{E}_\theta \left[\mathcal{J}_n(\theta)\right] = -\mathcal{E}_\theta \left(n \left[g_n(\theta)\right] \left[g_n(\theta)\right]'\right).$$

To see that this is equal to the expression in eq. 15.10, recall that the score vector is the average of the score contributions: $g_n = \frac{1}{n} \Sigma_t \, g_t$ (omitting the $\theta$). So, we have

$$\begin{aligned}
-\mathcal{E}_\theta \left(n \left[g_n(\theta)\right] \left[g_n(\theta)\right]'\right) &= -\mathcal{E}_\theta \frac{\left(\Sigma_t \, g_t\right) \left(\Sigma_t \, g_t'\right)}{n} \\
&= -\mathcal{E}_\theta \frac{\left(g_1 + g_2 + ... + g_n\right) \left(g_1' + g_2' + ... + g_n'\right)}{n} \\
&= -\mathcal{E}_\theta \left[\frac{1}{n} \sum_{t=1}^{n} \left[g_t(\theta)\right] \left[g_t(\theta)\right]'\right],
\end{aligned}$$

which is the original expression in eq. 15.10, because all cross products between different periods expect to zero. Finally take limits, we get

$$\mathcal{J}_\infty(\theta) = -\mathcal{I}_\infty(\theta). \tag{15.11}$$

This holds for all $\theta$, in particular, for $\theta_0$.

Using this,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \overset{a.s.}{\to} N\left[0, \mathcal{J}_\infty(\theta_0)^{-1}\mathcal{I}_\infty(\theta_0)\mathcal{J}_\infty(\theta_0)^{-1}\right]$$

simplifies to

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \overset{a.s.}{\to} N\left[0, \mathcal{I}_\infty(\theta_0)^{-1}\right] \tag{15.12}$$

or

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \overset{a.s.}{\to} N\left[0, -\mathcal{J}_\infty(\theta_0)^{-1}\right] \tag{15.13}$$

To estimate the asymptotic variance, we need estimators of $\mathcal{J}_\infty(\theta_0)$ and $\mathcal{I}_\infty(\theta_0)$.

We can use

$$\begin{aligned}\widehat{\mathcal{I}_\infty(\theta_0)} &= \frac{1}{n}\sum_{t=1}^{n} g_t(\hat{\theta})g_t(\hat{\theta})' \\ \widehat{\mathcal{J}_\infty(\theta_0)} &= \mathcal{J}_n(\hat{\theta}).\end{aligned}$$

as is intuitive if one considers equation 15.10.

Note, one can't use

$$\widehat{I_\infty(\theta_0)} = n\left[g_n(\hat{\theta})\right]\left[g_n(\hat{\theta})\right]'$$

to estimate the information matrix. Why not?

From this we see that there are alternative ways to estimate $V_\infty(\theta_0)$ that are all valid. These include

$$\begin{aligned}
\widehat{V_\infty(\theta_0)} &= -\widehat{\mathcal{J}_\infty(\theta_0)}^{-1} \\
\widehat{V_\infty(\theta_0)} &= \widehat{\mathcal{I}_\infty(\theta_0)}^{-1} \\
\widehat{V_\infty(\theta_0)} &= \widehat{\mathcal{J}_\infty(\theta_0)}^{-1}\widehat{\mathcal{I}_\infty(\theta_0)}\widehat{\mathcal{J}_\infty(\theta_0)}^{-1}
\end{aligned}$$

These are known as the *inverse Hessian, outer product of the gradient* (OPG) and *sandwich* estimators, respectively. The sandwich form is the most robust, since it coincides with the covariance estimator of the *quasi*-ML estimator.

With a little more detail, the methods are:

- The sandwich version:

$$\widehat{V_\infty} = n \left\{ \begin{array}{c} \left\{\Sigma_{t=1}^{n} D_\theta^2 \ln f(y_t|Y_{t-1},\hat{\theta})\right\} \times \\ \left\{\Sigma_{t=1}^{n} \left[D_\theta \ln f(y_t|Y_{t-1},\hat{\theta})\right] \left[D_\theta \ln f(y_t|Y_{t-1},\hat{\theta})\right]'\right\}^{-1} \times \\ \left\{\Sigma_{t=1}^{n} D_\theta^2 \ln f(y_t|Y_{t-1},\hat{\theta})\right\} \end{array} \right\}^{-1}$$

- or the inverse of the negative of the Hessian:

$$\widehat{V_\infty} = \left[-1/n \sum_{t=1}^{n} D_\theta^2 \ln f(y_t|Y_{t-1},\hat{\theta})\right]^{-1},$$

- or the inverse of the outer product of the gradient:

$$\widehat{V_\infty} = \left\{1/n \sum_{t=1}^{n} \left[D_\theta \ln f(y_t|Y_{t-1},\hat{\theta})\right] \left[D_\theta \ln f(y_t|Y_{t-1},\hat{\theta})\right]'\right\}^{-1}.$$

- This simplification is a special result for the MLE estimator - it **doesn't apply** to extremum estimators in general.

- Asymptotically, if the model is correctly specified, all of these forms converge to the same limit. In small samples they will differ. In particular, there is evidence that the outer product of the gradient formula does not perform very well in small samples (*e.g.,* see Davidson and MacKinnon, pg. 477).

- White's *Information matrix test* (White (1982)) is based upon comparing the two ways to estimate the information matrix: outer product of gradient or negative of the Hessian. If they differ by too much, this is evidence of misspecification of the model.

Once we have the estimated asymptotic variance, $\widehat{V_\infty(\theta_0)}$, the <u>approximate</u> small-sample distribution of the estimator is

$$\hat{\theta} \approx N(\theta_0, \frac{\widehat{V_\infty(\theta_0)}}{n})$$

and this can be used to compute confidence intervals, test hypotheses, etc.

**Exercise 44.** Examine the code in EstimatePoisson.jl and mle.jl to figure out how the variance-covariance of the parameters has been estimated. Figure 15.1 show the results using the sandwich and OPG forms. Note the radical changes in the t-statistics. This probably indicates that the Poisson model is not well-specified, following the logic of the information matrix test. When the t-statistics change so much, it means that the estimates of $I_\infty$ and $-J_\infty$ are far from one another.

Figure 15.1: Alternative variance computations for the OBDV Poisson model

```
File   Edit   View   Bookmarks   Settings   Help

julia> include("EstimatePoisson.jl");
*************************************************************
Poisson model, OBDV,  MEPS 1996 full data set
MLE Estimation Results     BFGS convergence: Normal
Average Log-L: -3.67109    Observations: 4564
OPG form covariance estimator

                 estimate      st. err       t-stat      p-value
    constant     -0.79055      0.01837    -43.04086      0.00000
   pub. ins.      0.84802      0.00851     99.63361      0.00000
  priv. ins.      0.29448      0.00858     34.32495      0.00000
         sex      0.48679      0.00552     88.13428      0.00000
         age      0.02402      0.00023    103.59860      0.00000
         edu      0.02917      0.00110     26.53989      0.00000
         inc     -0.00080      0.00010     -8.12865      0.00000

Information Criteria
                   Crit.       Crit/n
       CAIC  33575.68806      7.35664
        BIC  33568.68806      7.35510
        AIC  33523.70638      7.34525
*************************************************************

julia> include("EstimatePoisson.jl");
*************************************************************
Poisson model, OBDV,  MEPS 1996 full data set
MLE Estimation Results     BFGS convergence: Normal
Average Log-L: -3.67109    Observations: 4564
Sandwich form covariance estimator

                 estimate      st. err       t-stat      p-value
    constant     -0.79055      0.14945     -5.28980      0.00000
   pub. ins.      0.84802      0.07645     11.09203      0.00000
  priv. ins.      0.29448      0.07119      4.13640      0.00004
         sex      0.48679      0.05534      8.79580      0.00000
         age      0.02402      0.00209     11.46973      0.00000
         edu      0.02917      0.00953      3.06037      0.00222
         inc     -0.00080      0.00081     -0.97809      0.32808

Information Criteria
                   Crit.       Crit/n
       CAIC  33575.68806      7.35664
```

## 15.6 The Cramér-Rao lower bound

**Definition 45.** Consistent and asymptotically normal (CAN). An estimator $\hat{\theta}$ of a parameter $\theta_0$ is $\sqrt{n}$-consistent and asymptotically normally distributed if $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, V_\infty\right)$ where $V_\infty$ is a finite positive definite matrix.

There do exist, in special cases, estimators that are consistent such that $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{p} 0$. These are known as *superconsistent* estimators, since in ordinary circumstances with stationary data, $\sqrt{n}$ is the highest factor that we can multiply by and still get convergence to a stable limiting distribution.

**Definition 46.** Asymptotically unbiased. An estimator $\hat{\theta}$ of a parameter $\theta_0$ is asymptotically unbiased if

$\lim_{n\to\infty} \mathcal{E}_\theta(\hat{\theta}) = \theta.$

*Estimators that are CAN are asymptotically unbiased*, though not all consistent estimators are asymptotically unbiased. Such cases are unusual, though.

**Theorem 47.** [Cramer-Rao Lower Bound] *The limiting variance of a CAN estimator of $\theta_0$, say $\tilde{\theta}$, minus the inverse of the information matrix is a positive semidefinite matrix.*

Proof: Since the estimator is CAN, it is asymptotically unbiased, so

$$\lim_{n\to\infty} \mathcal{E}_\theta(\tilde{\theta} - \theta) = 0$$

Differentiate wrt $\theta'$, and make liberal use of dominated convergence:

$$\begin{aligned} D_{\theta'} \lim_{n\to\infty} \mathcal{E}_\theta(\tilde{\theta} - \theta) &= \lim_{n\to\infty} \int D_{\theta'} \left[ f(Y,\theta) \left( \tilde{\theta} - \theta \right) \right] dy \\ &= 0. \end{aligned}$$

(The RHS is zero, because we're differentiating something that is zero to start with).

Now, take the RHS, and differentiate. Noting that $D_{\theta'} f(Y,\theta) = f(\theta) D_{\theta'} \ln f(\theta)$ (a trick we have seen a few times already), and applying the product rule to differentiate the two parts that depend on $\theta$, we can write

$$\lim_{n\to\infty} \int \left( \tilde{\theta} - \theta \right) f(\theta) D_{\theta'} \ln f(\theta) dy + \lim_{n\to\infty} \int f(Y,\theta) D_{\theta'} \left( \tilde{\theta} - \theta \right) dy = 0.$$

Now note that $D_{\theta'}\left(\tilde{\theta}-\theta\right) = -I_K$, and $\int f(Y,\theta)(-I_K)dy = -I_K$. With this we have

$$\lim_{n\to\infty} \int \left(\tilde{\theta}-\theta\right) f(\theta) D_{\theta'} \ln f(\theta) dy = I_K.$$

Playing with powers of $n$ on the LHS, we get

$$\lim_{n\to\infty} \int \sqrt{n}\left(\tilde{\theta}-\theta\right) \sqrt{n}\, \underbrace{\frac{1}{n}\left[D_{\theta'} \ln f(\theta)\right]} f(\theta) dy \;=\; I_K$$

Note that the bracketed part is just the transpose of the score vector, $g(\theta)$, so we can write

$$\lim_{n\to\infty} \mathcal{E}_\theta \left[\sqrt{n}\left(\tilde{\theta}-\theta\right) \sqrt{n} g(\theta)'\right] = I_K$$

This means that the limiting covariance of the score function with $\sqrt{n}\left(\tilde{\theta}-\theta\right)$, for $\tilde{\theta}$ **any** CAN estimator, is an identity matrix. Using this, suppose the variance

of $\sqrt{n}\left(\tilde{\theta}-\theta\right)$ tends to $V_\infty(\tilde{\theta})$. Therefore,

$$V_\infty\begin{bmatrix}\sqrt{n}\left(\tilde{\theta}-\theta\right)\\ \sqrt{n}g(\theta)\end{bmatrix}=\begin{bmatrix}V_\infty(\tilde{\theta}) & I_K\\ I_K & \mathcal{I}_\infty(\theta)\end{bmatrix}. \tag{15.14}$$

Since this is a covariance matrix, it is positive semi-definite. Therefore, for any $K$ -vector $\alpha$,

$$\begin{bmatrix}\alpha' & -\alpha'\mathcal{I}_\infty^{-1}(\theta)\end{bmatrix}\begin{bmatrix}V_\infty(\tilde{\theta}) & I_K\\ I_K & \mathcal{I}_\infty(\theta)\end{bmatrix}\begin{bmatrix}\alpha\\ -\mathcal{I}_\infty(\theta)^{-1}\alpha\end{bmatrix}\geq 0.$$

This simplifies to

$$\alpha'\left[V_\infty(\tilde{\theta})-\mathcal{I}_\infty^{-1}(\theta)\right]\alpha\geq 0.$$

Since $\alpha$ is arbitrary, $V_\infty(\tilde{\theta})-\mathcal{I}_\infty^{-1}(\theta)$ must be positive semidefinite. This concludes the proof.

*Interpretation:*

- any linear combination of $\tilde{\theta}$ will have an asymptotic variance greater than or equal to the same linear combination of the ML estimator.
- *e.g.,* the individual variances of each $\tilde{\theta}_j$ will be no smaller than the corresponding element of the ML estimator, $j = 1, 2, ..., k$
- $\mathcal{I}_\infty^{-1}(\theta)$ is a *lower bound* for the asymptotic variance of a CAN estimator.

**Definition 48.** (*Asymptotic efficiency*) Given two CAN estimators of a parameter $\theta_0$, say $\tilde{\theta}$ and $\hat{\theta}$, $\hat{\theta}$ is asymptotically efficient with respect to $\tilde{\theta}$ if $V_\infty(\tilde{\theta}) - V_\infty(\hat{\theta})$ is a positive semidefinite matrix.

- *the MLE is asymptotically efficient with respect to any other CAN estimator.*
- this is the reason that the ML estimator is so important: it provides a benchmark for efficiency.
  - The strong assumptions on which it depends may be questionable, though.
  - If we can find another estimator that obtains an asymptotic variance similar to that of ML, but is consistent under weaker assumptions, we might choose to use it instead. But it's useful to know that this entails a potential loss of efficiency.

## 15.7 Likelihood ratio-type tests

Suppose we would like to test a set of $q$ possibly nonlinear restrictions $r(\theta) = 0$, where the $q \times k$ matrix $D_{\theta'} r(\theta)$ has rank $q$. The Wald test can be calculated using the unrestricted model. The score test can be calculated using only the restricted model. The likelihood ratio test, on the other hand, uses both the restricted and the unrestricted estimators. The test statistic is

$$LR = 2\left(\ln L(\hat{\theta}) - \ln L(\tilde{\theta})\right) = 2n\left[s_n(\hat{\theta}) - s_n(\tilde{\theta})\right]$$

where $\hat{\theta}$ is the unrestricted estimate and $\tilde{\theta}$ is the restricted estimate. We will show that, under the null hypothesis that $r(\theta) = 0$,

$$LR \xrightarrow{d} \chi^2(q).$$

*The following sketch of a proof should be read carefully by students with a special interest in econometric theory, but it can be scanned lightly by most students.*

To show that it is asymptotically $\chi^2$, take a second order Taylor's series expansion of $\ln L(\tilde{\theta})$ about $\hat{\theta}$ :

$$\ln L(\tilde{\theta}) \simeq \ln L(\hat{\theta}) + \frac{n}{2}\left(\tilde{\theta} - \hat{\theta}\right)' \mathcal{J}(\hat{\theta}) \left(\tilde{\theta} - \hat{\theta}\right)$$

(note, the first order term drops out since $D_\theta \ln L(\hat{\theta}) \equiv 0$ by the first order necessary conditions, and we need to multiply the second-order term by $n$ since $\mathcal{J}(\theta)$ is defined in terms of $\frac{1}{n}\ln L(\theta)$) so

$$LR \simeq -n\left(\tilde{\theta} - \hat{\theta}\right)' \mathcal{J}(\hat{\theta}) \left(\tilde{\theta} - \hat{\theta}\right)$$

As $n \to \infty, \mathcal{J}(\hat{\theta}) \to \mathcal{J}_\infty(\theta_0) = -\mathcal{I}(\theta_0)$, by the information matrix equality. So

$$LR \stackrel{a}{=} n\left(\tilde{\theta} - \hat{\theta}\right)' \mathcal{I}_\infty(\theta_0) \left(\tilde{\theta} - \hat{\theta}\right) \tag{15.15}$$

We also have that, from the theory on the asymptotic normality of the MLE and the information matrix equality

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \stackrel{a}{=} \sqrt{n}\mathcal{I}_\infty(\theta_0)^{-1} I_n g(\theta_0).$$

An analogous result for the restricted estimator is (this is unproven here, to prove this set up the Lagrangean for MLE subject to $r(\theta)=0$, and manipulate the first order conditions. Also note, $R$ is notation for $R = D_\theta r'(\theta)$):

$$\sqrt{n}\left(\tilde{\theta}-\theta_0\right) \stackrel{a}{=} \sqrt{n}\mathcal{I}_\infty(\theta_0)^{-1}\Big(I_n - R'\left(R\mathcal{I}_\infty(\theta_0)^{-1}R'\right)^{-1} R\mathcal{I}_\infty(\theta_0)^{-1}\Big)g(\theta_0).$$

Subtracting the penultimate equation from the last one, we get (the magenta colored term below is the difference between the two magenta terms in the last two lines, with the minus sign moved out to the front)

$$\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right) \stackrel{a}{=} -\sqrt{n}\mathcal{I}_\infty(\theta_0)^{-1} R'\left(R\mathcal{I}_\infty(\theta_0)^{-1}R'\right)^{-1} R\mathcal{I}_\infty(\theta_0)^{-1} g(\theta_0)$$

so, substituting into [15.15] , we get some nice cancellations, which lead to (note: *renewing use of colors here*)

$$LR \stackrel{a}{=} \left[n^{1/2} g(\theta_0)' \mathcal{I}_\infty(\theta_0)^{-1} R'\right] \left[R\mathcal{I}_\infty(\theta_0)^{-1} R'\right]^{-1} \left[R\mathcal{I}_\infty(\theta_0)^{-1} n^{1/2} g(\theta_0)\right]$$

But since

$$n^{1/2} g(\theta_0) \xrightarrow{d} N\left(0, \mathcal{I}_\infty(\theta_0)\right)$$

the linear function

$$R\mathcal{I}_\infty(\theta_0)^{-1} n^{1/2} g(\theta_0) \xrightarrow{d} N(0, R\mathcal{I}_\infty(\theta_0)^{-1} R').$$

We can see that LR is a quadratic form of this random variable, with the inverse of its variance in the middle, so, by the Continuous Mapping Theorem (also known as the Mann-Wald Theorem) (see Gallant (1997) Theorem 4.7 for a statement):

$$LR \xrightarrow{d} \chi^2(q).$$

**Example 49.** *Likelihood ratio test.* Continuing with the same code as was used in Example 41, LikelihoodRatioTest.jl computes the LR statistic for a simple linear $H_0$. This uses the Julia function fmincon to perform the restricted estimation.

- run the test a number of times to explore size
- change the value of $r$ in the null hypothesis $R\beta = r$ to make it false, and run the test a number of times, to check power.
- explore how power depends on the sample size

# Summary of MLE

- Consistent
- Asymptotically normal (CAN)
- Asymptotically efficient
- Asymptotically unbiased
- LR test is available for testing hypothesis
- The presentation is for general MLE: we haven't specified the distribution or the linearity/nonlinearity of the estimator

## 15.8 Examples

This section is quite long winded, with several examples of ML estimation. Just focusing on a couple of examples should be sufficient for most students.

### ML of Nerlove model, assuming normality

As we saw in Section 5.3, the ML and OLS estimators of $\beta$ in the linear model $y = X\beta + \epsilon$ coincide when $\epsilon$ is assumed to be i.i.d. normally distributed. The Julia script NerloveMLE.jl verifies this result, for the basic Nerlove model (eqn. 4.10). The output of the script follows:

```
*****************************************************
estimate Nerlove model by MLE
MLE Estimation Results
BFGS convergence: Normal convergence
Average Log-L: -0.46581
```

Observations: 145

| | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -3.527 | 1.695 | -2.081 | 0.039 |
| output | 0.720 | 0.032 | 22.413 | 0.000 |
| labor | 0.436 | 0.242 | 1.802 | 0.074 |
| fuel | 0.427 | 0.074 | 5.731 | 0.000 |
| capital | -0.220 | 0.319 | -0.689 | 0.492 |
| sig | 0.386 | 0.042 | 9.257 | 0.000 |

Information Criteria

| | Crit. | Crit/n |
|---|---|---|
| CAIC | 170.944 | 1.179 |
| BIC | 164.944 | 1.138 |
| AIC | 147.084 | 1.014 |

```
*****************************************************************
```

Compare the output to that of Nerlove.jl , which does OLS. The script also provides a basic example of how to use the MLE estimation routine `mleresults.jl`, which is in the Econometrics.jl package.

# Example: Binary response models: theory

This section extends the Bernoulli trial model to binary response models with conditioning variables, as such models arise in a variety of contexts.

Assume that

$$\begin{aligned} y^* &= x'\theta - \varepsilon \\ y &= 1(y^* > 0) \end{aligned}$$

Here, $y^*$ is an unobserved (latent) continuous variable, and $y$ is a binary variable that indicates whether $y^*$is negative or positive.

- If $\varepsilon \sim N(0,1)$ Then the *probit* model results, where $Pr(y=1|x) = Pr(\varepsilon < x'\theta) = \Phi(x'\theta)$, where

$$\Phi(\cdot) = \int_{-\infty}^{x\beta} (2\pi)^{-1/2} \exp(-\frac{\varepsilon^2}{2}) d\varepsilon$$

  is the standard normal distribution function.

- The *logit* model results if the errors $\epsilon$ are not normal, but rather have a logistic distribution. This distribution is similar to the standard normal, but has fatter tails. The probability, in this case, has the following parameterization

$$Pr(y=1|x) = \Lambda(x'\theta) = (1 + \exp(-x'\theta))^{-1}.$$

- In general, a binary response model will require that the choice probability be parameterized in some form which could be logit, probit, or something else. For a vector of explanatory variables $x$, the response probability will

be parameterized in some manner

$$Pr(y = 1|x) = p(x, \theta)$$

- Again, if $p(x, \theta) = \Lambda(x'\theta)$, we have a logit model. If $p(x, \theta) = \Phi(x'\theta)$, where $\Phi(\cdot)$ is the standard normal distribution function, then we have a probit model.

*The following, to the start of the next subsection, is another verification of the information matrix equality, skip this in lectures.*

Regardless of the parameterization, we are dealing with a Bernoulli density,

$$f_{Y_i}(y_i|x_i) = p(x_i, \theta)^{y_i}(1 - p(x, \theta))^{1-y_i}$$

so as long as the observations are independent, the maximum likelihood (ML) estimator, $\hat{\theta}$, is the maximizer of

$$\begin{aligned} s_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i \ln p(x_i, \theta) + (1 - y_i)\ln\left[1 - p(x_i, \theta)\right]\right) \\ &\equiv \frac{1}{n}\sum_{i=1}^{n} s(y_i, x_i, \theta). \end{aligned} \tag{15.16}$$

Following the above theoretical results, $\hat{\theta}$ tends in probability to the $\theta_0$ that maximizes the uniform almost sure limit of $s_n(\theta)$. Noting that $\mathcal{E}y_i = p(x_i, \theta_0)$, and following a SLLN for i.i.d. processes, $s_n(\theta)$ converges almost surely to the expectation of a representative term $s(y, x, \theta)$. First one can take the expectation

conditional on $x$ to get

$$\mathcal{E}_{y|x}\left\{y\ln p(x,\theta)+(1-y)\ln\left[1-p(x,\theta)\right]\right\}=p(x,\theta_0)\ln p(x,\theta)+[1-p(x,\theta_0)]\ln\left[1-p(x,\theta)\right]$$

Next taking expectation over $x$ we get the limiting objective function

$$s_\infty(\theta)=\int_{\mathcal{X}}\left\{p(x,\theta_0)\ln p(x,\theta)+\left[1-p(x,\theta_0)\right]\ln\left[1-p(x,\theta)\right]\right\}\mu(x)dx, \quad (15.17)$$

where $\mu(x)$ is the (joint - the integral is understood to be multiple, and $\mathcal{X}$ is the support of $x$) density function of the explanatory variables $x$. This is clearly continuous in $\theta$, as long as $p(x,\theta)$ is continuous, and if the parameter space is compact we therefore have uniform almost sure convergence. Note that $p(x,\theta)$ is continuous for the logit and probit models, for example. The maximizing element of $s_\infty(\theta)$, $\theta^*$, solves the first order conditions

$$\int_{\mathcal{X}}\left\{\frac{p(x,\theta_0)}{p(x,\theta^*)}\frac{\partial}{\partial\theta}p(x,\theta^*)-\frac{1-p(x,\theta_0)}{1-p(x,\theta^*)}\frac{\partial}{\partial\theta}p(x,\theta^*)\right\}\mu(x)dx=0$$

This is clearly solved by $\theta^* = \theta_0$. Provided the solution is unique, $\hat{\theta}$ is consistent. Question: what's needed to ensure that the solution is unique?

The asymptotic normality theorem tells us that

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right].$$

In the case of i.i.d. observations $\mathcal{I}_\infty(\theta_0) = \lim_{n\to\infty} Var\sqrt{n}D_\theta s_n(\theta_0)$ is simply the expectation of a typical element of the outer product of the gradient.

- There's no need to subtract the mean, since it's zero, following the f.o.c. in the consistency proof above and the fact that observations are i.i.d.

- The terms in $n$ also drop out by the same argument:

$$\begin{aligned}
\lim_{n\to\infty} Var\sqrt{n}D_\theta s_n(\theta_0) &= \lim_{n\to\infty} Var\sqrt{n}D_\theta \frac{1}{n}\sum_t s(\theta_0) \\
&= \lim_{n\to\infty} Var\frac{1}{\sqrt{n}}D_\theta \sum_t s(\theta_0) \\
&= \lim_{n\to\infty} \frac{1}{n}Var\sum_t D_\theta s(\theta_0) \\
&= \lim_{n\to\infty} Var D_\theta s(\theta_0) \\
&= Var D_\theta s(\theta_0)
\end{aligned}$$

So we get

$$\mathcal{I}_\infty(\theta_0) = \mathcal{E}\left\{\frac{\partial}{\partial\theta}s(y,x,\theta_0)\frac{\partial}{\partial\theta'}s(y,x,\theta_0)\right\}.$$

Likewise,

$$\mathcal{J}_\infty(\theta_0) = \mathcal{E}\frac{\partial^2}{\partial\theta\partial\theta'}s(y,x,\theta_0).$$

Expectations are jointly over $y$ and $x$, or equivalently, first over $y$ conditional on

$x$, then over $x$. From above, a typical element of the objective function is

$$s(y, x, \theta_0) = y \ln p(x, \theta_0) + (1 - y) \ln \left[1 - p(x, \theta_0)\right].$$

Now suppose that we are dealing with a correctly specified logit model:

$$p(x, \theta) = \left(1 + \exp(-\mathbf{x}'\theta)\right)^{-1}.$$

We can simplify the above results in this case. We have that

$$\begin{aligned}
\frac{\partial}{\partial \theta} p(x, \theta) &= \left(1 + \exp(-\mathbf{x}'\theta)\right)^{-2} \exp(-\mathbf{x}'\theta)\mathbf{x} \\
&= \left(1 + \exp(-\mathbf{x}'\theta)\right)^{-1} \frac{\exp(-\mathbf{x}'\theta)}{1 + \exp(-\mathbf{x}'\theta)} \mathbf{x} \\
&= p(x, \theta)\left(1 - p(x, \theta)\right)\mathbf{x} \\
&= \left(p(x, \theta) - p(x, \theta)^2\right)\mathbf{x}.
\end{aligned}$$

So

$$\begin{aligned}\frac{\partial}{\partial\theta}s(y,x,\theta_0) &= [y-p(x,\theta_0)]\,\mathbf{x} \qquad (15.18)\\ \frac{\partial^2}{\partial\theta\partial\theta'}s(\theta_0) &= -\left[p(x,\theta_0)-p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'.\end{aligned}$$

Taking expectations over $y$ then $\mathbf{x}$ gives

$$\begin{aligned}\mathcal{I}_\infty(\theta_0) &= \int E_Y\left[y^2-2p(x,\theta_0)p(x,\theta_0)+p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx \qquad (15.19)\\ &= \int\left[p(x,\theta_0)-p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx. \qquad (15.20)\end{aligned}$$

where we use the fact that $E_Y(y)=E_Y(y^2)=p(\mathbf{x},\theta_0)$. Likewise,

$$\mathcal{J}_\infty(\theta_0) = -\int\left[p(x,\theta_0)-p(x,\theta_0)^2\right]\mathbf{x}\mathbf{x}'\mu(x)dx. \qquad (15.21)$$

Note that we arrive at the expected result: the information matrix equality holds (that is, $\mathcal{J}_\infty(\theta_0)=-\mathcal{I}_\infty(\theta_0)$).

On a final note, the logit and standard normal CDF's are very similar - the logit

distribution is a bit more fat-tailed. While coefficients will vary slightly between the two models, functions of interest such as estimated probabilities $p(x, \hat{\theta})$ will be virtually identical for the two models.

# Estimation of the logit model

In this section we will consider maximum likelihood estimation of the logit model for binary 0/1 dependent variables.

A binary response is a variable that takes on only two values, customarily 0 and 1, which can be thought of as codes for whether or not a condisiton is satisfied. For example, 0=drive to work, 1=take the bus. Often the observed binary variable, say $y$, is related to an unobserved (latent) continuous varable, say $y^*$. We would like to know the effect of covariates, $x$, on $y$. The model can be represented as

$$\begin{aligned} y^* &= x'\theta - \varepsilon \\ y &= 1(y^* > 0) \\ Pr(y = 1) &= F_\varepsilon[x'\theta] \\ &\equiv p(x, \theta) \end{aligned}$$

The log-likelihood function is

$$s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i \ln p(x_i, \theta) + (1 - y_i) \ln \left[ 1 - p(x_i, \theta) \right] \right)$$

As above, for the logit model, the probability has the specific form

$$p(x, \theta) = \frac{1}{1 + \exp(-x'\theta)}$$

You should download and examine LogitDGP.jl , which generates data according to the logit model, logit.jl , which calculates the loglikelihood, and EstimateLogit.jl , which sets things up and calls the estimation routine.

Here are some estimation results with $n = 30$, and the true $\theta = (0, 1)'$.

```
julia> include("EstimateLogit.jl");
*****************************************************
estimate logit model
MLE Estimation Results
BFGS convergence: Normal convergence
Average Log-L: -0.62038
Observations: 30
```

```
        estimate st. err  t-stat p-value
       1  -0.557   0.443  -1.258   0.219
       2   0.944   0.503   1.876   0.071

Information Criteria
           Crit.Avg. Crit
    CAIC  46.025   1.534
     BIC  44.025   1.468
     AIC  41.223   1.374
*****************************************************
```

The estimation program is calling **mleresults**.jl, which in turn calls other routines. See mleresults.jl and mle.jl for the details.

## Example: the MEPS Data

We first saw the MEPS data in Section 12.4, where a Poisson model was estimated by maximum likelihood. To check the plausibility of the Poisson model for the MEPS data, we can compare the sample unconditional variance with the estimated unconditional variance that is implied by the Poisson model: $\widehat{V(y)} = \frac{\sum_{t=1}^{n} \hat{\lambda}_t}{n}$. (remember that, for the Poisson model, the parameter $\lambda$ is both the mean and the variance). Using the program PoissonVariance.jl, for OBDV and ERV, we get the results in Table 15.1. We see that even after conditioning, the overdispersion is

Table 15.1: Marginal Variances, Sample and Estimated (Poisson)

| | OBDV | ERV |
|---|---|---|
| Sample | 38.09 | 0.151 |
| Estimated | 3.28 | 0.086 |

not captured in either case. There is huge problem with OBDV, and a significant problem with ERV. In both cases the Poisson model does not appear to be

plausible. You can check this for the other use measures if you like.

## Infinite mixture models: the negative binomial model

Reference: Cameron and Trivedi (1998) *Regression analysis of count data,* chapter 4.

The two measures seem to exhibit extra-Poisson variation. To capture unobserved heterogeneity, a possibility is the *random parameters* approach. Consider the possibility that the parameter in a Poisson model were random:

$$f_Y(y|\lambda, v) \;=\; \frac{\exp(-\theta)\theta^y}{y!}$$

where $\theta = \lambda v$.

- Now $\nu$ is a multiplicative random term. The problem is that we don't observe $\nu$, so we will need to marginalize it to get a usable density.

- Suppose that $f_v(z;\psi)$ is the density of the latent term. We will need to

assume that this is known - it is part of the specification of the econometric model.

- Then, the marginal density of $Y$ is

$$f_Y(y|\lambda,\psi) = \int_{-\infty}^{\infty} \frac{\exp[-\lambda z]\,[\lambda z]^y}{y!} f_v(z;\psi)dz$$

This density *can* be used directly, perhaps using numerical integration to evaluate the likelihood function. Simulation-based approaches are another possibility.

- In some cases, though, the integral will have an analytic solution. For example, if $\nu$ follows a certain one parameter gamma density, then

$$f_Y(y|\lambda, \phi) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^r (1-p)^y \tag{15.22}$$

  where $\phi = (\lambda, \psi)$ and $p = \frac{\psi}{\psi+\lambda}$. $\psi$ appears since it is the parameter of the gamma density. The above density is the *negative binomial* density. See Gourieroux et al. (1984) for an influential paper on the topic.

- We usually parameterize $\lambda = \exp(\mathbf{x}'\beta)$, as before, for the Poisson model.

- The variance depends upon how $\psi$ is parameterized.

  - If $\psi = \lambda/\alpha$, where $\alpha > 0$, then $V(y|\mathbf{x}) = \lambda + \alpha\lambda$. Note that $\lambda$ is a function of $\mathbf{x}$, so that the variance is, too. This is referred to as the NB-I model.

  - If $\psi = 1/\alpha$, where $\alpha > 0$, then $V(y|\mathbf{x}) = \lambda + \alpha\lambda^2$. This is referred to

as the NB-II model.

- In both cases, the conditional mean is $E(y|\boldsymbol{x}) = \lambda = \exp(\mathbf{x}'\beta)$

So both forms of the NB model allow for overdispersion, with the NB-II model allowing for a more radical form. There are other parameterizations, but these two are the most widely-used.

Testing reduction of a NB model to a Poisson model cannot be done by testing $\alpha = 0$ using standard Wald or LR procedures. The critical values need to be adjusted to account for the fact that $\alpha = 0$ is on the boundary of the parameter space.

- Without getting into details, suppose that the data were in fact Poisson, so there is equidispersion and the true $\alpha = 0$.
- Then about half the time the sample data will be underdispersed, and about half the time overdispersed.
- When the data is underdispersed, the MLE of $\alpha$ will be $\hat{\alpha} = 0$.
- Thus, under the null, there will be a probability spike in the asymptotic distribution of $\sqrt{n}(\hat{\alpha} - \alpha) = \sqrt{n}\hat{\alpha}$ at 0, so the asymptotic distribution will not be normal, and standard testing methods will not be valid.

This program will do estimation using the NB model . Results for NB-I and NB-II models are

Negative Binomial-1 model, OBDV, MEPS 1996 full data set
MLE Estimation Results Convergence: true
Average log-$\mathcal{L}$: -2.18573 Observations: 4564
Sandwich form covariance estimator

| parameter | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -0.52279 | 0.10446 | -5.00493 | 0.00000 |
| pub. ins. | 0.76460 | 0.05385 | 14.19767 | 0.00000 |
| priv. ins. | 0.45061 | 0.04900 | 9.19606 | 0.00000 |
| sex | 0.45794 | 0.03389 | 13.51206 | 0.00000 |
| age | 0.01614 | 0.00136 | 11.86857 | 0.00000 |
| edu | 0.02656 | 0.00667 | 3.97930 | 0.00007 |
| inc | 0.00000 | 0.00054 | 0.00017 | 0.99986 |
| log(α) | 1.71461 | 0.05333 | 32.15222 | 0.00000 |

Information Criteria

| | Crit. | Crit/n |
|---|---|---|
| CAIC | 20026.75134 | 4.38798 |
| BIC | 20018.75134 | 4.38623 |
| AIC | 19967.34370 | 4.37497 |

Negative Binomial-2 model, OBDV, MEPS 1996 full data set
MLE Estimation Results  Convergence: true
Average log-$\mathcal{L}$: -2.18496  Observations: 4564
Sandwich form covariance estimator

| parameter | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -1.06808 | 0.16129 | -6.62229 | 0.00000 |
| pub. ins. | 1.10107 | 0.09483 | 11.61075 | 0.00000 |
| priv. ins. | 0.47557 | 0.08089 | 5.87952 | 0.00000 |
| sex | 0.56348 | 0.05047 | 11.16561 | 0.00000 |
| age | 0.02481 | 0.00203 | 12.23995 | 0.00000 |
| edu | 0.02912 | 0.00937 | 3.10682 | 0.00190 |
| inc | -0.00014 | 0.00079 | -0.17595 | 0.86034 |
| log(α) | 0.47802 | 0.03437 | 13.90994 | 0.00000 |

Information Criteria

| | Crit. | Crit/n |
|---|---|---|
| CAIC | 20019.74385 | 4.38645 |
| BIC | 20011.74385 | 4.38469 |
| AIC | 19960.33622 | 4.37343 |

- For the OBDV usage measure, the NB-II model does a slightly better job than the NB-I model, in terms of the average log-likelihood and the information criteria (more on this last in a moment).

- Note that both versions of the NB model fit much better than does the Poisson model (see 15.1).

- The estimated $\alpha$ is highly significant.

To check the plausibility of the NB-II model, we can compare the sample unconditional variance with the estimated unconditional variance according to the NB-II model: $\widehat{V(y)} = \frac{\Sigma_{t=1}^{n} \hat{\lambda}_t + \hat{\alpha}\left(\hat{\lambda}_t\right)^2}{n}$. For OBDV and ERV (estimation results not reported), we get For OBDV, the overdispersion problem is significantly better than in the Poisson case, but there is still some that is not captured. For ERV, the negative binomial model seems to capture the overdispersion adequately.

Table 15.2: Marginal Variances, Sample and Estimated (NB-II)

| | OBDV | ERV |
|---|---|---|
| Sample | 38.09 | 0.151 |
| Estimated | 30.58 | 0.182 |

## Finite mixture models: the mixed negative binomial model

Finite mixture models are quite widely used in statistical modeling. The finite mixture approach to fitting health care demand was introduced by Deb and Trivedi (1997). The mixture approach has the intuitive appeal of allowing for subgroups of the population with different health status. If individuals are classified as healthy or unhealthy then two subgroups are defined. A finer classification scheme would lead to more subgroups. Many studies have incorporated objective and/or subjective indicators of health status in an effort to capture this heterogeneity. The available objective measures, such as limitations on activity, are not necessarily very informative about a person's overall health status. Subjective, self-reported

measures may suffer from the same problem, and may also not be exogenous

Finite mixture models are conceptually simple. The density is

$$f_Y(y, \phi_1, ..., \phi_p, \pi_1, ..., \pi_{p-1}) = \sum_{i=1}^{p-1} \pi_i f_Y^{(i)}(y, \phi_i) + \pi_p f_Y^p(y, \phi_p),$$

where $\pi_i > 0, i = 1, 2, ..., p$, $\pi_p = 1 - \Sigma_{i=1}^{p-1} \pi_i$, and $\Sigma_{i=1}^{p} \pi_i = 1$. Identification requires that the $\pi_i$ are ordered in some way, for example, $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p$ and $\phi_i \neq \phi_j, i \neq j$. This is simple to accomplish post-estimation by rearrangement and possible elimination of redundant component densities.

- The properties of the mixture density follow in a straightforward way from those of the components. In particular, the moment generating function is the same mixture of the moment generating functions of the component densities, so, for example, $E(Y|x) = \Sigma_{i=1}^{p} \pi_i \mu_i(x)$, where $\mu_i(x)$ is the mean of the $i^{th}$ component density.

- Mixture densities may suffer from overparameterization, since the total num-

ber of parameters grows rapidly with the number of component densities. It is possible to constrain parameters across the mixtures.

- Testing for the number of component densities is a tricky issue. For example, testing for $p = 1$ (a single component, which is to say, no mixture) versus $p = 2$ (a mixture of two components) involves the restriction $\pi_1 = 1$, which is on the boundary of the parameter space. Not that when $\pi_1 = 1$, the parameters of the second component can take on any value without affecting the density. Usual methods such as the likelihood ratio test are not applicable when parameters are on the boundary under the null hypothesis. Information criteria means of choosing the model (see below) are valid.

The following results are for a mixture of 2 NB-II models, for the OBDV data, which you can replicate using this program .

OBDV

```
*****************************************************
Mixed Negative Binomial model, MEPS 1996 full data set

MLE Estimation Results
BFGS convergence: Normal convergence

Average Log-L: -2.164783
Observations: 4564
```

| | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | 0.127 | 0.512 | 0.247 | 0.805 |
| pub. ins. | 0.861 | 0.174 | 4.962 | 0.000 |
| priv. ins. | 0.146 | 0.193 | 0.755 | 0.450 |
| sex | 0.346 | 0.115 | 3.017 | 0.003 |
| age | 0.024 | 0.004 | 6.117 | 0.000 |
| edu | 0.025 | 0.016 | 1.590 | 0.112 |
| inc | -0.000 | 0.000 | -0.214 | 0.831 |

```
alpha            1.351       0.168       8.061       0.000
constant         0.525       0.196       2.678       0.007
pub. ins.        0.422       0.048       8.752       0.000
priv. ins.       0.377       0.087       4.349       0.000
sex              0.400       0.059       6.773       0.000
age              0.296       0.036       8.178       0.000
edu              0.111       0.042       2.634       0.008
inc              0.014       0.051       0.274       0.784
alpha            1.034       0.187       5.518       0.000
Mix              0.257       0.162       1.582       0.114

Information Criteria
CAIC : 19920.3807      Avg. CAIC:   4.3647
 BIC : 19903.3807       Avg. BIC:   4.3610
 AIC : 19794.1395       Avg. AIC:   4.3370
*****************************************************
```

It is worth noting that the mixture parameter is not significantly different from

zero, but also note that the coefficients of public insurance and age, for example, differ quite a bit between the two latent classes.

## Information criteria

As seen above, a Poisson model can't be tested (using standard methods) as a restriction of a negative binomial model. But it seems, based upon the values of the likelihood functions and the fact that the NB model fits the variance much better, that the NB model is more appropriate. How can we determine which of a set of competing models is the best?

The *information criteria* approach is one possibility. Information criteria are functions of the log-likelihood, with a penalty for the number of parameters used. The idea is to try to choose a model that fits well, but doesn't use an excessive number of parameters to obtain the good fit. Three popular information criteria are the Akaike (AIC), Bayes (BIC) and consistent Akaike (CAIC). The formulae

are

$$
\begin{aligned}
CAIC &= -2\ln L(\hat{\theta}) + k(\ln n + 1) \\
BIC &= -2\ln L(\hat{\theta}) + k\ln n \\
AIC &= -2\ln L(\hat{\theta}) + 2k
\end{aligned}
$$

- For a given criterion, the model that has the *lowest* value of the criterion is favored
- It can be shown that the CAIC and BIC will select the correctly specified model from a group of models, asymptotically.
- This doesn't mean, of course, that the correct model is necessarily in the group.
- The AIC is not consistent, and will asymptotically favor an over-parameterized model over the correctly specified model.

Here are information criteria values, normalized by $n$, for the models we've seen, for OBDV. Pretty clearly, the NB models are better than the Poisson. The

Table 15.3: Average Information Criteria, OBDV

| Model | AIC | BIC | CAIC |
|---|---|---|---|
| Poisson | 7.345 | 7.355 | 7.357 |
| NB-I | 4.375 | 4.386 | 4.388 |
| NB-II | 4.373 | 4.385 | 4.386 |
| MNB-II | 4.337 | 4.361 | 4.365 |

one additional parameter gives a very significant improvement in the likelihood function value. Between the NB-I and NB-II models, the NB-II is very slightly favored. But one should remember that information criteria values are statistics, with variances. With another sample, it may well be that the NB-I model would be favored, since the differences are so small. The MNB-II model is favored over the others, by all 3 information criteria, but it is less than 1% lower than the scores for the other NB models. Perhaps the additional complexity is not worth

the small improvement in fit?

To summarize:

- the Poisson model is clearly not appropriate, as it can't deal with the overdispersion that the data exhibits.
- the NB variants all perform in the same ballpark
- the issue of possible endogeneity of private insurance has been ignored. We will return to that once we have some more tools.

## 15.9 ML estimation of the DSGE model

Note: this section will not be converted to the Julia language at the present time, as there is presently no Julia language full equivalent of Dynare.

Chapter 14 introduced a simple dynamic stochastic general equilibrium model. The file CKml.mod allows you to explore maximum likelihood estimation of the model using Kalman or particle filtering, using the http://www.dynare.org/ package. To run it, start Octave/Matlab, and then enter `dynare CKml.mod`. Some output that can be obtained is:

**using c and n:**

```
RESULTS FROM MAXIMUM LIKELIHOOD ESTIMATION
parameters
       Estimate    s.d. t-stat

betta    0.9900  0.0021 469.4379
gam      1.8828  0.2127  8.8523
rho1     0.9012  0.0093 96.9574
sigma1   0.0189  0.0019 10.0407
rho2     0.7755  0.0377 20.5603
sigma2   0.0120  0.0011 10.7931
nss      0.3345  0.0021 157.2114
```

**using y and w:**

```
RESULTS FROM MAXIMUM LIKELIHOOD ESTIMATION
parameters
       Estimate    s.d. t-stat

betta    0.9919  0.0047 211.8497
gam      1.9313  1.0100  1.9122
rho1     0.9060  0.0191 47.5135
sigma1   0.0201  0.0017 11.6001
rho2     0.7964  0.0545 14.6188
sigma2   0.0114  0.0013  8.8456
nss      0.3318  0.0053 62.4283
```

- the parameter estimates are pretty good. However, the true values were used as start values, so in real life, it may be more difficult.

- note that the standard errors and t statistics change quite a bit depending on which variables are used.

- When order=1, the estimation process involves forming a linear approximation to the true model, which means that the estimator is not actually the true maximum likelihood estimator, it is actually a "quasi-ML" estimator (refer to 13.7). The quasi-likelihood is computed by putting the linearized model in state-space form, and then computing the likelihood iteratively using Kalman filtering, which relies on the assumption that shocks to the model are normally distributed. State space models and Kalman filtering are introduced in Section 17.5. Once the likelihood function is available, the methods studied in this Chapter may be applied.

- The linearization of the model, combined with the fact that it has only two

shocks, leads to a problem of "stochastic singularity", which means that at most two observed variables may be used to compute the likelihood function. The code lets you explore the choice. It seems to work best using c and n. It won't work using y and r. For some choices, the true parameter values will be outside the confidence intervals. If interested, do a Google search for this term, along with DSGE, and you'll find more information.

- Not using all of the observed variables for estimation is very likely to cause problems of lack of identification and inefficiency. This can be confirmed if you experiment with the estimation script. Using c and n, we get the above results. Using other variables, we get results that aren't so good. When you don't know the true parameters, how will you choose which results to believe?

- This is not a problem of the ML method, it is a problem due to the fact that we are not really estimating the true model, we're working with a linear approximation. Often, people artificially add measurement error to

the variables, which gets around the stochastic singularity problem, possibly at the cost of introducing a worse problem. I have not yet seen a careful study of the effect of estimating assuming measurement error when there really is no measurement error.

- The intention of presenting this example is to show that ML may be used for estimation of complex models. The problem here is that the econometric model is not complex enough: the linearization throws information, so that estimation may be unreliable. A better solution is to try to actually do ML estimation for the true nonlinear model: see papers by Fernández-Villaverde and Rubio-Ramírez, which use particle filtering~~. You can also modify the script to do this, by setting order=2 in the code, as Dynare supports the particle filter option~~ (this no longer works for this simple model, one would need to add measurement error, which is a way of dealing with stochastic singularity). This is very time-consuming, though.

## 15.10 Practical Summary

The practical summary for the Chapter is here.

## 15.11 Exercises

1. Consider coin tossing with a single possibly biased coin. The random variable $Y$ is equal to 1 if a heads results, or 0 if a tails results. The probability of a heads $P(Y = 1)$ is $p_0$. Thus, the density function for the random variable is

$$\begin{aligned} f_Y(y, p_0) &= p_0^y (1 - p_0)^{1-y}, y \in \{0, 1\} \\ &= 0, y \notin \{0, 1\} \end{aligned}$$

Suppose that we have a sample of size $n$. We know, or can show, that the ML estimator is $\widehat{p_0} = \bar{y}$. We also know from the theory above that

$$\sqrt{n}\,(\bar{y} - p_0) \overset{a}{\sim} N\left[0, \mathcal{J}_\infty(p_0)^{-1}\mathcal{I}_\infty(p_0)\mathcal{J}_\infty(p_0)^{-1}\right]$$

**a)** find the analytic expression for the score contribution $g_t(\theta)$ and show that $\mathcal{E}_\theta\left[g_t(\theta)\right] = 0$

**b)** find the analytical expressions for $\mathcal{J}_\infty(p_0)$ and $\mathcal{I}_\infty(p_0)$ for this problem

**c)** verify that the result for $\lim Var\sqrt{n}\,(\hat{p}-p)$ found in section 15.4 is equal to $\mathcal{J}_\infty(p_0)^{-1}\mathcal{I}_\infty(p_0)\mathcal{J}_\infty(p_0)^{-1}$

**d)** Write an Julia program that does a Monte Carlo study that shows that $\sqrt{n}\,(\bar{y}-p_0)$ is approximately normally distributed when $n$ is large. Please give me histograms that show the sampling frequency of $\sqrt{n}\,(\bar{y}-p_0)$ for several values of $n$.

2. The exponential density is

$$f_Y(y) = \begin{cases} \frac{e^{-\frac{y}{\lambda_0}}}{\lambda_0},\ y \geqslant 0 \\ 0,\ y < 0 \end{cases}$$

Suppose we have an independently and identically distributed sample of size $n$, $\{y_i\}, i = 1, 2, ..., n$, where each $y_i$ follows this exponential distribution.

(a) write the log likelihood function

(b) find an analytic expression for the maximum likelihood estimator of the

parameter $\lambda$.

(c) explain how to estimate the asymptotic variance of the ML estimator. That is, if $\sqrt{n}\left(\hat{\lambda} - \lambda_0\right) \to^d N(0, V_\infty)$, give a consistent estimator of $V_\infty$.

(d) explain how to compute an estimator of the standard error of $\hat{\lambda}$.

3. Generate a sample of 100 observations from the exponential model of the previous question, using the Julia commands

```
using Distributions
n=100
lambda0 = 4.
y = rand(Exponential(lambda0), n)
```

and use this data to:

(a) estimate the parameter by ML, using both an analytic formula, and by

numerically minimizing the negative log likelihood function.

(b) compute the estimated standard error of the estimated parameter, and give a 95% confidence interval for $\lambda_0$.

4. Now, assume the parameter of the exponential distribution depends on a regressor: $\lambda_0 = \exp(\beta_0 + \beta_1 x)$. Generate 100 observations from the exponential model of problem 3, using the commands

```
n = 100
x = [ones(n) randn(n) rand(n)]
beta = [-0.5, 1., 1.]
lambda0 = exp.(x*beta)
y = rand.(Exponential.(lambda0))
```

and use this data to estimate the parameter by ML by numerically minimizing the negative log likelihood function. You do not need to compute the variance or standard errors, only the estimates. To do this, modify your code for the previous problem. Note that element-by-element multiplication

or division of vectors uses the .* and ./ operators, respectively.

5. Suppose we have an i.i.d. sample of size $n$ from the Poisson density. The Poisson density is $f_y(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$. Verify that the ML estimator is asymptotically distributed as $\sqrt{n}\left(\hat{\lambda} - \lambda_0\right) \xrightarrow{d} N(0, \lambda_0)$, where $\lambda_0$ is the true parameter value. Hint: compute the asymptotic variance using $-\mathcal{J}_\infty(\lambda_0)^{-1}$.

6. **ML.** Estimate a Poisson model by ML using the 10 independent data points

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| y | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 3 |

For the Poisson model, the density $f_Y(y|\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}$, $y = 0, 1, 2, \ldots$.

   (a) create a data file that contains these observations

   (b) find the log-likelihood function

   (c) find the analytic expression for the ML estimator, and find an analytic expression for the asymptotic variance of $\sqrt{n}(\hat{\lambda} - \lambda^0)$.

   (d) write a function that computes the log-likelihood function, using the form `loglikelihood(theta, data)` and maximize it to find the ML

estimator. You will probably want to use an anonymous function for this.

(e) compute the ML estimator using your analytic expression. It should be very close to what you got using fminunc. Is it? If not, revise your code to make it work better.

(f) compute the estimated standard deviation of $\hat{\lambda}$ and report an asymptotic 95% confidence interval for $\lambda^0$.

(g) now, reparameterize the model as $f_Y(y|\alpha) = \frac{\exp(-\lambda)\lambda^y}{y!}$ where $\lambda = \exp\alpha$. The advantage of this is that $\alpha$ is unrestricted in sign, while the original $\lambda$ must be positive. This doesn't matter much at present, but it will when you allow the conditional mean to depend on other variables. Verify the invariance property of ML by estimating $\alpha$, and then showing that $\hat{\lambda}$ from part (d) is equal to $\exp\hat{\alpha}$.

7. Consider the model $y_t = x_t'\beta + \sigma\epsilon_t$. Find the score function $g_n(\theta)$ where

$\theta = \begin{pmatrix} \beta' & \sigma \end{pmatrix}'$ and

(a) the errors follow the Cauchy (Student-t with 1 degree of freedom) density:

$$f(\epsilon_t) = \frac{1}{\pi\left(1+\epsilon_t^2\right)}, -\infty < \epsilon_t < \infty$$

The Cauchy density has a shape similar to a normal density, but with much thicker tails. Thus, extremely small and large errors occur much more frequently with this density than would happen if the errors were normally distributed.

(b) where the errors are independent standard normal random variables: $\epsilon_t \sim N(0,1)$.

(c) Compare the first order conditions that define the ML estimators for these two cases, and interpret the differences. *Why* are the first order conditions that define an efficient estimator different in the two cases? How do the weights on observations differ?

8. Assume a d.g.p. follows the logit model: $\Pr(y = 1|x) = (1 + exp(-\beta^0 x))^{-1}$.

   (a) Assume that $x \sim$ uniform(-a,a). Find the asymptotic distribution of the ML estimator of $\beta^0$ (this is a scalar parameter).

   (b) Now assume that $x \sim$ uniform(-2a,2a). Again find the asymptotic distribution of the ML estimator of $\beta^0$.

   (c) Comment on the results

9. Estimate the simple Nerlove model discussed in section 4.8 by ML, assuming that the errors are i.i.d. $N(0, \sigma^2)$ and compare to the results you get from running Nerlove.jl .

10. Using the fmincon routine in Econometrics

    (a) estimate the Nerlove model with the restriction that $\beta_L + \beta_F + \beta_K = 1$ (the cost function satisfies homogeneity of degree one in factor prices). Test this restriction using the likelihood ratio test.

(b) test the restriction that $\beta_Q = 1$ (the model exhibits constant returns to scale) usir

(c) test homogeneity of degree 1 and constant returns to scale jointly, using the LR test.

11. Using logit.jl and EstimateLogit.jl as templates, write a function to calculate the probit log likelihood, and a script to estimate a probit model. Run it using data that actually follows a logit model (you can generate it in the same way that is done in the logit example).

12. Study `mleresults.jl` to see what it does. Examine the functions that `it` calls. Write a complete description of how thechain works.

13. In Subsection 12.4 a model is presented for data on health care usage, along with some Julia scripts. Look at the Poisson estimation results for the OBDV measure of health care use and give an economic interpretation. Estimate Poisson models for the other 5 measures of health care usage, using the provided scripts.

14. For practice using fminunc, estimate a Poisson model by ML using the 10 independent data points

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| y | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 3 |
| x | -1 | -1 | 1 | 0 | -1 | -1 | 1 | 1 | 2 | 2 |

.

For the Poisson model, the density $f_Y(y|x) = \frac{\exp(-\lambda)\lambda^y}{y!}$, $y = 0, 1, 2, \ldots$. To make the model depend on conditioning variables, use the parameterization $\lambda(x) = \exp(\theta_1 + \theta_2 x)$. The example EstimatePoisson.jl, in the notes, should be helpful

(a) create a data file that contains these observations

(b) find the log-likelihood function

(c) write a Julia function that computes the log-likelihood function.

(d) use fminunc to find the ML estimator. You need to use an anonymous function for this.

# Chapter 16

# Generalized method of moments

**Readings**: Cameron and Trivedi (2005), Ch. 6; Hamilton Ch. 14*; Davidson and MacKinnon, Ch. 17 (see pg. 587 for refs. to applications), Hansen (1982), Hansen and Singleton (1982), Newey and McFadden (1994).

## 16.1 Moment conditions

**Definition 50.** Moment condition: a moment condition $\bar{m}_n(\theta) = \bar{m}_n(Z_n, \theta)$ is a vector-valued function of the data $Z_n$ and the parameter $\theta$ that has mean zero, under the model, when evaluated at the true parameter value $\theta_0$, and expectation different from zero when evaluated at other parameter values:

$$E\bar{m}_n(Z_n, \theta_0) = 0$$
$$E\bar{m}_n(Z_n, \theta) \neq 0,\ \theta \neq \theta_0$$

- The expectation operator $E$ supposes that expectations are taken with respect to the true density of the data. This may depend on more parameters than appear in $\theta$, if the model is semi-parametric.
- The moment condition may be vector-valued, with dimension $G$, say.
- There are a couple of other details in the definition, which we'll get to.

