Econometrics Navigator Documentation

Release 0.0.5

Evgeny Pogrebnyak

Nov 10, 2019

CONTENTS

1 Analysis of variance (ANOVA) 1

2 Bias-variance tradeoff 3

3 Bootstrap 5

4 Causation, causality 7

5 Central limit theorem, CLT 9

6 Maximum likelihood 11

7 Mode 15

8 Ordinary least squares, OLS 17

9 Principal components analysis, PCA 19

10 Simulation 21

11 2. Textbooks and courses 23

12 3. The science of teaching 29

13 4. History of econometrics 33

14 Data 35

15 Software 37

16 Blogs 39

17 Good clues from Twitter 41

18 Acronyms 45

19 Roadmap 47

20 Welcome to Econometrics Navigator! 49

CHAPTER

ONE

ANALYSIS OF VARIANCE (ANOVA)

- ANOVA can mean several things: actual decomposition of variance, comparing the group means or representation of regression results
- Boils down to a regression with dummy (categorical) variables
- Heavy traction in terminolgy from design of exepriments (see definitons section here)
- Standartised result tables with `SS`, `DF`, `MSS`, `F`, `p`
- Frightening multitude of R packages
- May want to look at a simple reference case

Quote:

> *ANOVA can be seen as "syntactic sugar" for a special subgroup of linear regression models. ANOVA is regularly used by researchers who are not statisticians by training. They are now "institutionalized" and its hard to convert them back to using the more general representation* suncoolsu

1.1 Code examples

- https://stackoverflow.com/questions/25537399/anova-in-python-using-pandas-dataframe-with-statsmodels-or-scipy
- http://www.statsmodels.org/devel/anova.html
- https://stats.stackexchange.com/a/175265/211794
- also possibly in Think Stats and Hadley Wickham

1.2 Links

Intro by Jim

Cross-Validated has several general discussions:

- why-is-anova-taught-used-as-if-it-is-a-different-research-methodology-compared
- how-to-visualize-what-anova-does
- how-to-interpret-f-and-p-value-in-anova
- good-resource-to-understand-anova-and-ancova

. . . followed by ANOVA vs regression:

- difference-between-regression-analysis-and-analysis-of-variance

- why-is-anova-equivalent-to-linear-regression

NIST Handbook deals with ANOVA assumptions and interepations, as well as provides reference datasets:

- The one-way ANOVA model and assumptions
- Interpretation of the ANOVA table
- Reference datasets and regresion results

Very simple and illustrative case NIST reference case here.

Comoact Julia package ANOVA.jl at about 150 lines of code, but not as much documentation yet.

ANOVA is again a case where Russian wikipedia is more concise and clear on the subject.

'Types' of sum of squares and associated confusion:

- https://mcfromnz.wordpress.com/2011/03/02/anova-type-iiiiii-ss-explained/
- https://rcompanion.org/rcompanion/d_04.html
- https://stats.stackexchange.com/questions/20452/how-to-interpret-type-i-type-ii-and-type-iii-anova-and-manova

1.3 References

Gelman, A. (2005). Analysis of variance: why it is more important than ever (with discussion). Annals of Statistics 33, 1–53. doi:10.1214/009053604000001048

CHAPTER

TWO

BIAS-VARIANCE TRADEOFF

The bias variance trade-off is maybe not an ideal name, it should maybe have better been called interpolation/extrapolation trade-off. Anyway, the motivation for the name is that that when adding more parameters / complexity, you have

- Less systematic error (bias) in your model (supposedly, because it is more flexible, I would argue it depends on what you call error / bias)
- More variance in the estimation of the model parameters (because it is more flexible)

Florian Hartig at Cross Validated. What is the difference between bias and residuals?

CHAPTER

THREE

BOOTSTRAP

Bootstrapping is a method to construct empiric distributions of various statistics (mean, confidence intervals, deviation, etc) by using repreated sampling from an observed dataset.

A little magic is why exactly taking random samples like `[1,1,2]`, `[3,2,2]`, `[2,1,3]`, etc is a good idea to approximate properties of a dataset `[1,2,3]`.

Bootstrap originally proposed by Bradley Efron in 1979. For formal introduction see Horowitz chapter in Handbook of Econometrics (2001) and a usage overview by MacKinnon 2006.

3.1 Toy example

Bootstrap confidence intervals by Jeremy Orloff and Jonathan Bloom, pp. 4-6 provides the following basic code example for bootstrap. Their full code for this excercise is here.

```
# Bootstrap
# Adapted from https://math.mit.edu/~dav/05.dir/class24-empiricalbootstrap.r
cat("Example. Empirical boostrap confidence interval for the mean.",'\n')
x = c(30,37,36,43,42,43,43,46,41,42)
n = length(x)
set.seed(1)  # for repeatability

# sample mean
xbar = mean(x)
cat("data mean = ",xbar,'\n')
nboot = 20
# Generate 20 bootstrap samples, i.e. an n x 20 array of
# random resamples from x.
tmpdata = sample(x,n*nboot, replace=TRUE)
bootstrapsample = matrix(tmpdata, nrow=n, ncol=nboot)

# Compute the means xbar*
xbarstar = colMeans(bootstrapsample)

# Compute delta* for each bootstrap sample
deltastar = xbarstar - xbar

# Find the 0.1 and 0.9 quantiles for deltastar
d = quantile(deltastar,c(0.1,0.9))

# Calculate the 80\% confidence interval for the mean.
ci = xbar - c(d[2],d[1])
cat('Bootstrap confidence interval: [',ci,']','\n')
```

3.1.1 Bootstrap do's and don'ts by Anna Mikusheva

- If you have a pivotal statistic, bootstrap can give a refinement. So, if you have choice of statistics, bootstrap a pivotal one.
- Bootstrap may fix a finite-sample bias, but cannot help if you have inconsistent estimator.
- In general, if something does not work with traditional asymptotics, the bootstrap cannot fix your problem. For example, if we have an inconsistent estimate, the bootstrap bias correction does not fix anything.
- Bootstrap could not fix the following problems: weak instruments, parameter on a boundary, unit root, persistent regressors.
- Bootstrap requires re-centering (the null hypothesis to be true).

