Motivation for regression analysis

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

Before we start with more formal analysis….

- What is econometrics?
- Who is using it and for which purpose?
- Data structures
- The problem of causality

Econometrics is …

- estimating partial economic relationships
- testing economic theories
- evaluating and implementing government and business policies.

Examples:

- Evaluate the effectiveness of a publicly funded job training program
- Test different investment strategies of a bank to decide whether they comply with implied economic theory

- Formal economic modelling , e.g. a utility maximization framework, is often the starting point for empirical analysis.
- The economic model or our intuition provide us a mathematical relationship (equation).

Example:

Effect of training on the productivity of workers
(= higher wage)?

$$wage = f(educ, exper, training)$$

with *educ*=education, *exper*=experience
and a functional *f(.)*

- Turn the economic model into an econometric model:
  - Specify the functional *f(.).*
  - How to deal with unobserved variables?
  - Introduce parameters β of the econometric model.

Example (cont.): a complete econometric model might be

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u \ ,$$

where

- u contains all unobserved factors that can influence the person's wage. Examples?
- $f(.)$ is a linear functional
- there are four parameters $\beta_0, \beta_1, \beta_2, \beta_3$.

Formulate a hypothesis for the unknown parameters: $\beta_3 > 0$

Typical Data Structures

- A big variety of data structures and how data is generated (experiments, interviews, administrative purposes, business activity).
- In economics, data is typically nonexperimental, i.e. not collected in laboratory environments.
- In the following some common data structures are presented:
  - Cross section
  - Time series
  - Pooled cross section
  - Panel or longitudinal data

- Cross sectional data:
  - Consists of a sample of individuals, firms, states or a variety of other units.
  - At a given point in time.
  - Requirement: Random sample of the underlying population.

obsno	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

- Time series data:
  - Consists of observations on a variable or several variables over time.
  - Examples include stock prices, consumer price index and automobile sales figures.
  - Chronologial ordering of observations conveys potentially important information.

obsno	*year*	*avgmin*	*avgcov*	*unemp*	*gnp*
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

avgmin: average minimum wage, *avgcov*: average coverage rate

- Pooled cross section data
  - Has both cross sectional and time series features.
  - For example, several cross section with the same variables at different point of time are pooled to one data set.

- Panel or longitudinal data
  - Consists of a time series for each cross-sectional member in the data set.
  - This implies that one can follow each cross-sectional unit over time.

A Two Year Panel Data Set

obsno	*city*	*year*	*murders*	*population*	*unem*	*Police*
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Causality…

- In most cases one wants to identify whether one variable has a causal effect on another variable, such as
  - Education on worker productivity
  - Price on quantity demanded
  - Job training for unemployed on employability or wages

- Ceteris paribus means “all other relevant factors being equal”.
- Most economic problems are ceteris paribus by nature.
- If other factors are not held fixed, then we cannot identify the causal effect.
- In empirical work the key question is: have enough other factors been held fixed to make the case for causality?

- Example: **Effect of Fertilizer on Crop Yield.**

  - Keep in mind that many factors determine crop yield. Which?

  The following experiment is conducted:

  - Choose several one-acre plots of land.
  - Apply different amounts of fertilizer to each plot and measure the yields.
    - This produces what kind of data?
  - Measure the association between yields and fertilizer amounts by statistical methods.

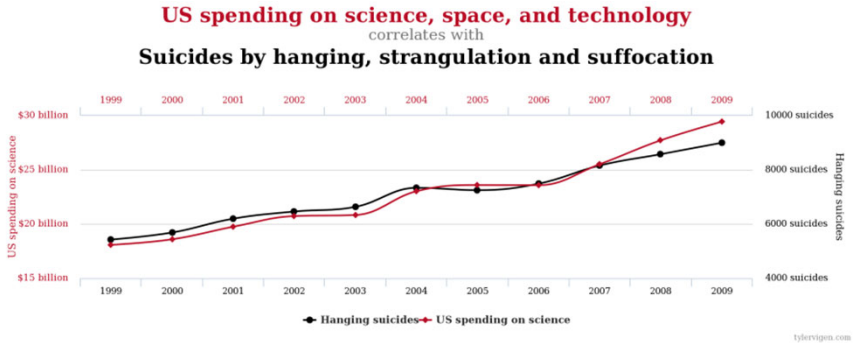
  - Why this may not be a good experiment?
    - Unclear how plots of land are chosen.
    - Other important factors are not observed.

  - When is this experiment useful to measure the causal effect?
    - If the levels of fertilizer are assigned to plots independently of other plots characteristics

Causality vs. correlation

- What can multivariate regression analysis add over competing approaches?

- When comparing only two variables, unrealistic bivariate correlations may be observed:
  - For example time series data: spurious correlations
  - Good resource: http://www.tylervigen.com/spurious-correlations



  - Solution: panel data analysis (if data available).

- Classification analysis is very common in business analytic.
  - May be based on high dimensional data structures.
  - Data fitting (maximise correlation between outcome and a number of predictors).
  - Does not reveal causation.
  - Classification may lead to discrimination and undesired inequalities based on spurious data artefacts. (Reference: Math Panic, Significance, 2016, Bursting Big Data Bubbles, 2017)

  - Solution: Define economic hypotheses and an economic model. Focus on consistent estimation of partial effects.

- By determining partial relationships, it is possible to dive deeper into the puzzle.

Econometrics and AI

- Econometrics is machine learning and therefore AI.
- Modern machine learning algorithm are increasingly used within classical econometric models to improve the accuracy of the fit of a model or to select model features such as variables.
- LLM can be used for code writing, interpretation of results (regression outputs, plots).
  - For example, the R-package **tidyllm** provides R an interface for interacting with the most common LLMs.

- As usual with AI tools, there is probably something right and something that is incorrect.
  - Code may not call the most appropriate methods, and it can be inefficient.
  - Result interpretation is superficial and does not inform about risks of the analysis (violation of model assumptions).

- To be able to assess the quality of these outputs, it is important to know the methods, whether assumptions hold and how to interpret the results.
  - This is what we do in this course.

Summary

The purpose and scope of econometric analysis:

- Used in all applied economic fields to test economic theories.
- Different data structures (cross section, time series).
- The notions of ceteris paribus and causal inference.
- While most hypothesis in the social sciences are ceteris paribus in nature, the nonexperimental nature of most data collected makes the estimation of causal relationships very challenging.

BA-BMECV2502U Econometrics

Ralf A. Wilke

rw.eco@cbs.dk

Kursuskatalog

https://kursuskatalog.cbs.dk/2025-2026/BA-BMECV2502U.aspx

- Week plan on Canvas.

Module aims

The main aims of this module are:

- to provide sufficient background in modern econometric methods to understand and implement appropriate econometric modelling techniques suited both to the data source and to the economic model;
- to understand how appropriate model specification choices can create a link between an economic problem and econometric estimation;
- to deepen the understanding of statistical properties of estimators and to acquire competencies with the related mathematical tools.
- to develop skills in the interpretation and critical appraisal of econometric estimates;
- to develop skills in applied economic work, exploiting the availability of computer techniques and packages for model solution.

Learning outcomes

- Understand econometric estimation and inference methods for higher dimensional data.
- Understand how to model, estimate and interpret the partial or causal relationship between two variables in models with many variables.
- Understand the relevance of assumptions on the econometric model for the properties of estimation and inference results.
- Appropriately choose an econometric model from those introduced in the course and assess its suitability.
- Understand estimation results and interpret them.
- Relate R-code and R-output to the econometric models introduced in the course.
- Conduct econometric analysis in R.

Course content

Lecture programme

- Part 1: Multiple Regressions Model

  Mid-term exam

- Part 2: Topics in Microeconometrics

  Final exam

Module content:

Lecture programme (Lecturer: Ralf Wilke)

- self-contained topics
- emphasise particular econometric issues and problems
- illustrative practical examples can be replicated

Exercises

- 7 times two hours classes (mix of theory and empirical questions).
- Content relevant for the exams.
- Work in R: Data, example files, output online
- Group work/presentations (groups on Canvas)

Lecture programme

Consists of 7 topics:

Emne A1. Estimation of the multiple regression model by OLS in matrix notation – Distribution and asymptotic properties

Prerequisite*: Estimation of the linear regressions model by OLS is known+ Standard inference methods. (Ch3+4, W2025), Fundamentals: AppA+D, W2025.*

Emne A2. OLS – topics

Emne A3. Policy Analysis

Emne B1. Endogeneity

Emne B2. Simultaneous Equation Models

Emne B3. Maximum Likelihood methods

Emne B4. Limited Dependent Variable Models

Leksions program

- *Lecture times:*
  - Check your Calendar on MyCBS.dk
- *Exercise times:*
  - Check your Calendar on MyCBS.dk
- Download teaching material from Canvas.
  - Lecture slides, revision material, problem sets
  - R code, data for lectures and exercises

Feedback

- Feedback hours:

  Mondays 14:00-16:00 in PH16, room 1.72
  From 1. september 2025. Teaching weeks only.

  Please book your 20 minutes appointment in advance.
  Go to course calendar in Canvas.

  2 bookings per student.
  If you need additional appointments, check my availability in Canvas.

Module texts

- The main course texts are:

  - **Wooldridge, J.** (2025) Introductory Econometrics (8th edition), Cengage.
    - eBook and blended learning environment by the publisher: MindTab: course environment: https://cengage.widen.net/s/q9wlchgl2q/emea_cbs_ba-bmecv1031u_econometrics_fall_2025
  - **Wooldridge, J.** (2010) Econometric Analysis of Cross-Section and Panel Data (2nd Edition), Cambridge: MIT Press.
    - eBook rental option: https://mitpress.ublish.com/book/econometric-analysis-cross-section-and-panel-data#purchase

  These books deal well with most aspects of this part of the course.

  To work with R:

  - **Heiss, F.** (2016) Using R for Introductory Econometrics

Review: Basics of the Linear Regression Model

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

- A simple model for explaining y given x is:

$$y = \beta_0 + \beta_1 x + u$$

- Some terminology:

  y: Dependent Variable, Explained Variable, Response Variable, Predicted Variable, Regressand

  x: Independent Variable, Explanatory Variable, Control Variable, Predictor Variable, Regressor

  u: Error Term or Disturbance, Key assumption: E(u|x)=0

  It represents factors other than *x* that affect *y*.

  β_1 slope parameter

  β_0 intercept

- The linearity of the model implies that a one-unit change in *x* has the same effect on *y* regardless of the initial value of *x*.
  - This is unrealistic in many economic applications.
- Does β_1 really measure the effect of *x* on *y* if we ignore all other factors? -> causality!
- How to obtain estimates for the coefficients of the population regression function (PRF)?
  - Use OLS!

- Example: CEO Salary and Return on Equity
- The dataset CEOSAL1.dta contains information on 209 CEOs for the year 1990.
- Let us analyse the association between the equity and the wage of a chief executive officer in the population of CEO's:
- *y*: annual *salary* in thousands
  - If *salary*=856.3 then the annual salary is $856,300.
- *x*: average return on equity (*roe*) for the CEO's firm for the previous three years.
  - If *roe*=10, then the average return is 10 percent

- We postulate the simple model:

$$salary = \beta_0 + \beta_1 roe + u$$

- β_1 measures the change in annual salary, in thousands of dollars, when return on equity increases by one percentage point.

- Using an econometric package, the OLS regression line relating *salary* to *roe* is

$$\widehat{salary} = 963.191 + 18.501 roe$$

- How to read these results?
  - If *roe*=0, the predicted salary is $963,191.
  - If roe=30, then

$$\widehat{salary} = 963.191 + 18.501 \times 30 = 1518.221$$

  - Next, $\widehat{\Delta salary} = 18.501 \times \Delta roe$. This means that if the return on equity increases by one percentage point, $\Delta roe = 1$, then *salary* is predicted to change by about $18,500.

- The model with 2 variables can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

where

β_0: intercept

β_1: measures the change in y with respect to x_1 holding other factors fixed

β_2: measures the change in y with respect to x_2 holding other factors fixed

- The key assumption of this model is: $E(u|x_1, x_2) = 0$
  - It means that for any value of x_1 and x_2 in the population, the average unobservable is zero.
  - How to read this? Suppose *u* is ability in a wage equation.

    Then for any combination of *exper* and *educ*, the average ability level is assumed to be the same.

The Model with *k* Independent Variables

- The general multiple regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- This equation contains *(k+1)* unknown population parameters.
- The key assumption for the general multiple regression model is: $E(u|x_1, x_2, \ldots, x_k) = 0$.
- We still require that *u* is uncorrelated with all independent variables $x_1, \ldots, x_k$.

