Linear Regression with Stata
Script

October 7, 2019

Lecturer

Paul Reimers
Office: HoF 3.42
E-mail: Reimers@econ.uni-frankfurt.de
Goethe University Frankfurt & GSEFM

Organization

Dates: October 2nd, 4th, 9th and 10th, Times: 9am - 12pm (sharp) and 1pm - 3pm.
Location: Seminarhaus 4.108, Campus Westend

Contents

I Introduction

Regression Analysis is one of the most heavily used statistical methods in social sciences. In empirical economics, we try to validate specific theories. Or, we try to understand and quantify causal relationships, in order to be able to predict economic outcomes & forecast future trends. Typically, we analyze the relationship between a dependent and an independent, or explanatory variable. That is, we impose some kind of causal relationship between the variables of interest. For example,

1. *Does higher education lead to higher wages?*
2. *What are the effects of labor tax increases on labor supply?*

The problem is that just because the variables of interest **correlate**, this does not imply **causality**! For example, higher education might be positively correlated with wages. But it is also positively correlated with ability, and ability is positively correlated with wages. So can we really deem only education to be causal on wages, or are other factors such as ability **confounding** our results? Empirical researchers spend a lot of effort in ensuring that they capture causal links. This effort goes into ensuring that

1. The data used is adequate:
   - Best case: Use of *population data*. Typically: Use of a *subsample*
   - Main assumption: ***Random Sampling/Assignment***
   - Ensure that data is representative, no selection
2. The ***identification strategy*** is adequate
   - From theory or by assumption, impose a *population model*
   - Does the specification **omit** important variables?
   - Does the specification suffer from **endogeneity** (simultaneity/reverse causality)?

These are the building blocks for all kinds of regression exercises. If you cannot ensure that the data you use is really randomly sampled and that your identification strategy is valid, then the causal links that you estimate might be **biased** and/or **inconsistent**. In small and/or large samples, you might then not estimate the correct (or true) causal links. In practice, you can basically always find limitations to the data that is used or think of reasons why an identification strategy might not capture what is going on in reality. This might sound somewhat discouraging. But it shouldn't, because econometricians have developed many ways to circumvent problematic data or identification strategies. Eventually, regression analysis has many good properties, even when the above assumptions cannot all be maintained. Plus, it is a fairly simple tool that yields easy-to-interpret, tractable results. This is why it is so useful and the main statistical method used in economics, or the social sciences in general.

A The simple OLS model

Let's start with the simplest case for linear regression models. We assume that in the population a dependent variable y is related to an explanatory variable x according to the linear specification

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

where β_0 is generally referred to as the *intercept* or the *constant*, β_1 is the *coefficient* on the explanatory variable x or the *slope parameter*. We are interested in estimating this β_1, i.e. how

an increase in x by one unit affects y. The term u is an *error term* that captures all kinds of other factors that might be relevant for y.

Based on Equation (1), imagine we draw a sample from the population: We **randomly** sample $N = 1000$ observations. In this data sample, we denote each observation using an index $i = 1, 2, ...N$. We now rewrite Equation (1) for each individual as

$$y_i = \beta_0 + \beta_1 x_i + u_i, \tag{2}$$

In Figure (1), I give you a glance what such a sample could look like. Now using our data sample, we can estimate the intercept β_0 and the coefficient β_1 using **Ordinary Least Squares**. These sample OLS-estimates that we retrieve are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$.

The OLS method chooses the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the distance between the observed data y_i and the **fitted value** $\hat{y}_i$

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3a}$$

The **fitted value** is defined as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3b}$$

The **residual** is defined as

$$\hat{u}_i = y_i - \hat{y}_i \tag{3c}$$

That is, equivalently to Equation (3a) you can think of OLS as **minimizing the sum of squared residuals (SSR):**

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} \hat{u}_i^2 \tag{3d}$$

Take the first derivatives on the objective (Eq. 3a)

$$\frac{\partial(.)}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3e}$$

$$\frac{\partial(.)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{N} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3f}$$

You can rewrite the partials as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3g}$$

$$\hat{\beta}_1 = \frac{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{\text{Cov}\,(x_i, y_i)}{\text{Var}\,(x_i)} \tag{3h}$$

Derivations:

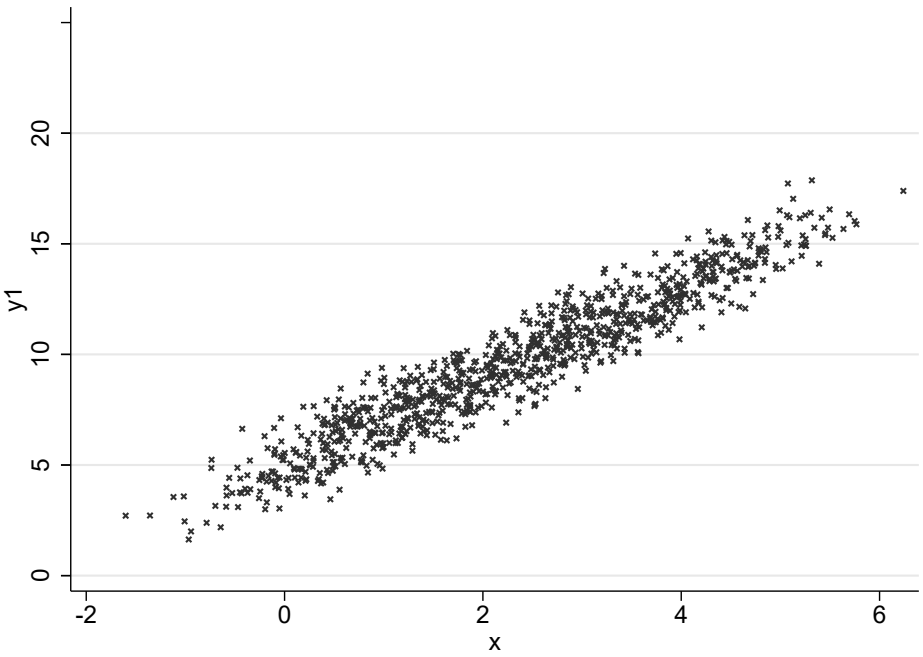
So you can see that the main idea in regression analysis builds on explaining variation. In our example we are essentially looking at how much of the variation in our dependent variable can be explained by the co-variation in x and y. A few terms do well in being clarified at this early stage already. Keep in mind the following definitions:

$$\textbf{Total Sum of Squares } \text{SST} = \sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{4a}$$

$$\textbf{Explained Sum of Squares } \text{SSE} = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \tag{4b}$$

$$\textbf{Residual Sum of Squares } \text{SSR} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\hat{u}_i^2 \tag{4c}$$

Figure 1: Data example



In example 1 of the Stata code accompanying this course, I generate a population of 10000 observations. X is computed as $x_i = (i/N + 2)^2 - 4$. For each observation, I also draw a random error term $u_i \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = 0, \sigma^2 = 1$. I then compute y following Equation (2), with $\beta_0 = 5, \beta_1 = 2.0$. This figure displays y, x for a random 10% sample drawn from the population.

We can use these terms to construct a commonly used *Goodness of Fit* measure, the R^2.

$$R^2 \equiv \frac{SSE}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5a}$$

or

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5b}$$

This statistic represents the share in total variation of y_i that can be explained through the lens of our regression. As such, it measures the explanatory power of the model - the higher R^2, the more powerful is the model. For R^2 very close to 0, you should really revisit the underlying population model that you assume. For R^2 close to 1, your model nearly fits all data points. But typically, this only happens if there exist some other issues - you will learn about these during the PhD.

II The Linear Regression Model with Multiple Regressors (Theory and Numerical Illustration in STATA)

So far, you have gained a first idea what linear regression is all about. In this section, I generalize the univariate linear model that we used so far. The idea is that we want to be able to make *ceteris-paribus* interpretation of the causal link in the dependent and an independent variable of interest. That is, what is the causal effect of a variable x on y, holding all other relevant factors fixed? To do so, we now allow for multiple regressors.

Let me change to matrix notation. Lower-case letters y, u, x_{ik}, β denote scalars. A lower case, bold letter $\mathbf{x}$ and $\mathbf{u}$ will denote a $1 \times k$ (row vector). $\mathbf{y}$ is an $N \times 1$, $\boldsymbol{\beta}$ is a $k \times 1$ column vector. Upper case bold letters denote matrices, f.e. $\boldsymbol{X}$ is an $N \times k$ matrix.

A Ordinary Least Squares (OLS) Estimation

Imagine we have a dependent variable y that is to be explained by k independent variables. The **population equation** reads

$$\underset{1\times1}{y} = \underset{1\times k}{\mathbf{x}}\underset{k\times1}{\boldsymbol{\beta}} + \underset{1\times1}{u} \tag{6}$$

Now we obtain a **random sample of size N** from the population in order to estimation $\boldsymbol{\beta}$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \tag{7}$$

In matrix notation, we rewrite the *sample counterpart* of the population equation as

$$\underset{N\times1}{\mathbf{y}} = \underset{N\times(k)}{\mathbf{X}}\underset{(k)\times1}{\boldsymbol{\beta}} + \underset{N\times1}{\boldsymbol{u}} \tag{8}$$

OLS estimation, the properties and asymptotics of OLS estimators are based on four main assumptions. Before we derive the OLS estimators, let's go through these assumptions and clarify a few points.

Assumptions of the Linear Regression model.

1. **OLS1: Linearity of the Regression Model.**

$$y = \mathbf{x}\boldsymbol{\beta} + u \tag{9}$$

   You assume that the true relationship between the dependent variable and the regressors is linear. Essentially, you want to make sure that you are estimating the right/correct model.

2. **OLS2: Full Rank.** (No perfect collinearity)

$$\text{The sample data matrix } \mathbf{X} \text{ is an } N \times k \text{ matrix with rank } k \tag{10}$$

   This ensures that $\mathbf{X}'\mathbf{X}$ and $\mathbb{E}(\boldsymbol{X'X})$ are invertible.

3. **OLS3: Exogeneity.** The error term is assumed to have *Zero Conditional Mean*, sometimes also referred to as *Mean Independence*:

$$\mathbb{E}[\boldsymbol{u}|\mathbf{X}] = 0 \tag{11}$$

Given the law of iterated expectations, this assumption implies $\mathbb{E}(\mathbf{X}'\boldsymbol{u}) = \mathbf{0}$.[1] Conditional on $\boldsymbol{X}$, the expected mean of the error term is zero. Furthermore, it has interesting implications for our population equation. It implies that the sample regression of $\mathbf{y}$ on $\mathbf{X}$ is the conditional expectation function (CEF) $\mathbb{E}[\mathbf{y}|\mathbf{X}]$:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u} \\ \mathbb{E}[\mathbf{y}|\mathbf{X}] &= \mathbb{E}[\mathbf{X}\boldsymbol{\beta}|\mathbf{X}] + \mathbb{E}[\boldsymbol{u}|\mathbf{X}] \\ \mathbb{E}[\mathbf{y}|\mathbf{X}] &= \mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{12}$$

Without the exogeneity assumption, $\mathbf{X}\boldsymbol{\beta}$ is not the **Conditional Expectation Function**.

4. **OLS4: Homoskedasticity.** We assume that the variance and covariance of the error term is constant:

$$\mathbb{E}[\boldsymbol{u}'\boldsymbol{u}|\mathbf{X}] = \sigma^2 \boldsymbol{I_N} \tag{13}$$

where $\boldsymbol{I_N}$ is the identity matrix. Homoskedasticity might not hold, when errors are heteroskedastic. We will come to this at later stages of this course. For now, just keep this homoskedasticity assumption in the back of your mind - we will use when deriving asymptotic properties of our OLS estimators.