**Definition 51.** Moment contribution: we will be dealing with moment conditions that are defined as averages: $\bar{m}_n(\theta) = \frac{1}{n}\Sigma_t\, m(Z_t, \theta) = \frac{1}{n}\Sigma_t\, m_t(\theta)$. The functions $m_t(\theta)$ are the *moment contributions.* The $t$-th moment contribution, $m_t$, is a function of the same observation's data. I'm casually using $m(Z_t, \theta)$, $m_t(\theta)$ and $m_t$ to all refer to the same thing. This first of these is the full expression, but I will suppress arguments when the context makes things clear enough, to reduce the notational burden. The main thing is that $\bar{m}_n$ refers to the average over the $n$ observations, and $m_t$ refers to the terms that are averaged.

**Example 52.** OLS. The classical linear model. Let $\bar{m}_n(\beta) = \frac{1}{n}\Sigma_t\, x_t(y_t - x_t'\beta)$. So the moment contributions are $m_t(\beta) = x_t(y_t - x_t'\beta)$. When $\beta = \beta_0$, $y_t - x_t'\beta_0 = \epsilon_t$, and $m_t = x_t\epsilon_t$. We know that $E(x_t\epsilon_t) = 0$, by the weak exogeneity assumption. Thus, the moment contributions, and the moment condition, which is their average, have expectation zero when evaluated at the true parameter value.

**Example 53.** ML. We have seen (see eqn. 15.4) that the score contributions of the ML estimator have mean zero: $E\left(D_\theta \ln f(y_t|x_x, \theta_0)\right) = 0$. So, we could set $m_t(\theta) = D_\theta \ln f(y_t|x_x, \theta)$.

Of course, you would need to know the density $f$ to compute this, so could only use these moment conditions if you had enough information to do ML estimation, so there wouldn't be much point in doing GMM. However, we gain the insight that the optimal moments would be the score vector.

**Example 54.** Sampling from $\chi^2$. Suppose we draw a random sample of $y_t$ from the $\chi^2(\theta_0)$ distribution. Here, $\theta_0$ is the parameter of interest. If $Y \sim \chi^2(\theta_0)$, then the mean $E(Y) = \theta_0$. Let the moment contribution be

$$m_t(\theta) = y_t - \theta$$

Then

$$\bar{m}_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} m_t(\theta) = \bar{y} - \theta$$

We know that $E(\bar{y}) = \theta_0$.

- Thus, $E\bar{m}_n(\theta_0) = 0$.
- However, $E\bar{m}_n(\theta) = \theta_0 - \theta \neq 0$ if $\theta \neq \theta_0$.

When the dimension of the moment conditions is the same as the dimension of the parameter vector $\theta$, the *method of moments principle* is to choose the estimator of the parameter to *set the moment condition equal to zero*: $\bar{m}_n(\hat{\theta}) \equiv 0$. Then the equation is solved for the estimator.

- In the case of OLS, this gives $\Sigma_t\, x_t(y_t - x_t'\hat{\beta}) = 0$, which gives a solution that you should already know.

- For the chi-squared example,

$$\bar{m}(\hat{\theta}) = \bar{y} - \hat{\theta} = 0$$

  is solved by $\hat{\theta} = \bar{y}$. Since $\bar{y} = \Sigma_{t=1}^{n}\, y_t/n \xrightarrow{p} \theta_0$ by the LLN, the estimator is consistent.

**Example 55.** $\chi^2$, version 2. The variance of a $\chi^2(\theta_0)$ r.v. is

$$V(Y) = E(Y - \theta_0)^2 = 2\theta_0.$$

Let

$$m_t(\theta) = \frac{n}{n-1}(y_t - \bar{y})^2 - 2\theta$$

Then

$$\bar{m}_n(\theta) = \frac{\Sigma_{t=1}^n (y_t - \bar{y})^2}{n-1} - 2\theta.$$

The first term is the unbiased formula for the sample variance, and thus has expectation equal to $2\theta_0$. So if we evaluate $\bar{m}_n(\theta)$ at $\theta_0$, the expectation is zero.

The MM estimator using the variance would set

$$\bar{m}_n(\hat{\theta}) = \frac{\Sigma_{t=1}^n (y_t - \bar{y})^2}{n-1} - 2\hat{\theta} \equiv 0.$$

Solving for the estimator, it is half the sample variance:

$$\hat{\theta} = \frac{1}{2}\frac{\Sigma_{t=1}^{n}\left(y_t - \bar{y}\right)^2}{n-1}.$$

Again, by the LLN, the sample variance is consistent for the true variance, that is,

$$\frac{\Sigma_{t=1}^{n}\left(y_t - \bar{y}\right)^2}{n} \xrightarrow{p} 2\theta_0.$$

So, this MM is also consistent for $\theta_0$.

**Example 56.** Try some MM estimation yourself: here's a Julia script that implements the two MM estimators discussed above: GMM/chi2mm.jl

Note that when you run the script, the two estimators give different results. Each of the two estimators is consistent.

- For the $\chi^2$ example, we have two alternative moment conditions and only one parameter: we have *overidentification,* which means that we have more information than is strictly necessary for consistent estimation of the parameter.

- The idea behind GMM is to combine information from the two moment conditions to form a new estimator which will be *more efficient,* in general (proof of this below).

- Note that the fact that the data has a chi-squared distribution is not used in estimation, it just as easily could have been normally distributed, sampling

from a $N(\theta_0, 2\theta_0)$ distribution. As long as the assumptions regarding the mean or variance are correct, the MM estimators are consistent. So, we don't make use of distributional assumptions when doing method of moment estimation, we only rely on certain moments being correctly specified.

- In this way, method of moments estimation is *more robust* than is maximum likelihood estimation: we obtain a consistent estimator with fewer assumptions.
- There being no free lunch, we should expect to pay something for this, of course. The cost will be a loss of efficiency, in general.

**To summarize**, a moment condition is a vector valued function which has expectation zero at the true parameter value. We have seen some examples of where we might get such functions, and more will follow. For now, let's take moment conditions as given, and work out the properties of the estimator.

## 16.2 Definition of GMM estimator

For the purposes of this course, the following definition of the GMM estimator is sufficiently general:

**Definition 57.** The GMM estimator of the $k$ -dimensional parameter vector $\theta^0$,

$$\hat{\theta} \equiv \arg\min_{\Theta} \bar{m}_n(\theta)' W_n \bar{m}_n(\theta),$$

- where $\bar{m}_n(\theta) = \frac{1}{n}\Sigma_{t=1}^{n} m(Z_t, \theta)$ is a $g$-vector valued function, $g \geq k$, with $\mathcal{E}m(Z_t, \theta^0) = 0$,$\mathcal{E}m(Z_t, \theta) = 0$, $\forall \theta \neq \theta_0$

- and $W_n$ converges almost surely to a finite $g \times g$ symmetric positive definite matrix $W_\infty$.

*What's the reason for using GMM if MLE is asymptotically efficient?*

- Robustness: GMM is based upon a limited set of moment conditions. For consistency, only these moment conditions need to be correctly specified, whereas MLE in effect requires correct specification of *every conceivable* moment condition. GMM is *robust with respect to distributional misspecification.* The price for robustness is usually a loss of efficiency with respect to the MLE estimator. Keep in mind that the true distribution is not known, so if we erroneously specify a distribution and estimate by "MLE", the estimator will in fact be what is known as a *quasi-ML* or *pseudo-ML* estimator (Gourieroux et al. (1984)), and it will be inconsistent in general (but not in some special cases).

- Feasibility: in some cases the MLE estimator is not available, because we are not able to deduce or compute the likelihood function. More on this in the section on simulation-based estimation. The GMM estimator may still be feasible even though MLE is not available.

**Example 58.** The Julia script GMM/chi2gmm.jl implements GMM using the same $\chi^2$ data as was using in Example 56, above. The two moment conditions, based on the sample mean and sample variance are combined. The weight matrix is an identity matrix, $I_2$.

## 16.3 Consistency

We simply assume that the assumptions of Theorem 35 hold, so the GMM estimator is strongly consistent. The main requirement is that the moment conditions have mean zero at the true parameter value, $\theta^0$. This will be the case if our moment conditions are correctly specified. With this, it is clear that the minimum of the limiting objective function occurs at the true parameter value. The only assumption that warrants additional comment is that of identification. In Theorem 35, the third assumption reads: (c) *Identification:* $s_\infty(\cdot)$ has a unique global maximum at $\theta^0$, *i.e.*, $s_\infty(\theta^0) > s_\infty(\theta)$, $\forall \theta \neq \theta^0$. Taking the case of a quadratic objective function $s_n(\theta) = \bar{m}_n(\theta)' W_n \bar{m}_n(\theta)$, first consider $\bar{m}_n(\theta)$.

- Applying a uniform law of large numbers, we get $\bar{m}_n(\theta) \xrightarrow{a.s.} m_\infty(\theta)$.
- Since $E\bar{m}_n(\theta^0) = 0$ by assumption, $m_\infty(\theta^0) = 0$.
- Since $s_\infty(\theta^0) = m_\infty(\theta^0)' W_\infty m_\infty(\theta^0) = 0$, in order for asymptotic identification, we need that $s_\infty(\theta) > 0$ for $\theta \neq \theta^0$. This requires that $m_\infty(\theta) \neq 0$

for $\theta \neq \theta^0$, for at least some element of the vector. There can be no other parameter value that sets the moment conditions to zero (at least, in the limit). *Draw picture here.* This and the assumption that $W_n \stackrel{a.s.}{\to} W_\infty$, a finite positive $g \times g$ definite $g \times g$ matrix guarantee that $\theta^0$ is asymptotically identified.

- an example of lack of identification is a MA(1) model $y_t = \epsilon_t + \theta\epsilon_{t-1}$ where the $\epsilon_t$ are i.i.d. $N(0,1)$.
  - consider the alternative MA(1) model $y_t = \epsilon_t + (1/\theta)\epsilon_{t-1}$ where the $\epsilon_t$ are i.i.d. $N(0,\theta^2)$.
  - All of the moments of $y_t$ are the same in the two cases. For example, for the first model, $V(y_t) = V(\epsilon_t)+\theta^2 V(\epsilon_{t-1}) = 1+\theta^2)$. For the alternative model, you can work out that the variance is the same.
  - So, a method of moments estimator can't distinguish between the two parameter vales. If $m_\infty(\theta)$ is equal to zero, then so is $m_\infty(1/\theta)$.
  - If we restrict the parameter of the model to be in the $(-1,1)$ interval, we eliminate the problem. This restriction makes the MA(1) model *invertible*, which is useful for theoretical reasons (more on this later).
- Note that asymptotic identification does not rule out the possibility of lack of identification for a given data set - there may be multiple minimizing

solutions in finite samples.

**Example 59.** Increase $n$ in the Julia script GMM/chi2gmm.jl to see evidence of the consistency of the GMM estimator.

## 16.4 Asymptotic normality

We also simply assume that the conditions of Theorem 37 hold, so we will have asymptotic normality. (Note that this implies that we are assuming that the moment conditions are continuous and differentiable.) However, we do need to find the structure of the asymptotic variance-covariance matrix of the estimator. From Theorem 37, we have

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$$

where $\mathcal{J}_\infty(\theta^0)$ is the almost sure limit of $\frac{\partial^2}{\partial\theta\partial\theta'}s_n(\theta)$ when evaluated at $\theta^0$ and

$$\mathcal{I}_\infty(\theta^0) = \lim_{n\to\infty} Var\sqrt{n}\frac{\partial}{\partial\theta}s_n(\theta^0).$$

We need to determine the form of these matrices given the objective function $s_n(\theta) = \bar{m}_n(\theta)'W_n\bar{m}_n(\theta)$.

Now, using the product rule from section 23.1,

$$\frac{\partial}{\partial\theta}s_n(\theta) = 2\left[\frac{\partial}{\partial\theta}\bar{m}_n'(\theta)\right]W_n\bar{m}_n(\theta)$$

(this is analogous to $\frac{\partial}{\partial\beta}\beta'X'X\beta = 2X'X\beta$ which appears when computing the first order conditions for the OLS estimator).

Define the $k \times g$ matrix

$$D_n(\theta) \equiv \frac{\partial}{\partial\theta}\bar{m}_n'(\theta), \tag{16.1}$$

so:

$$\frac{\partial}{\partial\theta}s(\theta) = 2D(\theta)W\bar{m}(\theta). \tag{16.2}$$

(Note that $s_n(\theta)$, $D_n(\theta)$, $W_n$ and $\bar{m}_n(\theta)$ all depend on the sample size $n$, but it is omitted to unclutter the notation).

To take second derivatives, let $D_i$ be the $i-$ th row of $D(\theta)$. This is a $1 \times g$ row vector, and

$$\frac{\partial}{\partial \theta_i} s(\theta) = 2D_i(\theta) W \bar{m}(\theta)$$

**is a scalar**. It element in the $i$th row of the column vector $\frac{\partial}{\partial \theta} s(\theta)$. The $i$th row of the matrix of second derivatives is (using the product rule in definition 23.1), is the derivative of this real-valued function, with respect to $\theta'$ :

$$\begin{aligned} \frac{\partial}{\partial \theta'} \frac{\partial}{\partial \theta_i} s(\theta) &= \frac{\partial}{\partial \theta'} \left[ 2D_i(\theta) W \bar{m}(\theta) \right] \\ &= 2D_i W D' + 2\bar{m}' W \left[ \frac{\partial}{\partial \theta'} D_i' \right] \end{aligned}$$

***Note to self for lectures:*** $D_i'$ is $g \times 1$, so $\left[\frac{\partial}{\partial \theta'} D_i'\right]$ is $g \times k$.
Note that the first term contains a $D'$, which appears due to $\frac{\partial}{\partial \theta'} \bar{m}_n(\theta)$, which is the transpose of what we defined in eqn. 16.1. When evaluating the second term:

$$2\bar{m}(\theta)' W \left[ \frac{\partial}{\partial \theta'} D(\theta)_i' \right]$$

(where the dependence of $D$ upon $\theta$ is emphasized) at $\theta^0$, assume that $\frac{\partial}{\partial\theta'}D(\theta)_i'$ satisfies a LLN (it is an average), so that it converges almost surely to a finite limit. In this case, we have

$$2\bar{m}(\theta^0)'W\left[\frac{\partial}{\partial\theta'}D(\theta^0)_i'\right] \overset{a.s.}{\to} 0,$$

because $\bar{m}(\theta^0) \overset{a.s.}{\to} 0$ and $W \overset{a.s.}{\to} W_\infty$.

Stacking these results over the $k$ rows of $D$, we get

$$\lim \frac{\partial^2}{\partial\theta\partial\theta'} s_n(\theta^0) = \mathcal{J}_\infty(\theta^0) = 2D_\infty W_\infty D'_\infty, a.s.,$$

where we define $\lim D = D_\infty$, $a.s.$, and $\lim W = W_\infty$, a.s. (we assume a LLN holds).

With regard to $\mathcal{I}_\infty(\theta^0)$, following equation 16.2, and noting that the scores have mean zero at $\theta^0$ (since $\mathcal{E}\bar{m}(\theta^0) = 0$ by assumption), we have

$$\begin{aligned}
\mathcal{I}_\infty(\theta^0) &= \lim_{n\to\infty} Var\sqrt{n}\frac{\partial}{\partial\theta}s_n(\theta^0) \\
&= \lim_{n\to\infty} \mathcal{E}4nDW\bar{m}(\theta^0)\bar{m}(\theta^0)'WD' \\
&= \lim_{n\to\infty} \mathcal{E}4DW\left\{\sqrt{n}\bar{m}(\theta^0)\right\}\left\{\sqrt{n}\bar{m}(\theta^0)'\right\}WD'
\end{aligned}$$

Now, given that $\bar{m}(\theta^0)$ is an average of centered (mean-zero) quantities, it is reasonable to expect a CLT to apply, after multiplication by $\sqrt{n}$. Assuming this,

$$\sqrt{n}\bar{m}(\theta^0) \xrightarrow{d} N(0, \Omega_\infty), \tag{16.3}$$

where

$$\Omega_\infty = \lim_{n\to\infty} \mathcal{E}\left[n\bar{m}(\theta^0)\bar{m}(\theta^0)'\right].$$

Using this, and the last equation, we get

$$\mathcal{I}_\infty(\theta^0) = 4D_\infty W_\infty \Omega_\infty W_\infty D'_\infty$$

Using these results, the asymptotic normality theorem (37) gives us

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty W_\infty D'_\infty\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D'_\infty \left(D_\infty W_\infty D'_\infty\right)^{-1}\right],$$

the asymptotic distribution of the GMM estimator for arbitrary weighting matrix $W_n$.

- Note that for $J_\infty$ to be positive definite, $D_\infty$ must have full row rank, $\rho(D_\infty) = k$.
- This is related to identification: we need that the parameters cause the moments to change, and each parameter must cause a change that is separate from the changes caused by the other parameters.
- Identification ($D_\infty$ being full row rank), plus $W_\infty$ being positive definite,

imply that $J_\infty$ is positive definite.

There are two things that affect the asymptotic variance:

- the choice of the moment conditions, $\bar{m}_n(\theta)$, which determines both $D_\infty$ and $\Omega_\infty$
- the choice of the weight matrix $W_n$, which determines $W_\infty$

We would probably like to know how to choose both $\bar{m}_n(\theta)$ and $W_n$ so that the asymptotic variance is a small as possible. That will be the topic of the next section.

**Example 60.** *Asymptotic normality of a GMM estimator.* The Julia script GMM/AsymptoticNormalityGMM.jl does a Monte Carlo of the GMM estimator for the $\chi^2$ data. Histograms for 1000 replications of $\sqrt{n}\left(\hat{\theta} - \theta^0\right)$ are given in Figure 16.1. On the left are results for $n = 10$, on the right are results for $n = 1000$. Note that the two distributions are more or less centered at 0. The distribution for the small sample size is somewhat asymmetric, which shows that the small sample distribution may be poorly approximated by the asymptotic distribution. This has mostly disappeared for the larger sample size.

Figure 16.1: Asymptotic Normality of GMM estimator, $\chi^2$ example

(a) $n = 10$ (b) $n = 1000$

## 16.5 Choosing the weighting matrix

$W$ is a *weighting matrix,* which determines the relative importance of violations of the individual moment conditions in the overall objective function. For example, if we are much more sure of the first moment condition, which is based upon the variance, than of the second, which is based upon the fourth moment, we could

set

$$W = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

with $a$ much larger than $b$. In this case, errors in the second moment condition have less weight in the objective function.

- Since moments are not independent, in general, we should expect that there be a correlation between the moment conditions, so it may not be desirable to set the off-diagonal elements to 0. $W$ may be a random, data dependent matrix.

- We have already seen that the choice of $W$ will influence the asymptotic distribution of the GMM estimator. Since the GMM estimator is already inefficient w.r.t. MLE, we might like to choose the $W$ matrix to make the GMM estimator efficient *within the class of GMM estimators* defined by $\bar{m}_n(\theta)$.

- To provide a little intuition, consider the linear model $y = \mathbf{x}'\beta + \varepsilon$, where $\varepsilon \sim N(0, \Omega)$. That is, he have heteroscedasticity and autocorrelation.
  - The generalized least square estimator minimizes the objective function $(y - \mathbf{X}\beta)'\Omega^{-1}(y - \mathbf{X}\beta)$. We have seen that the GLS estimator is efficient with respect to OLS, when there is het. and or aut.
  - The GLS optimal weighting matrix is seen to be the inverse of the covariance matrix of the errors. A similar result holds for GMM estimation.
  - Note: this presentation of GLS is not a GMM estimator as defined above, because if we take the errors as "moment conditions", the dimension is the sample size, $n$. Thus, the dimension is not fixed. Also, they are not averages, as we require - see definition 57. Later we'll see that GLS can be expressed in the GMM framework.

**Theorem 61.** *If $\hat{\theta}$ is a GMM estimator that minimizes $\bar{m}_n(\theta)'W_n\bar{m}_n(\theta)$, the asymptotic variance of $\hat{\theta}$ will be minimized by choosing $W_n$ so that $W_n \xrightarrow{a.s} W_\infty = \Omega_\infty^{-1}$, where $\Omega_\infty = \lim_{n\to\infty} \mathcal{E}\left[nm(\theta^0)m(\theta^0)'\right]$.*

***Proof:*** *For $W_\infty = \Omega_\infty^{-1}$, the asymptotic variance*

$$\left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' \left(D_\infty W_\infty D_\infty'\right)^{-1}$$

*simplifies to $\left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}$. Now, let $A$ be the difference between the general form and the simplified form:*

$$A = \left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' \left(D_\infty W_\infty D_\infty'\right)^{-1} \; - \; \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}$$

*Set $B = \left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty - \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1} D_\infty \Omega_\infty^{-1}$. One can show that $A = B\Omega_\infty B'$. This is a quadratic form in a p.d. matrix, so it is p.s.d., which concludes the proof.*

The result

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0,\left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}\right] \tag{16.4}$$

allows us to treat

$$\hat{\theta} \approx N\left(\theta^0, \frac{\left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}}{n}\right),$$

where the $\approx$ means "approximately distributed as." To operationalize this we need estimators of $D_\infty$ and $\Omega_\infty$.

- The obvious estimator of $\widehat{D_\infty}$ is simply $\frac{\partial}{\partial\theta}\bar{m}_n\left(\hat{\theta}\right)$, which is consistent by the consistency of $\hat{\theta}$, assuming that $\frac{\partial}{\partial\theta}\bar{m}_n$ is continuous in $\theta$. Stochastic equicontinuity results can give us this result even if $\frac{\partial}{\partial\theta}\bar{m}_n$ is not continuous.

- Estimating $\Omega_\infty$ is taken up below.

**Example 62.** To see the effect of using an efficient weight matrix, consider the Julia script GMM/EfficientGMM.jl. This modifies the previous Monte Carlo for the $\chi^2$ data. This new Monte Carlo computes the GMM estimator in two ways:

- based on an identity weight matrix
- using an estimated optimal weight matrix. The estimated efficient weight matrix is computed as the inverse of the estimated covariance of the moment conditions, using the inefficient estimator of the first step. See the next section for more on how to do this.
- The following figure shows the results, plotting densities for 1000 replications of $\sqrt{n}\left(\hat{\theta} - \theta^0\right)$. Note that the use of the estimated efficient weight matrix leads to much better results in this case. This is a simple case where it is possible to get a good estimate of the efficient weight matrix. This is not always so. See the next section.

Inefficient
0.100
0.075
0.050
0.025
0.000
−10
0
10
20
q05
90% non-rejection region
q95
mean
median
Efficient
0.15
0.10
0.05
0.00
−10
0
10
20
q05
90% non-rejection region
q95
mean
median

What's the intuition behind $\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0,\left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}\right]$?

- More variability of the moment conditions ($\Omega_\infty$ "large") leads to a more variable estimator, other things equal.

- The larger the magnitude of the elements of the Jacobian of the moment conditions, $D_\infty$, the smaller the variance of the estimator. Moment conditions which change more in response to changes in parameters lead to more precise estimates.

- *in* class*:* make a figure, much like Figure 13.5 .

## 16.6 Estimation of the variance-covariance matrix

**(See Hamilton Ch. 10, pp. 261-2 and 280-84)*.**

In the case that we wish to use the optimal weighting matrix, we need an estimate of $\Omega_\infty$, the limiting variance-covariance matrix of $\sqrt{n}\bar{m}_n(\theta^0)$. Remember that $\bar{m}_n$ is the average of the moment contributions, $m_t$, and, by assumption, $E(m_t(\theta^0) = 0$. In general, we expect that:

- $m_t$ will be autocorrelated ($\Gamma_{ts} = \mathcal{E}(m_t m'_{t-s}) \neq 0$). Note that this autocovariance matrix, which is of dimension $g \times g$, will not depend on $t$ if the moment conditions are covariance stationary.

- contemporaneously correlated, since the individual moment contributions will not in general be independent of one another ($\mathcal{E}(m_{it} m_{jt}) \neq 0$), where $i, j \in \{1, 2, ..., k\}$.

- and have different variances ($\mathcal{E}(m_{it}^2) = \sigma_{it}^2$ ).

Since we need to estimate so many components, it is unlikely that we would arrive at a correct parametric specification. For this reason, research has focused on consistent nonparametric estimators of $\Omega_\infty$.

Henceforth we assume that $m_t$ is *covariance stationary*, so the covariance between $m_t$ and $m_{t-s}$ does not depend on $t$. (See the first part of Chapter 17 for the definition). Thus,

**Definition 63.** (Autocovariance). Define the $s-th$ autocovariance of covariance stationary moment contributions as $\Gamma_s = \mathcal{E}(m_t m'_{t-s})$.

Because of stationarity, $\Gamma_s$ does not depend on $t$.

**Exercise 64.** Show that $\mathcal{E}(m_t m'_{t+s}) = \Gamma'_s$.

Recall that $m_t$ and $\bar{m}_n$ are functions of $\theta$, so, for now, assume that we have some consistent estimator of $\theta^0$. With this, a consistent estimator of $m_t(\theta^0)$ is $\hat{m}_t = m_t(\hat{\theta})$. Now

$$\begin{aligned}
\Omega_n &= \mathcal{E}\left[n\bar{m}_n(\theta^0)\bar{m}_n(\theta^0)'\right] = \mathcal{E}\left[n\left(1/n\sum_{t=1}^{n} m_t\right)\left(1/n\sum_{t=1}^{n} m_t'\right)\right] \\
&= \mathcal{E}\left[1/n\left(\sum_{t=1}^{n} m_t\right)\left(\sum_{t=1}^{n} m_t'\right)\right] \\
&= \mathcal{E}\left[1/n\left(m_1 + m_2 + ... + m_n\right)\left(m_1' + m_2' + ... + m_n'\right)\right] \\
&= \Gamma_0 + \frac{n-1}{n}\left(\Gamma_1 + \Gamma_1'\right) + \frac{n-2}{n}\left(\Gamma_2 + \Gamma_2'\right)\cdots + \frac{1}{n}\left(\Gamma_{n-1} + \Gamma_{n-1}'\right)
\end{aligned}$$

A natural, consistent estimator of $\Gamma_s$ is

$$\widehat{\Gamma_s} = 1/n\sum_{t=s+1}^{n} \hat{m}_t\hat{m}_{t-s}'.$$

(you might use $n - s$ in the denominator instead). This is consistent because of the LLN, and the fact that $\hat{\theta}$ is consistent for $\theta^0$ (Slutky theorem).

So, a natural, but inconsistent, estimator of $\Omega_\infty$ would be

$$\begin{aligned}\hat{\Omega} &= \widehat{\Gamma_0} + \frac{n-1}{n}\left(\widehat{\Gamma_1} + \widehat{\Gamma'_1}\right) + \frac{n-2}{n}\left(\widehat{\Gamma_2} + \widehat{\Gamma'_2}\right) + \cdots + \left(\widehat{\Gamma_{n-1}} + \widehat{\Gamma'_{n-1}}\right) \\ &= \widehat{\Gamma_0} + \sum_{s=1}^{n-1} \frac{n-s}{n}\left(\widehat{\Gamma_s} + \widehat{\Gamma'_s}\right).\end{aligned}$$

This estimator is inconsistent in general, since the number of parameters to estimate is more than the number of observations, and increases more rapidly than $n$, so information does not build up as $n \to \infty$. There is always only one observation to estimate the highest order autocovariance.

On the other hand, supposing that $\Gamma_s$ tends to zero *sufficiently rapidly* as $s$ tends to $\infty$, a modified estimator is

$$\hat{\Omega} = \widehat{\Gamma_0} + \sum_{s=1}^{q(n)} \left(\widehat{\Gamma_s} + \widehat{\Gamma'_s}\right).$$

- this will be consistent, provided that
  - $q(n) \xrightarrow{p} \infty$ as $n \to \infty$, so that the higher order terms are eventually

estimated, as the sample size becomes large enough (control the bias)‘

- $q(n)$ grows sufficiently slowly, so that we don't try to estimate terms unless we have enough informations (control the variance)
- The term $\frac{n-s}{n}$ can be dropped because it converges to 1.
- A disadvantage of this estimator is that it may not be positive definite. This could cause one to calculate a negative $\chi^2$ statistic, for example!

## Heteroscedasticity and autocorrelation consistent covariance estimators

The Newey-West estimator (Newey and West, 1987) solves the problem of possible non-positive definiteness of the above estimator. Their estimator is

$$\hat{\Omega} = \widehat{\Gamma}_0 + \sum_{s=1}^{q(n)} \left[1 - \frac{s}{q+1}\right] \left(\widehat{\Gamma}_s + \widehat{\Gamma}_s'\right).$$

This estimator is p.d. by construction. The condition for consistency is that $n^{-1/4}q \to 0$. Note that this is a very slow rate of growth for $q$. This estimator is nonparametric - we've placed no parametric restrictions on the form of $\Omega$. It is an example of a *kernel* estimator. Kernel estimators are discussed in more detail in Chapter 20.

- Around the same time as the paper by Newey and West, a number of other similar covariance matrix estimators were proposed, but the NW estimator seems to be the most widely used in empirical work.

- If there is no autocorrelation of the moments, then all $\Gamma_s$, $s > 0$ may be set to zero. The result is White's heteroscedastic consistent variance covariance estimator, White (1980a).

- A Julia implementation is at NeweyWest.jl.

- **in class:** Use Gretl for examples of both:

  - het: Nerlove model: look at t stats with ordinary and White standard errors.

  - aut: SP500 data (in Julia directory, load the csv file), rets and rets squared. Estimate an AR(4), with plain and HAC standard errors.

## 16.7 Two step and continuously updated GMM estimators

### Two step GMM estimator:

The most common way to do efficient GMM estimation is the two step GMM estimator:

1. Set the weight matrix to some positive definite matrix. Most commonly, one uses an identity matrix of order $g$. Obtain the GMM estimator that minimizes $s_n(\theta) = \bar{m}_n(\theta)' W \bar{m}_n(\theta)$

2. Based on this initial estimate, $\hat{\theta}$, compute the moment contributions $m_t(\hat{\theta})$, $t = 1, 2, ..., n$. Compute an estimate of $\Omega_\infty$ based on the moment contributions, say $\hat{\Omega}$. The exact way to do this will depend upon the assumptions of the model. For example, if moment conditions are suspected to be autocorrelated, one might use the Newey-West estimator. Given the estimate,

compute the efficient GMM estimator which minimizes

$$s_n(\theta) = \bar{m}_n(\theta)'\hat{\Omega}^{-1}\bar{m}_n(\theta).$$

- Note that $\hat{\Omega}^{-1}$ is fixed while numeric minimization finds the second step estimator. The result is the two step estimator.
- An example of this is given by running `gmmresults()`, using gmmresults.jl.
- Matlab/Octave code for GMM for Nerlove model. Examine and run TwoStepGMM.m , which illustrates how to do two step GMM for the Nerlove data. Note that the GMM results are the same as what you get estimating by OLS.

## Continuously updated GMM estimator:

The continuously updated estimator (Hansen et al. (1996)) solves a minimization problem where the efficient weight matrix is estimated at each iteration of the numeric optimization process, using the current value of $\theta$. The CUE estimator solves the minimization problem

$$s_n(\theta) = \bar{m}_n(\theta)'\hat{\Omega}(\theta)^{-1}\bar{m}_n(\theta).$$

- Note that the covariance of the moment conditions will be updated at each trial value of the objective function during the course of minimization.
- This estimator is equivalent to an iterated version of the two step estimator.
- The CUE estimator can be shown to have a smaller bias than does the two step estimator, which may have a large small sample bias (Newey and Smith (2003)).

- An example of CUE estimation is given by running **`gmmresults()`**, using gmmresults.jl.

## 16.8 Estimation using conditional moments

So far, the moment conditions have been presented as unconditional expectations. One common way of defining unconditional moment conditions is based upon conditional moment conditions.

This is important from the point of view of constructing an econometric model, since economic models often imply restrictions on *conditional* moments.

- Suppose a model tells us that the function $\epsilon(y_t, x_t, \theta^0)$ has expectation, conditional on the information set $I_t$, equal to zero

$$E\left(\epsilon(y_t, x_t, \theta^0)|I_t\right) = 0.$$

For example, consider the supply equation of eqn. 11.1:

$$q_t = \beta_1^0 + \beta_2^0 p_t + \varepsilon_{2t}$$

where income, $y_t$, is exogenous, and $E(\epsilon|y) = 0$.

- Set $\epsilon(y_t, x_t, \theta^0) = q_t - \beta_1^0 - \beta_2^0 p_t = \varepsilon_{2t}$
- So $\epsilon_t(\theta^0) = \epsilon_{2t}$, which has conditional expectation, given $y_t$, equal to zero, by assumption.
- At other parameter values, $\epsilon_t(\theta) = q_t - \beta_1 + \beta_2 p_t \neq \epsilon_{2t}$, and this will not have conditional expectation equal to zero.

Returning to the general case, if identification holds, then we will have

$$\mathcal{E}\epsilon_t(\theta)|I_t \neq 0,\ \theta \neq \theta^0.$$

This is a scalar moment condition, which isn't sufficient to identify a $K$ -dimensional parameter $\theta$ ($K > 1$). However, we can take any variables in the information set, say, $w_t$ (which may be vector-valued), and we may use tranformations of these variables, say, $Z(w_t)$,to obtain the moment contributions

$$m_t(\theta) = Z(w_t)\epsilon_t(\theta).$$

Now, let $Z(w_t)$ be a $g \times 1$-vector valued function of $w_t$ and $w_t$ is a set of variables drawn from the information set $I_t$. The $Z(w_t)$ are *instrumental variables.* We now have $g$ moment conditions, so as long as $g > K$ the necessary condition for identification holds.

- By the law of iterated expectation (43), $E(m_t(\theta^0) = 0$.

- The other thing we need, for identification, is that $E(m_t(\theta) \neq 0, \forall \theta \neq \theta^0$.

- In the above supply function example, $y_t$ is a variable in the information set. We can create instruments by using functions of $y_t$. Using only $y_t$ itself, by the law of iterated expectations, $E\left(y_t \epsilon_t(\theta^0)\right) = 0$.

  - In this example, we have only one moment condition, so far. We would need additional instruments to identify the parameters. We can use the constant as an instrument: $Z(y_t^0) = 1, \forall t$. Combining this with the previous instrument gives us two instruments, which is enough to estimate the two $\beta$ parameters.

Following the idea, one can form the $n \times g$ matrix

$$
\begin{aligned}
Z_n &= \begin{bmatrix} Z_1(w_1) & Z_2(w_1) & \cdots & Z_g(w_1) \\ Z_1(w_2) & Z_2(w_2) & & Z_g(w_2) \\ \vdots & & & \vdots \\ Z_1(w_n) & Z_2(w_n) & \cdots & Z_g(w_n) \end{bmatrix} \\
&= \begin{bmatrix} Z_1' \\ Z_2' \\ \\ Z_n' \end{bmatrix}
\end{aligned}
$$

With this we can form the $g$ moment conditions

$$
\bar{m}_n(\theta) = \frac{1}{n} Z_n' \begin{bmatrix} \epsilon_1(\theta) \\ \epsilon_2(\theta) \\ \vdots \\ \epsilon_n(\theta) \end{bmatrix}
$$

With this, we can write

$$\begin{aligned}\bar{m}_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n} Z_t\epsilon_t(\theta) \\ &= \frac{1}{n}\sum_{t=1}^{n} m_t(\theta)\end{aligned}$$

where $Z_{(t,\cdot)}$ is the $t^{th}$ row of $Z_n$.

- This fits the previous treatment. It is an example of how moment conditions may be defined from conditional expectations and instrumental variables created using the conditioning information.

## 16.9 Generalized instrumental variables estimator for linear models

The IV estimator may appear a bit unusual at first, but it will grow on you over time.

Let's look at the previous section's results in more detail, for the commonly encountered special case of a linear model with iid errors (the iid part is easy to generalize), but with correlation between regressors and errors (failure of weak exogeneity):

$$\begin{aligned} y_t &= x_t'\theta^0 + \varepsilon_t \\ \mathcal{E}(x_t\varepsilon_t) &\neq 0 \end{aligned}$$

- Let's assume, just to keep things simple, that the errors are iid
- The model in matrix form is $y = X\theta^0 + \epsilon$

**We have seen some cases where this problem arises:**

1. measurement error of regressors: Example 20

2. lagged dependent variable and autocorrelated errors: Example 24

3. simultaneous equations: Section 11.2

Let $K = dim(x_t)$. Consider some vector $z_t$ of dimension $G \times 1$, where $G \geq K$. Assume that $E(z_t \epsilon_t) = 0$. The variables $z_t$ are *instrumental variables.*

Consider the moment conditions

$$\begin{aligned} m_t(\theta) &= z_t \epsilon_t \\ &= z_t \left( y_t - x_t' \theta \right) \end{aligned}$$

We can arrange the instruments in the $n \times G$ matrix

$$Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_n' \end{bmatrix}$$

The average moment conditions are

$$\begin{aligned}\bar{m}_n(\theta) &= \frac{1}{n} Z'\epsilon \\ &= \frac{1}{n}(Z'y - Z'X\theta)\end{aligned}$$

- The *generalized instrumental variables* estimator is just the GMM estimator based upon these moment conditions.
- When $G = K$, we have exact identification, and it is referred to as the instrumental variables estimator.
- Given the form of the moment conditions, the general formulae for GMM lead to particular forms for the GIV estimator:

The first order conditions for GMM are $D_n W_n \bar{m}_n(\hat{\theta}) = 0$, which imply that

$$D_n W_n Z' X \hat{\theta}_{IV} = D_n W_n Z' y$$

**Exercise 65.** Verify that $D_n = -\frac{X'Z}{n}$. Remember that (assuming differentiability) identification of the GMM estimator requires that this matrix must converge to a matrix with full row rank. Can just any variable that is uncorrelated with the error be used as an instrument, or is there some other condition?

**Exercise 66.** Verify that the efficient weight matrix is $W_n = \left(\frac{Z'Z}{n}\right)^{-1}$ (up to a constant). Note, this would not hold if the errors were not i.i.d.

If we accept what is stated in these two exercises, then

$$D_n W_n Z' X \hat{\theta}_{IV} = D_n W_n Z' y$$

becomes

$$\frac{X'Z}{n} \left( \frac{Z'Z}{n} \right)^{-1} Z' X \hat{\theta}_{IV} = \frac{X'Z}{n} \left( \frac{Z'Z}{n} \right)^{-1} Z' y$$

Noting that the powers of $n$ cancel, we get

$$X'Z \left( Z'Z \right)^{-1} Z' X \hat{\theta}_{IV} = X'Z \left( Z'Z \right)^{-1} Z' y$$

or

$$\hat{\theta}_{IV} = \left( X'Z \left( Z'Z \right)^{-1} Z' X \right)^{-1} X'Z \left( Z'Z \right)^{-1} Z' y \tag{16.5}$$

Another way of arriving to the same point is to define the projection matrix $P_Z$

$$P_Z = Z(Z'Z)^{-1}Z'$$

Anything that is projected onto the space spanned by $Z$ will be uncorrelated with $\varepsilon$, by the definition of $Z$. Transforming the model with this projection matrix we get

$$P_Z y = P_Z X\beta + P_Z\varepsilon$$

or

$$y^* = X^*\theta + \varepsilon^*$$

Now we have that $\varepsilon^*$ and $X^*$ are uncorrelated, since this is simply

$$\begin{aligned}\mathcal{E}(X^{*\prime}\varepsilon^*) &= \mathcal{E}(X'P_Z'P_Z\varepsilon) \\ &= \mathcal{E}(X'P_Z\varepsilon)\end{aligned}$$

and

$$P_Z X = Z(Z'Z)^{-1}Z'X$$

is the fitted value from a regression of $X$ on $Z$. This is a linear combination of the columns of $Z$, so it must be uncorrelated with $\varepsilon$, because we initially assumed that $E(Z_t\epsilon_t) = 0$. This implies that applying OLS to the model

$$y^* = X^*\theta + \varepsilon^*$$

will lead to a consistent estimator, given a few more assumptions.

**Exercise 67.** Verify algebraically that applying OLS to the above model gives the IV estimator of equation 16.5.

With the definition of $P_Z$, we can write

$$\hat{\theta}_{IV} = (X'P_ZX)^{-1}X'P_Zy \tag{16.6}$$

from which we obtain

$$\begin{aligned}\hat{\theta}_{IV} &= (X'P_ZX)^{-1}X'P_Z(X\theta^0 + \varepsilon) \\ &= \theta^0 + (X'P_ZX)^{-1}X'P_Z\varepsilon\end{aligned}$$

so

$$\begin{aligned}\hat{\theta}_{IV} - \theta^0 &= (X'P_ZX)^{-1}X'P_Z\varepsilon \\ &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'\varepsilon\end{aligned}$$

Now we can introduce factors of $n$ to get

$$\hat{\theta}_{IV} - \theta^0 = \left(\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'X}{n}\right)\right)^{-1}\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'\varepsilon}{n}\right)$$

Assuming that each of the terms with a $n$ in the denominator satisfies a LLN, so that

- $\frac{Z'Z}{n} \xrightarrow{p} Q_{ZZ}$, a finite pd matrix
- $\frac{X'Z}{n} \xrightarrow{p} Q_{XZ}$, a finite matrix with rank $K$ $(= \text{cols}(X)\ )$. That is to say, the instruments must be correlated with the regressors. More precisely, each regressor must be correlated with at least one instrument. Otherwise, the row of $Q_{XZ}$ corresponding to that regressor would be all zeros, and thus the rank of the matrix would be less than $K$.
- $\frac{Z'\varepsilon}{n} \xrightarrow{p} 0$

then the plim of the rhs is zero. This last term has plim 0 because we started with the assumption that $Z$ and $\varepsilon$ are uncorrelated, e.g.,

$$\mathcal{E}(z_t'\varepsilon_t) = 0,$$

Given these assumptions, the IV estimator is consistent

$$\hat{\theta}_{IV} \xrightarrow{p} \theta^0.$$

Furthermore, scaling by $\sqrt{n}$, we have

$$\sqrt{n}\left(\hat{\theta}_{IV}-\theta^0\right)=\left(\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'X}{n}\right)\right)^{-1}\left(\frac{X'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'\varepsilon}{\sqrt{n}}\right) \tag{16.7}$$

Assuming that the far right term satisfies a CLT, so that

- $\frac{Z'\varepsilon}{\sqrt{n}} \xrightarrow{d} N(0, Q_{ZZ}\sigma^2)$

then we get (using some pleasing cancelations)

$$\sqrt{n}\left(\hat{\theta}_{IV}-\theta^0\right) \xrightarrow{d} N\left(0, (Q_{XZ}Q_{ZZ}^{-1}Q'_{XZ})^{-1}\sigma^2\right)$$

The adjustment for heteroscedastic or autocorrelated errors should be apparent: we just assume that $\frac{Z'\varepsilon}{\sqrt{n}} \xrightarrow{d} N(0,\Omega)$ and work out the algebra (also, see below, in the 2SLS section). And, we estimate $\Omega$ appropriately, according to the cases of heteroscedastic and/or autocorrelated errors.

The estimators for $Q_{XZ}$ and $Q_{ZZ}$ are the obvious ones. An estimator for $\sigma^2$ is

$$\widehat{\sigma^2_{IV}} = \frac{1}{n}\left(y - X\hat{\theta}_{IV}\right)'\left(y - X\hat{\theta}_{IV}\right).$$

- Note that his is computed using the real regressors, $X$, not the projected regressors, $X^*$.
- This estimator is consistent following the proof of consistency of the OLS estimator of $\sigma^2$, when the classical assumptions hold.

The formula used to estimate the variance of $\hat{\theta}_{IV}$ is

$$\hat{V}(\hat{\theta}_{IV}) = \left((X'Z)\,(Z'Z)^{-1}\,(Z'X)\right)^{-1}\widehat{\sigma^2_{IV}}$$

**The GIV estimator is**

1. Consistent
2. Asymptotically normally distributed

3. Biased in general, because even though $\mathcal{E}(X'P_Z\varepsilon) = 0$, $\mathcal{E}(X'P_ZX)^{-1}X'P_Z\varepsilon$ may not be zero, because $(X'P_ZX)^{-1}$ and $X'P_Z\varepsilon$ are not independent.

An important point is that the asymptotic distribution of $\hat{\beta}_{IV}$ depends upon $Q_{XZ}$ and $Q_{ZZ}$, and these depend upon the choice of $Z$. *The choice of instruments influences the efficiency of the estimator.*

- When we have two sets of instruments, $Z_1$ and $Z_2$ such that $Z_1 \subset Z_2$, then the IV estimator using $Z_2$ is at least as efficiently asymptotically as the estimator that used $Z_1$. More instruments leads to more asymptotically efficient estimation, in general. The same holds for GMM in general: adding moment conditions cannot cause the asymptotic variance to become larger.

- The penalty for indiscriminate use of instruments is that the small sample bias of the IV estimator rises as the number of instruments increases. The reason for this is that $P_Z X$ becomes closer and closer to $X$ itself as the number of instruments increases. Remember, OLS gets a perfect fit when $k = n$,

**Example 68.** GIV example. Recall Example 20 which deals with a dynamic model with measurement error. The model is

$$\begin{aligned} y_t^* &= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t \\ y_t &= y_t^* + \upsilon_t \end{aligned}$$

where $\epsilon_t$ and $\upsilon_t$ are independent Gaussian white noise errors. Suppose that $y_t^*$ is not observed, and instead we observe $y_t$. If we estimate the equation

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \nu_t$$

by OLS, we have seen in Example 20 that the estimator is biased and inconsistent. What about using the GIV estimator? Consider using as instruments $Z = [1\, x_t\, x_{t-1}\, x_{t-2}]$. The lags of $x_t$ are correlated with $y_{t-1}$ as long as $\beta$ is different from zero, and by assumption $x_t$ and its lags are uncorrelated with $\epsilon_t$ and $\upsilon_t$ (and thus they're also uncorrelated with $\nu_t$). Thus, these are legitimate instruments. As we have 4 instruments and 3 parameters, this is an overidentified situation. The

Julia script GMM/MeasurementErrorIV.jl does a Monte Carlo study using 1000 replications, with a sample size of 100. The results are comparable with those in Example 20. Using the GIV estimator, descriptive statistics for 1000 replications of the estimated parameters minus the true parameters are

| | mean | std | skew | kurt | min | max |
|---|---|---|---|---|---|---|
| 1 | -0.004 | 0.242 | -0.797 | 7.175 | -1.914 | 1.051 |
| 2 | -0.021 | 0.147 | -0.113 | 2.573 | -0.692 | 0.695 |
| 3 | 0.007 | 0.185 | 0.037 | 0.443 | -0.635 | 0.776 |

If you compare these with the results for the OLS estimator, you will see that the bias of the GIV estimator is much less for estimation of $\rho$. If you increase the sample size, you will see that the GIV estimator is consistent, but that the OLS estimator is not.

A histogram for $\hat{\rho} - \rho$ is in Figure 16.2. You can compare with the similar figure for the OLS estimator, Figure 8.5.

Figure 16.2: GIV estimation results for $\hat{\rho} - \rho$, dynamic model with measurement error

**Example 69.** The linear simultaneous equation model **Klein's model 1** (see Section 11.10) is estimated by GMM in the following code examples: Simeq/KleinGMM.jl and Simeq/KleinCUE.jl.

## 2SLS

We can give an alternative formulation of the GIV estimator. Let $\hat{X} = Z\left(Z'Z\right)^{-1} Z'X = P_Z X$. These are the fitted values from a regression of the regressors upon the instruments. Substitute this into eqn. 16.6, to get

$$\hat{\theta}_{IV} = (X'\hat{X})^{-1}\hat{X}'y$$

or

$$\hat{\theta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y.$$

These are equivalent. So, the GIV estimator can be obtained by

1. first regressing the regressors on the instruments, and obtaining the predicted values

2. then regressing the dependent variable on the predicted regressors.

It's clear why it's called 2SLS, no?

Eqn. 16.7 simplifies to

$$\sqrt{n}\left(\hat{\theta}_{IV} - \theta^0\right) = \left(\frac{\hat{X}'\hat{X}}{n}\right)^{-1}\left(\frac{\hat{X}'\varepsilon}{\sqrt{n}}\right) \tag{16.8}$$

From this, we can write (now allowing for possible HET/AUT)

$$\sqrt{n}\left(\hat{\theta}_{IV} - \theta^0\right) \xrightarrow{d} N\left(0, (Q_{\hat{X}}^{-1}\Omega Q_{\hat{X}}^{-1}\right)$$

where $\Omega = limV\left(\frac{\hat{X}'\varepsilon}{\sqrt{n}}\right)$.

- this can be estimated using White's or Newey-West estimators, as appropriate, or simplified further (as above) if the classical assumption regarding homoscedasticity and no autocorrelation hold.
- In either case, it is important to use the residuals $y - X\hat{\theta}_{IV}$, not $y - \hat{X}\hat{\theta}_{IV}$, to estimate $\Omega$ properly.
- Go to Section 11.10 for an example.

## 16.10 The Hansen-Sargan (or J) test

The first order conditions for minimization, using the an estimate of the optimal weighting matrix, are

$$\frac{\partial}{\partial\theta}s(\hat{\theta}) = 2\left[\frac{\partial}{\partial\theta}\bar{m}_n\left(\hat{\theta}\right)\right]\hat{\Omega}^{-1}\bar{m}_n\left(\hat{\theta}\right) \equiv 0$$

or

$$D(\hat{\theta})\hat{\Omega}^{-1}\bar{m}_n(\hat{\theta}) \equiv 0$$

Consider a Taylor expansion of $\bar{m}(\hat{\theta})$ about the true parameter value:

$$\bar{m}(\hat{\theta}) = \bar{m}_n(\theta^0) + D_n'(\theta^*)\left(\hat{\theta} - \theta^0\right) \tag{16.9}$$

where $\theta^*$ is between $\hat{\theta}$ and $\theta^0$. Multiplying by $D(\hat{\theta})\hat{\Omega}^{-1}$ we obtain

$$D(\hat{\theta})\hat{\Omega}^{-1}\bar{m}(\hat{\theta}) = D(\hat{\theta})\hat{\Omega}^{-1}\bar{m}_n(\theta^0) + D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\left(\hat{\theta} - \theta^0\right)$$

The lhs is zero, by the first order conditions for the GMM estimator, so

$$D(\hat{\theta})\hat{\Omega}^{-1}\bar{m}_n(\theta^0) = -\left[D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right]\left(\hat{\theta} - \theta^0\right)$$

or

$$\left(\hat{\theta} - \theta^0\right) = -\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1} D(\hat{\theta})\hat{\Omega}^{-1}\bar{m}_n(\theta^0)$$

Substitute the RHS into the last part of equation 16.9), and multiply by $\sqrt{n}$, to get

$$\sqrt{n}\bar{m}_n(\hat{\theta}) = \sqrt{n}\bar{m}_n(\theta^0) - \sqrt{n}D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1} D(\hat{\theta})\hat{\Omega}^{-1}\bar{m}_n(\theta^0).$$

With some factoring, this last can be written as

$$\sqrt{n}\bar{m}_n(\hat{\theta}) = \left(\hat{\Omega}^{1/2} - D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1} D(\hat{\theta})\hat{\Omega}^{-1/2}\right)\left(\sqrt{n}\hat{\Omega}^{-1/2}\bar{m}_n(\theta^0)\right)$$

(verify it by multiplying out the last expression). Also, a note: the matrix square

root of a matrix $A$ is any matrix $A^{1/2}$ such that $A = A^{1/2}A^{1/2}$. Any positive definite matrix has an invertible matrix square root.

**Example 70.** Aside on matrix square root

```
using LinearAlgebra
a = randn(3,3)
a = transpose(a)*a # make it PD, so it will have a matrix square root
s = sqrt(a)
eigvals(s) # check that it's invertible
s*s - a # verify that it's the square root
```

Next, multiply by $\hat{\Omega}^{-1/2}$ to get

$$\sqrt{n}\hat{\Omega}^{-1/2}\bar{m}_n(\hat{\theta}) = \left(I_g - \hat{\Omega}^{-1/2}D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1}D(\hat{\theta})\hat{\Omega}^{-1/2}\right)\left(\sqrt{n}\hat{\Omega}^{-1/2}\bar{m}_n(\theta^0)\right) \equiv \tag{16.10}$$

- the big matrix $P = I_g - \hat{\Omega}^{-1/2}D_n'(\theta^*)\left(D(\hat{\theta})\hat{\Omega}^{-1}D(\theta^*)'\right)^{-1}D(\hat{\theta})\hat{\Omega}^{-1/2}$ converges in probability to $P_\infty = I_g - \Omega_\infty^{-1/2}D_\infty'\left(D_\infty\Omega_\infty^{-1}D_\infty'\right)^{-1}D_\infty\Omega_\infty^{-1/2}$.
- One can easily verify that $P_\infty$ is idempotent and has rank $g - K$, (recall that the rank of an idempotent matrix is equal to its trace).
- The far right term, $X = \sqrt{n}\hat{\Omega}^{-1/2}\bar{m}_n(\theta^0)$ converges to a $g$ vector of i.i.d. standard normal random variables, by the LLN and the Slutsky theorem. This is because $\sqrt{n}\bar{m}(\theta^0) \xrightarrow{d} N(0, \Omega_\infty)$ (from 16.3) and we're multiplying by something that converges to $\Omega_\infty^{-1/2}$. Note that $\Omega_\infty^{-1/2}\Omega_\infty\Omega_\infty^{-1/2} = I_g$.
- Thus, $X'PX \xrightarrow{d} \chi^2(d)$, by the Continuous Mapping Theorem (Gallant

(1997), Theorem 4.7) . This is because, asymptotically, it is a quadratic form of standard normal variables, weighted by an idempotent matrix.

- So, the inner product of the r.h.s. of eq. 16.10 has an asymptotic chi-square distribution. The inner product using the l.h.s. must also have the same distribution, so we finally get

$$\left(\sqrt{n}\hat{\Omega}^{-1/2}\bar{m}_n(\hat{\theta})\right)'\left(\sqrt{n}\hat{\Omega}^{-1/2}\bar{m}_n(\hat{\theta})\right) = n\bar{m}_n(\hat{\theta})'\hat{\Omega}^{-1}\bar{m}_n(\hat{\theta}) \xrightarrow{d} \chi^2(g-K)$$

***Hansen-Sargan*** test*:* Supposing that the moment conditions actually have expectation zero at the true parameter value, and that we are using an estimate of the efficient weight matrix, then

$$n \cdot s_n(\hat{\theta}) \xrightarrow{d} \chi^2(g - K).$$

- This is a convenient test since we just multiply the optimized value of the objective function by $n$, and compare with a $\chi^2(g - K)$ critical value. The test is a general test of whether or not the moments used to estimate are correctly specified.

- This won't work when the estimator is just identified. The f.o.c. are

$$D_\theta s_n(\theta) = D\hat{\Omega}^{-1}\bar{m}_n(\hat{\theta}) \equiv 0.$$

But with exact identification, both $D$ and $\hat{\Omega}$ are square and invertible (at

least asymptotically, assuming that asymptotic normality hold), so

$$\bar{m}_n(\hat{\theta}) \equiv 0.$$

So the moment conditions are zero *regardless* of the weighting matrix used. As such, we might as well use an identity matrix and save trouble. Also $s_n(\hat{\theta}) = 0$, so the test breaks down.

- This sort of test often over-rejects in finite samples. One should be cautious in rejecting a model when this test rejects.

- This test goes by several names: Hansen test, Sargan test, Hansen-Sargan test, J test. I call it the GMM criterion test. An old name for GMM estimation is "minimum chi-square" estimation. This makes sense: the criterion function at the estimate (which makes the criterion as small as possible), scaled by $n$, has a $\chi^2$ distribution. GMM minimizes it.

The Julia script GMM/SpecTest.jl does a Monte Carlo study of the Hansen-Sargan

test, for same the dynamic model with measurement error as was discussed in Examples 20 and 68, which did GIV estimation, and shows that it over-rejects a correctly specified model, in this case. For example, if the significance level is set to 10%, the test rejects about 16% of the time. This is a common result for this test. Results from a run are:

```
rejection frequencies, nominal 10% 5% and 1%:
       10%        5%        1%
     0.155     0.088     0.020
```

## 16.11 Other estimators interpreted as GMM estimators

### Maximum likelihood

In the introduction we argued that ML will in general be more efficient than GMM since ML implicitly uses all of the moments of the distribution while GMM uses a limited number of moments. Actually, a distribution with $P$ parameters can be uniquely characterized by $P$ moment conditions. However, some sets of $P$ moment conditions may contain more information than others, since the moment conditions could be highly correlated. A GMM estimator that chose an optimal set of $P$ moment conditions would be fully efficient. The optimal moment conditions are simply the scores of the ML estimator.

In the chapter on maximum likelihood, we saw in eqn. 15.3 that the first

derivative of the average log likelihood function is

$$\frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f(y_t|x_x,\theta)$$

and that the ML estimator is obtained by setting this to zero, and solving. We also saw in eqn. 15.4 that the expectation of the score vector is zero, when evaluated at the true parameter values. Thus, the score vector satisfies the requirement to serve as moment conditions. Set

$$m_t(\theta) \equiv D_\theta \ln f(y_t|x_t,\theta)$$

- Recall that the score contributions are both conditionally and unconditionally uncorrelated. Conditional uncorrelation follows from the fact that $m_{t-s}$ if is a function of lagged endogenous variables, then they are included in $x_t$, which is what we are conditioning on at time $t$. Unconditional uncorrelation follows from the fact that conditional uncorrelation hold regardless of the realization of $y_{t-1}$, so marginalizing with respect to $Y_{t-1}$ preserves

uncorrelation (see the section on ML estimation, above).

- The fact that the scores are serially uncorrelated implies that $\Omega$ can be estimated by the estimator of the $0^{th}$ autocovariance of the moment conditions:

$$\widehat{\Omega} = 1/n \sum_{t=1}^{n} m_t(\hat{\theta}) m_t(\hat{\theta})' = 1/n \sum_{t=1}^{n} \left[ D_\theta \ln f(y_t|x_t, \hat{\theta}) \right] \left[ D_\theta \ln f(y_t|x_t, \hat{\theta}) \right]'$$

  - note that this is the estimator of the information matrix, from ML
  - There is no need for a Newey-West style estimator, the heteroscedastic-consistent estimator of White is sufficient.
  - Also, the fact that the scores of ML are uncorrelated suggests a means of testing the correct specification of the model: see if the fitted scores $(m_t(\hat{\theta})$ show evidence of serial correlation. If they do, the correctness of the specification of the model is subject to doubt.