Generalising functional relationships between variables

- A special case is for $x_2 = x_1^2$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$$

  which is a quadratic function in x_1.
- The multiple regression model therefore incorporates a variety of nonlinear functional forms.
- Please note that the interpretation of the parameter β_1 changes in this case because the other regressor cannot be held constant.
- The key assumption is in the case $x_2 = x_1^2$:

$$E(u|x_1, x_2) = E(u|x_1) = 0$$

- Models involving logarithms have interesting properties:

Model	*Dependent Variable*	*Independent Variable*	*Interpretation of* β_1
Level-level	*y*	*x*	$\Delta y = \beta_1 \Delta x$
Level-log	*y*	log(*x*)	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	log(*y*)	*x*	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	log(*y*)	log(*x*)	$\%\Delta y = \beta_1 \%\Delta x$

- These approximations rely on the properties of the natural logarithm. Less precise the larger the coefficient is.
- Exact formula for the effect of a change of a dummy variable in the log-level model is: $100[exp(\beta_1) - 1]$

- Interpreting the OLS Regression

  - Similar to the simple regression model
  - Now we can explicitly fix all observed regressors, which makes multiple regression analysis so useful.
  - $\hat{\beta}_0$ is the predicted *y* if all regressors *x* are 0.
  - The slope coefficients have partial effect, or ceteris paribus, interpretation:

$$\begin{aligned} \Delta\hat{y} &= \hat{\beta}_1 \Delta x_1 \\ &\vdots \\ \Delta\hat{y} &= \hat{\beta}_k \Delta x_k \end{aligned}$$

    when all other regressors are held constant, i.e. $\Delta x_l = 0$. This means, we control for the other variables.

- Example: Wage Equation (wage1.dta)

$$\widehat{log(wage)} = 0.284 + 0.092educ + 0.0041exper + 0.022tenure$$

with *tenure*: years with the current employer.

- Holding exper and tenure fixed, the coefficient 0.092 means that another year of education is predicted to increase *log(wage)* by 0.092, which translates into a 9.2% increase in *wage*.

Why is this?
Because:
$\%\Delta y \approx (100\beta_1)\Delta x_1$

- This estimate of the return to education at least keeps two important productivity factors fixed.
- We need to discuss in light of the statistical properties of OLS in order to see whether it is a good estimate or not.

- Changing more than one independent variable simultaneously:
  - What is the estimated effect on wage if the individual stays at the same firm for one more year? In this case *exper* and *tenure* increase:

$$\Delta log(\widehat{wage}) = 0.0041\Delta exper + 0.022\Delta tenure = 0.0263$$

This means that we should expect an increase in *wage* by 2.6%.

- The strength of the multiple regression model is that we can compare expected outcomes if we set the independent variables to different values.
- This does not require that the data is experimental, i.e. it can be a random sample with “arbitrary” values of the regressors.

- What is then the meaning of “linear” regression?
  - As we have seen, the general linear regression model also allows for certain nonlinear relationships.
  - What does “linear” mean?
  - The key is that the model is linear in the parameters β_0 and β_1…
  - Warning: the interpretation of the model coefficient does depend on the their definitions.

Inference

- You should be aware of the following:

- Hypothesis testing:
  - one sided alternatives $H_0 : \beta_j \leq 0 \text{ vs. } H_1 : \beta_j > 0$
  - two sided alternatives $H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$
    $H_0 : \beta_j = a_j \text{ vs. } H_1 : \beta_j \neq a_j$

  - testing for linear restrictions $H_0 : \beta_1 = \beta_2$

$$H_0 : \beta_i = \beta_j = \beta_k = 0 \text{ vs. } H_1 : H_0 \text{ is not true}$$

- T-values, F-values

- P-values

- Example: test score data: GPA1.dta (sample code inference)

- In the regression output, what is a p-value and a F-value?

- construct t-value, p-value, F-value. (tvalue_example.R, pvalue_code.R, ftest_code.R)

- test *educ*
- test educ=exper
- test_linear_restrictions_example.R

Revision

- Relevant textbook chapters in Wooldridge, Introductory Econometrics:
  - Chapter 2: Simple Regression Model
  - Chapter 3: Multiple Regression: Estimation
  - Chapter 4: Multiple Regression: Inference
- Parts of the content of chapters 3 and 4 will be touched during the next couple of weeks but we will adopt a higher level presentation using matrix notation (as in Appendix E). We will see several proofs of important properties.
- Problem Set 0: Revision of relevant mathematical & statistical tools. (Appendices C-D)

Multiple Linear Regression Model

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

OLS properties

Textbook:

Wooldridge (2025), Chapter 4.

Motivation for Multiple Regression

- Multiple regression analysis allows us to explicitly control for many factors that simultaneously affect the dependent variable.
- For this reason we can hope to infer causality in cases where simple regression analysis would be misleading.
- If we add more variables to explain *y*, then more of the variation of *y* can be explained. Thus, we can better predict the dependent variable.
- The model incorporates fairly general functional form relationships.

The Model with *k* Independent Variables

- The general multiple regression model is

$$\begin{aligned} y &= \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + u \\ &= \mathbf{x}\beta + u \end{aligned}$$

- x_1 is typically 1.
- This equation contains *K* unknown population parameters.
- The key assumption for the general multiple regression model is: $E(u|\mathbf{x}) = 0$.
- We still require that *u* is uncorrelated with all independent variables $x_1, \ldots, x_K$.

Estimation by OLS

- We seek estimates $\hat{\beta}_1, \ldots, \hat{\beta}_K$ in the equation

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_K x_K$$

- The OLS estimates are chosen to minimise the squared residuals:

$$\sum_{i=1}^{N} (y_i - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_K x_{iK})^2$$

  with observations *i=1,…,N*.
- Derive the *K* first order conditions which are linear and solve for the *K* unknowns.

Derivation of OLS estimates

- Using matrix notation.
- For

$$\begin{aligned} y_i &= \beta_1 x_{i,1} + \ldots + \beta_K x_{i,K} + u_i \\ &= \mathbf{x}_i \beta + u_i \end{aligned}$$

with

$$\mathbf{x}_i' = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,K} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$

- Define:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \text{ and } \quad \mathbf{X}_{N\times K} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \vdots & & & \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{pmatrix}.$$

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}$$

- This then altogether yields:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$\mathbf{X}\boldsymbol{\beta}$ is *(N* x *1)* because $\boldsymbol{\beta}$ is *K* x *1* and $\mathbf{X}$ is *N* x *K*.

- The *Kx1* vector of OLS estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K)'$, minimises *SSR(**b**)* over all possible vectors **b** and is:

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

- Define the fitted values and residuals as:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

OLS regression line
Sample regression function

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

- Then because of $\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$, we have

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$$

  - Because the first column of ***X*** consists of ones, the OLS residuals always sum to zero.
  - The covariance between each independent variable and the OLS residuals is zero

- Example: Determinants of College GPA

  - Data set: GPA1.dta
  - Contains: *colGPA*= college grade point average, *hsGPA*= high school GPA, *ACT*= achievement test score.
  - *n*=141 students from a large university
  - Both GPAs are on a four point scheme
  - We obtain the following OLS regression to predict college GPA:

$$\widehat{colGPA} = 1.29 + 0.453 hsGPA + 0.0094 ACT$$

  - How do we interpret this?
  - The predicted college GPA if *hsGPA* and *ACT* are zero, is 1.29.
    - Since *hsGPA*=0 or *ACT*=0 does not exist in the sample, the intercept is not meaningful.

- More interesting are the slope coefficients.
  - Positive relationship between *colGPA* and *hsGPA*, holding *ACT* fixed. (a point increase in *hsGPA* predicts an increase of *colGPA by 0.453*)
  - Positive relation between *ACT* and *colGPA*, holding *hsGPA* fixed. The estimated effect is however, very small (an increase of 1000 in *ACT* predicts an increase of *colGPA* by less than one point). Note, that the sample average of *ACT* is about 24.
- What happens if we ignore *hsGPA* in the regression?

$$\widehat{colGPA} = 2.40 + 0.0271 ACT$$

thus, the coefficient on *ACT* is now almost three time larger, suggesting a stronger relationship.

  - In this model, we cannot, however, compare two people with the same high school GPA.
  - We will later formally discuss the implication if variables are omitted.

Comparison of Simple and Multiple Regression estimates

- Suppose *k*=2.
- In which cases will a simple regression of *y* on x_1 produce the same results as the regression of *y* on x_1 and x_2?
- Define:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \\ \tilde{x}_2 &= \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \end{aligned}$$

  One can show that: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$

- There is equality in two cases:
  - $\hat{\beta}_2 = 0$.
  - $\tilde{\delta}_1 = 0$. Why?

- Example: Determinants of College GPA

$$\widehat{colGPA} = 1.29 + 0.453 hsGPA + 0.0094 ACT$$

$$\tilde{colGPA} = \tilde{\beta}_0 + 0.482 hsGPA$$

  - The correlation between *hsgpa* and *ACT* is about 0.346 but the coefficient $\hat{\beta}_2$is very little.
  - For this reason the two slope estimates for *hsGPA* are quite similar.

- This reasoning can be extended to *k*-independent regressors. Estimates for β_1 are just identical if:
  - $\hat{\beta}_j = 0$ for $j = 2, \ldots, k$
  - if x_1 is uncorrelated with each of $x_2, \ldots, x_k$.

Goodness-of-Fit

- Same as in the simple regression model, because definitions only depend upon y_i, $\hat{y}_i$ and $\hat{u}_i$:
  - Total sum of squares (SST): $\mathrm{SST}=\sum_i (y_i - \bar{y})^2$
  - Explained sum of squares (SSE): $\mathrm{SSE}=\sum_i (\hat{y}_i - \bar{y})^2$
  - Residual sum of squares (SSR):

$$\mathrm{SSR}=\sum_i \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

    with SST=SSE+SSR.

  - $R^2=\mathrm{SSE}/\mathrm{SST}=1\text{-}\mathrm{SSR}/\mathrm{SST}$
    which is between *0* and *1*.

The expected value of the OLS estimators

- We state and discuss four assumptions, which are direct extensions of the simple regression model assumptions, under which the OLS estimators are unbiased for the population parameters.

Assumption 1: Linear in Parameters

The model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \ ,$$

where ***y*** is an observed (*n x 1*) vector, ***X*** is an *n x k* observed matrix, and ***u*** is an (*n x 1)* vector of unobserved errors or disturbances.

- This is the population or true model.

Assumption 2: Random Sampling

We have a random sample of size *n*, $\{(x_{i1}, x_{i2}, \ldots, x_{iK}, y_i)_{i=1,\ldots,n}\}$, following the population model in Assumption 1.

- This assumption implies that the selection into the sample is random. In particular that it is not related to the error term *u.*

- OPTIONAL MATERIAL:

  Hirschauer et al. (2021), Inference using non-random samples? Stop right there! Significance.

Assumption 3: No Perfect Collinearity

The matrix ***X*** has rank(*K*).

- Under this assumption, ***X'X*** is nonsingular and we can write $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- This assumption does not preclude some correlation between the independent variables, it rules out perfect correlation.

Assumption 4: Population orthogonality condition

$$E(\mathbf{x}'u) = 0$$

- This assumption rules out that there are independent variables which are correlated with u.
- It is implied by the zero conditional mean assumption:

$$E(u|\mathbf{x}) = 0 \qquad \text{(Assumption 4')}$$

- If ***x*** contains a constant, *u* has zero mean.
- The independent variables are said to be exogenous.

Theorem 1.1 (Unbiasedness of OLS)

Using Assumptions 1 through 4', the OLS estimator $\hat{\beta}$ is unbiased for β.

- Proof:
  - First rewrite $\hat{\beta}$ as a function of β:

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

$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = I_K$

  - Then take the expectation conditional on ***X***

$$\begin{aligned}E(\hat{\beta}|\mathbf{X}) &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{0} \\ &= \beta\end{aligned}$$

Assumption 4'

The Variance of the OLS Estimators

- In order to derive the simplest form of the variance-covariance matrix of $\hat{\beta}$, we make an additional assumption.

Assumption 5 (Homoscedasticity)

(i) $\mathrm{Var}(u_i|\mathbf{X}) = \sigma^2$ for *i=1,…,n*.

(ii) $\mathrm{Cov}(u_i, u_j|\mathbf{X}) = 0$ for all $i \neq j$.

In matrix form this is

$$\mathrm{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 I_N ,$$

where I_N is the (*N x N*) identity matrix.

- Part i) says that the variance of *u* cannot depend on any element of ***X***.
- Part ii) says that the errors cannot be correlated across observations. It is implied by Assumption 2.

Theorem 1.2 (Variance Covariance Matrix of the OLS Estimator)

Under Assumptions 1 through 5,

$$\text{var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

- Proof: First, remark that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$.