The Sum of Squared Residuals Minimizer. There exist multiple ways to derive estimators for the coefficients. Let us first go through the multivariate analogue of the univariate estimator that we derived previously. Starting from sample data, OLS minimizes the sum of squared residuals:

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\hat{\boldsymbol{\beta}}_{OLS}} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{OLS}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}) \tag{14a}$$

FOCs leads to:

$$\frac{\partial(.)}{\partial \hat{\boldsymbol{\beta}}_{OLS}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} = \mathbf{0} \tag{14b}$$

You rearrange and get

$$\begin{aligned} \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} &= \mathbf{X}'\mathbf{y} \\ \hat{\boldsymbol{\beta}}_{OLS} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \end{aligned} \tag{14c}$$

[1] ***Law of iterated Expectations***

$$\mathbb{E}(y) = \mathbb{E}[\mathbb{E}(y|\mathbf{x})]$$

The outer expectation operator is over x, while the inner expectation operator is over y.

One thing is crucial: For $\hat{\boldsymbol{\beta}}_{OLS}$ to be well-defined, $(\boldsymbol{X'X})$ must have full rank: $rank(\boldsymbol{X'X})) = k$. That is, the $N \times k$ matrix must be invertible. Otherwise, you will not be able to estimate $\hat{\boldsymbol{\beta}}_{OLS}$. You will encounter situations in which full rank breaks down - this can happen when two variables have exact linear relationships. I.e., they are very strongly correlated. We then refer to this as a multicollinearity problem.

The Method of Moments Estimator. For another way to derive the OLS estimators $\hat{\boldsymbol{\beta}}_{OLS}$, consider the following starting point:

$$\begin{aligned} \mathbb{E}\left(\mathbf{X}'\boldsymbol{u}\right) &= \mathbf{0} && \text{(OLS3)} \\ \mathbb{E}\left(\mathbf{X}'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right) &= \mathbf{0} && \text{(OLS1)} \\ \mathbb{E}\left(\mathbf{X}'\boldsymbol{y}\right) &= \mathbb{E}\left(\mathbf{X}'\mathbf{X}\right)\boldsymbol{\beta} \\ \boldsymbol{\beta} &= [\mathbb{E}\left(\mathbf{X}'\mathbf{X}\right)]^{-1}\mathbb{E}\left(\mathbf{X}'\boldsymbol{y}\right) && \text{(OLS2)} \end{aligned} \tag{15}$$

Using sample data to estimate the population terms: replacing the moments with the corresponding sample averages:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \tag{16}$$

In one of the exercises at the end of this section, I ask you to derive the variance of the least square estimator:

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right) = \sigma^2(\boldsymbol{X'X})^{-1} \tag{17}$$

What we actually need in order to be able to do statistical inference is an estimate for σ^2 in that equation, say $\hat{\sigma}^2$. One intuitive approach to estimate $\hat{\sigma}^2$ would be as

$$\hat{\sigma}^2 = \frac{1}{N}\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} \tag{18}$$

However, it can be shown that under this formula, $\hat{\sigma}^2$ is biased downwards (see Greene, p. 102 for details). The correct formula to compute an estimate for the variance of the error term is

$$\hat{\sigma}^2 = \frac{1}{N-k-1}\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = \frac{1}{N-k-1}\sum_{i=1}^{N}\hat{u}_i^2. \tag{19}$$

Our estimate for the sample variance of $\hat{\boldsymbol{\beta}}_{OLS}$ then is

$$\widehat{\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right)} = \hat{\sigma}^2(\boldsymbol{X'X})^{-1}. \tag{20}$$

I will refer to this as the *sample estimate of the sampling variance of the estimator* $\hat{\boldsymbol{\beta}}_{OLS}$.

B Regression Anatomy

Let us now come to a theorem that helps us understand the source of identification of our regression estimates, when we have multiple regressors. Some of those might be our variables of interest, others are *controls*. Suppose we have a sample of size N and want to estimate the following linear model with two stacks of regressors:

$$\underset{N\times 1}{\mathbf{y}} = \underset{N\times k_1}{\mathbf{X}}\underset{k_1\times 1}{\boldsymbol{\beta}} + \underset{N\times k_2}{\mathbf{Z}}\underset{k_2\times 1}{\boldsymbol{\gamma}} + \mathbf{u} \qquad \text{where } \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I_N}) \tag{21}$$

So far, we have derived our coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ using the covariance/variance formula. Now think of $\mathbf{Z}$ as control variables. According to the theorem by Firsch-Waugh-Lovell, controlling for $\mathbf{Z}$ is equivalent to regressing residuals of $\boldsymbol{y}$ (from partialling out all controls $\mathbf{Z}$) on residuals of the regressor of interest $\mathbf{X}$ (after partialling out all controls $\mathbf{Z}$, too).

Theorem 1 (Frisch-Waugh-Lovell theorem) *Let Equation* (21) *be the equation of interest. Then,*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{y} \tag{22a}$$

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{M}_{\mathbf{X}}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_{\mathbf{X}}\mathbf{y} \tag{22b}$$

Where $\mathbf{M}_{\mathbf{Q}} = \mathbf{I} - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$, $\mathbf{Q} = \mathbf{Z}, \mathbf{X}$

We call the matrix $\mathbf{M}_{\mathbf{Q}}$ the residual maker, and it is an idempotent matrix. I.e., it has the property $\mathbf{M}_{\mathbf{Q}} = \mathbf{M}_{\mathbf{Q}}{}'$. When $\mathbf{M}_{\mathbf{Q}} = \mathbf{I} - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$, then $\mathbf{M}_{\mathbf{X}}\mathbf{y}$ is the residual of regressing $\mathbf{y}$ on $\mathbf{X}$, since

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

So equivalently in the theorem, the matrix $\mathbf{X}'\mathbf{M}_{\mathbf{Z}}$ contains the residuals of $\mathbf{X}$ after regressing $\mathbf{X}$ on $\mathbf{Z}$. Mechanically, we can make use of $\mathbf{M}_{\mathbf{Z}}$ being idempotent to rewrite

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

as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}_{\mathbf{Z}}{}'\mathbf{M}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}{}'\mathbf{M}_{\mathbf{Z}}\mathbf{y} \tag{23}$$

$$= [(\mathbf{M}_{\mathbf{Z}}\mathbf{X})'\mathbf{M}_{\mathbf{Z}}\mathbf{X}]^{-1}(\mathbf{M}_{\mathbf{Z}}\mathbf{X})'\mathbf{M}_{\mathbf{Z}}\mathbf{y} \tag{24}$$

Therefore, $\hat{\boldsymbol{\beta}}$ is the coefficient estimation when we regress the residual of regressing $\mathbf{y}$ on $\mathbf{Z}$ on the residual of regressing $\mathbf{X}$ on $\mathbf{Z}$.

Graphically, this would look as follows:

***Exercise:** Proof of Frisch-Waugh-Lovell Theorem on the equivalence of *Partialling out and regressing residuals* to the procedure of *Controlling*.

Derivations:

Cont.

C Coefficient Interpretation

Our main interest in estimating Equation (6): $y = \mathbf{x}\boldsymbol{\beta} + u$ using sample data is to quantify the effect of a variable x_j on $\mathbb{E}(y)$. The sample estimate for the partial effect of x_j, $\hat{\beta}_j$, is a ceteris-paribus effect (i.e. the partial effect of x_j, holding all other $x_k, k \neq j$ constant):

$$\frac{\partial \mathbb{E}(y|\mathbf{x})}{\partial x_j} = \beta_j; \qquad j = 1, 2 \cdots k \tag{26}$$

How to interpret $\hat{\beta}_j$ depends on the types of the variables. Generally there are **three types of variables** we consider in social sciences. Continuous variable (working experience, year of age, year of education $\cdots$), variables in log form ($\ln wage$, $\ln price$), and dummy variables (0,1) or categorical variables (0,1,2 $\cdots$) (gender, nationality, Univ major $\cdots$). Let us now distinguish how to interpret the partial effect of each kind of explanatory variable, depending on the kind of the dependent variable y in the underlying population equation $y = \mathbf{x}\boldsymbol{\beta} + u$. For this purpose, we will consider a model with three explanatory variables, i.e. in the sample estimation $\mathbf{X}$ is $N \times 3$ and $\boldsymbol{\beta}$ is 3×1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \tag{27a}$$

1. The dependent variable ***y is continuous***. The result from our estimation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \tag{27b}$$

   Imagine all $\hat{\beta}$s are significantly positive. Analogously, change *increase* to *decrease* if you prefer to imagine that $\hat{\beta}$ is significantly negative.

   - x_1 **is continuous:** if x_1 increases by one unit, y changes by $\hat{\beta}_1$ unit(s), *ceteris paribus*.
   - x_2 **is in log form:** if x_2 increases by 100%, y changes by $\hat{\beta}_2$ unit(s),*ceteris paribus*.
   - x_3 **is dummy variable:** if x_3 turns from 0 to 1, y changes by $\hat{\beta}_3$ unit(s),*ceteris paribus*.

   ceteris paribus means all other things held constant.

2. The dependent variable ***y is in log form***. The result from our estimation is

$$\widehat{\log y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \tag{27c}$$

   - x_1 **is continuous:** if x_1 increases by one unit, y changes by $\hat{\beta}_1 \times 100\%$, *ceteris paribus*.
   - x_2 **is in log form:** if x_2 increases by 100%, y changes by $\hat{\beta}_2 \times 100\%$,*ceteris paribus*.
   - x_3 **is dummy variable:** if x_3 turns from 0 to 1, y changes by $\hat{\beta}_3 \times 100\%$,*ceteris paribus*.

3. The dependent variable ***y is a dummy variable***. The result from our estimation is a linear probability model (LPM: you will see this in Econometrics I.2):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \tag{27d}$$

   - x_1 **is continuous:** if x_1 increases by one unit, the probability that y is equal to 1 increases by $\hat{\beta}_1 \times 100$ percentage point, *ceteris paribus*.
   - x_2 **is in log form:** if x_2 increases by 100%, the probability that y is equal to 1 increases by $\hat{\beta}_2 \times 100$ percentage point,*ceteris paribus*.