## OLS as a GMM estimator - the Nerlove model again

**Example 71.** Matlab/Octave code for GMM for Nerlove model. Examine and run TwoStepGMM.m , which illustrates how to do two step GMM for the Nerlove data. Note that the GMM results are the same as what you get estimating by OLS.

The simple Nerlove model can be estimated using GMM, as we've seen. So, OLS is a special case of GMM.

## 16.12 The Hausman Test

This section discusses the Hausman test (Hausman (1978)).

Consider the simple linear regression model $y_t = x_t'\beta + \epsilon_t$. We assume that the functional form and the choice of regressors is correct, but that the some of the regressors may be correlated with the error term, which as you know will produce inconsistency of $\hat{\beta}$. For example, this will be a problem if

- if some regressors are endogenous
- some regressors are measured with error
- some relevant regressors are omitted (equivalent to imposing false restrictions)
- lagged values of the dependent variable are used as regressors and $\epsilon_t$ is autocorrelated.

To illustrate, the Julia program OLSvsIV.jl performs a Monte Carlo experiment where errors are correlated with regressors, and estimation is by OLS and IV.

- The true value of the slope coefficient used to generate the data is $\beta = 2$.
- Figure 16.3 shows that the OLS estimator is quite biased and that the IV estimator is on average much closer to the true value.
- If you play with the program, increasing the sample size, you can see evidence that the OLS estimator is asymptotically biased, while the IV estimator is consistent.
- You can also play with the covariances of the instrument and regressor, and the covariance of the regressor and the error.

Figure 16.3: OLS and IV ($n = 1000$,cov(X,$\epsilon$) $= 0.5$,cov(W,$\epsilon$) $= 0.5$)

We have seen that inconsistent and the consistent estimators converge to different probability limits. This is the idea behind the Hausman test - a pair of consistent estimators converge to the same probability limit, while if one is consistent and the other is not they converge to different limits. If we accept that one is consistent (*e.g.*, the IV estimator), but we are doubting if the other is consistent (*e.g.*, the OLS estimator), we might try to check if the difference between the

estimators is significantly different from zero.

- If we're doubting about the consistency of OLS (or QML, *etc.*), why should we be interested in testing - why not just use the IV estimator? Because the OLS estimator is *more efficient* when the regressors are exogenous and the other classical assumptions (including normality of the errors) hold.

- Play with the above script to convince yourself of this point: make exogeneity hold, and compare the variances of OLS and IV

- When we have a more efficient estimator that relies on stronger assumptions (such as exogeneity) than the IV estimator, we might prefer to use it, unless we have evidence that the assumptions are false.

  - When exogeneity holds, OLS is better.
  - When exogeneity is sufficiently false, IV is better.
  - How do we choose, if we're not sure about whether or not exogeneity holds?

So, let's consider the covariance between the MLE estimator $\hat{\theta}$ (or any other fully efficient estimator) and some other CAN estimator, say $\tilde{\theta}$. Now, let's recall some results from MLE. Equation 15.6 implies:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} -\mathcal{J}_\infty(\theta_0)^{-1}\sqrt{n}g(\theta_0).$$

Equation 15.11 is

$$\mathcal{J}_\infty(\theta) = -\mathcal{I}_\infty(\theta).$$

Combining these two equations, we get

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{I}_\infty(\theta_0)^{-1}\sqrt{n}g(\theta_0).$$

Also, equation 15.14 tells us that the asymptotic covariance between any CAN estimator and the MLE score vector is

$$V_\infty \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n} g(\theta) \end{bmatrix} = \begin{bmatrix} V_\infty(\tilde{\theta}) & I_K \\ I_K & \mathcal{I}_\infty(\theta) \end{bmatrix}.$$

These results imply that

$$\begin{bmatrix} I_K & 0_K \\ 0_K & I_\infty(\theta)^{-1} \end{bmatrix} \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n} g(\theta) \end{bmatrix} \to^d \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n}\left(\hat{\theta} - \theta\right) \end{bmatrix}.$$

The asymptotic covariance of the RHS is the same as the asymptotic variance of the LHS, so

$$\begin{aligned} V_\infty \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n}\left(\hat{\theta} - \theta\right) \end{bmatrix} &= \begin{bmatrix} I_K & 0_K \\ 0_K & I_\infty(\theta)^{-1} \end{bmatrix} \begin{bmatrix} V_\infty(\tilde{\theta}) & I_K \\ I_K & \mathcal{I}_\infty(\theta) \end{bmatrix} \begin{bmatrix} I_K & 0_K \\ 0_K & I_\infty(\theta)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} V_\infty(\tilde{\theta}) & I_\infty(\theta)^{-1} \\ I_\infty(\theta)^{-1} & I_\infty(\theta)^{-1} \end{bmatrix}, \end{aligned}$$

which, for clarity in what follows, we might write as (note to self for lectures: the

2,2 element has changed)

$$V_\infty \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta\right) \\ \sqrt{n}\left(\hat{\theta} - \theta\right) \end{bmatrix} = \begin{bmatrix} V_\infty(\tilde{\theta}) & I_\infty(\theta)^{-1} \\ I_\infty(\theta)^{-1} & V_\infty(\hat{\theta}) \end{bmatrix}.$$

The main point of interest here are the off diagonal entries. So, the asymptotic covariance between the MLE and any other CAN estimator is equal to the MLE asymptotic variance (the inverse of the information matrix).

Now, suppose we wish to test whether the the two estimators are in fact both converging to $\theta_0$, versus the alternative hypothesis that the "MLE" estimator is not in fact consistent (the consistency of $\tilde{\theta}$ is a maintained hypothesis). Under the null hypothesis that they are equal, we have

$$\begin{bmatrix} I_K & -I_K \end{bmatrix} \begin{bmatrix} \sqrt{n}\left(\tilde{\theta} - \theta_0\right) \\ \sqrt{n}\left(\hat{\theta} - \theta_0\right) \end{bmatrix} = \sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right),$$

will be asymptotically normally distributed as (work out on blackboard)

$$\sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right) \xrightarrow{d} N\left(0, V_\infty(\tilde{\theta}) - V_\infty(\hat{\theta})\right).$$

So,

$$n\left(\tilde{\theta} - \hat{\theta}\right)' \left(V_\infty(\tilde{\theta}) - V_\infty(\hat{\theta})\right)^{-1} \left(\tilde{\theta} - \hat{\theta}\right) \xrightarrow{d} \chi^2(\rho),$$

where $\rho$ is the rank of the difference of the asymptotic variances. A statistic that

has the same asymptotic distribution is

$$\left(\tilde{\theta}-\hat{\theta}\right)'\left(\hat{V}(\tilde{\theta})-\hat{V}(\hat{\theta})\right)^{-1}\left(\tilde{\theta}-\hat{\theta}\right)\xrightarrow{d}\chi^2(\rho),$$

where the $n$ has disappeared because we are using the small sample estimates of the covariances of the two estimators. This is the Hausman test statistic, in its original form. The reason that this test has power under the alternative hypothesis is that in that case the "MLE" estimator will not be consistent, and will converge to $\theta_A$, say, where $\theta_A \neq \theta_0$. Then the mean of the asymptotic distribution of vector $\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right)$ will be $\theta_0-\theta_A$, a non-zero vector, so the test statistic will eventually reject, regardless of how small a significance level is used.

- The quantity $V_\infty(\tilde{\theta})-V_\infty(\hat{\theta})$ may be a singular matrix, in which case the inverse in $\left(V_\infty(\tilde{\theta})-V_\infty(\hat{\theta})\right)^{-1}$ must be replaced with a generalized inverse. This can occur when the two estimators are defined using some common moment conditions, which can introduce some linear dependencies between the estimators.

- When this is the case, the rank, $\rho$, of the difference of the asymptotic variances will be less than the dimension of the matrices, and it may be difficult to determine what the true rank is. If the true rank is lower than what is taken to be true, the test will be biased against rejection of the null hypothesis. The null is that both estimators are consistent. Failure to reject when this hypothesis is false would cause us to use an inconsistent estimator: not a desirable outcome! The contrary holds if we underestimate the rank.

- A solution to this problem is to use a rank 1 test, by comparing only a single coefficient. For example, if a variable is suspected of possibly being endogenous, that variable's coefficients may be compared.

- Note: if the test is based on a sub-vector of the entire parameter vector of the MLE, it is possible that the inconsistency of the MLE will not show up in the portion of the vector that has been used. If this is the case, the test may not have power to detect the inconsistency. This may occur, for example, when the consistent but inefficient estimator is not identified for all

Figure 16.4: Incorrect rank and the Hausman test

the parameters of the model, so that we estimate only some of the parameters using the inefficient estimator, and the test does not include the others.

- This simple formula only holds when the estimator that is being tested for consistency is *fully* efficient under the null hypothesis. This means that it must be a ML estimator or a fully efficient estimator that has the same asymptotic distribution as the ML estimator. This is quite restrictive since modern estimators such as GMM, QML, or even OLS with heteroscedastic consistent standard errors are not in general fully efficient.

Following up on this last point, let's think of two not necessarily efficient estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, where one is assumed to be consistent, but the other may not be. We assume for expositional simplicity that both $\hat{\theta}_1$ and $\hat{\theta}_2$ belong to the same parameter space, and that each estimator can be expressed as generalized method of moments (GMM) estimator. The estimators are defined (suppressing

the dependence upon data) by

$$\hat{\theta}_i \;=\; \arg\min_{\theta_i \in \Theta} \bar{m}_i(\theta_i)' \, W_i \, \bar{m}_i(\theta_i)$$

where $\bar{m}_i(\theta_i)$ is a $g_i \times 1$ vector of moment conditions, and $W_i$ is a $g_i \times g_i$ positive definite weighting matrix, $i = 1, 2$. Consider the omnibus GMM estimator

$$\left(\hat{\theta}_1, \hat{\theta}_2\right) = \arg\min_{\Theta \times \Theta} \left[\, \bar{m}_1(\theta_1)' \;\; \bar{m}_2(\theta_2)' \,\right] \begin{bmatrix} W_1 & \mathbf{0}_{(g_1 \times g_2)} \\ \mathbf{0}_{(g_2 \times g_1)} & W_2 \end{bmatrix} \begin{bmatrix} \bar{m}_1(\theta_1) \\ \bar{m}_2(\theta_2) \end{bmatrix}. \tag{16.11}$$

The minimizer just gives the two separate GMM estimators.

Suppose that the asymptotic covariance of the omnibus moment vector is

$$\begin{aligned} \Sigma \;&=\; \lim_{n \to \infty} Var \left\{ \sqrt{n} \begin{bmatrix} \bar{m}_1(\theta_1) \\ \bar{m}_2(\theta_2) \end{bmatrix} \right\} \\ &\equiv \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \cdot & \Sigma_2 \end{pmatrix}. \end{aligned} \tag{16.12}$$

The standard Hausman test is equivalent to a Wald test of the equality of $\theta_1$ and $\theta_2$ (or subvectors of the two) applied to the omnibus GMM estimator, but with the covariance of the moment conditions estimated as

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma_1} & \mathbf{0}_{(g_1 \times g_2)} \\ \mathbf{0}_{(g_2 \times g_1)} & \widehat{\Sigma_2} \end{pmatrix}.$$

While this is clearly an inconsistent estimator in general, the omitted $\Sigma_{12}$ term cancels out of the test statistic when one of the estimators is asymptotically efficient, as we have seen above, and thus it need not be estimated.

The general solution when neither of the estimators is efficient is clear: the entire $\Sigma$ matrix must be estimated consistently, since the $\Sigma_{12}$ term will not cancel out. Methods for consistently estimating the asymptotic covariance of a vector of moment conditions are well-known, *e.g.*, the Newey-West estimator discussed previously. The Hausman test using a proper estimator of the overall covariance matrix will now have an asymptotic $\chi^2$ distribution when neither estimator is

efficient.

However, the test suffers from a loss of power due to the fact that the omnibus GMM estimator of equation 16.11 is defined using an inefficient weight matrix. A new test can be defined by using an alternative omnibus GMM estimator

$$\left(\hat{\theta}_1, \hat{\theta}_2\right) = \arg\min_{\Theta\times\Theta} \left[\, \bar{m}_1(\theta_1)' \;\; \bar{m}_2(\theta_2)' \,\right] \left(\widetilde{\Sigma}\right)^{-1} \begin{bmatrix} \bar{m}_1(\theta_1) \\ \bar{m}_2(\theta_2) \end{bmatrix}, \tag{16.13}$$

where $\widetilde{\Sigma}$ is a consistent estimator of the overall covariance matrix $\Sigma$ of equation 16.12. By standard arguments, this is a more efficient estimator than that defined by equation 16.11, so the Wald test using this alternative is more powerful. See my article in *Applied Economics*, 2004, for more details, including simulation results. The Octave script hausman.m calculates the Wald test corresponding to the efficient joint GMM estimator (the "H2" test in my paper), for a simple linear model, and compares to the standard Hausman test.

## 16.13 More moments are better (?)

A similar framework allows us to show that an overidentified GMM estimator will be asymptotically efficient relative to a GMM estimator that only uses a subset of moment conditions that still maintain identification. Let $\bar{m}_n(\theta)$ be partitioned as

$$\bar{m}_n(\theta) = \begin{bmatrix} \bar{m}_n^1(\theta) \\ \bar{m}_n^2(\theta) \end{bmatrix}$$

where $\bar{m}_n^1(\theta)$ is a subset of moment conditions that still identifies the parameters. Suppose that the limiting covariance of $\sqrt{n}\bar{m}_n(\theta)$ is

$$\Omega_\infty = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}' & \Omega_{22} \end{bmatrix}.$$

The efficient weight matrix is something that converges to the inverse of this. The GMM estimator that uses only the subset of moment conditions, $\bar{m}_n^1(\theta)$, is

equivalent to a GMM estimator that uses all of the moment conditions, but with the weight matrix being

$$W = \begin{bmatrix} \Omega_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

where the zeros are appropriately sized matrices of zeros. This is not the efficient weight matrix for the full set of moment conditions, so this GMM estimator will not be efficient with respect to the efficiently weighted GMM estimator that uses all of the moment conditions.

- so, comparing two GMM estimators, A and B, where B uses a subset of the moment conditions that A uses, and both are identified, A will be asymptotically efficient with respect to B.

- however, the efficiency gain may be zero, or very little (recall that the definition of "efficient with respect to" uses positive semi-definiteness).

- the small sample bias of the GMM estimator tends to increase as the number of moment conditions increase (Donald and Newey (2001); Donald et al. (2009))

- so, estimator A may have worse small sample performance than version B, in spite of being relatively asymptotically efficient

- experience has shown that using only reasonably informative moments, with a small degree of overidentification, is usually better than the "kitchen sink" approach. See the literature on weak instruments, etc.

## 16.14 Examples

### Linear IV: The Card returns to schooling data

Card (1993) presents an analysis of returns to schooling using the data from the National Longitudinal Survey of Young Men, for those interviewed in 1976. Card presents OLS and instrumental variables estimates for a number of specifications, using college proximity as an instrument for years of education, and age as an instrument for experience. Here, we work with the simple model from column (1) of Card's Table 2. The model is

$$\begin{aligned}\ln W = \beta_0 + \beta_{EDUC} EDUC + \beta_X EXP + \beta_{EXP^2} \frac{EXP^2}{100} \\ + \beta_{BLACK} BLACK + \beta_{SMSA} SMSA + \beta_{SOUTH} SOUTH + \epsilon\end{aligned}$$

- the dependent variable $\ln W$ is log hourly earnings (in cents)
- the regressors are years of education (EDUC), experience (EXP), experience

squared divided by 100, a black indicator (BLACK), a metropolitan area indicator (SMSA), and a South indicator (SOUTH).

- We explore estimation treating all variables as exogenous, or treating education and experience as endogenous, and the others as exogenous.
- If uncontrolled for factors that affect wages also affect education, then education will be endogenous.
- EXPER is defined as EXPER = AGE-EDUC-6. So, if EDUC is endogenous, so is EXPER.
- Instruments:
  - we use proximity to an accredited four year college (NEARC4) as an instrumental variable that should be correlated with EDUC
  - We use AGE as an instrument for EXPER, and AGE squared for EXP squared.

- The Card data set is provided with the Wooldridge data set for GRETL, see the GRETL web page. A version prepared for the model used here is card.gdt .
- The data is also here: cooked.csv , ready for use with Julia.
- Card.jl does OLS, GMM-CUE, and 2 step GMM.
- The effect of an additional year of education on wages is about 7%, according to OLS, and about 13%, according to IV.

# Application: Hansen-Singleton, 1982

**Readings:** (Hansen and Singleton, 1982); (Tauchen, 1986)

Though GMM estimation has many applications, application to rational expectations models is elegant, since theory directly suggests the moment conditions. Hansen and Singleton's 1982 paper is also a classic worth studying in itself. Though I strongly recommend reading the paper, I'll use a simplified model with notation similar to Hamilton's. The literature on estimation of these models has grown a lot since these early papers. After work like the cited papers, people moved to ML estimation of linearized models, using Kalman filtering. Current methods are usually Bayesian, and involve sophisticated filtering methods to compute the likelihood function for nonlinear models with non-normal shocks. There is a lot of interesting stuff that is beyond the scope of this course. I have done some work using simulation-based estimation methods applied to such models. The methods explained in this section are intended to provide an example of GMM estimation. They are not the state of the art for estimation of such models.

We assume a representative consumer maximizes expected discounted utility over an infinite horizon. Expectations are rational, and the agent has full information (is fully aware of the history of the world up to the current time period - how's that for an assumption!). Utility is temporally additive, and the expected utility hypothesis holds. The future consumption stream is the stochastic sequence $\{c_t\}_{t=0}^{\infty}$. The objective function at time $t$ is the discounted expected utility

$$\sum_{s=0}^{\infty} \beta^s \mathcal{E}\left(u(c_{t+s})|I_t\right). \tag{16.14}$$

- The parameter $\beta$ is between 0 and 1, and reflects discounting.
- $I_t$ is the *information set* at time $t$, and includes the all realizations of all random variables that are known at time $t$.
- The choice variable is $c_t$ - current consumption, which is constrained to be less than or equal to current wealth $w_t$.

- Suppose the consumer can invest in a risky asset. A dollar invested in the asset yields a gross return

$$(1 + r_{t+1}) = \frac{p_{t+1} + d_{t+1}}{p_t}$$

  where $p_t$ is the price and $d_t$ is the dividend in period $t$. Thus, $r_{t+1}$ is the net return on a dollar invested in period $t$.

- The price of $c_t$ is normalized to 1.

- Current wealth $w_t = (1+r_t)i_{t-1}$, where $i_{t-1}$ is investment in period $t-1$. At time $t$, the consumer observes $r_t$, which determines wealth. So the problem is to allocate current wealth between current consumption and investment to finance future consumption: $w_t = c_t + i_t$. The choice variable can be taken as $c_t$,which determines $i_t$.

- Future net rates of return $r_{t+s}, s > 0$ are *not known* in period $t$: the asset is risky.

- The information set $I_t$ includes $r_t$,$w_t$, $p_r$,$d_t$ and $c_{t-1}$ and $i_{t-1}$, as well as all lags of these variables.

A partial set of necessary conditions for utility maximization have the form:

$$u'(c_t) = \beta \mathcal{E} \left\{ (1 + r_{t+1}) \, u'(c_{t+1}) | I_t \right\} . \tag{16.15}$$

To see that the condition is necessary, suppose that the lhs $<$ rhs. Then by reducing current consumption marginally would cause equation 16.14 to drop by $u'(c_t)$, since there is no discounting of the current period. At the same time, the marginal reduction in consumption finances investment, which has gross return $(1 + r_{t+1})$ , which could finance consumption in period $t + 1$. This increase in consumption would cause the objective function to increase by $\beta \mathcal{E} \left\{ (1 + r_{t+1}) \, u'(c_{t+1}) | I_t \right\}$ . Therefore, unless the condition holds, the expected discounted utility function is not maximized.

- To use this we need to choose the functional form of utility. A constant relative risk aversion (CRRA) form is

$$u(c_t) = \frac{c_t^{1-\gamma} - 1}{1 - \gamma}$$

where $\gamma$ is the coefficient of relative risk aversion. With this form,

$$u'(c_t) = c_t^{-\gamma}$$

so the foc are

$$c_t^{-\gamma} = \beta \mathcal{E} \left\{ (1 + r_{t+1}) \, c_{t+1}^{-\gamma} | I_t \right\}$$

While it is true that

$$\mathcal{E}\left(c_t^{-\gamma} - \beta\left\{\left(1 + r_{t+1}\right) c_{t+1}^{-\gamma}\right\}\right) | I_t = 0$$

so that we could use this to define moment conditions, it is unlikely that $c_t$ is stationary, even though it is in real terms, and our theory requires stationarity. To solve this, divide though by $c_t^{-\gamma}$

$$\mathcal{E}\left(1\text{-}\beta\left\{\left(1 + r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right\}\right) | I_t = 0$$

(note that $c_t$ can be passed though the conditional expectation since $c_t$ is chosen based only upon information available in time $t$). That is to say, $c_t$ is in the information set $I_t$.

Now

$$1\text{-}\beta\left\{\left(1 + r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right\}$$

is analogous to $h_t(\theta)$ defined above: it's a scalar moment condition that has con-

ditional expectation equal to zero. To get a vector of moment conditions we need some instruments. Suppose that $\mathbf{z}_t$ is a vector of variables drawn from the information set $I_t$. We can use the necessary conditions to form the expressions

$$\left[1 - \beta\left(1 + r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right]\mathbf{z}_t \equiv m_t(\theta)$$

- $\theta$ represents $\beta$ and $\gamma$.
- Therefore, the above expression may be interpreted as a moment condition which can be used for GMM estimation of the parameters $\theta^0$.

Note that at time $t$, $m_{t-s}$ has been observed, and is therefore an element of the information set. By rational expectations, the autocovariances of the moment conditions other than $\Gamma_0$ should be zero. The optimal weighting matrix is therefore the inverse of the variance of the moment conditions:

$$\Omega_\infty = \lim E\left[n\bar{m}(\theta^0)\bar{m}(\theta^0)'\right]$$

which can be consistently estimated by

$$\hat{\Omega} = 1/n \sum_{t=1}^{n} m_t(\hat{\theta})m_t(\hat{\theta})'$$

As before, this estimate depends on an initial consistent estimate of $\theta$, which can be obtained by setting the weighting matrix $W$ arbitrarily (to an identity matrix, for example). After obtaining $\hat{\theta}$, we then minimize

$$s(\theta) = \bar{m}(\theta)'\hat{\Omega}^{-1}\bar{m}(\theta).$$

This process can be iterated, e.g., use the new estimate to re-estimate $\Omega$, use this to estimate $\theta^0$, and repeat until the estimates don't change.

- In principle, we could use a very large number of moment conditions in estimation, because we can use current period exogenous variables, and lags of endogenous and exogenous variables. Since use of more moment conditions will lead to a more (asymptotically) efficient estimator, one might be tempted to use many instrumental variables. We will do a computer lab that will show that this may not be a good idea with finite samples. This issue has been studied using Monte Carlos (Tauchen, *JBES*, 1986). The reason for poor performance when using many instruments is that the estimate of $\Omega$ becomes very imprecise.

- Empirical papers that use this approach often have serious problems in obtaining precise estimates of the parameters, and identification can be problematic. Note that we are basing everything on a single partial first order condition. Probably this f.o.c. is simply not informative enough.

## Empirical example of a portfolio model

The Julia program portfolio.jl performs GMM estimation of a portfolio model of the sort presented in this section, using the data file tauchen.data. The columns of this data file are $c$, $p$, and $d$ in that order. There are 95 observations (source: Tauchen (1986)). As instruments we use lags of $c$ and $r$, as well as a constant.

For a single lag the estimation results are

```
julia> include("portfolio.jl")
*****************************************************
Two step GMM estimation of portfolio model
GMM Estimation Results    BFGS convergence: Normal
Observations: 94
Hansen-Sargan statistic: 7.81392
Hansen-Sargan p-value: 0.05002

                 estimate      st. err       t-stat      p-value
        beta      0.91712      0.00933     98.34885      0.00000
       gamma      0.46783      0.32089      1.45789      0.14828

*****************************************************
*****************************************************
```

```
CUE GMM estimation of portfolio model
GMM Estimation Results    BFGS convergence: Normal
Observations: 94
Hansen-Sargan statistic: 8.02361
Hansen-Sargan p-value: 0.04553

                 estimate     st. err      t-stat     p-value
        beta      0.91677     0.00781   117.31563     0.00000
       gamma      0.49275     0.26499     1.85949     0.06615

*****************************************************************
```

For two lags the estimation results are

```
julia> include("portfolio.jl")
WARNING: imported binding for names overwritten in module Main
*****************************************************
Two step GMM estimation of portfolio model
GMM Estimation Results    BFGS convergence: Normal
Observations: 93
Hansen-Sargan statistic: 10.44107
Hansen-Sargan p-value: 0.16493

                 estimate     st. err      t-stat     p-value
        beta      0.86394     0.02155    40.09031     0.00000
       gamma     -2.20139     0.42680    -5.15785     0.00000

*****************************************************
```

```
*****************************************************
CUE GMM estimation of portfolio model
GMM Estimation Results    BFGS convergence: Normal
Observations: 93
Hansen-Sargan statistic: 24.33766
Hansen-Sargan p-value: 0.00099

                 estimate     st. err      t-stat     p-value
        beta      0.91947     0.01489    61.73498     0.00000
       gamma      1.03452     0.70396     1.46957     0.14513

*****************************************************
```

Pretty clearly, the results are sensitive to the choice of instruments. Also, if you examine the objective function values, it seems unlikely that the global minimum was found in all cases, probably multiple start values or global minimization are needed. Maybe there is some problem here: poor instruments, or possibly a conditional moment that is not very informative. Moment conditions formed from Euler conditions sometimes do not identify the parameter of a model. See Hansen et al. (1996). I believe that this is the case here, though I haven't checked it carefully.

The Octave/Matlab program HallGMM.m estimates a very similar model, following Chapter 23 of the Gretl Users Guide. I encourage you to verify that you can obtain the same results using Gretl and Octave.

**Exercise 72.** Translate the HallGMM.m code to run on Julia.

# GMM estimation of the DSGE example

Here we return to the DSGE model of Chapter 14, and derive some moment conditions that can be used for estimation.

- this example shows how moment conditions can be derived from the structure of a model
- it will also illustrate the care that is sometimes needed when doing numeric optimization

**MRS and wage**

From the first order conditions of the model (see eq. 14.2 and following), we have

$$\begin{aligned} w_t &= \psi \eta_t c_t^{\gamma} \\ \eta_t &= \frac{w_t}{\psi c_t^{\gamma}} \\ \ln \eta_t &= \ln w_t - \ln \psi - \gamma \ln c_t \end{aligned}$$

The real values of this shock $\eta_t$ are not observed, but, given a guess for the parameters $\psi$ and $\gamma$, and the data, the right hand side of the above equation can be calculated. Also, we have

$$\ln \eta_t = \rho \ln \eta_{t-1} + \sigma_\eta \epsilon_t.$$

So, we can regress the calculated $\ln \eta_t$ on their lags. The FOC for the OLS

estimator set the mean of

$$u_t = \ln \eta_{t-1}[\ln \eta_t - \rho_\eta \ln \eta_{t-1}]$$

to zero. At the true parameter values, this expression has mean zero, so it can be used to define a moment condition. We also have that

$$E\left(u_t^2 - \sigma_\eta^2\right) = 0$$

at the true parameters, so this gives us a second moment condition. These two moment conditions are informative for all of the parameters that enter into their definitions: $\gamma, \rho_\eta, \sigma_\eta$ and $\alpha$,$\beta$,$\delta$ and $\bar{n}$ (because $\psi$ depends on them, see above). We're only missing $\rho_z$ and $\sigma_z$.

**Euler equation**

The Euler equation is

$$c_t^{-\gamma} = E\left(\beta \cdot c_{t+1}^{-\gamma}\left[1 + MPK_{t+1} - \delta\right]\right),$$

where the expectation is taken conditional on the information available in period $t$ (which include variables indexed $t$ and before). But $r = MPK$, so

$$E\left(\beta \cdot c_{t+1}^{-\gamma}\left[1 + r_{t+1} - \delta\right]\right) - c_t^{-\gamma} = 0$$

Thus,

$$v_t = \beta \cdot c_{t+1}^{-\gamma}\left[1 + r_{t+1} - \delta\right] - c_t^{-\gamma} \tag{16.16}$$

has mean zero, conditional on information available in period $t$. Moment conditions that use this error should be informative for $\gamma$,$\delta$ and $\beta$.

**Estimation by GMM**

A sample of size 160, generated from the model at the true parameter values, above, is at dsgedata.txt. The columns are y, c, n, r, w. Have another look at the data, if you like: Figure 14.1.

A Julia function to compute the moment conditions discussed above, and others, is at DSGEmoments.jl.

The script DoGMM.jl implements CUE-GMM estimation of the model, using the selected moment conditions, using simulated annealing to do the minimization.

The final estimates, standard errors, and 95% CI bounds, are

| estimate | std. err. | CI lower | CI upper |
|---|---|---|---|
| 0.99032 | 0.00071 | 0.98892 | 0.99172 |
| 2.01726 | 0.24349 | 1.54002 | 2.49449 |
| 0.90145 | 0.02800 | 0.84656 | 0.95633 |
| 0.01903 | 0.00834 | 0.00267 | 0.03539 |
| 0.71185 | 0.24502 | 0.23161 | 1.19208 |
| 0.01031 | 0.00726 | -0.00392 | 0.02453 |
| 0.33476 | 0.00165 | 0.33152 | 0.33800 |

- the point estimates are very good, but the standard errors of $\gamma$ and $\rho_2$ are quite high. Some parameters are precisely estimated, but others, especially $\rho_\eta$ have fairly large standard errors. Perhaps better moments could be found to better identify these parameters.

- The true parameters are inside the 95% confidence intervals, in all cases. (See Table 14.4).

- care is needed to obtain a real global minimum. The attempt to use ordinary gradient-based minimization fails, as you will see if you run the code. A person who tried these methods might conclude that the moments don't identify the parameters well, but this is not the case: it is possible to obtain good results using GMM.

- Simulated annealing, on the other hand, converges to the same value in repeated runs. It is possible that on a given run, a different outcome might be obtained, if the cooling rate is too rapid, but I have yet to see this with

the current setup. SA requires many function evaluations, about 30000 with the setting in the example code. However, it doesn't take too long, only about 11 seconds. This doesn't seem like too much time to get a reliable answer.

- The take home conclusions here are:
  - that GMM can give reliable estimates, but perhaps we should try to improve estimation of some parameters by looking for better moment conditions.
  - multiple local minima and irregular objective functions really can be a problem, even with simple models like this one. Imagine what would happen with a large scale DSGE model! For similar problems with a model that is much more simple, see Hansen et al. (1996).
  - the difficulties with extremum estimation may motivate other computational methods, such as using a Bayesian approach to compute classical

estimators as was proposed by Chernozhukov and Hong (2003). We will return to this idea in Chapter 18.

## 16.15 Practical Summary

The practical summary for the Chapter is here, but the previous examples, especially the Card example, are important, too, as those concepts are not repeated in this summary.

## 16.16 Exercises

1. Suppose you have data on a dependent variable $y_i$ and a column vector of regressors $x_i$. Consider the model

$$y_i = x_i'\beta_0 + \epsilon_i$$

   (a) Suppose that $E[\epsilon_i|x_i] = 0$. Use this information to propose a GMM estimator that is equivalent to the OLS estimator. Your answer should include:

      i. state the moment conditions and the GMM objective function clearly

      ii. compute the first order conditions for minimization of the GMM criterion function and solve them to find the expression for the estimator

   (b) Now, suppose that $E[\epsilon_i|x_i] \neq 0$ but that there is another vector $z_i$ with

$\dim z = \dim x$ such that $E[\epsilon_i | z_i] = 0$.

i. Show that the OLS estimator of $\beta_0$ is not consistent, given this information.

ii. Propose a consistent GMM estimator of the parameter vector $\beta_0$ that uses this information. Your answer should include:

A. a clear statement of the moment conditions and the GMM objective function to be minimized which defines the estimator.

B. a closed-form expression (that is, an explicit formula) for the estimator.

C. a proof that the estimator is consistent. If you need to make additional assumptions to prove consistency, state them.

2. Consider the linear regression model $y_i = x_i' \beta_0 + \epsilon_i$. Suppose that the $n$ observations satisfy all assumptions of the classical linear regression model, except they are heteroscedastic, so that $V(\epsilon_i | x_i) = \sigma_i^2$. Suppose that $\sigma_i^2$ is

**known**, for all $i$. As such, we can compute the generalized least squares (GLS) estimator, which can be expressed as $\hat{\beta}_{GLS} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}y$, where $\Sigma$ is a diagonal matrix with $\sigma_i^2$ in position $(i, i)$, $i = 1, 2, ..., n$.

(a) Define moment conditions which define a GMM estimator that is equivalent to the GLS estimator. Prove that the estimators are equivalent.

(b) Is your GMM estimator exactly identified, or overidentified?

(c) For efficient estimation using the moment conditions which you have defined, is single step GMM estimation using an identity matrix as the weight matrix fully efficient, or will two step estimation be more efficient? Explain.

3. Do the exercises in section 16.9.

4. Show how the GIV estimator presented in section 16.9 can be adapted to account for an error term with HET and/or AUT.

5. For the GIV estimator presented in section 16.9, find the form of the expressions $\mathcal{I}_\infty(\theta^0)$ and $\mathcal{J}_\infty(\theta^0)$ that appear in the asymptotic distribution of the estimator, assuming that an efficient weight matrix is used.

6. Using Julia, generate data from the logit dgp. The script EstimateLogit.jl should prove quite helpful.

   (a) Recall that $E(y_t|\mathbf{x}_t) = \mathbf{p}(\mathbf{x}_t, \theta) = [1+\exp(-\mathbf{x}_t\prime\theta)]^{-1}$. Consider the moment condtions (exactly identified) $m_t(\theta) = [y_t - p(\mathbf{x}_t, \theta)]\mathbf{x}_t$. Estimate by GMM (using `gmmresults`), using these moments.

   (b) Estimate by ML (using `mleresults`).

   (c) The two estimators should coincide. Prove analytically that the estimators coincide.

7. When working out the structure of $\Omega_n$, show that $\mathcal{E}(m_t m_{t+s}') = \Gamma_s'$.

8. Verify the missing steps needed to show that $n \cdot \bar{m}(\hat{\theta})'\hat{\Omega}^{-1}\bar{m}(\hat{\theta})$ has a $\chi^2(g - K)$ distribution. That is, show that the monster matrix is idempotent and has trace equal to $g - K$.

9. For the portfolio example, experiment with the program using lags of 3 and 4 periods to define instruments

   (a) Iterate the estimation of $\theta = (\beta, \gamma)$ and $\Omega$ to convergence.

   (b) Comment on the results. Are the results sensitive to the set of instruments used? Look at $\hat{\Omega}$ as well as $\hat{\theta}$. Are these good instruments? Are the instruments highly correlated with one another? Is there something analogous to collinearity going on here?

10. Run the Julia script GMM/chi2gmm.jl with several sample sizes. Do the results you obtain seem to agree with the consistency of the GMM estimator? Explain.

11. The GMM estimator with an arbitrary weight matrix has the asymptotic distribution

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' \left(D_\infty W_\infty D_\infty'\right)^{-1}\right]$$

Supposing that you compute a GMM estimator using an arbitrary weight matrix, so that this result applies. Carefully explain how you could test the hypothesis $H_0 : R\theta^0 = r$ versus $H_A : R\theta^0 \neq r$, where $R$ is a given $q \times k$ matrix, and $r$ is a given $q \times 1$ vector. I suggest that you use a Wald test. Explain exactly what is the test statistic, and how to compute every quantity that appears in the statistic.

12. (proof that the GMM optimal weight matrix is one such that $W_\infty = \Omega_\infty^{-1}$) Consider the difference of the asymptotic variance using an arbitrary weight matrix, minus the asymptotic variance using the optimal weight matrix:

$$A = \left(D_\infty W_\infty D_\infty'\right)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D_\infty' \left(D_\infty W_\infty D_\infty'\right)^{-1} - \left(D_\infty \Omega_\infty^{-1} D_\infty'\right)^{-1}$$

Set $B = (D_\infty W_\infty D'_\infty)^{-1} D_\infty W_\infty - (D_\infty \Omega_\infty^{-1} D'_\infty)^{-1} D_\infty \Omega_\infty^{-1}$. Verify that $A = B\Omega_\infty B'$. What is the implication of this? Explain.

13. The asymptotic distribution of the GMM estimator, using a non-optimal weight matrix, is

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left[0, (D_\infty W_\infty D'_\infty)^{-1} D_\infty W_\infty \Omega_\infty W_\infty D'_\infty (D_\infty W_\infty D'_\infty)^{-1}\right]$$

We know that in the case of exact identification, the GMM estimator does not depend on the weight matrix, $W$. If this is the case, the asymptotic covariance matrix must not depend on $W_\infty$, either. Prove that this is true, by showing that $W$ cancels out of the asymptotic variance. Hint: $(AB)^{-1} = B^{-1}A^{-1}$ if both $A$ and $B$ are invertible matrices.

14. In the context of the Hansen-Sargan test for correct specification of moments, discussed in Section 16.10, prove that the matrix $P_\infty = I_g - \Omega_\infty^{-1/2} D'_\infty (D_\infty \Omega_\infty^{-1} D'_\infty)^{-1} D$ is idempotent and that its rank is $g - K$, where $g$ is the number of moment

conditions and $K$ is the number of parameters.

15. Consider the two equation model

$$\begin{aligned}
\text{Demand: } q_t &= \alpha_1 + \alpha_2 p_t + \alpha_3 y_t + \varepsilon_{1t} \\
\text{Supply: } q_t &= \beta_1 + \beta_2 p_t + \varepsilon_{2t} \\
\mathcal{E}\left( \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} | y_t \right) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
\mathcal{E}\left( \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} & \varepsilon_{2t} \end{bmatrix} | y_t \right) &= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \forall t
\end{aligned}$$

The variables $q_t$ and $p_t$ are endogenous, and the variable $y_t$ is weakly exogenous. Assume that the observations are independent over time. Consider GMM estimation of the parameters of the two equations implemented as two stage least squares (2SLS). Recall that the 2SLS estimator uses $\widehat{p_t}$ as an instrument for the endogenous regressor $p_t$, where $\widehat{p_t}$ is the fitted value from OLS applied to the equation $p_t = \pi_1 + \pi_2 y_t + v_t$.

(a) Show that the regressor $p_t$ is correlated with each of the structural errors $\varepsilon_{1t}$ and $\varepsilon_{2t}$.

(b) Will OLS give a consistent estimator of the parameters of the supply equation? Explain your answer.

(c) Give the exact expression for the 2SLS estimator of the parameters of the supply equation, and explain why the estimator is consistent.

(d) Give the exact expression for the 2SLS estimator of the parameters of the demand equation, and carefully explain why 2SLS will *not* give a consistent estimator of these parameters. Note that the 2SLS estimator is a particular GMM estimator, and it is a particular instrumental variables (IV) estimator. Keeping this in mind may help you to answer the question.

16. Prove that the GMM estimator based upon the $g$ moment conditions $\bar{m}_n(\theta) = \begin{bmatrix} p_n'(\theta) & q_n'(\theta) \end{bmatrix}'$ and the corresponding true optimal weight matrix is asymp-

totically efficient with respect to the GMM estimator based upon the $h < g$ moment conditions $p_n(\theta)$ and the corresponding true optimal weight matrix.

(a) Interpret the result

(b) Discuss the importance of the result from an empirical point of view. Are there any cautions one should observe when doing applied GMM work? Describe any problems you can imagine.

17. Recall the dynamic model with measurement error that was discussed in class:

$$\begin{aligned} y_t^* &= \alpha + \rho y_{t-1}^* + \beta x_t + \epsilon_t \\ y_t &= y_t^* + \upsilon_t \end{aligned}$$

where $\epsilon_t$ and $\upsilon_t$ are independent Gaussian white noise errors. Suppose that

$y_t^*$ is not observed, and instead we observe $y_t$. We can estimate the equation

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \nu_t$$

using GIV, as was done above. The Julia script GMM/SpecTest.jl performs a Monte Carlo study of the performance of the GMM criterion test,

$$n \cdot s_n(\hat{\theta}) \xrightarrow{d} \chi^2(g - K)$$

Examine the script and describe what it does. Run this script to verify that the test over-rejects. Increase the sample size, to determine if the over-rejection problem becomes less severe. Discuss your findings.

18. Suppose we have two equations

$$\begin{aligned} y_{t1} &= \alpha_1 + \alpha_2 y_{t2} + \epsilon_{t1} \\ y_{t2} &= \beta_1 + \beta_2 x_t + \epsilon_{t2} \end{aligned}$$

where $V(\epsilon_{t1}) = \sigma_1^2 > 0$, $V(\epsilon_{t2}) = \sigma_2^2 > 0$, $E(\epsilon_{t1}\epsilon_{t2}) = \sigma_{12} \neq 0$. The observations are independent over time, and the errors have zero mean. The variable $x_t$ is strictly exogenous: it is uncorrelated with the two epsilons at all time periods.

(a) Is the OLS estimator of the parameters of the first equation consistent or not? Explain.

(b) Is the OLS estimator of the parameters of the second equation consistent or not? Explain.

(c) If the OLS estimator of the parameters of the first equation is not consistent, propose a consistent estimator of the parameters of the first equation and explain why the proposed estimator is consistent.

(d) If the OLS estimator of the parameters of the second equation is not consistent, propose a consistent estimator of the parameters of the second equation and explain why the proposed estimator is consistent.

19. Estimate a logit model by GMM using the 10 independent data points

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| y | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| x | -1 | -1 | 1 | 0 | -1 | -1 | 1 | 1 | 2 | 2 |

.

For the logit model, the probability $P(y_t = 1|x_t) = (1+\exp(-\theta_1 - \theta_2 x_t))^{-1}$, and the probability that $y_t = 0$ is the complement.

(a) create a data file that contains these observations

(b) find the conditional mean $E(y|x)$ and the conditional variance $V(y|x)$

(c) propose at least 2 moment conditions, using the mean and the variance you found in (b)

(d) write a Julia function that computes the GMM estimator using your two moment conditions

(e) compute the two step efficient GMM estimator

(f) comment on the results

20. Given the 10 independent data points

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| y | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 3 |
| x | -1 | -1 | 1 | 0 | -1 | -1 | 1 | 1 | 2 | 2 |

.

For the Poisson model, the density $f_Y(y|x) = \frac{\exp(-\lambda)\lambda^y}{y!}$, $y = 0, 1, 2, \ldots$. To make the model depend on conditioning variables, use the parameterization $\lambda(x) = \exp(\theta_1 + \theta_2 x)$.

(a) The mean of a Poisson distribution with parameter $\lambda$ is equal to $\lambda$, and so is the variance. Propose moment conditions to an overidentified ($g > k$) GMM estimator of $\theta_1$ and $\theta_2$.

(b) Discuss how your proposed moment conditions relate to the score function of the maximum likelihood estimator.

(c) Estimate the parameters using two-step efficient GMM, using the moment conditions you have proposed.

(d) Discuss the results, and compare them to your ML estimates for the

similar problem in the chapter on ML estimation.

21. Suppose that we have a rational expectations model. At time $t$, the representative agent knows the values of all variables indexed $t-1$ and earlier. Suppose the production function is $y_t = k_t^{\alpha} n_t^{1-\alpha} \exp(\epsilon_t)$, where $y$ is output, $k$ is capital, $n$ is labor, and $\epsilon$ is a white noise shock. One of the first order conditions tells us that the wage rate, $w_t$, is equal to the marginal product of labor (recall: MPL=$\partial y/\partial n$). The econometrician has a data set that includes only the variables $k_t$,$n_t$ and $w_t$. The econometrician knows the correct specification of the production function, except for the value of $\alpha$, and also knows that $\epsilon_t$ is a white noise shock. At time $t$, the agent observes the shock $\epsilon_t$, and the capital level, $k_t$, which is a non-stochastic function of variables that are pre-determined at time $t$, and then makes choices that determine the values of the variables $y_t$,$n_t$,$w_t$,$r_t$, among others, using the information that has been given, plus some other unspecified equations.

(a) Consider the linear regression $\log w_t = c + \alpha(\log k_t - \log w_t) + \epsilon_t$. Will the OLS estimate of $\alpha$ using this regression be consistent or not? Explain.

(b) Provide at least two moment conditions that can be used to consistently estimate the parameter $\alpha$ by GMM.

(c) Explain why the proposed moment conditions have mean zero at the true parameter value.

(d) Explain how you could estimate $\alpha$ using a two step efficient GMM estimator. Give the explicit form of the optimal weight matrix, for your chosen moment conditions.

22. Consider the model

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \beta x_t + \epsilon_t$$

where $\epsilon_t$ is a $N(0,1)$ white noise error. This is an autoregressive model of

order 2 (AR2) model, with an additional exogenous regressor. Suppose that data is generated from the AR2 model, but the econometrician mistakenly decides to estimate an AR1 model, $y_t = \alpha + \rho_1 y_{t-1} + \beta x_t + v_t$. This is a case of omitted relevant variables.

(a) show that weak exogeneity fails for the AR1 model.

(b) Consider IV estimation of the AR1 model, using lags of $x_t$ as instruments. Is this a consistent estimator?

(c) simulate data from the correct AR2 model, using $\alpha = 0$, $\rho_1 = 0.5$, $\rho_2 = 0.4$, $\beta = 2$, and $x_t \sim IIN(0, 1)$. Use a sample size of 30 observations.

   i. estimate the incorrectly specified AR1 model by OLS

   ii. estimate the correctly specified AR2 model by OLS

   iii. implement your proposed IV estimator of the AR1 model

   iv. embed the simulations and estimations in a loop, to do a Monte Carlo study using 1000 replications. Provide histograms for the

distribtions of the estimators of the parameter $\rho_1$ for the 3 estimators.

(d) discuss all results thoroughly, focusing on bias and standard errors of the estimators of the autoregressive parameters

23. Estimate the investment equation of the Klein Model 1 (see Section 11.10) using GMM. See the example at the end of the discussion of 2SLS for a good hint.

24. Verify the missing steps needed to show that $n \cdot m(\hat{\theta})'\hat{\Omega}^{-1}m(\hat{\theta})$ has a $\chi^2(g-K)$ distribution. That is, show that the big ugly matrix is idempotent and has trace equal to $g-K$.

# Chapter 17

# Models for time series data

Hamilton, *Time Series Analysis* is a good reference for this chapter.

Up to now we've considered the behavior of the dependent variable $y_t$ as a function of other variables $x_t$. These variables can of course contain lagged dependent variables, e.g., $x_t = (w_t, y_{t-1}, ..., y_{t-j})$. Pure time series methods consider the

behavior of $y_t$ as a function only of its own lagged values, unconditional on other observable variables. One can think of this as modeling the behavior of $y_t$ after marginalizing out all other variables. But, of course, general models will include lagged dependent variables and other explanatory variables. This Chapter gives a brief description of some of the widely used models.

# Basic concepts

**Definition 73.** [Stochastic process] A stochastic process is a sequence of random variables, indexed by time: $\{Y_t\}_{t=-\infty}^{\infty}$

**Definition 74.** [Time series] A time series is **one** observation of a stochastic process, over a specific interval: $\{y_t\}_{t=1}^{n}$.

So a time series is a sample of size 1 of $n$ particular elements of a stochastic process. Each of the $Y_t$ is sampled only once, for $t = 1, 2, ..., n$. It's important to keep in mind that conceptually, one could draw another sample, and that the values would be different.

**Definition 75.** [Autocovariance] The $j^{th}$ autocovariance of a stochastic process is $\gamma_{jt} = \mathcal{E}(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j})$ where $\mu_t = \mathcal{E}(Y_t)$.

**Definition 76.** [Covariance (weak) stationarity] A stochastic process is covariance stationary if it has time constant mean and autocovariances of all orders:

$$\begin{aligned} \mu_t &= \mu, \ \forall t \\ \gamma_{jt} &= \gamma_j, \ \forall t \end{aligned}$$

As we've seen, this implies that $\gamma_j = \gamma_{-j}$ : the autocovariances depend only one the interval between observations, but not the time of the observations.

**Definition 77.** [Strong stationarity] A stochastic process is strongly stationary if the joint distribution of an arbitrary collection of the $\{Y_t\}$, e.g.,

$$(Y_{t-j}, Y_{t-k}, ..., Y_t, ..., Y_{t+l}, Y_{t+m}\},$$

doesn't depend on $t$.

Since moments are determined by the distribution, strong stationarity$\Rightarrow$weak

stationarity.

How can we estimate the mean of $Y_t$?

- Here's where we need stationarity. Without it, each $Y_t$ will have a different mean, $\mu_t$, and with a time series, we have only one observation to work with.

- The time series is one sample from the stochastic process, and each of the random variables over the sample interval is sampled only once.

- One could think of $M$ repeated samples from the stoch. proc., e.g., $\{y_{tm}\}_{m=1}^{M}$ By a LLN, we would expect that

$$\frac{1}{M}\sum_{m=1}^{M} y_{tm} \xrightarrow{p} \mathcal{E}(Y_t)$$

  as $M$ gets large. We would need $M$ universes, each observed at time $t$.

- The problem is, we have only one sample to work with, since we can't go back in time, or to another universe, and collect another. How can $\mathcal{E}(Y_t)$ be estimated then? It turns out that *ergodicity* is the needed property.

**Definition 78.** [Ergodicity]. A stationary stochastic process is ergodic (for the mean) if the time average converges to the mean

$$\frac{1}{n}\sum_{t=1}^{n} y_t \xrightarrow{p} \mu \tag{17.1}$$

A sufficient condition for ergodicity is that the autocovariances be absolutely summable:

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty$$

This implies that the autocovariances die off, so that the $Y_t$ are not so strongly dependent that they don't satisfy a LLN.

**Definition 79.** [Autocorrelation] The $j^{th}$ autocorrelation, $\rho_j$ is just the $j^{th}$ autocovariance divided by the variance:

$$\rho_j = \frac{\gamma_j}{\gamma_0} \tag{17.2}$$

**Definition 80.** [White noise] White noise is just the time series literature term for a classical error. $\epsilon_t$ is white noise if i) $\mathcal{E}(\epsilon_t) = 0, \forall t$, ii) $V(\epsilon_t) = \sigma^2, \forall t$ and iii) $\epsilon_t$ and $\epsilon_s$ are independent, $t \neq s$. Gaussian white noise just adds a normality assumption.

**Example 81.** US quarterly macro data, used in Stock and Watson (2011), Chapter 14. The original materials are at `http://wps.pearsoned.co.uk/ema_ge_stock_ieupdate_3/251/64413/16489878.cw/index.html`. The data file , in GRETL format. The data description file. Use GRETL to:

- plot the GDP data, and notice that it's nonstationary.
- Plot the growth rate, and note that it's stationary.
- compute the autocorrelations for the annual growth rate of GDP, using the GRETL correlogram option: they die off fairly quickly, so ergodicity seems to hold
- compute the autocorrelations of GDP. HIGHLY PERSISTENT. Doubtful that the ergodicity condition will hold.
- we are going to want to work with stationary data, if we want to apply standard regression methods and inference.

- working with nonstationary data can give very misleading results, if we rely on standard theory for stationary data, as we will see.

## 17.1 ARMA models

With these concepts, we can discuss ARMA models. These are closely related to the AR and MA error processes that we've already discussed. The main difference is that the lhs variable is observed directly now.

# MA(q) processes

A $q^{th}$ order moving average (MA) process is

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where $\varepsilon_t$ is white noise. The variance is

$$\begin{aligned}
\gamma_0 &= \mathcal{E}\left(y_t - \mu\right)^2 \\
&= \mathcal{E}\left(\varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}\right)^2 \\
&= \sigma^2 \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)
\end{aligned}$$

Similarly, the autocovariances are

$$\begin{aligned}
\gamma_j &= \mathcal{E}\left[\left(y_t - \mu\right)\left(y_{t-j} - \mu\right)\right] \\
&= \sigma^2 (\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \cdots + \theta_q \theta_{q-j}), j \leq q \\
&= 0, j > q
\end{aligned}$$

Therefore an MA(q) process is necessarily covariance stationary and ergodic, as long as $\sigma^2$ and all of the $\theta_j$ are finite.

For example, if we have an MA(1) model, then $E(y_t) = \mu$, $V(y_t) = \sigma^2(1+\theta_1^2)$, and $\gamma_1 = \sigma^2\theta_1$. The higher order autocovariances are zero.

- Thus, if the model is MA(1) with normally distributed shocks, the density of the vector of $n$ observations, $y$, is

$$f_Y(y|\rho) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(y-\mu)' \Sigma^{-1} (y-\mu)\right) \tag{17.3}$$

  where

$$\Sigma = \sigma^2 \begin{bmatrix} 1+\theta_1^2 & \theta_1 & 0 & \cdots & 0 \\ \theta_1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \theta_1 \\ 0 & \cdots & 0 & \theta_1 & 1+\theta_1^2 \end{bmatrix}.$$

  - With this, it is very easy to program the log-likelihood function. For higher order MA models, the only difference is the structure of $\Sigma$ becomes more complicated. In this form, one needs a lot of computer

memory. A more economical approach uses the Kalman filter, which we'll see in the discussion of state space models.

- If we don't make assumptions on the distribution of the shocks, then method of moments estimation can be used.

**Exercise 82.** Generate data that follows a simple MA(1) model: $y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}$ for $\mu = 0$ and $\theta_1 = 0.5$, and with $\epsilon_t + 1$ distributed independently and identically $\chi^2(1)$. Do estimation by GMM, and verify experimentally (by increasing the sample size) that the estimator is consistent. Hint: generate $\epsilon_t$ as the square of a standard normal, minus 1.

- An issue to be aware of is that MA models are not identified, in that there exist multiple parameter values that give the same value of the likelihood function.

- For example, the MA(1) model with $\tilde{\sigma}^2 = \theta^2\sigma^2$ and $\tilde{\theta}_1 = \frac{1}{\theta_1}$ has identical first and second moments to the original model, so the likelihood function has the same value.

- Normally, the parameterization that leads to an *invertible* MA model is the one that is selected. An invertible MA model is one that has a representation as a AR($\infty$) model. For the MA(1) model, the invertible parameterization has $|\theta_1| < 1$.

- This implies that parameter restrictions will need to be imposed when estimating the MA model, to enforce selection of the invertible model.

- Maximization of the conditional likelihood is also used for estimation, sometimes. Assuming that $\epsilon_0$ is known (for example, equal to zero), then one can

compute $\epsilon_1$, given the parameters. Then one works forward recursively to get all of the $\epsilon_t$. With these, the likelihood function is very easy to compute. This is a convenient shortcut, but it's not recommended if the sample is not large, especially since it's not hard to compute the exact likelihood function.

# AR(p) processes

An AR(p) process can be represented as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

where $\epsilon_t$ is white noise. This is just a linear regression model, and assuming stationarity, we can estimate the parameters by OLS. What is needed for stationarity?

The dynamic behavior of an AR(p) process can be studied by writing this $p^{th}$ order difference equation as a vector first order difference equation (this is known as the companion form):

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} \phi_1 & \phi_2 & \cdots & & \phi_p \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \cdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

This is just the equation of the AR(p) process in the first row, and identities in the remaining rows. In matrix form:

$$Y_t = C + FY_{t-1} + E_t$$

- There are $p$ variables in this set of $p$ equations, so $C$ is $p \times 1$,$F$ is $p \times p$, etc.

With this, we can recursively work forward in time:

$$\begin{aligned} Y_{t+1} &= C + FY_t + E_{t+1} \\ &= C + F\left(C + FY_{t-1} + E_t\right) + E_{t+1} \\ &= C + FC + F^2Y_{t-1} + FE_t + E_{t+1} \end{aligned}$$

and

$$\begin{aligned} Y_{t+2} &= C + FY_{t+1} + E_{t+2} \\ &= C + F\left(C + FC + F^2Y_{t-1} + FE_t + E_{t+1}\right) + E_{t+2} \\ &= C + FC + F^2C + F^3Y_{t-1} + F^2E_t + FE_{t+1} + E_{t+2} \end{aligned}$$

or in general

$$Y_{t+j} = C + FC + \cdots + F^jC + F^{j+1}Y_{t-1} + \underbrace{F^jE_t} + F^{j-1}E_{t+1} + \cdots + FE_{t+j-1} + E_{t+j}$$

Consider the impact of a shock in period $t$ on $y_{t+j}$. This is simply the partial of $Y_{t+j}$ with respect to the (1,1) element of the shocks in the bracketed term of the previous equation:

$$\frac{\partial Y_{t+j}}{\partial E_t'}_{(1,1)} = F^j_{(1,1)}$$

If the system is to be stationary, then as we move forward in time this impact must die off. Otherwise a shock causes a permanent change in the mean of $y_t$. Therefore, stationarity requires that

$$\lim_{j \to \infty} F^j_{(1,1)} = 0$$

- Save this result, we'll need it in a minute.

Consider the eigenvalues of the matrix $F$. These are the $\lambda$ such that

$$|F - \lambda I_P| = 0$$

The determinant here can be expressed as a polynomial. For example, for $p = 1$, the matrix $F$ is simply

$$F = \phi_1$$

so

$$|\phi_1 - \lambda| = 0$$

can be written as

$$\phi_1 - \lambda = 0$$

When $p = 2$, the matrix $F$ is

$$F = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}$$

so

$$F - \lambda I_P = \begin{bmatrix} \phi_1 - \lambda & \phi_2 \\ 1 & -\lambda \end{bmatrix}$$

and

$$|F - \lambda I_P| = \lambda^2 - \lambda\phi_1 - \phi_2$$

So the eigenvalues are the roots of the polynomial

$$\lambda^2 - \lambda\phi_1 - \phi_2$$

which can be found using the quadratic equation. This generalizes. For a $p^{th}$ order AR process, the eigenvalues are the roots of

$$\lambda^p - \lambda^{p-1}\phi_1 - \lambda^{p-2}\phi_2 - \cdots - \lambda\phi_{p-1} - \phi_p = 0$$

Supposing that all of the roots of this polynomial are distinct, then the matrix $F$ can be factored as

$$F = T\Lambda T^{-1}$$

where $T$ is the matrix which has as its columns the eigenvectors of $F$, and $\Lambda$ is a diagonal matrix with the eigenvalues on the main diagonal. Using this decomposition, we can write

$$F^j = \left(T\Lambda T^{-1}\right)\left(T\Lambda T^{-1}\right)\cdots\left(T\Lambda T^{-1}\right)$$

where $T\Lambda T^{-1}$ is repeated $j$ times. This gives

$$F^j = T\Lambda^j T^{-1}$$

and

$$\Lambda^j = \begin{bmatrix} \lambda_1^j & 0 & & 0 \\ 0 & \lambda_2^j & & \\ & & \ddots & \\ 0 & & & \lambda_p^j \end{bmatrix}$$

Supposing that the $\lambda_i$ $i = 1, 2, ..., p$ are all real valued, it is clear that

$$\lim_{j \to \infty} F^j_{(1,1)} = 0$$

requires that

$$|\lambda_i| < 1, i = 1, 2, ..., p$$

e.g., the eigenvalues must be less than one in absolute value.