  Then $\text{var}(\hat{\boldsymbol{\beta}}) = \text{var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})$.

  and conditional on ***X*** :

$$\begin{aligned} var(\hat{\beta}|\mathbf{X}) &= var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[var(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

$var(\mathbf{A}'\mathbf{X}) = \mathbf{A}'var(\mathbf{X})\mathbf{A}$

Assumption 5

$\mathbf{X}'I_n = \mathbf{X}'$

$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = I_K$

$$\mathrm{Var}(\hat{\beta}_j|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}$$

- the conditional variance of $\hat{\beta}_j$ is obtained by multiplying σ^2 by the j'th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

$SST_j = \sum_i (x_{ij} - \bar{x}_j)^2$
R_j^2: R-squared of x_j on other x

- One can show that:

$$\mathrm{Var}(\hat{\beta}_j|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj} = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

  - The conditional variance is large if
    - σ^2 is large.
    - the sample variation in x_j is small (SST_j).
    - there is a high correlation between x_j and any other x_l, $l \neq j$: R_j^2 is large. *"Multicollinearity"*
- The conditional variance decreases with *N.*

Multicollinearity

- If there is high (but not perfect) correlation between two or more variables.
- This implies that the R^2 of a regression of x_j on all other independent variables is large R_j^2.
- In this case $\text{var}(\hat{\beta}_j|\mathbf{X})$ is large. It explodes as R_j^2 goes to one, i.e. $R_j^2 \to 1 \Longrightarrow \text{var}(\hat{\beta}_j|\mathbf{X}) \to \infty$
- Note: if x_j is uncorrelated with all other x_l, $l \neq j$: $R_j^2 = 0$.
- Why is $R_j^2 \neq 1$?

- In an application it is better to have less correlation between the regressors.

- As $\text{var}(\hat{\beta}_j|\mathbf{X})$ explodes when $R_j^2 \to 1$ it could be useful to define an upper ”acceptable” level of multicollinearity.
- This is sometimes by considering the so-called variance inflation factor (VIF):

$$\text{Var}(\hat{\beta}_j|\mathbf{X}) = \frac{\sigma^2}{SST_j(1-R_j^2)} = \frac{\sigma^2}{SST_j} VIF_j$$

with

$$VIF_j = 1/(1-R_j^2).$$

  - Evidently, $VIF_j = 1$ whenever there is no correlation between x_j and the other regressors and VIF_j explodes as $R_j^2 \to 1$.
  - $VIF_j > 10$ indicates a high a degree of multicollinearity and can be used to explain large variance of OLS estimates.
    - This does not mean, however, that omitting some of the variables will “improve” estimates as this normally leads to omitted variable bias. Trade-off! (Code: multiple_reg_vif.R)

- Alternatively, statistical regularisation methods can be applied that automatically select the relevant regressors.
  - Useful techniques if the regressor set is large.
  - For consistency, all variables of the population model are observed plus some that may not belong to the model.
  - Superior than sequential elimination of variables (Oracle property).
  - Corresponds to an OLS regression under additional inequality constraints on the parameters (Penalised regression).

Example:
What makes employees change their employer?

- To some extent there is some theory suggesting important variables:
  - Pay (absolute/relative)
  - Performance
  - Contract duration

- What else?
  - Possibly a vast number of variables and sources.

- Classical variable selection approaches based on VIF and sequential elimination are not optimal.
- Human resources analytic software produces scores or fitted probabilities for the probability of leaving.
  - Typically based on fitting methods such as Neural Networks. Limited interpretability of results (black box).
- What are the relevant factors and how to catch them?

We wish to have an oracle that tells us what to include.

Each data source may comprise of a large number of variables.

Variable selection techniques

- High-dimensional data cause problems for estimation when
  - K>N (more candidate variables than observations).
  - High degree of multicollinearity.
- Drawbacks of sequential elimination methods (subset selection based on likelihood ratio test, stepwise selection based on AIC/BIC,etc.). Code: stepwise_aic.R
  - Only a small number of variables can be tested.
  - Results are highly affected by sequence of the test.
  - Overfitting.
- Desired statistical property
  - Oracle property: Only the relevant variables are selected, and the estimates of those variables are asymptotically equal to the estimates from a model that only includes the relevant variables.

Penalised Regression

- A penalty is added to the objective function that penalises the use of too many variables.
- Objective function: $min_\beta L(\beta|X,y)$
- Add penalty: $P_\lambda(\beta)$, where λ is a penalisation/tuning parameter.
- Penalised regression: $min_\beta L(\beta|X,y) + P_\lambda(\beta)$
- Also called:
  - Shrinkage methods
  - Statistical regularisation
  - Unsupervised learning

e.g.
OLS: $L(\beta|X,y) = \sum(y_i - X_i\beta)^2$

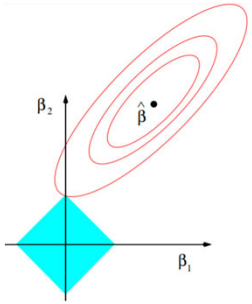
Shrinkage methods

Various methods. Differ in the choice of penalty.

- Individual variable selection
  - Lasso (ℓ_1-type penalty)

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K} |\beta_k|$$

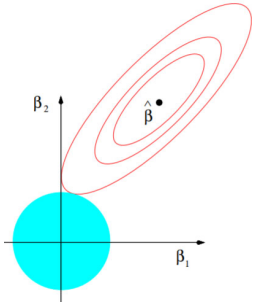
$$min\, L(\beta|X,y)\, s.t. \sum_{k=1}^{K} |\beta_k| < t$$



  - Ridge (ℓ_2-type penalty)

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K} \beta_k^2$$

$$min\, L(\beta|X,y)\, s.t. \sum_{k=1}^{K} \beta_k^2 < t$$



- Group variable selection
  - Group Lasso

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{J} \sqrt{A_j} \sqrt{\sum_{k=1}^{A_j} \beta_{jk}^2}$$

- Bi-level variable selection
  - (Adaptive) group bridge

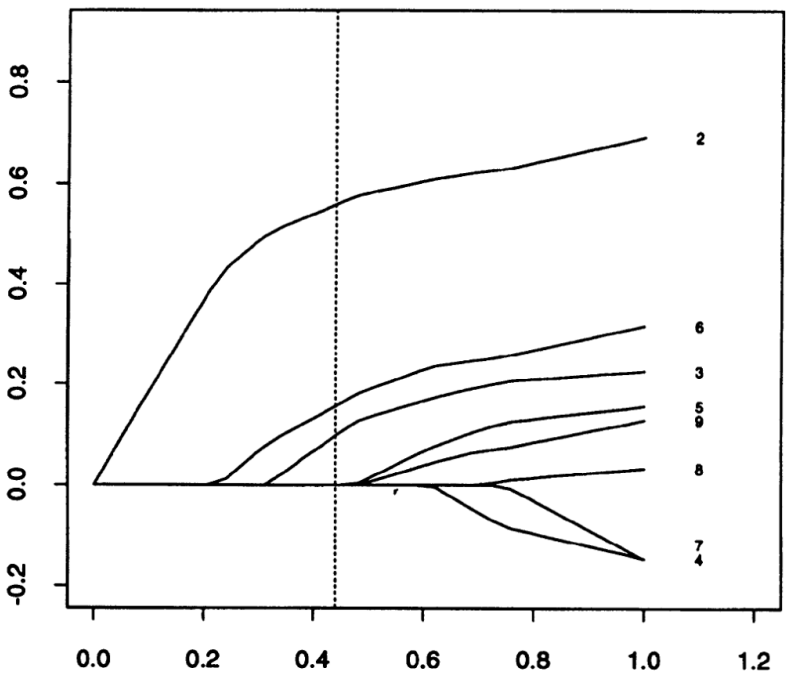
$$P_\lambda(\beta) = \lambda \sum_{j=1}^{J} \sqrt{A_j} \sqrt{\sum_{k=1}^{A_j} w_{jk} |\beta_{jk}|}$$

Source: Hastie et al (2009)
K=2, t is tuning parameter.

A_j: number of variables in a group
w_{jk}: weights

The tuning parameter for LASSO

- The number of selected variables increases with t (penalty becomes smaller=λ decreases).
- The choice of the tuning parameter is important:
  - Various methods (optional): AIC (some models), BIC, Cross validation (CV), Generalised Cross Validation (GCV)
- Example code: shrinkage.R



Source: Tibshirani (1996)

Penalty terms (literature optional)

- Penalty terms incorporate different beliefs on the structure and magnitude of the variables and result in different models
  - Individual variable selection: Lasso (Tibshirani, 1996), Elasticity net (Zou and Hastie, 2003), Adaptive Lasso (Zou, 2006), Fused Lasso (Tibshirani, 2005)
  - Group-level variable selection: Group Lasso (Yuan and Lin, 2006), Hierarchical Lasso (Zhao et al., 2009)
  - Bi-level variable selection: Group bridge (Huang et al., 2009), Sparse group lasso (Simon et al., 2013)
- Simultaneous variable selection and inference is challenging. Still a developing field.
  - Sample splitting (Meinshausen et al., 2009), covariance test (Lockhart et al., 2014), exact post-selection inference (Lee et al., 2016), OLS post-Lasso (Belloni and Chernozhukov, 2013), etc.

Selected References (optional)

- Journal references
  - Huang, J., Breheny, P., & Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models
  - Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference
- Book references
  - Bühlmann, P., & Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications
  - Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations.

Variances in misspecified models

- How do the variances change if we omit a variable?
- Remember: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

  $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

- We know: $\mathrm{Var}(\hat{\beta}_1|\mathbf{X}) = \sigma^2 / SST_1(1 - R_1^2)$
- It can be shown: $\mathrm{Var}(\tilde{\beta}_1|\mathbf{X}) = \sigma^2 / SST_1$
- This implies that $\mathrm{Var}(\tilde{\beta}_1|\mathbf{X}) \leq \mathrm{Var}(\hat{\beta}_1|\mathbf{X})$
- They are equal if all independent variables are uncorrelated, otherwise not.

- If $\beta_2 = 0$, do not include x_2 in the regression, because variance of the OLS estimator may increase, while both estimators are unbiased.

- The case $\beta_2 \neq 0$ is more difficult:
  - There is a trade-off between bias and variance
  - For large samples we may, however, prefer to include in $\hat{\beta}$, because the variance becomes less important, while the bias does not depend on the sample size.

Estimating σ^2

- The sampling variance of $\hat{\beta}_j$ depends on σ^2.
- Since $E(u_i^2) = \sigma^2$, it would be natural to estimate σ^2 by $N^{-1} \sum_i u_i^2$ this is, however, not possible because u_i is unknown.
- Instead, we use the estimated u_i, which are unbiased.

Theorem 1.3 (Unbiased Estimation of σ^2)

Under Assumptions 1 through 5,

$$E(\hat{\sigma}^2) = \sigma^2$$

With $$\hat{\sigma}^2 = (N-K)^{-1}\sum_i \hat{u}_i^2 = (N-K)^{-1}\hat{\mathbf{u}}'\hat{\mathbf{u}}$$
$$= SSR/(N-K) .$$

- The denominator is *(N-K)* and not *N* because the residuals have to satisfy the *K* conditions:

$$\sum_i x_{il}\hat{u}_i = 0 \text{ for } l = 1, \ldots, K$$

- For this reason we have only (*N-K)* degrees of freedom.

- In contrast to the SSR, $\hat{\sigma}$ can increase or decrease when another variable is added, since degrees of freedom decrease.

- $\hat{\sigma}$ is called the standard error of the regression (**SER**) and typically reported by econometric packages.

- The standard error of $\hat{\beta}_j$ is therefore estimated by

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{[SST_j(1-R_j^2)]^{1/2}} = \hat{\sigma}[(\mathbf{X}'\mathbf{X})_{jj}^{-1}]^{1/2}$$

- Since it relies on the homoscedasticity assumption, it is not a valid estimator if Assumption 5 does not hold.

Efficiency of OLS

- It can be shown that under the above assumptions that OLS has another nice property.

> **Theorem 1.4 (Gauss-Markov Theorem)**
>
> Under Assumptions 1 through 5, $\hat{\beta}$ is the best linear unbiased estimator (BLUE).

- This means that for any estimator $\tilde{\beta}_j$ that is linear and unbiased, $\text{var}\hat{\beta}_j \leq \text{var}(\tilde{\beta}_j)$, i.e. OLS has the smallest variance among all unbiased linear estimators.
- For this reason, we don't need to look for a better estimator under Assumptions 1 – 5.
- Assumptions 1 - 5 are known as Gauss-Markov assumptions.

- Finite sample or exact properties of the OLS estimators:
  - Unbiasedness holds for any sample size *N* if the four Assumptions 1-4' hold.
  - Also, the fact that OLS is the best linear unbiased estimator under Assumptions 1-5 is a finite sample property

- It is also important to know the large sample or asymptotic properties. This is if sample size grows without bound ($N \to \infty$).
  - OLS estimators have nice asymptotic properties (consistent and asymptotically normal distributed).

Theorem 1.5 (Consistency of OLS)

Under Assumptions 1 through 4, the OLS estimator $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

Theorem 1.6 (Asymptotic Normality of OLS)

Under the Gauss-Markov Assumptions 1 through 5,

(i) $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$ with $\mathbf{A} = E(\mathbf{x}_i' \mathbf{x}_i)$

(ii) $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 = \text{var}(u)$.