Source: MIT lecture notes

3.2 More links (preliminary)

- https://core.ac.uk/download/pdf/6494253.pdf
- https://github.com/wmutschl/GMMIndirectInferenceBootstrap
- https://www.schmidheiny.name/teaching/bootstrap2up.pdf
- http://rosetta.ahmedmoustafa.io/bootstrap/
- http://www.cs.cornell.edu/courses/cs1380/2018sp/textbook/chapters/11/2/bootstrap.html
- http://economics.fundamentalfinance.com/bootstrap.php

3.3 Editor notes

- Russian article in Wikipedia on bootstrap is much more concise and understandable than English one.

CHAPTER

FOUR

CAUSATION, CAUSALITY

Important: Correlation is not causation.

- causality (not 'casuality')

4.1 Book of Why by Judea Pearl

4.2 History

- Pearl, J. (2014). TRYGVE HAAVELMO AND THE EMERGENCE OF CAUSAL CALCULUS. Econometric Theory, 31(1), 152–179. https://doi.org/10.1017/s0266466614000231

CHAPTER

FIVE

CENTRAL LIMIT THEOREM, CLT

CHAPTER

SIX

MAXIMUM LIKELIHOOD

The probability density function `p = f(x, )` tells you a probability of occurrence of a random value near `x`. Likelihood is essentially a reverse operation of estimating unknown paramter from the same equation using `p` and `x`.

6.1 Lead by example

- Observations
- Probability of observations
- Observed sample is considered the most likely one
- Maximisation of probability allows to compute distribution parameters

6.2 Generalisation

We usualy denote a set of parameters like and as , a vector of parameters. Our task is to estimate parameter given:

- a sample of observations of random variable $X = (x_1, x_2, ..., x)$, and
- a pre-defined probability density function `f(x, )`.

Solution steps:

1. collect observations $X = (x_1, x_2, ..., x)$
2. make a decision which probability density function `f(x, )` is appropriate for this data
3. compose joint probability of observations as a function of : $L() = f(x_1,) \cdot f(x_2,) \cdot ... \cdot f(x,)$.
4. Come to terms with a principle “if we observed this event, we consider it was the most probable outcome of all possible events in this distribution”
5. Find which maiximises joint probability of observations

6.3 Code

Python code below below relies on `scipy.optimixe.minimize` solver to find parameters of a normal distribution based on two measurements of mice weights. It can be easily applied to more observations.

```
"""
Maximum likelihood with two mice.
"""

import numpy as np
from scipy.optimize import minimize

def dnorm(x, mu=0, sigma=1):
    """Normal distribution probability density fucntion."""
    const = 1 / (sigma * np.sqrt(2 * np.pi))
    power = - (x - mu)**2 / (2 * sigma**2)
    return  const * np.exp(power)

def log_likelihood(observed_x):
    """Sum of logs of probability densities at *observed_x*.
    Return:
       function of mu and sigma
    """
    def foo(mu, sigma):
        logs = [np.log(dnorm(x, mu, sigma)) for x in observed_x]
        return sum(logs)
    return foo

def maximise(f, start_mu, start_sigma):
    """Return mu and lambda, which maximise *f*."""
    f = lambda p: -1 * l_func(mu=p[0], sigma=p[1])
    res = minimize(f, x0=[start_mu, start_sigma])
    return res.x[0], res.x[1]

# two mice weigths are given, similar to https://www.ncbi.nlm.nih.gov/pmc/articles/
↪PMC6143748/
events = [30, 50]
# construct likelihood as a function of unknown mu and sigma
l_func = log_likelihood(events)
# run maximisation procedure
# attention: need a sensible pick for start variables, eg (0, 1) will fail
estimated_mu, estimated_sd = maximise(l_func, start_mu=30, start_sigma=3)

# test outcomes
#estimated_mu is 39.99999527669165
assert np.isclose(estimated_mu, 40)
#estimated_sd is 9.999976480910071
assert np.isclose(estimated_sd, 10)

# Result: observed values [30, 50] were most likely coming from
#         normal distribution with parameters =40 and =10.
```

Other code examples:

- Annotated R code by Andrew Collier (2013)
- Doing Maximum Likelihood Estimation by Hand in R by John Myles White (2010)
- Julia vs R vs Python Simple Optimization by Zhuo Jiadai (2018)

6.4 Links

- Nice video with weight of mice
- Maximum likelihood estimation in layman terms
- Why is maximum likelihood estimation considered to be a frequentist technique
- Very accessible math treatment (in Russian)
- Tourist sees a fountain (also in Russian)

CHAPTER

SEVEN

MODE

CHAPTER

EIGHT

ORDINARY LEAST SQUARES, OLS

OLS is at the core of econometrics curriculum, it is easily derived and highly practical to familiarise a learner with regression possibilites and limitations.