Figure 2: Summary statistics: Wage regressions

```
. sum wage educ exper insur lwage lwage_indus
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1,158	19737.69	19505.28	2544.23	362800.1
educ	1,158	12.82038	2.611496	7	19
exper	1,158	3.036269	2.365935	0	10
insur	1,158	.4896373	.5001086	0	1
lwage	1,158	9.703928	.5810174	7.841583	12.80161
lwage_indus	1,158	10.78193	3.575144	5.571751	23.28384

- x_3 **is dummy variable:** if x_3 turns from 0 to 1, the probability that y is equal to 1 increases by $\hat{\beta}_3 \times 100$ percentage point,*ceteris paribus*.

Figure 3: Coefficient interpretation

(a) Case 1: y is a continuous variable

```
. reg wage educ exper afri_amer lwage_indus
```

Source	SS	df	MS	Number of obs	=	1,158
				F(4, 1153)	=	37.23
Model	5.0353e+10	4	1.2588e+10	Prob > F	=	0.0000
Residual	3.8983e+11	1,153	338104933	R-squared	=	0.1144
				Adj R-squared	=	0.1113
Total	4.4019e+11	1,157	380456015	Root MSE	=	18388

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	1655.687	231.7009	7.15	0.000	1201.084	2110.289
exper	1487.815	248.9207	5.98	0.000	999.4262	1976.203
afri_amer	-5013.192	1184.11	-4.23	0.000	-7336.445	-2689.94
lwage_indus	669.4993	151.9099	4.41	0.000	371.4485	967.5501
_cons	-11272.28	3840.593	-2.94	0.003	-18807.62	-3736.948

(b) Case 2: y is in log-form

```
. reg lwage educ exper afri_amer lwage_indus
```

Source	SS	df	MS	Number of obs	=	1,158
				F(4, 1153)	=	145.91
Model	131.264735	4	32.8161837	Prob > F	=	0.0000
Residual	259.316694	1,153	.224906066	R-squared	=	0.3361
				Adj R-squared	=	0.3338
Total	390.581429	1,157	.337581183	Root MSE	=	.47424

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0997011	.0059759	16.68	0.000	.0879763	.1114259
exper	.080874	.00642	12.60	0.000	.0682778	.0934702
afri_amer	-.2229976	.0305399	-7.30	0.000	-.2829175	-.1630777
lwage_indus	.0193245	.003918	4.93	0.000	.0116373	.0270116
_cons	8.058661	.0990542	81.36	0.000	7.864314	8.253007

(c) Case 3: y is a dummy