(iii) for each j, $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \overset{a}{\sim} N(0,1)$

- From now we mainly focus on asymptotic properties.

- Under the Gauss-Markov assumptions, the OLS estimators are best linear unbiased.
  - OLS is also asymptotically efficient among a certain class of estimators under the Gauss-Markov assumptions.
  - A wide class of estimators are unbiased for β but OLS has the smallest asymptotic variance in this class.
- Suppose *g(x)* is any function of x such that *g(x)* and *u* are uncorrelated. Let $\tilde{\beta}$ be the solution to the *K* conditions:

$$\sum_i g_j(\mathbf{x}_i)(y_i - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_K x_{iK}) = 0 \text{ for } j = 0, 1, \ldots, K$$

Theorem 1.7 (Asymptotic Efficiency of OLS)

Under the Gauss-Markov Assumptions 1 through 5, let $\tilde{\beta}_j$ denote estimators that solve the above equations and let $\hat{\beta}_j$ denote the OLS estimators. Then for j=1,2,…,K, the OLS estimators have the smallest asymptotic variances:

$$\text{avar } \sqrt{n}(\hat{\beta}_j - \beta_j) \leq \text{avar } \sqrt{n}(\tilde{\beta}_j - \beta_j)$$

Large Sample Test: Lagrange Multiplier Statistic

- For most purposes there is little reason to go beyond the usual *t* and *F* statistics.
- There are, however, other ways to test multiple exclusion restrictions.
- The Lagrange Multiplier (**LM**) Statistics has achieved some popularity in modern econometrics.
- It does not require estimation of the unrestricted model.
- It does not require the normality of errors.

- A guide to perform the LM Test for q exclusion restrictions:
  1. Regress *y* on the restricted set of independent variables and save the residuals $\tilde{u}$.
  2. Regress $\tilde{u}$ on all independent variables and obtain the R-squared, say R_u^2.
  3. Compute $LM = NR_u^2$.
  4. Compare *LM* to the appropriate critical value, *c*, in a χ_q^2 distribution; if *LM*>*c*, the null hypothesis is rejected. (similarly, one can also compute the p-value and reject if it is too low.) Otherwise, H_0 cannot be rejected.

- Often results are similar compared to the F-test.
- The *F* statistic is usually automatically computed by econometric packages.

Summary

- In the multiple regression model:
  - We can hold several factors fixed while looking at partial effects.
  - Independent variables can be correlated.
  - Selection of independent variables important.
  - Captures a variety of nonlinear relationships between x_j and *y.*
  - OLS is easy to calculate and has nice properties under five assumptions: unbiasedness & efficiency
  - Sample and asymptotic distribution of the OLS estimator, which can be used for inference.

Multiple Regression Model: Violations of the G-M assumptions

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

Heteroskedasticity

Consequences of Heteroskedasticity for OLS

- The homoscedasticity assumption 5 is required for the derivation of the sampling and the asymptotic distribution of OLS in the “classical” regression model.
- Not required for consistency and unbiasedness of OLS.
- If there is no homoskedasticity the estimated variances of $\hat{\beta}_j$ are biased and t- and F-statistics and CIs are not valid anymore.
- OLS no longer efficient (smallest variance).

- A generalised model is by relaxing assumption 5.

Assumption 5' (Heteroscedasticity)

(i) $\mathrm{Var}(u_i|\mathbf{X}) = \sigma_i^2 = \sigma^2\omega_i$ for *i=1,…,N*.

(ii) $\mathrm{Cov}(u_i, u_j|\mathbf{X}) = 0$ for all $i \neq j$.

In matrix form this is

$$\mathrm{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{\Omega},$$

where $\sigma^2\mathbf{\Omega}$ is a (*N x N*) diagonal matrix with diagonal elements σ_i^2. $\mathbf{\Omega}$ is a diagonal matrix with some positive elements $\omega_i > 0$.

- $\mathbf{\Omega}$ is assumed to be positive definite.

- The exact sampling distribution of $\hat{\boldsymbol{\beta}}$ becomes:

Theorem 1.8 (Variance Covariance Matrix of the OLS Estimator)

Under Assumptions 1 through 4 and 5':

$$\text{var}(\hat{\boldsymbol{\beta}} \,|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}.$$

Proof: First, remark that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$.

Then $\text{var}(\hat{\boldsymbol{\beta}}) = \text{var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})$.

and conditional on $\boldsymbol{X}$:

$$\begin{aligned} var(\hat{\beta}|\mathbf{X}) &= var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[var(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\boldsymbol{\Omega})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Similarly, the asymptotic variance is:

$$\text{avar}(\hat{\beta}|\mathbf{X}) = (E(\mathbf{x}'\mathbf{x}))^{-1}(E(u^2\mathbf{x}'\mathbf{x}))(E(\mathbf{x}'\mathbf{x}))^{-1}$$

for which a valid estimator is:

$$\widehat{avar}(\hat{\beta}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_i \hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i\right)(\mathbf{X}'\mathbf{X})^{-1}$$

- This is called the heteroscedasticity robust standard error for $\hat{\beta}_j$.
- Also called Huber or White standard error.
- One can show that the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is:

Theorem 1.9 (Asymptotic Distribution of the OLS Estimator)

Under Assumptions 1 through 4 - 5':

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \sigma^2(E(\mathbf{x}'\mathbf{x}))^{-1}(E(u^2\mathbf{x}'\mathbf{x}))(E(\mathbf{x}'\mathbf{x}))^{-1}).$$

- Common test for individual and joint significance applicable under some modifications.
- For example:
  - A heteroskedastic-robust *t* statistic is:

$$t = \frac{estimate - hypothesized \quad value}{standard \quad error}$$

  - Heteroskedasticity robust versions of Wald test and F test.

- Example: log wage equation (wage1.dta)

$$\begin{aligned}\widehat{log(wage)} &= 0.321 + 0.213marrmale - 0.198marrfem - 0.110singfem \\ &\quad (0.100)\ (0.055)\ (0.058)\ (0.056) \\ &\quad [0.109]\ [0.057]\ [0.058]\ [0.057] \\ &\quad +0.079educ + 0.027exper - 0.00054exper^2 + 0.029tenure - 0.00053tenure^2 \\ &\quad (0.0067)\ (0.0055)\ (0.00011)\ (0.0068)\ (0.00023) \\ &\quad [0.0074]\ [0.0051]\ [0.00011]\ [0.0069]\ [0.00024] \\ &\quad n = 526,\quad R^2 = 0.461\end{aligned}$$

  - Usual OLS standard errors are in parentheses, ().
  - Heteroskedasticity-robust standard errors are in brackets, [].

  - The two sets of standard errors are similar, no changes in significance.
  - Heteroskedasticity-robust standard errors can be larger or smaller than the usual OLS standard errors.
  - In applications, the heteroskedastic standard errors are often larger.

Testing for Heteroskedasticity

- The null is that Assumption 5 is true: $H_0 : \mathrm{var}(u|\mathbf{x}) = \sigma^2$
  - The heteroskedasticity function is independent of ***x***.
- Because of Assumption 4 the null is equivalent to:

$$H_0 : E(u^2|\mathbf{x}) = E(u^2) = \sigma^2$$

- This means we want to test, whether u^2 is related to one or more of the explanatory variables. The null is false, if u^2 is any function of $\mathbf{x}$.
- The various tests differ in the specification of the heteroskedasticity function.

- A simple approach is to assume a linear function:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_K x_K + \nu$$

- The null of homoskedasticity is: $H_0 : \delta_1 = ... = \delta_K = 0$
- However, the problem is that we do not observe u.
- Instead, use the OLS residuals as a consistent estimate for the error:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_K x_K + \nu$$

- One can show that the OLS residuals asymptotically do not affect the *F* and *LM* statistics.
- For this reason compute the *F* or *LM* statistic for the significance of this regression.

Breusch Pagan Test for Heteroskedasticity

1. Estimate the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ by OLS and obtain $\tilde{u}^2$.
2. Run the regression

$$\tilde{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_K + \nu$$

   and compute $R^2_{\tilde{u}^2}$.
3. Form either the F or the LM statistic and compare it to the critical value (or compute its P-value). This determines whether we can reject the hypothesis of homoscedasticity.

- Example: Housing price equation (hprice1.dta)

$$\begin{aligned} \widehat{price} &= -21.77 + 0.00207 lotsize + 0.123 sqrft + 13.85 bdrms \\ &\quad (29.48) \quad (0.00064) \qquad (0.013) \qquad (9.01) \\ &\quad n = 88, \quad R^2 = 0.672 \end{aligned}$$

  - In this case $R^2_{\tilde{u}^2} = 0.1601$, with N=88 and K=3:

$$F = [0.1601/(1 - 0.1601)](84/3) \approx 5.34$$
$$LM = 88 * 0.1601 \approx 14.10$$

This implies that the assumption of homoscedasticity is rejected and that OLS statistics are not reliable.

White Test for Heteroskedasticity

- It can be shown that the homoscedasticity assumption $\mathrm{var}(u|\mathbf{x}) = \sigma^2$ can be replaced by the weaker assumption that u^2 is uncorrelated with all x_j, x_j^2 and $x_j x_l \ (j \neq l)$.
- In this case we can regress $\hat{u}^2$ on all x_j, x_j^2 and $x_j x_l \ (j \neq l)$.

$$\begin{aligned}\hat{u}^2 \quad = \quad & \delta_0 + \delta_1 x_1 + ... + \delta_K x_K + \delta_{K+1} x_1^2 + ... + \delta_{K+K} x_K^2 \\ & + \delta_{K+K+1} x_1 x_2 + \delta_{K+K+2} x_1 x_3 + ... + \nu\end{aligned}$$

- The White test for heteroskedasticity is the *LM* (or *F*) statistic for testing that all the δ_j in the equation are zero.
- A disadvantage of the White test is that it uses many degrees of freedom.

- An interesting alternative which preserves the spirit of the White test is to use the OLS fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, instead of the independent variables, their squares and cross products:

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \nu$$

- We use the fitted values because they are linear functions of the independent variables.
- We can use the F or LM statistic for the null hypothesis

$$H_0 : \delta_1 = 0, \delta_2 = 0$$

- This test conserves more degrees of freedom and it is easy to implement.
- Since $\hat{y}$ is an estimate of *y* given x_j, the test is in particular useful in situations where we suspect that the variance changes with the level of $E(y|\mathbf{x})$.

A special case of the White Test

1. Estimate the model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ by OLS and obtain $\tilde{u}^2$.
2. Run the regression $\tilde{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \nu$ and compute $R^2_{\tilde{u}^2}$.
3. Form either the *F* or the *LM* statistic and compute the p-value or critical value using the $F_{2,N-3}$ or the χ^2_2 distribution.

- Note: results between the special form of the white test and Breusch-Pagan test can differ.
- Test for heteroskedasticity are sometimes also considered as test for specification (e.g. omitted variables) because in this case there is also no homoskedasticity. This is, however, not the best approach to test for this.

Generalised Least Squares Estimation

- If there is heteroskedasticity present, OLS is not efficient anymore.
  - Use **GLS/FGLS**.
- This will also lead to correct s.e. and new F and t statistics.

The heteroskedasticity function.

- We assume that $\mathrm{var}(u|x) = \sigma^2\omega(x)$, where *ω(x)>0* is some positive function.
- For a random drawing *i* from the population, we can write: $\sigma_i^2 = var(u_i|\mathbf{x}_i) = \sigma^2\omega(\mathbf{x}_i) = \sigma^2\omega_i$.
- For example, consider the simple savings function:

$$sav_i = \beta_1 + \beta_2 inc_i + u_i \qquad \mathrm{var}(u_i|inc_i) = \sigma^2 inc_i$$

  - Here: *ω(x)=ω(inc)=inc*. The error variance is proportional to the level of income.
- How, can we estimate such a model?
  - We transform it into a model which satisfies the Gauss-Markov assumptions and which has homoskedastic errors.

- Since $\omega_i = \omega(\mathbf{x}_i)$ is just a function of $\mathbf{x}_i$, $u_i/\sqrt{\omega_i}$ has zero expected value and its variance is σ^2 (both conditional on $\mathbf{x}_i$):

$$E\left((u_i/\sqrt{\omega_i})^2|\mathbf{x}_i\right) = E(u_i^2|\mathbf{x}_i)/\omega_i = (\sigma^2\omega_i)/\omega_i = \sigma^2$$

- Thus, we divide our model by $\sqrt{\omega_i}$:

$$y_i/\sqrt{\omega_i} = \beta_1(x_1/\sqrt{\omega_i}) + ... + \beta_K(x_{iK}/\sqrt{\omega_i}) + u_i/\sqrt{\omega_i}$$

or

$$\begin{aligned} y_i^* &= \beta_1 x_{i1}^* + ... + \beta_K x_{iK}^* + u_i^* \\ &= \mathbf{x}_i^*\beta + u_i^* \end{aligned}$$

- We can estimate this model efficiently by OLS. Why?
  - Because if the original model satisfies Assumptions 1-4, then also this model does, plus Assumption 5 holds for the latter.