The usual way to teach OLS is to present assumptions and show how to deal with their violations as indicated below in a review chart from Kennedy’s textbook.

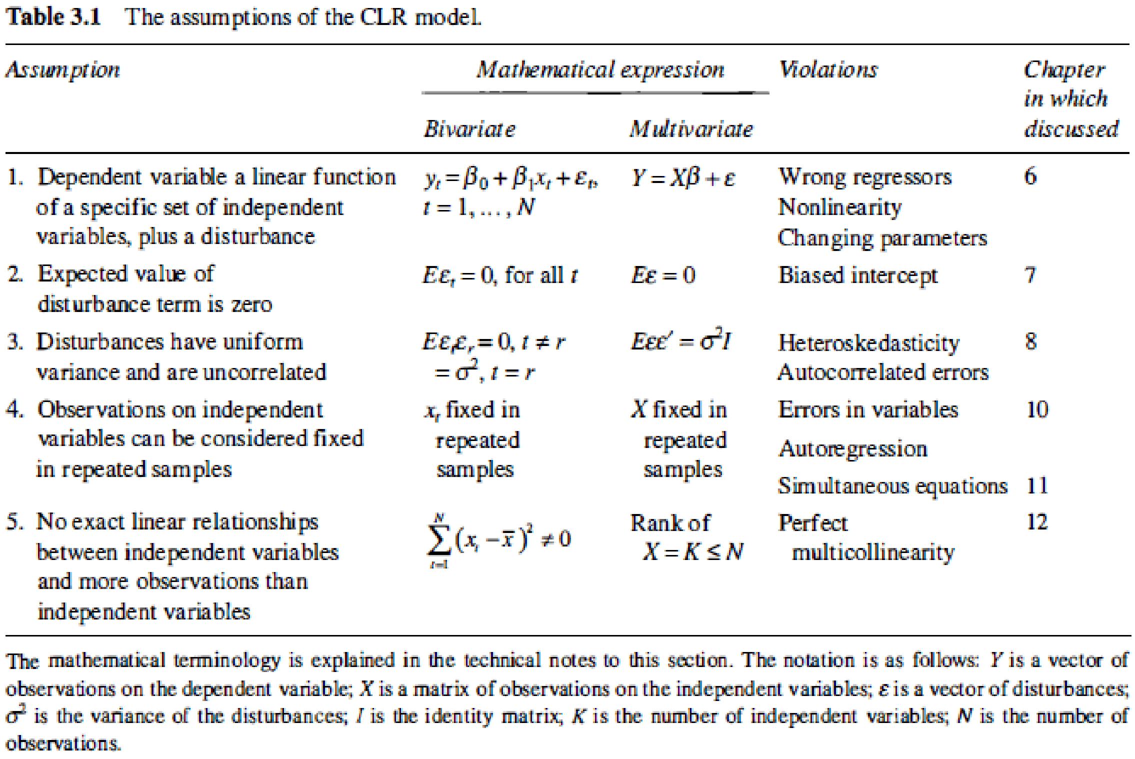


Table 3.1 The assumptions of the CLR model.

Assumption	*Mathematical expression*		*Violations*	*Chapter in which discussed*
	Bivariate	*Multivariate*		
1. Dependent variable a linear function of a specific set of independent variables, plus a disturbance	$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, $t = 1, \ldots, N$	$Y = X\beta + \varepsilon$	Wrong regressors Nonlinearity Changing parameters	6
2. Expected value of disturbance term is zero	$E\varepsilon_t = 0$, for all t	$E\varepsilon = 0$	Biased intercept	7
3. Disturbances have uniform variance and are uncorrelated	$E\varepsilon_t\varepsilon_r = 0, t \neq r$ $= \sigma^2, t = r$	$E\varepsilon\varepsilon' = \sigma^2 I$	Heteroskedasticity Autocorrelated errors	8
4. Observations on independent variables can be considered fixed in repeated samples	x_t fixed in repeated samples	X fixed in repeated samples	Errors in variables Autoregression Simultaneous equations	10 11
5. No exact linear relationships between independent variables and more observations than independent variables	$\sum_{t=1}^{N}(x_t - \bar{x})^2 \neq 0$	Rank of $X = K \leq N$	Perfect multicollinearity	12

The mathematical terminology is explained in the technical notes to this section. The notation is as follows: Y is a vector of observations on the dependent variable; X is a matrix of observations on the independent variables; ε is a vector of disturbances; σ^2 is the variance of the disturbances; I is the identity matrix; K is the number of independent variables; N is the number of observations.

Math:

$Y = \beta X + \epsilon$, ϵ is iid, normal with finite variance.

Common steps:

1. specify model: select explanatory variables, transform them if needed
2. estimate coefficients
3. elaborate on model quality (the hardest part)

4. go to 1 if needed
5. know what model *does not* show (next hardeer part)

What may go wrong:

- residuals are not random
- variables are cointegrated
- multicollinearity in regressors
- residuals depend on x (heteroscedasticity)
- inference is not causality
- wrong signs, insignificant coefficients
- variable normalisation was not described

Discussion:

- why sum of squares as a loss function?
- connections to bayesian estimation
- is R2 useful or dangerous?