- It may be the case that some eigenvalues are complex-valued. The previous result generalizes to the requirement that the eigenvalues be less than one in *modulus*, where the modulus of a complex number $a + bi$ is

  $$mod(a + bi) = \sqrt{a^2 + b^2}$$

  This leads to the famous statement that "stationarity requires the roots of the determinantal polynomial to lie inside the complex unit circle." *draw picture here.*

- When there are roots on the unit circle (unit roots) or outside the unit circle, we leave the world of stationary processes.

- Dynamic multipliers: $\partial y_{t+j}/\partial \varepsilon_t = F^j_{(1,1)}$ is a *dynamic multiplier* or an *impulse-response* function. Real eigenvalues lead to steady movements, whereas complex eigenvalues lead to oscillatory behavior. Of course, when there are multiple eigenvalues the overall effect can be a mixture. *pictures*

## Moments of AR(p) process

The AR(p) process is

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Assuming stationarity, $\mathcal{E}(y_t) = \mu, \forall t$, so

$$\mu = c + \phi_1 \mu + \phi_2 \mu + ... + \phi_p \mu$$

so

$$\mu = \frac{c}{1 - \phi_1 - \phi_2 - ... - \phi_p}$$

and

$$c = \mu - \phi_1 \mu - ... - \phi_p \mu$$

SO

$$
\begin{aligned}
y_t - \mu &= \mu - \phi_1\mu - \ldots - \phi_p\mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \mu \\
&= \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t
\end{aligned}
$$

With this, the second moments are easy to find: The variance is

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + ... + \phi_p\gamma_p + \sigma^2$$

The autocovariances of orders $j \geq 1$ follow the rule

$$\begin{aligned}
\gamma_j &= \mathcal{E}\left[(y_t - \mu)\left(y_{t-j} - \mu\right)\right] \\
&= \mathcal{E}\left[\left(\phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t\right)\left(y_{t-j} - \mu\right)\right] \\
&= \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + ... + \phi_p\gamma_{j-p}
\end{aligned}$$

Using the fact that $\gamma_{-j} = \gamma_j$, one can take the $p+1$ equations for $j = 0, 1, ..., p$, which have $p+1$ unknowns ($\sigma^2$, $\gamma_0, \gamma_1, ..., \gamma_p$) and solve for the unknowns. With these, the $\gamma_j$ for $j > p$ can be solved for recursively.

# ARMA model

An ARMA$(p, q)$ model is $(1 + \phi_1 L + \phi_2 L^2 + ... + \phi_p L^p) y_t = c + (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q)\epsilon_t$. These are popular in applied time series analysis.

- A high order AR process *may* be well approximated by a low order MA process, and a high order MA process *may* be well approximated by a low order AR process.

- By combining low order AR and MA processes in the same model, one can hope to fit a wide variety of time series using a parsimonious number of parameters.

- There is much literature on how to choose $p$ and $q$, which is outside the scope of this course. This involves comparing the ACF and PACFs.

- Estimation can be done using the Kalman filter, assuming that the errors are normally distributed.

**Example 83.** Using GRETL, try out various models to explain the unemployment rate, using the S&W US quarterly macro data.

- estimate a MA(4) model for the unemployment rate
- estimate a AR(4) model for the unemployment rate
- estimate an ARMAX(1,1) model using 12 lags of the GDP growth rate (don't use current value)
  - interpret the estimated coefficients. What can we say about persistence and speed of transmission of effects in the economy?
  - look at fit and residuals. Observe the "Great Moderation" of the 1990's, and the return to volatility after the 2007 Great Recession.
- look at the BIC to help to decide which model to use

## 17.2 VAR models

Consider the model

$$
\begin{aligned}
y_t &= C + A_1 y_{t-1} + \epsilon_t \qquad (17.4)\\
E(\epsilon_t \epsilon_t') &= \Sigma \\
E(\epsilon_t \epsilon_s') &= 0, t \neq s
\end{aligned}
$$

where $y_t$ and $\epsilon_t$ are $G \times 1$ vectors, $C$ is a $G \times 1$ of constants, and $A_1$ is a $G \times G$ matrix of parameters. The matrix $\Sigma$ is a $G \times G$ covariance matrix. Assume that we have $n$ observations. This is a *vector autoregressive* model, of order 1 - commonly referred to as a VAR(1) model. It is a collection of $G$ AR(1) models, augmented to include lags of other endogenous variables, and the $G$ equations are contemporaneously correlated. The extension to a VAR(p) model is quite obvious.

- As shown in Section 11.3, it is efficient to estimate a VAR model using OLS equation by equation, there is no need to use GLS, in spite of the cross

equation correlations.

A VAR model of this form can be thought of as the reduced form of a dynamic simultaneous equations system, with all of the variables treated as endogenous, and with lags of all of the endogenous variables present:

- The simultaneous equations model is (see equation 11.2)

$$Y_t'\Gamma = X_t'B + E_t'$$

- this can be written after transposing (and adapting notation to use small case, pulling the constant out of $X_t$ and using $v_t$ for the error) as $\Gamma' y_t = a + B' x_t + v_t$.

- Let $x_t = y_{t-1}$. Then we have $\Gamma' y_t = a + B' y_{t-1} + v_t$.

- Premultiplying by the inverse of $\Gamma'$ gives

$$y_t = (\Gamma')^{-1} a + (\Gamma')^{-1} B' y_{t-1} + (\Gamma')^{-1} v_t.$$

- Finally define $C = (\Gamma')^{-1} a$, $A_1 = (\Gamma')^{-1} B'$ and $\epsilon_t = (\Gamma')^{-1} v_t$, and we have the VAR(1) model of equation 17.4.

- C. Sims originally proposed reduced form VAR models as an alternative to structural simultaneous equations models, which were perceived to require too many unrealistic assumptions for their identification. He showed that VARs often forecast better than structural models, which cast doubt upon the identification restrictions of the structural model.

- However, the search for structural interpretations of VAR models slowly crept back into the literature, leading to "structural VARs".

- A structural VAR model is really just a certain form of dynamic linear simultaneous equations model, with other imaginative and hopefully more realistic methods used for identification.

- The issue of identifying the structural parameters $\Gamma$ and $B$ is more or less the same problem that was studied in the context of simultaneous equations.

- There, identification was obtained through zero restrictions. In the structural VAR literature, zero restrictions are often used, but other information may

also be used, such as covariance matrix restrictions or sign restrictions.

- Interest often focuses on the impulse-response functions. Identification of the impact of structural shocks (how to estimate the impact-response functions) is complicated, with many alternative methodologies, and is often a topic of much disagreement among practitioners. The estimated impulse response functions are often sensitive to the identification strategy that is used. There is a large literature.

- Papers by C. Sims are a good place to start, if one wants to learn more. He also offers a good deal of useful software on his web page.

An issue which arises when a VAR(p) model $y_t = C + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t$ is contemplated is that the number of parameters increases rapidly in p, which introduces severe collinearity problems.

- One can use Bayesian methods such as the "Minnesota prior" (search for papers by Litterman), which is a prior that each variable separately follows a random walk (an AR(1) model with $\rho = 1$).
  - The prior on $A_1$ is that it is an identity matrix
  - and the prior on the $A_j$, $j > 1$ is that they are zero matrices
  - thus, each variable follows a random walk, according to the prior
- This can be done using stochastic restrictions similar to what was in the discussion of collinearity and ridge regression. For example, a VAR(2) model in de-meaned variables, with $G$ variables, can be written as

$$Y = \begin{bmatrix} Y_{-1} & Y_{-2} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} + \epsilon$$

We can impose the stochastic restriction that $A_1 = I_2 - v_1$ and that $A_2 = 0_2 - v_2$. Augmenting the data with these 4 "artificial observations", we get

$$\begin{bmatrix} Y \\ I_G \\ 0_G \end{bmatrix} = \begin{bmatrix} Y_{-1} & Y_{-2} \\ I_G & 0_G \\ 0_G & I_G \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} + \begin{bmatrix} \epsilon \\ v_1 \\ v_2 \end{bmatrix}$$

Then we can impose how important the restrictions are by weighting the stochastic restrictions, along the lines of a GLS heteroscedasticity correction:

$$\begin{bmatrix} Y \\ k_1 I_G \\ 0_G \end{bmatrix} = \begin{bmatrix} Y_{-1} & Y_{-2} \\ k_1 I_G & 0_G \\ 0_G & k_2 I_G \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} + \begin{bmatrix} \epsilon \\ k_1 v_1 \\ k_2 v_2 \end{bmatrix}$$

Then we fit by OLS. When $k_1$ is large, the estimated $A_1$ will be forced to be close to an identity matrix. When $k_2$ is large, the second lag coefficients are all forced to zero. Jointly, these restrictions push the model in the direction

of separate random walks for each variable. The degree to which the model is pushed depends on the $ks$. When the $ks$ are small, the fit is close to the unrestricted OLS fit, when they are large, it is close to separate random walks.

"Bayesian VARs" is a now a substantial body of literature. An introduction to more formal Bayesian methods is given in a chapter that follows. For highly parameterized models, Bayesian methods can help to impose structure.

**Example 84.** Using GRETL, using the Stock and Watson US quarterly macro data.

- compute the term spread using the US Macro data. (term spread is GS10 - TB3MS)
- estimate a VAR(1) model for unemployment rate, GDP growth rate, and the term spread
- examine the impulse-response functions
- using the BIC, is the equation for the unemployment rate preferred, compared to the models of the previous example?
- See Stock and Watson, Ch. 14 for more discussion

**Exercise 85.** Get the simulation data from the example DSGE model . Recall that this simulated data intends to be representative of 40 years of quarterly data.

1. Estimate a VAR(1) model. Do an analysis of collinearity. Compute impulse-response functions.

2. Estimate an AR(1) model for output (by using VAR, but only select output). Compare $R^2$ and impulse-responses to the VAR(1) model. Note that matching impulse response functions has sometimes been used for estimation of DSGE models. Perhaps we'll see this idea again.

## 17.3 ARCH, GARCH and Stochastic volatility

ARCH (autoregressive conditionally heteroscedastic) models appeared in the literature in 1982, in Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation", *Econometrica* 50:987-1008. This paper stimulated a very large growth in the literature for a number of years afterward. The related GARCH (generalized ARCH) model is now one of the most widely used models for financial time series.

Financial time series often exhibit several type of behavior:

- volatility clustering: periods of low variation can be followed by periods of high variation
- fat tails, or excess kurtosis: the marginal density of a series is more strongly peaked and has fatter tails than does a normal distribution with the same mean and variance.

- leverage (negative correlation between returns and volatility), which often shows up as negative skewness of returns
- perhaps slight autocorrelation within the bounds allowed by arbitrage

The data set "nysewk.gdt", which is provided with Gretl, provides an example (or, use the SP500 data ) . If we compute 100 times the growth rate of the series, using log differences, we can obtain the plots in Figure 17.1 (Julia code for this is here ). In the first we clearly see volatility clusters, and in the second, we see excess kurtosis, skew, and tails fatter than the normal distribution. The skewness suggests that leverage may be present. We'll see how the third plot was made in the chapter on nonparametric estimation.

- compute descriptive statistics: negative skew and positive excess kurtosis
- regress returns on its own lag and on squared returns and lags: low predictability
- regress squared returns on its own lags and on returns: more predictable, evidence of leverage

Figure 17.1: NYSE weekly close price, 100 $\times$log differences

- The presence of volatility clusters indicates that the variance of the series is not constant over time, conditional on past events. Engle's ARCH paper was the first to model this feature.

- The frequency plot shows excess kurtosis and skew (leverage)

# ARCH

A basic ARCH specification is

$$
\begin{aligned}
y_t &= \mu + \rho y_{t-1} + \epsilon_t \\
&= g_t + \epsilon_t \\
\epsilon_t &= \sigma_t u_t \\
\sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2
\end{aligned}
$$

where the $u_t$ are Gaussian white noise shocks. The ARCH variance is a moving average process. Previous large shocks to the series cause the conditional variance of the series to increase. There is no leverage: negative shocks have the same impact on the future variance as do positive shocks..

- for $\sigma_t^2$ to be positive for all realizations of $\{\epsilon_t\}$, we need $\omega > 0$, $\alpha_i \geq 0$, $\forall i$.
- to ensure that the model is covariance stationary, we need $\Sigma_i \alpha_i < 1$. Otherwise, the variances will explode off to infinity.

Given that $\epsilon_t$ is normally distributed, to find the likelihood in terms of the observable $y_t$ instead of the unobservable $\epsilon_t$, first note that the series $u_t = (y_t - g_t)/\sigma_t = \frac{\epsilon_t}{\sigma_t}$ is iid Gaussian, so the likelihood is simply the product of standard normal densities.

$$\begin{aligned} u &\sim N(0, I), \text{ so} \\ f(u) &= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_t^2}{2}\right) \end{aligned}$$

The joint density for $y$ can be constructed using a change of variables:

- We have $u_t = (y_t - \mu - \rho y_{t-1})/\sigma_t$, so $\frac{\partial u_t}{\partial y_t} = \frac{1}{\sigma_t}$ and $|\frac{\partial u}{\partial y'}| = \Pi_{t=1}^{n} \frac{1}{\sigma_t}$,
- doing a change of variables,

$$f(y;\theta) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_t} \exp\left(-\frac{1}{2}\left(\frac{y_t - \mu - \rho y_{t-1}}{\sigma_t}\right)^2\right)$$

  where $\theta$ is the vector of all parameters (the parameters in $g_t$, and the $\omega$ and

alpha parameters of the ARCH specification. Taking logs,

$$\ln L(\theta) = -n \ln \sqrt{2\pi} - \sum_{t=1}^{n} \ln \sigma_t - \frac{1}{2} \sum_{t=1}^{n} \left( \frac{y_t - \mu - \rho y_{t-1}}{\sigma_t} \right)^2 .$$

In principle, this is easy to maximize. Some complications can arise when the restrictions for positivity and stationarity are imposed. Consider a fairly short data series with low volatility in the initial part, and high volatility at the end. This data appears to have a nonstationary variance sequence. If one attempts to estimate and ARCH model with stationarity imposed, the data and the restrictions are saying two different things, which can make maximization of the likelihood function difficult.

- use GRETL to estimate ARCH(1) and ARCH(4)
- if interested, adapt the Julia code for GARCH(1,1), below, to estimate an ARCH model.

# GARCH

Note that an ARCH model specifies the variance process as a moving average. For the same reason that an ARMA model may be used to parsimoniously model a series instead of a high order AR or MA, one can do the same thing for the variance series. A basic GARCH(p,q) (Bollerslev, Tim (1986). "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, 31:307-327) specification is

$$
\begin{aligned}
y_t &= \mu + \rho y_{t-1} + \epsilon_t \\
\epsilon_t &= \sigma_t u_t \\
\sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2
\end{aligned}
$$

It's just an ARCH model, with an autoregressive part added to the specification of the conditional variance. The idea is that a GARCH model with low values of p and q may fit the data as well or better than an ARCH model with large q.

- the model also requires restrictions for positive variance and stationarity, which are:

  - $\omega > 0$
  - $\alpha_i \geq 0,\ i = 1, ..., q$
  - $\beta_i \geq 0,\ i = 1, ..., p$
  - $\Sigma_{i=1}^{q} \alpha_i + \Sigma_{i=1}^{p} \beta_i < 1.$

- to estimate a GARCH model, you need to initialize $\sigma_0^2$ at some value. The sample unconditional variance is one possibility. Another choice could be the sample variance of the initial elements of the sequence. One can also "backcast" the conditional variance.

- The GARCH model also requires restrictions on the parameters to ensure stationarity and positivity of the variance.

- A useful modification is the EGARCH model (exponential GARCH, Nelson, D. B. (1991). "Conditional heteroskedasticity in asset returns: A new approach", Econometrica 59: 347-370). This model treats the logarithm of the variance as an ARMA process, so the variance will be positive without restrictions on the parameters.

- There are many variants that introduce asymmetry (leverage) and non-normality.

- GARCH(1,1) is a highly popular model in financial analysis.

The Julia script Garch11Example.jl illustrates estimation of a GARCH(1,1) model, using the NYSE closing price data. Results:

```
julia> include("Garch11Example.jl")
_____________________________________________________________
Garch(1,1) results
MLE Estimation Results     Convergence: true
Average Log-L: -2.07862    Observations: 2115
Sandwich form covariance estimator
```

| parameter | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| μ | 0.17649 | 0.04030 | 4.37928 | 0.00001 |
| ρ | 0.00167 | 0.02288 | 0.07312 | 0.94171 |
| ω | 0.15814 | 0.05640 | 2.80368 | 0.00510 |
| α | 0.11191 | 0.02599 | 4.30618 | 0.00002 |
| β | 0.85432 | 0.03039 | 28.11443 | 0.00000 |

Information Criteria

| | Crit. | Crit/n |
|---|---|---|
| CAIC | 8835.84587 | 4.17770 |
| BIC | 8830.84587 | 4.17534 |
| AIC | 8802.56182 | 4.16197 |

- examine the code to see how start values were determined, and how the variance loop was initialized.
- The AR(1) in the mean is probably not needed.

- Compare BIC (see subsection 15.8) to ARCH(1) and ARCH(4), which you can obtain using GRETL.

You can get the same results quickly and easily using Gretl:

```
Model 1: GARCH, using observations 670078-672192 (T = 2115)
Dependent variable: y
Standard errors based on Hessian

             coefficient  std. error    z       p-value
  ---------------------------------------------------------
  const      0.177119     0.0387575    4.570    4.88e-06 ***
  y_1        0.00148067   0.0232384    0.06372  0.9492

  alpha(0)   0.155435     0.0451241    3.445    0.0006   ***
  alpha(1)   0.111397     0.0171598    6.492    8.48e-11 ***
  beta(1)    0.855317     0.0228815   37.38     8.18e-306 ***

Mean dependent var 0.129001  S.D. dependent var 2.061158
Log-likelihood    4396.923  Akaike criterion   8805.846
Schwarz criterion  8839.786  Hannan-Quinn       8818.273
```

- There are some minor differences, because the Julia code initializes the variance in a different way, using only the first 10 observations. Also, the Julia code uses sandwich standard errors, while GRETL uses the Hessian, which tends to inflate t-statistics.

- Note that the $\beta_1$ parameter is highly significant. If you compare likelihood values or information criteria values with the ARCH results, you'll see that this model is favored - it fits better with fewer parameters.

- Gretl has a number of other ARCH/GARCH style models available.

- With Gretl, run the GARCH variants GJR(1,1) with skewed t shocks.

  - Do a density plot
  - note the BIC value
  - there are a lot of options to explore

- Note that the test of homoscedasticity against ARCH or GARCH involves parameters being on the boundary of the parameter space, which means that standard asymptotics do not apply. Also, the reduction of GARCH to ARCH has the same problem. Testing needs to be done taking this into account. See Demos and Sentana (1998) *Journal of Econometrics*.

## Stochastic volatility

In ARCH and GARCH models, the same shocks that affect the level also affect the variance. The stochastic volatility model allows the variance to have its own random component. A simple example is

$$
\begin{aligned}
y_t &= \phi \exp(h_t/2)\epsilon_t \\
h_t &= \rho h_{t-1} + \sigma u_t
\end{aligned}
$$

In this model, the log of the variance of the observed sequence follows an AR(1) model. Once can introduce leverage by allowing correlation between $\epsilon_t$ and $u_t$. This model is used as an example in the SimulatedNeuralMoments package, which can be used to generate data from the model. Typical data and a nonparametric density plot look like what we see in Figure 17.2. Note the volatility clusters, leptokurtosis, and the fat tails of the density.

- While the ARCH and GARCH models have a link between the shocks to $y_t$

Figure 17.2: SV model, typical data and density

and the dynamics of the variance of $y_t$, the stochastic volatility model has latent shocks to the variance which are not directly linked to the observed dependent variable. This may be perfectly reasonable: even when volatility is high, the mean of shocks to the observables may be zero. An ARCH model could not account for an increase in volatility without having a realized extreme shock to the level. The SV model can allow for this.

- The latent shocks complicate estimation. Many estimation methods have

been proposed, and this sort of model helped to popularize Bayesian methods in econometrics: see Jacquier, E., Polson, N.G. and Rossi, P.E., 2002. Bayesian analysis of stochastic volatility models. Journal of Business & Economic Statistics, 20(1), pp.69-87. We will see an examples of estimation in the chapter on simulation-based estimation.

- Variants of this sort of model are widely used to model financial data, competing with the GARCH(1,1) model for being the most popular choice.

## 17.4 Diffusion models

Financial data is often modeled using a continuous time specification. An example is the following model, taken from a paper of mine (JEF, 2015, with D. Kristensen).

A basic model is a simple continuous time stochastic volatility model with leverage. Log price $p_t = \log(P_t)$, solves the following pure diffusion model,

$$dp_t = (\mu_0 + \mu_1 \exp(h_t - \alpha))\, dt + \exp\left(\frac{h_t}{2}\right) dW_{1,t}$$

where the spot volatility (the instantaneous variance of returns), $\exp(h_t)$ is modeled using its logarithm:

$$dh_t = \kappa(\alpha - h_t)dt + \sigma dW_{2,t}.$$

Here, $W_{1,t}$ and $W_{2,t}$ are two standard Brownian motions with instantaneous correlation $\rho = Cov\,(dW_{1,t}, dW_{2,t})$. The parameters are interpreted as follows: $\mu_0$ is the baseline drift of returns; $\mu_1$ allows drift to depend upon spot volatility; $\alpha$ is the

mean of log volatility; $\kappa$ is the speed of mean reversion of log volatility, such that low values of $\kappa$ imply high persistence of log volatility; $\sigma$ is the so-called volatility of volatility; and $\rho$ is a leverage parameter that affects the correlation between returns and log volatility. We collect the parameters in $\theta = (\mu_0, \mu_1, \alpha, \kappa, \sigma, \rho)$.

An extension is to add jumps to the above model. These occur with Poisson frequency, and are conditionally normally distributed. More specifically, log-price $p_t$ solves the following continuous-time jump-diffusion model,

$$dp_t = (\mu_0 + \mu_1 \exp(h_t/2))\, dt + \exp\left(\frac{h_t}{2}\right) dW_{1,t} + J_t dN_t.$$

The Poisson process $N_t$ counts the number of jumps up to time $t$, and has jump intensity $\lambda_t = \lambda_0 + \lambda_1 \exp(h_t - \alpha)$ that varies with the volatility, while jump sizes, conditional on the occurrence of a jump, are independent and conditionally normally distributed: $J_t \sim N(\mu_J, \sigma_J^2)$. The inclusion of the jump component adds four parameters to $\theta$ as defined above, $\mu_J$, $\sigma_J^2$ and $\lambda_0$, and $\lambda_1$. This jump model was considered in, for example, Andersen, Benzoni and Lund (2002).

Figure 17.3: Returns from jump-diffusion model

An example of how returns, $r_t = 100(p_t - p_{t-1})$, generated by such a model might look is given in Figure 17.3. The spot volatility is plotted in Figure 17.4. Returns are observable, but spot volatility is not.

Figure 17.4: Spot volatility, jump-diffusion model

One might want to try to infer the parameters of the model, and also the latent spot volatility, using the observable data.

- Estimation of the parameters of such models is complicated by the fact that data is available in discrete time: $p_1, p_2, ...p_n$, but the model is in continuous time.

- One can "discretize" the model, to obtain something like the discrete time SV model of the previous section, but the discrete time transition density implied by the approximating model is not the same as the true transition density

$$p_t \sim f_p\left(p_t | p_{t-1}, h_{t-1}; \theta\right),$$

  induced by the continuous time model.

- This true density is unknown, however, so using it for ML estimation is not possible. If one estimates the discrete time version treating it as the actual density, there is an approximation misspecification that causes the estimates

to be inconsistent: we're not doing ML, we're doing quasi-ML, which is in general an inconsistent estimator.

- Estimation of parameters can be done using simulation-based estimation, discussed in following chapter. A means of learning about spot volatility, $h_t$, given estimated parameters and the history of observable variables, is discussed in the chapter on nonparametric inference, in Section 20.3.

## 17.5 State space models

For Kalman filtering, see Hamilton, *Time Series Analysis,* Chapter 13 and Mikusheva's MIT Open Courseware notes, lectures 21 and 22. A tutorial with Julia code is here: Quantitative Economics Kalman filter. Another source is the summary in the introduction of Lopes and Tsay (2011).

For nonlinear state space models, or non-Gaussian state space models, the basic Kalman filter cannot be used, and the particle filter is becoming a widely-used means of computing the likelihood. This is a fairly new, computationally demanding technique, and is currently (this was written in 2013) an active area of research. See Lopes and Tsay (2011) for a review. Papers by Fernández-Villaverde and Rubio-Ramírez provide interesting and reasonably accessible applications in the context of estimating macroeconomic (DSGE) models.

Here's a document that gives a nice summary of Kalman filtering to compute the likelihood function of a state space model: R.D. Peng time series.

Linear models with Gaussian shocks can often be put into what is known as state-space form. A simple state-space model, using Peng's notation, is

$$
\begin{aligned}
y_t &= A_t x_t + V_t \\
x_t &= B_t x_{t-1} + W_t \\
V_t &\sim N(0, S) \\
W_t &\sim N(0, R)
\end{aligned}
$$

- The first equation is called the measurement equation, and the second is the state equation.
- The variables $y_t$ are observed at time $t$.
- The state variables may or may not be observed at time $t$ (that is, some or all of them may be latent).
- The shocks, $V_t$ and $W_t$ are vector white noise, and are uncorrelated with each other at all lags.

**Example: MA(1)**

As we saw above, for an MA model, we can compute the likelihood function using eqn. 17.3. That form of writing the likelihood uses a lot of computer memory, as the entire $\Sigma$ matrix must be stored. A more efficient method is to write the MA model as a linear Gaussian state-space model, and to use Kalman filtering to compute the likelihood.

The MA(1) model

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1}$$

can be put into state space form, using

$$x_t = \begin{bmatrix} \epsilon_t \\ \epsilon_{t-1} \end{bmatrix}, \; A_t = \begin{bmatrix} 1 & \theta_1 \end{bmatrix}, \; W_t = 0, \; B_t = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \; V_t = \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix}$$

- many more complicated linear models, including ARMA(p,q) models, for example, can be put into this (or a slightly extended) form.
- Adding exogenous variables is straightforward.

Using Kalman filtering, one can compute the likelihood by computing it as the product of the conditional densities of the observed variables:

$$f(y_t|y_{t-1}, y_{t-1}, ..., y_1)$$

starting with $f(y_1)$, then $f(y_2|y_1)$, and working forward to the last measurement observation, $y_t$.

- It turns out that marginal and conditional densities for joint normal random variables are also normal.
  - We use marginalization to get rid of the unobserved state variables
  - conditioning is used to update densities as information arrives in the form of measurement variables
- So, if we can keep track of conditional means and variances, we can compute all of the conditional densities, starting with $f(y_1)$,then getting $f(y_2|y_1)$, and working up to the last.

- There are convenient formulae for updating conditional means and variances, as new information (the measurement variables) arrives.
- All of the densities depend on the parameters, of course. We will estimate those by maximum likelihood.
- See the other references for more information on how the recursive updating works.
- This is easy to program, and is fast, as all equations are closed form. So, this is often the way that likelihoods are computed in practice, for models that can be put into state space form.
- For this to work, the model must be linear, and the shocks must be Gaussian. Without these conditions, other methods, more complicated and computationally demanding, must be used. So, there is a natural tendency to try to fit models into the state-space form, even if they don't actually fit exactly.

An example would be working with a linearized model. That can lead to inconsistency, as we saw.... (harp, harp).

## 17.6 Nonstationarity and cointegration

I'm going to follow Karl Whelan's notes, which are available at Whelan notes.

- A Gretl script file which generates data following the random walk with drift example is Drift.inp .
- A similar script generates random walks without drift. RandomWalks.inp
- There is also one for a deterministic time trend. Trend.inp
- run the script to generate data. Then set the data set structure to time series.
- do a time series plot of the y and x series
- run an OLS of y on x

## $CO_2$ concentration over time.

- search for "Keeling-Whorf"
- Here's Keeling and Whorf data on $CO_2$ concentrations at Mona Loa.
  - run model with constant and trend, observe residuals
  - add monthly indicators for the seasonality: looks at fit and t-stats
  - form of $X'X/T$
  - do model in differences

## 17.7 Exercises

1. Use Matlab/Octave to estimate the same GARCH(1,1) model as in the GarchExample.jl script provided above. Also, estimate an ARCH(4) model for the same data. If unconstrained estimation does not satisfy stationarity restrictions, then do contrained estimation. Compare likelihood values. Which of the two models do you prefer? But do the models have the same number of parameters? Find out what is the "consistent Akaike information criterion" or the "Bayes information criterion" and what they are used for. Compute one or the other, or both, and discuss what they tell you about selecting between the two models.

2. Use Gretl to estimate (by ML) the same Garch(1,1) model as in the previous problem using the nysewk.gdt data set. Do you get the same parameter estimates?

3. Consider the data generating process

$$Y_t = \alpha t + u_t$$
$$X_t = \gamma t + v_t$$

where $t = 1, 2, ..., T$, and $u_t$ and $v_t$ are two independent $N(0, 1)$ white noise processes. That is, both $Y_t$ and $X_t$ follow time trends, each with a different white noise shock. Consider the regression model *without constant* $Y_t = \beta X_t + \epsilon_t$.

(a) To what value does the OLS estimator of $\beta$ tend? Prove your result. Hints: recall that $\frac{1}{T^2} \Sigma_{t=1}^{T} t \to \frac{1}{2}$ and $\frac{1}{T^3} \Sigma_{t=1}^{T} t^2 \to \frac{1}{3}$. Also, the formula for the OLS estimator when there is no constant is $\hat{\beta} = \frac{\Sigma_t Y_t X_t}{\Sigma_t X_t^2}$.

(b) Does this regression model exhibit the *spurious regression* problem? Explain why or why not.

4. Write a Matlab/Julia/your favorite package script that generates two independent random walks, $x_t = x_{t-1} + u_t$ and $y_t = y_{t-1} + u_t$, where the initial conditions are $x_0 = 0$ and $y_0 = 0$, and the two errors are both iid N(0,1). Use a sample size of 1000: $t = 1, 2, ..., 1000$.

   (a) regress y upon x and a constant.

   (b) discuss your findings, especially the slope coefficient, the t statistic of the slope, and $R^2$. Are the findings sensible, given that we know that $x$ has nothing to do with $y$?

   (c) compute the variance of $y_t$ and $x_t$ conditional on the initial conditions $y_0 = 0$ and $x_0 = 0$. Does the variance depend on $t$?

   (d) which of the assumptions of the classical linear regression model are not satisfied by this data generating process?

   (e) present estimation results using transformation(s) of y and/or x so that the regression using the transformed variables confirms that there is

no relationship between the variables. Explain why the transformation(s) you use are successful in eliminating the problem of a spurious relationship.

# Chapter 18

# Bayesian methods

This chapter provides a brief introduction to Bayesian methods, which form a large part of econometric research, especially in the last two decades. Advances in computational methods (e.g., MCMC, particle filtering), combined with practical advantages of Bayesian methods (e.g., no need for minimization and improved identification coming from the prior) have contributed to the popularity of this approach. References I have used to prepare these notes: Cameron and

Trivedi (2005), Chapter 13; Chernozhukov and Hong (2003); Gallant and Tauchen (1996),Gallant and Tauchen (2010); Hoogerheide, van Dijk and van Oest (2007) "Simulation Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances". You might also like to read Mikusheva's MIT OpenCourseWare notes, lectures 23-26: Bayesian notes.

## 18.1 Definitions

The Bayesian approach summarizes beliefs about parameters using a density function:

- There is a true unknown parameter vector, $\theta_0$, and the density $\pi(\theta)$, which is known as the *prior*, reflects current beliefs about the parameter, before observing the sample. It is assumed that the econometrician can provide this density.

- We also have sample information, $y$=$\{y_1, y_2, ...y_n\}$. We're already familiar with the *likelihood function,* $f(y|\theta)$, which is the density of the sample given a parameter value.

Given these two pieces, we can write the joint density of the sample and the beliefs:

$$f(y, \theta) = f(y|\theta)\pi(\theta)$$

We can get the *marginal likelihood* by integrating out the parameter, integrating over its support $\Theta$:

$$f(y) = \int_{\Theta} f(y, \theta) d\theta$$

The last step is to get the *posterior* of the parameter. This is simply the density of the parameter conditional on the sample, and we get it in the normal way we get a conditional density, using Bayes' theorem:

$$f(\theta|y) = \frac{f(y, \theta)}{f(y)} = \frac{f(y|\theta)\pi(\theta)}{f(y)}$$

- The movement from the prior to the posterior reflects the learning that occurs about the parameter when one receives the sample information.
- The sources of information used to make the posterior are the prior and the likelihood function.
- Once we have the posterior, one can provide a complete probabilistic de-

scription about our updated beliefs about the parameter, using quantiles or moments of the posterior.

- The posterior mean or median provide the Bayesian analogue of the frequentist point estimator, in the form of the ML estimator.
- One can show that these point estimators converge to the true $\theta_0$.
- We can define regions analogous to confidence intervals by using quantiles of the posterior, or the marginal posterior.

So far, this is pretty straightforward. The complications are mostly computational. To illustrate, the posterior mean is

$$\begin{aligned} E(\theta|y) &= \int_\Theta \theta f(\theta|y) d\theta \\ &= \frac{\int_\Theta \theta f(y|\theta)\pi(\theta) d\theta}{\int_\Theta f(y,\theta) d\theta} \end{aligned}$$

- One can see that a means of integrating will be needed.
- note that $\theta$ is a vector, so that the integrals are multiple
- Only in very special cases will the integrals have analytic solutions.
- Otherwise, computational methods will be needed. Advances in computational methods are what have lead to the increased use of Bayesian methods.

## 18.2 Philosophy, etc.

So, the classical paradigm views the data as generated by a data generating process, which is a perhaps unknown model characterized by a parameter vector, and the data is generated from the model at a particular value of the parameter vector, $\theta_0$. Bayesians view data as given, and update beliefs about a parameter using the information about the parameter contained in the data. There's nothing obviously contradictory in these views. Nevertheless, it's not hard to find discussions where there are disagreements.

Here, I'm trying to address a model with a fixed non-random parameter about which we would like to learn. As long as the object of interest - the dgp and it's parameter - is agreed upon, then we can contemplate using any convenient tool.

Even if one is a strict frequentist, one shouldn't reinvent the wheel each time we get a new sample: previous samples have information about the parameter, and we should use all of the available information. A pure frequentist "full information" approach would require writing the joint likelihood of all samples, which would al-

most certainly constitute an impossible task. The Bayesian approach concentrates all of the information coming from previous work in the form of a prior. A fairly simple, easy to use prior may not *exactly* capture all previous information, but it could offer a handy and reasonably accurate summary, and it's almost certainly better than simply pretending that all of that previous information simply doesn't exist. So, the idea of a prior as a summary of what we have learned may simply be viewed as a practical solution to the problem of using all the available information. Given that it's a summary, one may as well use a convenient form, as long as it's plausible and the results don't depend too exaggeratedly upon particular form used.

As long as one takes the view that there is a fixed unknown parameter value $\theta_0$ which generates all samples, then frequentist and Bayesian methods are trying to inform us about the same object, and the choice between tools may become one of convenience. It turns out that one can analyze Bayesian estimators from a classical (frequentist) perspective. It also turns out that Bayesian estimators

may be easier to compute reliably than analogous classical estimators. These computational advantages, combined with the ability to use information from previous work in an intelligent way, make the study of Bayesian methods attractive for frequentists. If a Bayesian takes the view that there is a fixed data generating process, and Bayesian learning leads in the limit to the same fixed true value that frequentists posit, then the study of frequentist theory will be useful to a Bayesian practitioner. For example, the GMM estimator is closely related to some versions of Approximate Bayesian Computing (ABC). Thus, knowledge of theory and practical experience with GMM can be a useful guide to implementing ABC estimators.

- For the rest of this, I will adopt the classical, frequentist perspective, and study the behavior of Bayesian estimators in this context.

- One should note that the traditional Bayesian approach requires the likelihood function, just as is the case with ML. Thus, it uses *strong assumptions*, for a given model.

- There are Bayesian methods for choosing between models, which we will not get into.

- There are also recent Bayesian-inspired methods that attempt to work without knowledge of the likelihood function. For instance, Chernozhukov and Hong (2003) use Bayesian methods to compute a GMM estimator. Some such methods, e.g. Approximate Bayesian Computing require the model to be simulable, in which case, essentially the same strong assumptions as underlie ML are being used.

## 18.3 Example

Suppose data is generated by i.i.d. sampling from an exponential distribution with mean $\theta$. An exponential random variable takes values on the positive real numbers. Waiting times are often modeled using the exponential distribution.

- The density of a typical sample element is $f(y|\theta) = \frac{1}{\theta}e^{-y/\theta}$. The likelihood is simply the product of the sample contributions.

- Suppose the prior for $\theta$ is $\theta \sim \text{lognormal}(1, 1)$. This means that the logarithm of $\theta$ is normally distributed with mean 1 and standard deviation 1. We use a lognormal prior because it enforces the requirement that the parameter of the exponential density be positive.

- The Julia script BayesExample1.jl implements Bayesian estimation for this setup.

With a sample of 10 observations, we obtain the results in panel (a) of Figure 18.1,

while with a sample of size 50 we obtain the results in panel (b). Note how the posterior is more concentrated around the true parameter value in panel (b). Also note how the posterior mean is closer to the prior mean when the sample is small. When the sample is small, the likelihood function has less weight, and more of the information comes from the prior. When the sample is larger, the likelihood function will have more weight, and its effect will dominate the prior's.

Figure 18.1: Bayesian estimation, exponential likelihood, lognormal prior

(a) N=10 (b) N=50

## 18.4 Theory

Chernozhukov and Hong (2003) "An MCMC Approach to Classical Estimation" http://www.sciencedirect.com/science/article/pii/S0304407603001003 is a very interesting article that shows how Bayesian methods may be used with criterion functions that are associated with classical estimation techniques. For example, it is possible to compute a posterior mean version of a GMM estimator. Chernozhukov and Hong provide their Theorem 2, which proves consistency and

Figure 18.2: Chernozhukov and Hong, Theorem 2

**Theorem 2** (LTE in large samples). *Under Assumptions* 1–4,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \xi_{J_n(\theta_0)} + U_n + \mathrm{o}_p(1), \quad \Omega_n^{-1/2}(\theta_0) J_n(\theta_0) U_n \to_{\mathrm{d}} \mathcal{N}(0, I).$$

*Hence*

$$\Omega_n^{-1/2}(\theta_0) J_n(\theta_0)(\sqrt{n}(\hat{\theta} - \theta_0) - \xi_{J_n(\theta_0)}) \to_{\mathrm{d}} \mathcal{N}(0, I).$$

*If the loss function* $\rho_n$ *is symmetric, i.e.* $\rho_n(h) = \rho_n(-h)$ *for all* $h$, $\xi_{J_n(\theta_0)} = 0$ *for each* $n$.

asymptotic normality for a general class of such estimators. When the criterion function $L_n(\theta)$ in their paper is set to the log-likelihood function, the pseudo-prior $\pi(\theta)$ is a real Bayesian prior, and the penalty function $\rho_n$ is the squared loss function (see the paper), then the class of estimators discussed by CH reduces to the ordinary Bayesian posterior mean. As such, their Theorem 2, in Figure 18.2 tells us that this estimator is consistent and asymptotically normally distributed. In particular, the Bayesian posterior mean has the same asymptotic distribution as does the ordinary maximum likelihood estimator.

- the intuition is clear: as the amount of information coming from the sample increases, the likelihood function brings an increasing amount of information, relative to the prior. Eventually, the prior is no longer important for determining the shape of the posterior.

- when the sample is large, the shape of the posterior depends on the likelihood function. The likelihood function collapses around $\theta_0$ when the sample is generated at $\theta_0$. The same is true of the posterior, it narrows around $\theta_0$. This causes the posterior mean to converge to the true parameter value. In fact, all quantiles of the posterior converge to $\theta_0$. Chernozhukov and Hong discuss estimators defined using quantiles.

- For an econometrician coming from the frequentist perspective, this is attractive. The Bayesian estimator has the same asymptotic behavior as the MLE. There may be computational advantages to using the Bayesian approach, because there is no need for optimization. If the objective function that defines the classical estimator is irregular (multiple local optima, nondifferentiabil-

ities, noncontinuities...), then optimization may be very difficult. However, Bayesian methods that use integration may be more tractable. This is the main motivation of CH's paper.

- Additional advantages include the benefits if an informative prior is available. When this is the case, the Bayesian estimator can have better small sample performance than the maximum likelihood estimator. The ML estimator uses only the information in the sample. Adding outside information can help.

## 18.5 Computational methods

- To compute the posterior mean, we need to evaluate

$$\begin{aligned} E(\theta|y) &= \int_\Theta \theta f(\theta|y) d\theta \\ &= \frac{\int_\Theta \theta f(y|\theta)\pi(\theta) d\theta}{\int_\Theta f(y,\theta) d\theta}. \end{aligned}$$

- Note that both of the integrals are multiple integrals, with the dimension given by that of the parameter, $\theta$.

- Under some special circumstances (conjugate priors), the integrals may have analytic solutions: e.g., Gaussian likelihood with a Gaussian prior leads to a Gaussian posterior.

- When the dimension of the parameter is low, quadrature methods may be used. What was done in BayesExample1.jl is an unsophisticated example of this. More sophisticated methods use an intelligently chosen grid to re-

duce the number of function evaluations. Still, these methods only work for dimensions up to 3 or so.

- Otherwise, some form of simulation-based "Monte Carlo" integration must be used. The basic idea is that $E(\theta|y)$ can be approximated by $(1/S)\,\Sigma_{s=1}^{S}\,\theta^s$, where $\theta^s$ is a random draw from the posterior distribution $f(\theta|y)$. The trick is *how to make draws from the posterior* when in general we can't compute the posterior.

  - the law of large numbers tells us that this average will converge to the desired expectation as $S$ gets large
  - convergence will be more rapid if the random draws are independent of one another, but insisting on independence may have computational drawbacks.

Monte Carlo methods include:

- importance sampling
- Markov chain Monte Carlo (MCMC)
- sequential Monte Carlo (SMC, also known as particle filtering)
- Hamiltonian Monte Carlo (HMC).
- The great expansion of these methods over the years has caused Bayesian econometrics to become much more widely used than it was in the not so distant (for some of us) past. There is much literature - here we will only look at a basic example that captures the main ideas.

# MCMC

Variants of Markov chain Monte Carlo have become a very widely used means of computing Bayesian estimates. See Tierney (1994) "Markov Chains for Exploring Posterior Distributions" *Annals of Statistics* and Chib and Greenberg (1995) "Understanding the Metropolis-Hastings algorithm" *The American Statistician.*

Let's consider the basic **Metropolis-Hastings MCMC** algorithm. This is good to get the basic idea of Monte Carlo methods, but it is often not the best choice in practice.

We will generate a long realization of a Markov chain process for $\theta$, as follows:

- The prior density is $\pi(\theta)$, as above.

- Let $g(\theta^*; \theta^s)$ be a proposal density, which describes the density of a trial value $\theta^*$ conditional on starting at $\theta^s$. It must be possible to sample from the proposal. This gives a new trial parameter value $\theta^*$, given the most

recently accepted parameter value $\theta^s$. A proposal will be accepted if

$$\frac{f(\theta^*|y)}{f(\theta^s|y)}\frac{g(\theta^s;\theta^*)}{g(\theta^*;\theta^s)} > \alpha$$

where $\alpha$ is a $U(0,1)$ random variate.

There are two parts to the numerator and denominator: the posterior, and the proposal density.

- Focusing on the numerator, when the trial value of the proposal has a higher posterior, acceptance is favored.

- The other factor is the density associated with returning to $\theta^s$ (where we are) when starting at $\theta^*$ (where we will be if the proposal is accepted), which has to do with the reversibility of the Markov chain. If this is too low, acceptance is not favored. We don't want to jump to a new region if we will never get back, as we need to sample from the entire support of the posterior. Falling into a region we can never get out of will not let us explore the whole support.

- The two together mean that we will jump to a new area only if we are able to eventually jump back with a reasonably high probability. The probability of jumping is higher when the new area has a higher posterior density, but lower if it's hard to get back.

- The idea is to sample from all regions of the posterior, those with high and low density, sampling more heavily from regions of high density. We want to go occasionally to regions of low density, but it is important not to get stuck there.

- Consider a bimodal density: we want to explore the area around both modes. To be able to do that, it is important that the proposal density allows us to be able to jump between modes.

- Understanding in detail why this makes sense is the tricky and elegant part of the theory, see the references for more information.

- Note that the ratio of posteriors is equal to the ratio of likelihoods times the ratio of priors:

$$\frac{f(\theta^*|y)}{f(\theta^s|y)} = \frac{f(y|\theta^*)}{f(y|\theta^s)}\frac{\pi(\theta^*)}{\pi(\theta^s)}$$

  because the marginal likelihood $f(y)$ is the same in both cases. We don't need to compute that integral! We don't need to know the posterior, either. The acceptance criterion can be written as: accept if

$$\frac{f(y|\theta^*)}{f(y|\theta^s)}\frac{\pi(\theta^*)}{\pi(\theta^s)}\frac{g(\theta^s;\theta^*)}{g(\theta^*;\theta^s)} > \alpha$$

  otherwise, reject

- From this, we see that the information needed to determine if a proposal is accepted or rejected is the prior, the proposal density, and the likelihood function $f(y|\theta)$.

  - in principle, the prior is non-negotiable. In practice, people often chose

priors with convenience in mind

  - the likelihood function is what it is, it comes from the model
  - the place where artistry comes to bear is the choice of the proposal density

- when the proposal density is *symmetric,* so that $g(\theta^s;\theta^*) = g(\theta^*;\theta^s)$, the acceptance criterion simplifies to

$$\frac{f(y|\theta^*)}{f(y|\theta^s)}\frac{\pi(\theta^*)}{\pi(\theta^s)} > \alpha$$

A random walk proposal, where the trial value is the current value plus a shock that doesn't depend on the current value, satisfies symmetry.

- the steps are:

1. the algorithm is initialized at some $\theta^1$
2. for $s = 2, ..., S$,
   (a) draw $\theta^*$ from $g(\theta^*; \theta^s)$
   (b) according to the acceptance/rejection criterion, if the result is acceptance, set $\theta^{s+1} = \theta^*$, otherwise set $\theta^{s+1} = \theta^s$
   (c) iterate

- Once the chain is considered to have stabilized, say at iteration $r$, the values of $\theta^s$ for $s > r$ are taken to be draws from the posterior. The posterior mean is computed as the simple average of the value. Quantiles, etc., can be computed in the appropriate fashion.

- the art of applying these methods consists of providing a good proposal density so that the acceptance rate is reasonably high, but not too high. There is a vast literature on this, and the vastness of the literature should serve as a warning that getting this to work in practice is not necessarily a simple matter. If it were, there would be fewer papers on the topic. A good proposal is one that looks like the posterior, because we are trying to sample from the posterior. However, we don't know the posterior, which is why we are doing MCMC in the first place....

  - too high acceptance rate: this is usually due to a proposal density that gives proposals very close to the current value, e.g, a random walk with

very low variance. This means that the posterior is being explored inefficiently, we travel around through the support at a very low rate, which means the chain will have to run for a (very, very...) long time to do a thorough exploration.

- too low acceptance rate: this means that the steps are too large, and we attempt to move to low posterior density regions too frequently. The chain will become highly autocorrelated, as it stays in the same place due to rejections, so long periods convey little additional information relative to a subset of the values in the interval
- different ways to construct proposals forms a large part of the MCMC literature.
- the newer Hamiltonian Monte Carlo method has the advantage that one does not need to provide a proposal. However, the likelihood has to be differentiable to use HMC.

- look at Geoff Gordon's mh.h

## 18.6 Examples

### MCMC for the simple example

The simple exponential example with log-normal prior can be implemented using MH MCMC, and this is done in the Julia script BayesExample2.jl . Play around with the sample size and the tuning parameter, and note the effects on the computed posterior mean and on the acceptance rate. An example of output is given in Figure 18.3, which shows the final draws of the chain, and the posterior density (computed using non-parametric density estimation, more on that later). In that Figure, the chain is probably too spiky: too many draws are being accepted (it's around 0.6, which you'll see if you run the code), meaning that the tuning parameter needs to be increased, to lower the acceptance rate. If you increase the sample size, you'll see how the posterior concentrates around the true value, 3.

An example of estimating this same simple exponential model, but using a specialized package (Turing.jl) and a more sophisticated sampling method (NUTS:

Figure 18.3: Metropolis-Hastings MCMC, exponential likelihood, lognormal prior

(a) chain, last 1000 draws

(b) posterior

no U-turn Hamiltonian Monte Carlo) that does not require that a proposal be specified is BayesExample3.jl .

## 18.7 Full sample Bayesian estimation of the DSGE model

In Section 15.9, a simple DSGE model was estimated by ML. EstimateCGHK_Bayes.m is a script which estimates the same model, using Bayesian methods, with MCMC or particle filtering. Adjust the .mod file mentioned in the script to change options. Run it in Octave/Dynare using `dynare CKmcmc.mod.` We can obtain the results in Figure 18.4.

Some conclusions we can draw:

- Estimation of the parameters is quite good, when order=1 (linearized, which allows for Kalman filtering), The first panel uses the variables c and n, and the second uses two others (forgot which). Note that the results for the parameter $\gamma$ change quite a bit, depending on which observable variables are used. This is probably due to the stochastic singularity problem, when a first order solution is used. When we don't know the true parameter values,

how do we choose which results to use?

- The first panel gives results that are substantially similar to those obtained in the Section on ML estimation, which also used c and n as the observables. This is not surprising, it just means that the sample is large enough so that the prior does not have a large impact on the posterior.

- One still has the stochastic singularity problem, and the results that are obtained will depend on which variables are selected for estimation.

- The GMM results (see 16.14) are substantially closer to the true parameter values, for the parameters $\gamma$ and $\rho_2$, but the estimated standard error for $\rho_2$ is larger than what we get from ML and MCMC. GMM does use all of the observable variables: c, n, y, r, w. Perhaps the use of all variables improves the results in some ways. Remember, all of this is for just one sample, so we don't want to take any firm conclusions, it's just an observation. It would not be hard to perform a Monte Carlo study to see what we can actually say

about the performance of the various methods....

- MCMC can be a little time consuming, and estimation by particle filtering (set order=2) will take a looong time...

- Figure 18.5 plots the priors and posteriors. Note that the posterior is substantially different than the prior: we learn a lot from the sample. That's why MCMC and ML are substantially similar.

For tips on using Dynare for MCMC estimation, see these notes by Wouter den Haan.

Figure 18.4: MCMC results for simple DSGE example model (two different runs using different observed variables), plus the GMM results for comparison

MCMC, ver. 1

```
ESTIMATION RESULTS

Log data density is 1201.836940.

parameters
          prior mean   post. mean        90% HPD interval    prior       pstdev

betta         0.975       0.9900      0.9859      0.9938    unif       0.0144
gam           2.500       1.8805      1.4102      2.2771    unif       1.4434
rho1          0.500       0.9047      0.8873      0.9209    unif       0.2887
sigma1        0.050       0.0186      0.0150      0.0223    unif       0.0289
rho2          0.500       0.7929      0.7329      0.8550    unif       0.2887
sigma2        0.050       0.0128      0.0104      0.0150    unif       0.0289
nss           0.312       0.3347      0.3309      0.3389    unif       0.0361
Total computing time : 0h02m32s
Note: warning(s) encountered in MATLAB/Octave code
octave:2>
```

MCMC, ver. 2

```
ESTIMATION RESULTS

Log data density is 791.328414.

parameters
          prior mean   post. mean        90% HPD interval    prior       pstdev

betta         0.975       0.9922      0.9885      0.9956    unif       0.0144
gam           2.500       2.1686      1.2519      2.9315    unif       1.4434
rho1          0.500       0.9039      0.8765      0.9312    unif       0.2887
sigma1        0.050       0.0203      0.0186      0.0225    unif       0.0289
rho2          0.500       0.8234      0.7368      0.9035    unif       0.2887
```

Figure 18.5: CGHK model, posteriors

## 18.8 Bayesian GMM for the DSGE model

A script which shows how to do Bayesian GMM as proposed by Chernozhukov and Hong (2003) is DoMCMC.jl . The issue of how to come up with an effective proposal density is always important when doing MH MCMC. Examine the code to see what was done, which perhaps could be improved. Once we have the chain, it can be used to compute posterior densities for the parameters, for example, the estimated posterior for $\gamma$ follows. Recall that the true value that generated the sample is $\gamma = 2$, so, for this sample, the method worked reasonably well for point estimation.

q05
90% non-rejection region
q95
mean
median
20
15
10
5
0
1.95
2.00
2.05
2.10

The results for all parameters follow. The point estimates (mean and 50% quantile) are very close to the true values. The true values of most parameters are close to the center of the 95% confidence intervals, with $\beta$ and nss being the exceptions. A notable difference with respect to extremum-based GMM, which computes confidence intervals using asymptotic theory, is that the confidence intervals for $\gamma$ and $\rho_2$ are much tighter here.

Summary Statistics

| parameters | mean | std | naive_se | mcse | ess | rhat |
|---|---|---|---|---|---|---|
| Symbol | Float64 | Float64 | Float64 | Float64 | Float64 | Float64 |
| $\beta$ | 0.9903 | 0.0001 | 0.0000 | 0.0000 | 868.2780 | 1.0004 |
| $\gamma$ | 2.0169 | 0.0176 | 0.0001 | 0.0004 | 1203.1861 | 1.0000 |
| $\rho_1$ | 0.9013 | 0.0019 | 0.0000 | 0.0000 | 1615.3781 | 1.0001 |
| $\sigma_1$ | 0.0193 | 0.0008 | 0.0000 | 0.0000 | 1317.3935 | 1.0000 |
| $\rho_2$ | 0.7134 | 0.0188 | 0.0001 | 0.0004 | 894.3037 | 1.0020 |
| $\sigma_2$ | 0.0105 | 0.0006 | 0.0000 | 0.0000 | 1382.1438 | 1.0003 |
| nss | 0.3346 | 0.0005 | 0.0000 | 0.0000 | 930.8899 | 1.0003 |

Quantiles

| parameters | 2.5% | 25.0% | 50.0% | 75.0% | 97.5% |
|---|---|---|---|---|---|
| Symbol | Float64 | Float64 | Float64 | Float64 | Float64 |
| $\beta$ | 0.9900 | 0.9902 | 0.9903 | 0.9904 | 0.9905 |
| $\gamma$ | 1.9820 | 2.0052 | 2.0169 | 2.0287 | 2.0513 |
| $\rho_1$ | 0.8975 | 0.9001 | 0.9014 | 0.9026 | 0.9051 |
| $\sigma_1$ | 0.0176 | 0.0187 | 0.0193 | 0.0198 | 0.0209 |
| $\rho_2$ | 0.6772 | 0.7009 | 0.7130 | 0.7255 | 0.7523 |
| $\sigma_2$ | 0.0093 | 0.0101 | 0.0105 | 0.0109 | 0.0116 |
| nss | 0.3336 | 0.3343 | 0.3346 | 0.3350 | 0.3357 |

For reference, the results from ordinary CUE-GMM (extremum estimator), from Section 16.14 are in the panel above. Note that the point estimates are very close to the Bayesian GMM point estimates, but the standard deviations and confidence intervals are quite different. The standard errors from the asymptotics are considerably larger than what we get from the Bayesian version. Which of the two versions of confidence intervals are more accurate is still an open question. We will return to this issue later.

Here is a summary of the chain, using the very nice MCMCChains.jl package.

β
Iteration
Density
Sample value
γ
ρ₁
σ₁
ρ₂
σ₂
nss

## 18.9 Exercises

1. Experiment with the examples to learn about tuning, etc.

# Chapter 19

# Introduction to panel data

References: Cameron and Trivedi (2005), Part V, Chapters 21 and 22 (plus 23 if you have special interest in the topic) and a basic discussion in Stock and Watson, *Introduction to Econometrics*, Chapter 10. The GRETL manual also has two chapters, which are a nice reference.

In this chapter we'll look at panel data. Panel data is an important area in applied econometrics, simply because much of the available data has this structure.

Also, it provides an example where things we've already studied (GLS, endogeneity, GMM, Hausman test) come into play. There has been much work in this area, and the intention is not to give a complete overview, but rather to highlight the issues and see how the tools we have studied can be applied.

## 19.1 Generalities

Panel data combines cross sectional and time series data: we have a time series for each of the agents observed in a cross section.

- The addition of temporal information to a cross sectional model can in principle allow us to investigate issues such as persistence, habit formation, and dynamics.

- Starting from the perspective of a single time series, the addition of cross-sectional information allows investigation of heterogeneity.