- Example: savings function (cont.)
  - $sav_i = \beta_0 + \beta_1 inc_i + u_i \qquad \text{var}(u_i|inc_i) = \sigma^2 inc_i$
  - The transformed model is:

$$sav_i/\sqrt{inc_i} = \beta_0(1/\sqrt{inc_i}) + \beta_1\sqrt{inc_i} + u_i^*$$

  - The coefficient β_1 in this model is still the marginal propensity to save, the interpretation does not change.

- The estimators of the transformed model are examples of the generalized least squares estimators (**GLS**).
- All statistics (except the R^2) and estimators of the transformed model will be valid, the interpretation of the coefficients is as in the original model.

- The GLS estimator for correcting heteroskedasticity is:

$$\begin{aligned}\beta^* &= (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}(\mathbf{X}^{*'}\mathbf{y}^*) \\ &= (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}) \\ &= \beta + (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u})\end{aligned}$$

  - When we assume:

$$E(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}) = A \text{ is nonsingular } KxK.$$

$$E(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}) = \mathbf{0}$$

Consistency follows from an application of a law of large numbers and Slutsky's theorem:

$$plim\beta^* = \beta + \mathbf{A}^{-1}\mathbf{0} = \beta$$

- β^* is asymptotically normal with:

$$\sqrt{N}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \overset{a}{\sim} N(\mathbf{0}, \sigma^2(E(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}))^{-1})$$

  - Asymptotically efficient because the OLS estimator on the transformed data has minimum variance (Gauss-Markov).

- β^* is unbiased under a zero conditional mean condition on the error.

- Requires known $\boldsymbol{\Omega}$.

The heteroskedasticity Function must be estimated: Feasible GLS (FGLS)

- If we do not have an idea about the function $\omega(\mathbf{x}_i)$, we need to estimate it from data.
- Using $\hat{\omega}_i$ instead of ω_i in GLS, yields an estimator called **FGLS** estimator.

- Example of a positive variance function:

$$\text{var}(u|\mathbf{x}) = \sigma^2 exp(\delta_0 + \delta_1 x_1 + ... + \delta_k x_k)$$

with $\delta_0, ...\delta_k$ as unknown parameters.

- We can write using $E(\nu|\mathbf{x}) = 1$:

$$u^2 = \sigma^2 exp(\delta_0 + \delta_1 x_1 + ... + \delta_k x_k)\nu$$

and we assume $\text{cov}(\mathbf{x}, \nu) = 0$. We obtain:

$$\log(u^2) = \alpha_0 + \delta_1 x_1 + ... + \delta_k x_k + e$$

with *e* has zero conditional mean and is independent of ***x***. In general $\delta_0 \neq \alpha_0$ but this is not important.

- Since the Gauss-Markov assumptions are satisfied in this regression, we obtain consistent and unbiased estimates of δ_j.
- In an application, we replace *u* by its estimate $\hat{u}$ and regress $log(\hat{u}^2)$ on $\mathbf{x}$.
- Denote the fitted values of this regression as $\hat{g}_i$ and estimate ω_i by $\hat{\omega}_i = exp(\hat{g}_i)$.

- A Feasible GLS procedure to correct for heteroskedasticity.
  1. Run the regression of y on ***x*** and obtain the residuals $\hat{u}$.
  2. Create $\log(\hat{u}^2)$.
  3. Run the regression $log(\hat{u}^2)$ on $\mathbf{x}$ and compute the fitted values $\hat{g}_i$
  4. Compute $\hat{\omega}_i = exp(\hat{g}_i)$.
  5. Estimate the equation

     $$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

     by WLS, using weights $1/\hat{\omega}$. Each equation is divided by $\sqrt{\hat{w}}$.

- Since ω_i is now estimated, FGLS is not an unbiased estimator, but it is still consistent. Moreover, one can show that it is more efficient than OLS.
- FGLS is an attractive alternative to OLS if there is evidence for unknown heteroscedasticity.

- Example: Demand for cigarettes (smoke.dta)
  - We estimate the demand function for daily cigarette consumption.
  - The dependent variable is often zero, because there are non smokers in the data. A linear model is therefore not optimal because it predicts negative values but we can still learn something from this model.
  - We estimate by OLS:

$$\begin{aligned}\widehat{cigs} &= -3.64 + 0.880 log(income) - 0.751 log(cigpric) \\ &\quad (24.08) \quad (0.728) \qquad\qquad (5.773) \\ &\quad -0.501 educ + 0.771 age - 0.0090 age^2 - 2.83 restaurn \\ &\quad (0.167) \qquad (0.160) \qquad (0.0017) \qquad (1.11) \\ &\quad n = 807, \quad R^2 = 0.0526\end{aligned}$$

with *restaurn*: smoking restrictions in restaurant at state level.

  - 2% of the fitted values are below 0.
  - Is heteroskedasticity present? We compute the Breusch-Pagan test and obtain LM=32.26, which is strong evidence of heteroskedasticity.

- We estimate FGLS according to our specification of the conditional variance function:

$$\begin{aligned}\widehat{cigs} \quad = \quad & 5.64 + 1.30log(income) - 2.94log(cigpric) \\ & (17.80) \quad (0.44) \qquad\qquad (4.46) \\ & -0.463educ + 0.482age - 0.0056age^2 - 3.46restaurn \\ & (0.120) \qquad (0.097) \qquad (0.0009) \qquad (0.80) \\ & n = 807, \quad R^2 = 0.1134\end{aligned}$$

- Compared to the OLS regression:
- The income effect is now statistically significant and larger.
- The price effect is bigger but still insignificant.

- Remark: F statistic for testing multiple exclusion restrictions based on WLS estimates cannot directly be used. They have to be reweighed.

Summary

- We analysed the properties of OLS in presence of heteroskedasticity.
  - It does not cause bias or inconsistency, just inefficiency
  - Standard errors and t- and F-statistics are not valid anymore.
  - We have seen a heteroskedasticity robust estimator for the standard error. Many computer packages compute heteroskedasticity robust statistics.
- We have seen two tests for heteroskedasticity: the Breusch-Pagan and the White test.
- We have introduced GLS/FGLS as an efficient alternative to OLS.

Serial Correlation

Main text: parts of Chapter 12 in Wooldridge “Introductory Econometrics”, 2025.

- Error terms are correlated across observations/periods.
  - Common in panel data where the same unit has repeated observations.
  - Common in time series data where the same unit is observed repeatedly.
  - Also called Autocorrelation.
- Given that serial correlation is normally due to the longitudinal dimension (time), we use the index $t=1,\ldots,N$.
- Serial correlation can be sometimes removed by using differenced data (between periods) instead of the levels.
- Similar to heteroscedasticity, OLS is still consistent but no longer BLUE.
  - Gauss-Markov Theorem requires homoskedasticity and serially uncorrelated errors.

- The variance of the OLS estimator may be smaller or larger than the usual OLS variance under homoskedasticity, depending on the direction of the serial correlation.

- Testing for Serial correlation:
  - Strictly exogenous regressors: $E(u_t|\mathbf{X}) = 0$
    - Error in period *t* not correlated with regressors from any period.
    - T-test for AR(1), F/LM test for higher order serial correlation

- Serial error correlation in models with lagged dependent variable as regressor require special care (also called AR-models).

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

  - Dynamic structure.
  - Time series/financial econometrics.
  - **Dynamic panel models.**

  - Breusch-Godfrey Test (F-/LM test) for higher order serial correlation

- T-test for AR(1)
  - Large sample test
  - We consider the AR(1) error model

$$u_t = \rho u_{t-1} + e_t, t = 1, 2, ..., n$$

with $E(e_t|u_{t-1}, u_{t-2}, ...) = 0$ and $var(e_t) = \sigma_e^2$. $|\rho| \leq 1$.

  - Null Hypothesis: $H_0 : \rho = 0$
  - How can we test for this?
    - Under the assumptions above, the OLS estimator for the model above is asymptotically normal.
    - Run the regression of u_t on u_{t-1} (without intercept) and conduct a t-test!
    - Since u_t are not observed, use a consistent estimate, the OLS residuals of the main model: $\hat{u}_t$.
      - The large sample distribution of the t-statistic is not affected by this (when regressors are strictly exogenous).

- Step by step procedure:

1. Run the OLS regression of y_t on $x_{t1}, \ldots, x_{tk}$ and obtain the OLS residuals, $\hat{u}_t$, for all *t=1,…,n*.
2. Run the regression of $\hat{u}_t$ on $\hat{u}_{t-1}$ for all *t=2,…,n* and construct the t-statistic $t_{\hat{\rho}}$ for the coefficient $\hat{\rho}$ on $\hat{u}_{t-1}$. (the regression may contain an intercept or not. Asymptotically this does not affect $t_{\hat{\rho}}$).
3. Use $t_{\hat{\rho}}$ to test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ in the usual way. Normally, conclude that there is evidence for serial correlation when H_0 is rejected at the 5% level.

- Variation: To make this test robust to heteroskedasticity in e_t, use the usual heteroskedasticity robust version for $t_{\hat{\rho}}$.

Example: Testing for AR(1) serial correlation in the Phillips curve.

- Macro relation between inflation and unemployment rate.
- Data: phillips.dta , sample code: ar1_ttest.R

- Testing for higher order serial correlation AR(q)
  - An extension of the t-test approach for AR(1) is straightforward
  - Suppose there is serial correlation of order *q* in errors :

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + ... + \rho_q u_{t-q} + e_t, t = 1, 2, ..., N$$

  - The null hypothesis is: $H_0 : \rho_1 = 0, \rho_2 = 0, ..., \rho_q = 0$
  - The idea is to apply a test for joint significance of $\hat{u}_{t-1}, \hat{u}_{t-2}, ..., \hat{u}_{t-q}$.
  - This can be either the F-test or LM-test.
  - Heteroskedasticity robust version as for *q=1*.
  - As for *q=1*, it is an asymptotic test.

- [For finite sample test: Durbin-Watson test]

- Step by step procedure

1. Run the OLS regression of y_t on $x_{t1}, \ldots, x_{tk}$ and obtain the OLS residuals, $\hat{u}_t$, for all *t=1,…,n*.
2. Run the regression of $\hat{u}_t$ on $\hat{u}_{t-1}, \ldots, \hat{u}_{t-q}$ for all *t=q+1,…,n* .
3. Compute the F-test or LM test for joint significance of $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-q}$.

Reminder: The LM statistic in 3. is $LM = (n-q)R^2_{\hat{u}} \sim \chi^2_q$, where is the R-squared of the model in 2.

- Example: arq_test.R

- An interesting variation is to test only for selected lagged errors to be serially correlated, for example u_t and u_{t-4} if the data are quarterly.
  - Then only use the relevant lagged residuals on the right hand side in the model in 2.

- The LM test version of this test is a special case of a more general model without strictly exogenous regressors.
  - This is for example the case if one uses a lagged dependent variable as regressor (quite common in time series analysis).
  - It can be shown that if one uses $x_{t1}, x_{t2}, \ldots, x_{tk}$ as additional regressors in model 2, removes the "endogeneity distortion" in the statistic in 3.
  - The LM version of this test is called **Breusch-Godfrey** test for AR(q) serial correlation.

- Step by step procedure of Breusch-Godfrey test

1. Run the OLS regression of y_t on $x_{t1}, \ldots, x_{tk}$ and obtain the OLS residuals, $\hat{u}_t$, for all *t=1,…,n*.
2. Run the regression of $\hat{u}_t$ on $x_{t1}, x_{t2}, \ldots, x_{tk}, \hat{u}_{t-1}, \ldots, \hat{u}_{t-q}$ for all *t=q+1,…,n* .
3. Compute the LM test for joint significance of $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-q}$.

- Suppose there is evidence of serial correlation.
  - OLS estimates are still consistent but no longer BLUE and standard errors are incorrect.
  - What can we do about this?

1. Transforming data:

   - Similar to heteroskedasticity there exist GLS transformations. Application of OLS to transformed data produces estimates that are BLUE. FGLS requires pre-estimation of ρ and transformation of data in t=1 requires special care. Transformation depends on order of serial correlation.
   - If ρ is positive, large and even possibly =1 (random walk), first differencing of the data is appealing.

2. Serial correlation robust standard errors

- (First-)Differencing and serial correlation:
  - Consider the differenced version of the model

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, ..., n$$

that is

$$\Delta y_t = \beta_1 \Delta x_t + \Delta u_t, \quad t = 2, ..., n$$

with $\Delta y_t = y_t - y_{t-1}, \Delta x_t = x_t - x_{t-1}$ and $\Delta u_t = u_t - u_{t-1}$.