Implementations:

- lm function in R
- OLS class in python statsmodels
- python scypi least squares
- julia Alistair, GLM.jl, Regression.jl
- Replication examples
- check unsorted links about OLS - but it is not better than googling on your own

CHAPTER

NINE

PRINCIPAL COMPONENTS ANALYSIS, PCA

Math:

Assumptions:

Usual steps:

What may go wrong:

Discussion:

Replication examples:

Links:

- https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues/2700#2700
- https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_3d.html#sphx-glr-auto-examples-decomposition-plot-pca-3d-py
- https://stats.stackexchange.com/questions/48214/replicating-shalizis-new-york-times-pca-example?rq=1

Projection, rejection, PCA:

- https://stackoverflow.com/questions/52288029/function-that-computes-projection-and-recostruction-error-using-numpy-python/52290082#52290082
- http://www.cs.cmu.edu/~guestrin/Class/15781/slides/pca-mdps-annotated.pdf
- https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/least-squares-determinants-and-eigenvalues/projections-onto-subspaces/MIT18_06SCF11_Ses2.2sum.pdf

CHAPTER TEN

SIMULATION

> A very useful tool is simulation – with that we can examine the properties of our tools in situations very like those it appears our data may have arisen from, and so either use them in the comforting knowledge that they have good properties in those cases (or, sometimes, see that they don't work as well as we might hope).

From an answer to Why do we care so much about normally distributed error terms by Glen_b.

To cover next: Concepts

- indentification
- inference
- overfitting
- spurious regression

To cover next: Theorems

- Bayes theorem
- Gauss-Markov theorem

To cover next: Research design

- replication, replicability
- complexity bias
- publishing and promotion

CHAPTER

ELEVEN

2. TEXTBOOKS AND COURSES

11.1 Mathematic preliminaries

Typical prerequisites for statistics and econometrics are:

- linear algebra
- calculus
- probability

They usually take 2-4 semester in college. Linear algebra is fully covered by VMLS, and probability is exposed in PSC, even though it is a compact reference, not formally a textbook. Scipy lectures are a great one-stop resource for numerical computing basics.

I do not have a one single source to recommend for calculus yet.

11.1.1 Linear Algebra

- Vectors, Matrices, and Least Squares (VLMS)
- You can also check Computational Linear Algebra repository from *fast.ai* here.

11.1.2 Calculus

The Matrix Calculus You Need For Deep Learning recommends Khan Academy differential calculus course, but it is not a single downloadable reference. *fast.ai* also has a calculus intro, going from start to derivatives in deep learning quickly.

11.1.3 Probability and statistics

- Probability and Statistics Cookbook (PSC)

11.1.4 Numerical computing

- Scipy lectures: one document to learn numerics, science, and data with Python

11.2 Statistical inference

This page is a draft.

Some discussions about a good book:

- Math Overflow: Statistics for Mathematicians
- Stack Exchange: What to learn after Casella/Berger

11.2.1 Topics

Review of probability:

- random variables, outcomes, probability
- distributions, pdf/cdf

Mathematic statistics itself:

- idea: learn about data generating process (DGP) from a sample
- sample/realisation as a random vector
- statistic as a function of sample
- parameter inference:
  - point estimation and methods (estimators)
  - estimator quality (bias, consistency, efficiency)
- confidence intervals (CI)
- hypothesis testing (HT), types of errors
- analysis of variance and regression

See an example of stated learning objectives

11.2.2 Key textbook

Casella, Berger. Statistical Inference.

- link

11.2.3 Other choices

Larry Wasserman. All of Statistics: A Concise Course in Statistical Inference

```
- [text 1](https://www.ic.unicamp.br/~wainer/cursos/1s2013/ml/livro.pdf)
- [text 2](http://static.stevereads.com/papers_to_read/all_of_statistics.pdf)
- [Amazon](https://www.amazon.com/All-Statistics-Statistical-Inference-Springer/dp/
↪0387402721)
```

Efron, Hastie. Computer Age Statistical Inference: Algorithms, Evidence and Data Science.

```
- https://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf
```

11.2.4 Supplements

James Gentle. A theory of statistics.

De Groot and Shervish%20(1).pdf)

Herman J. Bierens. Introduction to the Mathematical and Statistical Foundations of Econometrics

Aris Spanos. Probability Theory and Statistical Inference: Econometric Modeling with Observational Data.

```
- [link](https://wilfridomtz.files.wordpress.com/2014/08/cambridge-university-press-
↪probability-theory-and-statistical-inference-842pg.pdf)
```

Probability and Statistics for Engineers and Scientists (PSES)

11.2.5 Other reading

Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars

```
- [Amazon](https://www.amazon.com/Statistical-Inference-Severe-Testing-Statistics/dp/
↪1107664640)
- very philosophic reading
```

Peter M. Aronow, Benjamin T. Miller. Foundations of Agnostic Statistics.

```
- [Google Book (fragments)](https://books.google.ru/books?id=u1N-DwAAQBAJ&
↪printsec=frontcover&hl=ru&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false)
- [book proposal](http://aronow.research.yale.edu/aronowmillerproposal.pdf)
- simplictic
```

11.2.6 Papers

Statistical Inference: The Big Picture. Robert E. Kass (2006)

A short history of probability theory and its applications. International Journal of Mathematical Education. January 2015. Kanadpriya Basu.

11.2.7 In Russian

Cramer, 1946, Russian translation, timeless! also has dataset if wheat yeilds.