```
. reg insur educ afri_amer lwage
```

Source	SS	df	MS	Number of obs	=	1,158
				F(3, 1154)	=	346.48
Model	137.130722	3	45.7102408	Prob > F	=	0.0000
Residual	152.244925	1,154	.131928011	R-squared	=	0.4739
				Adj R-squared	=	0.4725
Total	289.375648	1,157	.250108598	Root MSE	=	.36322

insur	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1038214	.0045011	23.07	0.000	.0949901	.1126527
afri_amer	-.0934717	.0236744	-3.95	0.000	-.1399213	-.0470221
lwage	.1782535	.0209048	8.53	0.000	.1372379	.2192691
_cons	-2.534747	.1949533	-13.00	0.000	-2.91725	-2.152244

An additional case you encounter often in research is when the model we want to estimate

includes an interaction term. Think for example of an equation to estimate the gender gap in wages for men and women as schooling years improve. We will estimate wages w_i on a constant, schooling years s, a dummy for gender (1 = female) to capture the gender gap in overall wages, and an interaction term of being a women and higher educated:

$$w_i = \beta_0 + \beta_1 s_i + \beta_2 Gender_i + \beta_{12} s_i \times Gender_i + u_i \tag{28}$$

The partial effect of being a women on wages is estimated as $\hat{\beta}_2 + \hat{\beta}_{12} s_i$. Hence we need to choose a value for $Schl_i$ when interpreting the partial effect of gender.

1. Partial effect at average (PEA): $\hat{\beta}_2 + \hat{\beta}_{12}\bar{s}$ where $\bar{s} = N^{-1}\sum_{i=1}^{N} s_i$
2. Average partial effect (APE): $N^{-1}\sum_{i=1}^{N}(\hat{\beta}_2 + \hat{\beta}_{12} s_i)$
3. Measure at some certain value: for example, $(\hat{\beta}_2 + \hat{\beta}_{12} s | s = 10)$

Notice: here PEA and APE produces the same result. But in nonlinear cases they are different.

D Hypothesis Testing

One of the main purposes of regression analysis is quantify relationships between variables of interest. We do this in order to validate economic theories or to improve our understanding of economic outcomes. Once we have our regression estimates, we can do statistical inference: We perform statistical tests to see if we can confirm theories/model predictions/causal relationships empirically. Note that this has nothing to do with the economic significance of an estimate.

The main tool we use to test this is the *Hypothesis test*. Let's assume that we run an estimation of

$$y = \mathbf{x}\boldsymbol{\beta} + u; \tag{29}$$

using a sample of size N to retrieve $\hat{\boldsymbol{\beta}}$. Simply put, now we want to know whether we can "trust" one of our coefficients, β_j, in statistical terms. Can we actually rely on that estimate, or is it too *noisy*, such that we should not rely on it? We check this in a ***Two-sided hypothesis test*** by assuming

$$\begin{aligned} H_0: &\quad \beta_j = 0 \qquad (\textbf{Null Hypothesis}) \\ H_1: &\quad \beta_j \neq 0 \qquad (\textbf{Alternative Hypothesis}) \end{aligned}$$

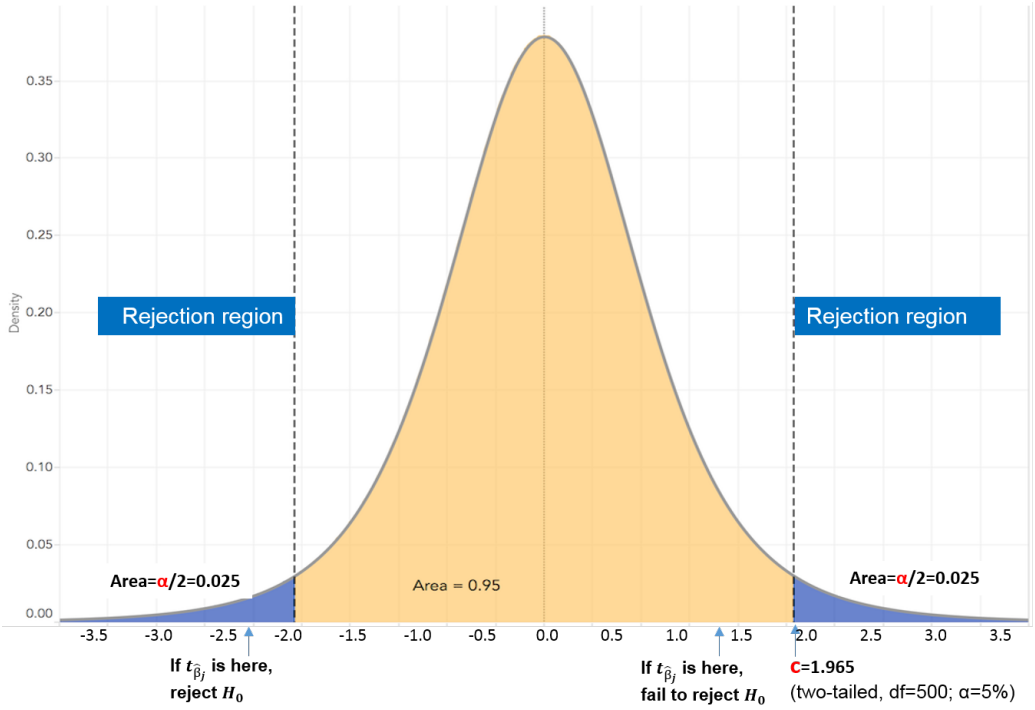
Everything boils down to being able to reject the null hypothesis. This implies that there a relationship between y and x_j - irrespective of its economic significance (i.e the size of β_j)! To check if we can reject the null, we compute the so called ***t*-statistic**:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - 0}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{df} \tag{30}$$

where $\hat{\sigma}_{\hat{\beta}_j}$ is the sample estimate for the sampling variance of our regression coefficient $\hat{\beta}_j$ following from Equation (20). *df* refers to degree of freedom ($df = n-(k+1)$). Think of this measure as telling you how far away from 0 your estimate is in terms of its own standard deviation.

Then, choose a **significance level** $\alpha = 1\%, 5\%, \text{or} 10\%$ (usually of course, you set α before you compute the t-stat!). Simply put, this significance level governs the level of *trust* you can put into your decision. The lower α, the more you can trust your decision if you are able to

Figure 4: Critical values for for a two-sided t-test at 5% significance level (in the normal distribution)



reject H_0. You can also interprete α as the probability of rejecting H_0, when in fact it is true (**Type I error!**). The lower is α, the lower the probability that H_0 is falsely rejected, i.e. that we infer statistical significance of our estimate for β_j, eventhough actually $\beta_j = 0$. When you set $\alpha = 5\%$, in one of twenty cases you will mistakenly find a statistical effect eventhough in reality, you should not. On the other hand, we call it a **Type II error** when we do not reject H_0 eventhough in reality H_1 is true.
Find the critical value z associated to the significance level α (two-tailed, $df = n - (k+1)$, say $\alpha = 5\%$). When the sample is large enough and we thus have enough degrees of freedom (typically $df > 120$), we retrieve these from the normal distribution. Example given, for a two-sided test with a significance level of 5%, the critical z-value is 1.96.

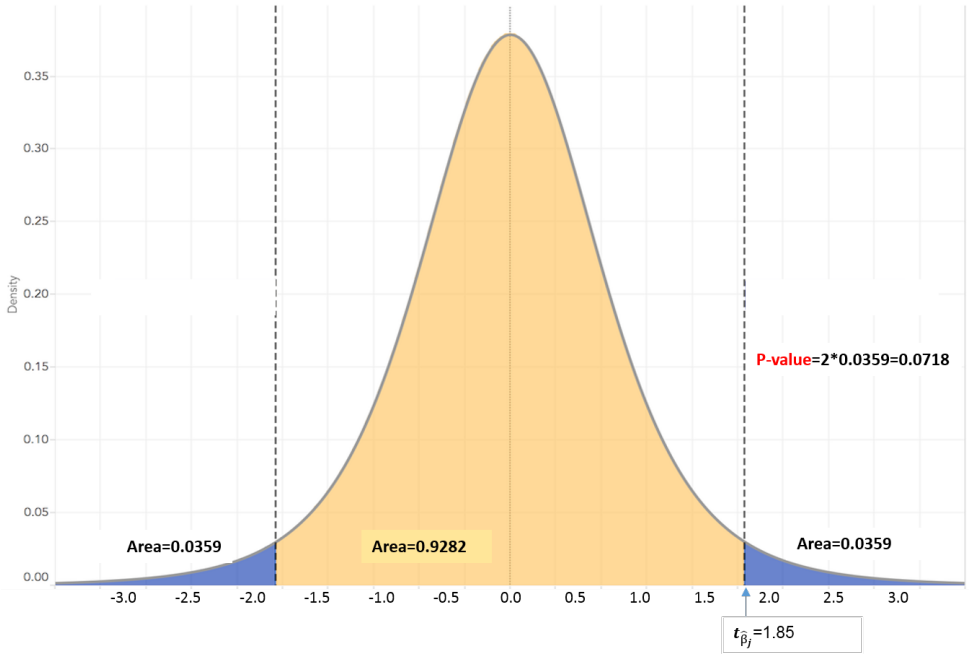
Test: If $|t_{\hat{\beta}_j}| > z$, we reject H_0. Otherwise we fail to reject H_0.

In Figure 4, this boils down to checking whether in terms of normalized standard deviations, our estimate is far enough from zero. The critical value z that we infer from α in a two-sided test gives us three regions. If our t-stat lies within the middle region, we can not reject the null. If it lies in either of the two outer regions, we reject H_0, and say x_j is statistically significant at the α% level.
Note that you can also think of the hypothesis test the other way round: Given a specific $t_{\hat{\beta}_j}$, what is the *lowest* significance level (the strictest criteria) at which H_0 would be rejected. This level is known as **p-value**

- If $\alpha \geq p$, then H_0 should be rejected at the level of α.

For example, imagine your t-test yields $t_{\hat{\beta}_j} = 1.85$, $df = 40$. What is the probability that your estimate lies more than 1.85 normalized standard deviations above or below zero? See Figure 5.

Figure 5: P-Value for a t-test of 1.85



One-sided hypothesis test So far, we discussed the two-sided hypothesis test. Of course, it is also possible to check alternative assumptions. For example, you could check

$$\begin{aligned} H_0: &\quad \beta_j \leq 0 \\ H_1: &\quad \beta_j > 0 \end{aligned}$$

1. Compute t **statistic**:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - 0}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{df} \tag{31}$$

2. Choose a **significance level** α and find critical value c (**one-tailed**, $df = n - (k+1)$, say $\alpha = 5\%$). In terms of standard deviations, is your coefficient so far away from zero that you can indeed trust to be really positive?

3. Test: if $t_{\hat{\beta}_j} > c$, we reject H_0. Otherwise we fail to reject H_0

When H_0 is rejected, we say β_j is statistically significant positive at the α % level.

Other hypotheses. What if you want to check if your estimate confirms a specific value?

$$\begin{aligned} H_0: &\quad \beta_j = b_j \\ H_1: &\quad \beta_j \neq b_j \end{aligned}$$

t **statistic** becomes:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - b_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{df} \tag{32}$$

Or, if you want to check whether two estimates from your regression are actually the same?

$$\begin{aligned} H_0: &\quad \beta_j = \beta_s \\ H_1: &\quad \beta_j \neq \beta_s \end{aligned}$$

t **statistic** becomes:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \hat{\beta}_s}{\hat{\sigma}_{(\hat{\beta}_j - \hat{\beta}_s)}} \sim t_{df} \tag{33}$$

$$\hat{\sigma}_{(\hat{\beta}_j - \hat{\beta}_s)} = \sqrt{\hat{\sigma}^2_{\hat{\beta}_j} + \hat{\sigma}^2_{\hat{\beta}_s} - 2cov(\hat{\beta}_j, \hat{\beta}_s)} \tag{34}$$

Finally, the ***test of joint significance***. There are situations where you want to test whether a group of coefficients of size q has a partial effect on the dependent variable. To check this, you do a joint hypothesis test, where the alternative is that at least one of your coefficients is not equal to zero:

$$\begin{aligned} H_0 &: \beta_{k-q} = \beta_{k-q+1} = ... = \beta_k = 0 \\ H_1 &: \beta_j \neq 0 \text{ for some j = k-q, k-q+1,..k} \end{aligned} \tag{35}$$

Using each coefficients' t-stat to see whether they individually are significant is not appropriate. Inidividually, all your parameters might not be statistically significant. But together, they might have a statistically significant influence on y.

In order to test for common significance, you compare the explanatory power of an unrestricted model (that includes all regressors) to the explanatory of a restricted model (that omits the regressors you want to check). Keep in mind that decreasing the number of variables, the sum of squared residuals (SSR) cannot go down, it can only stay the same or go up (in other terms, R^2 cannot go up!). What you do then is to reject the null hypothesis if the increase in SSR is large enough. This, you measure through the F-stat:

$$F \equiv \frac{N-k-1}{q} \frac{SSR_r - SSR_{ur}}{SSR_{ur}}, \tag{36}$$

where N is the sample size, k is the number of regressors, q is the number of restrictions imposed. The F-stat is distributed as

$$F \sim F_{q,N-k-1}. \tag{37}$$

with q degrees of freedom from the input in the denominator and $n-k-1$ degrees of freedom from the numerator.

Confidence Intervals. Now let's take as a starting point our sample estimate $\hat{\beta}_j$. We do not know the true value of β_j, but using our estimate and its standard deviation, we can infer a range within which the true value should actually lie. We specify the *Confidence Interval (CI)* as

