LECTURE #3

Econometrics I

OLS PROPERTIES

Jiri Kukacka, Ph.D.
Institute of Economic Studies, Faculty of Social Sciences, Charles University

Summer semester 2024, March 5

In the previous lecture #2

- We discussed the types of data analyzed in econometrics.
- We defined the **simple linear regression model**

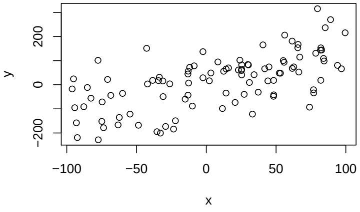
$$y = \beta_0 + \beta_1 x + u.$$

- From $\mathbb{E}(u) = 0$ and the **zero conditional mean assumption** $\mathbb{E}(u|x) = 0$, we got $\text{Cov}(x, u) = \mathbb{E}(xu) = 0$.
- We derived the **OLS estimators** (MM or LS approach):

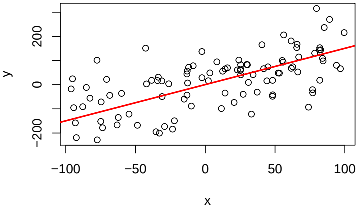
$$\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})} \quad \text{and} \quad \hat{\beta}_0^{OLS} = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- Alternatively: $\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.
- Readings for lecture #3:
  - Chapter 2: 2.3, 2.5, **2.6 (mandatory for/after seminars)**

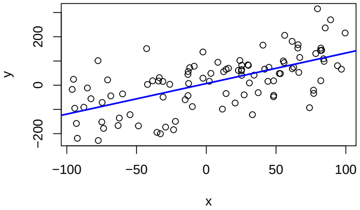
(a) $y = 0.5 + 1.5x + u, \; n = 100$



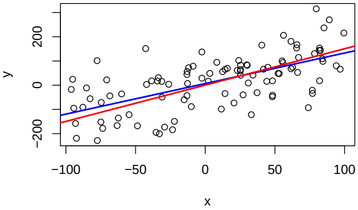
(b) PRF: $\mathbb{E}(y|x) = 0.5 + 1.5x$



(c) SRF (OLS RL): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



(d) SRF $\neq$ PRF



Outline

Outline

Basic OLS properties

Expected values and variances of the OLS estimators

Unbiasedness

Variance

Regression through the origin

Algebraic properties of the OLS statistics

1. Sum of the OLS residuals is zero:
$$\sum_{i=1}^{n} \hat{u}_i = 0.$$
2. Sample covariance between the explanatory variables and the OLS residuals is zero:
$$\sum_{i=1}^{n} x_i \hat{u}_i = 0.$$
3. Observed value of y can be split into two uncorrelated parts, such that
$$y_i = \hat{y}_i + \hat{u}_i \quad \text{and} \quad \sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0.$$
4. Sample means/averages of the observed and fitted values are equal:
$$\bar{y} = \bar{\hat{y}} \quad \text{or alternatively} \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$$
5. Point $(\bar{x}, \bar{y})$ is always on the OLS regression line.

Algebraic properties of the OLS statistics: Proofs

1.,2. First two are given by the MM and LS derivations of the estimators. In fact, $\hat{\beta}_1$ and $\hat{\beta}_1$ chosen to make them hold.

3. Observed value of y can be split into two uncorrelated parts:

$$\boxed{\hat{y}_i + \hat{u}_i} = \hat{y}_i + (y_i - \hat{y}_i) = \boxed{y_i},$$

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{u}_i = \hat{\beta}_0 \sum \hat{u}_i + \hat{\beta}_1 \sum x_i \hat{u}_i = 0.$$

4. Sample means/averages of the observed and fitted values are equal:

$$\boxed{\sum \hat{y}_i} = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) = n\bar{y} - n\hat{\beta}_1 \bar{x} + \hat{\beta}_1 \sum x_i =$$

$$= n\bar{y} - n\hat{\beta}_1 \bar{x} + n\hat{\beta}_1 \bar{x} = n\bar{y} = \boxed{\sum y_i}.$$

5. Point $(\bar{x}, \bar{y})$ is always on the OLS regression line:

$$\boxed{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \boxed{\bar{y}}.$$

Various 'sums of squares'

- Total sum of squares (SST)

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Explained sum of squares (SSE)

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residual sum of squares (SSR)

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

- It holds that

$$SST = SSE + SSR.$$

SST = SSE + SSR

- We need to use a little trick of 'adding zero' to the sum:

$$SST = \sum (y_i - \bar{y})^2 = \sum \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right)^2 =$$
$$= \sum (y_i - \hat{y}_i)^2 + 2 \sum \underbrace{(y_i - \hat{y}_i)}_{\hat{u}_i} (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 =$$
$$= SSR + 2 \sum \hat{u}_i (\hat{y}_i - \bar{y}) + SSE.$$