- In both cases, if parameters are common across units or over time, the additional data allows for more precise estimation. This is simply an example of estimation subject to restrictions, which improves efficiency *if the restrictions are correct*

The basic idea is to allow variables to have two indices, $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$. The simple linear model

$$y_i = \alpha + x_i\beta + \epsilon_i$$

becomes

$$y_{it} = \alpha + x_{it}\beta + \epsilon_{it}$$

We could think of allowing the parameters to change over time and over cross sectional units. This would give

$$y_{it} = \alpha_{it} + x_{it}\beta_{it} + \epsilon_{it}$$

The problem here is that there are more parameters than observations, so the model is not identified. We need some restraint! The proper restrictions to use of course depend on the problem at hand, and a single model is unlikely to be appropriate for all situations. For example, one could have time and cross-sectional

indicator variables, and slopes that are constant over time and across agents:

$$y_{it} = \alpha_i + \gamma_t + x_{it}\beta + \epsilon_{it}$$

(in class, write out the form of the model in matrix notation).

There is a lot of room for playing around here. We also need to consider whether or not $n$ and $T$ are fixed or growing. We'll need at least one of them to be growing in order to do asymptotics.

To provide some focus, we'll consider common slope parameters, but agent-specific intercepts, which:

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it} \tag{19.1}$$

- I will refer to this as the "simple linear panel model". We assume that the regressors $x_{it}$ are exogenous, with no correlation with the error term.

- This is the model most often encountered in the applied literature. It is like the original cross-sectional model, in that the $\beta's$ are constant over time for all $i$. However we're now allowing for the constant to vary across $i$ (some individual heterogeneity).

- We can consider what happens as $n \to \infty$ but $T$ is fixed. This would be relevant for microeconometric panels, (e.g., the PSID data) where a survey of a large number of individuals may be done for a limited number of time periods.

- Macroeconometric applications might look at longer time series for a small number of cross-sectional units (e.g., 40 years of quarterly data for 15 European countries). For that case, we could keep $n$ fixed (seems appropriate when dealing with the EU countries), and do asymptotics as $T$ increases, as is normal for time series.

- The asymptotic results depend on how we do this, of course.

**Why bother using panel data, what are the benefits?**

The model

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$$

is a restricted version of

$$y_{it} = \alpha_i + x_{it}\beta_i + \epsilon_{it}$$

which could be estimated for each $i$ in turn, using time series data. Why use the panel approach?

- Because the restrictions that $\beta_i = \beta_j = ... = \beta$, if true, lead to more efficient estimation. Estimation for each $i$ in turn will be very uninformative if $T$ is small.

- Another reason is that panel data allows us to estimate parameters that are not identified by cross sectional (time series) data.

- For example, if the model is

$$y_{it} = \alpha_i + \gamma_t + x_{it}\beta + \epsilon_{it}$$

  and we have only cross sectional data, we cannot estimate the $\alpha_i$.

- If we have only time series data on a single cross sectional unit $i = 1$, we cannot estimate the $\gamma_t$.

- Cross-sectional variation allows us to estimate parameters indexed by time, and time series variation allows us to estimate parameters indexed by cross-sectional unit. Parameters indexed by both $i$ and $t$ will require other forms of restrictions in order to be estimable.

- A **very important reason** is that $\alpha_i$ can absorb any missing variables in the regression that don't change over time, and $\gamma_t$ can absorb missing variables that don't change across $i$. For example, suppose we have the model

$$y_{it} = \alpha + x_{it}\beta + z_i\gamma + \epsilon_{it} \tag{19.2}$$

where the variables in $z_i$ are unobserved, but are <u>constant over time</u>.

- Assume that, as is <u>usually the case</u>, there is some correlation between the variables in $x_{it}$ and $z_i$. That is to say, there is some ordinary collinearity of the regressors.

- If we have only one time period, then we have to estimate the model

$$y_i = \alpha + x_i\beta + z_i\gamma + \epsilon_i$$

using the observations $i = 1, 2, ..., n$,

- Because $z_i$ is unobserved, we have to let it be absorbed in the error term. For convenience, and to keep the notation simple, assume that the mean of $z_i\gamma$ is zero (this does not affect the argument in any important way), so the model we can actually estimate is

$$y_i = \alpha + x_i\beta + v_i$$

  where $v_i = z_i\gamma + \epsilon_i$.

- This model has correlation between the regressors and the error, so the OLS estimates would be inconsistent. Furthermore, we don't have any natural instruments to estimate the model by IV.

- However, **suppose we have at least two time periods** of data, and $n$ cross-sectional observations. Then, we can let $z_i\gamma$ move into the constant, and we get the model

$$y_{it} = \alpha + x_{it}\beta + z_i\gamma + \epsilon_{it}$$
$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$$

  where $\alpha_i = \alpha + z_i\gamma$. This is the simple linear panel data model.

  - Notice that the problematic $z_i$ have now disappeared!
  - It turns out that OLS estimation of this model will give consistent estimates of the $\beta$ parameters, as the cross sectional size of the sample, $n$ increases, as long as the regressors are exogenous. If it's not clear how this can be estimated by OLS, then consider estimating it using first differences: that model is pretty obviously consistently estimable using OLS.

Returing to panel data in general, **the main issues are:**

- can $\beta$ be estimated consistently? This is almost always a goal.
- can the $\alpha_i$ be estimated consistently? This is often of secondary interest.
- sometimes, we're interested in estimating the distribution of $\alpha_i$ across $i$.
- are the $\alpha_i$ correlated with $x_{it}$? This is very likely the case.
- does the presence of $\alpha_i$ complicate estimation of $\beta$?

what about the covariance structure?

- We're likely to have both HET and AUT, in the original model, so GLS issues will probably be relevant.
  - Potential for efficiency gains
  - need to take care of it to obtain valid standard errors.

## 19.2 Static models and correlations between variables

To begin with, assume that:

- the $x_{it}$ are weakly exogenous variables (uncorrelated with $\epsilon_{it}$)
- the model is static: $x_{it}$ does not contain lags of $y_{it}$.
- then the basic problem we have in the panel data model $y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$ is the presence of the $\alpha_i$. These are individual-specific parameters. Or, possibly more accurately, they can be thought of as individual-specific variables that are not observed (latent variables). The reason for thinking of them as variables is because the agent may choose their values following some process, or may choose other variable taking these ones as given.

Define $\alpha = E(\alpha_i)$, so $E(\alpha_i - \alpha) = 0$, where the expectation is with respect to the density that describes the distribution of the $\alpha_i$ in the population. Our model

$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$ may be written

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}\beta + \epsilon_{it} \\ &= \alpha + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ &= \alpha + x_{it}\beta + \eta_{it} \end{aligned}$$

Note that $E(\eta_{it}) = 0$. A way of thinking about the data generating process is this:

- First, $\alpha_i$ is drawn, from the population density
- then $T$ values of $x_{it}$ are drawn from $f_X(z|\alpha_i)$.
- the important point is that the distribution of $x$ *may vary depending on the realization of* $\alpha_i$.
- For example, if $y$ is the quantity demanded of a luxury good, then a high value of $\alpha_i$ means that agent $i$ will buy a large quantity, on average. This may be possible only when the agent's income is also high. Thus, it may be

possible to draw high values of $\alpha_i$ only when income is also high, otherwise, the budget constraint would be violated. If income is one of the variables in $x_{it}$, then $\alpha_i$ and $x_{it}$ are not independent.

- Another example: consider returns to education, modeling wage as a function of education. $\alpha_i$ could be an individual specific measure of ability. Ability could affect wages, but it could also affect the number of years of education. When education is a regressor and ability is a component of the error, we may expect an endogeneity problem.

- Thus, there may be correlation between $\alpha_i$ and $x_{it}$, in which case $E(x_{it}\eta_{it}) \neq 0$ in the above equation.

  - This means that OLS estimation of the model would lead to biased and inconsistent estimates.

  - However, it is possible (but unlikely for economic data) that $x_{it}$ and $\eta_{it}$ are independent or at least uncorrelated, if the distribution of $x_{it}$

is constant with respect to the realization of $\alpha_i$. In this case OLS estimation would be consistent.

**Fixed effects**: when $E(x_{it}\eta_{it}) \neq 0$, the model is called the "fixed effects model"

**Random effects**: when $E(x_{it}\eta_{it}) = 0$, the model is called the "random effects model".

I find this to be pretty poor nomenclature, because the issue is not whether "effects" are fixed or random (they are always random, unconditional on $i$). The issue is whether or not the "effects" are correlated with the other regressors.

- In economics, it seems likely that the unobserved variable $\alpha$ is probably correlated with the observed regressors, $x$ (this is simply the presence of collinearity between observed and unobserved variables, and collinearity is usually the rule rather than the exception).

- So, we expect that the "fixed effects" model is probably the relevant one unless special circumstances (e.g., if one were working with experimental data) imply that the $\alpha_i$ are uncorrelated with the $x_{it}$.

## 19.3 Estimation of the simple linear panel model

### "Fixed effects": The "within" estimator

How can we estimate the parameters of the simple linear panel model (equation 19.1) and what properties do the estimators have? First, we assume that the $\alpha_i$ are correlated with the $x_{it}$ ("fixed effects" model ). The model can be written as $y_{it} = \alpha + x_{it}\beta + \eta_{it}$, and we have that $E(x_{it}\eta_{it}) \neq 0$. As such, OLS estimation of this model will give biased an inconsistent estimated of the parameters $\alpha$ and $\beta$.

The "within" estimator is a solution. First, go back to the original formulation of the model: $y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$ . The within estimator involves subtracting the

time series average from each cross sectional unit.

$$
\begin{aligned}
\overline{x}_i &= \frac{1}{T}\sum_{t=1}^{T} x_{it} \\
\overline{\epsilon_i} &= \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} \\
\overline{y}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} = \alpha_i + \frac{1}{T}\sum_{t=1}^{T} x_{it}\beta + \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} \\
\overline{y}_i &= \alpha_i + \overline{x}_i\beta + \overline{\epsilon_i} \qquad (19.3)
\end{aligned}
$$

The transformed model is

$$
\begin{aligned}
y_{it} - \overline{y}_i &= \alpha_i + x_{it}\beta + \epsilon_{it} - \alpha_i - \overline{x}_i\beta - \overline{\epsilon_i} \qquad (19.4) \\
y^*_{it} &= x^*_{it}\beta + \epsilon^*_{it}
\end{aligned}
$$

where $x^*_{it} = x_{it} - \overline{x}_i$ and $\epsilon^*_{it} = \epsilon_{it} - \overline{\epsilon}_i$.

- In this model, it is clear that $x^*_{it}$ and $\epsilon^*_{it}$ are uncorrelated, as long as the original regressors $x_{it}$ are *strongly* exogenous with respect to the original error

$\epsilon_{it}$ ($E(x_{it}\epsilon_{is}) = 0$, $\forall t, s$). In this case, OLS will give consistent estimates of the parameters of this model, $\beta$.

**Exercise 86.** Explain why we need strong exogeneity of the $x_{it}$ with respect to $\epsilon_{it}$.

What about the $\alpha_i$? Can they be consistently estimated? An estimator is

$$\hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^{T} \left( y_{it} - x_{it}\hat{\beta} \right)$$

It's fairly obvious that this is a consistent estimator *if* $T \to \infty$. For a short panel with fixed $T$, this estimator is not consistent. Nevertheless, the variation in the $\hat{\alpha}_i$ can be fairly informative about the heterogeneity. A couple of notes:

- an equivalent approach is to estimate the model

$$y_{it} = \sum_{j=1}^{n} d_{j,it}\alpha_j + x_{it}\beta + \epsilon_{it}$$

  by OLS. The $d_j$, $j = 1, 2, ..., n$ are $n$ indicator variables that take on the value 1 if $j = i$, zero otherwise. They are indicators of the cross sectional unit of the observation. For example, with 3 cross sectional units and 3 time

periods, and a single $x$ regressor, the model in matrix form would look like

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & x_{11} \\ 1 & 0 & 0 & x_{12} \\ 1 & 0 & 0 & x_{13} \\ 0 & 1 & 0 & x_{21} \\ 0 & 1 & 0 & x_{22} \\ 0 & 1 & 0 & x_{23} \\ 0 & 0 & 1 & x_{31} \\ 0 & 0 & 1 & x_{32} \\ 0 & 0 & 1 & x_{33} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{bmatrix}$$

Estimating this model directly by OLS gives numerically exactly the same results as the OLS version of the "within" estimator, and you get the $\hat{\alpha}_i$ automatically. See Cameron and Trivedi, section 21.6.4 for details. An interesting and important result known as the Frisch-Waugh-Lovell Theorem

can be used to show that the two means of estimation give identical results.

- This last expression makes it clear why the "within" estimator cannot estimate slope coefficients corresponding to variables that have no time variation. Such variables are perfectly collinear with the cross sectional indicators $d_j$. The corresponding coefficients are not identified.

- OLS estimation of the "within" model is consistent, but probably not efficient, because it is highly probable that the $\epsilon_{it}$ are not iid. There is very likely heteroscedasticity across the $i$ and autocorrelation between the $T$ observations corresponding to a given $i$. *One needs to estimate the covariance matrix of the parameter estimates taking this into account.*

  - at a minimum, robust standard errors will be needed, to be able to get valid standard errors and to be able to test restrictions. We need panel robust standard errors, which are also referred to as cluster robust standard errors.

- It is possible to use GLS corrections if you make assumptions regarding the het. and autocor. Quasi-GLS, using a possibly misspecified model of the error covariance, can lead to more efficient estimates than simple OLS. One can then combine it with subsequent panel-robust covariance estimation to deal with the misspecification of the error covariance, which would invalidate inferences if ignored. The White heteroscedasticity consistent covariance estimator is easily extended to panel data with independence across $i$, but with heteroscedasticity and autocorrelation within $i$, and heteroscedasticity between $i$. See Cameron and Trivedi, Section 21.2.3.

## Estimation with random effects

The original model is

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$$

This can be written as

$$\begin{aligned} y_{it} &= \alpha + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ y_{it} &= \alpha + x_{it}\beta + \eta_{it} \end{aligned} \tag{19.5}$$

Under random effects, the $\alpha_i$ are assumed not to be correlated with the $x_{it}$, so $E(\eta_{it}) = 0$, and $E(x_{it}\eta_{it}) = 0$. As such, the OLS estimator of this model is consistent. We can recover estimates of the $\alpha_i$ as discussed above. It is to be noted that the error $\eta_{it}$ is almost certainly heteroscedastic and autocorrelated, so OLS will not be efficient, and inferences based on OLS need to be done taking this into account. One could attempt to use GLS, or panel-robust covariance matrix estimation, or both, as above.

There are other estimators when we have random effects, a well-known example being the "between" estimator, which operates on the time averages of the cross sectional units. There is no advantage to doing this, as the overall estimator is already consistent, and averaging looses information (efficiency loss). One would still need to deal with cross sectional heteroscedasticity when using the between estimator, so there is no gain in simplicity, either.

It is to be emphasized that "random effects" *is not a plausible assumption* with most economic data, so use of this estimator is discouraged, even if your statistical package offers it as an option. Think carefully about whether the assumption is warranted before trusting the results of this estimator.

## Hausman test

Suppose you're doubting about whether fixed or random effects are present.

- If we have correlation between $x_{it}$ and $\alpha_i$ (fixed effects), then the "within" estimator will be consistent, but the random effects estimator of the previous section will not.

- Evidence that the two estimators are converging to different limits is evidence in favor of fixed effects, not random effects.

- A Hausman test statistic can be computed, using the difference between the two estimators.

  - The null hypothesis is that the effects are uncorrelated with the regressors in $x_{it}$ ("random effects") so that both estimators are consistent under the null.

- When the test rejects, we conclude that fixed effects are present, so the "within" estimator should be used.
- Now, what happens if the test does not reject? One could optimistically turn to the random effects model, but it's probably more realistic to conclude that the test may have low power. Failure to reject does not mean that the null hypothesis is true. After all, estimation of the covariance matrices needed to compute the Hausman test is a non-trivial issue, and is a source of considerable noise in the test statistic (noise=low power).
- Finally, the simple version of the Hausman test requires that the estimator under the null be fully efficient. Achieving this goal is probably a utopian prospect. A conservative approach would acknowledge that neither estimator is likely to be efficient, and to operate accordingly. I have a little paper on this topic, Creel, *Applied Economics,* 2004. See also Cameron and Trivedi, section 21.4.3.

**In class, do the first part of the example at the end of the chapter at this time**

## 19.4 Dynamic panel data

When we have panel data, we have information on both $y_{it}$ as well as $y_{i,t-1}$. One may naturally think of including $y_{i,t-1}$ as a regressor, to capture dynamic effects that can't be analyed with only cross-sectional data. Excluding dynamic effects is often the reason for detection of spurious AUT of the errors. With dynamics, there is likely to be less of a problem of autocorrelation, but one should still be concerned that some might still be present. The model, using a single lag of the dependent variable, becomes

$$\begin{aligned} y_{it} &= \alpha_i + \gamma y_{i,t-1} + x_{it}\beta + \epsilon_{it} \\ y_{it} &= \alpha + \gamma y_{i,t-1} + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ y_{it} &= \alpha + \gamma y_{i,t-1} + x_{it}\beta + \eta_{it} \end{aligned}$$

We assume that the $x_{it}$ are uncorrelated with $\epsilon_{it}$.

- Note that $\alpha_i$ is a component that determines both $y_{it}$ and its lag, $y_{i,t-1}$.

Thus, $\alpha_i$ and $y_{i,t-1}$ are correlated, even if the $\alpha_i$ are pure random effects (uncorrelated with $x_{it}$).

- So, $y_{i,t-1}$ is correlated with $\eta_{it}$.
- For this reason, OLS estimation is inconsistent even for the random effects model, and it's also of course still inconsistent for the fixed effects model.
- When regressors are correlated with the errors, the natural thing to do is start thinking of instrumental variables estimation, or GMM.

# Arellano-Bond estimator

The first thing is to realize that the $\alpha_i$ that are a component of the error are correlated with all regressors in the general case of fixed effects. Getting rid of the $\alpha_i$ is a step in the direction of solving the problem. We could subtract the time averages, as above for the "within" estimator, but this would give us problems later when we need to define instruments. Instead, consider the model in first differences

$$\begin{aligned} y_{it} - y_{i,t-1} &= (\alpha_i + \gamma y_{i,t-1} + x_{it}\beta + \epsilon_{it}) - (\alpha_i + \gamma y_{i,t-2} + x_{i,t-1}\beta + \epsilon_{i,t-1}) \\ &= \gamma\left(y_{i,t-1} - y_{i,t-2}\right) + \left(x_{it} - x_{i,t-1}\right)\beta + \epsilon_{it} - \epsilon_{i,t-1} \end{aligned}$$

or

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta x_{it}\beta + \Delta\epsilon_{it}$$

- Now the pesky $\alpha_i$ are no longer in the picture.
- Note that we loose one observation when doing first differencing.

- OLS estimation of this model will still be inconsistent, because $y_{i,t-1}$is clearly correlated with $\epsilon_{i,t-1}$.

- Note also that the error $\Delta\epsilon_{it}$ is serially correlated even if the $\epsilon_{it}$ are not.

- There is no problem of correlation between $\Delta x_{it}$ and $\Delta\epsilon_{it}$. Thus, to do GMM, we need to find instruments for $\Delta y_{i,t-1}$, but the variables in $\Delta x_{it}$ can serve as their own instruments.

How about using $y_{i.t-2}$ as an instrument?

- It is clearly correlated with $\Delta y_{i,t-1} = (y_{i,t-1} - y_{i,t-2})$
- *as long as the $\epsilon_{it}$ are not serially correlated*, then $y_{i,t-2}$ is not correlated with $\Delta\epsilon_{it} = \epsilon_{it} - \epsilon_{i,t-1}$.
- We can also use additional lags $y_{i,t-s}$, $s \geq 2$ to increase efficiency, because GMM with additional instruments is asymptotically more efficient than with less instruments (but small sample bias may become a serious problem).

This sort of estimator is widely known in the literature as an Arellano-Bond estimator, due to the influential 1991 paper of Arellano and Bond (1991).

- Note that this sort of estimators requires $T = 3$ at a minimum.
- For $t = 1$ and $t = 2$, we cannot compute the moment conditions.
  - for $t = 1$, we do not have $y_{i,t-1} = y_{i,0}$, so we can't compute dependent variable.

- for $t = 2$, we can compute the dependent variable $\Delta y_{i2}$, but not the regressor $\Delta y_{i,2-1} = y_{i,1} - y_{i,0}$.

- for $t = 3$, we can compute the dep. var. $\Delta y_{i,3}$, the regressor $\Delta y_{i,2} = y_{i,2} - y_{i,1}$, and we have $y_{i,1,}$ to serve as an instrument for $\Delta y_{i,2}$

- If $T > 3$, then when $t = 4$, we can use both $y_{i,1}$ and $y_{i,2}$ as instruments. This sort of unbalancedness in the instruments requires a bit of care when programming. Also, additional instruments increase asymptotic efficiency but can lead to increased small sample bias, so one should be a little careful with using too many instruments. Some robustness checks, looking at the stability of the estimates are a way to proceed.

One should note that serial correlation of the $\epsilon_{it}$ will cause this estimator to be inconsistent. Serial correlation of the errors *may* be due to dynamic misspecification, and this can be solved by including additional lags of the dependent variable. However, too many lags leads to a reduction of the sample size, so there's a limit to what can be done without having variances explode. However, serial correlation may also be due to factors not captured in lags of the dependent variable. If this is a possibility, then the validity of the Arellano-Bond type instruments is in question.

- A final note is that the error $\Delta\epsilon_{it}$ is serially correlated even when the $\epsilon_{it}$ are not, and very likely heteroscedastic across $i$. One needs to take this into account when computing the covariance of the GMM estimator. One can also attempt to use GLS style weighting to improve efficiency. There are many possibilities.

- there is a "system" version of this sort of estimator that adds additional moment conditions, to improve efficiency

## 19.5 Examples

### Arellano-Bond data

Use the GRETL data set abdata.gdt to illustrate fixed effects, random effects, and DPD estimation For FE and RE, use the model

$$n_{it} = \alpha_i + \beta_t + \gamma w_{it} + \delta k_{it} + \phi ys_{it} + \epsilon_{it}$$

- open abdata.gdt in GRETL
- read dataset info: 9 years of data on 140 companies in manufacturing sector (different industries).
  - examine the variables: note that the data set is not "balanced": some companies are not observed in some years
  - taking care of this problem is annoying without using a well written panel data package.

- estimate fixed effects
  - note the pattern of the coefficients of the yearly indicators: the "Margaret Thatcher effect"
  - signs of coefficients seem ok. Exogeneity to be trusted?
  - save fixed effects, save residuals
  - do residuals appear to be normally distributed? Test, and nonparametric density. If not normal, then random effects is not fully efficient, even if exogeneity of effects is valid.
  - is there evidence of serial correlation of residuals? Run AR(1) on residuals: significant autocorrelation. Suggests an omitted dynamic effect.
  - do nonparametric density plot of fixed effects: mean is 1, but significant variation across companies (different industries have different labor intensity)

- run random effects (tradition, but not logic, demands that we do it)
  - Hausman test: rejects RE (unsurprisingly): we should favor FE.
    - However, if errors are not normal (they are not, according to test)
    - or if there is serial correlation (there is, according to AR1 fit)
    - then the test is not valid.
  - Nevertheless, FE is probably favored on strictly theoretical grounds.

- Given that the residuals seem to be serially correlated, we probably need to introduce dynamic structure. For DPD, use the model

$$n_{it} = \alpha_i + \beta_t + \rho_1 n_{i,t-1} + \gamma w_{it} + \delta k_{it} + \phi ys_{it} + \epsilon_{it}$$

- the estimate of $\rho_1$ is economically and statistically significant
- note the important differences in the other coefficients compared to the FE model

check the serial correlation of the residuals: if it exists, the instruments are not valid. Possible solution is to augment the AR order, but the sample size gets smaller with every additional lag.

# Stock and Watson Beer tax example

Stock and Watson "Introduction to Econometrics", 3rd Edition, discuss panel data in Chapter 10. The book is available from the UAB library as an e-text (if you happen to be at the UAB). They use the data set fatality.xlsx, which is explained in fatality.docx. The data is available as a GRETL file here. Using this data, and the software of your choice (hint: Gretl is easy)

1. replicate the results in Stock and Watson, Ch. 10, Eqn. 10.2 and Figure 10.1

2. replicate the results in Eqn 10.8 and Figure 10.2.

3. replicate the fixed effects panel data results in column 2 of Table 10.1

A video which shows how to do this is here.

## 19.6 Practical summary

The practical summary for the Chapter is here. This does FE and DPD for the Arellano-Bond data, using Julia.

## 19.7 Exercises

1. Stock and Watson, in their textbook "Introduction to Econometrics", 3rd Edition, discuss panel data, in Chapter 10. They use the data set fatality.xlsx, which is explained in fatality.docx. Using this data, and the software of your choice (hint: Gretl is easy)

   (a) replicate the results in Eqn. 10.2 and Figure 10.1.

   (b) replicate the results in Eqn 10.8 and Figure 10.2.

   (c) replicate the fixed effects panel data results in column 2 of Table 10.1.

   (d) give an economic interpretation of the results.

2. In the context of a dynamic model with fixed effects, why is the differencing used in the "within" estimation approach (equation 19.4) problematic? That is, why does the Arellano-Bond estimator operate on the model in first differences instead of using the within approach?

3. Consider the simple linear panel data model with random effects (equation 19.5). Suppose that the $\epsilon_{it}$ are independent across cross sectional units, so that $E(\epsilon_{it}\epsilon_{js}) = 0$, $i \neq j$, $\forall t, s$. With a cross sectional unit, the errors are independently and identically distributed, so $E(\epsilon_{it}^2) = \sigma_i^2$, but $E(\epsilon_{it}\epsilon_{is}) = 0$, $t \neq s$. More compactly, let $\epsilon_i = \begin{bmatrix} \epsilon_{i1} & \epsilon_{i2} & \cdots & \epsilon_{iT} \end{bmatrix}'$. Then the assumptions are that $E(\epsilon_i\epsilon_i') = \sigma_i^2 I_T$, and $E(\epsilon_i\epsilon_j') = 0$, $i \neq j$.

   (a) write out the form of the entire covariance matrix ($nT \times nT$) of all errors, $\Sigma = E(\epsilon\epsilon')$, where $\epsilon = \begin{bmatrix} \epsilon_1' & \epsilon_2' & \cdots & \epsilon_T' \end{bmatrix}'$ is the column vector of $nT$ errors.

   (b) suppose that $n$ is fixed, and consider asymptotics as $T$ grows. Is it possible to estimate the $\Sigma_i$ consistently? If so, how?

   (c) suppose that $T$ is fixed, and consider asymptotics an $n$ grows. Is it possible to estimate the $\Sigma_i$ consistently? If so, how?

   (d) For one of the two preceeding parts (b) and (c), consistent estimation

is possible. For that case, outline how to do "within" estimation using a GLS correction.

# Chapter 20

# Nonparametric inference

Cameron and Trivedi (2005), Ch. 9; Li and Racine (2007).

What do we mean by the term "nonparametric inference"? Simply, this means inferences that are possible without restricting the functions of interest to belong to a parametric family.

**Example 87.** A parametric demand function: $x = \alpha + \beta p + \gamma m + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Here, the functional form of the conditional mean is restricted to

be linear in the parameter and the regressors, and the distribution of the error is restricted to the set of mean zero normal distributions.

**Example 88.** A nonparametric demand function: $x = x(p, m) + \epsilon$, where $E(\epsilon|p, m) = 0$. The conditional mean is the function $x(p, m)$, but the form is not restricted. Also, the error has conditional mean zero, but may have any distribution that follows this restriction.

- Normally, it is good to use parametric restrictions *if we are confident that they are at least approximately true*, as this will lead to low variance, low bias estimation.

- If we impose parametric restrictions for which we have little or no justification, we may provoke serious biases which can lead to incorrect conclusions.

**Motivation** (see White (1980b)).

In this section we return to an example which we've already seen: approximating a nonlinear in the variables regression line using a linear in the variables regression line.

We suppose that data is generated by random sampling of $(y, x)$, where $y = f(x) + \varepsilon$, $x$ is uniformly distributed on $(0, 2\pi)$, and $\varepsilon$ is a classical error with variance equal to 1. Suppose that the regression function is truly a quadratic function:

$$f(x) = 1 + \frac{3x}{2\pi} - \left(\frac{x}{2\pi}\right)^2$$

- If we knew the functional form but not the coefficients, we could just estimate by least squares.

- But, let's assume that we do not know the functional form, to make things interesting

- Suppose that the problem of interest is to estimate the elasticity of $f(x)$ with respect to $x$, throughout the range of $x$. Recall that the elasticity is an elasticity is the marginal function divided by the average function:

$$\varepsilon(x) = \frac{f'(x)}{f(x)/x}$$

- We would like to be able to estimate this quantity well for any arbitrary value $x$.

In general, the functional form of $f(x)$ is unknown. One idea is to take a Taylor's series approximation to $f(x)$ about some point $x_0$. Flexible functional forms such as the transcendental logarithmic (usually known as the translog) can be interpreted as second order Taylor's series approximations. We'll work with a first order approximation, for simplicity. Approximating about $x_0$:

$$h(x) = f(x_0) + D_x f(x_0)\,(x - x_0)$$

If the approximation point is $x_0 = 0$, we can write

$$h(x) = a + bx$$

The coefficient $a$ is the value of the function at $x = 0$, and the slope is the value of the derivative at $x = 0$. These are of course not known. One might try estimation by ordinary least squares. The objective function is

$$s(a, b) = 1/n \sum_{t=1}^{n} \left(y_t - h(x_t)\right)^2 .$$

The limiting objective function, following the argument we used to get equations 13.1 and 24.8 is

$$s_\infty(a,b) = \int_0^{2\pi} \left(f(x) - h(x)\right)^2 dx + C$$

The theorem regarding the consistency of extremum estimators (Theorem 35) tells us that $\hat{a}$ and $\hat{b}$ will converge almost surely to the values that minimize the limiting objective function. Solving the first order conditions[1] reveals that $s_\infty(a,b)$ obtains its minimum at $\left\{a^0 = \frac{7}{6}, b^0 = \frac{1}{\pi}\right\}$. The estimated approximating function $\hat{h}(x)$ therefore tends almost surely to

$$h_\infty(x) = 7/6 + x/\pi$$

[1]The following results were obtained using the free computer algebra system (CAS) Maxima. Unfortunately, I have lost the source code to get the results. It's not hard to do, though: see 13.3.

Figure 20.1: True and simple approximating functions

In Figure 20.1 we see the true function and the limit of the approximation to see the asymptotic bias as a function of $x$.

(The approximating model is the straight line, the true model has curvature.) Note that the approximating model is in general inconsistent, even at the approximation point. This shows that "flexible functional forms" based upon Taylor's series approximations do not in general lead to consistent estimation of functions.

- The approximating model seems to fit the true model fairly well, asymptotically, so maybe the approximation problem is not too important?

- However, we are interested in the elasticity of the function. Recall that the elasticity is an elasticity is the marginal function divided by the average function:

$$\varepsilon(x) = \frac{f'(x)}{f(x)/x}$$

- Good approximation of the elasticity over the range of $x$ will require a good approximation of both $f(x)$ and $f'(x)$ over the range of $x$. The approximating elasticity is

$$\eta(x) = \frac{h'(x)}{h(x)/x}$$

- The question is: how well does $\eta(x)$ approximate $\varepsilon(x)$?

Figure 20.2: True and approximating elasticities

In Figure 20.2 we see the true elasticity and the elasticity obtained from the limiting approximating model.

The true elasticity is the line that has negative slope for large $x$. Visually we see that the elasticity is not approximated so well. Root mean squared error in the

approximation of the elasticity is

$$\left(\int_0^{2\pi} \left(\varepsilon(x) - \eta(x)\right)^2 dx\right)^{1/2} = 0.31546$$

Now suppose we use the leading terms of a trigonometric series as the approximating model. The reason for using a trigonometric series as an approximating model is motivated by the asymptotic properties of the Fourier flexible functional form (Gallant, 1981, 1982), which is an example of a *sieve estimator*. Normally with this type of model the number of basis functions is an increasing function of the sample size. Here we hold the set of basis function fixed. We will consider the asymptotic behavior of a fixed model, which we interpret as an approximation to the estimator's behavior in finite samples. Consider the set of basis functions:

$$Z(x) = \left[\, 1 \;\; x \;\; \cos(x) \;\; \sin(x) \;\; \cos(2x) \;\; \sin(2x) \,\right].$$

The approximating model is

$$g_K(x) = Z(x)\alpha.$$

Maintaining these basis functions as the sample size increases, we find that the

limiting objective function is minimized at

$$\left\{a_1 = \frac{7}{6}, a_2 = \frac{1}{\pi}, a_3 = -\frac{1}{\pi^2}, a_4 = 0, a_5 = -\frac{1}{4\pi^2}, a_6 = 0\right\}.$$

Substituting these values into $g_K(x)$ we obtain the almost sure limit of the approximation

$$g_\infty(x) = 7/6 + x/\pi + (\cos x)\left(-\frac{1}{\pi^2}\right) + (\sin x)\,0 + (\cos 2x)\left(-\frac{1}{4\pi^2}\right) + (\sin 2x)\,0 \quad (20.1)$$

Figure 20.3: True function and more flexible approximation

In Figure 20.3 we have the approximation and the true function: Clearly the truncated trigonometric series model offers a better approximation, asymptotically, than does the linear model.

Figure 20.4: True elasticity and more flexible approximation

In Figure 20.4 we have the more flexible approximation's elasticity and that of the true function: On average, the fit is better, though there is some implausible waviness in the estimate.

Root mean squared error in the approximation of the elasticity is

$$\left(\int_0^{2\pi}\left(\varepsilon(x)-\frac{g'_\infty(x)x}{g_\infty(x)}\right)^2 dx\right)^{1/2} = 0.16213,$$

about half that of the RMSE when the first order approximation is used.

- Sieve estimators allow the number of regressors to grow as the sample size grows.
  - this must be done in a controlled way: there is a variance/bias tradeoff:
    - more regressors -> less bias
    - more regressors-> more variance
- It can be shown that if we introduce regressors at a slow rate, it is possible to drive the RMSE of the approximation to the regression function and to the elasticity to zero, as the sample size grows.

- This is why sieve estimators (and other nonparametric regression estimators) are of interest: we can obtain consistent estimates of regression functions without knowledge of the functional form of the true regression line.

## 20.1 Estimation of regression functions

Here, we will see two examples of methods of estimating regression functions without knowledge of the true functional form: kernel regression and neural nets. There are other methods, for example sieve estimators, nearest neighbors, etc.

### Kernel regression estimators

**Readings**: Li and Racine (2007), Ch. 2; Cameron and Trivedi (2005), Ch. 9; Bierens (1987).

Kernel regression estimation is an example of fully nonparametric estimation (others are splines, nearest neighbors, etc.). We'll consider the Nadaraya-Watson kernel regression estimator in a simple case.

- Suppose we have an iid sample from the joint density $f(x, y)$, where $x$ is $k$ -dimensional. The model is

$$y_t = g(x_t) + \varepsilon_t,$$

where

$$E(\varepsilon_t | x_t) = 0.$$

- The conditional expectation of $y$ given $x$ is $g(x)$. By definition of the condi-

tional expectation, we have

$$\begin{aligned} g(x) &= \int y \frac{f(x,y)}{h(x)} dy \\ &= \frac{1}{h(x)} \int y f(x,y) dy, \end{aligned}$$

where $h(x)$ is the marginal density of $x$ :

$$h(x) = \int f(x,y) dy.$$

- This suggests that we could estimate $g(x)$ by estimating $h(x)$ and $\int y f(x,y) dy$.

## Estimation of the denominator

A kernel estimator for $h(x)$ has the form

$$\hat{h}(x) = \frac{1}{n} \sum_{t=1}^{n} \frac{K\left[(x - x_t)/\gamma_n\right]}{\gamma_n^k},$$

where $n$ is the sample size and $k$ is the dimension of $x$.

- The function $K(\cdot)$ (the kernel) is absolutely integrable:

  $$\int |K(x)| dx < \infty,$$

  and $K(\cdot)$ integrates to 1 :

  $$\int K(x) dx = 1.$$

  In this respect, $K(\cdot)$ is like a density function, but we do not necessarily restrict $K(\cdot)$ to be nonnegative.

- The *window width* parameter, $\gamma_n$ is a sequence of positive numbers that

satisfies

$$\begin{aligned} \lim_{n\to\infty} \gamma_n &= 0 \\ \lim_{n\to\infty} n\gamma_n^k &= \infty \end{aligned}$$

So, the window width must tend to zero, but not too quickly.

- To show pointwise consistency of $\hat{h}(x)$ for $h(x)$, first consider the expectation of the estimator (because the estimator is an average of iid terms, we only need to consider the expectation of a representative term):

$$E\left[\hat{h}(x)\right] = \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right] h(z) dz.$$

Change variables as $z^* = (x-z)/\gamma_n$, so $z = x - \gamma_n z^*$ and $|\frac{dz}{dz^{*\prime}}| = \gamma_n^k$, we obtain

$$\begin{aligned} E\left[\hat{h}(x)\right] &= \int \gamma_n^{-k} K\left(z^*\right) h(x - \gamma_n z^*) \gamma_n^k dz^* \\ &= \int K\left(z^*\right) h(x - \gamma_n z^*) dz^*. \end{aligned}$$

Now, asymptotically,

$$\begin{aligned}
\lim_{n\to\infty} E\left[\hat{h}(x)\right] &= \lim_{n\to\infty} \int K\left(z^*\right) h(x-\gamma_n z^*) dz^* \\
&= \int \lim_{n\to\infty} K\left(z^*\right) h(x-\gamma_n z^*) dz^* \\
&= \int K\left(z^*\right) h(x) dz^* \\
&= h(x) \int K\left(z^*\right) dz^* \\
&= h(x),
\end{aligned}$$

since $\gamma_n \to 0$ and $\int K\left(z^*\right) dz^* = 1$ by assumption. (Note: that we can pass the limit through the integral is a result of the dominated convergence theorem. For this to hold we need that $h(\cdot)$ be dominated by an absolutely integrable function.)

- Next, considering the variance of $\hat{h}(x)$, we have, due to the iid assumption

$$\begin{aligned} n\gamma_n^k V\left[\hat{h}(x)\right] &= n\gamma_n^k \frac{1}{n^2} \sum_{t=1}^{n} V\left\{\frac{K\left[(x-x_t)/\gamma_n\right]}{\gamma_n^k}\right\} \\ &= \gamma_n^{-k} \frac{1}{n} \sum_{t=1}^{n} V\left\{K\left[(x-x_t)/\gamma_n\right]\right\} \end{aligned}$$

- By the representative term argument, this is

$$n\gamma_n^k V\left[\hat{h}(x)\right] = \gamma_n^{-k} V\left\{K\left[(x-z)/\gamma_n\right]\right\}$$

- Also, since $V(x) = E(x^2) - E(x)^2$ we have

$$\begin{aligned} n\gamma_n^k V\left[\hat{h}(x)\right] &= \gamma_n^{-k} E\left\{\left(K\left[(x-z)/\gamma_n\right]\right)^2\right\} - \gamma_n^{-k}\left\{E\left(K\left[(x-z)/\gamma_n\right]\right)\right\}^2 \\ &= \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz - \gamma_n^k \left\{\int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right] h(z)dz\right\}^2 \\ &= \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz - \gamma_n^k E\left[\widehat{h}(x)\right]^2 \end{aligned}$$

The second term converges to zero:

$$\gamma_n^k E\left[\widehat{h}(x)\right]^2 \to 0,$$

by the previous result regarding the expectation and the fact that $\gamma_n \to 0$. Therefore,

$$\lim_{n\to\infty} n\gamma_n^k V\left[\hat{h}(x)\right] = \lim_{n\to\infty} \int \gamma_n^{-k} K\left[(x-z)/\gamma_n\right]^2 h(z)dz.$$

Using exactly the same change of variables as before, this can be shown to be

$$\lim_{n\to\infty} n\gamma_n^k V\left[\hat{h}(x)\right] = h(x)\int \left[K(z^*)\right]^2 dz^*.$$

Since both $\int \left[K(z^*)\right]^2 dz^*$ and $h(x)$ are bounded, the RHS is bounded, and since $n\gamma_n^k \to \infty$ by assumption, we have that

$$V\left[\hat{h}(x)\right] \to 0.$$

- Since the bias and the variance both go to zero, we have pointwise consistency (convergence in quadratic mean implies convergence in probability).

## Estimation of the numerator

To estimate $\int yf(x,y)dy$, we need an estimator of $f(x,y)$. The estimator has the same form as the estimator for $h(x)$, only with one dimension more:

$$\hat{f}(x,y) = \frac{1}{n}\sum_{t=1}^{n} \frac{K_*\left[\left(y-y_t\right)/\gamma_n, \left(x-x_t\right)/\gamma_n\right]}{\gamma_n^{k+1}}$$

The kernel $K_*(\cdot)$ is required to have mean zero:

$$\int yK_*(y,x)\,dy = 0$$

and to marginalize to the previous kernel for $h(x)$ :

$$\int K_*(y,x)\,dy = K(x).$$

With this kernel, we have (not obviously, see Li and Racine, Ch. 2, section 2.1)

$$\int y\hat{f}(y,x)dy = \frac{1}{n}\sum_{t=1}^{n} y_t \frac{K\left[\left(x-x_t\right)/\gamma_n\right]}{\gamma_n^k}$$

by marginalization of the kernel, so we obtain

$$\begin{aligned}
\hat{g}(x) &:= \frac{1}{\hat{h}(x)}\int y\hat{f}(y,x)dy \\
&= \frac{\frac{1}{n}\Sigma_{t=1}^{n} y_t \frac{K[(x-x_t)/\gamma_n]}{\gamma_n^k}}{\frac{1}{n}\Sigma_{t=1}^{n} \frac{K[(x-x_t)/\gamma_n]}{\gamma_n^k}} \\
&= \frac{\Sigma_{t=1}^{n} y_t K\left[\left(x-x_t\right)/\gamma_n\right]}{\Sigma_{t=1}^{n} K\left[\left(x-x_t\right)/\gamma_n\right]}
\end{aligned}$$

This is the Nadaraya-Watson kernel regression estimator.

## Discussion

- defining

$$w_t = \frac{K\left[\left(x - x_t\right)/\gamma_n\right]}{\Sigma_{t=1}^{n} K\left[\left(x - x_t\right)/\gamma_n\right]},$$

  the kernel regression estimator for $g(x_t)$ can be written as

$$\hat{g}(x) = \sum_{t=1}^{n} y_t w_t,$$

  a weighted average of the $y_j$, $j = 1, 2, ..., n$, where higher weights are associated with points that are closer to $x_t$. The weights sum to 1. See this link for a graphic interpretation.

- The window width parameter $\gamma_n$ imposes smoothness. The estimator is increasingly flat as $\gamma_n \to \infty$, since in this case each weight tends to $1/n$.

- A large window width reduces the variance (strong imposition of flatness), but increases the bias.

- A small window width reduces the bias, but makes very little use of information except points that are in a small neighborhood of $x_t$. Since relatively little information is used, the variance is large when the window width is small.

- The standard normal density is a popular choice for $K(.)$ and $K_*(y, x)$, though there are possibly better alternatives.

## Choice of the window width: Cross-validation

The selection of an appropriate window width is important. One popular method is cross validation. This consists of splitting the sample into two parts (e.g., 50%-50%). The first part is the "in sample" data, which is used for estimation, and the second part is the "out of sample" data, used for evaluation of the fit though RMSE or some other criterion. The steps are:

1. Split the data. The out of sample data is $y^{out}$ and $x^{out}$ (these are the first $n_{out}$ observations, say-

2. Choose a window width $\gamma$.

3. With the in sample data, fit $\hat{y}_t^{out}(\gamma)$ corresponding to each $x_t^{out}$. This fitted value is a function of the window width, the in sample data, as well as the evaluation point $x_t^{out}$, but it does not involve $y_t^{out}$.

4. Repeat for all out of sample points.

5. Calculate $\text{RMSE}(\gamma) = \sqrt{\frac{1}{n_{out}} \Sigma_{t=1}^{n_{out}} \left(y_t^{out} - \hat{y}_t^{out}(\gamma)\right)^2}$

6. Go to step 2, or to the next step if enough window widths have been tried.

7. Select the $\gamma$ that minimizes $\text{RMSE}(\gamma)$ (Verify that a minimum has been found, for example by plotting RMSE as a function of $\gamma$).

8. Re-estimate using the best $\gamma$ and all of the data.

- there is a variation known as leave-one-out cross validation, where each $y_t^{out}$ is fit in turn using all of the remaining observations, omitting the $t^{th}$ observation. This is the recommended procedure. It is somewhat more demanding computationally, but works better.

**Example:** from Julia, and after doing `using Econometrics`, run npreg(). It will give you a figure similar to

You see that the kernel regression mean and median are close to the true function, and the 90% CI defined by nonparametric quantiles contains the true function, everywhere, in this case. Edit the code to see what's going on.

## Neural nets

Neural networks are a well known tool in many fields, and there are many presentations, both academic and more informal, of various structures that can be used. For this reason, the presentation here is brief. For more details and references, see Kuan and White (1994). A very useful practical guide is given by LeCun et al. (2012). A good practical introduction is here. Papers by Gallant and White (1988) and Hornik et al. (1989) show that some types of neural networks can be thought of as as nonparametric regression estimators, but this discussion seems to still be open, in the case of the "deep learning" nets that are popular nowadays. The discussion below is based on Creel (2017), and code for the example below is at https://github.com/mcreel/NeuralNetsForIndirectInference.jl.

- Suppose we are interested in the regression model $y = g(x) + \epsilon$, where $x$ is a $K$-vector and $y$ is a $G$-vector.
  - This is a multivariate (more than one dependent variable) multiple (more than one regressor) regression model.
  - Because we don't specify the form of $g(x) = E(y|x)$, it is a nonparametric regression model.
  - Let's model this using a neural net.

- Consider a simple feed forward neural net for regression of an output in $R^K$ upon an input in $R^G$. A typical feed forward net is depicted in Figure 20.5, which maps 3 inputs to 2 outputs.
  - The inputs to the net, $I_1$, $I_2$, and $I_3$, are scalar real numbers, as are the outputs $O_1$ and $O_2$.
  - The net has two hidden layers, formed by 4 and 3 hidden nodes or neurons, $h_1$,$h_2$,...,$h_7$,
  - and an output layer, which gives the values of the two outputs $O_1$ and $O_2$.
  - The values $\alpha_1$, $\alpha_2$ and $\alpha_3$ are vectors of "bias" parameters, which are discussed below.

Figure 20.5: A simple neural net

- In general, a net is a series of transformations of the inputs.
- Each of the transformations is referred to as a layer.
- The inputs themselves constitute the zero-th layer, and the final result of the transformations is the output layer.
- A layer, $H_j$, is a vector of real numbers, which is the result of the $j^{th}$ in the series of transformations.
  - Let $H_0$ be the $G$ dimensional vector of inputs.
  - Suppose that there are $P$ layers.
  - Let $n_j$ be the number of neurons in the $j^{th}$ layer, $j = 1, 2, ..., P$.
- The value taken by a neuron in the $j^{th}$ layer is the result of the layer's "activation function", $f_j(\cdot)$, applied on an element-by-element basis to an affine function of the inputs to the layer. The relationship between the

layers is given by

$$H_j = f_j(\alpha_j + \beta_j H_{j-1}),\ j = 1, 2, ..., P, \tag{20.2}$$

- $\alpha_j$ is a $n_j$ dimensional vector of parameters (these are known as bias parameters in the neural net literature)
- $\beta_j$ is a $n_j \times n_{j-1}$ matrix of parameters.
- The layers 1, 2, ..., $P-1$ are referred to as hidden layers
- layer $P$ is the output layer.
- The input to the first hidden layer, known as the input layer, is simply the input data, $H_0 \in \mathbb{R}^G$. The output of the net is the final layer, $H_P \in \mathbb{R}^K$. When using a net for regression, the last activation function, $f_P(\cdot)$, is simply an identity function, so that $H_P = \alpha_P + \beta_P H_{P-1}$. The reason that an activation function is used with the hidden layers is that this is what allows the net to approximate a nonlinear mapping. If

all activation functions were identity functions, the entire net would reduce to an over-parameterized linear regression model. In this paper, the activation function that is used for the hidden layers is the "rectified linear unit" (ReLU) function, $f(x) = \max(0, x)$, a very widely used choice in modern deep learning applications.

A neural net may contain many, many hidden parameters

- Suppose the number of inputs, $G$, is 40, and the number of outputs, $K$, is 9.
- Suppose the net has two hidden layers, of size 300 and 40, respectively.
- Then there are $300 \times 40$ parameters in the $\beta_1$ matrix of the first layer and 40 elements in the $\alpha_1$ vector.
- Similarly, in the second hidden layer, there are $40 \times 300 + 40$ parameters
- there are $9 \times 40 + 9$ parameters corresponding to the output layer.
- Thus, the total number of parameters is 24449.

- A neural net is a nonlinear regression model that may be highly parameterized
  - may be more parameters than observations in a single sample
  - lack of identification: neurons can be reordered
  - multiple local minima

Partial solutions:

- for a simulable model, we can generate multiple data sets to train the net. With much data, even a large net can be trained well.
- For the multiple local minima problem, "stochastic gradient descent" and techniques related to cross validation can help a lot:
  - compute the gradient using a small number of observations from the training set. This is called a stochastic gradient, because it depends on the observations that were chosen.

- take a small step in that direction. The step size is called the "learning rate" in the NN literature.
- evaluate the new fit using a testing set
- iterate gradient/learning, with the learning rate (step size) getting smaller as learning proceeds, until the fit to the testing set no longer improves.

Modern software exists to make this quite easy to do. For Julia, see this page to get started.

- For this to work well, you need a lot of data, to train the net.
- Simulation based econometric methods can give us a lot of simulated data, so using neural nets when doing simulation based estimation is very natural
- A neural net indirect inference estimator is not an extremum estimator: how to test hypotheses?
  - bootstrapping? *Update*: based on my experimentation, no. Inference requires accurate estimation of tail quantiles, and this is difficult to do based on a training sample drawn from the prior. Likewise, neural quantile regression does not lead to good estimation of tail quantiles, for the same reason (one would need an enormous sample from the prior).
  - One can use the NN estimator as a statistic for indirect inference or related methods, and then use the asymptotic theory for those methods. This works pretty well - see https://github.com/mcreel/SNM and the working paper referenced there .

## 20.2 Density function estimation

### Kernel density estimation

The previous discussion suggests that a kernel density estimator may easily be constructed. We have already seen how joint densities may be estimated. If were interested in a conditional density, for example of $y$ conditional on $x$, then the kernel estimate of the conditional density is simply

$$\begin{aligned}
\widehat{f}_{y|x} &= \frac{\hat{f}(x,y)}{\hat{h}(x)} \\
&= \frac{\frac{1}{n}\Sigma_{t=1}^{n}\frac{K_*[(y-y_t)/\gamma_n,(x-x_t)/\gamma_n]}{\gamma_n^{k+1}}}{\frac{1}{n}\Sigma_{t=1}^{n}\frac{K[(x-x_t)/\gamma_n]}{\gamma_n^{k}}} \\
&= \frac{1}{\gamma_n}\frac{\Sigma_{t=1}^{n}K_*\left[(y-y_t)/\gamma_n,(x-x_t)/\gamma_n\right]}{\Sigma_{t=1}^{n}K\left[(x-x_t)/\gamma_n\right]}
\end{aligned}$$

where we obtain the expressions for the joint and marginal densities from the section on kernel regression.

**Example 89.** The Julia script ExampleKernelDensity.jl draws data from a $\chi^2(3)$ distribution and plots a kernel density fit, plus the true density. We see that they're pretty close, when the sample size is large enough for the kernel estimate to be precise. Try playing around with a smaller sample, and see what happens.

# Semi-nonparametric maximum likelihood

**Readings:** Gallant and Nychka, *Econometrica*, 1987. For a Fortran program to do this and a useful discussion in the user's guide, see this link. See also Cameron and Johansson, *Journal of Applied Econometrics*, V. 12, 1997.

MLE is the estimation method of choice when we are confident about specifying the density. Is is possible to obtain the benefits of MLE when we're not so confident about the specification? In part, yes.

Suppose we're interested in the density of $y$ conditional on $x$ (both may be vectors). Suppose that the density $f(y|x, \phi)$ is a reasonable starting approximation to the true density. This density can be reshaped by multiplying it by a squared polynomial. The new density is

$$g_p(y|x, \phi, \gamma) = \frac{h_p^2(y|\gamma) f(y|x, \phi)}{\eta_p(x, \phi, \gamma)}$$

where

$$h_p(y|\gamma) = \sum_{k=0}^{p} \gamma_k y^k$$

and $\eta_p(x, \phi, \gamma)$ is a normalizing factor to make the density integrate (sum) to one. Because $h_p^2(y|\gamma)/\eta_p(x, \phi, \gamma)$ is a homogenous function of $\theta$ it is necessary to impose a normalization to identify the parameters: $\gamma_0$ is set to 1. The normalization factor $\eta_p(\phi, \gamma)$ is calculated (following Cameron and Johansson) using

$$\begin{aligned}
E(Y^r) &= \sum_{y=0}^{\infty} y^r f_Y(y|\phi, \gamma) \\
&= \sum_{y=0}^{\infty} y^r \frac{\left[h_p\left(y|\gamma\right)\right]^2}{\eta_p(\phi, \gamma)} f_Y(y|\phi) \\
&= \sum_{y=0}^{\infty} \sum_{k=0}^{p} \sum_{l=0}^{p} y^r f_Y(y|\phi) \gamma_k \gamma_l y^k y^l / \eta_p(\phi, \gamma) \\
&= \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k \gamma_l \left\{ \sum_{y=0}^{\infty} y^{r+k+l} f_Y(y|\phi) \right\} / \eta_p(\phi, \gamma) \\
&= \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k \gamma_l m_{k+l+r} / \eta_p(\phi, \gamma).
\end{aligned}$$

By setting $r = 0$ we get that the normalizing factor is

20.3

$$\eta_p(\phi, \gamma) = \sum_{k=0}^{p} \sum_{l=0}^{p} \gamma_k \gamma_l m_{k+l} \tag{20.3}$$

Recall that $\gamma_0$ is set to 1 to achieve identification. The $m_r$ in equation 20.3 are the raw moments of the baseline density. Gallant and Nychka (1987) give conditions under which such a density may be treated as correctly specified, asymptotically. Basically, the order of the polynomial must increase as the sample size increases. However, there are technicalities.

Similarly to Cameron and Johannson (1997), we may develop a negative binomial polynomial (NBP) density for count data. The negative binomial baseline density may be written (see equation 15.22) as

$$f_Y(y|\phi) = \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y}$$

where $\phi = \{\lambda, \psi\}$, $\lambda > 0$ and $\psi > 0$. The usual means of incorporating condi-

tioning variables $\mathbf{x}$ is the parameterization $\lambda = e^{\mathbf{x}'\beta}$. When $\psi = \lambda/\alpha$ we have the negative binomial-I model (NB-I). When $\psi = 1/\alpha$ we have the negative binomial-II (NP-II) model. For the NB-I density, $V(Y) = \lambda + \alpha\lambda$. In the case of the NB-II model, we have $V(Y) = \lambda + \alpha\lambda^2$. For both forms, $E(Y) = \lambda$.

The reshaped density, with normalization to sum to one, is

$$f_Y(y|\phi,\gamma) = \frac{[h_p(y|\gamma)]^2}{\eta_p(\phi,\gamma)} \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y}. \tag{20.4}$$

To get the normalization factor, we need the moment generating function:

$$M_Y(t) = \psi^{\psi}\left(\lambda - e^t\lambda + \psi\right)^{-\psi}. \tag{20.5}$$

To illustrate, Figure 20.6 shows calculation of the first four raw moments of the NB density, calculated using MuPAD, which is a Computer Algebra System that (used to be?) free for personal use. These are the moments you would need to use a second order polynomial ($p = 2$). MuPAD will output these results in the

form of C code, which is relatively easy to edit to write the likelihood function for the model. This has been done in NegBinSNP.cc, which is a C++ version of this model that can be compiled to use with octave using the `mkoctfile` command. Note the impressive length of the expressions when the degree of the expansion is 4 or 5! This is an example of a model that would be difficult to formulate without the help of a program like *MuPAD.*

It is possible that there is conditional heterogeneity such that the appropriate reshaping should be more local. This can be accomodated by allowing the $\gamma_k$ parameters to depend upon the conditioning variables, for example using polynomials.

Gallant and Nychka, *Econometrica*, 1987 prove that this sort of density can approximate a wide variety of densities arbitrarily well as the degree of the polynomial increases with the sample size. This approach is not without its drawbacks: the sample objective function can have an *extremely* large number of local maxima that can lead to numeric difficulties. If someone could figure out how to do in

Figure 20.6: Negative binomial raw moments

a way such that the sample objective function was nice and smooth, they would probably get the paper published in a good journal. Any ideas?

Here's a plot of true and the limiting SNP approximations (with the order of the polynomial fixed) to four different count data densities, which variously exhibit over and underdispersion, as well as excess zeros. The baseline model is a negative binomial density.

Case 1
.5
.4
.3
.2
.1
0
5
10
15
20
Case 2
.1
.05
0
5
10
15
20
25
Case 3
.25
.2
.15
.1
.05
1
2
3
4
5
6
7
Case 4
.2
.15
.1
.05
2.5
5
7.5
10
12.5
15

## 20.3 Examples

Some of these examples are old, using Octave code. I may try to get around to translating them.

### MEPS health care usage data

We'll use the MEPS OBDV data to illustrate kernel regression and semi-nonparametric maximum likelihood.

#### Kernel regression estimation

Let's try a kernel regression fit for the OBDV data. The program OBDVkernel.m loads the MEPS OBDV data, computes kernel regression estimates using the same conditioning variables as in subsection 12.4, and plots the fitted OBDV usage versus AGE and INCOME. The plots are in Figure 20.7.

- Note that usage increases with age, just as we've seen with the parametric models.
- Note that for income, there is a U shape. Previously, we found that income appeared to be insignificant (run EstimatePoisson to see it again).
  - Perhaps that insignificance was due to omitting a nonlinear effect (e.g., quadratic).
  - The U shape could also be due to ignoring endogeneity of income. If a person is seriously ill, they may make more doctor visits, but may also suffer loss of income due to reduces work hours.
  - Another explanation might be that kernel regression has a high variance in regions of data sparseness, so that for very low or high incomes, an outlier or two can have a big impact
- Nonparametric analysis can help us to learn what might be appropriate parametric models, by helping to identify potential problems with a parametric

model

- Once could use bootstrapping or other methods to generate a confidence intervals for the fits.