Also concise version of lectures here (in Russian)

11.3 Econometrics

11.3.1 General textbooks

- Thread about picking a textbook
- Favorite introductory book: Peter Kennedy. A Guide To Econometrics
- Very condensed synopsis of undergrad econometrics

Some well-known texts:

- Greene
- Dougherty
- Hayashi
- Stock and Watson
- Verbeek
- Hansen. Econometircs
- Sargent-Hansen

11.3.2 Cross Section and Panel Data

- Wooldridge. Econometric Analysis of Cross Section and Panel Data
- , with criticisms by:
  - Diebold: "sub-sub-sub-area of applied econometrics"
  - J. Pearl

Note that MHE is one of few books with excercises reproduced on github.

Criticism of MHE:

> Rather, it's a companion for a highly-specialized group of applied non-structural micro-econometricians hoping to estimate causal effects using non-experimental data and largely-static, linear, regression-based methods. It's a novel treatment of that sub-sub-sub-area of applied econometrics, but pretending to be anything more is most definitely harmful, particularly to students, who have no way to recognize the charade as a charade.

https://fxdiebold.blogspot.com/2015/01/mostly-harmless-econometrics.html

- Lecture Notes in Microeconometrics - nicely presented topics in econometrics for microeconomics from OLS to Bootstrap. Guides to software.

11.3.3 Time Series

- Hamilton
- Diebold
- Granger/Watson time series chapter in Handbook of Econometrics
- Enders
- Jeffrey Parker. Fundamental Concepts of Time-Series Econometrics in *Theory and Practice of Econometrics. Reed College course*

Macroeconomic time series:

- John H. Cochrane. Time Series for Macroeconomics and Finance.
- Eric Sims. Graduate Macro Theory II: Notes on Time Series
- Lars Peter Hansen. Time Series Econometrics in Macroeconomics and Finance
- Hilde C. Bjørnland and Leif Anders Thorsrud. Applied time series for macroeconomics. (not open source, but good table of contents and has MATLAB code)
- Karl Whelan. Time Series and Macroeconomics. (also see RATS code and other lectures on the web site)

11.3.4 ayesian Methods

- Bayesian Methods for Hackers has programming-first approach.
- SciPy 2019 Lecture on Bayesian Model Evaluation and Criticism.
- A simple explanation of Naive Bayes Classification, an overwhelmingly popular StackOverflow answer.

11.4 . . . and its history

11.4.1 Landmark events

- Galton and Pearson
- Tinbergen, Haavelmo
- Cowles Commission
- Lucas and Sims critique

11.4.2 Overviews

- Econometrics: A Bird's Eye View
- Econometrics: An Historical Guide for the Uninitiated
- The First Fifty Years of Modern Econometrics

11.4.3 By topic

- Pearl, J. (2014). Tygve Haavelmo and the emergence of causal calculus. Econometric Theory, 31(1), 152–179.
- A Short History of Markov Chain Monte Carlo (arxiv)
- Working on 1960s macroeconometrics (blog)
- Criticizing the Lucas Critique: Macroeconometricians' Response to Robert Lucas
- W N Venables. Exegeses on Linear Models (1998)