- We thus need to show that $\sum \hat{u}_i(\hat{y}_i - \bar{y}) = 0$:
  - in the algebraic properties of OLS, we have already shown that $\sum \hat{u}_i \hat{y}_i = 0$
  - and also $\sum \hat{u}_i = 0$ so that $\sum \hat{u}_i \bar{y} = \bar{y} \sum \hat{u}_i = \bar{y} \cdot 0 = 0$

Goodness-of-fit

- We need to measure how well our model (or now specifically variable x) explains the variation in y.
- **Coefficient of determination**, or **R-squared**, is defined

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

- R^2 can be interpreted (for the simple regression) as a fraction of the sample variation in y explained by x.
- R^2 ranges between 0 and 1 and is sometimes reported in percentages.
- Threshold values differ across disciplines and even across branches of economics and finance (usually data-type dependent).

Outline

Basic OLS properties

Expected values and variances of the OLS estimators

- Unbiasedness
- Variance

Regression through the origin

Outline

Unbiasedness of OLS

Simple linear regression (SLR) assumptions:

- **SLR.1 Linear in parameters:** We have the population model

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

  where β_0 is the population intercept and β_1 is the population slope parameter. The inclusion of β_0 implies $\mathbb{E}(u) = 0$.
- **SLR.2 Random sampling:** We have a random sample of size n following the population model.
- **SLR.3 Sample variation in the explanatory variable:** The sample outcomes on x are not all the same value.
- **SLR.4 Zero conditional mean:** The error u has an expected value of zero given any value of the explanatory variable, i.e., $\mathbb{E}(u|x) = 0$.

Unbiasedness of the OLS estimators

Assuming SLR.1 through SLR.4, $\mathbb{E}(\hat{\beta}_0^{OLS}) = \beta_0$ and $\mathbb{E}(\hat{\beta}_1^{OLS}) = \beta_1$ for any values of β_0 and β_1. In other words, $\hat{\beta}_0^{OLS}$ **is unbiased for** β_0 **and** $\hat{\beta}_1^{OLS}$ **is unbiased for** β_1.

Unbiasedness of the OLS estimator $\hat{\beta}_1$: Proof

- We first need to rewrite the OLS estimator (see lecture #2 Appendix) as

$$
\begin{aligned}
\boxed{\hat{\beta}_1} &= \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum (\beta_0 + \beta_1 x_i + u_i)(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \\
&= \beta_0 \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + \beta_1 \frac{\sum x_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + \frac{\sum u_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \\
&= 0 + \boxed{\beta_1 + \frac{\sum u_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}}.
\end{aligned}
$$

- OLS estimator $\hat{\beta}_1$ can thus be expressed as the true parameter β_1 plus an additional term, a linear combination of errors $\{u_1, u_2, \ldots, u_n\}$. This is where its stochasticity comes from.
- We now need to show its expected value.

Unbiasedness of the OLS estimator $\hat{\beta}_1$: Proof

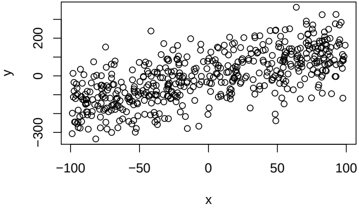
- We rewrite the OLS estimator as

$$\boxed{\mathbb{E}(\hat{\beta}_1)} = \beta_1 + \mathbb{E}\left(\frac{\sum u_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right) =$$

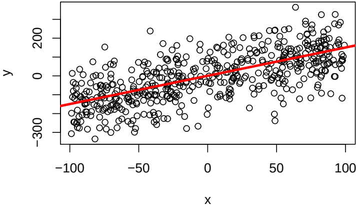
$$= \beta_1 + \frac{\sum \overbrace{\mathbb{E}(u_i x_i)}^{=0\ (SLR.4)}}{\sum (x_i - \bar{x})^2} - \frac{\bar{x} \sum \overbrace{\mathbb{E}(u_i)}^{=0\ (SLR.1)}}{\sum (x_i - \bar{x})^2} = \boxed{\beta_1}.$$

- OLS estimator $\hat{\beta}_1$ is thus unbiased (a feature of the sampling distribution!).
- Unbiasedness generally fails if any of the four assumptions SLR.1 through <u>SLR.4</u> fail!