Figure 20.7: Kernel regression fits, OBDV health care usage versus AGE and INCOME

(a) Kernel fitted OBDV usage versus AGE

(b) Kernel fitted OBDV usage versus INCOME

Figure 20.8: Dollar-Euro

## Financial data and volatility

The data set rates contains the growth rate (100×log difference) of the daily spot \$/euro and \$/yen exchange rates at New York, noon, from January 04, 1999 to February 12, 2008. There are 2291 observations. See the README file for details. Figures 20.8 and 20.9 show the data and their histograms.

- at the center of the histograms, the bars extend above the normal density that best fits the data, and the tails are fatter than those of the best fit normal density. This feature of the data is known as *leptokurtosis*.

Figure 20.9: Dollar-Yen

- in the series plots, we can see that the variance of the growth rates is not constant over time. Volatility clusters are apparent, alternating between periods of stability and periods of more wild swings. This is known as *conditional heteroscedasticity*. ARCH and GARCH well-known models that are often applied to this sort of data.

- Many structural economic models often cannot generate data that exhibits conditional heteroscedasticity without directly assuming shocks that are conditionally heteroscedastic. It would be nice to have an economic explanation for how conditional heteroscedasticity, leptokurtosis, and other (leverage,

etc.) features of financial data result from the behavior of economic agents, rather than from a black box that provides shocks.

The Octave script kernelfit.m performs kernel regression to fit $E(y_t^2|y_{t-1,}^2 y_{t-2}^2)$, and generates the plots in Figure 20.10.

- From the point of view of learning the practical aspects of kernel regression, note how the data is compactified in the example script.

- In the Figure, note how current volatility depends on lags of the squared return rate - it is high when both of the lags are high, but drops off quickly when either of the lags is low.

- The fact that the plots are not flat suggests that this conditional moment contain information about the process that generates the data. Perhaps attempting to match this moment might be a means of estimating the parameters of the dgp. We'll come back to this later.

Figure 20.10: Kernel regression fitted conditional second moments, Yen/Dollar and Euro/Dollar

(a) Yen/Dollar

(b) Euro/Dollar

## Additional kernel regression examples

There is a basic example of kernel regression and kernel density estimation in kernel_example.m .

There is another example of local constant and local linear kernel regression in kernel_local_linear_example.m . With that, you can experiment with different bandwidths.

## Seminonparametric ML estimation and the MEPS data

Now let's estimate a seminonparametric density for the OBDV data. We'll reshape a negative binomial density, as discussed above. The program EstimateNBSNP.m loads the MEPS OBDV data and estimates the model, using a NB-I baseline density and a 2nd order polynomial expansion. The output is:

```
OBDV

=======================================================
BFGSMIN final results

Used numeric gradient

-------------------------------------------------------
STRONG CONVERGENCE
Function conv 1  Param conv 1  Gradient conv 1
-------------------------------------------------------
```

```
Objective function value 2.17061
Stepsize 0.0065
24 iterations
-----------------------------------------------------

 param    gradient  change
  1.3826   0.0000  -0.0000
  0.2317  -0.0000   0.0000
  0.1839   0.0000   0.0000
  0.2214   0.0000  -0.0000
  0.1898   0.0000  -0.0000
  0.0722   0.0000  -0.0000
 -0.0002   0.0000  -0.0000
  1.7853  -0.0000  -0.0000
 -0.4358   0.0000  -0.0000
  0.1129   0.0000   0.0000


*****************************************************
```

NegBin SNP model, MEPS full data set

MLE Estimation Results
BFGS convergence: Normal convergence

Average Log-L: -2.170614
Observations: 4564

| | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -0.147 | 0.126 | -1.173 | 0.241 |
| pub. ins. | 0.695 | 0.050 | 13.936 | 0.000 |
| priv. ins. | 0.409 | 0.046 | 8.833 | 0.000 |
| sex | 0.443 | 0.034 | 13.148 | 0.000 |
| age | 0.016 | 0.001 | 11.880 | 0.000 |
| edu | 0.025 | 0.006 | 3.903 | 0.000 |
| inc | -0.000 | 0.000 | -0.011 | 0.991 |
| gam1 | 1.785 | 0.141 | 12.629 | 0.000 |
| gam2 | -0.436 | 0.029 | -14.786 | 0.000 |
| lnalpha | 0.113 | 0.027 | 4.166 | 0.000 |

```
Information Criteria
CAIC : 19907.6244      Avg. CAIC:   4.3619
 BIC : 19897.6244       Avg. BIC:   4.3597
 AIC : 19833.3649       Avg. AIC:   4.3456
*************************************************
```

Note that the CAIC and BIC are lower for this model than for the models presented in Table 15.3. This model fits well, still being parsimonious. You can play around trying other use measures, using a NP-II baseline density, and using other orders of expansions. Density functions formed in this way may have **MANY** local maxima, so you need to be careful before accepting the results of a casual run. To guard against having converged to a local maximum, one can try using multiple starting values, or one could try simulated annealing as an optimization method. If you uncomment the relevant lines in the program, you can use SA to do the minimization. This will take a *lot* of time, compared to the default BFGS

minimization. The chapter on parallel computations might be interesting to read before trying this.

# Limited information nonparametric filtering

Add discussion from JEF paper.

## 20.4 Exercises

1. In Octave, type "`edit kernel_example`".

   (a) Look this script over, and describe in words what it does.

   (b) Run the script and interpret the output.

   (c) Experiment with different bandwidths, and comment on the effects of choosing small and large values.

2. In Octave, type "`help kernel_regression`".

   (a) How can a kernel fit be done without supplying a bandwidth?

   (b) How is the bandwidth chosen if a value is not provided?

   (c) What is the default kernel used?

3. Using the Octave script OBDVkernel.m as a model, plot kernel regression fits for OBDV visits as a function of income and education.

# Chapter 21

# Quantile regression

References: Cameron and Trivedi (2005), Chapter 4, Koenker and Bassett (1978), Koenker and Hallock (2001), Chernozhukov and Hansen (2005), and Chernozhukov's MIT OpenCourseWare notes, lecture 8 Chernozhukov's quantile reg notes.

This chapter gives an outline of quantile regression. The quantile IV estimator provides an opportunity to explore MCMC methods.

**Conditional quantile, definition**

The $\alpha$ quantile of a random variable $Y$, conditional on $X = x$ (notation: $Y_{\alpha|X=x}$) is the smallest value $z$ such that $Pr(Y \leq z|X = x) = \alpha$.

- If $F_{Y|X=x}$ is the conditional CDF of $Y$, then the $\alpha$-conditional quantile is

$$Y_{\alpha|X=x} = \inf y : \alpha \leq F_{Y|X=x}(y).$$

- When $\alpha = 0.5$, we are talking about the conditional median $Y_{0.5|X=x}$, but we could be interested in other quantiles, too.

- The linear regression model is focused on the conditional mean of the dependent variable.
- However, when looking at economic policies, we're often interested in distributional effects:
  - we may like to know how the rich and poor may be differentially affected by a policy that provides a public good
  - or we might like to know how a training program affects low-performing students compared to high-performing students
- For these sorts of issues, we're not concerned with the average agent: we want to know about the extremes, too.

## 21.1 Quantiles of the linear regression model

The classical linear regression model $y_t = x_t'\beta + \epsilon_t$ with normal errors implies that the distribution of $y_t$ conditional on $x_t$ is

$$y_t \sim N(x_t'\beta, \sigma^2)$$

- Note that $Pr(Y < x'\beta | X = x) = 0.5$ when the model follows the classical assumptions with normal errors, because the normal distribution is symmetric about the mean, so the mean and the median are the same, that is, $Y_{0.5|X=x} = x'\beta$.

- One can estimate the conditional median just by using the fitted conditional mean, because the mean and median are the same, given normality.

How about other quantiles? We have $y = x'\beta + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$.

- Conditional on $x$, $x'\beta$ is given, and the distribution of $\epsilon$ does not depend on $x$.

- Note that $\epsilon/\sigma$ is standard normal, and the $\alpha$ quantile of $\epsilon/\sigma$ is simply the inverse of the standard normal CDF evaulated at $\alpha$, $\Phi^{-1}(\alpha)$, where $\Phi$ is the standard normal CDF function.

- The probit function $\Phi^{-1}(\alpha)$ is tabulated (or can be found in Julia using `using Distributions; y = quantile.(Normal(),range(0.001,stop=0.999,` It is plotted in Figure 21.1.

The $\alpha$ quantile of $\epsilon$ is $\sigma\Phi^{-1}(\alpha)$. Thus, the $\alpha$ conditional quantile of $y$ is $Y_{\alpha|X=x} = x'\beta + \sigma\Phi^{-1}(\alpha)$. Some quantiles are pictured in Figure 21.2. These give confidence intervals for the the fitted value, $x'\beta$.

- The conditional quantiles for the classical model are *parallel*, *linear* functions of $x$

Figure 21.1: Inverse CDF for N(0,1)

Figure 21.2: Quantiles of classical linear regression model

- all have the same slope: the only thing that changes with $\alpha$ is the intercept $\sigma\Phi^{-1}(\alpha)$.

- If the error is heteroscedastic, so that $\sigma = \sigma(x)$, quantiles can have <u>different slopes</u>, and given quantiles may be nonlinear functions of $x$, depending on the form of the heteroscedasticity. *Draw a picture.*

## 21.2 Fully nonparametric conditional quantiles

To compute conditional quantiles for the classical linear model, we used the assumption of normality. Can we estimate conditional quantiles without making distributional assumptions? Yes, we can! (nod to Obama) (a note from 2018: those were the good old days!). You can do fully nonparametric conditional density estimation, as in Chapter 20, and use the fitted conditional density to compute quantiles.

- Note that estimating quantiles where $\alpha$ is close to 0 or 1 is difficult, because you have few observations that lie in the neighborhood of the quantile, so you should expect a large variance if you go the nonparametric route. For more central quantiles, like the median, this will be less of a problem.

- For this reason, we may go the *semi-parametric* route, which imposes more structure. When people talk about quantile regression, they usually mean

the semi-parametric approach.

## 21.3 Quantile regression as a semi-parametric estimator

The most widely used method does not take either of the extreme positions, it is not fully parametric, like the linear regression model with known distribution of errors, but some parametric restrictions are made, to improve efficiency compared to the fully nonparametric approach.

- The assumption is that the $\alpha$-conditional quantile of the dependent variable $Y$ is a linear function of the conditioning variables $X$: $Y_{\alpha|X=x} = x'\beta_\alpha$.
- This is a generalization of what we get from the classical model with normality, where the slopes of the quantiles with respect to the regressors are constant for all $\alpha$:
  - For the classical model with normality, $\frac{\partial}{\partial x} Y_{\alpha|X=x} = \beta$.

- With the assumption of linear quantiles without distributional assumptions, $\frac{\partial}{\partial x} Y_{\alpha|X=x} = \beta_\alpha$, so the slopes (and constants) are allowed to change with $\alpha$.

- This is a step in the direction of flexibility, but it also means we need to estimate many parameters if we're interested in many quantiles: there may be an efficiency loss due to using many parameters to avoid distributional assumptions.

- The question is how to estimate $\beta_\alpha$ when we don't make distributional assumptions.

It turns out that the problem can be expressed as an extremum estimator: $\widehat{\beta_\alpha} = \arg\min s_n(\beta)$ where

$$s_n(\beta) = \sum_{i=1}^{n} \left[1(y_i \geq x_i'\beta_\alpha)\alpha + 1(y_i < x_i'\beta_\alpha)(1-\alpha)\right] |y_i - x_i'\beta_\alpha|$$

First, suppose that $\alpha = 0.5$, so we are estimating the median. Then the objective simplifies to minimizing the absolute deviations:

$$s_n(\beta) = \sum_{i=1}^{n} |y_i - x_i'\beta_\alpha|$$

The presence of the weights in the general version accounts for the fact that if we're estimating the $\alpha = 0.1$ quantile, we expect 90% of the $y_i$ to be greater than $x_i'\beta_\alpha$, and only 10% to be smaller. We need to down-weight the likely events and up-weight the unlikely events so that the objective function minimizes at the appropriate place.

- One note is that median regression may be a useful means of dealing with data that satisfies the classical assumptions, except for contamination by outliers. *In class, use Gretl to show this.*

- Note that the quantile regression objective function is discontinuous. Minimization can be done quickly using linear programming. BFGS won't work.

- the asymptotic distribution is normal, with the sandwich form typical of extremum estimators. Estimation of the terms is not completely straightforward, so methods like bootstrapping may be preferable.

- the asymptotic variance depends upon which quantile we're estimating. When $\alpha$ is close to 0 or 1, the asymptotic variance becomes large, and the asymptotic approximation is unreliable for the small sample distribution.

- Extreme quantiles are hard to estimate with precision, because the data is sparse in those regions.

The artificial data set quantile.gdt allows you to explore quantile regression with GRETL, and to see how median regression can help to deal with data contamination.

- If you do quantile regression of the variable y versus x, we are in a situation where the assumptions of the classical model hold. Quantiles all have approximately the same slope (the true value is 1).
- With heteroscedastic data, the quantiles have different slopes.
- see Figure 21.3

Figure 21.3: Quantile regression results

(a) homoscedastic data

(b) heteroscedastic data

**Exercise 90.** Suppose that $y$ depends on a single regressor, $x$. Think about how you could do quantile regression estimation where the quantiles are nonlinear functions of $x$.

## 21.4 Returns to schooling: quantile regression, quantile IV regression, and Bayesian GMM via MCMC

Card (1993) presents an analysis of returns to schooling using the data from the National Longitudinal Survey of Young Men, for those interviewed in 1976. Card presents OLS and instrumental variables estimates for a number of specifications, using college proximity as an instrument for years of education, and age as an instrument for experience. Here, we work with the simple model from column (1) of Card's Table 2. Let's consider estimation of conditional quantiles for the model. The model is

$$\begin{aligned} Q_{\ln W|X}(\tau) &= \beta_0(\tau) + \beta_{EDUC}(\tau)EDUC + \beta_X(\tau)EXP + \beta_{EXP^2}(\tau)\frac{EXP^2}{100} \\ &+ \beta_{BLACK}(\tau)BLACK + \beta_{SMSA}(\tau)SMSA + \beta_{SOUTH}(\tau)SOUTH \\ &\equiv X\beta(\tau) \end{aligned}$$

- the dependent variable $\ln W$ is log hourly earnings (in cents)
- the regressors are years of education (EDUC), experience (EXP), experience squared divided by 100, a black indicator (BLACK), a metropolitan area indicator (SMSA), and a South indicator (SOUTH).
- We explore estimation of quantiles treating all variables as exogenous, or treating education and experience as endogenous, and the others as exogenous.
- When education and experience are treated as endogenous, we use proximity to an accredited four year college (NEARC4) as an instrumental variable. EXPER is defined as EXPER = AGE-EDUC-6, so if EDUC is endogenous, so is EXPER. We use AGE as an instrument for EXPER.

- If all variables are taken as exogenous, then quantile regression (QR) estimates may be obtained by standard methods, as implemented in the GRETL software package.

- The Card data set is provided with the Wooldridge data set for GRETL, see the GRETL web page. A version prepared for the model used here is card.gdt .

- QR results from GRETL for EDUC are in Figure 21.4. Note that the QR results are pretty close to the OLS results, for all quantiles, and there's no clear pattern of the effect of education differing across quantiles.

- The effect of an additional year of education on earnings is about 7-8%, all across the distribution.

Figure 21.4: QR results for the Card data, $\tau$ sequence

(a) Estimated $\beta_{EDUC}(\tau)$ as a function of $\tau$, with 95% confidence band

(b) Estimated $\beta_0(\tau)$ as a function of $\tau$, with 95% confidence band

- If education and experience are taken as endogenous, ordinary quantile regression will give biased estimates, just as OLS is biased and inconsistent with endogenous regressors.

- We may use an instrumental variables version of quantile regression, due to

Chernozhukov and Hansen (2005). They show that the moment conditions

$$m_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} Z_i \left(\tau - 1\left[y_i \leq X_i \beta(\tau)\right]\right)$$

(where $\theta = \beta(\tau)$) have expectation zero at the true parameter values, and thus can be used for GMM estimation.

- We can show that, at the true parameter values

$$\sqrt{n} m_n(\theta_0) \to^d N(0, (\tau - \tau^2) Q_Z)$$

so an estimate of the efficient weight matrix is the inverse of $\hat{\Sigma} = \frac{(\tau - \tau^2)}{n} \Sigma_i Z_i Z_i'$.

- The problem is that these moment conditions are discontinuous functions of the parameters, due to the indicator function, so gradient-based optimization methods will not work for computing the GMM estimates.

- To deal with this problem, we can consider using the MCMC methods proposed by Chernozhukov and Hong (2003) to compute a Bayesian version of the GMM estimator.

- This estimator works with the asymptotic distribution of the moment conditions to define the likelihood used in MCMC, rather than the full sample likelihood function, but otherwise, it is standard MCMC.

  - the use of moment conditions is a dimension reducing operation: the full sample likelihood requires knowing the distribution of $n$ (growing with the sample size) random variables, while the use of moment conditions and their asymptotic distribution only requires knowing the (asymptotic) distribution of $G$ (fixed and finite) random variables

  - thus, GMM is like a limited information ML estimator, with the asymptotic distribution substituting the actual small sample distribution.

- The model is implemented in QIVmodel.jl , and the estimation by MCMC

is done in QIVbyMCMC . It may be of interest to examine the code to see how posterior means and 90% confidence intervals are computed using the Chernozhukhov-Hong method.

- For those of you interested in MCMC, there is the file PlayWithMCMC.jl , which studies the MCMC chain a bit. The basic proposal draws the parameter from independent normal densities. The second proposal draws the parameters from a joint normal density that accounts for correlations in the posterior. The empirical results reported below don't depend on which proposal is used, though, as a long enough chain was used so that the difference washes out. This is an issue of computational efficiency, not one of statistical reliability. To obtain reliable results with a shorter chain, the proposal density should be chosen to ensure good mixing (sampling from the whole support of the posterior).

Figure 21.5: Two chains for $\beta_0(\tau = 0.5)$, independent and correlated proposals

(a) Independent proposals, poor mixing

(b) Correlated proposals, better mixing

- The results are in Figure 21.6, for $\beta_{EDUC}(\tau)$ and $\beta_0(\tau)$.
- We can see that the IVQR results are substantially different from the ordinary QR results in Figure 21.4.
  - The effect of education, according to the IVQR results, is substantially larger, for all quantiles, with an additional year of education increasing all quantiles, except the 40th, by more than 20%. This is good news for the people in the U.S. that have to take out enormous student loans. Given the cost of college tuition in the U.S., the miserable 7% return that OLS and ordinary QR find would probably not be enough to induce people to take out loans. So, we have external reasons to believe that this higher number may be more realistic. It would be interesting to study the evolution of returns over time, and compare them to the cost of education.
  - There is a U shape, with a greater effect at the lower and higher quan-

tiles.

- The confidence bands are broader for the IV version, which is to be expected. This is similar to what happens with ordinary IV and OLS.

- The results are quite similar to those of Chernozhukov and Hansen (2006), who estimate a similar model using the Angrist and Keueger (1991) data (this is the influential paper that used quarter of birth as an instrument for education).

Figure 21.6: IV-QR results

(a) Estimated $\beta_{EDUC}(\tau)$ as a function of $\tau$, with 90% confidence band

(b) Estimated $\beta_0(\tau)$ as a function of $\tau$, with 90% confidence band

# Chapter 22

# Simulation-based methods for estimation and inference

**Readings**: Cameron and Trivedi (2005), Ch. 12; Gourieroux and Monfort (1996). There are many articles. Some of the seminal papers are McFadden (1989), Pakes

and Pollard (1989), Gouriéroux et al. (1993), Smith (1993), Duffie and Singleton (1993), Gallant and Tauchen (1996).

- Human brain power is perhaps growing over time, but not as fast as thumb dexterity, I would argue.
- On the other hand, computing power is growing more or less exponentially, according to Moore's Law.
- Any economist would argue that we need to use inputs in proportion to their relative prices, which means that we should increasingly be using computers to make advancements in econometrics (and, maybe, our thumbs, too).
- Simulation-based methods do just that. When intensive use of computer power is contemplated, it is possible to do things that are otherwise infeasible:
  - obtaining more accurate results that what asymptotic theory gives us, using methods like bootstrapping,
  - performing estimation of models that are complex enough so that analytic expressions for objective functions that define conventional estimators (e.g., ML, GMM) are not available. Once you go down this

rabbit hole, you can estimate *very* complex models. Keeping in mind that a model is used to extract the essential, interesting features of a problem, we don't necessarily want to go too far in adding complexity.

- Simulation based estimation, especially the method of simulated moments, has become quite standard in applied research, so it is important to understand how it works.

## 22.1 Motivation

Simulation methods are of interest when the DGP is fully characterized by a parameter vector, so that simulated data can be generated, but the likelihood function and/or analytic moments of the observable variables are not calculable, so that ordinary MLE or GMM estimation is not possible.

- Many moderately complex models result in intractable likelihoods or moments.
- Simulation-based estimation methods open up the possibility to estimate truly complex models.
- The desirability of introducing a great deal of complexity may be an issue[1], but it least it becomes a possibility.

[1] Remember that a model is an abstraction from reality, and abstraction helps us to isolate the important features of a phenomenon.

## Example: Multinomial and/or dynamic discrete response models

(following McFadden (1989), which is one of the seminal articles behind the simulation-based estimation boom)

Let $y_i^*$ be a latent random vector of dimension $m$. Suppose that

$$y_i^* = X_i\beta + \varepsilon_i$$

where $X_i$ is $m \times K$. Suppose that

$$\varepsilon_i \sim N(0, \Omega) \tag{22.1}$$

Henceforth drop the $i$ subscript when it is not needed for clarity.

- $y^*$ is not observed. Rather, we observe a the result of applying the mapping $\tau : R^m \to \{0,1\}^m$

$$y = \tau(y^*)$$

such that each element of $y$ (also an $m$-vector) is either zero or one (in some cases only one element will be one). The original $y^*$ is a vector of continuous random variables, but $y$ is a vector of binary discrete random variables.

- Define

$$A_i = A(y_i) = \{y^* | \tau(y^*) = y_i\}$$

So, $A_i$ is the set of $y^*$ such that the vector $y_i$ is the outcome of the mapping. Suppose data is generated by random sampling of $(y_i, X_i)$. In this case the elements of $y_i$ may not be independent of one another (and clearly are not if $\Omega$ is not diagonal). However, $y_i$ is independent of $y_j$, $i \neq j$.

- Let $\theta = (\beta', \text{vech}(\Omega)')'$ be the vector of parameters of the model (vechfunction vectorizes the lower triangle of a square matrix). The contribution of the $i^{th}$ observation to the likelihood function is

$$p_i(\theta) = \int_{A_i} \phi(y_i^* - X_i\beta, \Omega) dy_i^*$$

where

$$\phi(\varepsilon, \Omega) = (2\pi)^{-M/2} \left|\Omega\right|^{-1/2} \exp\left[\frac{-\varepsilon'\Omega^{-1}\varepsilon}{2}\right]$$

is the multivariate normal density of an $M$ -dimensional random vector. The log-likelihood function is

$$\ln \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln p_i(\theta).$$

- The problem is that evaluation of $\mathcal{L}_i(\theta)$ and its derivative w.r.t. $\theta$ by standard methods of numeric integration such as quadrature is computationally infeasible when $m$ (the dimension of $y$) is higher than 3 or 4 (as long as there are no restrictions on $\Omega$).

- The mapping $\tau(y^*)$ has not been made specific. This setup is quite general: for different choices of $\tau(y^*)$ it nests the case of dynamic binary discrete choice models as well as the case of multinomial discrete choice (the choice of one out of a finite set of alternatives). See McFadden's paper for more details.

- The point here is that it is not too hard to come up with models for which we can draw data, given parameters, but for which we can't compute the likelihood function.

## Example: Marginalization of latent variables

Economic data often presents substantial heterogeneity that may be difficult to model. A possibility is to introduce latent random variables. This can cause the problem that there may be no known closed form for the distribution of observable variables after marginalizing out the unobservable latent variables. For example, count data (that takes values $0, 1, 2, 3, ...$) is often modeled using the Poisson distribution

$$\Pr(Y = y_i) = \frac{\exp(-\lambda)\lambda^i}{y_i!}$$

The mean and variance of the Poisson distribution are both equal to $\lambda$ :

$$\mathcal{E}(y) = V(y) = \lambda.$$

Often, one parameterizes the conditional mean as

$$\lambda_i = \exp(X_i\beta).$$

This ensures that the mean is positive (as it must be). Estimation by ML is straightforward.

Often, count data exhibits "overdispersion" which simply means that

$$V(y) > \mathcal{E}(y).$$

If this is the case, a solution is to use the negative binomial distribution rather than the Poisson. An alternative is to introduce a latent variable that reflects heterogeneity into the specification:

$$\lambda_i = \exp(X_i\beta + \eta_i)$$

where $\eta_i$ has some specified density with support $S$ (this density may depend on additional parameters). Let $d\mu(\eta_i)$ be the density of $\eta_i$. The marginal density of $y$ is

$$\Pr(Y = y_i|X_i) = \int_S \frac{\exp\left[-\exp(X_i\beta + \eta_i)\right]\left[\exp(X_i\beta + \eta_i)\right]^{y_i}}{y_i!} d\mu(\eta_i)$$

- In some cases, this will have a closed-form solution (one can derive the negative binomial distribution in this way if $\eta$ has a certain exponential distribution - see equation 15.22)

- Often this will not be possible. In this case, simulation is a means of calculating $\Pr(Y = y_i|X_i)$, which is then used to do ML estimation.

  - we draw $\eta$ from its marginal distribution, then we draw $y$ from its distribution conditional on $\eta$ and $X$, in this case the Poisson distribution. We make a number of such draws, and average them. That gives a simulated $\Pr(Y = y_i|X_i)$, where $\eta$ has been marginalized out.

  - This would be an example of the Simulated Maximum Likelihood (SML) estimation.

- In this case, since there is only one latent variable, using numeric integration methods (quadrature) is probably a better choice. But with more random

parameters, and thus more latent variables, quadrature becomes too costly and/or inaccurate.

## 22.2 Simulated maximum likelihood (SML)

For simplicity, consider cross-sectional data. An ML estimator solves

$$\hat{\theta}_{ML} = \arg\max s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \ln f(y_t|X_t, \theta)$$

where $f(y_t|X_t, \theta)$ is the density function of the $t^{th}$ observation. When $f(y_t|X_t, \theta)$ does not have a known closed form, $\hat{\theta}_{ML}$ is an infeasible estimator. However, it may be possible to define a random function such that

$$\mathcal{E}_\nu p(\nu, y_t|X_t, \theta) = f(y_t|X_t, \theta)$$

where the density of $\nu$ is known. If this is the case, the simulator

$$\tilde{f}\,(y_t|X_t, \theta) = \frac{1}{H} \sum_{s=1}^{H} p(\nu_{ts}, y_t|X_t, \theta)$$

is unbiased for $f(y_t|X_t, \theta)$.

- The SML simply substitutes $\tilde{f}\left(y_t|X_t,\theta\right)$ in place of $f(y_t|X_t,\theta)$ in the log-likelihood function, that is

$$\hat{\theta}_{SML} = \arg\max s_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln \tilde{f}\left(y_t|X_t,\theta\right)$$

## Properties

The properties of the SML estimator depend on how $H$ is set. The following is taken from Lee (1995).

**Theorem 91.** *[Lee] 1) if* $\lim_{n\to\infty} n^{1/2}/H = 0$, *then*

$$\sqrt{n}\left(\hat{\theta}_{SML} - \theta^0\right) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\theta^0))$$

*2) if* $\lim_{n\to\infty} n^{1/2}/H = \lambda$, $\lambda$ *a finite constant, then*

$$\sqrt{n}\left(\hat{\theta}_{SML} - \theta^0\right) \xrightarrow{d} N(B, \mathcal{I}^{-1}(\theta^0))$$

*where $B$ is a finite vector of constants.*

- This means that the SML estimator is asymptotically biased if $H$ doesn't grow faster than $n^{1/2}$.
- The covariance matrix is the typical inverse of the information matrix, so

that as long as $H$ grows fast enough, the estimator is consistent and fully asymptotically efficient.

- SML is actually not used nearly as often as is the method of simulated moments, in one of its variations, probably because one needs to use a large number of simulations to drive bias down to acceptable levels.

## 22.3 Method of simulated moments (MSM)

Suppose we have a data generating process DGP$(y|x, \theta)$ which is simulable, given $\theta$ and exogenous variables $x$, but is such that the density of $y$ is not calculable.

A formulation of the GMM estimator which we have studied is based upon the moment conditions $m_t = z_t \epsilon_t(\theta)$, where $\epsilon_t(\theta)$ has conditional expectation equal to zero when evaluated at the true parameter value. Consider

$$\epsilon_t(\theta) = K(y_t, x_t) - k(x_t, \theta)$$

where $k(x_t, \theta) = E_\theta K(y_t, x_t | I_t)$, where $I_t$ is the information set at time $t$. Then, at the true parameter $\theta^0$ that generated the data, $E_{\theta^0} K(y_t, x_t | I_t) = k(x_t, \theta^0)$. From this, we could base GMM estimation on $\epsilon_t(\theta)$, crossed with instrumental variables drawn from $I_t$.

- However, assume that we can't compute $k(x_t, \theta) = E_\theta K(y_t, x_t | I_t)$, for some reason.

- Nevertheless, $k(x_t, \theta)$ is readily simulated (meaning that we can make random draws of $k(x_t, \theta)$) using

$$\widetilde{k}\left(x_t, \theta\right) = \frac{1}{H} \sum_{h=1}^{H} K(\widetilde{y}_t^h, x_t)$$

  where $\widetilde{y}_t^h$ is drawn from DGP$(y|x, \theta)$.

- Note that $E_\theta K(\widetilde{y}_t^h, x_t | I_t) = k(x_t, \theta)$, and, by the law of large numbers, $\widetilde{k}\left(x_t, \theta\right) \overset{a.s.}{\to} k\left(x_t, \theta\right)$, as $H \to \infty$.

- This allows us to form the moment conditions

$$\widetilde{m}_t(\theta) = \left[K(y_t, x_t) - \widetilde{k}\left(x_t, \theta\right)\right] z_t \tag{22.2}$$

where $z_t$ is drawn from the information set. As before, form

$$\begin{aligned}\widetilde{m}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{m}_t(\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[K(y_t, x_t) - \frac{1}{H}\sum_{h=1}^{H}k(\widetilde{y}_t^h, x_t)\right]z_t \qquad (22.3)\end{aligned}$$

– note: I'm using $\widetilde{m}_n(\theta)$ for the average moment conditions, in place of $\bar{m}_n(\theta)$, which appears in Definition 57, to highlight that simulations are used. With this simple difference, we form the GMM criterion and estimate as usual. Note that the unbiased simulator $k(\widetilde{y}_t^h, x_t)$ appears linearly within the sums.

## Properties

Suppose that the optimal weighting matrix is used. McFadden (1989) and Pakes and Pollard (1989) show that the asymptotic distribution of the MSM estimator is very similar to that of the infeasible GMM estimator. In particular, assuming that the optimal weighting matrix is used, and for $H$ finite,

$$\sqrt{n}\left(\hat{\theta}_{MSM} - \theta^0\right) \xrightarrow{d} N\left[0, \left(1 + \frac{1}{H}\right)\left(D_\infty \Omega^{-1} D_\infty'\right)^{-1}\right] \tag{22.4}$$

where $\left(D_\infty \Omega^{-1} D_\infty'\right)^{-1}$ is the asymptotic variance of the infeasible GMM estimator.

- That is, the asymptotic variance is inflated by a factor $1 + 1/H$. For this reason the MSM estimator is not fully asymptotically efficient relative to the infeasible GMM estimator, for $H$ finite, but the efficiency loss is small and controllable, by setting $H$ reasonably large.

- The estimator is asymptotically unbiased even for $H = 1$. This is an advantage relative to SML.

- If one doesn't use the optimal weighting matrix, the asymptotic varcov is just the ordinary GMM varcov, inflated by $1 + 1/H$. Even for moderate $H$, the variance inflation can be quite small.

The above presentation is in terms of a specific set of moment conditions based upon the conditional mean and instruments. The MSM can be applied to moment conditions of other forms, too.

- A leading example is **Indirect Inference** (Gouriéroux et al. (2010)) where we set $\widetilde{m}_n(\theta) = \hat{\phi}\text{-}\frac{1}{S}\,\Sigma\,\tilde{\phi}^s(\theta)$, and then we just do ordinary GMM. Note that this is an average over $S$, not over $n$, so this is a departure from the standard presentation of GMM.

- Here, $\hat{\phi}$ is an extremum estimator corresponding to some auxiliary model. The $\tilde{\phi}^s(\theta)$ are the same extremum estimator, applied to simulated data generated from the model. The logic is that $\hat{\phi}$ will converge to a pseudo-true value, and $\tilde{\phi}^s(\theta)$ will converge to another pseudo-true value, depending on the value of $\theta$ that generated the simulated data. When $\theta = \theta^0$, the two pseudo-true values will be the same. Trying to make the average of the simulated estimators as close as possible to the estimator generated by the real data will cause the MSM estimator to be consistent, given identification.

- For such an estimator to have good efficiency, we need the auxiliary model to "fit well": it should pick up the relevant features of the data.

- one can combine moment conditions using indirect inference-type moments with the usual MSM moments.

- a drawback of the II estimator is that the auxiliary model must be estimated many times. This is not a problem if it's a simple linear model, but it could be a problem if it's more complicated. For efficiency, we need a good fit, and a simple linear model may not provide this. The EMM (Gallant and Tauchen (1996)) estimator discussed below is asymptotically equivalent to II, and it requires the auxiliary model to be estimated only once. See Section 24.1 for some more discussion, or go directly to the article).

- So, as Gallant and Tauchen ask, "Which Moments to Match?" is the fundamental question when doing MSM or EMM (or ordinary GMM, for that matter).

## Comments

Why is SML inconsistent when the number of simulations is finite, while MSM is? The reason is that SML is based upon an average of **logarithms** of an unbiased simulator (the densities of the observations). To use the multinomial probit model as an example, the log-likelihood function is

$$\ln \mathcal{L}(\beta, \Omega) = \frac{1}{n} \sum_{i=1}^{n} y_i' \ln p_i(\beta, \Omega)$$

The SML version is

$$\ln \mathcal{L}(\beta, \Omega) = \frac{1}{n} \sum_{i=1}^{n} y_i' \ln \tilde{p}_i(\beta, \Omega)$$

The problem is that

$$E \ln(\tilde{p}_i(\beta, \Omega)) \neq \ln(\mathcal{E}\tilde{p}_i(\beta, \Omega))$$

in spite of the fact that

$$\mathcal{E}\tilde{p}_i(\beta, \Omega) = p_i(\beta, \Omega)$$

due to the fact that $\ln(\cdot)$ is a nonlinear transformation. The only way for the two to be equal (in the limit) is if $H$ tends to infinite so that $\tilde{p}(\cdot)$ tends to $p(\cdot)$.

The reason that MSM does not suffer from this problem is that in this case the unbiased simulator appears *linearly* within every sum of terms, and it appears within a sum over $n$ (see equation [22.3]). Therefore the SLLN applies to cancel out simulation errors, from which we get consistency. That is, using simple notation for the random sampling case, the moment conditions

$$\begin{aligned}\tilde{m}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left[K(y_t, x_t) - \frac{1}{H}\sum_{h=1}^{H} k(\tilde{y}_t^h, x_t)\right] z_t && (22.5)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[k(x_t, \theta^0) + \varepsilon_t - \frac{1}{H}\sum_{h=1}^{H}\left[k(x_t, \theta) + \tilde{\varepsilon}_{ht}\right]\right] z_t && (22.6)\end{aligned}$$

converge almost surely to

$$\tilde{m}_\infty(\theta) = \int \left[k(x, \theta^0) - k(x, \theta)\right] z(x) d\mu(x).$$

(note: $z_t$ is assume to be made up of functions of $x_t$). The objective function converges to

$$s_\infty(\theta) = \tilde{m}_\infty(\theta)' \Omega_\infty^{-1} \tilde{m}_\infty(\theta)$$

which obviously has a minimum at $\theta^0$, henceforth consistency.

- If you look at equation 22.6 a bit, you will see why the variance inflation factor is $(1 + \frac{1}{H})$.

# Numerical issues: "Chatter"

When a function of interest is approximated with simulations, for example, approximating $E_\theta K(y_t, x_t|I_t)$ with the average of simulations $\frac{1}{H}\Sigma_{h=1}^{H} k(\widetilde{y}_t^h, x_t)$, where each $\widetilde{y}_t^h$ is drawn at $\theta$, we need to recognize that if we repeat the simulation process, drawing $H$ new samples, the value of $\frac{1}{H}\Sigma_{h=1}^{H} k(\widetilde{y}_t^h, x_t)$ will change, even if $\theta$ is held constant, unless special care is taken. When the random elements of the simulation process are not held fixed for the simulations, this function is everywhere discontinuous.

- From a theoretical point of view, this is not a problem, as the function will be stochastically equicontinuous, under reasonable assumptions, and this still allows for consistency and asymptotic normality.

- However, from the point of view of computing an extremum estimator, using gradient-based optimization, or for computing a covariance matrix that depends on taking derivatives, it is a problem.

- In certain cases, when $K(y_t, x_t)$ is a continuous function in $y$, is is possible to control chatter by keeping the underlying random draws of latent variables fixed across simulations, as $\theta$ varies. The example in the next section does this.

- In other cases, this is harder. For example, if $y$ is a discrete random variable, then the function will be discontinuous in $\theta$, even if random draws of latent variables are held fixed. In these cases, extremum estimators have been devised that rely on forms of smoothing of the functions. However, it is quite easy to move to a Bayesian perspective, where we don't need to optimize, and this will also be easier to program, as we won't need to keep draws of latent variable fixed. The practical example at the end of the chapter gives an example.

## 22.4 Example: stochastic volatility

### MSM and Bayesian MSM

The simple stochastic volatility model from Section 17.3 is

$$
\begin{aligned}
y_t &= \phi \exp(h_t/2)\epsilon_t \\
h_t &= \rho h_{t-1} + \sigma u_t
\end{aligned}
$$

Typical data and a nonparametric density plot looks like what we see in Figure 22.1. Note the volatility clusters, leptokurtosis, and the fat tails of the density.

The Julia script EstimateSV.jl estimates the stochastic volatility model by MSM, implemented as a two step GMM estimator, and then goes on to do the Chernozhukhov-Hong MCMC version of MSM.

- The minimization is by simulated annealing, to ensure robustness against numerical problems. I haven't tried with gradient-based methods.

Figure 22.1: SV model, typical data and density

- Examine the script to see how the objective function is "bullet-proofed"
- Study what auxiliary statistics are used to define the moments, in the file SVlib.jl (this is the key to success or failure when doing moment-based estimation), and think about the problem to try to come up with some better ones.
- Also in that file, see how latent variables are held fixed across simulations,

to eliminate chatter.

- The results use the sample that is pictured in Figure 22.1 (or, you have the option of generating a new sample).
- check the script to see how to compute standard errors, etc.

- the results for extremum MSM are in Figure 22.2.
- MCMC gives us a full posterior for the parameters, and we can use posterior quantiles to define alternative confidence intervals. For this sample, we get the results in Figure 22.3.
- If you compare the MSM and Bayesian MSM results, you'll see that the point estimates are very similar, but there are some differences in the CIs: the extremum estimator CIs are broader, in this case.

Figure 22.2: MSM for SV model

```
extremum estimation results:
true values, estimates, st. error, and limits of 90% CI
```

| true value | estimate | std. err. | CI lower | CI upper |
|---|---|---|---|---|
| 0.69212 | 0.73287 | 0.06437 | 0.62706 | 0.83869 |
| 0.90000 | 0.91615 | 0.04561 | 0.84117 | 0.99113 |
| 0.36300 | 0.30680 | 0.08892 | 0.16062 | 0.45298 |

Unfortunately, inferences using overidentified GMM-type estimators are often not reliable in finite samples (Donald et al. (2009), Hansen et al. (1996), Tauchen (1986)). The script SV_MonteCarlo.jl runs 100 replications of estimation of the same model and computes confidence intervals using both asymptotic theory for MSM estimators, and using quantiles of the MCMC chain. Confidence interval coverage, which is the proportion of times the true parameters are inside the computed confidence intervals, is reported in Figure 22.4. In all cases, the true parameters are over-rejected, which is to say, the confidence intervals are tighter than they should be, and Type-I error will occur more often than it should. This

Figure 22.3: MCMC estimation using simulated moments and limited information quasi-likelihood

applies to both the extremum and Baysesian versions. The Bayesian version could perhaps be improved somewhat by more careful tuning of the MCMC algorithm (see below) but the extremum version does not use tuning, so the results are not dependent on this sort of qualification.

Figure 22.4: CI coverage, SV model, MSM and Bayesian MSM

```
julia> include("SV_MonteCarlo.jl")
99% CI, extremum
```

| ϕ | ρ | σ |
|---|---|---|
| 0.97000 | 0.91000 | 0.92000 |

99% CI, Bayes

| ϕ | ρ | σ |
|---|---|---|
| 0.88000 | 0.80000 | 0.80000 |

95% CI, extremum

| ϕ | ρ | σ |
|---|---|---|
| 0.90000 | 0.85000 | 0.84000 |

95% CI, Bayes

| ϕ | ρ | σ |
|---|---|---|
| 0.81000 | 0.73000 | 0.72000 |

90% CI, extremum

| ϕ | ρ | σ |
|---|---|---|
| 0.85000 | 0.79000 | 0.78000 |

90% CI, Bayes

| ϕ | ρ | σ |
|---|---|---|
| 0.69000 | 0.64000 | 0.61000 |

## Simulated Neural Moments

A solution to this problem is to reduce the dimension of the statistics by passing them through a trained neural net. This can be done using the package SimulatedNeuralMoments.jl. Using methods discussed there, one can obtain the CI coverage for the same SV model presented above that is seen in Figure 22.5. Note that when the neural statistics are used (columns labeled Z), the coverage is statistically correct. The results using the original statistics, without the neural net (the column labeled W) are a little better than what's in the previous figure, but are still far from correct. There are some differences in the Monte Carlo design which may explain these differences. The main difference is that, for the results here, the criterion function was the continuous updating version of GMM, whereas the previous results used two step GMM. The results, though, are qualitatively the same: without use of the neural net, the confidence intervals over-reject the true values, in some cases by serious margins. Creel (2021) reports Monte Carlo results for other models which confirm that processing using a neural net leads to

correct coverage of confidence intervals.

Figure 22.5: 90, 95 and 99% CI coverage, SV model, using simulated neural moments

**Table 2.** 90% confidence interval coverage for SV and ARMA models, using raw ($W$) or neural net ($Z$) statistics. Correct coverage rejected when outside 0.864–0.932.

| Model | Parameter | $W$ (CUE) | $Z$ (Two-Step) | $Z$ (CUE) |
|---|---|---|---|---|
| SV | $\phi$ | 0.876 | 0.884 | 0.912 |
| | $\rho$ | 0.732 | 0.976 | 0.910 |
| | $\sigma$ | 0.762 | 0.956 | 0.928 |

**Table 3.** 95% confidence interval coverage for SV and ARMA models, using raw ($W$) or neural net ($Z$) statistics. Correct coverage rejected when outside 0.924–0.974.

| Model | Parameter | $W$ (CUE) | $Z$ (Two-Step) | $Z$ (CUE) |
|---|---|---|---|---|
| SV | $\phi$ | 0.916 | 0.938 | 0.954 |
| | $\rho$ | 0.796 | 0.990 | 0.944 |
| | $\sigma$ | 0.824 | 0.976 | 0.958 |

**Table 4.** 99% confidence interval coverage for SV and ARMA models, using raw ($W$) or neural net ($Z$) statistics. Correct coverage rejected when outside 0.976–1.000.

| Model | Parameter | $W$ (CUE) | $Z$ (Two-Step) | $Z$ (CUE) |
|---|---|---|---|---|
| SV | $\phi$ | 0.936 | 0.968 | 0.990 |
| | $\rho$ | 0.848 | 0.998 | 0.978 |
| | $\sigma$ | 0.888 | 0.994 | 0.986 |

## 22.5 Simulated Neural Moments estimation of the DSGE model

Please see this page for a summary of estimation and inference using MCMC-MSM, where moments are based on a neural net that is fitted using simulations from the model, at parameter vectors drawn from the prior. This procedure leads to parameter estimates that have low bias and RMSE, and confidence intervals that have approximately proper coverage.

## 22.6 Exercises

1. (advanced, but even if you don't do this you should be able to describe what needs to be done) Write code to do SML estimation of the probit model. Do an estimation using data generated by a probit model. Compare the SML estimates to ML estimates.

2. do the same, but computing a MSM estimator of the probit model.

3. (more advanced) Do a little Monte Carlo study to compare ML, SML and MSMestimation of the probit model. Investigate how the number of simulations affect the two simulation-based estimators.

## 22.7 Practical summary

The practical summary for the Chapter is here.

# Chapter 23

# Notation and Review

- All vectors will be column vectors, unless they have a transpose symbol (or I forget to apply this rule - your help catching typos and er0rors is much appreciated). For example, if $x_t$ is a $p \times 1$ vector, $x_t'$ is a $1 \times p$ vector. When I refer to a $p$-vector, I mean a column vector.

## 23.1 Notation for differentiation of vectors and matrices

Gallant (1987a)

Let $s(\cdot) : \Re^p \to \Re$ be a real valued function of the $p$-vector $\theta$. Then $\frac{\partial s(\theta)}{\partial \theta}$ is organized as a $p$-vector,

$$\frac{\partial s(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial s(\theta)}{\partial \theta_1} \\ \frac{\partial s(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial s(\theta)}{\partial \theta_p} \end{bmatrix}$$

Following this convention, $\frac{\partial s(\theta)}{\partial \theta'}$ is a $1 \times p$ vector, and $\frac{\partial^2 s(\theta)}{\partial \theta \partial \theta'}$ is a $p \times p$ matrix. Also,

$$\frac{\partial^2 s(\theta)}{\partial \theta \partial \theta'} = \frac{\partial}{\partial \theta}\left(\frac{\partial s(\theta)}{\partial \theta'}\right) = \frac{\partial}{\partial \theta'}\left(\frac{\partial s(\theta)}{\partial \theta}\right).$$

**Exercise 92.** For $a$ and $x$ both $p$-vectors, show that $\frac{\partial a'x}{\partial x} = a$.

Let $f(\theta)$:$\Re^p \to \Re^n$ be a $n$-vector valued function of the $p$-vector $\theta$. Let $f(\theta)'$ be the $1 \times n$ valued transpose of $f$ . Then $\left(\frac{\partial}{\partial\theta} f(\theta)'\right)' = \frac{\partial}{\partial\theta'} f(\theta)$.

**Definition.** Product rule. Let $f(\theta)$:$\Re^p \to \Re^n$ and $h(\theta)$:$\Re^p \to \Re^n$ be $n$-vector valued functions of the $p$-vector $\theta$. Then

$$\frac{\partial}{\partial\theta'} h(\theta)' f(\theta) = h'\left(\frac{\partial}{\partial\theta'} f\right) + f'\left(\frac{\partial}{\partial\theta'} h\right)$$

has dimension $1 \times p$. Applying the transposition rule we get

$$\frac{\partial}{\partial\theta} h(\theta)' f(\theta) = \left(\frac{\partial}{\partial\theta} f'\right) h + \left(\frac{\partial}{\partial\theta} h'\right) f$$

which has dimension $p \times 1$.

**Exercise 93.** For $A$ a $p \times p$ matrix and $x$ a $p \times 1$ vector, show that $\frac{\partial x' A x}{\partial x} = (A + A')\, x$. Also, what is the result if $A$ is symmetric?

**Definition 94.** Chain rule. Let $f(\cdot)$:$\Re^p \to \Re^n$ a $n$-vector valued function of a $p$-vector argument, and let $g()$:$\Re^r \to \Re^p$ be a $p$-vector valued function of an

$r$-vector valued argument $\rho$. Then

$$\frac{\partial}{\partial \rho'} f\left[g\left(\rho\right)\right] = \frac{\partial}{\partial \theta'} f(\theta)\Big|_{\theta=g(\rho)} \frac{\partial}{\partial \rho'} g(\rho)$$

has dimension $n \times r$.

**Exercise 95.** For $x$ and $\beta$ both $p \times 1$ vectors, show that $\frac{\partial \exp(x'\beta)}{\partial \beta} = \exp(x'\beta)x$.

## 23.2 Convergence modes

**Readings:** Davidson, R. and J.G. MacKinnon, *Econometric Theory and Methods*, Ch. 4; Gallant, A.R., *An Introduction to Econometric Theory,* Ch. 4.

We will consider several modes of convergence. The first three modes discussed are simply for background. The stochastic modes are those which will be used later in the course.

**Definition 96.** A sequence is a mapping from the natural numbers $\{1, 2, ...\} =$

$\{n\}_{n=1}^{\infty} = \{n\}$ to some other set, so that the set is ordered according to the natural numbers associated with its elements.

## Real-valued sequences:

**Definition 97.** *[Convergence]* A real-valued sequence of vectors $\{a_n\}$ *converges* to the vector $a$ if for any $\varepsilon > 0$ there exists an integer $N_\varepsilon$ such that for all $n > N_\varepsilon$, $\| a_n - a \| < \varepsilon$ . $a$ is the *limit* of $a_n$, written $a_n \to a$.

## Deterministic real-valued functions

Consider a sequence of functions $\{f_n(\omega)\}$ where

$$f_n : \Omega \to T \subseteq \Re.$$

$\Omega$ may be an arbitrary set.

**Definition 98.** *[Pointwise convergence]* A sequence of functions $\{f_n(\omega)\}$ *con-*

*verges pointwise* on $\Omega$ to the function $f(\omega)$ if for all $\varepsilon > 0$ and $\omega \in \Omega$ there exists an integer $N_{\varepsilon\omega}$ such that

$$|f_n(\omega) - f(\omega)| < \varepsilon, \forall n > N_{\varepsilon\omega}.$$

It's important to note that $N_{\varepsilon\omega}$ depends upon $\omega$, so that converge may be much more rapid for certain $\omega$ than for others. Uniform convergence requires a similar rate of convergence throughout $\Omega$.

**Definition 99.** *[Uniform convergence]* A sequence of functions $\{f_n(\omega)\}$ *converges uniformly* on $\Omega$ to the function $f(\omega)$ if for any $\varepsilon > 0$ there exists an integer $N$ such that

$$\sup_{\omega \in \Omega} |f_n(\omega) - f(\omega)| < \varepsilon, \forall n > N.$$

(insert a diagram here showing the envelope around $f(\omega)$ in which $f_n(\omega)$ must lie).

## Stochastic sequences

In econometrics, we typically deal with stochastic sequences. Given a probability space $(\Omega, \mathcal{F}, P)$, recall that a random variable maps the sample space to the real line, i.e., $X(\omega) : \Omega \to \Re$. A sequence of random variables $\{X_n(\omega)\}$ is a collection of such mappings, *i.e.,* each $X_n(\omega)$ is a random variable with respect to the probability space $(\Omega, \mathcal{F}, P)$. For example, given the model $Y = X\beta^0 + \varepsilon$, the OLS estimator $\hat{\beta}_n = (X'X)^{-1} X'Y$, where $n$ is the sample size, can be used to form a sequence of random vectors $\{\hat{\beta}_n\}$. A number of modes of convergence are in use when dealing with sequences of random variables. Several such modes of convergence should already be familiar:

**Definition 100.** *[Convergence in probability]* Let $X_n(\omega)$ be a sequence of random variables, and let $X(\omega)$ be a random variable. Let $\mathcal{A}_n = \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}$. Then $\{X_n(\omega)\}$ converges in probability to $X(\omega)$ if

$$\lim_{n\to\infty} P(\mathcal{A}_n) = 0, \forall \varepsilon > 0.$$

Convergence in probability is written as $X_n \xrightarrow{p} X$, or $\text{plim}\, X_n = X$.

**Definition 101.** *[Almost sure convergence]* Let $X_n(\omega)$ be a sequence of random variables, and let $X(\omega)$ be a random variable. Let $\mathcal{A} = \{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$. Then $\{X_n(\omega)\}$ converges almost surely to $X(\omega)$ if

$$P(\mathcal{A}) = 1.$$

In other words, $X_n(\omega) \to X(\omega)$ (ordinary convergence of the two functions) except on a set $C = \Omega - \mathcal{A}$ such that $P(C) = 0$. Almost sure convergence is written as $X_n \xrightarrow{a.s.} X$, or $X_n \to X$, $a.s.$ One can show that

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X.$$

**Definition 102.** *[Convergence in distribution]* Let the r.v. $X_n$ have distribution function $F_n$ and the r.v. $X_n$ have distribution function $F$. If $F_n \to F$ at every continuity point of $F$, then $X_n$ converges in distribution to $X$.

Convergence in distribution is written as $X_n \xrightarrow{d} X$. It can be shown that convergence in probability implies convergence in distribution.

## Stochastic functions

Simple laws of large numbers (LLN's) allow us to directly conclude that $\hat{\beta}_n \xrightarrow{a.s.} \beta^0$ in the OLS example, since

$$\hat{\beta}_n = \beta^0 + \left(\frac{X'X}{n}\right)^{-1} \left(\frac{X'\varepsilon}{n}\right),$$

and $\frac{X'\varepsilon}{n} \xrightarrow{a.s.} 0$ by a SLLN. Note that this term is not a function of the parameter $\beta$. This easy proof is a result of the linearity of the model, which allows us to express the estimator in a way that separates parameters from random functions. In general, this is not possible. We often deal with the more complicated situation where the stochastic sequence depends on parameters in a manner that is not reducible to a simple sequence of random variables. In this case, we have a sequence

of random functions that depend on $\theta$: $\{X_n(\omega, \theta)\}$, where each $X_n(\omega, \theta)$ is a random variable with respect to a probability space $(\Omega, \mathcal{F}, P)$ and the parameter $\theta$ belongs to a parameter space $\theta \in \Theta$.

**Definition 103.** *[Uniform almost sure convergence]* $\{X_n(\omega, \theta)\}$ converges uniformly almost surely in $\Theta$ to $X(\omega, \theta)$ if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |X_n(\omega, \theta) - X(\omega, \theta)| = 0, \text{(a.s.)}$$

Implicit is the assumption that all $X_n(\omega, \theta)$ and $X(\omega, \theta)$ are random variables w.r.t. $(\Omega, \mathcal{F}, P)$ for all $\theta \in \Theta$. We'll indicate uniform almost sure convergence by $\overset{u.a.s.}{\to}$ and uniform convergence in probability by $\overset{u.p.}{\to}$ .

- An equivalent definition, based on the fact that "almost sure" means "with probability one" is

$$\Pr\left(\lim_{n\to\infty} \sup_{\theta\in\Theta} |X_n(\omega, \theta) - X(\omega, \theta)| = 0\right) = 1$$

This has a form similar to that of the definition of a.s. convergence - the essential difference is the addition of the sup.

## 23.3 Rates of convergence and asymptotic equality

It's often useful to have notation for the relative magnitudes of quantities. Quantities that are small relative to others can often be ignored, which simplifies analysis.

**Definition 104.** *[Little-o]* Let $f(n)$ and $g(n)$ be two real-valued functions. The notation $f(n) = o(g(n))$ means $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

**Definition 105.** *[Big-O]* Let $f(n)$ and $g(n)$ be two real-valued functions. The notation $f(n) = O(g(n))$ means there exists some $N$ such that for $n > N$, $\left|\frac{f(n)}{g(n)}\right| < K$, where $K$ is a finite constant.

This definition doesn't require that $\frac{f(n)}{g(n)}$ have a limit (it may fluctuate boundedly).

If $\{f_n\}$ and $\{g_n\}$ are sequences of random variables analogous definitions are

**Definition 106.** The notation $f(n) = o_p(g(n))$ means $\frac{f(n)}{g(n)} \xrightarrow{p} 0$.

**Example 107.** The least squares estimator $\hat{\theta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\theta^0 + \varepsilon) = \theta^0 + (X'X)^{-1}X'\varepsilon$. Since $\text{plim}\frac{(X'X)^{-1}X'\varepsilon}{1} = 0$, we can write $(X'X)^{-1}X'\varepsilon = o_p(1)$ and $\hat{\theta} = \theta^0 + o_p(1)$. Asymptotically, the term $o_p(1)$ is negligible. This is just a way of indicating that the LS estimator is consistent.

**Definition 108.** The notation $f(n) = O_p(g(n))$ means there exists some $N_\varepsilon$ such that for $\varepsilon > 0$ and all $n > N_\varepsilon$,

$$P\left(\left|\frac{f(n)}{g(n)}\right| < K_\varepsilon\right) > 1 - \varepsilon,$$

where $K_\varepsilon$ is a finite constant.

**Example 109.** If $X_n \sim N(0,1)$ then $X_n = O_p(1)$, since, given $\varepsilon$, there is always some $K_\varepsilon$ such that $P\left(|X_n| < K_\varepsilon\right) > 1 - \varepsilon$.

Useful rules:

- $O_p(n^p)O_p(n^q) = O_p(n^{p+q})$
- $o_p(n^p)o_p(n^q) = o_p(n^{p+q})$

**Example 110.** Consider a random sample of iid r.v.'s with mean 0 and variance $\sigma^2$. The estimator of the mean $\hat{\theta} = 1/n\,\Sigma_{i=1}^n x_i$ is asymptotically normally distributed, e.g., $n^{1/2}\hat{\theta} \overset{A}{\sim} N(0,\sigma^2)$. So $n^{1/2}\hat{\theta} = O_p(1)$, so $\hat{\theta} = O_p(n^{-1/2})$. Before we had $\hat{\theta} = o_p(1)$, now we have have the stronger result that relates the rate of convergence to the sample size..