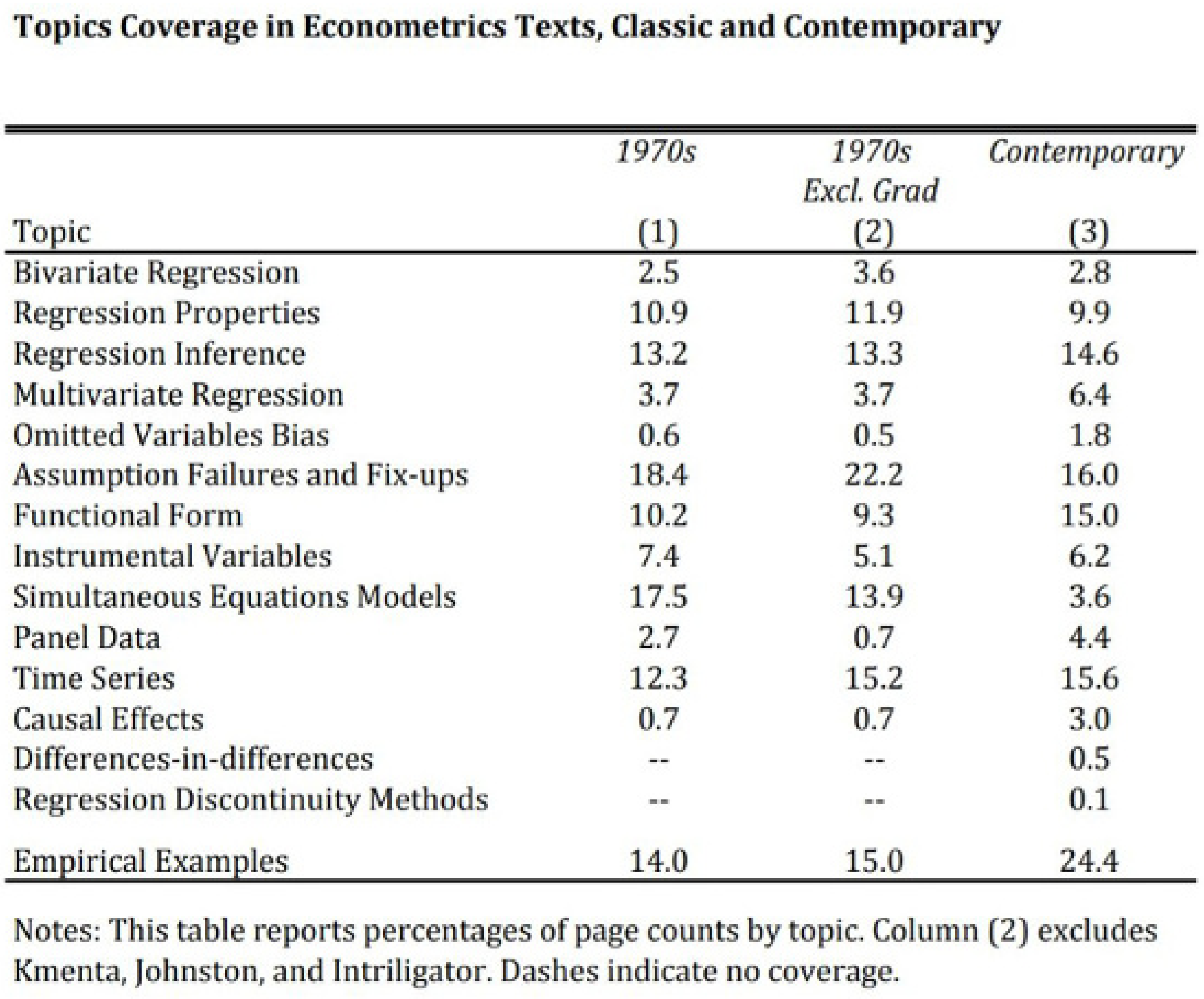
11.5 . . . body of knowledge

What exactly is a body of knowledge of econometrics? Surely it is a set of teaching curricula at universities and accompaigning textbooks and reading lists, but what if you needed construct a review of the field rather quickly? The following approaches might be useful.

11.5.1 1. Textbook structure

The undergraduate textbook structure often repeats itself: OLS, estimator properties, some of OLS deviations/extensions, logit/probit + maximum likelihood, time series and maybe a bit of panels and simultaneous equations.

More textbook analysis may be found in 2017 Angrist and Pischke article Undergraduate Econometrics Instruction: Through Our Classes, Darkly, below is a summary table (some of it does not escape criticism).



Topics Coverage in Econometrics Texts, Classic and Contemporary

Topic	*1970s* (1)	*1970s* *Excl. Grad* (2)	*Contemporary* (3)
Bivariate Regression	2.5	3.6	2.8
Regression Properties	10.9	11.9	9.9
Regression Inference	13.2	13.3	14.6
Multivariate Regression	3.7	3.7	6.4
Omitted Variables Bias	0.6	0.5	1.8
Assumption Failures and Fix-ups	18.4	22.2	16.0
Functional Form	10.2	9.3	15.0
Instrumental Variables	7.4	5.1	6.2
Simultaneous Equations Models	17.5	13.9	3.6
Panel Data	2.7	0.7	4.4
Time Series	12.3	15.2	15.6
Causal Effects	0.7	0.7	3.0
Differences-in-differences	--	--	0.5
Regression Discontinuity Methods	--	--	0.1
Empirical Examples	14.0	15.0	24.4

Notes: This table reports percentages of page counts by topic. Column (2) excludes Kmenta, Johnston, and Intriligator. Dashes indicate no coverage.

11.5.2 2. Econometric software manuals

Gretl and EViews have quite comprehensive manuals covering principal applications of the software. They both qualify as textbooks in econometrics:

- gretl
- course based on gretl
- Eviews I
- Eviews II

Additionally one can look into [R package system for econometrics] (https://cran.r-project.org/web/views/Econometrics.html), MATLAB manual and course.

Manulas of some less popular packages:

- RATS
- PcGive
- Shazam
- econtools (STATA flavour)

11.5.3 3. Handbook of Econometrics

Elsevier Handbook of Econometrics is a publication series running since 1983. It now features 77 chapters in 6 volumes. Many earlier articles are foundational, but quite a few recent ones are about some really narrow subjects areas. I think the volume TOC is great, but publications are overpriced (it's Elsevier).

11.6 . . . mindmap

It would be great to show a modern roadmap in econometrics starting from mathematic foundations (linear algebra, calculus, probability) to econometrics to computationally intensive data processing tasks. I've seen this being approached as clusters of courses, Khan Academy has goals by subject, but I think there is more that can be done.

11.6.1 An (over)simplified view of econometrics curriculum

- linear algebra, calculus, probability and statistical inference
- OLS (assumptions, violations, fixes + estimatore quality)
- limited depenedent variables + maximum likelihood
- intrumental variables
- time series, state space representation
- panel data
- classifications
- systems of equations

11.6.2 Key areas

- data structures (crosssection, time series, panel)
- inference methods
  - model specification
  - estimation procedure
  - model evaluation
- use cases

11.6.3 Additional topics

- simulation (Monte Carlo, bootstrap)
- transformations (PCA)

OLS Extensions:

- GMM
- 2,3 stage OLS
- quantile regressions
- lasso, rigde

Estimation:

- maximum likelihood
- bayesian estimation
- mcmc (see reddit post)

Time series:

- time series, stationarity, unit root
- state space representation, Kalman filter
- fractional integration
- seasonal adjustment
- (vector) error correction model, VECM
- structural breaks

11.6.4 User profiles

1. "Numerate biologists" - solve a domain problem in biology, psychology, social sciences
2. "Want to hit a 'Run' button" - quick results without thinking, typical of students
3. "I'm doing XYZ now!" - excited adopters, writing a piece on Medium full of acronyms
4. "Sane econometrics" - appropriate methods with clear, accessible explaination, rare trait
5. "Asymptotics" - publish evermore sophisticated articles to secure academic career