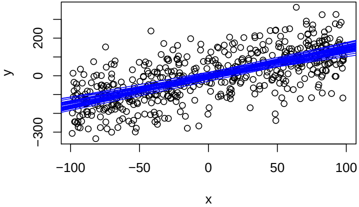
(a) $y = 0.5 + 1.5x + u, \ n = 500$



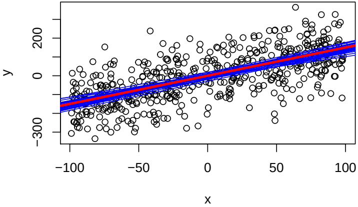
(b) PRF: $\mathbb{E}(y|x) = 0.5 + 1.5x$



(c) $30\times$SRF: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \ nn = 100$



(d) $\mathbb{E}(\hat{\beta}_0) = 0.5$ and $\mathbb{E}(\hat{\beta}_1) = 1.5$



Outline

Basic OLS properties

Expected values and variances of the OLS estimators

- Unbiasedness
- Variance

Regression through the origin

Variance of the OLS estimators

Additional assumption:

- **SLR.5 Homoskedasticity:** The error u has the same variance given any value of the explanatory variable, i.e.,

$$\mathrm{Var}(u|x) = \sigma^2.$$

- Homoskedasticity vs. heteroskedasticity
- SLR.5 implies $\mathrm{Var}(y|x) = \sigma^2$.

Variance of the OLS estimators

- It is also crucial to know how far we can expect $\hat{\beta}_1$ to be away from β_1 on average, i.e., how precise the estimator is.
- Assuming SLR.1 through SLR.5,

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \\
\mathrm{Var}(\hat{\beta}_0) &= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}.
\end{aligned}
$$

Variance of the OLS estimator $\hat{\beta}_1$: Derivation

- We will use the rewritten estimator $\hat{\beta}_1 = \beta_1 + \frac{\sum u_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$ as a starting point.
- As the variance of a parameter (constant) is zero, we can write

$$\boxed{\mathrm{Var}(\hat{\beta}_1)} = \mathrm{Var}\left(\frac{\sum u_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right) =$$

$$= \frac{1}{\left(\sum (x_i - \bar{x})^2\right)^2} \mathrm{Var}\left(\sum u_i(x_i - \bar{x})\right) \overset{SLR.4}{=}$$

$$= \frac{1}{\left(\sum (x_i - \bar{x})^2\right)^2} \sum \mathrm{Var}\big(u_i(x_i - \bar{x})\big) \overset{SLR.4}{=}$$

$$= \frac{1}{\left(\sum (x_i - \bar{x})^2\right)^2} \sum (x_i - \bar{x})^2 \mathrm{Var}(u_i) \overset{SLR.5}{=}$$

$$= \sigma^2 \frac{\sum (x_i - \bar{x})^2}{\left(\sum (x_i - \bar{x})^2\right)^2} = \boxed{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}.$$

Estimating the error variance

- σ^2 from the previous slides is not observed and hardly ever known $\Rightarrow$ it also needs to be estimated from data.
- Errors u (unknown) vs. residuals $\hat{u}$ (outcomes of the estimation procedure) $\Rightarrow$ we cannot use $\frac{\sum_{i=1}^{n} u_i^2}{n}$ as an estimator of σ^2.
- Under SLR.1 through SLR.5, **the unbiased estimator of σ^2**,

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2.$$

- $n-2$ because we lose two degrees of freedom due to two restrictions on residuals:

$$\sum_{i=1}^{n} \hat{u}_i = 0,$$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0.$$

Estimating the error variance

- $\hat{\sigma}$ is called the **standard error of the regression**.
- **Standard error of** $\hat{\beta}_1$ is then

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

- $se(\hat{\beta}_1)$ is necessary to construct test statistics and confidence intervals.

- Note: you can consult the attached R code (not mandatory) that compares the theoretical $sd(\beta_1)$ and estimated $se(\hat{\beta}_1)$ in simulations.

Outline

Regression through the origin

- In rare cases, assuming $\beta_0 = 0$, we are interested in a model

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

- Both the method of moments and the least squares estimation via minimizing $SSR = \sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_i)^2$ lead to

$$\sum_{i=1}^{n} x_i(y_i - \tilde{\beta}_1 x_i) = 0. \tag{1}$$

- Solving Eq. 1 leads to

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

- Iff $\bar{x} = 0$, then $\tilde{\beta}_1 = \hat{\beta}_1$.
- If $\beta_0 \neq 0$ then $\tilde{\beta}_1$ is biased $\Rightarrow$ not often used in practice.
- Mind the difference between R^2 of a standard regression and a regression through the origin!

Seminars and the next lecture

- Seminars:
  - interpretation of estimates and causality recap
  - SLR.5 (homoskedasticity) violation
  - regression through the origin: consequences
  - computer exercise with simulated data (BYOD?)
- Next lecture #4:
  - multiple regression model and OLS
  - expected value of the OLS estimators
    - unbiasedness
    - irrelevant variables
    - omitted variables
  - variance of the OLS estimators (multicollinearity)
- Readings for lecture #4:
  - Chapter 3: 3.1–3.4, 3.6 (**3.1** and **3.4, sections 'Multicollinearity' and 'Misspecified models' mandatory**)