**Example 111.** Now consider a random sample of iid r.v.'s with mean $\mu$ and variance $\sigma^2$. The estimator of the mean $\hat{\theta} = 1/n\,\Sigma_{i=1}^n x_i$ is asymptotically normally

distributed, e.g., $n^{1/2}\left(\hat{\theta}-\mu\right) \overset{A}{\sim} N(0,\sigma^2)$. So $n^{1/2}\left(\hat{\theta}-\mu\right) = O_p(1)$, so $\hat{\theta}-\mu = O_p(n^{-1/2})$, so $\hat{\theta} = O_p(1)$.

These two examples show that averages of centered (mean zero) quantities typically have plim 0, while averages of uncentered quantities have finite nonzero plims. Note that the definition of $O_p$ does not mean that $f(n)$ and $g(n)$ are of the same order. Asymptotic equality ensures that this is the case.

**Definition 112.** Two sequences of random variables $\{f_n\}$ and $\{g_n\}$ are asymptotically equal (written $f_n \overset{a}{=} g_n$) if

$$plim\left(\frac{f(n)}{g(n)}\right) = 1$$

Finally, analogous almost sure versions of $o_p$ and $O_p$ are defined in the obvious way.

## 23.4 Slutsky Theorem and Continuous Mapping Theorem

The following two theorems are important for getting the asymptotic distribution of estimators, and for test statistics that are derived from transformations of estimators. See Gallant (1997), Theorems 4.6 and 4.7. Statement of the theorems are here:

Slutsky Theorem

Continuous Mapping Theorem

1. For $a$ and $x$ both $p \times 1$ vectors, show that $D_x a'x = a$.

2. For $A$ a $p \times p$ matrix and $x$ a $p \times 1$ vector, show that $D_x^2 x'Ax = A + A'$.

3. For $x$ and $\beta$ both $p \times 1$ vectors, show that $D_\beta \exp x'\beta = \exp(x'\beta)x$.

4. For $x$ and $\beta$ both $p \times 1$ vectors, find the analytic expression for $D_\beta^2 \exp x'\beta$.

5. Write an Octave program that verifies each of the previous results by taking numeric derivatives. For a hint, type `help numgradient` and `help numhessian` inside octave.

# Chapter 24

# The attic

This holds material that is not really ready to be incorporated into the main body, or that I believe distracts from the flow, but that I don't want to lose. Basically, ignore it.

## 24.1 Efficient method of moments (EMM)

Note: this is a specific type of MSM estimator. I moved this out of the main text, as it will be of interest to a reduced group of students.

The choice of which moments upon which to base a GMM estimator can have very pronounced effects upon the efficiency of the estimator.

- A poor choice of moment conditions may lead to very inefficient estimators, and can even cause identification problems (as we've seen with the GMM problem set).

- The drawback of the above approach MSM is that the moment conditions used in estimation are selected arbitrarily. The asymptotic efficiency of the estimator may be low.

- The asymptotically optimal choice of moments would be the score vector of

the likelihood function,

$$m_t(\theta) = D_\theta \ln p_t(\theta \mid I_t)$$

As before, this choice is unavailable.

The efficient method of moments (EMM) (see Gallant and Tauchen (1996)) seeks to provide moment conditions that closely mimic the score vector. If the approximation is very good, the resulting estimator will be very nearly fully efficient.

The DGP is characterized by random sampling from the density

$$p(y_t|x_t, \theta^0) \equiv p_t(\theta^0)$$

We can define an auxiliary model, called the "score generator", which simply provides a (misspecified) parametric density

$$f(y|x_t, \lambda) \equiv f_t(\lambda)$$

- This density is known up to a parameter $\lambda$. We assume that this density function *is* calculable. Therefore quasi-ML estimation is possible. Specifically,

$$\hat{\lambda} = \arg\max_{\Lambda} s_n(\lambda) = \frac{1}{n} \sum_{t=1}^{n} \ln f_t(\lambda).$$

- After determining $\hat{\lambda}$ we can calculate the score functions $D_\lambda \ln f(y_t|x_t, \hat{\lambda})$.

- The important point is that even if the density is misspecified, there is a pseudo-true $\lambda^0$ for which the true expectation, taken with respect to the true but unknown density of $y$, $p(y|x_t, \theta^0)$, and then marginalized over $x$ is zero:

$$\exists \lambda^0 : \mathcal{E}_X \mathcal{E}_{Y|X} \left[ D_\lambda \ln f(y|x, \lambda^0) \right] = \int_X \int_{Y|X} D_\lambda \ln f(y|x, \lambda^0) p(y|x, \theta^0) dy d\mu(x) = 0$$

- We have seen in the section on QML that $\hat{\lambda} \xrightarrow{p} \lambda^0$; this suggests using the

moment conditions

$$\bar{m}_n(\theta, \hat{\lambda}) = \frac{1}{n} \sum_{t=1}^{n} \int D_\lambda \ln f_t(\hat{\lambda}) p_t(\theta) dy \tag{24.1}$$

- These moment conditions are not calculable, since $p_t(\theta)$ is not available, but they are simulable using

$$\widetilde{\bar{m}}_n(\theta, \hat{\lambda}) = \frac{1}{n} \sum_{t=1}^{n} \frac{1}{H} \sum_{h=1}^{H} D_\lambda \ln f(\tilde{y}_t^h | x_t, \hat{\lambda})$$

  where $\tilde{y}_t^h$ is a draw from $DGP(\theta)$, holding $x_t$ fixed. By the LLN and the fact that $\hat{\lambda}$ converges to $\lambda^0$,

$$\widetilde{m}_\infty(\theta^0, \lambda^0) = 0.$$

  This is not the case for other values of $\theta$, assuming that $\lambda^0$ is identified.

- The advantage of this procedure is that if $f(y_t | x_t, \lambda)$ closely approximates $p(y | x_t, \theta)$, then $\widetilde{m}_n(\theta, \hat{\lambda})$ will closely approximate the optimal moment con-

ditions which characterize maximum likelihood estimation, which is fully efficient.

- If one has prior information that a certain density approximates the data well, it would be a good choice for $f(\cdot)$.

- If one has no density in mind, there exist good ways of approximating unknown distributions parametrically: Philips' ERA's (*Econometrica*, 1983) and Gallant and Nychka's (*Econometrica, 1987)* SNP density estimator which we saw before. Since the SNP density is consistent, the efficiency of the indirect estimator is the same as the infeasible ML estimator.

## Optimal weighting matrix

I will present the theory for $H$ finite, and possibly small. This is done because it is sometimes impractical to estimate with $H$ very large. Gallant and Tauchen give the theory for the case of $H$ so large that it may be treated as infinite (the difference

being irrelevant given the numerical precision of a computer). The theory for the case of $H$ infinite follows directly from the results presented here.

The moment condition $\widetilde{m}(\theta, \hat{\lambda})$ depends on the pseudo-ML estimate $\hat{\lambda}$. We can apply Theorem 37 to conclude that

$$\sqrt{n}\left(\hat{\lambda} - \lambda^0\right) \xrightarrow{d} N\left[0, \mathcal{J}(\lambda^0)^{-1}\mathcal{I}(\lambda^0)\mathcal{J}(\lambda^0)^{-1}\right] \tag{24.2}$$

If the density $f(y_t|x_t, \hat{\lambda})$ were in fact the true density $p(y|x_t, \theta)$, then $\hat{\lambda}$ would be the maximum likelihood estimator, and $\mathcal{J}(\lambda^0)^{-1}\mathcal{I}(\lambda^0)$ would be an identity matrix, due to the information matrix equality. However, in the present case we assume that $f(y_t|x_t, \hat{\lambda})$ is only an approximation to $p(y|x_t, \theta)$, so there is no cancellation.

Recall that $\mathcal{J}(\lambda^0) \equiv p\lim\left(\frac{\partial^2}{\partial\lambda\partial\lambda'}s_n(\lambda^0)\right)$. Comparing the definition of $s_n(\lambda)$ with the definition of the moment condition in Equation 24.1, we see that

$$\mathcal{J}(\lambda^0) = D_{\lambda'}m(\theta^0, \lambda^0).$$

As in Theorem 37,

$$\mathcal{I}(\lambda^0) = \lim_{n\to\infty} \mathcal{E}\left[ n \left.\frac{\partial s_n(\lambda)}{\partial \lambda}\right|_{\lambda^0} \left.\frac{\partial s_n(\lambda)}{\partial \lambda'}\right|_{\lambda^0} \right].$$

In this case, this is simply the asymptotic variance covariance matrix of the moment conditions, $\Omega$. Now take a first order Taylor's series approximation to $\sqrt{n}\bar{m}_n(\theta^0, \hat{\lambda})$ about $\lambda^0$ :

$$\sqrt{n}\tilde{m}_n(\theta^0, \hat{\lambda}) = \sqrt{n}\tilde{m}_n(\theta^0, \lambda^0) + \sqrt{n}D_{\lambda'}\tilde{m}(\theta^0, \lambda^0)\left(\hat{\lambda} - \lambda^0\right) + o_p(1)$$

First consider $\sqrt{n}\tilde{m}_n(\theta^0, \lambda^0)$. It is straightforward but somewhat tedious to show that the asymptotic variance of this term is $\frac{1}{H}I_\infty(\lambda^0)$.

Next consider the second term $\sqrt{n}D_{\lambda'}\tilde{m}(\theta^0, \lambda^0)\left(\hat{\lambda} - \lambda^0\right)$. Note that $D_{\lambda'}\tilde{m}_n(\theta^0, \lambda^0) \overset{a.s.}{\to} \mathcal{J}(\lambda^0)$, so we have

$$\sqrt{n}D_{\lambda'}\tilde{m}(\theta^0, \lambda^0)\left(\hat{\lambda} - \lambda^0\right) = \sqrt{n}\mathcal{J}(\lambda^0)\left(\hat{\lambda} - \lambda^0\right), a.s.$$

But noting equation 24.2

$$\sqrt{n}\mathcal{J}(\lambda^0)\left(\hat{\lambda}-\lambda^0\right) \overset{a}{\sim} N\left[0,\mathcal{I}(\lambda^0)\right]$$

Now, combining the results for the first and second terms,

$$\sqrt{n}\tilde{m}_n(\theta^0,\hat{\lambda}) \overset{a}{\sim} N\left[0,\left(1+\frac{1}{H}\right)\mathcal{I}(\lambda^0)\right]$$

Suppose that $\widehat{\mathcal{I}(\lambda^0)}$ is a consistent estimator of the asymptotic variance-covariance matrix of the moment conditions. This may be complicated if the score generator is a poor approximator, since the individual score contributions may not have mean zero in this case (see the section on QML) . Even if this is the case, the individuals means can be calculated by simulation, so it is always possible to consistently estimate $\mathcal{I}(\lambda^0)$ when the model is simulable. On the other hand, if the score generator is taken to be correctly specified, the ordinary estimator of the information matrix is consistent. Combining this with the result on the efficient

GMM weighting matrix in Theorem 61, we see that defining $\hat{\theta}$ as

$$\hat{\theta} = \arg\min_{\Theta} \bar{m}_n(\theta, \hat{\lambda})' \left[ \left(1 + \frac{1}{H}\right) \widehat{\mathcal{I}(\lambda^0)} \right]^{-1} \bar{m}_n(\theta, \hat{\lambda})$$

is the GMM estimator with the efficient choice of weighting matrix.

- If one has used the Gallant-Nychka ML estimator as the auxiliary model, the appropriate weighting matrix is simply the information matrix of the auxiliary model, since the scores are uncorrelated. (e.g., it really is ML estimation asymptotically, since the score generator can approximate the unknown density arbitrarily well).

## Asymptotic distribution

Since we use the optimal weighting matrix, the asymptotic distribution is as in Equation 16.4, so we have (using the result in Equation 24.2):

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0, \left(D_\infty\left[\left(1+\frac{1}{H}\right)\mathcal{I}(\lambda^0)\right]^{-1} D_\infty'\right)^{-1}\right],$$

where

$$D_\infty = \lim_{n\to\infty} \mathcal{E}\left[D_\theta \bar{m}_n(\theta^0, \lambda^0)\right].$$

This can be consistently estimated using

$$\hat{D} = D_\theta \bar{m}_n(\hat{\theta}, \hat{\lambda})$$

## Diagnostic testing

The fact that

$$\sqrt{n}\bar{m}_n(\theta^0, \hat{\lambda}) \overset{a}{\sim} N\left[0, \left(1 + \frac{1}{H}\right)\mathcal{I}(\lambda^0)\right]$$

implies that

$$n\bar{m}_n(\hat{\theta}, \hat{\lambda})' \left[\left(1 + \frac{1}{H}\right)\mathcal{I}(\hat{\lambda})\right]^{-1} \bar{m}_n(\hat{\theta}, \hat{\lambda}) \overset{a}{\sim} \chi^2(q)$$

where $q$ is $\dim(\lambda) - \dim(\theta)$, since without $\dim(\theta)$ moment conditions the model is not identified, so testing is impossible. One test of the model is simply based on this statistic: if it exceeds the $\chi^2(q)$ critical point, something may be wrong (the small sample performance of this sort of test would be a topic worth investigating).

- Information about what is wrong can be gotten from the pseudo-t-statistics:

$$\left(\operatorname{diag}\left[\left(1 + \frac{1}{H}\right)\mathcal{I}(\hat{\lambda})\right]^{1/2}\right)^{-1} \sqrt{n}\bar{m}_n(\hat{\theta}, \hat{\lambda})$$

  can be used to test which moments are not well modeled. Since these mo-

ments are related to parameters of the score generator, which are usually related to certain features of the model, this information can be used to revise the model. These aren't actually distributed as $N(0,1)$, since $\sqrt{n}\bar{m}_n(\theta^0, \hat{\lambda})$ and $\sqrt{n}\bar{m}_n(\hat{\theta}, \hat{\lambda})$ have different distributions (that of $\sqrt{n}\bar{m}_n(\hat{\theta}, \hat{\lambda})$ is somewhat more complicated). It can be shown that the pseudo-t statistics are biased toward nonrejection. See Gourieroux *et. al.* or Gallant and Long, 1995, for more details.

## 24.2 Parallel programming for econometrics

The following borrows heavily from Creel (2005).

Parallel computing can offer an important reduction in the time to complete computations. This is well-known, but it bears emphasis since it is the main reason that parallel computing may be attractive to users. To illustrate, the Intel Pentium IV (Willamette) processor, running at 1.5GHz, was introduced in November of

2000. The Pentium IV (Northwood-HT) processor, running at 3.06GHz, was introduced in November of 2002. An approximate doubling of the performance of a commodity CPU took place in two years. Extrapolating this admittedly rough snapshot of the evolution of the performance of commodity processors, one would need to wait more than 6.6 years and then purchase a new computer to obtain a 10-fold improvement in computational performance. The examples in this chapter show that a 10-fold improvement in performance can be achieved immediately, using distributed parallel computing on available computers.

Recent (this is written in 2005) developments that may make parallel computing attractive to a broader spectrum of researchers who do computations. The first is the fact that setting up a cluster of computers for distributed parallel computing is not difficult. If you are using the http://pareto.uab.es/mcreel/ParallelKnoppix bootable CD that accompanies these notes, you are less than 10 minutes away from creating a cluster, supposing you have a second computer at hand and a crossover ethernet cable. See the http://pareto.uab.es/mcreel/ParallelKnoppix/ParallelKnoppixTutori

A second development is the existence of extensions to some of the high-level matrix programming (HLMP) languages[1] that allow the incorporation of parallelism into programs written in these languages. A third is the spread of dual and quad-core CPUs, so that an ordinary desktop or laptop computer can be made into a mini-cluster. Those cores won't work together on a single problem unless they are told how to.

Following are examples of parallel implementations of several mainstream problems in econometrics. A focus of the examples is on the possibility of hiding parallelization from end users of programs. If programs that run in parallel have an interface that is nearly identical to the interface of equivalent serial versions, end users will find it easy to take advantage of parallel computing's performance. We continue to use Octave, taking advantage of the http://atc.ugr.es/javier-bin/mpitb, by by Fernández Baldomero *et al.* (2004). There are also parallel packages for Ox, R, and Python which may be of interest to econometricians, but as of this

[1]By "high-level matrix programming language" I mean languages such as MATLAB (TM the Mathworks, Inc.), Ox (TM OxMetrics Technologies, Ltd.), and GNU Octave (`www.octave.org`), for example.

writing, the following examples are the most accessible introduction to parallel programming for econometricians.

## Example problems

This section introduces example problems from econometrics, and shows how they can be parallelized in a natural way.

### Monte Carlo

A Monte Carlo study involves repeating a random experiment many times under identical conditions. Several authors have noted that Monte Carlo studies are obvious candidates for parallelization (Doornik *et al.* 2002; Bruche, 2003) since blocks of replications can be done independently on different computers. To illustrate the parallelization of a Monte Carlo study, we use same trace test example as do Doornik, *et. al.* (2002). ./Examples/Parallel/montecarlo/tracetest.m is a function that calculates the trace test statistic for the lack of cointegration of

integrated time series. This function is illustrative of the format that we adopt for Monte Carlo simulation of a function: it receives a single argument of cell type, and it returns a row vector that holds the results of one random simulation. The single argument in this case is a cell array that holds the length of the series in its first position, and the number of series in the second position. It generates a random result though a process that is internal to the function, and it reports some output in a row vector (in this case the result is a scalar).

./Examples/Parallel/montecarlo/mc_example1.m is an Octave script that executes a Monte Carlo study of the trace test by repeatedly evaluating the `tracetest.m` function. The main thing to notice about this script is that lines 7 and 10 call the function `montecarlo.m.` When called with 3 arguments, as in line 7, `montecarlo.m` executes serially on the computer it is called from. In line 10, there is a fourth argument. When called with four arguments, the last argument is the number of slave hosts to use. We see that running the Monte Carlo study on one or more processors is transparent to the user - he or she must only indicate

the number of slave computers to be used.

## ML

For a sample $\{(y_t, x_t)\}_n$ of $n$ observations of a set of dependent and explanatory variables, the maximum likelihood estimator of the parameter $\theta$ can be defined as

$$\hat{\theta} = \arg\max s_n(\theta)$$

where

$$s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \ln f(y_t|x_t, \theta)$$

Here, $y_t$ may be a vector of random variables, and the model may be dynamic since $x_t$ may contain lags of $y_t$. As Swann (2002) points out, this can be broken into sums over blocks of observations, for example two blocks:

$$s_n(\theta) = \frac{1}{n} \left\{ \left( \sum_{t=1}^{n_1} \ln f(y_t|x_t, \theta) \right) + \left( \sum_{t=n_1+1}^{n} \ln f(y_t|x_t, \theta) \right) \right\}$$

Analogously, we can define up to $n$ blocks. Again following Swann, parallelization can be done by calculating each block on separate computers.

./Examples/Parallel/mle/mle_example1.m is an Octave script that calculates the maximum likelihood estimator of the parameter vector of a model that assumes that the dependent variable is distributed as a Poisson random variable, conditional on some explanatory variables. In lines 1-3 the data is read, the name of the density function is provided in the variable `model`, and the initial value of the parameter vector is set. In line 5, the function `mle_estimate` performs ordinary serial calculation of the ML estimator, while in line 7 the same function is called with 6 arguments. The fourth and fifth arguments are empty placeholders where options to `mle_estimate` may be set, while the sixth argument is the number of slave computers to use for parallel execution, 1 in this case. A person who runs the program sees no parallel programming code - the parallelization is transparent to the end user, beyond having to select the number of slave computers. When executed, this script prints out the estimates `theta_s` and `theta_p`, which are

identical.

It is worth noting that a different likelihood function may be used by making the `model` variable point to a different function. The likelihood function itself is an ordinary Octave function that is not parallelized. The `mle_estimate` function is a generic function that can call any likelihood function that has the appropriate input/output syntax for evaluation either serially or in parallel. Users need only learn how to write the likelihood function using the Octave language.

## GMM

For a sample as above, the GMM estimator of the parameter $\theta$ can be defined as

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\theta)$$

where

$$s_n(\theta) = \bar{m}_n(\theta)' W_n \bar{m}_n(\theta)$$

and

$$\bar{m}_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} m_t(y_t|x_t, \theta)$$

Since $\bar{m}_n(\theta)$ is an average, it can obviously be computed blockwise, using for example 2 blocks:

$$\bar{m}_n(\theta) = \frac{1}{n} \left\{ \left( \sum_{t=1}^{n_1} m_t(y_t|x_t, \theta) \right) + \left( \sum_{t=n_1+1}^{n} m_t(y_t|x_t, \theta) \right) \right\} \tag{24.3}$$

Likewise, we may define up to $n$ blocks, each of which could potentially be computed on a different machine.

./Examples/Parallel/gmm/gmm_example1.m is a script that illustrates how GMM estimation may be done serially or in parallel. When this is run, `theta_s` and `theta_p` are identical up to the tolerance for convergence of the minimization routine. The point to notice here is that an end user can perform the estimation in parallel in virtually the same way as it is done serially. Again, `gmm_estimate`, used in lines 8 and 10, is a generic function that will estimate any model specified by the `moments` variable - a different model can be estimated by changing the value of

the `moments` variable. The function that `moments` points to is an ordinary Octave function that uses no parallel programming, so users can write their models using the simple and intuitive HLMP syntax of Octave. Whether estimation is done in parallel or serially depends only the seventh argument to `gmm_estimate` - when it is missing or zero, estimation is by default done serially with one processor. When it is positive, it specifies the number of slave nodes to use.

## Kernel regression

The Nadaraya-Watson kernel regression estimator of a function $g(x)$ at a point $x$ is

$$\begin{aligned}\hat{g}(x) &= \frac{\Sigma_{t=1}^{n} y_t K\left[\left(x - x_t\right)/\gamma_n\right]}{\Sigma_{t=1}^{n} K\left[\left(x - x_t\right)/\gamma_n\right]} \\ &\equiv \sum_{t=1}^{n} w_t y_y\end{aligned}$$

We see that the weight depends upon every data point in the sample. To calculate the fit at every point in a sample of size $n$, on the order of $n^2k$ calculations must be done, where $k$ is the dimension of the vector of explanatory variables, $x$. Racine (2002) demonstrates that MPI parallelization can be used to speed up calculation of the kernel regression estimator by calculating the fits for portions of the sample on different computers. We follow this implementation here. ./Examples/Parallel/kernel/kernel_example1.m is a script for serial and parallel kernel regression.

The example programs show that parallelization may be mostly hidden from end users. Users can benefit from parallelization without having to write or understand parallel code. The speedups one can obtain are highly dependent upon the specific problem at hand, as well as the size of the cluster, the efficiency of the network, *etc.* Some examples of speedups are presented in Creel (2005). Figure 24.1 reproduces speedups for some econometric problems on a cluster of 12 desktop computers. The speedup for $k$ nodes is the time to finish the problem on a single

Figure 24.1: Speedups from parallelization

node divided by the time to finish the problem on $k$ nodes. Note that you can get 10X speedups, as claimed in the introduction. It's pretty obvious that much greater speedups could be obtained using a larger cluster, for the "embarrassingly parallel" problems.

# Duration data and the Weibull model

In some cases the dependent variable may be the time that passes between the occurrence of two events. For example, it may be the duration of a strike, or the time needed to find a job once one is unemployed. Such variables take on values on the positive real line, and are referred to as duration data.

A *spell* is the period of time between the occurrence of initial event and the concluding event. For example, the initial event could be the loss of a job, and the final event is the finding of a new job. The spell is the period of unemployment.

Let $t_0$ be the time the initial event occurs, and $t_1$ be the time the concluding event occurs. For simplicity, assume that time is measured in years. The random variable $D$ is the duration of the spell, $D = t_1 - t_0$. Define the density function of $D$, $f_D(t)$, with distribution function $F_D(t) = \Pr(D < t)$.

Several questions may be of interest. For example, one might wish to know the expected time one has to wait to find a job given that one has already waited

$s$ years. The probability that a spell lasts more than $s$ years is

$$\Pr(D > s) = 1 - \Pr(D \leq s) = 1 - F_D(s).$$

The density of $D$ conditional on the spell being longer than $s$ years is

$$f_D(t|D > s) = \frac{f_D(t)}{1 - F_D(s)}.$$

The expected additional time required for the spell to end given that is has already lasted $s$ years is the expectation of $D$ with respect to this density, minus $s$.

$$E = \mathcal{E}(D|D > s) - s = \left( \int_t^\infty z \frac{f_D(z)}{1 - F_D(s)} dz \right) - s$$

To estimate this function, one needs to specify the density $f_D(t)$ as a parametric density, then estimate by maximum likelihood. There are a number of possibilities including the exponential density, the lognormal, *etc.* A reasonably flexible model that is a generalization of the exponential density is the Weibull density

$$f_D(t|\theta) = e^{-(\lambda t)^\gamma} \lambda\gamma(\lambda t)^{\gamma-1}.$$

According to this model, $\mathcal{E}(D) = \lambda^{-\gamma}$. The log-likelihood is just the product of the log densities.

To illustrate application of this model, 402 observations on the lifespan of dwarf mongooses in Serengeti National Park (Tanzania) were used to fit a Weibull model. The "spell" in this case is the lifetime of an individual mongoose. The parameter estimates and standard errors are $\hat{\lambda} = 0.559\,(0.034)$ and $\hat{\gamma} = 0.867\,(0.033)$ and the log-likelihood value is -659.3. Figure 24.2 presents fitted life expectancy (expected additional years of life) as a function of age, with 95% confidence bands. The plot is accompanied by a nonparametric Kaplan-Meier estimate of life-expectancy. This nonparametric estimator simply averages all spell lengths greater than age, and then subtracts age. This is consistent by the LLN.

In the figure one can see that the model doesn't fit the data well, in that it predicts life expectancy quite differently than does the nonparametric model.

Figure 24.2: Life expectancy of mongooses, Weibull model

For ages 4-6, the nonparametric estimate is outside the confidence interval that results from the parametric model, which casts doubt upon the parametric model. Mongooses that are between 2-6 years old seem to have a lower life expectancy than is predicted by the Weibull model, whereas young mongooses that survive beyond infancy have a higher life expectancy, up to a bit beyond 2 years. Due to the dramatic change in the death rate as a function of $t$, one might specify $f_D(t)$ as a mixture of two Weibull densities,

$$f_D(t|\theta) = \delta \left( e^{-(\lambda_1 t)^{\gamma_1}} \lambda_1 \gamma_1 (\lambda_1 t)^{\gamma_1 - 1} \right) + (1-\delta) \left( e^{-(\lambda_2 t)^{\gamma_2}} \lambda_2 \gamma_2 (\lambda_2 t)^{\gamma_2 - 1} \right).$$

The parameters $\gamma_i$ and $\lambda_i$, $i = 1, 2$ are the parameters of the two Weibull densities, and $\delta$ is the parameter that mixes the two.

With the same data, $\theta$ can be estimated using the mixed model. The results are a log-likelihood = -623.17. Note that a standard likelihood ratio test cannot be used to chose between the two models, since under the null that $\delta = 1$ (single density), the two parameters $\lambda_2$ and $\gamma_2$ are not identified. It is possible to take

this into account, but this topic is out of the scope of this course. Nevertheless, the improvement in the likelihood function is considerable. The parameter estimates are

| Parameter | Estimate | St. Error |
|---|---|---|
| $\lambda_1$ | 0.233 | 0.016 |
| $\gamma_1$ | 1.722 | 0.166 |
| $\lambda_2$ | 1.731 | 0.101 |
| $\gamma_2$ | 1.522 | 0.096 |
| $\delta$ | 0.428 | 0.035 |

Note that the mixture parameter is highly significant. This model leads to the fit in Figure 24.3. Note that the parametric and nonparametric fits are quite close to one another, up to around 6 years. The disagreement after this point is not too important, since less than 5% of mongooses live more than 6 years, which implies that the Kaplan-Meier nonparametric estimate has a high variance (since it's an average of a small number of observations).

Figure 24.3: Life expectancy of mongooses, mixed Weibull model

Mixture models are often an effective way to model complex responses, though they can suffer from overparameterization. Alternatives will be discussed later.

For examples of MLE using the Poisson model applied to count data, see Section 12.4 in the chapter on Numerical Optimization. You should examine the scripts and run them to see how MLE is actually done, and how parameter standard errors are estimated.

## 24.3 Quasi-ML

Quasi-ML is the estimator one obtains when a misspecified probability model is used to calculate an "ML" estimator.

Given a sample of size $n$ of a random vector $\mathbf{y}$ and a vector of conditioning variables $\mathbf{x}$, suppose the joint density of $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \ldots & \mathbf{y}_n \end{pmatrix}$ conditional on $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{pmatrix}$ is a member of the parametric family $p_{\mathcal{Y}}(\mathbf{Y}|\mathbf{X}, \rho)$, $\rho \in \Xi$. The true

joint density is associated with the vector $\rho^0$ :

$$p_{\mathcal{Y}}(\mathbf{Y}|\mathbf{X}, \rho^0).$$

As long as the marginal density of $\mathbf{X}$ doesn't depend on $\rho^0$, this conditional density fully characterizes the random characteristics of samples: i.e., it fully describes the probabilistically important features of the d.g.p. The *likelihood function* is just this density evaluated at other values $\rho$

$$L(\mathbf{Y}|\mathbf{X}, \rho) = p_{\mathcal{Y}}(\mathbf{Y}|\mathbf{X}, \rho), \rho \in \Xi.$$

- Let $\mathbf{Y}_{t-1} = \begin{pmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_{t-1} \end{pmatrix}$, $\mathbf{Y}_0 = 0$, and let $\mathbf{X}_t = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_t \end{pmatrix}$ The likelihood function, taking into account possible dependence of observations,

can be written as

$$\begin{aligned} L(\mathbf{Y}|\mathbf{X},\rho) &= \prod_{t=1}^{n} p_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\rho) \\ &\equiv \prod_{t=1}^{n} p_t(\rho) \end{aligned}$$

- The average log-likelihood function is:

$$s_n(\rho) = \frac{1}{n}\ln L(\mathbf{Y}|\mathbf{X},\rho) = \frac{1}{n}\sum_{t=1}^{n}\ln p_t(\rho)$$

- Suppose that we do not have knowledge of the family of densities $p_t(\rho)$. Mistakenly, we may assume that the conditional density of $\mathbf{y}_t$ is a member of the family $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\theta)$, $\theta \in \Theta$, where there is no $\theta^0$ such that $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\theta^0) = p_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\mathbf{X}_t,\rho^0), \forall t$ (this is what we mean by "misspecified").

- This setup allows for heterogeneous time series data, with dynamic misspecification.

The QML estimator is the argument that maximizes the **misspecified** average log likelihood, which we refer to as the quasi-log likelihood function. This objective function is

$$\begin{aligned} s_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n} \ln f_t(\mathbf{y}_t|\mathbf{Y}_{t-1}, \mathbf{X}_t, \theta^0) \\ &\equiv \frac{1}{n}\sum_{t=1}^{n} \ln f_t(\theta) \end{aligned}$$

and the QML is

$$\hat{\theta}_n = \arg\max_{\Theta} s_n(\theta)$$

A SLLN for dependent sequences applies (we assume), so that

$$s_n(\theta) \stackrel{a.s.}{\to} \lim_{n\to\infty} \mathcal{E}\frac{1}{n}\sum_{t=1}^{n} \ln f_t(\theta) \equiv s_\infty(\theta)$$

We assume that this can be strengthened to uniform convergence, a.s., following the previous arguments. The "pseudo-true" value of $\theta$ is the value that maximizes

$\bar{s}(\theta)$:

$$\theta^0 = \arg\max_{\Theta} s_\infty(\theta)$$

Given assumptions so that theorem 35 is applicable, we obtain

$$\lim_{n\to\infty} \hat{\theta}_n = \theta^0, \text{a.s.}$$

- Applying the asymptotic normality theorem,

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right]$$

  where

$$\mathcal{J}_\infty(\theta^0) = \lim_{n\to\infty} \mathcal{E} D_\theta^2 s_n(\theta^0)$$

  and

$$\mathcal{I}_\infty(\theta^0) = \lim_{n\to\infty} Var\sqrt{n} D_\theta s_n(\theta^0).$$

- Note that asymptotic normality only requires that the additional assumptions regarding $\mathcal{J}$ and $\mathcal{I}$ hold in a neighborhood of $\theta^0$ for $\mathcal{J}$ and at $\theta^0$, for

$\mathcal{I}$, not throughout $\Theta$. In this sense, asymptotic normality is a local property.

## Consistent Estimation of Variance Components

Consistent estimation of $\mathcal{J}_\infty(\theta^0)$ is straightforward. Assumption (b) of Theorem 37 implies that

$$\mathcal{J}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\hat{\theta}_n) \overset{a.s.}{\to} \lim_{n\to\infty} \mathcal{E}\frac{1}{n}\sum_{t=1}^{n} D_\theta^2 \ln f_t(\theta^0) = \mathcal{J}_\infty(\theta^0).$$

That is, just calculate the Hessian using the estimate $\hat{\theta}_n$ in place of $\theta^0$.

Consistent estimation of $\mathcal{I}_\infty(\theta^0)$ is more difficult, and may be impossible.

- **Notation**: Let $g_t \equiv D_\theta f_t(\theta^0)$

We need to estimate

$$\begin{aligned}
\mathcal{I}_\infty(\theta^0) &= \lim_{n\to\infty} Var\sqrt{n}D_\theta s_n(\theta^0) \\
&= \lim_{n\to\infty} Var\sqrt{n}\frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f_t(\theta^0) \\
&= \lim_{n\to\infty} \frac{1}{n}Var\sum_{t=1}^{n} g_t \\
&= \lim_{n\to\infty} \frac{1}{n}\mathcal{E}\left\{\left(\sum_{t=1}^{n}(g_t-\mathcal{E}g_t)\right)\left(\sum_{t=1}^{n}(g_t-\mathcal{E}g_t)\right)'\right\}
\end{aligned}$$

This is going to contain a term

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}(\mathcal{E}g_t)(\mathcal{E}g_t)'$$

which will not tend to zero, in general. This term is not consistently estimable in general, since it requires calculating an expectation using the true density under the d.g.p., which is unknown.

- There are important cases where $\mathcal{I}_\infty(\theta^0)$ *is* consistently estimable. For ex-

ample, suppose that the data come from a random sample (*i.e.,* they are iid). This would be the case with cross sectional data, for example. (Note: under i.i.d. sampling, the joint distribution of $(y_t, x_t)$ is identical. This does not imply that the conditional density $f(y_t|x_t)$ is identical).

- With random sampling, the limiting objective function is simply

$$s_\infty(\theta^0) = \mathcal{E}_X \mathcal{E}_0 \ln f(y|x, \theta^0)$$

where $\mathcal{E}_0$ means expectation of $y|x$ and $\mathcal{E}_X$ means expectation respect to the marginal density of $x$.

- By the requirement that the limiting objective function be maximized at $\theta^0$ we have

$$D_\theta \mathcal{E}_X \mathcal{E}_0 \ln f(y|x, \theta^0) = D_\theta s_\infty(\theta^0) = 0$$

- The dominated convergence theorem allows switching the order of expecta-

tion and differentiation, so

$$D_\theta \mathcal{E}_X \mathcal{E}_0 \ln f(y|x,\theta^0) = \mathcal{E}_X \mathcal{E}_0 D_\theta \ln f(y|x,\theta^0) = 0$$

The CLT implies that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} D_\theta \ln f(y|x,\theta^0) \xrightarrow{d} N(0, \mathcal{I}_\infty(\theta^0)).$$

That is, it's not necessary to subtract the individual means, since they are zero. Given this, and due to independent observations, a consistent estimator is

$$\widehat{\mathcal{I}} = \frac{1}{n} \sum_{t=1}^{n} D_\theta \ln f_t(\hat{\theta}) D_{\theta'} \ln f_t(\hat{\theta})$$

This is an important case where consistent estimation of the covariance matrix is possible. Other cases exist, even for dynamically misspecified time series models.

## 24.4 Nonlinear simultaneous equations

Taken out of GMM chapter. GMM provides a convenient way to estimate nonlinear systems of simultaneous equations. We have a system of equations of the form

$$\begin{aligned} y_{1t} &= f_1(\mathbf{z}_t, \theta_1^0) + \varepsilon_{1t} \\ y_{2t} &= f_2(\mathbf{z}_t, \theta_2^0) + \varepsilon_{2t} \\ &\vdots \\ y_{Gt} &= f_G(\mathbf{z}_t, \theta_G^0) + \varepsilon_{Gt}, \end{aligned}$$

or in compact notation

$$y_t = f(\mathbf{z}_t, \theta^0) + \varepsilon_t,$$

where $f(\cdot)$ is a $G$ -vector valued function, and $\theta^0 = (\theta_1^{0\prime}, \theta_2^{0\prime}, \cdots, \theta_G^{0\prime})'$. We assume that $\mathbf{z}_t$ contains the current period endogenous variables, so we have a simultaneity problem.

We need to find an $A_i \times 1$ vector of instruments $\mathbf{x}_{it}$, for each equation, that are uncorrelated with $\varepsilon_{it}$. Typical instruments would be low order monomials in the exogenous variables in $\mathbf{z}_t$, with their lagged values. Then we can define the $\left(\Sigma_{i=1}^{G} A_i\right) \times 1$ orthogonality conditions

$$m_t(\theta) = \begin{bmatrix} (y_{1t} - f_1(\mathbf{z}_t, \theta_1))\, \mathbf{x}_{1t} \\ (y_{2t} - f_2(\mathbf{z}_t, \theta_2))\, \mathbf{x}_{2t} \\ \vdots \\ (y_{Gt} - f_G(\mathbf{z}_t, \theta_G))\, \mathbf{x}_{Gt} \end{bmatrix}.$$

- once we have gotten this far, we can just proceed with GMM estimation, one-step, two-step, CUE, or whatever.

- A note on identification: selection of instruments that ensure identification is a non-trivial problem. Identification in nonlinear models is not as easy to check as it is with linear models, where counting zero restrictions works.

- A note on efficiency: the selected set of instruments has important effects on

the efficiency of estimation. There are some papers that study this problem, but the results are fairly complicated and difficult to implement. I think it's safe to say that the great majority of applied work does not attempt to use optimal instruments.

## 24.5 Example: The MEPS data

Taken out of the GMM chapter, distracting, and not a great example. The MEPS data on health care usage discussed in section 12.4 estimated a Poisson model by "maximum likelihood" (probably misspecified). Perhaps the same latent factors (e.g., chronic illness) that induce one to make doctor visits also influence the decision of whether or not to purchase insurance. If this is the case, the PRIV variable could well be endogenous, in which case, the Poisson "ML" estimator would be inconsistent, even if the conditional mean were correctly specified. Suppose that

$$y = \exp(X\beta + Z\gamma)v$$

where $E(v|X) = 1$ but $v$ may be related to $Z$, so $Z$ is endogenous. Then $E(y/\exp(X\beta + Z\gamma) - 1|X) = 0$. This expression can be used to define moment conditions. The Octave script ./Examples/GMM/MEPS/meps.m estimates the parameters of the model presented in equation 12.1, using Poisson "ML" (better thought of as quasi-ML), and IV estimation[2]. Both estimation methods are implemented using a GMM form. Running that script gives the output

```
OBDV



*****************************************************
IV

GMM Estimation Results
BFGS convergence: Normal convergence
```

[2]The validity of the instruments used may be debatable, but real data sets often don't contain ideal instruments.

Objective function value: 0.004273

Observations: 4564

No moment covariance supplied, assuming efficient weight matrix

| | Value | df | p-value |
|---|---|---|---|
| X^2 test | 19.502 | 3.000 | 0.000 |

| | estimate | st. err | t-stat | p-value |
|---|---|---|---|---|
| constant | -0.441 | 0.213 | -2.072 | 0.038 |
| pub. ins. | -0.127 | 0.149 | -0.851 | 0.395 |
| priv. ins. | -1.429 | 0.254 | -5.624 | 0.000 |
| sex | 0.537 | 0.053 | 10.133 | 0.000 |
| age | 0.031 | 0.002 | 13.431 | 0.000 |
| edu | 0.072 | 0.011 | 6.535 | 0.000 |

```
inc                0.000       0.000       4.500       0.000
*****************************************************


*****************************************************
Poisson QML

GMM Estimation Results
BFGS convergence: Normal convergence

Objective function value: 0.000000
Observations: 4564
No moment covariance supplied, assuming efficient weight matrix

Exactly identified, no spec. test
```

```
                estimate      st. err       t-stat      p-value
constant          -0.791        0.149       -5.289        0.000
pub. ins.          0.848        0.076       11.092        0.000
priv. ins.         0.294        0.071        4.136        0.000
sex                0.487        0.055        8.796        0.000
age                0.024        0.002       11.469        0.000
edu                0.029        0.010        3.060        0.002
inc               -0.000        0.000       -0.978        0.328
*****************************************************
```

Note how the Poisson QML results, estimated here using a GMM routine, are the same as were obtained using the ML estimation routine (see 15.1). This is an example of how (Q)ML may be represented as a GMM estimator. Also note that the IV and QML results are considerably different. Treating PRIV as potentially

endogenous causes the sign of its coefficient to change. Perhaps it is logical that people who own private insurance make fewer visits, if they have to make a co-payment. Note that income becomes positive and significant when PRIV is treated as endogenous.

Perhaps the difference in the results depending upon whether or not PRIV is treated as endogenous can suggest a method for testing exogeneity....

## Invertibility of AR process

To begin with, define the lag operator $L$

$$Ly_t = y_{t-1}$$

The lag operator is defined to behave just as an algebraic quantity, e.g.,

$$\begin{aligned} L^2 y_t &= L(L y_t) \\ &= L y_{t-1} \\ &= y_{t-2} \end{aligned}$$

or

$$\begin{aligned} (1-L)(1+L)y_t &= 1 - L y_t + L y_t - L^2 y_t \\ &= 1 - y_{t-2} \end{aligned}$$

A mean-zero AR(p) process can be written as

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} = \varepsilon_t$$

or

$$y_t(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) = \varepsilon_t$$

Factor this polynomial as

$$1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p = (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L)$$

For the moment, just assume that the $\lambda_i$ are coefficients to be determined. Since $L$ is defined to operate as an algebraic quantity, determination of the $\lambda_i$ is the same as determination of the $\lambda_i$ such that the following two expressions are the same for all $z$ :

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = (1 - \lambda_1 z)(1 - \lambda_2 z) \cdots (1 - \lambda_p z)$$

Multiply both sides by $z^{-p}$

$$z^{-p} - \phi_1 z^{1-p} - \phi_2 z^{2-p} - \cdots \phi_{p-1} z^{-1} - \phi_p = (z^{-1} - \lambda_1)(z^{-1} - \lambda_2) \cdots (z^{-1} - \lambda_p)$$

and now define $\lambda = z^{-1}$ so we get

$$\lambda^p - \phi_1 \lambda^{p-1} - \phi_2 \lambda^{p-2} - \cdots - \phi_{p-1} \lambda - \phi_p = (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_p)$$

The LHS is precisely the determinantal polynomial that gives the eigenvalues of $F$. Therefore, the $\lambda_i$ that are the coefficients of the factorization are simply the eigenvalues of the matrix $F$.

Now consider a different stationary process

$$(1 - \phi L)y_t = \varepsilon_t$$

- Stationarity, as above, implies that $|\phi| < 1$.

Multiply both sides by $1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j$ to get

$$\left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right)(1 - \phi L)y_t = \left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j\right)\varepsilon_t$$

or, multiplying the polynomials on the LHS, we get

$$\left(1 + \phi L + \phi^2 L^2 + ... + \phi^j L^j - \phi L - \phi^2 L^2 - ... - \phi^j L^j - \phi^{j+1} L^{j+1}\right) y_t = \left(1 + \phi L + \phi^2 L^2 -\right.$$

and with cancellations we have

$$\left(1-\phi^{j+1}L^{j+1}\right)y_t = \left(1+\phi L+\phi^2L^2+...+\phi^jL^j\right)\varepsilon_t$$

so

$$y_t = \phi^{j+1}L^{j+1}y_t + \left(1+\phi L+\phi^2L^2+...+\phi^jL^j\right)\varepsilon_t$$

Now as $j \to \infty$, $\phi^{j+1}L^{j+1}y_t \to 0$, since $|\phi| < 1$, so

$$y_t \cong \left(1+\phi L+\phi^2L^2+...+\phi^jL^j\right)\varepsilon_t$$

and the approximation becomes better and better as $j$ increases. However, we started with

$$(1-\phi L)y_t = \varepsilon_t$$

Substituting this into the above equation we have

$$y_t \cong \left(1+\phi L+\phi^2L^2+...+\phi^jL^j\right)(1-\phi L)y_t$$

so

$$\left(1+\phi L+\phi^2 L^2+...+\phi^j L^j\right)(1-\phi L)\cong 1$$

and the approximation becomes arbitrarily good as $j$ increases arbitrarily. Therefore, for $|\phi|<1$, define

$$(1-\phi L)^{-1}=\sum_{j=0}^{\infty}\phi^j L^j$$

Recall that our mean zero AR(p) process

$$y_t(1-\phi_1 L-\phi_2 L^2-\cdots-\phi_p L^p)=\varepsilon_t$$

can be written using the factorization

$$y_t(1-\lambda_1 L)(1-\lambda_2 L)\cdots(1-\lambda_p L)=\varepsilon_t$$

where the $\lambda$ are the eigenvalues of $F$, and given stationarity, all the $|\lambda_i|<1$.

Therefore, we can invert each first order polynomial on the LHS to get

$$y_t = \left(\sum_{j=0}^{\infty} \lambda_1^j L^j\right)\left(\sum_{j=0}^{\infty} \lambda_2^j L^j\right)\cdots\left(\sum_{j=0}^{\infty} \lambda_p^j L^j\right)\varepsilon_t$$

The RHS is a product of infinite-order polynomials in $L$, which can be represented as

$$y_t = (1 + \psi_1 L + \psi_2 L^2 + \cdots)\varepsilon_t$$

where the $\psi_i$ are real-valued and absolutely summable.

- The $\psi_i$ are formed of products of powers of the $\lambda_i$, which are in turn functions of the $\phi_i$.

- The $\psi_i$ are real-valued because any complex-valued $\lambda_i$ always occur in conjugate pairs. This means that if $a+bi$ is an eigenvalue of $F$, then so is $a-bi$.

In multiplication

$$\begin{aligned}(a+bi)(a-bi) &= a^2 - abi + abi - b^2 i^2 \\ &= a^2 + b^2\end{aligned}$$

which is real-valued.

- This shows that an AR(p) process is representable as an infinite-order MA(q) process.

- Recall before that by recursive substitution, an AR(p) process can be written as

$$Y_{t+j} = C + FC + \cdots + F^j C + F^{j+1} Y_{t-1} + F^j E_t + F^{j-1} E_{t+1} + \cdots + F E_{t+j-1} + E_{t+j}$$

If the process is mean zero, then everything with a $C$ drops out. Take this

and lag it by $j$ periods to get

$$Y_t = F^{j+1}Y_{t-j-1} + F^j E_{t-j} + F^{j-1}E_{t-j+1} + \cdots + FE_{t-1} + E_t$$

As $j \to \infty$, the lagged $Y$ on the RHS drops out. The $E_{t-s}$ are vectors of zeros except for their first element, so we see that the first equation here, in the limit, is just

$$y_t = \sum_{j=0}^{\infty} \left(F^j\right)_{1,1} \varepsilon_{t-j}$$

which makes explicit the relationship between the $\psi_i$ and the $\phi_i$ (and the $\lambda_i$ as well, recalling the previous factorization of $F^j$).

## Invertibility of MA(q) process

An MA(q) can be written as

$$y_t - \mu = (1 + \theta_1 L + ... + \theta_q L^q)\varepsilon_t$$

As before, the polynomial on the RHS can be factored as

$$(1 + \theta_1 L + ... + \theta_q L^q) = (1 - \eta_1 L)(1 - \eta_2 L)...(1 - \eta_q L)$$

and each of the $(1 - \eta_i L)$ can be inverted as long as each of the $|\eta_i| < 1$. If this is the case, then we can write

$$(1 + \theta_1 L + ... + \theta_q L^q)^{-1}(y_t - \mu) = \varepsilon_t$$

where

$$(1 + \theta_1 L + ... + \theta_q L^q)^{-1}$$

will be an infinite-order polynomial in $L$, so we get

$$\sum_{j=0}^{\infty} -\delta_j L^j (y_{t-j} - \mu) = \varepsilon_t$$

with $\delta_0 = -1$, or

$$(y_t - \mu) - \delta_1(y_{t-1} - \mu) - \delta_2(y_{t-2} - \mu) + ... = \varepsilon_t$$

or

$$y_t = c + \delta_1 y_{t-1} + \delta_2 y_{t-2} + ... + \varepsilon_t$$

where

$$c = \mu + \delta_1\mu + \delta_2\mu + ...$$

So we see that an MA(q) has an infinite AR representation, as long as the $|\eta_i| < 1$, $i = 1, 2, ..., q$.

- It turns out that one can always manipulate the parameters of an MA(q) process to find an invertible representation. For example, the two MA(1) processes

$$y_t - \mu = (1 - \theta L)\varepsilon_t$$

and

$$y_t^* - \mu = (1 - \theta^{-1}L)\varepsilon_t^*$$

have exactly the same moments if

$$\sigma_{\varepsilon^*}^2 = \sigma_\varepsilon^2 \theta^2$$

For example, we've seen that

$$\gamma_0 = \sigma^2(1 + \theta^2).$$

Given the above relationships amongst the parameters,

$$\gamma_0^* = \sigma_\varepsilon^2 \theta^2 (1 + \theta^{-2}) = \sigma^2(1 + \theta^2)$$

so the variances are the same. It turns out that *all* the autocovariances will be the same, as is easily checked. This means that the two MA processes are *observationally equivalent*. As before, it's impossible to distinguish

between observationally equivalent processes on the basis of data.

- For a given MA(q) process, it's always possible to manipulate the parameters to find an invertible representation (which is unique).

- It's important to find an invertible representation, since it's the only representation that allows one to represent $\varepsilon_t$ as a function of past $y's$. The other representations express $\epsilon_t$ as a function of future $y's$

- Why is invertibility important? The most important reason is that it provides a justification for the use of parsimonious models. Since an AR(1) process has an MA($\infty$) representation, one can reverse the argument and note that at least some MA($\infty$) processes have an AR(1) representation. Likewise, some AR($\infty$) processes have an MA(1) representation. At the time of estimation, it's a lot easier to estimate the single AR(1) or MA(1) coefficient rather than the infinite number of coefficients associated with the MA($\infty$) or AR($\infty$) representation.

- This is the reason that ARMA models are popular. Combining low-order AR and MA models can usually offer a satisfactory representation of univariate time series data using a reasonable number of parameters.

- Stationarity and invertibility of ARMA models is similar to what we've seen - we won't go into the details. Likewise, calculating moments is similar.

**Exercise 113.** Calculate the autocovariances of an ARMA(1,1) model:$(1+\phi L)y_t = c + (1+\theta L)\epsilon_t$

## Optimal instruments for GMM

PLEASE IGNORE THE REST OF THIS SECTION: there is a flaw in the argument that needs correction. In particular, it may be the case that $E(Z_t\epsilon_t) \neq 0$ if instruments are chosen in the way suggested here.

An interesting question that arises is how one should choose the instrumental variables $Z(w_t)$ to achieve maximum efficiency.

Note that with this choice of moment conditions, we have that $D_n \equiv \frac{\partial}{\partial\theta} m'(\theta)$ (a $K \times g$ matrix) is

$$\begin{aligned} D_n(\theta) &= \frac{\partial}{\partial\theta}\frac{1}{n}\left(Z_n' h_n(\theta)\right)' \\ &= \frac{1}{n}\left(\frac{\partial}{\partial\theta} h_n'(\theta)\right) Z_n \end{aligned}$$

which we can define to be

$$D_n(\theta) = \frac{1}{n} H_n Z_n.$$

where $H_n$ is a $K \times n$ matrix that has the derivatives of the individual moment

conditions as its columns. Likewise, define the var-cov. of the moment conditions

$$\begin{aligned}
\Omega_n &= \mathcal{E}\left[n\bar{m}_n(\theta^0)\bar{m}_n(\theta^0)'\right] \\
&= \mathcal{E}\left[\frac{1}{n}Z_n' h_n(\theta^0)h_n(\theta^0)'Z_n\right] \\
&= Z_n'\mathcal{E}\left(\frac{1}{n}h_n(\theta^0)h_n(\theta^0)'\right)Z_n \\
&\equiv Z_n'\frac{\Phi_n}{n}Z_n
\end{aligned}$$

where we have defined $\Phi_n = V\left(h_n(\theta^0)\right)$. Note that the dimension of this matrix is growing with the sample size, so it is not consistently estimable without additional assumptions.

The asymptotic normality theorem above says that the GMM estimator using the optimal weighting matrix is distributed as

$$\sqrt{n}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} N(0, V_\infty)$$

where

$$V_\infty = \lim_{n\to\infty} \left( \left( \frac{H_n Z_n}{n} \right) \left( \frac{Z_n' \Phi_n Z_n}{n} \right)^{-1} \left( \frac{Z_n' H_n'}{n} \right) \right)^{-1} . \tag{24.4}$$

Using an argument similar to that used to prove that $\Omega_\infty^{-1}$ is the efficient weighting matrix, we can show that putting

$$Z_n = \Phi_n^{-1} H_n'$$

causes the above var-cov matrix to simplify to

$$V_\infty = \lim_{n\to\infty} \left( \frac{H_n \Phi_n^{-1} H_n'}{n} \right)^{-1} . \tag{24.5}$$

and furthermore, this matrix is smaller that the limiting var-cov for any other choice of instrumental variables. (To prove this, examine the difference of the inverses of the var-cov matrices with the optimal intruments and with non-optimal instruments. As above, you can show that the difference is positive semi-definite).

- Note that both $H_n$, which we should write more properly as $H_n(\theta^0)$, since it depends on $\theta^0$, and $\Phi$ must be consistently estimated to apply this.

- Usually, estimation of $H_n$ is straightforward - one just uses

$$\widehat{H} = \frac{\partial}{\partial \theta} h_n' \left(\tilde{\theta}\right),$$

  where $\tilde{\theta}$ is some initial consistent estimator based on non-optimal instruments.

- Estimation of $\Phi_n$ may not be possible. It is an $n \times n$ matrix, so it has more unique elements than $n$, the sample size, so without restrictions on the parameters it can't be estimated consistently. Basically, you need to provide a parametric specification of the covariances of the $h_t(\theta)$ in order to be able to use optimal instruments. A solution is to approximate this matrix parametrically to define the instruments. Note that the simplified var-cov matrix in equation 24.5 will not apply if approximately optimal instruments

are used - it will be necessary to use an estimator based upon equation 24.4, where the term $n^{-1}Z_n'\Phi_n Z_n$ must be estimated consistently apart, for example by the Newey-West procedure.

## 24.6 Hurdle models

Returning to the Poisson model, lets look at actual and fitted count probabilities. Actual relative frequencies are $f(y=j) = \Sigma_i 1(y_i = j)/n$ and fitted frequencies are $\hat{f}(y=j) = \Sigma_{i=1}^n f_Y(j|x_i, \hat{\theta})/n$ We see that for the OBDV measure, there are many more actual zeros than predicted. For ERV, there are somewhat more actual zeros than fitted, but the difference is not too important.

Why might OBDV not fit the zeros well? What if people made the decision to contact the doctor for a first visit, they are sick, then the *doctor* decides on whether or not follow-up visits are needed. This is a principal/agent type situation, where the total number of visits depends upon the decision of both the patient

Table 24.1: Actual and Poisson fitted frequencies

| Count | OBDV | | ERV | |
|---|---|---|---|---|
| Count | Actual | Fitted | Actual | Fitted |
| 0 | 0.32 | 0.06 | 0.86 | 0.83 |
| 1 | 0.18 | 0.15 | 0.10 | 0.14 |
| 2 | 0.11 | 0.19 | 0.02 | 0.02 |
| 3 | 0.10 | 0.18 | 0.004 | 0.002 |
| 4 | 0.052 | 0.15 | 0.002 | 0.0002 |
| 5 | 0.032 | 0.10 | 0 | 2.4e-5 |

and the doctor. Since different parameters may govern the two decision-makers choices, we might expect that different parameters govern the probability of zeros versus the other counts. Let $\lambda_p$ be the parameters of the patient's demand for visits, and let $\lambda_d$ be the paramter of the doctor's "demand" for visits. The patient will initiate visits according to a discrete choice model, for example, a logit model:

$$\begin{aligned}
\Pr(Y = 0) &= f_Y(0, \lambda_p) = 1 - 1/\left[1 + \exp(-\lambda_p)\right] \\
\Pr(Y > 0) &= 1/\left[1 + \exp(-\lambda_p)\right],
\end{aligned}$$

The above probabilities are used to estimate the binary 0/1 hurdle process. Then, for the observations where visits are positive, a truncated Poisson density is esti-

mated. This density is

$$
\begin{aligned}
f_Y(y, \lambda_d | y > 0) &= \frac{f_Y(y, \lambda_d)}{\Pr(y > 0)} \\
&= \frac{f_Y(y, \lambda_d)}{1 - \exp(-\lambda_d)}
\end{aligned}
$$

since according to the Poisson model with the doctor's paramaters,

$$
\Pr(y = 0) = \frac{\exp(-\lambda_d)\lambda_d^0}{0!}.
$$

Since the hurdle and truncated components of the overall density for $Y$ share no parameters, they may be estimated separately, which is computationally more efficient than estimating the overall model. (Recall that the BFGS algorithm, for example, will have to invert the approximated Hessian. The computational overhead is of order $K^2$ where $K$ is the number of parameters to be estimated) .