$$\left[\hat{\beta}_j - z \cdot \hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + z \cdot \hat{\sigma}_{\hat{\beta}_j}\right] \tag{38}$$

The *df* and α is implied by z, given the distribution we impose. If we were to obtain random samples over and over, then the unknown population value β_j would lie within that range for (1-α)*100% of the samples. Or think of it another way: The CI implies that the population β_j has a probability of $(1-\alpha)\cdot 100\%$ to lie within the lower bound $\hat{\beta}_j - z \cdot \hat{\sigma}_{\hat{\beta}_j}$ and the upper bound $\hat{\beta}_j + z \cdot \hat{\sigma}_{\hat{\beta}_j}$.

Exercise Section

1. **Exercise*:** Assume $\mathbb{E}[\mathbb{E}(u|\mathbf{x})] = 0$. Show that then, $\mathbb{E}[ux] = 0$.

2. **Exercise*:** Derive the sampling variance $\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right)$ as well as the asymptotic variance $A\text{Var}\left(\hat{\boldsymbol{\beta}}\right)$ of the OLS-estimator :

3. **Exercise**:** Show that the CEF $\mathbb{E}(y_i|\boldsymbol{x_i})$ is the best predictor of y given $\boldsymbol{x}$, since it solves the minimum square prediction error:
$\mathbb{E}(y_i|\boldsymbol{x_i}) = \min_{m(\boldsymbol{x_i})}(y_i - m(\boldsymbol{x_i}))^2$

III Properties & Asymptotics of OLS Estimates

So far, I have shown you how to derive a sample OLS-estimator, and how to interprete it. Let us now turn to asymptotics and properties of the OLS estimator. It is really important to understand that there are certain steps and assumptions involved in linking sample and populations estimates. We impose a population model that is assumed to be true (Asspt. 1). We also assume exogeneity (Asspt. 3) and Homoskedasticity (Asspt. 4). These three assumptions are in *population nature* - but we use random samples of some size N to compute sample estimates $\hat{\boldsymbol{\beta}}$. Essentially, $\hat{\boldsymbol{\beta}}$ gives us an idea of the true population parameters $\boldsymbol{\beta}$.

But what are the properties of those estimators? I.e., how do these estimators behave when applied to a sample of data? Are they ***unbiased***? This concept implies that on average, an estimator will correctly estimate the parameter in question. A biased estimator will be systematically too high, or too low, **irrespective of the sample size!**. That is, you persistently "miss the target" - even if you increase data sample! Reasons: Endogeneity/Omitted Variable/Measurement Error. On the rarest occasions, you will be able to persuade other empirical researchers that you really have an unbiased estimator - all you can do is to try and improve your estimate by including more information (other variables - not add more observations). We will go into the mathematical details of biasedness and its sources in this chapter.

But in terms of properties, we are also interested in how our estimator behaves when we increase the sample size. Under unbiasedness, it doesn't matter how many data points you have: If your sample estimate is unbiased, then that's it - irrespective of whether you have 2, 1000 or a million observations. Think of it: Shouldn't it make sense that a larger sample is better than a small one? More data points mean you have more information to draw on for your estimation. This is why econometricians also look at the asymptotic (or large-sample) properties of a sample estimator. Under the concept of ***consistency***, I like to think that as we get more and more data points, we eventually know the true population estimate. Under the concept of ***asymptotic normality of OLS-parameters***, I like to think that as we get more and more data, our OLS estimates behave more and more like normally distributed random variables. But before we go into the details, we need to look at a few more mathematical concepts that will prove useful lateron.

A Basic Asymptotic Theory

1. Definition: **Convergence**

   A sequence of nonrandom numbers $\{a_N : N = 1, 2 \cdots\}$ **converges to** a (or has limit a) if for all $\epsilon > 0$, there exists N_ϵ such that if $N > N_\epsilon$, then $|a_N - a| < \epsilon$. We write:

   $$a_N \to a, \text{as } N \to \infty \tag{39}$$

   Examples:

   - $a_N = \frac{1}{N}$
   - $a_N = (1 + \frac{1}{N})^N$

2. Definition: **Boundedness**

   A sequence $\{a_N : N = 1, 2 \cdots\}$ is **bounded** if and only if there exists some $b < \infty$ such that $|a_N| \leq b$ for all $N = 1, 2, \cdots$. Otherwise $\{a_N\}$ is **unbounded**.

Convergence implies boundedness, but not the other way round.

Example:

- $a_N = (-1)^N$

3. Definition: **Convergence in probability**

A sequence of random variables $\{x_N : N = 1, 2 \cdots\}$ **converges in probability** to the constant a if for all $\epsilon > 0$,

$$Prob[|x_N - a| > \epsilon] \to 0, \quad \text{as } N \to \infty \tag{40}$$

We write $x_N \xrightarrow{\mathbb{P}} a$ or plim $x_N = a$

4. Definition: **Convergence in distribution**

A sequence of random variables $\{x_N : N = 1, 2 \cdots\}$ **converges in distribution** to the continuous random variable x if and only if

$$F_N(\xi) \to F_x(\xi), \quad \text{as } N \to \infty \text{ for all } \xi \in \mathbb{R} \tag{41}$$

where F_N is the cdf of x_N and F_x is the (continuous) cdf of x. We write:

$$x_N \xrightarrow{d} x \tag{42}$$

If $x \sim Normal(\mu, \sigma^2)$, we write $x_N \xrightarrow{d} N(\mu, \sigma^2)$ or $x_N \overset{a}{\sim} N(\mu, \sigma^2)$ (asymptotically normal). Note the difference between convergence to a constant versus convergence to a distribution: When a random variable converges to a a constant, its variance converges to zero - the random variable converges to something that is not random.

Let us now come to a theorem that we will use lateron when deriving Asymptotics of our OLS estimators:

Theorem 2 (Slutsky's theorem) *Let $g : R^k \to R^J$ be a function continuous at some point $\mathbf{c} \in R^k$. Let $\mathbf{x_1}, \mathbf{x_2} \cdots \mathbf{x_N}$ be sequence of $k \times 1$ random vector such that plim$(\mathbf{x_N}) = \mathbf{c}$, then*

$$plim[g(\mathbf{x_N})] = g(plim(\mathbf{x_N})) = g(\mathbf{c}) \tag{43}$$

It shows that the operator plim passes through nonlinear functions, provided they are continuous. **Corollaries:** Let x_n and y_n be two sequences of random variables with probability limits θ and ν. Then,

$$\begin{aligned} \text{plim}(x_n + / - y_n) &= \theta + \nu \\ \text{plim}(x_n \times y_n) &= \theta \times \nu \\ \text{plim}(x_n / y_n) &= \theta / \nu \end{aligned}$$

Laws of Large numbers and Central Limit Theorems

Theorem 3 (Weak Law of Large Numbers) *$X_1, X_2, ...$ is an infinite sequence of i.i.d. random samples drawn from a distribution whose expected value is μ, then the sample average converges in probability towards this expected value:*

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 \cdots + X_n) \xrightarrow{\mathbb{P}} \mu; \qquad \text{when } n \to \infty \tag{44a}$$

or for any arbitrarily small positive number ϵ

$$\lim_{n\to\infty} Pr(|\bar{X}_n - \mu| > \epsilon) = 0 \tag{44b}$$

Note that writing $plim(\bar{X}_n) = \mu$ is equivalent to writing $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.

This theorem is weak in the sense that it leaves open the possibility that on many occasions in which we draw a sample and compute a sample average $\bar{X}_n$, the difference between the sample average and the true population average μ is larger than some pre-specified tolerance term ϵ. While this theorem converges in probability, there is another one that refers to *almost sure convergence*:

Theorem 4 (Strong Law of Large Numbers) *$X_1, X_2, ...$ is an infinite sequence of i.i.d. random samples drawn from a distribution whose expected value is μ. Under the strong law of large numbers, the sample average converges almost surely towards this expected value:*

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 \cdots + X_n) \xrightarrow{a.s.} \mu; \qquad \text{when } n \to \infty \tag{45a}$$

In other terms,

$$Pr(\lim_{n\to\infty} \bar{X}_n = \mu) = 1 \tag{45b}$$

Almost sure convergence implies convergence in probability, but the converse is not true (that is why the laws of large numbers are called strong and weak respectively).

The laws of large numbers state that sample means converge to a constant in large samples. But in order to be able to speak to the *distribution* of sample means, we recur to (one of many) central limit theorems:

Theorem 5 (Lindeberg-Levy Central Limit Theorem) *$X_1, X_2, ...$ is an infinite sequence of i.i.d. random samples drawn from a distribution whose expected value given by μ and variance given by $\sigma^2 < \infty$.*
Then the random variable $\sqrt{N}(\bar{X}_n - \mu)$ converges in distribution towards a normal distribution $N(0, \sigma^2)$:

$$\sqrt{N}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \qquad \text{when } n \to \infty \tag{46a}$$

where $\bar{X}_n = N^{-1} \sum_i^N X_i$

In most applications, we let random samples converge in distribution to the normal distribution as stated in the above CLT. Otherwise, you might also encounter applications in which random variables converge in distribution to the chi-square distribution.

Finally, the **Delta method**. This is useful when you want to derive the variance of a function that has, as arguments, asymptotically normal random variables of which you know the variance.

Let $\{\hat{\boldsymbol{\theta}}_N : N = 1, 2 \cdots\}$ be a sequence of estimators of the $K \times 1$ vector $\boldsymbol{\theta} \in \Theta$. Suppose that

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} Normal(\mathbf{0}, \mathbf{V}) \tag{47}$$