  - It can be shown that the differencing removes strong positive serial correlation but increases the amount of serial correlation when it is negative.

  - The interpretation of β_1 is the same in the two models.
  - OLS of differenced data produces (more) correct standard errors when u_t possesses strong positive serial correlation.

Example of first differencing: FD.R

- Data: (intdef.dta) time series of US macroeconomic data on three-months T-bill rate (*i3* – a measure for short-term interest rates), annual inflation rate (*inf*) and federal budget deficit as % of GDP (*def*).
- Estimate level and first differenced version of the model:

$$i3_t = \beta_0 + \beta_1 inf_t + \beta_2 def_t + u_t$$

- Do test of first order serial correlation in the two models.
- It is found that $\hat{\rho} = 0.623$ and highly significant in the level model and $\hat{\rho} = 0.073$ and insignificant in the differenced model. Better to use FD model!
- [Another observation: If regressors are exogenous (uncorrelated with the error term), both estimates are consistent. The results, however, point to significant changes. This likely points to endogeneity in the level model. (Unit root or omitted heterogeneity.) Another reason for FD!]

- Serial correlation robust standard errors and statistics
  - This is appealing because it does not require knowledge of the form of serial correlation (it applies to fairly arbitrary forms of serial correlation).
  - A general approach to correct the standard errors for fairly arbitrary forms of heteroskedasticity and serial correlation has been suggested by Newey and West (1987).
  - No details here but example code: nw_se.R
    - Estimation of Phillips curve with serial correlation robust standard errors.
  - S.E. can both increase or decrease but likely to increase if $\rho > 0$.

Policy Analysis

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

Readings:

Wooldridge, J. (2025), Introductory Econometrics, 8th Edition, South-Western, Chapter 3.7, 7.6, 13.2.

Potential Outcomes, Treatment Effects, Policy Analysis

- To estimate the casual effect of for example policy interventions
  - Does job training increase earnings?
  - Do more elective courses increase student outcomes?
- *w* is a binary policy indicator (=1: treated, =0: not treated).
- Potential outcomes: y(0) in absence of treatment, y(1) in presence of treatment
  - In reality we will only observe one of the outcomes.
- If the effect of the intervention is constant, say τ, we have for any observation *i*:

$$y_i(1) = \tau + y_i(0)$$

- More generally, if $\mathcal{T}$ is not constant, we consider the average treatment effect:

$$\tau_{ATE} = E[y_i(1) - y_i(0)]$$

where the expectation is over the entire population.

- For the i'th observation, the observed outcome y_i can be written

$$y_i = (1 - w_i)y_i(0) + w_i y_i(1)$$

- Remark: If we regress y on w, the resulting OLS estimate is only unbiased for τ_{ATE} if we have random assignment of *w*, that is *w* is independent of [y(0),y(1)] conditional on ***x*** *(*if there are further **x** in the model*):*

$$y_i = \tau_{ATE} w_i (+\mathbf{x}_i \beta) + u_i$$

  The point is here that for all values of *w*, it must be unrelated with *u*, so *w* is not permitted to be related with y(0),y(1) (conditional on ***x***).
- Random assignment is rare in economic and business data but can be ensured by doing randomized control trials (RCT).
  - In presence of a RCT, the econometrics becomes easy!
  - By designing policy changes or interventions in a clever way it is easy to learn something about their effects.

$$y_i = \tau_{ATE} w_i + \mathbf{x}_i \beta + u_i$$

- Random assignment conditional on ***x*** is more likely to hold the more variables there are in the data.
  - The more we take out of *u*, the less likely it will contain something that is correlated with *w* conditional on ***x***.
- Random is also called unconfounded assignment or ignorable assignment.
- Unfortunately, in most empirical situation there is no random assignment as the decision to introduce an intervention is related to outcomes.
  - E.g. individuals with poor labour market performance receive treatment. Hopefully, this can be mitigated by using a rich conditioning set.

- We show how random assignment conditional on ***x*** leads to unbiased estimation.
- We use $y = (1 - w)y(0) + wy(1)$ and let

$$E[y|w, x] = \alpha + \tau w + \mathbf{x}\gamma$$

- To make the point, we consider models with covariates centered about their means $\eta_j = E(x_j)$, where we consider two separate equations for the two outcomes:

$$\begin{aligned} y(0) &= \phi_0 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma}_0 + u(0) \\ y(1) &= \phi_1 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma}_1 + u(1) \end{aligned}$$

with $\phi_0 = E(y(0))$ and $\phi_1 = E(y(1))$.

- The treatment effect for observation *i* is

$$\begin{aligned} te_i &= y_i(1) - y_i(0) \\ &= (\phi_1 - \phi_0) + (\mathbf{x}_i - \boldsymbol{\eta})(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + [u_i(1) - u_i(0)] \end{aligned}$$

  which is a function of ***x*** and the error terms.

- The average treatment effect is

$$\begin{aligned} E(te_i) &= \tau_{ATE} \\ &= (\phi_1 - \phi_0) + E\{(\mathbf{x}_i - \boldsymbol{\eta})(\boldsymbol{\gamma}_1) - \boldsymbol{\gamma}_0) + [u_i(1) - u_i(0)]\} \\ &= (\phi_1 - \phi_0) + \mathbf{0}(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + 0 \\ &= \phi_1 - \phi_0 \end{aligned}$$

- The observed outcome $y_i = (1 - w_i)y_i(0) + w_i y_i(1)$ can be written as

$$\begin{aligned} y_i &= \phi_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\eta})\boldsymbol{\gamma}_0 + w_i(\mathbf{x}_i - \boldsymbol{\eta})\boldsymbol{\delta} \\ &\quad + u(0) + w_i[u_i(1) - u_i(0)] \end{aligned}$$

where $\boldsymbol{\delta} = (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0)$.

- The realised (but unobserved error) is

$$u_i = u_i(0) + w_i[u_i(1) - u_i(0)]$$

- Random assignment or unconfoundedness implies:

$$\begin{aligned} E(u_i|w_i, \mathbf{x_i}) &= E[u_i(0)|w_i, \mathbf{x}_i] + w_i E\{[u_i(1) - u_i(0)]|w_i, \mathbf{x}_i\} \\ &= E[u_i(0)|\mathbf{x}_i] + w_i E\{[u_i(1) - u_i(0)]|\mathbf{x}_i\} \\ &= 0 \end{aligned}$$

This means the error is conditional zero mean if treatment assignment is random cond. on ***x***.

- In practice, regress:

$$y_i \text{ on } w_i, \mathbf{x}_i, w_i(\mathbf{x}_i - \bar{\mathbf{x}}_i)$$

The coefficient on w_i is $\hat{\tau}_{ATE}$.

- The $\hat{\tau}_{ATE}$ obtained is typically different from the one obtained from the restricted regression of

$$y_i \text{ on } w_i, x_i$$

- The latter approach is called restricted regression adjustment (RRA), while the unrestricted model is unrestricted regression adjustment (URA).
- Without showing this, $\hat{\tau}_{ATE}$ for the unrestricted model can be obtained from two separate regressions, which is typically more convenient:
  - Split the sample by $w_i = 0$ and $w_i = 1$. Run the regression $y_i \text{ on } \mathbf{x}_i$ for each sample.
  - Compute fitted values $\hat{y}_i^{(1)}$ and $\hat{y}_i^{(0)}$ and $\hat{\tau}_{ATE} = n^{-1} \sum_i [\hat{y}_i^{(1)} - \hat{y}_i^{(0)}]$
- Example: tau_ATE.R

Policy Analysis with Pooled Cross Sections

- If cross sectional data is collected before and after an event, it can be used to determine the effect on economic outcomes.
- Example: Effect of Garbage Incinerator's Location on House Prices
  - New incinerator in North Andover, Massachusetts
  - Construction started in 1981, operation in 1985
  - Rumors that it will be built began in 1978
  - Data about prices of houses sold in 1978 and 1981 (KIELMC.dta), prices in 1978.

  - The hypothesis is that house prices near the incinerator will drop relative to prices of more distant houses.

- Example: cont.
  - A naïve approach to estimate the effect of the incinerator would be:

$$rprice = \gamma_0 + \gamma_1 nearinc + u$$

    where *nearinc* is 1 if the house is within 3 miles from the incinerator.
  - When we estimate it for the 1981 data:

$$\begin{aligned} \widehat{rprice} &= 101,307.5 - 30,688.27 nearinc \\ &\quad (3,093.0) \quad (5,827.71) \\ &\quad n = 142, \quad R^2 = 0.165 \end{aligned}$$

  - The intercept is the average selling price for homes not near the incinerator
  - Average selling price for home near the incinerator is $30,688 lower. Statistically significant.
  - Is this already evidence for an effect of the incinerator?

- Not really! When we repeat the analysis for the 1978 data, we obtain:

$$\begin{aligned} \widehat{rprice} &= 82,517.23 - 18,824.37 nearinc \\ & \quad (2,653.79) \quad (5,827.71) \\ & \quad n = 179, \quad R^2 = 0.082 \end{aligned}$$

- Home prices were already statistically significant lower around the future location of the incinerator before the planning process was started.
- How can we then assess whether there was a decrease in response to the construction of the incinerator?
- Simply take the difference in the two coefficients on *nearinc*:

$$\begin{aligned} \hat{\delta}_1 &= -30,688.27 - (-18,824.37) = -11,863.9 \\ &= (\overline{rprice}_{81,n} - \overline{rprice}_{81,f}) - (\overline{rprice}_{78,n} - \overline{rprice}_{78,f}) \end{aligned}$$

is an estimate of the effect of the incinerator on values of homes, where _*n* is near and _*f* is farther away from the site.

$$\hat{\delta}_1 = (\overline{rprice}_{81,n} - \overline{rprice}_{81,f}) - (\overline{rprice}_{78,n} - \overline{rprice}_{78,f})$$

- This is known as the difference-in-differences (DID) estimator.
- It is the difference over time in the differences of housing prices in the two locations.
- We can directly estimate $\hat{\delta}_1$ and its standard error by using regression analysis:

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 * nearinc + u$$

- β_0 is the average home price not near the site in 1978.
- δ_0 captures changes in prices not near between 1978 and 1981
- β_1 measures the location effect that is not due to the incinerator.
- δ_1 measures the decline in home prices due to the new incinerator, provided we assume everything else equal.

- Example (cont.) Estimation results for three sets of regressors:

Dependent Variable: rprice			
Independent Variable	*(1)*	*(2)*	*(3)*
Constant	*82,517.23* *(2,726.91)*	*89,116.54* *(2,406.05)*	13,807.67 (11,166.59)
y81	*18,790.29* *(4,050.07)*	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	-18,824.37 (4,875.32)	*9,397.94* *(4,812.22)*	3,780.34 (4,453.42)
*y81*nearinc*	-11,863.90 (7,456.65)	-21,920.27 (6,359.75)	-14,177.93 (4,987.27)
Other controls	No	Age, age^2	Full Set
Observations *R-squared*	321 0.174	321 0.414	321 0.660

  - DID estimator only marginally significant in (1), magnitude depends on model specification.

- In the previous example it would have been better to use the log of price as dependent variable. Then the DID estimator tells us the percentage change in prices in response to the new site.
- A DID estimator can be applied if the data arises from a natural experiment (or quasi experiment):
  - An exogenous event (e.g. change of policy) changes the environment or labour market outcome.
  - A natural experiment has always a control group (C) and always a treatment group (T).
  - We need (at least) two periods of data, one before (T0) and one after the policy change (T1).
  - Denote d2 as the dummy for the second period and dT as the dummy for the treatment group, then the following regression directly estimates the DID treatment effect:

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 * dT + other\ \ factors$$

  - Also called the average treatment effect as it measures the effect on the average outcome of y.

Summary

- Multiple regression analysis is frequently used as a tool to analyse the effect of a policy change or intervention on an outcome.
- Under some restrictions it is possible to estimate the average treatment effect.
- Knowledge about how to estimate treatment effects is also useful for designing the treatment or change in a way that it is possible to empirically assess its implications.

Endogeneity

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

- Text: Wooldridge (2010), Econometric Analysis of Cross Section and Panel Data, Chapter 5, parts of Chapter 6
- Less formal: Several (sub) chapters of Wooldridge (2025), Introductory Econometrics, 8th Edition, Cengage, mainly Chapter 15

- Suppose we have a linear regression model:

$$y = \mathbf{x}\beta + u$$

- Definition: Exogeneity and Endogeneity of Independent Variables.
  - x_j is exogenous if it is uncorrelated with *u*.
  - x_j is endogenous if it is correlated with *u*.

- OLS estimation of the linear regression model requires exogeneity of x_j .

- Endogeneity can be caused by many things.
  - An important variable that is not observed and omitted
  - Functional form misspecification
  - Simultaneity
  - Measurement error in the regressors
  - ...
- Endogeneity is present in most applications in applied economic research.