The expectation of $Y$ is

$$
\begin{aligned}
E(Y|x) &= \Pr(Y > 0|x) E(Y|Y > 0, x) \\
&= \left(\frac{1}{1 + \exp(-\lambda_p)}\right)\left(\frac{\lambda_d}{1 - \exp(-\lambda_d)}\right)
\end{aligned}
$$

Here are hurdle Poisson estimation results for OBDV, obtained from ./Examples/MEPS-II/estimate_hpoisson.ox

```
*****************************************************************
MEPS data, OBDV
logit results
Strong convergence
Observations = 500
Function value      -0.58939
t-Stats
                    params       t(OPG)     t(Sand.)       t(Hess)
constant           -1.5502      -2.5709      -2.5269      -2.5560
pub_ins             1.0519       3.0520       3.0027       3.0384
priv_ins           0.45867       1.7289       1.6924       1.7166
sex                0.63570       3.0873       3.1677       3.1366
age               0.018614       2.1547       2.1969       2.1807
```

```
educ                0.039606        1.0467       0.98710        1.0222
inc                 0.077446        1.7655        2.1672        1.9601
Information Criteria
Consistent Akaike
       639.89
Schwartz
       632.89
Hannan-Quinn
       614.96
Akaike
       603.39
*****************************************************************************
```

The results for the truncated part:

```
*****************************************************************
MEPS data, OBDV
tpoisson results
Strong convergence
Observations = 500
Function value       -2.7042
t-Stats
```

| | params | t(OPG) | t(Sand.) | t(Hess) |
|---|---|---|---|---|
| constant | 0.54254 | 7.4291 | 1.1747 | 3.2323 |
| pub_ins | 0.31001 | 6.5708 | 1.7573 | 3.7183 |
| priv_ins | 0.014382 | 0.29433 | 0.10438 | 0.18112 |
| sex | 0.19075 | 10.293 | 1.1890 | 3.6942 |
| age | 0.016683 | 16.148 | 3.5262 | 7.9814 |
| educ | 0.016286 | 4.2144 | 0.56547 | 1.6353 |

```
inc              -0.0079016      -2.3186     -0.35309     -0.96078
Information Criteria
Consistent Akaike
       2754.7
Schwartz
       2747.7
Hannan-Quinn
       2729.8
Akaike
       2718.2
*****************************************************************
```

Fitted and actual probabilites (NB-II fits are provided as well) are:

Table 24.2: Actual and Hurdle Poisson fitted frequencies

| Count | OBDV | | | ERV | | |
|---|---|---|---|---|---|---|
| Count | Actual | Fitted HP | Fitted NB-II | Actual | Fitted HP | Fitted NB-II |
| 0 | 0.32 | 0.32 | 0.34 | 0.86 | 0.86 | 0.86 |
| 1 | 0.18 | 0.035 | 0.16 | 0.10 | 0.10 | 0.10 |
| 2 | 0.11 | 0.071 | 0.11 | 0.02 | 0.02 | 0.02 |
| 3 | 0.10 | 0.10 | 0.08 | 0.004 | 0.006 | 0.006 |
| 4 | 0.052 | 0.11 | 0.06 | 0.002 | 0.002 | 0.002 |
| 5 | 0.032 | 0.10 | 0.05 | 0 | 0.0005 | 0.001 |

For the Hurdle Poisson models, the ERV fit is very accurate. The OBDV fit is not so good. Zeros are exact, but 1's and 2's are underestimated, and higher counts are overestimated. For the NB-II fits, performance is at least as good as the hurdle Poisson model, and one should recall that many fewer parameters are used. Hurdle version of the negative binomial model are also widely used.

## 24.7 Finite mixture models

The following are results for a mixture of 2 negative binomial (NB-I) models, for the OBDV data, which you can replicate using ./Examples/MEPS-II/estimate_mixnegbin.ox

**********

MEPS data, OBDV

mixnegbin results

Strong convergence

Observations = 500

Function value -2.2312

t-Stats

| | params | t(OPG) | t(Sand.) | t(Hess) |
|---|---|---|---|---|
| constant | 0.64852 | 1.3851 | 1.3226 | 1.4358 |
| pub_ins | -0.062139 | -0.23188 | -0.13802 | -0.18729 |
| priv_ins | 0.093396 | 0.46948 | 0.33046 | 0.40854 |
| sex | 0.39785 | 2.6121 | 2.2148 | 2.4882 |
| age | 0.015969 | 2.5173 | 2.5475 | 2.7151 |
| educ | -0.049175 | -1.8013 | -1.7061 | -1.8036 |
| inc | 0.015880 | 0.58386 | 0.76782 | 0.73281 |

| | | | | |
|---|---|---|---|---|
| ln_alpha | 0.69961 | 2.3456 | 2.0396 | 2.4029 |
| constant | -3.6130 | -1.6126 | -1.7365 | -1.8411 |
| pub_ins | 2.3456 | 1.7527 | 3.7677 | 2.6519 |
| priv_ins | 0.77431 | 0.73854 | 1.1366 | 0.97338 |
| sex | 0.34886 | 0.80035 | 0.74016 | 0.81892 |
| age | 0.021425 | 1.1354 | 1.3032 | 1.3387 |
| educ | 0.22461 | 2.0922 | 1.7826 | 2.1470 |
| inc | 0.019227 | 0.20453 | 0.40854 | 0.36313 |
| ln_alpha | 2.8419 | 6.2497 | 6.8702 | 7.6182 |
| logit_inv_mix | 0.85186 | 1.7096 | 1.4827 | 1.7883 |

Information Criteria

Consistent Akaike

2353.8

Schwartz

2336.8

```
Hannan-Quinn
       2293.3
Akaike
       2265.2
*****************************************************************
Delta method for mix parameter st.  err.
          mix       se_mix
      0.70096      0.12043
```

- The 95% confidence interval for the mix parameter is perilously close to 1, which suggests that there may really be only one component density, rather than a mixture. Again, this is *not* the way to test this - it is merely suggestive.

- Education is interesting. For the subpopulation that is "healthy", i.e., that makes relatively few visits, education seems to have a positive effect on visits.

For the "unhealthy" group, education has a negative effect on visits. The other results are more mixed. A larger sample could help clarify things.

The following are results for a 2 component constrained mixture negative binomial model where all the slope parameters in $\lambda_j = e^{\mathbf{x}\beta_j}$ are the same across the two components. The constants and the overdispersion parameters $\alpha_j$ are allowed to differ for the two components.

```
*****************************************************************
MEPS data, OBDV
cmixnegbin results
Strong convergence
Observations = 500
Function value      -2.2441
t-Stats
```

| | params | t(OPG) | t(Sand.) | t(Hess) |
|---|---|---|---|---|
| constant | -0.34153 | -0.94203 | -0.91456 | -0.97943 |
| pub_ins | 0.45320 | 2.6206 | 2.5088 | 2.7067 |
| priv_ins | 0.20663 | 1.4258 | 1.3105 | 1.3895 |
| sex | 0.37714 | 3.1948 | 3.4929 | 3.5319 |
| age | 0.015822 | 3.1212 | 3.7806 | 3.7042 |
| educ | 0.011784 | 0.65887 | 0.50362 | 0.58331 |
| inc | 0.014088 | 0.69088 | 0.96831 | 0.83408 |

| | | | | |
|---|---|---|---|---|
| ln_alpha | 1.1798 | 4.6140 | 7.2462 | 6.4293 |
| const_2 | 1.2621 | 0.47525 | 2.5219 | 1.5060 |
| lnalpha_2 | 2.7769 | 1.5539 | 6.4918 | 4.2243 |
| logit_inv_mix | 2.4888 | 0.60073 | 3.7224 | 1.9693 |

Information Criteria

Consistent Akaike

2323.5

Schwartz

2312.5

Hannan-Quinn

2284.3

Akaike

2266.1

**********

```
Delta method for mix parameter st.  err.
          mix       se_mix
      0.92335     0.047318
```

- Now the mixture parameter is even closer to 1.
- The slope parameter estimates are pretty close to what we got with the NB-I model.

## 24.8 Nonlinear least squares (NLS)

**Readings**: Davidson and MacKinnon, Ch. 2* and 5*; Gallant, Ch. 1

### Introduction and definition

Nonlinear least squares (NLS) is a means of estimating the parameter of the model

$$y_t = f(\mathbf{x}_t, \theta^0) + \varepsilon_t.$$

- In general, $\varepsilon_t$ will be heteroscedastic and autocorrelated, and possibly non-normally distributed. However, dealing with this is exactly as in the case of linear models, so we'll just treat the iid case here,

$$\varepsilon_t \sim iid(0, \sigma^2)$$

If we stack the observations vertically, defining

$$\mathbf{y} = (y_1, y_2, ..., y_n)'$$

$$\mathbf{f} = (f(x_1, \theta), f(x_1, \theta), ..., f(x_1, \theta))'$$

and

$$\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$$

we can write the $n$ observations as

$$\mathbf{y} = \mathbf{f}(\theta) + \varepsilon$$

Using this notation, the NLS estimator can be defined as

$$\hat{\theta} \equiv \arg\min_{\Theta} s_n(\theta) = \frac{1}{n}\left[\mathbf{y} - \mathbf{f}(\theta)\right]'\left[\mathbf{y} - \mathbf{f}(\theta)\right] = \frac{1}{n} \parallel \mathbf{y} - \mathbf{f}(\theta) \parallel^2$$

- The estimator minimizes the weighted sum of squared errors, which is the same as minimizing the Euclidean distance between $\mathbf{y}$ and $\mathbf{f}(\theta)$.

The objective function can be written as

$$s_n(\theta) = \frac{1}{n}\left[\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{f}(\theta) + \mathbf{f}(\theta)'\mathbf{f}(\theta)\right],$$

which gives the first order conditions

$$-\left[\frac{\partial}{\partial\theta}\mathbf{f}(\hat{\theta})'\right]\mathbf{y} + \left[\frac{\partial}{\partial\theta}\mathbf{f}(\hat{\theta})'\right]\mathbf{f}(\hat{\theta}) \equiv 0.$$

Define the $n \times K$ matrix

$$\mathbf{F}(\hat{\theta}) \equiv D_{\theta'}\mathbf{f}(\hat{\theta}). \tag{24.6}$$

In shorthand, use $\hat{\mathbf{F}}$ in place of $\mathbf{F}(\hat{\theta})$. Using this, the first order conditions can be written as

$$-\hat{\mathbf{F}}'\mathbf{y} + \hat{\mathbf{F}}'\mathbf{f}(\hat{\theta}) \equiv 0,$$

or

$$\hat{\mathbf{F}}'\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right] \equiv 0. \tag{24.7}$$

This bears a good deal of similarity to the f.o.c. for the linear model - the derivative of the prediction is orthogonal to the prediction error. If $\mathbf{f}(\theta) = \mathbf{X}\theta$, then $\hat{\mathbf{F}}$ is simply $\mathbf{X}$, so the f.o.c. (with spherical errors) simplify to

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta = 0,$$

the usual 0LS f.o.c.

We can interpret this geometrically: *INSERT drawings of geometrical de-*

*piction of OLS and NLS (see Davidson and MacKinnon, pgs. 8,13 and 46).*

- Note that the nonlinearity of the manifold leads to potential multiple local maxima, minima and saddlepoints: the objective function $s_n(\theta)$ is not necessarily well-behaved and may be difficult to minimize.

## Identification

As before, identification can be considered conditional on the sample, and asymptotically. The condition for asymptotic identification is that $s_n(\theta)$ tend to a limiting function $s_\infty(\theta)$ such that $s_\infty(\theta^0) < s_\infty(\theta)$, $\forall \theta \neq \theta^0$. This will be the case if $s_\infty(\theta^0)$ is strictly convex at $\theta^0$, which requires that $D^2_\theta s_\infty(\theta^0)$ be positive definite.

Consider the objective function:

$$
\begin{aligned}
s_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n}\left[y_t - f(\mathbf{x}_t,\theta)\right]^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[f(\mathbf{x}_t,\theta^0) + \varepsilon_t - f_t(\mathbf{x}_t,\theta)\right]^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[f_t(\theta^0) - f_t(\theta)\right]^2 + \frac{1}{n}\sum_{t=1}^{n}(\varepsilon_t)^2 \\
&- \frac{2}{n}\sum_{t=1}^{n}\left[f_t(\theta^0) - f_t(\theta)\right]\varepsilon_t
\end{aligned}
$$

- As in example 13.4, which illustrated the consistency of extremum estimators using OLS, we conclude that the second term will converge to a constant which does not depend upon $\theta$.

- A LLN can be applied to the third term to conclude that it converges pointwise to 0, as long as $\mathbf{f}(\theta)$ and $\varepsilon$ are uncorrelated.

- Next, pointwise convergence needs to be strengthened to uniform almost sure convergence. There are a number of possible assumptions one could

use. Here, we'll just assume it holds.

- Turning to the first term, we'll assume a pointwise law of large numbers applies, so

$$\frac{1}{n}\sum_{t=1}^{n}\left[f_t(\theta^0)-f_t(\theta)\right]^2 \overset{a.s.}{\to} \int\left[f(z,\theta^0)-f(z,\theta)\right]^2 d\mu(z), \tag{24.8}$$

  where $\mu(x)$ is the distribution function of $x$. In many cases, $f(x,\theta)$ *will* be bounded and continuous, for all $\theta \in \Theta$, so strengthening to uniform almost sure convergence is immediate. For example if $f(x,\theta) = [1+\exp(-x\theta)]^{-1}$, $f : \Re^K \to (0,1)$, a bounded range, and the function is continuous in $\theta$.

Given these results, it is clear that a minimizer is $\theta^0$. When considering identification (asymptotic), the question is whether or not there may be some other minimizer. A local condition for identification is that

$$\frac{\partial^2}{\partial\theta\partial\theta'}s_\infty(\theta) = \frac{\partial^2}{\partial\theta\partial\theta'}\int\left[f(x,\theta^0)-f(x,\theta)\right]^2 d\mu(x)$$

be positive definite at $\theta^0$. Evaluating this derivative, we obtain (after a little work)

$$\frac{\partial^2}{\partial\theta\partial\theta'}\int\left[f(x,\theta^0)-f(x,\theta)\right]^2 d\mu(x)\Big|_{\theta^0} = 2\int\left[D_\theta f(z,\theta^0)'\right]\left[D_{\theta'} f(z,\theta^0)\right]' d\mu(z)$$

the expectation of the outer product of the gradient of the regression function evaluated at $\theta^0$. (Note: the uniform boundedness we have already assumed allows passing the derivative through the integral, by the dominated convergence theorem.) This matrix will be positive definite (wp1) as long as the gradient vector is of full rank (wp1). The tangent space to the regression manifold must span a $K$ -dimensional space if we are to consistently estimate a $K$ -dimensional parameter vector. This is analogous to the requirement that there be no perfect colinearity in a linear model. This is a necessary condition for identification. Note that the LLN implies that the above expectation is equal to

$$\mathcal{J}_\infty(\theta^0) = 2\lim \mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n}$$

## Consistency

We simply assume that the conditions of Theorem 35 hold, so the estimator is consistent. Given that the strong stochastic equicontinuity conditions hold, as discussed above, and given the above identification conditions an a compact estimation space (the closure of the parameter space $\Theta$), the consistency proof's assumptions are satisfied.

## Asymptotic normality

As in the case of GMM, we also simply assume that the conditions for asymptotic normality as in Theorem 37 hold. The only remaining problem is to determine the form of the asymptotic variance-covariance matrix. Recall that the result of the asymptotic normality theorem is

$$\sqrt{n}\left(\hat{\theta}-\theta^0\right) \xrightarrow{d} N\left[0, \mathcal{J}_\infty(\theta^0)^{-1}\mathcal{I}_\infty(\theta^0)\mathcal{J}_\infty(\theta^0)^{-1}\right],$$

where $\mathcal{J}_\infty(\theta^0)$ is the almost sure limit of $\frac{\partial^2}{\partial\theta\partial\theta'}s_n(\theta)$ evaluated at $\theta^0$, and

$$\mathcal{I}_\infty(\theta^0) = \lim Var\sqrt{n}D_\theta s_n(\theta^0)$$

The objective function is

$$s_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\left[y_t - f(\mathbf{x}_t,\theta)\right]^2$$

So

$$D_\theta s_n(\theta) = -\frac{2}{n}\sum_{t=1}^{n}\left[y_t - f(\mathbf{x}_t,\theta)\right]D_\theta f(\mathbf{x}_t,\theta).$$

Evaluating at $\theta^0$,

$$D_\theta s_n(\theta^0) = -\frac{2}{n}\sum_{t=1}^{n}\varepsilon_t D_\theta f(\mathbf{x}_t,\theta^0).$$

Note that the expectation of this is zero, since $\epsilon_t$ and $\mathbf{x}_t$ are assumed to be uncorrelated. So to calculate the variance, we can simply calculate the second moment

about zero. Also note that

$$\begin{aligned}\sum_{t=1}^{n} \varepsilon_t D_\theta f(\mathbf{x}_t, \theta^0) &= \frac{\partial}{\partial \theta}\left[\mathbf{f}(\theta^0)\right]' \varepsilon \\ &= \mathbf{F}'\varepsilon\end{aligned}$$

With this we obtain

$$\begin{aligned}\mathcal{I}_\infty(\theta^0) &= \lim Var\sqrt{n} D_\theta s_n(\theta^0) \\ &= \lim n\mathcal{E}\frac{4}{n^2}\mathbf{F}'\varepsilon\varepsilon^{'}\mathbf{F} \\ &= 4\sigma^2 \lim \mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n}\end{aligned}$$

We've already seen that

$$\mathcal{J}_\infty(\theta^0) = 2\lim \mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n},$$

where the expectation is with respect to the joint density of $x$ and $\varepsilon$. Combining these expressions for $\mathcal{J}_\infty(\theta^0)$ and $\mathcal{I}_\infty(\theta^0)$, and the result of the asymptotic

normality theorem, we get

$$\sqrt{n}\left(\hat{\theta}-\theta^{0}\right) \xrightarrow{d} N\left(0,\left(\lim \mathcal{E}\frac{\mathbf{F}'\mathbf{F}}{n}\right)^{-1}\sigma^{2}\right).$$

We can consistently estimate the variance covariance matrix using

$$\left(\frac{\hat{\mathbf{F}}'\hat{\mathbf{F}}}{n}\right)^{-1}\hat{\sigma}^{2}, \tag{24.9}$$

where $\hat{\mathbf{F}}$ is defined as in equation 24.6 and

$$\hat{\sigma}^{2}=\frac{\left[\mathbf{y}-\mathbf{f}(\hat{\theta})\right]'\left[\mathbf{y}-\mathbf{f}(\hat{\theta})\right]}{n},$$

the obvious estimator. Note the close correspondence to the results for the linear model.

# Example: The Poisson model for count data

Suppose that $y_t$ conditional on $\mathbf{x}_t$ is independently distributed Poisson. A Poisson random variable is a *count data* variable, which means it can take the values {0,1,2,...}. This sort of model has been used to study visits to doctors per year, number of patents registered by businesses per year, *etc.*

The Poisson density is

$$f(y_t) = \frac{\exp(-\lambda_t)\lambda_t^{y_t}}{y_t!}, y_t \in \{0, 1, 2, ...\}.$$

The mean of $y_t$ is $\lambda_t$, as is the variance. Note that $\lambda_t$ must be positive. Suppose that the true mean is

$$\lambda_t^0 = \exp(\mathbf{x}_t'\beta^0),$$

which enforces the positivity of $\lambda_t$. Suppose we estimate $\beta^0$ by nonlinear least squares:

$$\hat{\beta} = \arg\min s_n(\beta) = \frac{1}{T}\sum_{t=1}^{n}\left(y_t - \exp(\mathbf{x}_t'\beta)\right)^2$$

We can write

$$
\begin{aligned}
s_n(\beta) &= \frac{1}{T}\sum_{t=1}^{n}\left(\exp(\mathbf{x}_t'\beta^0+\varepsilon_t-\exp(\mathbf{x}_t'\beta)\right)^2 \\
&= \frac{1}{T}\sum_{t=1}^{n}\left(\exp(\mathbf{x}_t'\beta^0-\exp(\mathbf{x}_t'\beta)\right)^2+\frac{1}{T}\sum_{t=1}^{n}\varepsilon_t^2+2\frac{1}{T}\sum_{t=1}^{n}\varepsilon_t\left(\exp(\mathbf{x}_t'\beta^0-\exp(\mathbf{x}_t'\beta)\right)
\end{aligned}
$$

The last term has expectation zero since the assumption that $\mathcal{E}(y_t|\mathbf{x}_t)=\exp(\mathbf{x}_t'\beta^0)$ implies that $\mathcal{E}\left(\varepsilon_t|\mathbf{x}_t\right)=0$, which in turn implies that functions of $\mathbf{x}_t$ are uncorrelated with $\varepsilon_t$. Applying a strong LLN, and noting that the objective function is continuous on a compact parameter space, we get

$$
s_\infty(\beta)=\mathcal{E}_{\mathbf{x}}\left(\exp(\mathbf{x}'\beta^0-\exp(\mathbf{x}'\beta)\right)^2+\mathcal{E}_{\mathbf{x}}\exp(\mathbf{x}'\beta^0)
$$

where the last term comes from the fact that the conditional variance of $\varepsilon$ is the same as the variance of $y$. This function is clearly minimized at $\beta=\beta^0$, so the NLS estimator is consistent as long as identification holds.

**Exercise 114.** Determine the limiting distribution of $\sqrt{n}\left(\hat{\beta}-\beta^0\right)$. This means

finding the the specific forms of $\frac{\partial^2}{\partial\beta\partial\beta'}s_n(\beta)$, $\mathcal{J}(\beta^0)$, $\left.\frac{\partial s_n(\beta)}{\partial\beta}\right|$, and $\mathcal{I}(\beta^0)$. Again, use a CLT as needed, no need to verify that it can be applied.

## The Gauss-Newton algorithm

**Readings:** Davidson and MacKinnon, Chapter 6, pgs. 201-207*.

The Gauss-Newton optimization technique is specifically designed for nonlinear least squares. The idea is to linearize the nonlinear model, rather than the objective function. The model is

$$\mathbf{y} = \mathbf{f}(\theta^0) + \varepsilon.$$

At some $\theta$ in the parameter space, not equal to $\theta^0$, we have

$$\mathbf{y} = \mathbf{f}(\theta) + \nu$$

where $\nu$ is a combination of the fundamental error term $\varepsilon$ and the error due to evaluating the regression function at $\theta$ rather than the true value $\theta^0$. Take a first

order Taylor's series approximation around a point $\theta^1$ :

$$\mathbf{y} = \mathbf{f}(\theta^1) + \left[D_{\theta'}\mathbf{f}\left(\theta^1\right)\right]\left(\theta - \theta^1\right) + \nu + \text{approximation error.}$$

Define $\mathbf{z} \equiv \mathbf{y} - \mathbf{f}(\theta^1)$ and $b \equiv (\theta - \theta^1)$. Then the last equation can be written as

$$\mathbf{z} = \mathbf{F}(\theta^1)b + \omega,$$

where, as above, $\mathbf{F}(\theta^1) \equiv D_{\theta'}\mathbf{f}(\theta^1)$ is the $n \times K$ matrix of derivatives of the regression function, evaluated at $\theta^1$, and $\omega$ is $\nu$ plus approximation error from the truncated Taylor's series.

- Note that $\mathbf{F}$ is known, given $\theta^1$.
- Note that one could estimate $b$ simply by performing OLS on the above equation.
- Given $\hat{b}$, we calculate a new round estimate of $\theta^0$ as $\theta^2 = \hat{b} + \theta^1$. With this,

take a new Taylor's series expansion around $\theta^2$ and repeat the process. Stop when $\hat{b} = 0$ (to within a specified tolerance).

To see why this might work, consider the above approximation, but evaluated at the NLS estimator:

$$\mathbf{y} = \mathbf{f}(\hat{\theta}) + \mathbf{F}(\hat{\theta})\left(\theta - \hat{\theta}\right) + \omega$$

The OLS estimate of $b \equiv \theta - \hat{\theta}$ is

$$\hat{b} = \left(\hat{\mathbf{F}}'\hat{\mathbf{F}}\right)^{-1}\hat{\mathbf{F}}'\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right].$$

This must be zero, since

$$\hat{\mathbf{F}}'\left(\hat{\theta}\right)\left[\mathbf{y} - \mathbf{f}(\hat{\theta})\right] \equiv 0$$

by definition of the NLS estimator (these are the normal equations as in equation 24.7, Since $\hat{b} \equiv 0$ when we evaluate at $\hat{\theta}$, updating would stop.

- The Gauss-Newton method doesn't require second derivatives, as does the Newton-Raphson method, so it's faster.

- The varcov estimator, as in equation 24.9 is simple to calculate, since we have $\hat{\mathbf{F}}$ as a by-product of the estimation process (*i.e.,* it's just the last round "regressor matrix"). In fact, a normal OLS program will give the NLS varcov estimator directly, since it's just the OLS varcov estimator from the last iteration.

- The method can suffer from convergence problems since $\mathbf{F}(\theta)'\mathbf{F}(\theta)$, may be very nearly singular, even with an asymptotically identified model, especially if $\theta$ is very far from $\hat{\theta}$. Consider the example

$$y = \beta_1 + \beta_2 x_t \beta^3 + \varepsilon_t$$

  When evaluated at $\beta_2 \approx 0$, $\beta_3$ has virtually no effect on the NLS objective function, so $\mathbf{F}$ will have rank that is "essentially" 2, rather than 3. In this case, $\mathbf{F}'\mathbf{F}$ will be nearly singular, so $(\mathbf{F}'\mathbf{F})^{-1}$ will be subject to large roundoff errors.

# Application: Limited dependent variables and sample selection

**Readings**: Davidson and MacKinnon, Ch. 15* (a quick reading is sufficient), J. Heckman, "Sample Selection Bias as a Specification Error", *Econometrica*, 1979 (This is a classic article, not required for reading, and which is a bit out-dated. Nevertheless it's a good place to start if you encounter sample selection problems in your research).

Sample selection is a common problem in applied research. The problem occurs when observations used in estimation are sampled non-randomly, according to some selection scheme.

# Example: Labor Supply

Labor supply of a person is a positive number of hours per unit time supposing the offer wage is higher than the reservation wage, which is the wage at which the

person prefers not to work. The model (very simple, with $t$ subscripts suppressed):

- Characteristics of individual: $\mathbf{x}$
- Latent labor supply: $s^* = \mathbf{x}'\beta + \omega$
- Offer wage: $w^o = \mathbf{z}'\gamma + \nu$
- Reservation wage: $w^r = \mathbf{q}'\delta + \eta$

Write the wage differential as

$$\begin{aligned} w^* &= (\mathbf{z}'\gamma + \nu) - (\mathbf{q}'\delta + \eta) \\ &\equiv \mathbf{r}'\theta + \varepsilon \end{aligned}$$

We have the set of equations

$$\begin{aligned} s^* &= \mathbf{x}'\beta + \omega \\ w^* &= \mathbf{r}'\theta + \varepsilon. \end{aligned}$$

Assume that

$$\begin{bmatrix} \omega \\ \varepsilon \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \right).$$

We assume that the offer wage and the reservation wage, as well as the latent variable $s^*$ are unobservable. What is observed is

$$\begin{aligned} w &= 1\left[w^* > 0\right] \\ s &= ws^*. \end{aligned}$$

In other words, we observe whether or not a person is working. If the person is working, we observe labor supply, which is equal to latent labor supply, $s^*$. Otherwise, $s = 0 \neq s^*$. Note that we are using a simplifying assumption that individuals can freely choose their weekly hours of work.

Suppose we estimated the model

$$s^* = \mathbf{x}'\beta + \text{residual}$$

using only observations for which $s > 0$. The problem is that these observations are those for which $w^* > 0$, or equivalently, $-\varepsilon < \mathbf{r}'\theta$ and

$$\mathcal{E}\left[\omega | -\varepsilon < \mathbf{r}'\theta\right] \neq 0,$$

since $\varepsilon$ and $\omega$ are dependent. Furthermore, this expectation will in general depend on $\mathbf{x}$ since elements of $\mathbf{x}$ can enter in $\mathbf{r}$. Because of these two facts, least squares estimation is biased and inconsistent.

Consider more carefully $\mathcal{E}\left[\omega | -\varepsilon < \mathbf{r}'\theta\right]$. Given the joint normality of $\omega$ and $\varepsilon$, we can write (see for example Spanos *Statistical Foundations of Econometric Modelling,* pg. 122)

$$\omega = \rho\sigma\varepsilon + \eta,$$

where $\eta$ has mean zero and is independent of $\varepsilon$. With this we can write

$$s^* = \mathbf{x}'\beta + \rho\sigma\varepsilon + \eta.$$

If we condition this equation on $-\varepsilon < \mathbf{r}'\theta$ we get

$$s = \mathbf{x}'\beta + \rho\sigma\mathcal{E}(\varepsilon| - \varepsilon < \mathbf{r}'\theta) + \eta$$

which may be written as

$$s = \mathbf{x}'\beta + \rho\sigma\mathcal{E}(\varepsilon|\varepsilon > -\mathbf{r}'\theta) + \eta$$

- A useful result is that for

$$z \sim N(0, 1)$$

$$E(z|z > z^*) = \frac{\phi(z^*)}{\Phi(-z^*)},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution function, respectively. The quantity on the RHS above is known as the *inverse Mill's ratio:*

$$IMR(\mathbf{z}^*) = \frac{\phi(z^*)}{\Phi(-z^*)}$$

With this we can write (making use of the fact that the standard normal density is symmetric about zero, so that $\phi(-a) = \phi(a)$):

$$s = \mathbf{x}'\beta + \rho\sigma\frac{\phi\left(\mathbf{r}'\theta\right)}{\Phi\left(\mathbf{r}'\theta\right)} + \eta \tag{24.10}$$

$$\equiv \left[\begin{array}{cc} \mathbf{x}' & \frac{\phi\left(\mathbf{r}'\theta\right)}{\Phi\left(\mathbf{r}'\theta\right)} \end{array}\right]\left[\begin{array}{c} \beta \\ \zeta \end{array}\right] + \eta. \tag{24.11}$$

where $\zeta = \rho\sigma$. The error term $\eta$ has conditional mean zero, and is uncorrelated with the regressors $\mathbf{x}' \ \frac{\phi\left(\mathbf{r}'\theta\right)}{\Phi\left(\mathbf{r}'\theta\right)}$. At this point, we can estimate the equation by NLS.

- Heckman showed how one can estimate this in a two step procedure where first $\theta$ is estimated, then equation 24.11 is estimated by least squares using the estimated value of $\theta$ to form the regressors. This is inefficient and estimation of the covariance is a tricky issue. It is probably easier (and more efficient) just to do MLE.

- The model presented above depends strongly on joint normality. There exist

many alternative models which weaken the maintained assumptions. It is possible to estimate consistently without distributional assumptions. See Ahn and Powell, *Journal of Econometrics*, 1994.

## 24.9 The Fourier functional form

This material was removed from the chapter on nonparametric regression, to make that chapter easier to read, and to focus on the main ideas.

**Readings**: Gallant, 1987, "Identification and consistency in semi-nonparametric regression," in *Advances in Econometrics, Fifth World Congress,* V. 1, Truman Bewley, ed., Cambridge.

Suppose we have a multivariate model

$$y = f(\mathbf{x}) + \varepsilon,$$

where $f(x)$ is of unknown form and $x$ is a $P-$dimensional vector. For simplicity,

assume that $\varepsilon$ is a classical error. Let us take the estimation of the vector of elasticities with typical element

$$\xi_{x_i} = \frac{\mathbf{x}_i}{f(\mathbf{x})} \frac{\partial f(\mathbf{x})}{\partial x_i f(x)},$$

at an arbitrary point $\mathbf{x}_i$.

The Fourier form, following Gallant (1982), but with a somewhat different parameterization, may be written as

$$g_K(\mathbf{x} \mid \theta_K) = \alpha + \mathbf{x}'\beta + 1/2\mathbf{x}'\mathbf{C}\mathbf{x} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left(u_{j\alpha} \cos(j\mathbf{k}'_\alpha \mathbf{x}) - v_{j\alpha} \sin(j\mathbf{k}'_\alpha \mathbf{x})\right). \tag{24.12}$$

where the $K$-dimensional parameter vector

$$\theta_K = \{\alpha, \beta', vec^*(C)', u_{11}, v_{11}, \ldots, u_{JA}, v_{JA}\}'. \tag{24.13}$$

- We assume that the conditioning variables $\mathbf{x}$ have each been transformed to

lie in an interval that is shorter than $2\pi$. This is required to avoid periodic behavior of the approximation, which is desirable since economic functions aren't periodic. For example, subtract sample means, divide by the maxima of the conditioning variables, and multiply by $2\pi - eps$, where $eps$ is some positive number less than $2\pi$ in value.

- The $k_\alpha$ are "elementary multi-indices" which are simply $P-$ vectors formed of integers (negative, positive and zero). The $k_\alpha$, $\alpha = 1, 2, ..., A$ are required to be linearly independent, and we follow the convention that the first non-zero element be positive. For example

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 1 \end{bmatrix}'$$

is a potential multi-index to be used, but

$$\begin{bmatrix} 0 & -1 & -1 & 0 & 1 \end{bmatrix}'$$

is not since its first nonzero element is negative. Nor is

$$\begin{bmatrix} 0 & 2 & -2 & 0 & 2 \end{bmatrix}'$$

a multi-index we would use, since it is a scalar multiple of the original multi-index.

- We parameterize the matrix $C$ differently than does Gallant because it simplifies things in practice. The cost of this is that we are no longer able to test a quadratic specification using nested testing.

The vector of first partial derivatives is

$$D_x g_K(\mathbf{x} \mid \theta_K) = \beta + \mathbf{C}\mathbf{x} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left[ \left( -u_{j\alpha} \sin(j\mathbf{k}_\alpha' \mathbf{x}) - v_{j\alpha} \cos(j\mathbf{k}_\alpha' \mathbf{x}) \right) j\mathbf{k}_\alpha \right] \tag{24.14}$$

and the matrix of second partial derivatives is

$$D_x^2 g_K(\mathbf{x}|\theta_K) = \mathbf{C} + \sum_{\alpha=1}^{A} \sum_{j=1}^{J} \left[ \left(-u_{j\alpha} \cos(j\mathbf{k}'_\alpha \mathbf{x}) + v_{j\alpha} \sin(j\mathbf{k}'_\alpha \mathbf{x})\right) j^2 \mathbf{k}_\alpha \mathbf{k}'_\alpha \right] \quad (24.15)$$

To define a compact notation for partial derivatives, let $\lambda$ be an $N$-dimensional multi-index with no negative elements. Define $\mid \lambda \mid^*$ as the sum of the elements of $\lambda$. If we have $N$ arguments $\mathbf{x}$ of the (arbitrary) function $h(\mathbf{x})$, use $D^\lambda h(\mathbf{x})$ to indicate a certain partial derivative:

$$D^\lambda h(\mathbf{x}) \equiv \frac{\partial^{|\lambda|^*}}{\partial x_1^{\lambda_1} \partial x_2^{\lambda_2} \cdots \partial x_N^{\lambda_N}} h(\mathbf{x})$$

When $\lambda$ is the zero vector, $D^\lambda h(\mathbf{x}) \equiv h(\mathbf{x})$. Taking this definition and the last few equations into account, we see that it is possible to define $(1 \times K)$ vector $Z^\lambda(\mathbf{x})$ so that

$$D^\lambda g_K(\mathbf{x}|\theta_K) = \mathbf{z}^\lambda(\mathbf{x})' \theta_K. \quad (24.16)$$

- Both the approximating model and the derivatives of the approximating

model are linear in the parameters.

- For the approximating model to the function (not derivatives), write $g_K(\mathbf{x}|\theta_K) = \mathbf{z}'\theta_K$ for simplicity.

The following theorem can be used to prove the consistency of the Fourier form.

**Theorem 115.** *[Gallant and Nychka, 1987] Suppose that $\hat{h}_n$ is obtained by maximizing a sample objective function $s_n(h)$ over $\mathcal{H}_{K_n}$ where $\mathcal{H}_K$ is a subset of some function space $\mathcal{H}$ on which is defined a norm $\| h \|$. Consider the following conditions:*

*(a) Compactness: The closure of $\mathcal{H}$ with respect to $\| h \|$ is compact in the relative topology defined by $\| h \|$.*

*(b) Denseness: $\cup_K \mathcal{H}_K$, $K = 1, 2, 3, ...$ is a dense subset of the closure of $\mathcal{H}$ with respect to $\| h \|$ and $\mathcal{H}_K \subset \mathcal{H}_{K+1}$.*

*(c) Uniform convergence: There is a point $h^*$ in $\mathcal{H}$ and there is a function*

*$s_\infty(h, h^*)$ that is continuous in $h$ with respect to $\| h \|$ such that*

$$\lim_{n\to\infty} \sup_{\overline{\mathcal{H}}} \mid s_n(h) - s_\infty(h, h^*) \mid = 0$$

*almost surely.*

*(d) Identification: Any point $h$ in the closure of $\mathcal{H}$ with $s_\infty(h, h^*) \geq s_\infty(h^*, h^*)$ must have $\| h - h^* \| = 0$.*

*Under these conditions $\lim_{n\to\infty} \| h^* - \hat{h}_n \| = 0$ almost surely, provided that $\lim_{n\to\infty} K_n = \infty$ almost surely.*

The modification of the original statement of the theorem that has been made is to set the parameter space $\Theta$ in Gallant and Nychka's (1987) Theorem 0 to a single point and to state the theorem in terms of maximization rather than minimization.

This theorem is very similar in form to Theorem 35. The main differences are:

1. A generic norm $\| h \|$ is used in place of the Euclidean norm. This norm

may be stronger than the Euclidean norm, so that convergence with respect to $\| h \|$ implies convergence w.r.t the Euclidean norm. Typically we will want to make sure that the norm is strong enough to imply convergence of all functions of interest.

2. The "estimation space" $\mathcal{H}$ is a function space. It plays the role of the parameter space $\Theta$ in our discussion of parametric estimators. There is no restriction to a parametric family, only a restriction to a space of functions that satisfy certain conditions. This formulation is much less restrictive than the restriction to a parametric family.

3. There is a denseness assumption that was not present in the other theorem.

We will not prove this theorem (the proof is quite similar to the proof of theorem [35], see Gallant, 1987) but we will discuss its assumptions, in relation to the Fourier form as the approximating model.

**Sobolev norm** Since all of the assumptions involve the norm $\| h \|$ , we need to make explicit what norm we wish to use. We need a norm that guarantees that the errors in approximation of the functions we are interested in are accounted for. Since we are interested in first-order elasticities in the present case, we need close approximation of both the function $f(x)$ and its first derivative $f'(x)$, throughout the range of $x$. Let $\mathcal{X}$ be an open set that contains all values of $x$ that we're interested in. The Sobolev norm is appropriate in this case. It is defined, making use of our notation for partial derivatives, as:

$$\| h \|_{m,\mathcal{X}} = \max_{|\lambda^*| \leq m} \sup_{\mathcal{X}} \left| D^{\lambda} h(x) \right|$$

To see whether or not the function $f(x)$ is well approximated by an approximating model $g_K(x \mid \theta_K)$, we would evaluate

$$\| f(\mathbf{x}) - g_K(\mathbf{x} \mid \theta_K) \|_{m,\mathcal{X}} \ .$$

We see that this norm takes into account errors in approximating the function and partial derivatives up to order $m$. If we want to estimate first order elasticities, as is the case in this example, the relevant $m$ would be $m = 1$. Furthermore, since we examine the sup over $\mathcal{X}$, convergence w.r.t. the Sobolev means *uniform* convergence, so that we obtain consistent estimates for all values of $x$.

**Compactness** Verifying compactness with respect to this norm is quite technical and unenlightening. It is proven by Elbadawi, Gallant and Souza, *Econometrica*, 1983. The basic requirement is that if we need consistency w.r.t. $\| h \|_{m,\mathcal{X}}$, then the functions of interest must belong to a Sobolev space which takes into account derivatives of order $m + 1$. A Sobolev space is the set of functions

$$\mathcal{W}_{m,\mathcal{X}}(D) = \{h(\mathbf{x}) : \| h(\mathbf{x}) \|_{m,\mathcal{X}} < D\},$$

where $D$ is a finite constant. In plain words, the functions must have bounded partial derivatives of one order higher than the derivatives we seek to estimate.

**The estimation space and the estimation subspace** Since in our case we're interested in consistent estimation of first-order elasticities, we'll define the estimation space as follows:

**Definition 116.** [Estimation space] The estimation space $\mathcal{H} = \mathcal{W}_{2,\mathcal{X}}(D)$. The estimation space is an open set, and we presume that $h^* \in \mathcal{H}$.

So we are assuming that the function to be estimated has bounded second derivatives throughout $\mathcal{X}$.

With seminonparametric estimators, we don't actually optimize over the estimation space. Rather, we optimize over a subspace, $\mathcal{H}_{K_n}$, defined as:

**Definition 117.** [Estimation subspace] The estimation subspace $\mathcal{H}_K$ is defined as

$$\mathcal{H}_K = \{g_K(\mathbf{x}|\theta_K) : g_K(\mathbf{x}|\theta_K) \in \mathcal{W}_{2,\mathcal{Z}}(D), \theta_K \in \Re^K\},$$

where $g_K(\mathbf{x}, \theta_K)$ is the Fourier form approximation as defined in Equation 24.12.

**Denseness** The important point here is that $\mathcal{H}_K$ is a space of functions that is indexed by a finite dimensional parameter ($\theta_K$ has $K$ elements, as in equation 24.13). With $n$ observations, $n > K$, this parameter is estimable. Note that the true function $h^*$ is not necessarily an element of $\mathcal{H}_K$, so optimization over $\mathcal{H}_K$ may not lead to a consistent estimator. In order for optimization over $\mathcal{H}_K$ to be equivalent to optimization over $\mathcal{H}$, at least asymptotically, we need that:

1. The dimension of the parameter vector, $\dim \theta_{K_n} \to \infty$ as $n \to \infty$. This is achieved by making $A$ and $J$ in equation 24.12 increasing functions of $n$, the sample size. It is clear that $K$ will have to grow more slowly than $n$. The second requirement is:

2. We need that the $\mathcal{H}_K$ be dense subsets of $\mathcal{H}$.

The estimation subspace $\mathcal{H}_K$, defined above, is a subset of the closure of the estimation space, $\overline{\mathcal{H}}$ . A set of subsets $\mathcal{A}_a$ of a set $\mathcal{A}$ is "dense" if the closure of

the countable union of the subsets is equal to the closure of $\mathcal{A}$:

$$\overline{\cup_{a=1}^{\infty}\mathcal{A}_a} = \overline{\mathcal{A}}$$

*Use a picture here. The rest of the discussion of denseness is provided just for completeness: there's no need to study it in detail.* To show that $\mathcal{H}_K$ is a dense subset of $\overline{\mathcal{H}}$ with respect to $\| h \|_{1,\mathcal{X}}$, it is useful to apply Theorem 1 of Gallant (1982), who in turn cites Edmunds and Moscatelli (1977). We reproduce the theorem as presented by Gallant, with minor notational changes, for convenience of reference:

**Theorem 118.** *[Edmunds and Moscatelli, 1977] Let the real-valued function $h^*(\mathbf{x})$ be continuously differentiable up to order $m$ on an open set containing the closure of $\mathcal{X}$. Then it is possible to choose a triangular array of coefficients $\theta_1, \theta_2, \ldots \theta_K, \ldots,$ such that for every $q$ with $0 \leq q < m$, and every $\varepsilon > 0$, $\| h^*(\mathbf{x}) - h_K(\mathbf{x}|\theta_K) \|_{q,\mathcal{X}} = o(K^{-m+q+\varepsilon})$ as $K \to \infty$.*

In the present application, $q = 1$, and $m = 2$. By definition of the estimation

space, the elements of $\mathcal{H}$ are once continuously differentiable on $\mathcal{X}$, which is open and contains the closure of $\mathcal{X}$, so the theorem is applicable. Closely following Gallant and Nychka (1987), $\cup_\infty \mathcal{H}_K$ is the countable union of the $\mathcal{H}_K$. The implication of Theorem 118 is that there is a sequence of $\{h_K\}$ from $\cup_\infty \mathcal{H}_K$ such that

$$\lim_{K\to\infty} \| h^* - h_K \|_{1,\mathcal{X}} = 0,$$

for all $h^* \in \mathcal{H}$. Therefore,

$$\mathcal{H} \subset \overline{\cup_\infty \mathcal{H}_K}.$$

However,

$$\cup_\infty \mathcal{H}_K \subset \mathcal{H},$$

so

$$\overline{\cup_\infty \mathcal{H}_K} \subset \overline{\mathcal{H}}.$$

Therefore

$$\overline{\mathcal{H}} = \overline{\cup_\infty \mathcal{H}_K},$$

so $\cup_\infty \mathcal{H}_K$ is a dense subset of $\mathcal{H}$, with respect to the norm $\| h \|_{1,\mathcal{X}}$.

**Uniform convergence** We now turn to the limiting objective function. We estimate by OLS. The sample objective function stated in terms of maximization is

$$s_n(\theta_K) = -\frac{1}{n} \sum_{t=1}^{n} \left(y_t - g_K(\mathbf{x}_t \mid \theta_K)\right)^2$$

With random sampling, as in the case of Equations 13.1 and 24.8, the limiting objective function is

$$s_\infty(g, f) = -\int_{\mathcal{X}} \left(f(\mathbf{x}) - g(\mathbf{x})\right)^2 d\mu x - \sigma_\varepsilon^2. \tag{24.17}$$

where the true function $f(x)$ takes the place of the generic function $h^*$ in the presentation of the theorem. Both $g(x)$ and $f(x)$ are elements of $\overline{\cup_\infty \mathcal{H}_K}$.

The pointwise convergence of the objective function needs to be strengthened to uniform convergence. We will simply assume that this holds, since the way to verify this depends upon the specific application. We also have continuity of the

objective function in $g$, with respect to the norm $\| h \|_{1,\mathcal{X}}$ since

$$\begin{aligned} & \lim_{\|g^1-g^0\|_{1,\mathcal{X}}\to 0} \left\{ s_\infty\left(g^1, f\right) - s_\infty\left(g^0, f\right) \right\} \\ = & \lim_{\|g^1-g^0\|_{1,\mathcal{X}}\to 0} \int_{\mathcal{X}} \left[ \left(g^1(\mathbf{x}) - f(\mathbf{x})\right)^2 - \left(g^0(\mathbf{x}) - f(\mathbf{x})\right)^2 \right] d\mu x. \end{aligned}$$

By the dominated convergence theorem (which applies since the finite bound $D$ used to define $\mathcal{W}_{2,\mathcal{Z}}(D)$ is dominated by an integrable function), the limit and the integral can be interchanged, so by inspection, the limit is zero.

**Identification** The identification condition requires that for any point $(g, f)$ in $\overline{\mathcal{H}} \times \overline{\mathcal{H}}$, $s_\infty(g, f) \geq s_\infty(f, f) \Rightarrow \| g - f \|_{1,\mathcal{X}} = 0$. This condition is clearly satisfied given that $g$ and $f$ are once continuously differentiable (by the assumption that defines the estimation space).

**Review of concepts** For the example of estimation of first-order elasticities, the relevant concepts are:

- Estimation space $\mathcal{H} = \mathcal{W}_{2,\mathcal{X}}(D)$: the function space in the closure of which the true function must lie.

- Consistency norm $\| \ h \ \|_{1,\mathcal{X}}$ . The closure of $\mathcal{H}$ is compact with respect to this norm.

- Estimation subspace $\mathcal{H}_K$. The estimation subspace is the subset of $\mathcal{H}$ that is representable by a Fourier form with parameter $\theta_K$. These are dense subsets of $\mathcal{H}$.

- Sample objective function $s_n(\theta_K)$, the negative of the sum of squares. By standard arguments this converges uniformly to the

- Limiting objective function $s_\infty(\ g, f)$, which is continuous in $g$ and has a global maximum in its first argument, over the closure of the infinite union of the estimation subpaces, at $g = f$.

- As a result of this, first order elasticities

$$\frac{\mathbf{x}_i}{f(\mathbf{x})}\frac{\partial f(\mathbf{x})}{\partial x_i f(x)}$$

  are consistently estimated for all $\mathbf{x} \in \mathcal{X}$.

**Discussion** Consistency requires that the number of parameters used in the expansion increase with the sample size, tending to infinity. If parameters are added at a high rate, the bias tends relatively rapidly to zero. A basic problem is that a high rate of inclusion of additional parameters causes the variance to tend more slowly to zero. The issue of how to chose the rate at which parameters are added and which to add first is fairly complex. A problem is that the allowable rates for asymptotic normality to obtain (Andrews 1991; Gallant and Souza, 1991) are very strict. Supposing we stick to these rates, our approximating model is:

$$g_K(\mathbf{x}|\theta_K) = \mathbf{z}'\theta_K.$$

- Define $\mathbf{Z}_K$ as the $n \times K$ matrix of regressors obtained by stacking observations. The LS estimator is

$$\hat{\theta}_K = \left(\mathbf{Z}_K'\mathbf{Z}_K\right)^+ \mathbf{Z}_K' y,$$

where $(\cdot)^+$ is the Moore-Penrose generalized inverse.

  - This is used since $\mathbf{Z}_K'\mathbf{Z}_K$ may be singular, as would be the case for $K(n)$ large enough when some indicator variables are included.

- . The prediction, $\mathbf{z}'\hat{\theta}_K$, of the unknown function $f(\mathbf{x})$ is asymptotically normally distributed:

$$\sqrt{n}\left(\mathbf{z}'\hat{\theta}_K - f(x)\right) \xrightarrow{d} N(0, AV),$$

where

$$AV = \lim_{n\to\infty} E\left[\mathbf{z}'\left(\frac{\mathbf{Z}_K'\mathbf{Z}_K}{n}\right)^+ \mathbf{z}\hat{\sigma}^2\right].$$

Formally, this is exactly the same as if we were dealing with a parametric linear model. I emphasize, though, that this is only valid if $K$ grows very slowly as $n$ grows. If we can't stick to acceptable rates, we should probably use some other method of approximating the small sample distribution. Bootstrapping is a possibility. We'll discuss this in the section on simulation.

# Bibliography

Joshua D Angrist and Alan B Keueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991. 1137

S. Borağan Aruoba, Jesús Fernández-Villaverde, and Juan F. Rubio-Ramírez. Comparing solution methods for dynamic equilibrium economies. *Journal of Economic Dynamics and Control*, 30(12):2477–2508, dec 2006. doi: 10.1016/j.jedc.2005.07.008. URL http://dx.doi.org/10.1016/j.jedc.2005.07.008. 570

Herman J Bierens. Kernel estimators of regression functions. In *Advances in econometrics: Fifth world congress*, volume 1, pages 99–144, 1987. 1057

A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005. 26, 439, 495, 582, 712, 953, 995, 1041, 1057, 1113, 1140

David Card. Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research, 1993. URL http://www.nber.org/papers/w4483.pdf. 822, 1128

Victor Chernozhukov and Christian Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005. 1113, 1132

Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006. 1137

Victor Chernozhukov and Han Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, aug 2003. doi: 10.1016/s0304-4076(03)00100-3. URL http://dx.doi.org/10.1016/s0304-4076(03)00100-3. 851, 954, 962, 965, 988, 1133

Michael Creel. Neural nets for indirect inference. *Econometrics and Statistics*, 2:36–49, 2017. URL https://doi.org/10.1016/j.ecosta.2016.11.008. 1074

Michael Creel. Inference using simulated neural moments. *Econometrics*, 9(4): 35, 2021. URL https://doi.org/10.3390/econometrics9040035. 1178

Stephen G Donald and Whitney K Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001. URL https://doi.org/10.1111/1468-0262.00238. 821

Stephen G. Donald, Guido W. Imbens, and Whitney K. Newey. Choosing instrumental variables in conditional moment restriction models. *Journal of Econo-*

*metrics*, 152(1):28–36, 2009. ISSN 0304-4076. doi: 10.1016/j.jeconom.2008.10.013. URL http://www.sciencedirect.com/science/article/pii/S0304407609000566. Recent Adavances in Nonparametric and Semiparametric Econometrics: A Volume Honouring Peter M. Robinson. 821, 1174

Darrell Duffie and Kenneth J Singleton. Simulated moments estimation of markov models of asset prices. *Econometrica (1986-1998)*, 61(4):929, 1993. 1141

Robert F Engle, David F Hendry, and Jean-Francois Richard. Exogeneity. *Econometrica: Journal of the Econometric Society*, pages 277–304, 1983. URL http://www.jstor.org/stable/1911990. 589

A. R. Gallant and H. White. There exists a neural network that does not make avoidable mistakes. In *Neural Networks, 1988., IEEE International Conference on*, pages 657–664 vol.1, July 1988. doi: 10.1109/ICNN.1988.23903. URL https://doi.org/10.1109/ICNN.1988.23903. 1074

A. Ronald Gallant. *Nonlinear statistical models*. Wiley series in probability and

mathematical statistics. Wiley, New York [u.a.], 1987a. ISBN 978-0-471-80260-0. 1185

A. Ronald Gallant. *Identification and consistency in semi-nonparametric regression*, volume 1 of *Econometric Society Monographs*, page 145â170. Cambridge University Press, 1987b. 510

A. Ronald Gallant. An introduction to econometric theory. *Princeton University*, 1997. 545, 614, 656, 795, 1198

A. Ronald Gallant and George Tauchen. Which moments to match? *Econometric Theory*, 12:363–390, 1996. doi: 10.2139/ssrn.37760. URL http://dx.doi.org/10.2139/ssrn.37760. 954, 1141, 1163, 1202

A. Ronald Gallant and George Tauchen. Emm: A program for efficient method of moments estimation, version 2.6, user's guide. *Manuscript, University of North Carolina*, 2010. URL https://aronaldg.org/courses/compecon/papers/emmguide.pdf. 954

C. Gouriéroux, A. Monfort, and E. Renault. Indirect inference. *Journal of Applied Econometrics*, pages S85–S118, 1993. doi: 10.1002/jae.3950080507. URL http://dx.doi.org/10.1002/jae.3950080507. 1141

Christian Gourieroux and Alain Monfort. *Simulation-based econometric methods*. Oxford university press, 1996. 1140

Christian Gourieroux, Alain Monfort, and Alain Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica: Journal of the Econometric Society*, pages 681–700, 1984. URL https://doi.org/10.2307/1913471. 678, 725

Christian Gouriéroux, Peter CB Phillips, and Jun Yu. Indirect inference for dynamic panel models. *Journal of Econometrics*, 157(1):68–77, 2010. 1162

Pablo A. Guerrón-Quintana. What you match does matter: the effects of data on DSGE estimation. *Journal of Applied Econometrics*, 25(5):774–804, 2010. ISSN 1099-1255. doi: 10.1002/jae.1106. URL http://dx.doi.org/10.1002/jae.1106. 579

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029, jul 1982. doi: 10.2307/1912775. URL http://dx.doi.org/10.2307/1912775. 712

Lars Peter Hansen and Kenneth J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5): 1269, sep 1982. doi: 10.2307/1911873. URL http://dx.doi.org/10.2307/1911873. 712, 826

Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, jul 1996. doi: 10.1080/07350015.1996.10524656. URL http://dx.doi.org/10.1080/07350015.1996.10524656. 760, 842, 850, 1174

J. A. Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251, nov

1978. doi: 10.2307/1913827. URL http://dx.doi.org/10.2307/1913827. 804

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8. URL http://dx.doi.org/10.1016/0893-6080(89)90020-8. 1074

Roger Koenker and Gilbert S Bassett. Regression quantiles. *Econometrica*, 46: 33–50, 1978. 1113

Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001. 1113

Chung-Ming Kuan and Halbert White. Artificial neural networks: an econometric perspective. *Econometric Reviews*, 13(1):1–91, 1994. doi: 10.1080/07474939408800273. URL http://dx.doi.org/10.1080/07474939408800273. 1074

Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In *Lecture Notes in Computer Science*, pages 9–48. Springer Science Mathplus Business Media, 2012. doi: 10.1007/978-3-642-35289-8_3. URL http://dx.doi.org/10.1007/978-3-642-35289-8_3. 1074

Lung-Fei Lee. Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. *Econometric Theory*, 11(3):437–483, 1995. 1155

Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007. 1041, 1057

Hedibert F. Lopes and Ruey S. Tsay. Particle filters and bayesian inference in financial econometrics. *Journal of Forecasting*, 30(1):168–209, 2011. ISSN 1099-131X. doi: 10.1002/for.1195. URL http://dx.doi.org/10.1002/for.1195. 941

Daniel McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026,

1989. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913621. 1140, 1145, 1160

W. Newey and D. McFadden. Large sample estimation and hypothesis testing. In R. Engle and D. McFadden, editors, *Handbook of Econometrics, Vol. 4*, pages 2113–2241. North Holland, 1994. URL http://dx.doi.org/10.1016/S1573-4412(05)80005-4. 495, 712

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *ECONOMETRICA*, pages 219–255, 2003. 760

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703, may 1987. doi: 10.2307/1913610. URL http://dx.doi.org/10.2307/1913610. 756

A. Pakes and D. Pollard. Simulation and the asymptotics of optimization esti-

mators. *Econometrica*, 57(5):1027–1057, 1989. doi: 10.2307/1913622. URL http://dx.doi.org/10.2307/1913622. 1140, 1160

Frank Smets and Rafael Wouters. Shocks and frictions in US business cycles: a Bayesian DSGE approach. *American Economic Review*, 97(3):586–606, jun 2007. doi: 10.1257/aer.97.3.586. URL http://dx.doi.org/10.1257/aer.97.3.586. 579

Anthony A Smith. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8(S1), 1993. URL https://doi.org/10.1002/jae.3950080506. 1141

James H Stock and Mark W Watson. *Introduction to Econometrics, 3rd International edition edition, Ed.* Pearson/Education, 2011. 879

George Tauchen. Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data. *Journal of*

*Business & Economic Statistics*, 4(4):397, oct 1986. doi: 10.2307/1391493. URL http://dx.doi.org/10.2307/1391493. 826, 837, 1174

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980a. 757

Halbert White. Using least squares to approximate unknown regression functions. *International Economic Review*, pages 149–170, 1980b. URL http://www.jstor.org/stable/2526245. 535, 1043

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526. 618, 642

# Index

ARCH, 1100
asymptotic equality, 1195

Cobb-Douglas model, 64
conditional heteroscedasticity, 1100
convergence, almost sure, 1189
convergence, in distribution, 1189
convergence, in probability, 1188
Convergence, ordinary, 1186
convergence, pointwise, 1186
convergence, uniform, 1187
convergence, uniform almost sure, 1191

estimator, linear, 76, 94
estimator, OLS, 66
extremum estimator, 498

fitted values, 68

GARCH, 1100

leptokurtosis, 1099
leverage, 77
likelihood function, 584

matrix, idempotent, 75
matrix, projection, 73

matrix, symmetric, 75

observations, influential, 76
outliers, 76
own influence, 78

parameter space, 584

R- squared, uncentered, 82
R-squared, centered, 83
residuals, 68