where $\mathbf{V}$ is $K \times K$ positive semidefinite matrix. Let $\mathbf{c} : \Theta \to \mathbb{R}^Q$ be a continously differentiable function on the parameter space $\Theta \subset \mathbb{R}^K$, where $Q \leq K$ and assume that $\boldsymbol{\theta}$ is in the interior of the parameter space. Define the $Q \times K$ **Jacobian** of $\mathbf{c}$ is $\mathbf{J}(\boldsymbol{\theta}) \equiv \nabla_\theta \mathbf{c}(\boldsymbol{\theta})$. Then

$$\sqrt{N}(\mathbf{c}(\hat{\boldsymbol{\theta}}_N) - \mathbf{c}(\boldsymbol{\theta})) \xrightarrow{d} Normal[\mathbf{0}, \mathbf{J}(\boldsymbol{\theta})\mathbf{V}\mathbf{J}(\boldsymbol{\theta})'] \tag{48}$$

B Properties of the OLS Estimator

Unbiasedness of an Estimator

An estimator $\hat{\boldsymbol{\beta}}$ is unbiased if $\mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$.
Let's start off with our population model in Asspt. 1 and the according sample estimate for $\boldsymbol{\beta}$:

$$y = \mathbf{x}\boldsymbol{\beta} + u \qquad \text{(pop. equation)} \tag{49}$$

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad \text{(sample estimate of } \boldsymbol{\beta}) \tag{50}$$

Derivations on board.

Graphically, this would look as follows:

Consistency of an Estimator

As I increase my sample size, my frequency distribution gets closer and closer to the population value of $\hat{\beta}$. So basically, in probability limit you hit the true population estimate everytime. Let's now derive consistency of an estimator.

Derivations on board.

Graphically, consistency means the following:

Efficiency of the OLS Estimator

Graphical example for unbiased and normally distributed estimator $\hat{\boldsymbol{\beta}}$ versus $\tilde{\boldsymbol{\beta}}$ that has less sampling variance and is thus more efficient.

Graphically, this would look as follows:

Let's show efficiency.
Derivations on board.

Asymptotic Normality of the OLS Estimator
I now show that $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \sigma^2(\mathbb{E}(\mathbf{x}'\mathbf{x}))^{-1})$ Starting point - what else - is again the population model in Asspt. 1 and the according sample estimate for $\boldsymbol{\beta}$:

$$y = \mathbf{x}\boldsymbol{\beta} + u \tag{51a}$$
$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{51b}$$
$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \tag{51c}$$

Rearrange,

$$\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} = (\frac{1}{N}\boldsymbol{X}'\boldsymbol{X})^{-1}(\frac{1}{N}\boldsymbol{X}'\boldsymbol{u})$$
$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) = (\frac{1}{N}\boldsymbol{X}'\boldsymbol{X})^{-1}(\frac{1}{\sqrt{N}}\boldsymbol{X}'\boldsymbol{u})$$

Now,

1. By WLLN & Slutzky, $\left(\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \xrightarrow{\mathbb{P}} \mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$

   $\Rightarrow \left(\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ converges in probability limit to constant w/out a variance!

2. By Lindeberg-Levy CLT: $\frac{1}{\sqrt{N}}\boldsymbol{X}'\boldsymbol{u} \xrightarrow{d} N\Big(0, \underbrace{\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u}'\boldsymbol{u}\right)}_{=\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\sigma^2, \text{ Asspt. 4}}\Big)$

such that we get

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} - \boldsymbol{\beta}) &\xrightarrow{d} N\Big(0, \mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\right)\sigma^2\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\Big) \\ &\xrightarrow{d} N\Big(0, \mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2\Big).\end{aligned} \tag{51d}$$

C Model Selection: Irrelevant & omitted variables

What happens if the underlying population model that you assume to be true (Asspt. 1 is incorrect? Think of two possible mistakes you make: For one, your assumed population model omits relevant variables? This will actually turn out to be a problem, as it affects *unbiasedness* and *consistency* of your sample regression estimates. What about a second case, in which your assumed population model includes regressors that are *irrelevant*. Does this affect the sample estimates for the relevant regressors?

Omitted variable bias
Examples: Wage (y) on education (x_1). Omitted: Ability x_2.
As a starting point, consider the two following population models:

$$y = x_1\beta_1 + x_2\beta_2 + u \qquad \text{(true pop. model)}$$
$$y = x_1\beta_1 + \epsilon \qquad \text{(assumed, but incorrect pop. model)}$$

Your sample estimate for β_1, $\hat{\beta}_1$ using data samples $\boldsymbol{y}, \boldsymbol{x_1}$ will equal

$$\hat{\beta}_1 = (\boldsymbol{x_1}'\boldsymbol{x_1})^{-1}\boldsymbol{x_1}'\boldsymbol{y} \tag{52a}$$

Now, substitute $\boldsymbol{y}$ using the *correct* sample counterpart of the population equation:

$$\begin{aligned}\hat{\beta}_1 &= (\boldsymbol{x_1'x_1})^{-1}\boldsymbol{x_1'}\Big(\mathbf{x_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u}\Big) \\ &= \beta_1 + (\boldsymbol{x_1'x_1})^{-1}\boldsymbol{x_1'}\mathbf{x_2}\beta_2 + (\boldsymbol{x_1'x_1})^{-1}\boldsymbol{x_1'}\mathbf{u}\end{aligned}$$

Taking conditional expectations (denoting $\boldsymbol{X = (x_1, x_2)}$), we get that

$$\mathbb{E}\left(\hat{\beta}_1|\boldsymbol{X}\right) = \beta_1 + \mathbb{E}\left(\boldsymbol{x_1'x_1}|\boldsymbol{X}\right)^{-1}\mathbb{E}\left(\boldsymbol{x_1'x_2}|\boldsymbol{X}\right)\beta_2 + \underbrace{\mathbb{E}\left(\boldsymbol{x_1'x_1}|\boldsymbol{X}\right)^{-1}\mathbb{E}\left(\boldsymbol{x_1'u}|\boldsymbol{X}\right)}_{=0 \text{ by Asspt. 3}}$$

such that

$$\mathbb{E}\left(\hat{\beta}_1\right) = \beta_1 + \frac{\text{Cov}\left(\boldsymbol{X_1, X_2}\right)}{\text{Var}\left(\boldsymbol{X_1, X_1}\right)}\beta_2 \tag{52b}$$

Note that the third term can be thought of as a sample estimate from a regression of $\boldsymbol{x_2}$ on $\boldsymbol{x_1}$, formally:

$$\begin{aligned} x_2 &= \boldsymbol{x_1}\delta_1 + \nu \\ \text{s.t. } \mathbb{E}\left(\boldsymbol{x_2}|\boldsymbol{x_1}\right) &= \mathbb{E}\left(\boldsymbol{x_1}\delta_1 + \boldsymbol{\nu}|\boldsymbol{x_1}\right) \qquad (52c) \\ \text{with } \hat{\delta}_1 &= (\boldsymbol{x_1'x_1})^{-1}\boldsymbol{x_1'}\mathbf{x_2} \qquad (52d)\end{aligned}$$

Baseline: Your estimate for β_1 is ***biased***, if the included and the omitted regressors are correlated and the omitted regressor has an effect on y. The larger the correlation (be it positive or negative!) of both regressors, and the larger β_2 as the marginal effect of the omitted regressor is on y, the larger is the *upward* or *downward* bias, respectively. **Exercise:** Think about this in the wage/education/ability example!

If either the marginal effect of the omitted regressor on y or the correlation of the two regressors is zero, then the estimator is unbiased and consistent. However, if $\beta_2 = 0$, then the omitted variable is irrelevant. Should we actually care about including variables that are irrelevant? At first thought, one might think that including "more variables is always better": More information can yield more explanatory power. Why not include as many variables as possible, irrespective of whether these variables have an effect on the dependent variable y?

Irrelevant variables

The answer to this question is not straightforward. Consider a model such as

$$y_i = \mathbf{x_{1i}}\beta + \mathbf{z_i}\boldsymbol{\gamma} + u_i, \tag{53}$$

where $\boldsymbol{z_i} = (x_{2i}, x_{3i}, .., x_{ki})$ is a $1 \times (k-1)$ vector. The asymptotic variance of the estimator $\hat{\beta}$, $A\text{Var}\left(\hat{\beta}\right)$ is:

$$A\text{Var}\left(\hat{\beta}\right) = \frac{\sigma^2}{N\sigma_{x_1}^2(1 - R_{x_1}^2)}, \tag{54}$$

where the term σ^2 is the sum of squared residuals, $\boldsymbol{u'u}$, $\sigma_{x_1}^2$ is the variance of the regressor $\boldsymbol{x_1}$, and $R_{x_1}^2$ is the *Goodness of fit measure* in a regression of $\boldsymbol{x_1}$ on all other regressors $\boldsymbol{Z}$. I actually ask you to derive this formula in the exercise section. From this it becomes clear that the Var $\left(\hat{\beta}\right) \uparrow$ with higher collinearity ($R_X^2 \uparrow$ under multicollinearity). So if you add more and more variables that unfortunately correlate too much with your explanatory variable x_1, you might end up having problems in your inference. This is because the estimators sampling variance then increases.

Otherwise note:

- $\uparrow$ with σ^2 (the variance of the regression residual)
- $\downarrow$ with sample size $N \uparrow$
- $\downarrow$ with variance of regressor $\boldsymbol{X}$

Exercise Section

1. **Exercise**:** Does consistency imply unbiasedness, or does unbiasedness imply consistency?

2. **Exercise**:** Attenuation bias. Consider the following model

$$y = \beta_0 + x^* \beta_1 + u. \tag{55}$$

   You do not observe x^*, but rather x where $x = x^* + e$. Derive the estimators $\tilde{\beta}_0, \tilde{\beta}_1$ from a regression of y on x. Are they biased, and if so under which conditions?
   Hint: Assume that the error e is related to x by $e = \delta_0 + x\delta_1 + \epsilon$, with $\mathbb{E}(\epsilon|x) = 0$.

3. ****Exercise:** Consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}; \tag{56a}$$

   Derive the asymptotic variance of the estimator $\hat{\boldsymbol{\beta}}$, $A\text{Var}\left(\hat{\boldsymbol{\beta}}\right)$ and show that you can rewrite it as:

$$A\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \frac{\sigma^2}{N\sigma_X^2(1 - R_X^2)}, \tag{56b}$$

   where the term σ^2 is the sum of squared residuals, $\boldsymbol{u'u}$, σ_X^2 is the variance of the regressor $\boldsymbol{X}$, and R_X^2 is the *Goodness of fit measure* in a regression of $\boldsymbol{X}$ on all other regressors $\boldsymbol{Z}$.