11.6.5 Discussion

Undergraduate Econometrics Instruction: Through Our Classes, Darkly. NBER/IZA and a criticism of G1/G2 goals

11.7 Machine learning (ML) and deep learning (DL)

11.7.1 Books

ML

- An Introduction to Statistical Learning. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

DL

- Deep Learning. Ian Goodfellow and Yoshua Bengio and Aaron Courville

Slightly overcomplicated extras

- Foundations of Data Science
- Understanding Machine Learning: From Theory to Algorithms

11.7.2 Courses

ML

- ods
- Andrew Ng course
- QuantEcon ML application

DL

- fast.ai
- deeplearning.ai

CHAPTER

TWELVE

3. THE SCIENCE OF TEACHING

Notes on technical pedagogy.

12.1 Ideas

- "Overarching education" (plenty of math and fundimentals, aimed at systemic thinking, practical applications derived later) is expensive method of teaching.
- Student motivation is scarce, information is abundant. Teaching is a guidance (esp masters level).
- Can teach programming first, and followup with more solid math second, if ever (programming allows experimenting).
- Need open source textbooks, able to update and share parts of text as well as interactive excercises. Static site generators are not fully there yet.

12.2 Links

1. 10 rules of teaching by Greg Wilson (@gvwilson) starts with rule #1 "Be kind: all else is details." More detail is at teachtogether.tech.
2. Nick Huntington-Klein proposes a course structure based on statistical programming, and causal inference/research design, with regressions postponed.
3. Allen Downey has a presentation about teaching physical modelling and sequencing of math and programming.

4. Very introductory courses are good for building student confidence and making simple things simple. They prevent gate-keeping (maybe a reason why they are attacked). See a thread by Rochelle Terman on teaching computational social science.

CHAPTER

THIRTEEN

DATA

13.1 Large collections

- Python libraries statsmodels, sci-learn and seaborn and R itself have build-in datasets.
- gretl has extensive selection of datasets, including data from key textbooks.
- Well known dataset repository is UCI.
- Kaggle obviously has plenty of datasets.
- Sheffield University made a good listing of datasets for teaching.
- Australian National Centre for Econometric Research website has a useful Data and code section.
- Google Dataset search is not yet as good as main Google search, but will surely improve.
- FRED, Quandl and dbnomics are standard sources for macroeconomic data. My own effort for clean Russian macro time series - mini-kep.

13.2 Individual datasets

- Tidy Tuesday publishes weekly datasets and accompanying articles.
- Nick Huntington-Klein provides a variety of extra examples of data importable into R.
- I extracted a small, but illustartive dataset about lightbulb survival rate.
- Sunspot data is listed at variety of sources.

CHAPTER

FOURTEEN

SOFTWARE

Below is a small chart that outlines common options for statistics / econometrics software. gretl, EViews and MATLAB tend to have better-organised documentation. My personal choices:

- if I had to choose just one: python;
- if I had to choose other two: R (libraries) and gretl (documentation);
- before open source era: Eviews;
- if I had more time: Julia.

Open source	Proprietary
• R (RStudio), derived from S • Python (Anaconda) • Julia	
• gretl	• EViews (derived from TSP)
• Octave	• MATLAB
	• SAS • SPSS • Stata
	RATS, Ox, PcGive
Stan, PyMC3, Turing.jl	
JASP	

14.1 SAS and terminology

> SAS, it seems, has become the gold standard, the output of SAS programs the ultimate point of reference for correct and appropriate statistical calculations and the SAS terminology israpidly taking over as standard terminology. This is very Microsoft-like indeed and very worrying for anyone who cares about the profession.

WN Venables. Exegeses on Linear Models - on early adoption of S Plus

14.2 Example: software used in black hole discovery

Reading *First M87 Event Horizon Telescope Results. III. Data Processing and Calibration*:

CHAPTER

FIFTEEN

GOOD CLUES FROM TWITTER

A secret subtitle for this publication is *"Can you learn econometrics from Twitter and Stack Overflow alone without distracting yourself to data science tutorials"*.

Links collcted in no particular order, some will show in other sections of the Navigator.

15.1 No links between leading macrotextbooks

15.2 OLS interactively exposed

15.3 R language guide is an econometrics guide

15.4 OLS, MML, Bayes and MCMC(!) for linear regression

https://peterroelants.github.io/posts/linear-regression-four-ways/

15.5 Very true on R2

15.6 Instrumental variables

15.7 Interpreting coefficients 1

15.8 Interpreting coefficients 2

15.9 Doing PCA approach

15.10 OLS explained for social scientists

15.11 Suggested stat excercises

15.12 Coding a tree

15.13 Consumer demand modelling

15.14 Scott Cameron learning method

15.15 Model evaluation compendium (on classifier)

15.16 Causality by Judea Pearl

15.17 A thesis turned tutorial on probabilistic programming and MC inference by Tom Rainforth

15.18 Value of logit

15.19 Traditional statistics vs ML

15.20 Program evaluation by John Holbein

15.21 218 pages on probabilistic programming

15.22 A4 econometric art

15.23 A 1910 must-read

15.24 Amazing statistics animation

15.25 Undergrad econometrics condensed to 3 pages

CHAPTER

SIXTEEN

BLOGS

Blogs may seem as an outdated fashion of communication, but they often present a wider story than a twitter post.