Omitted Variable Bias

- What happens if we omit variables that actually belong in the true model?

  Let $K_1 + K_2 = K$ with $1 \leq K_2 < K$ and

$$y = \mathbf{x}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\boldsymbol{\beta}_2 + u$$

  [Full regression: $y = \mathbf{x}\beta + u$]

  Regress y on $x_1, ..., x_{K_1}$ only: $y = \mathbf{x}_1\beta_1 + v$

  The estimator is:

$$\begin{aligned}\hat{\beta}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'y \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + u) \\ &= \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'u\end{aligned}$$

- Then because of $E(\mathbf{x}_1' u) = 0$:

$$E(\hat{\beta}_1 | \mathbf{X}) = \beta_1 + (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2 \beta_2$$

there is an omitted variable bias if:

- $\mathbf{x}_1' \mathbf{x}_2 \neq 0$, i.e. the two regressors sets are not orthogonal.
- $\beta_2 \neq \mathbf{0}$, i.e. the omitted variables play a role.

- The magnitude of the bias depends on the magnitude of the elements of β_2 and on how strongly the independent variables are correlated.

- Solutions:

  Instrumental Variables, Proxy Variables, Panel Data

Using a Proxy Variable for Unobserved Explanatory Variables

- A more difficult problem arises when a model excludes a key variable, usually because of data unavailability.
- Example: Return to Education
  - the population model is:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

  - Suppose we do not observe the ability. Ignoring *abil* would generally give biased and inconsistent estimates of the return to education.
    - We expect an upward bias for the estimated return to education. Why?
  - How can we solve or at least mitigate this omitted variable problem?

- One possibility is to use a proxy variable for the omitted variable.
  - Something that is related to the unobserved variable.

- In the wage equation one could use the intelligence quotient, or IQ as a proxy for ability. IQ and ability do not need to be the same, but they need to be correlated.
- Suppose we have the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

with x_3^* being unobserved. We have a proxy variable, which we call x_3

- What do we require of x_3?
  - When we would run the regression $x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$, we should obtain $\delta_3 > 0$. Otherwise the proxy is not good.

- The proposal is just to regress *y* on x_1, x_2, x_3 as x_3 and x_3^* would be the same. This is called the **plug-in solution** to the omitted variables problem.
- Since x_3 and x_3^* are not the same: when this procedure does in fact give consistent estimators for β_1 and β_2?

- The assumptions are with respect to *u* and ν_3 :

1. In addition to assuming that u and x_1, x_2, x_3^* are uncorrelated, we need that u and x_3 are uncorrelated. This means that x_3 is irrelevant in the population model once x_1, x_2, x_3^* are included.

2. The error ν_3 is uncorrelated with x_1, x_2 and x_3. This means that x_3 is a good proxy for x_3^*: $E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3)$

- From the latter assumption follows, that
  $E(x_3^*|x_3) = \delta_0 + \delta_3 x_3$.

- In terms of our wage equation this means:
  $$E(abil|educ, exper, IQ) = E(abil|IQ) = \delta_0 + \delta_3 IQ$$
  thus the average value of ability only changes with *IQ*.

- More formally, what are the implications of the two assumptions?

- By combining

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$
$$x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$$

  we obtain:

$$y = (\beta_0 + \beta_3 \delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \delta_3 x_3 + u + \beta_3 \nu_3$$

  - now let us denote $e = u + \beta_3 \nu_3$ as the composite error.
  - And note that *u* and ν_3 have both zero mean and each is uncorrelated with x_1, x_2 and x_3. Then *e* has also zero mean and is uncorrelated with x_1, x_2 and x_3 .

- For this reason, we can write

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

  - This gives unbiased (or consistent) estimates of $\alpha_0, \beta_1, \beta_2$ and α_3
  - We do not get unbiased estimates for β_0, β_3.
  - In an application α_3 may even be of more interest than β_3.

- Example: Return to education
  - Wage2.dta
  - We estimate a log wage equation without *IQ* (1) and with *IQ* (2).
  - Our primary interest is in what happens to the estimated return to education.

	log(wage)	
Indep. Variables	*(1)*	(2)
educ	0.065 (0.006)	0.054 (0.007)
exper	0.014 (0.003)	0.014 (0.003)
tenure	0.012 (0.002)	0.011 (0.002)
…	…	…
IQ	-	0.0036 (0.0010)
intercept	5.395 (0.113)	5.176 (0.128)
Observations *R-Squared*	935 0.253	935 0.263

- In model (1) the estimated return to education is 6.5%, while in model (2) it is just 5.4%. This corresponds to our beliefs about omitted variable bias.

- In particular the estimate decreases but it is still large.

- In the data wage2.dta there are also other measures of ability, such as *Knowledge of the World of Work (KWW)* test.

Functional form misspecification

- Special case: omission of a relevant variable x_1^2.
- Suppose $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$

  $y = \beta_0 + \beta_1 x_1 + \tilde{u}$

  with $E(u|x_1, x_1^2) = 0$ and $\tilde{u} = \beta_2 x_1^2 + u$.

- Now, since $cov(x_1, x_1^2) \neq 0$ and if $\beta_2 \neq 0$,

  we do not have $E(\tilde{u}|x_1) = 0$ and we would have a bias due to functional form misspecification.
- Solution: Test for functional form (RESET), Non- and semiparametric methods.

RESET test for model specification

$$y = \mathbf{x}\beta + u$$

- How do we know whether we have assumed the correct functional form?
  - For example: have we included all relevant quadratics and interaction terms?
- By noting that y^2 and y^3 are highly nonlinear functions of all regressors and their interactions, we could use the fitted values of the model above to compute $\hat{y}^2$ and $\hat{y}^3$.
- Then we estimate

$$y = \mathbf{x}\beta + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$$

and perform an F-test for joint significance of $\hat{y}^2$ and $\hat{y}^3$:

$$H_0 : \delta_1 = \delta_2 = 0$$

Simultaneity

- If an explanatory variable is determined simultaneously with the dependent variable, it is generally correlated with the error terms.
- In this case OLS is biased and inconsistent.
- Will be done in Part B2 “Simultaneous Equation Models” of the course.

Measurement error in an explanatory variable

- We consider the simple regression model:

$$y = \beta_0 + \beta_1 x_1^* + u$$

  and assume that it satisfies the Gauss Markov assumptions.
- We do not observe x_1^* but x_1 (e.g. actual and reported income).
- The measurement error in the population is: $e_1 = x_1 - x_1^*$
- We assume: $E(e_1) = 0$
- Moreover, we assume that *u* is uncorrelated with x_1 and x_1^* :

$$E(y|x_1, x_1^*) = E(y|x_1^*)$$

- The model can be written as: $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$

- The **classical errors-in-variables** (**CEV**) assumption is that the measurement error is uncorrelated with the unobserved explanatory variable: $cov(x_1^*, e_1) = 0$

  - This has the meaning that the observed measure x_1 consists of two uncorrelated components: $x_1 = x_1^* + e_1$
  - (We still assume that *u* is uncorrelated with x_1 and x_1^*.)
  - The above assumption implies that x_1 and e_1 must be correlated:

$$cov(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$

  - This correlation causes problems for the OLS estimation.

- This implies for our model $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$ that since *u* and x_1 are uncorrelated, the covariance between x_1 and the composite error $u - \beta_1 e_1$ is:

$$cov(x_1, u - \beta_1 e_1) = -\beta_1 cov(x_1, e_1) = -\beta_1 \sigma^2_{e_1}$$

- Note also that $var(x_1) = var(x_1^*) + var(e_1)$
- Then one can show:

$$\begin{aligned} plim(\hat{\beta}_1) &= \beta_1 + \frac{cov(x_1, u - \beta_1 e_1)}{var(x_1)} \\ &= \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} = \beta_1 \left(1 - \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} \right) \\ &= \beta_1 \left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} \right) \end{aligned}$$

- This equation is very interesting: $plim(\hat{\beta}_1)$ is always closer to zero than β_1: **attenuation bias**

- OLS is biased in the classical error in variables model:
  - If β_1 is positive, it will underestimate β_1 and vice versa.
- Things are more complicated if we look at the multiple regression model but again OLS will be biased and inconsistent.

- Solution: Instrumental Variable estimation, ...

IV Estimation of the Multiple Regression Model

- The model is: $y = \mathbf{x}\beta + u$
  with $E(u) = 0$, β is *Kx1* and $\mathbf{x} = (1, x_2, ..., x_K)$ and x_K is endogenous with $cov(x_K, u) \neq 0$.
- We call the above equation **structural equation** as we are interested in the coefficients.
- We will use an instrument for x_K to obtain consistent estimates.
- We need another exogenous variable z_1 with
$$cov(z_1, u) = 0 \ .$$
- Then $E(\mathbf{z}'u) = \mathbf{0}$ with $\mathbf{z} = (1, x_2, ..., x_{K-1}, z_1)$.

- A variable z_1 is a candidate for an instrument for a variable x_K if it satisfies:

$$cov(z_1, u) = 0$$

- Some remarks on the choice of an instrument:
  - It is often difficult to find a good instrument.
  - A proxy variable does not make a good instrument as it is supposed to be correlated with the error term.
  - Example: Ability is not observed and IQ is highly correlated with ability. Then it is a candidate for a proxy but clearly violates the condition.
  - If one is not sure about the quality of an instrument, it may be better to use a proxy variable (if available).

- The **reduced form equation** is

$$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K$$

with $E(r_K) = 0$ and r_K is uncorrelated with all exogenous variables.

- Rank condition: We require a non-zero partial correlation between x_K and z_1: $\theta_1 \neq 0$. (use t-test to check this)
- This means x_K and z_1 need to be partially related.
- This implies $\text{rank } E(\mathbf{z}'\mathbf{x}) = K$ which makes it invertible.
- We obtain:

$$\begin{aligned} y &= \mathbf{x}\beta + u \\ \mathbf{z}'y &= \mathbf{z}'\mathbf{x}\beta + \mathbf{z}'u \\ E[\mathbf{z}'y] &= E[\mathbf{z}'\mathbf{x}\beta] + E[\mathbf{z}'u] \\ E[\mathbf{z}'y] &= E[\mathbf{z}'\mathbf{x}]\beta \\ \beta &= [E(\mathbf{z}'\mathbf{x})]^{-1} E(\mathbf{z}'y) \end{aligned}$$

- $E[\mathbf{z}'\mathbf{x}]$ and $E[\mathbf{z}'y]$ can be consistently estimated.

- Given a random sample *i=1,....,N* the instrumental variables estimator of β is

$$\begin{aligned}\hat{\beta} &= \left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'y_i\right) \\ &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}\end{aligned}$$

with ***Z*** and ***X*** are *N x K* and ***Y*** is *N x 1*.

$\mathbf{x}_i$ and $\mathbf{z}_i$ are the *i*'th row of ***X*** and ***Z*** respectively.

- Example: College Proximity as IV for Education
  - Data: Card.dta
  - Log(wage) is dependent variable, several controls
  - Instrument for education: dummy if someone grew up near a four year old college (*nearc4*).
  - We assume that *nearc4* is uncorrelated with the error. Moreover, to be a valid instrument it has to be partially correlated with *educ.*
  - We can test this by estimating the reduced form equation:

$$\begin{aligned} \widehat{educ} &= 16.64 + 0.320 nearc4 + ... \\ &\quad (0.24) \quad (0.088) \\ &\quad n = 3,010, \quad R^2 = 0.477 \end{aligned}$$

  - The *t*-statistic is 3.64 and therefore if *nearc4* is uncorrelated with the error term, we can use it as IV for *educ*.

- The following table reports OLS and IV estimates.

Dependent Variable: log(wage)		
Independent Variable	*(1) OLS*	*(2) IV*
Educ	*0.075* *(0.003)*	*0.132* *(0.055)*
Exper	*0.085* *(0.007)*	0.108 (0.024)
Exper^2	-0.0023 (0.0003)	*-0.0023* *(0.0003)*
…other controls	…	…
Observations *R-squared*	3,010 0.300	3,010 0.238

- IV estimate is almost twice as large as the OLS estimate.
- SE of the IV estimate is 18 times larger. This is the price we have to pay if we use an instrument to obtain a consistent estimator.

Two Stage Least Squares (2SLS)

- Sometimes there are multiple valid IVs for an endogenous explanatory variable.
- Suppose the variables $z_1, \ldots, z_M$ satisfy

$$\text{cov}(z_h, u) = 0 \text{ for } h = 1, ..., M$$

- We could simply use both of them as instruments and obtain multiple IV estimators.
- The idea is to use both together to obtain a more efficient estimator:
  - Let $\mathbf{z} = (1, x_2, ..., x_{K-1}, z_1, ..., z_M)$ be *1x L* with *L=K-1+M*.
  - As each element of ***z*** is uncorrelated with *u*, any linear combination is also uncorrelated with *u*.