IV Heteroskedasticity & Generalized least squares (GLS)

So far, we have worked with a multiple regression model where disturbances are homoskedastic. Remember that one of the main assumptions we work with when estimating regressions models is the assumption of homoskedasticity (see Asspt. 4). We require that

$$\text{Var}\left(u_i|x_{1i}, x_{2i}, ...x_{ki}\right) = \sigma^2 \qquad \forall i = 1, ...N \tag{57a}$$

or in matrix notation:

$$\mathbb{E}\left(\boldsymbol{u'u}|\boldsymbol{X}\right) = \sigma^2\boldsymbol{I}, \tag{57b}$$

i.e. that for all observations, conditional on the regressors in $\boldsymbol{X}$ the disturbance has the same variance σ^2. What if the assumption of homoskedasticity cannot be maintained?

In order to discuss this, it is useful to set up a more general version of the regression model we discussed so far. The generalized linear regression model will be

$$y = \boldsymbol{x\beta} + u, \tag{58a}$$

with

$$\mathbb{E}[\boldsymbol{u}|\mathbf{X}] = 0 \qquad \textit{(Exogeneity)} \tag{58b}$$

$$\mathbb{E}\left(\boldsymbol{u'u}|\boldsymbol{X}\right) = \sigma^2\boldsymbol{\Omega} \qquad \textit{(Allow for Heteroskedasticity)} \tag{58c}$$

This version is more general in the sense that it allows for error terms to be heteroskedastic. That is, the disturbances can exhibit differing variation, even after controlling for $\boldsymbol{X}$. I want you to think of $\boldsymbol{\Omega}$ as a scaling matrix such as

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & ... & 0 \\ 0 & \omega_2 & ... & 0 \\ & & ... & \\ 0 & ... & 0 & \omega_N \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ & & ... & \\ 0 & ... & 0 & \sigma_N^2 \end{bmatrix}$$

That is, for $i = 1, ...N$ you can write $\sigma_i^2 = \sigma^2\omega_i$. Heteroskedasticity is what we will focus on in this section. Let me just point out that under serial correlation of the error term (or *autocorrelation*) in time series regressions, $\boldsymbol{\Omega}$ looks more like

$$\sigma^2\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho_1 & ... & \rho_N \\ \rho_1 & 1 & ... & 0 \\ & & ... & \\ \rho_N & \rho_{N-1} & ... & 1 \end{bmatrix}$$

You can think of the data as displaying a "memory", with the variation in a variable being dependent across time. You will encounter this in the second part of the Ph.D. graduate courses in econometrics.

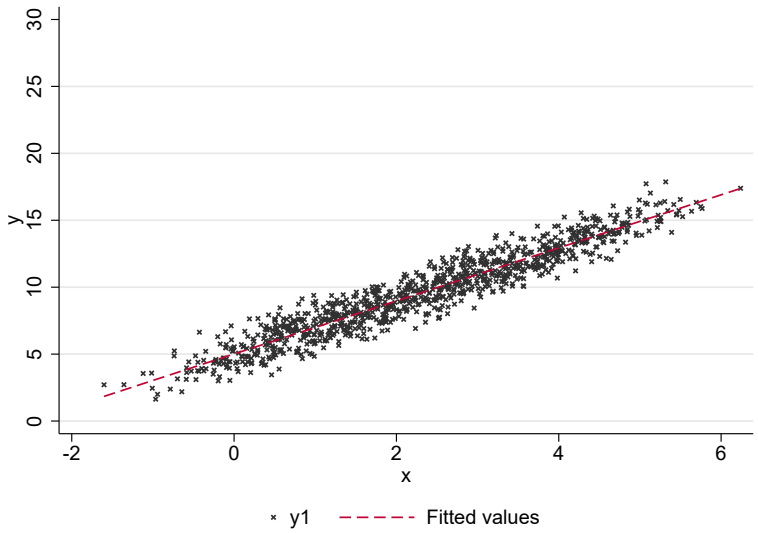
The assumption of homoskedasticity, i.e. that $\boldsymbol{\Omega} = \boldsymbol{I}$ did not play a role in showing that the OLS estimators are consistent and unbiased. But heteroskedasticity has the consequence that the estimators for the *variance of* $\beta_j, j = 1, ..k$ *are biased* (NOT: The estimated coefficients!). And what we need for inference (like our t-tests) are unbiased standard errors! In this section, we will look at how to detect and how to deal with heteroscedasticity in this generalized regression model.

A Detecting Heteroskedasticity

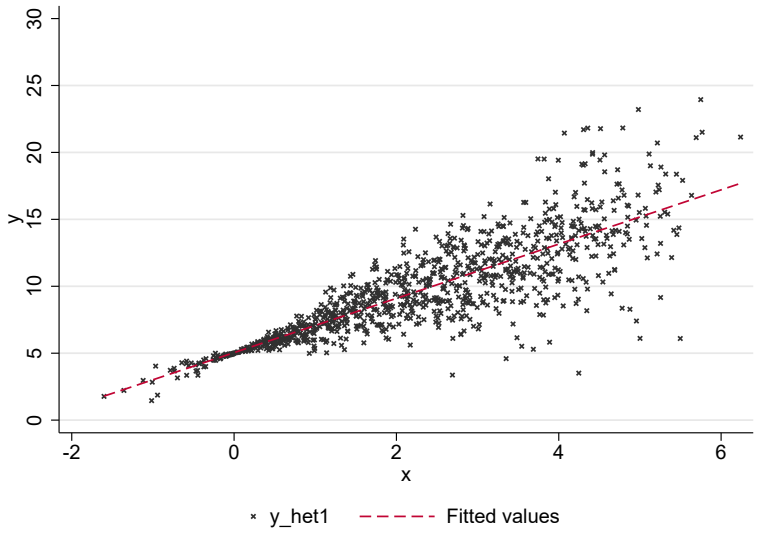
Let's go back to the first graph I showed you, and compare it to a different one where I set up the dependent variable y with an error term that is heteroskedastic. The first thing you should always do, before you run any regression, is a graphical analysis of your variables y and x. This might already be helpful in signalling that there might be some issue related to heteroskedasticity. Take for example the two following graphs:

Figure 6: Data

(a) Homoskedasticity



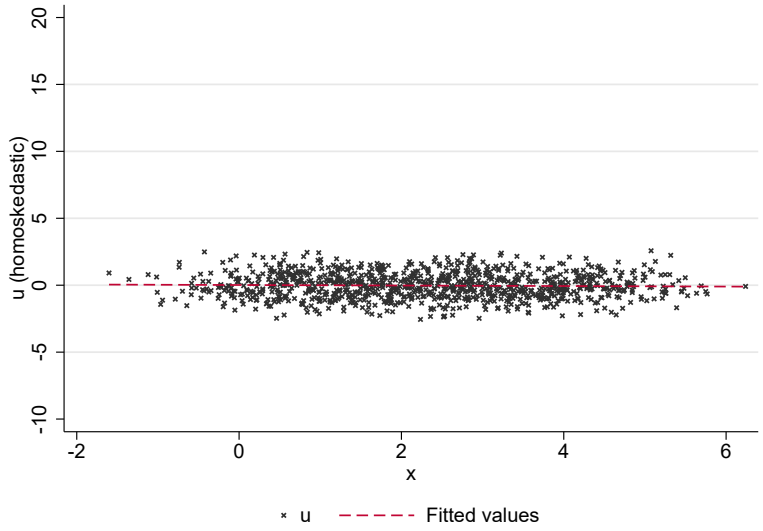
(b) Heteroskedasticity



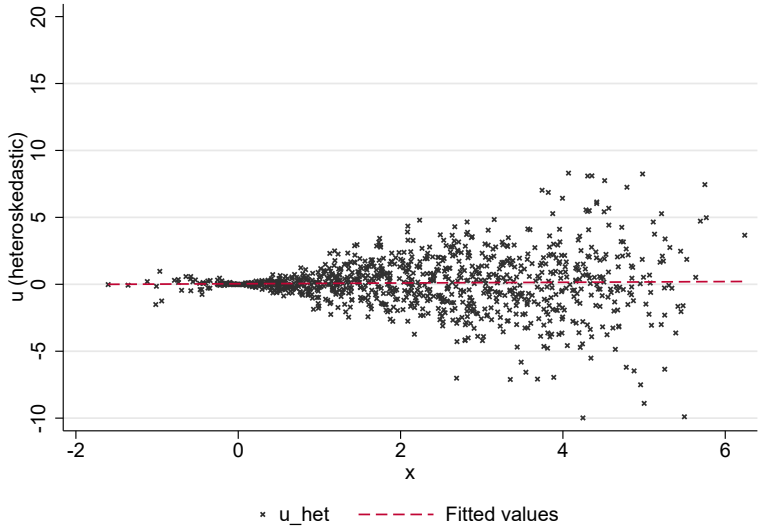
Notes: The lhs-graph displays the values in a scenario where I draw y using homoskedastic error terms. In the rhs-panel, y is computed using heteroskedastic error terms. For details on how to set this up, see the DoFile accompanying this course.

Figure 7: Error term

(a) Homoskedasticity



(b) Heteroskedasticity



Examples include Firm profits (dep. var) and firm size(indep. var), or vacation expenditures (dep. var) on family income (indep. var). Let me come back to how heteroscedasticity affects our regression results. Again, let me highlight that homoskedasticity is not needed for our parameter estimate to be unbiased and consistent. To see this, consider the following sample regression outputs for the underlying population model

$$y = \beta_0 + \beta_1 x + u \tag{59}$$

Table 1: Regression under Homo- & Heteroskedasticity

	Homoskedastic u	Heteroskedastic u	Het. u, robust Std. Errors
β_1	2.0^{***}	2.0^{***}	2.0^{***}
	(0.021)	(0.043)	(0.051)
β_0	5.0^{***}	5.0^{***}	5.0^{***}
	(0.057)	(0.12)	(0.079)
N	975	975	975

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: All three column displays regression results from a sample estimation of Equation (56d). In the first column, the error term is homoskedastic. In the second and third column the error term is heteroskedastic, but in column (2) I run a regression without adjusting standard errors for heteroskedasticity. In row (3), the adjustment for heteroskedastic errors is done.

What I want you to see from this is that the parameter estimates for β_1 do not change - neither when the error terms turn from being homo- to heteroskedastic, and neither when we do an adjustment of the standard errors for heteroskedasticity. Also, and this is the crucial danger arising from heteroskedasticity: The standard error of our parameter estimate β_1 is biased downwards. That is, when we do our inference and compute the test-statistic

$$t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} \tag{60}$$

then the t-stat is larger than it should be. We run into the danger of falsely rejecting the null hypothesis that $\beta_1 = 0$! I already include the results from an adjustment for heteroskedasticity robust standard errors for you to see that with the adjustment, $se(\hat{\beta}_1)$ goes up. We will come to the details of this adjustment lateron.

Of course, there also exist formal tests for heteroskedasticity. One of them is the ***Generalized White Test for heteroskedasticity***. Essentially, what it does is to test whether the residuals obtained from a fitted regression correlate with the explanatory variables. Consider the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \tag{61}$$

Broadly speaking, the test-statistic is computed from the following steps:

1. Run a sample estimation of Equation (56f).
2. Compute the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$ and retrieve the fitted residuals $\hat{u}_i = y_i - \hat{y}_i$
3. Run a regression of $\hat{u}_i^2$ on $\hat{y}_i, \hat{y}_i^2, ...$, retrieve $R^2_{\hat{u}_i^2}$
4. Compute the F-statistic to test

$$H_0 : \delta_1 = \delta_2 = ... = \delta_k = 0 \text{ (Homoskedasticity)} \tag{62}$$
$$H_1 : \delta_j \neq 0 \text{ for some j = 1,2,..,k} \tag{63}$$

with

$$F \equiv \frac{N-k-1}{k} \frac{R^2_{\hat{u}_i^2}}{1 - R^2_{\hat{u}_i^2}} \sim F_{k,N-k-1}, \tag{64}$$

with

Note that you regress the fitted residuals on the fitted values of y, not on the actual regressors. Implicitly though, by regressing the fitted residuals on a linear combination of the fitted dependent variable, you're testing whether $\hat{u}_i^2$ is correlated to all kinds of combinations of the explanatory variables $x_1, ...x_k$. For example, you could simply run $\hat{u}_i^2 = \delta_0 + \delta_1 x_i + \delta_2 x_{2i} + ... + \delta_k x_{ki}$. But it would be even more general to actually allow for non-linear relations in u_i^2 and the regressors in $x_1, ...x_k$, i.e. you could include $\gamma_1 x_{1i}^2 + ... + \gamma_k x_{ki}^2$, or even cross products like $x_1 i \times x_2 i$! However, then you're loosing degrees of freedom and thus power of your statistical tests. Nevertheless, this kind of regression forms a similar basis for the so-called ***Breusch-Pagan Lagrange Multiplier Test***, which is another test for heteroskedasticity. There, you test the hypothesis that $\sigma^2 = \sigma^2 f(\boldsymbol{x\delta})$ f.e. based on

$$\hat{u}_i^2 = \delta_0 + \delta_1 x_i + \delta_2 x_{2i} + ... + \delta_k x_{ki} \tag{65}$$

and compute the LM- test statistic as $LM = N \times R^2_{\hat{u}_i^2} \sim X_k^2$. The intuition behind this LM-test statistic is that it captures in how far in the sample, the explanatory variables of the model explain a lot of the variation in the residuals. Under homoskedasticity, $R^2_{\hat{u}_i^2}$ should be zero! Thus, the higher the LM-stat, the more likely it is that I have a heteroscedasticity problem.

B Robust Standard Errors

So, how do you deal with a heteroskedasticity problem? That's actually not too much of an issue. Remember that we derived the sampling variance of $\hat{\beta}_{OLS}$

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right) = (\boldsymbol{X'X})^{-1}\boldsymbol{X}\text{Var}\left(\boldsymbol{u}|\boldsymbol{X}\right)\boldsymbol{X}(\boldsymbol{X'X})^{-1}$$

which under homoskedasticity reduces to

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right) = \sigma^2(\boldsymbol{X'X})^{-1}. \tag{66a}$$

We use

$$\hat{\sigma}^2 = \frac{1}{N-k-1}\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}. \tag{66b}$$

to estimate the sampling variance of $\hat{\boldsymbol{\beta}}_{OLS}$ using our sample as

$$\widehat{\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right)} = \hat{\sigma}^2(\boldsymbol{X'X})^{-1}. \tag{66c}$$

Now under heteroskedasticity, $\text{Var}\left(\boldsymbol{u}|\boldsymbol{X}\right) = \sigma^2\boldsymbol{\Omega}$, such that in the generalized model from Equation (56ca) the sampling variance of $\hat{\beta}_{OLS}$