Below there are some personal blogs that enlight and inspire about statistics and econometrics. They are regularly updated.

- Matt Bogard
- Francis Diebold
- Dave Giles
- Rob Hyndman

Simply Statistics by biostatistics professors Rafa Irizarry, Roger Peng, and Jeff Leek cover data management, analysis and teaching stats.

Medium has a variety of posts on statistics, but rarely tags econometrics. PCA is a widely popular topic.

CHAPTER

SEVENTEEN

ACRONYMS

Econometrics is full of fancy abbreviations that one can juggle with.

ANOVA analysis of variance

ARIMA autoregression, integration, moving average

DID difference-in-differences

FE feature engineering

GARCH generalised (a)uto(r)egressive conditional heteroscedasticity

GLS generalised least squares

GMM generalised method of moments

iid independent identicaly distributed

IV instrumental variable

OLS ordinary least squares

PCA principal components analysis

RDD regression discontinuity design

VECM vector error correction model

WLS weighted least squares

CHAPTER

EIGHTEEN

CHANGELOG

v.0.0.5 (November 2019):

- new TOC and flatter stucture
- generated a rough pdf
- minimised errors for sphinx builds
- tasks.py for invoke renewed

v.0.0.4 (May 2019):

- Science of teaching: quoting @gvwilson, @nickchk, @AllenDowney, @RochelleTerman at https://tinyurl.com/em-nav-teach
- Data: added data from @stlouisfed/@quandl/@DBnomics along with several R data sources by @nickchk at https://tinyurl.com/em-nav-da
- Books:
  - WM Venables. Exegeses on Linear Models
  - Walpole, Myers, Myers, Ye. Probability and Statistics for Engineers and Scientists

v.0.0.3 (April 2019) scraps several unfinished articles, including a section on applications (hard to fill it quickly). Three main parts in content established (own articles, textbook annotations and how to teach resources).

v.0.0.2 (November 2018) original version of EN nobody understood what it is good for, had sample articles on max likelihood, bootstrap, ANOVA.

18.1 Roadmap

18.2 November 09, 2019

- drafts for cases and excercises
- add more tweets
- add from twitter personalities - links to them
- Section 4 History may go somewhere else
- logit models, tweets about them
- add presentation about reproducibility
- Statistical inference section is still a draft

18.3 May 13, 2019

18.3.1 General

- draw a mindmap for econometrics (as described in text)
- put key textbooks on a roadmap
- pay more attention to bayesian / causality

18.3.2 Data science textbooks

- Data Science fo Economists

18.3.3 Articles and code

- write more articles and code for the main section
- translate some RATS/MATLAB code to open source (especially time series)

Specific tasks:

- clean ols
- run pca example

18.3.4 Reproducibility

Add resources on reproducible research and why it has such a poor traction in economics

- this - 1
- this - 2

DAG tools:

- waf.io
- invoke

18.3.5 How to publish a textbook

Need an opensource textbook with interactive code examples, translation into Russian desired. Hard to see a combination of a static site generator, good theme and PDf export working smoothly.

- Allen Downey on Medium
- jupyter-book
- bookdown

18.3.6 Our publishing process

- things mentioned in todo.txt

18.3.7 Other goals

- review ‘depreciated’ folder, and ‘history’ page
- ‘vednorize’ LessOLS.jl

CHAPTER

NINETEEN

WELCOME TO ECONOMETRICS NAVIGATOR!

19.1 Goals

The Econometrics Navigator (EN) goal is to make quality instruction in statictics and econometrics accessible.

Let's lower the barrier for entry and prevent gate-keeping!

19.2 Types of content

We aim to provide you with:

- open access textbooks and community knowledge (reddit, StackOverflow, Twitter threads),
- minimal code examples in Python, R, gretl or Julia,
- datasets and cases for quantative analysis,
- ideas on how to structure your learning paths.

19.2.1 1. Own articles

The articles are in Concepts and techniques section, organised alphabetically. Finished examples are:

- Maximum likelihood
- Bootstrap
- ANOVA

19.2.2 2. Textbooks guide

Textbooks review attempts to sort out and annotate textbooks and references by several categories, starting from math preliminaries and up to ML/DL applications.

In econometrics the categories are 'general' textbooks, cross-section/panel, time series and bayesian texts. Whatever I could not document well, I did put in the mindmap section.

The backstage workings of the Navigator are History of econometrics, Ways to review econometrics.

Review of resources about mathematic statistics are still a draft.

19.2.3 3. Better teaching

These documents organise thinking about better teaching of econometrics in terms of sequence of topics, better analogies for the learner and a faster bridge to coding and working with real data from formulas and concepts. Specifically I collected the links about technical pedagogy here.

19.3 Twitter

So far twitter has been an enormously valuable source of demos, links and opinion for me. I keep a separate page with twitter posts, some of my favorites are:

- Undergrad Econometrics Cheatsheet by Tyler Ransom
- Casual graphs and XY plane animations by Nick Huntington-Klein
- Investigative time series example by @cubic_logic
- Common statistical tests are linear models (or: how to teach stats) by Jonas Kristoffer Lindeløv

19.4 Changelog

Changelog and future steps outlined are outlined in roadmap.

19.5 Contacts

This publication is edited by @PogrebnyakE.

19.6 Source

The source of this publication is available at https://github.com/epogrebnyak/econometrics-navigator and the short URL for this page is https://tinyurl.com/emnavig.