- The linear combination of ***z*** which is most highly correlated with x_K is the linear projection of x_K on ***z***. The reduced form equation is

  $$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M z_M + r_K$$

  where $E(r_K) = 0$ and r_K is uncorrelated with all elements of ***z***.
- r_K is correlated with *u* if x_K is endogenous.
- x_K^* is not correlated with *u* with

  $$x_K^* = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M z_M$$
- Consistently estimate the parameters by OLS and use fitted values as an estimator for x_{iK}^*

  $$\hat{x}_{iK} = \hat{\delta}_1 + \hat{\delta}_2 x_{i2} + ... + \hat{\delta}_{K-1} x_{i,K-1} + \hat{\theta}_1 z_{i1} + ... + \hat{\theta}_M z_{iM}$$
- We require that at least one θ_j is non-zero. Use F-test.

- Now, let $\hat{\mathbf{x}}_i = (1, x_{i1}, ..., x_{i,K-1}, \hat{x}_{iK})$ and use it as the instruments for $\mathbf{x}_i$:

$$\begin{aligned} \hat{\beta} &= \left(N^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i' y_i \right) \\ &= (\hat{\mathbf{X}}'\mathbf{X})^{-1} \hat{\mathbf{X}}'\mathbf{y} \end{aligned}$$

- It can be shown that $\hat{\mathbf{X}}'\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$ and thus

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

- This estimator is consistent under the conditions:

$$E(\mathbf{z}'u) = \mathbf{0} \quad , \text{rank } E(\mathbf{z}'\mathbf{x}) = K \, , \text{ rank } E(\mathbf{z}'\mathbf{z}) = L \, , \; L \geq K$$

- The last condition suggests that we need at least as many instruments as explanatory variables in the model (**order condition**)

- By noting $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ the 2SLS estimator can be written as:

$$
\begin{aligned}
\hat{\beta} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\
&= \left[\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1} \\
&\quad \times \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{z}_i'y_i\right)
\end{aligned}
$$

(AB)'=B'A'

  - One step estimator.
  - In the case of *L*=*K* (just identified): Replace $\hat{\mathbf{X}} = \mathbf{Z}$ in the equation for $\hat{\beta}$.

- By using $y_i = \mathbf{x}_i\beta + u_i$, we obtain:

$$\hat{\beta} = \beta + \left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1}$$
$$\times\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'u_i\right)$$

  - By application of a law of large numbers to the various terms and the Slutsky theorem, we obtain consistency provided $E(\mathbf{z}_i'u_i) = 0$.

- Under homoscedasticity $E(u^2\mathbf{z}'\mathbf{z}) = \sigma^2 E(\mathbf{z}'\mathbf{z})$, which is slightly weaker than RW5 $E(u^2|\mathbf{z}) = \sigma^2$, it is possible to show that asymptotically:

$$\sqrt{N}(\hat{\beta} - \beta) \to N\left(\mathbf{0}, \sigma^2([E(\mathbf{x}'\mathbf{z})][E(\mathbf{z}'\mathbf{z})]^{-1}E(\mathbf{z}'\mathbf{x})])^{-1}\right)$$

  - The more unrelated (orthogonal) ***x*** and ***z***, the smaller is $E(\mathbf{x}'\mathbf{z})$, and the larger the variance of $\hat{\beta}$ becomes.

- Under homoskedasticity the 2SLS estimator is asymptotically efficient.

- The variance matrix can be estimated by $\hat{\sigma}^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$ with

$$\hat{\sigma}^2 = (N-K)^{-1}\sum_{i=1}^{N}\hat{u}_i^2 = (N-K)^{-1}\sum_{i=1}^{N}(y_i - \mathbf{x}_i\hat{\beta})^2$$

  - The more instruments we use the more variation we will have in $\hat{\mathbf{X}}$ and the smaller the variance of $\hat{\beta}$.

RW5 it is equivalent to u^2 is uncorrelated with all x_j, x_j^2 and cross products x_j*x_k

Ralf Wilke; 12-03-2018

- The IV estimator with multiple instruments is called two stage least squares (2SLS) estimator:
  - One can show that the IV estimates are identical to OLS estimates from the regression of y on $1, x_2, ..., x_{K-1}$ and $\hat{x}_K$. This is the second stage.
  - The first stage is the regression of

$$x_K \text{ on } 1, x_2, ..., x_{K-1}, z_1, ..., z_M.$$

- 2SLS standard errors are larger than for OLS. This is because:
  - $\hat{x}_K$ has less variation than x_K.
  - $\hat{x}_K$ has more correlation with $x_2, ...x_{K-1}$ than x_K. (multicollinearity in the second stage)

IV Estimation with a poor Instrumental Variable

- Simple regression: $y = \beta_0 + \beta_1 x + u$, instrument: *z*
- IV estimates can have large standard errors if *x* and *z* are only weakly correlated. (Don't use IV in this case.)
- IV estimates can have a large asymptotic bias if *z* and *u* are only weakly correlated:
  - For illustration: model with one regressor (*x*) and one instrument (*z*):

$$plim\hat{\beta}_1 = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)}\frac{\sigma_u}{\sigma_x}$$

  This implies that the bias can be large if the population correlation between *z* and *x* is small even if the population correlation between *z* and *u* is small.
- For this reason IV can be worse in terms of consistency than OLS even if *Corr(z,u)* is small (provided that *Corr(z,x)* is also small).
- One can show that IV is only superior in terms of asymptotic bias if

$$Corr(z,u)/Corr(z,x) < Corr(x,u)$$

R-Squared and IV Estimation

$$R^2 = 1 - SSR/SST$$

- SSR (sum of squared IV residuals) can be larger than SST. For this reason the R-squared can become negative and it is smaller than for OLS.
- It is not clear whether the IV R-squared should be reported after IV estimation.
- If you try to maximise the R-squared, use OLS as IV tries to improve the quality of ceteris paribus effects.

Some Remarks:

- If the R-squared of $\hat{x}_K$ on the exogenous variables appearing in the structural equation (without instruments) is very large, the standard error of 2SLS explodes. Can be verified with data at hand.
- 2SLS can be also used in models with more than one endogenous variable.
  - We need more candidates for instruments to achieve identification.
  - The sufficient condition for identification is the **rank condition**.
- Since R-squared after 2SLS cannot be compared to OLS R-squared we must be careful when using the F-test.
- It is possible to derive a statistic with an approximate F-distribution in large samples. Use econometric packages to test multiple hypothesis after 2SLS as commands are available.

IV Solutions to Errors in Variables Problems

- Suppose we have the model

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u$$

with x_1^* is not observed but $x_1 = x_1^* + e_1$, with e_1 is the measurement error.

  - x_1 and e_1 are correlated and therefore OLS when regressing *y* on x_1 and x_2 is biased and inconsistent.

- In the case of the classical errors-in-variable model, we have seen that the OLS estimator is biased towards zero.
- It is possible to use an IV procedure to overcome the measurement error problem.

- After plugging in the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1)$$

- We assume that e_1 is uncorrelated with x_1^* and x_2.
- We also assume that *u* is uncorrelated with x_1, x_1^* and x_2.
- Therefore x_2 is exogenous, but x_1 is correlated with e_1.
- We need an instrument for x_1 that is correlated with x_1, but uncorrelated with u and e_1.
  - One possibility could be a second measurement of x_1^*:

    $$z_1 = x_1^* + a_1$$

    where we need that a_1 is uncorrelated with e_1 and *u*. This means the two measurement errors need to be uncorrelated.
  - Another possibility is to use another exogenous variable as IV for x_1 as with the usual IV procedure.

- Example: Wage regression with two erroneous measures of ability. (wage2.dta,IV_ErrorsInVariables.R)
- Continued example from proxy variable model.
- We use *IQ* as a mismeasured observed variable for ability.
  - But now *IQ* is endogenous. Given that *IQ* is correlated with *educ,* the estimate for the return to education might be biased as well.
- We use *KWW* as the IV for IQ.
  - KWW is another mismeasured ability variable.
- Resulting IV estimate for *educ* is smaller and insignificant.
  - Statistically not different from OLS estimate.
  - Large standard errors due to multicollinearity in second stage regression.

Testing for Endogeneity

- 2SLS is less efficient than OLS and can have large standard errors.
- It is therefore useful to have a test for endogeneity of explanatory variables to show whether 2SLS is even necessary:

1. Regression based test (Hausman, 1978)

2. Durbin, Wu and Hausman suggest a test which directly compares OLS and 2SLS estimates and determines whether differences are statistically significant (DWH Test):

H0: all regressors are exogeneous

Regression based test (Hausman, 1978)

- Suppose we have the structural equation

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

with y_2 being endogenous and there are two exogenous variables z_3, z_4 which are not included in the model.

- The idea behind the test is as follows:
  - We have:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

    and

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + \nu_2$$

  - Each z_j is uncorrelated with u_1.
  - y_2 is uncorrelated with u_1 if and only if, ν_2 is uncorrelated with u_1 and has zero mean. This is what we want to test.
  - Write $u_1 = \delta_1 \nu_2 + e_1$, where e_1 is uncorrelated with ν_2 and $E(e_1) = 0$.
  - Then u_1 and ν_2 are uncorrelated if and only if $\delta_1 = 0$.
  - Simply plug this into the structural equation and do a *t* test.
  - Since ν_2 is not observed, use instead the residuals from the reduced form equation as a regressor:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{\nu}_2 + error$$

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{\nu}_2 + error$$

- We then test: $H_0 : \delta_1 = 0$ using a t-test.
- if we reject it at a small significance level, we conclude y_2 is endogenous because u_1 and ν_2 are correlated.

- Practical guideline for the **Hausman test**:

1. Estimate the reduced form for y_2 and obtain $\hat{\nu}_2$.
2. Add $\hat{\nu}_2$ to the structural regression and estimate it by OLS. You may want to use a heteroscedasticity robust version of the t-test for testing whether the coefficient on $\hat{\nu}_2$ is significant. If it is statistically significant from zero, we conclude that y_2 is indeed endogenous.

- This test requires the availability of valid instruments.

- It can be easily extended to the case of multiple endogenous variables:
  - The reduced form of step 1 is then estimated for each endogenous regressor.
  - The regression in step 2 then includes the residuals obtained by all regressions of step 1. The test is then to test for joint significance of all residuals using F- or LM test in this regression.

Durbin-Wu-Hausman (DWH) test

- If under the H0 all regressors are exogenous but some are endogenous under H1, we can base a test directly on the difference between 2SLS and OLS estimators.

- The DWH statistic

$$(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})/\hat{\sigma}^2_{OLS}$$

  is asymptotically $\chi^2_{(G_1)}$ distributed with G_1 is the number of endogenous regressors and $\mathbf{H}^{-}$ is a generalised RW6 inverse of **H**.
- Statistic may be cumbersome to compute. Wald test.

RW6 Generalised inverse A^- is such that: A A^- A=A.
A useful concept in the case the regular inverve A^-1 does not exist. If A^-1 exists it is the unique generalised inverve A^-.
Ralf Wilke; 12-03-2018

Testing Overidentifying restrictions

- If we have more instruments for one endogenous explanatory variable, we can test whether at least some of them are not correlated with u_1 (validity of the instrument). We need that at least one of the IVs is exogenous and we need to know which one.
- Then we can test the overidentifying restrictions that are used in 2SLS:
  - E(**z**'u)=**0** is Lx1
  - Suppose we estimate the same model by 2SLS but with a different number of instruments. Say model 1 is just identified and model 2 is overidentified. Then L in the second model is bigger than in the first and thus we have imposed additional moment conditions. These conditions can be tested.
  - Under H0: E(**z**'u)=**0** is true for model 2.
- Regression based version (LM test): ***Sargan Test***

More remarks on IV estimation:

- Heteroscedasticity in the context of 2SLS raises the same issues as with OLS.
- There are standard errors and test statistics available which are robust with respect to heteroscedasticity. R: iv_robust in `estimatr`
- There are also tests for heteroscedasticity available.

- Lasso IV variable Selection for IV estimation:

Belloni et al. (2012) "Sparse models and methods for optimal instruments with an application to eminent domain." Econometrica 80: 2369-2429.

- In more general regression models, IV models are not identified (set estimation).
  - Due to support restrictions on the dependent variable, ordered data, interval censoring of regressors etc.
  - Compare Chesher/Rosen (2017), Generalised Instrumental Variable Models, Econometrica, 959-989 or Chesher/Rosen (2013), What Do Instrumental Variable Models Deliver with Discrete Dependent Variables? American Economic Review.
  - Check for potential issues before you apply IV methods outside the standard linear regression world.

Summary

- We have seen the method of instrumental variables as a way to consistently estimate the parameters in a linear model when there are endogenous explanatory variables.
- When instruments are poor, IV estimates can be worse than OLS.
- 2SLS is routinely used in economics and social sciences alike.
- Tests for endogeneity.