$$\begin{aligned}\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right) &= (\boldsymbol{X'X})^{-1}\boldsymbol{X}\text{Var}\left(\boldsymbol{u}|\boldsymbol{X}\right)\boldsymbol{X}(\boldsymbol{X'X})^{-1} \\ &= \sigma^2(\boldsymbol{X'X})^{-1}\boldsymbol{X'\Omega X}(\boldsymbol{X'X})^{-1}\end{aligned} \tag{67a}$$

or put differently:

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right) = (\boldsymbol{X'X})^{-1}\left(\sigma^2\sum_{i=1}^{N}\omega_i\boldsymbol{x_i}\boldsymbol{x_i'}\right)(\boldsymbol{X'X})^{-1}. \tag{67b}$$

We neither know the true weights ω_i nor the true errors u_i that form σ^2. So we rely on sample estimates. From our sample, we compute the *White heteroskedasticity consistent estimator* for the sampling variance of the estimator $\hat{\boldsymbol{\beta}}_{OLS}$ as

$$\widehat{\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}\right)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\boldsymbol{x_i}\boldsymbol{x_i'}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{68}$$

i.e. we draw on the fitted residuals from our regression to compute this so called *Sandwich estimator*.

C Generalized Least Squares Estimation

The aim of GLS is to provide efficient estimation of our parameters $\boldsymbol{\beta}$ in the generalized model. Now, what we do under GLS is factorize our matrix $\boldsymbol{\Omega}$ to obtain a matrix $\boldsymbol{P}$ s.t.

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{P'P}. \tag{69}$$

Now, if you take our generalized regression model in Equation (56ca) and multiply both sides by $\boldsymbol{P}$, you get a transformed model

$$\boldsymbol{Py} = \boldsymbol{PX\beta} + \boldsymbol{Pu}, \tag{70}$$

$$\text{or } \boldsymbol{y^*} = \boldsymbol{X^*\beta} + \boldsymbol{u^*} \tag{71}$$

with the characteristic that the error terms of the transformed model are homoskedastic:

$$\mathbb{E}\left(\boldsymbol{Pu}(\boldsymbol{uP})'|\boldsymbol{X^*}\right) = \boldsymbol{P}\sigma^2\boldsymbol{\Omega P'} = \sigma^2\boldsymbol{I} \tag{72}$$

That is, if we apply OLS on the transformed model, the resulting GLS estimate

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X'P'PX})^{-1}\boldsymbol{X'P'}P\boldsymbol{y} \tag{73}$$

$$= (\boldsymbol{X'\Omega^{-1}X})^{-1}\boldsymbol{X'\Omega^{-1}y} \tag{74}$$

is efficient and results from applying the "weighting matrix" $\boldsymbol{\Omega^{-1}}$ to the data. It is obtained by regressing

$$\boldsymbol{Py} = \begin{bmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ . \\ . \\ y_N/\sqrt{\omega_N} \end{bmatrix}$$

on

$$\boldsymbol{PX} = \begin{bmatrix} \boldsymbol{x_1'}/\sqrt{\omega_1} \\ \boldsymbol{x_2'}/\sqrt{\omega_2} \\ . \\ . \\ \boldsymbol{x_N'}/\sqrt{\omega_N} \end{bmatrix}$$

The fact that you are weighting your variables might be more obvious if you look at the *Weighted least squares estimator* analogue to the $\boldsymbol{\beta}_{GLS}$. In summation terms:

$$\hat{\boldsymbol{\beta}}_{WLS} = \left[\sum_{i=1}^{N} w_i\boldsymbol{x_i}\boldsymbol{x_i'}\right]^{-1}\left[\sum_{i=1}^{N} w_i\boldsymbol{x_i}\boldsymbol{y_i}\right], \tag{75}$$

where $w_i = 1/\omega_i$. Intuitively, you compute an efficient estimate by weighting observations with smaller variance of the error term ω_i more than observations with a larger variation of their error term.

Weighted least squares in practice - or feasible GLS. Typically, we do not know the true values of the weights - or the true relationship between our explanatory variables $\boldsymbol{X}$ and $\boldsymbol{u}$ that describes the heteroskedasticity in $\boldsymbol{u}$. So what we do in practice is to impose a relationship between those variables. We assume

$$\text{Var}\left(u_i|\boldsymbol{x_i}\right) = \sigma^2 \exp\left(\boldsymbol{x_i'\delta}\right) \tag{76}$$

We then compute *two-step estimators*, where

1. Step: Estimate $y_i = \boldsymbol{x_i\beta} + u_i$ using data

   $\Rightarrow$ retrieve $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$, i.e. $\hat{u}_i$ for $i = 1, ..., N$

2. Step: Estimate $log(\hat{u}_i^2) = \delta_0 + \delta_1\boldsymbol{x_{1i}} + ... + \delta_k\boldsymbol{x_{ki}}$

   $\Rightarrow$ retrieve $\hat{g}_i = \hat{\delta}_0 + \hat{\delta}_1\boldsymbol{x_{1i}} + ... + \hat{\delta}_k\boldsymbol{x_{ki}}$
   where $\hat{g}_i$ is the fitted value for the log of the conditional variance of u_i. This is what we can use to transform our data.

3. Step: Retransform $\hat{g}_i$ to obtain $\exp(\hat{g}_i) = \hat{h}_i = \hat{u}_i^2$, the estimated conditional variance.

4. Step: Estimate

$$\frac{y_i}{\sqrt{\hat{h}_i}} = \frac{\boldsymbol{x_i}}{\sqrt{\hat{h}_i}}\boldsymbol{\beta} + \nu_i \tag{77}$$

Properties: Feasible GLS is biased in finite samples, but consistent.

V Monte Carlo Experiments in Econometrics: Key Ideas and Numerical Illustration in STATA

The properties of estimators and statistical tests we apply on them are sometimes harder to establish. Large-sample asymptotics can be derived in theory, but what when it comes to characterizing distributions and properties of estimates from finite samples? Exact finite-sample properties are typically not very tractable - but Monte Carlo simulations are a tool that helps us learn about the behavior of our estimator experimentally.

In your exercises, you typically have only one sample. Keep in mind that any statistic you derive from a sample is a random variable, given that the act of obtaining a sample is (hopefully) a random experiment. Of course, there is a true data generating process (DGP) underlying the population you draw your sample from and that forms the baseline for your population equation. But each sample that you obtain will deliver different values to your statistics. Now we are interested in how much we can "trust" our estimates and/or inference exercises, given our sample characteristics. In order to find out about this, we can recur to experiments. For example, we can repeatedly generate samples for a known DGP. Each time, we re-calculate the statistics we are interested in, and check its propertiers. Or, we can check how the properties of our estimators change when we allow for different sampling conditions (like sample size N).

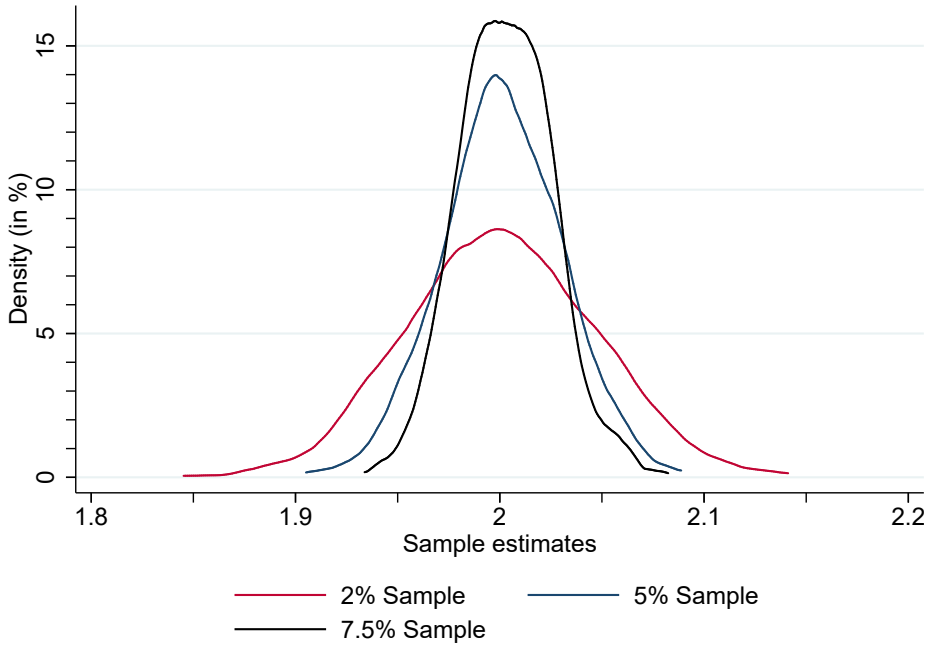
Example: Suppose you are interested in how the distribution of some estimator $\hat{\beta}_1$ changes as you increase the sample size. The key steps to a Monte-Carlo simulation are

1. Specify the data generating process. In our Stata example, I generate a population of $N = 10K$ observations.

$$y = 5 + 2x_1 + u \tag{78}$$

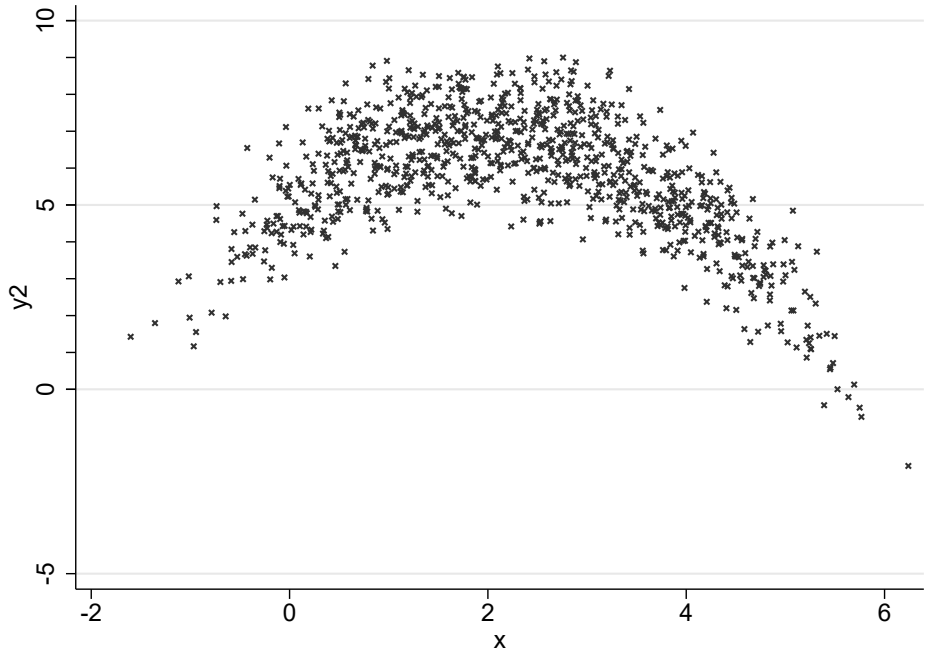
2. Draw a sample from the DGP (here: 2%, 5% and 7.5% samples)
3. Calculate and store $\hat{\beta}_1$ based on this sample
4. Repeat steps (2)-(3) R times, where R is large and each repetition is a *replication*.
5. Repeat step (4) for each of the three sample sizes
6. Evaluate the distribution of your coefficient given different sample sizes (here: density graphs to check efficiency, see Figure 8)

Figure 8: Monte Carlo example



VI Appendix

Figure 9: Data example: Quadratic relationships



In example 1 of the Stata code accompanying this course, I generate a population of 10000 observations. X is computed as $x_i = (i/N + 2)^2 - 4$. For each observation, I also draw a random error term $u_i \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = 0, \sigma^2 = 1$. I then compute y as $y_i = 5 + 2 * x_i - 0.5 * x_i^2 + u_i$. This figure displays y, x for a random 10% sample drawn from the population.

Figure 10: Hours per week in market work

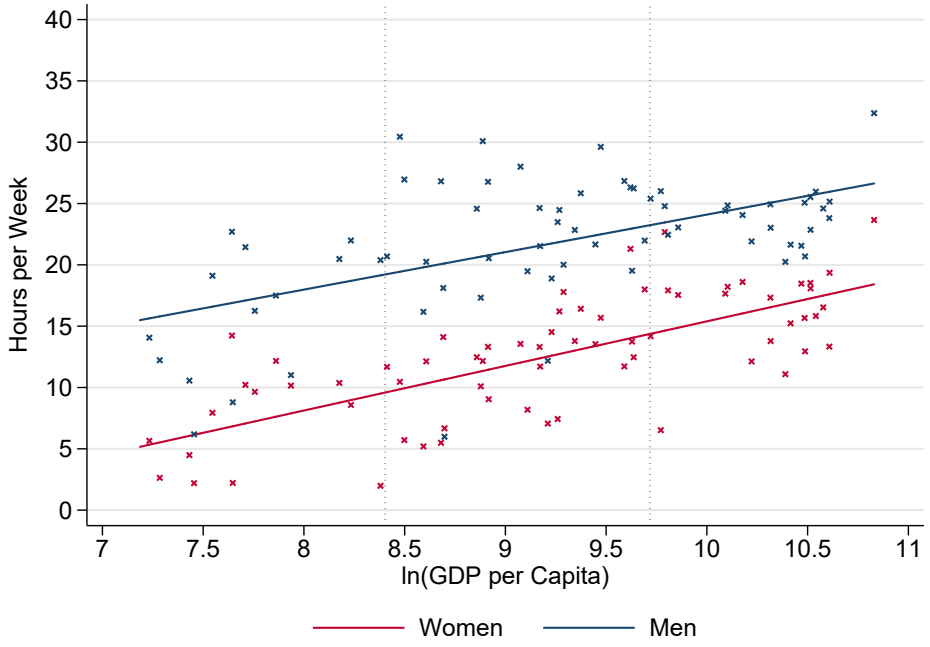
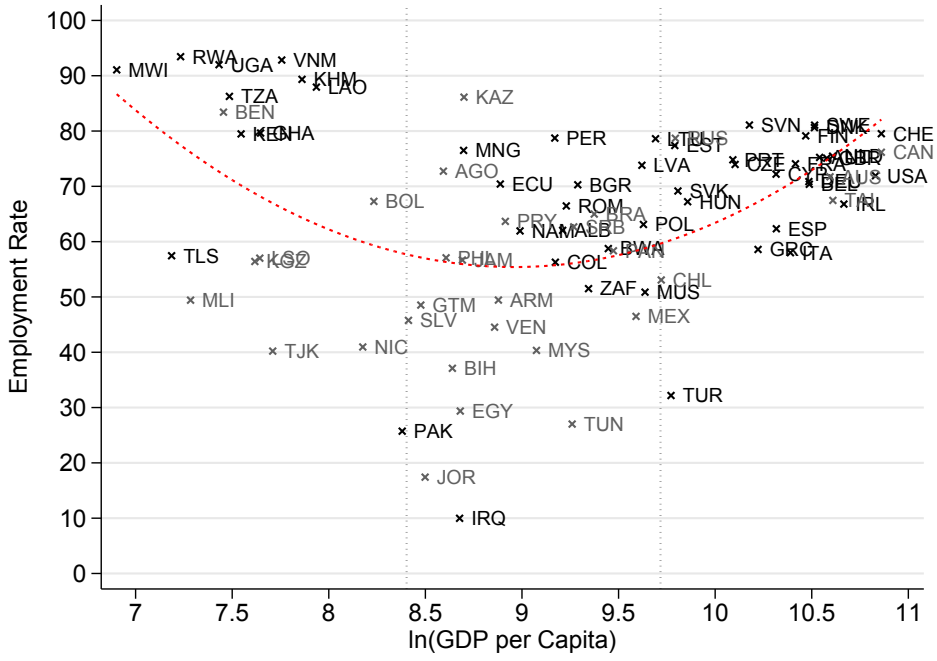


Figure 11: Employment Rates, prime aged women



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Table 3: Coefficient interpretation

	Cont.-Cont.	Log-Cont.	Linear Prob. Model
1.edu	0	0	0
	(.)	(.)	(.)
2.edu	5.7***	0.3***	0.07***
	(0.29)	(0.0077)	(0.0033)
3.edu	14.9***	0.7***	0.2***
	(0.28)	(0.0075)	(0.0033)
age	1.2***	0.07***	0.04***
	(0.022)	(0.00058)	(0.00021)
age2	-0.01***	-0.0007***	-0.0005***
	(0.00026)	(0.0000070)	(0.0000022)
male	5.4***	0.2***	0.1***
	(0.10)	(0.0027)	(0.0015)
_cons	-19.5***	0.8***	-0.2***
	(0.51)	(0.014)	(0.0055)
N	165789	165522	321576

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: See Dofile, Example5. Numbers in brackets report standard deviations of the coefficients.

Table 2: Coefficient interpretation

	y is Continuous	y is in log-terms	y is a Dummy
Dep.: Continuous	-0.5***	-2.2***	-0.5***
	(0.031)	(0.031)	(0.031)
Dep.: Dummy	-0.5***	0.3**	-0.5***
	(0.093)	(0.093)	(0.093)
Dep.: Log	0.3***	0.5***	0.3***
	(0.043)	(0.043)	(0.043)
_cons	4.9***	14.9***	4.9***
	(0.098)	(0.098)	(0.098)
N	472	472	472

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: In the Dofile, Example2 (Coefficient Interpretation), I generate three dependent variables. The first is continuous, the second is to be thought of as a log variable, and the third a dummy. See the respective equations for the underlying coefficient parameters I assume in the population equations. Numbers in brackets report standard deviations of the coefficients.

References

[1] J. D. Angrist and J. S. Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2008.

[2] A. C. Cameron and P. K. Trivedie. *Microeconometrics Using Stata*. Stata Press, 2010.

[3] R. Davidson and J. MacKinnon. *Econometric Theory and Methods*. Oxford University Press, 2004.

[4] W. H. Greene. *Econometric Analysis*. Prentice Hall International, 2011.

[5] D. Gujarati. *Basic Econometrics*. McGraw-Hill, 2003.

[6] J. H. Stock and M. W. Watson. *Introduction to Econometrics*. Addison Wesley, 2006.

[7] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.

[8] J. M. Wooldridge. *Introductory Econometrics*. Cengage Learning, 2009.