# Market Analysis with Econometrics and Machine Learning

# 1b The simple linear regression model and the endogeniety problem

Uni Ulm

Prof. Dr. Sebastian Kranz

SoSe 2020



## The simple linear regression model

- A simple linear regression model satisfies the following relationship for all observations $t = 1, \ldots, T$

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

- $y = (y_1, \ldots y_T)$ is the dependent variable.
- $x = (x_1, \ldots x_T)$ is the explanatory variable.
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)$ (epsilon) is a random variable that describes unobserved influences on $y$, sometimes called *disturbance*. We will typically make some assumptions on the distribution of $\varepsilon$. We will also use the letters $u$ and $\eta$ (eta) to denote disturbances.
- $\beta = (\beta_0, \beta_1)$ is the vector of true coefficients.



## Estimate, Predicted Value and Residuum

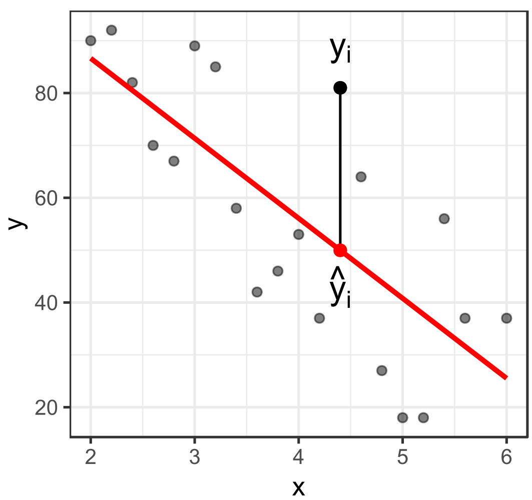
- Let $\hat{\beta}$ be an *estimate* of the true parameter vector $\beta$.
- The *predicted values* (also called fitted values) of $y$ are given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The *residuals* (estimated values of the disturbance) are given by

$$\hat{\varepsilon} = y - \hat{y} = y - \hat{\beta}_0 - \hat{\beta}_1 x$$

  - The residuals $\hat{\varepsilon}$ are close to the true disturbances $\varepsilon$ if our estimate $\hat{\beta}$ is close to the true parameters $\beta$.



## Ordinary Least Squares Estimation (OLS)



- An ordinary least squares (OLS) estimate minimizes the sum of squared residuals

$$\hat{\beta} = \arg\min \sum_{t=1}^{T} \hat{\varepsilon}_t^2$$

- For the simple linear regression (one explanatory variable), the OLS estimator $\hat{\beta}_1$ has the following formula

$$\hat{\beta}_1 = \frac{Cov(x_t, y_t)}{Var(x_t)} = cor(x_t, y_t)\frac{sd(y)}{sd(x)}$$

  where $cor$ denotes an empirical correlation and $sd$ an empirical standard deviation for our sample data.

## Linear Regression Model in Matrix Notation



- One often writes a linear regression model in matrix notation:

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

  with

$$X = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_T \end{pmatrix} = \begin{pmatrix} \mathbf{1} & x \end{pmatrix}$$

- The OLS estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ is then given by

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

- You don't have to understand the matrix notation for this lecture, but for those who do, we will sometimes mention it.

## Estimators and estimates



- Since $y = X\beta + \varepsilon$, the OLS estimator can be rewritten as

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

  - This means $\hat{\beta}$ is a linear transformation of the true parameters $\beta$ and the disturbance $\varepsilon$
- As a function of a random variable $\varepsilon$ the OLS *estimator* $\hat{\beta}$ is itself a random variable
- The OLS *estimate* $\hat{\beta}$ is a realization of the OLS estimator, i.e. the value for particular draws of $\varepsilon$ and $X$.
- To understand what econometrics and most of statistics is doing, one should keep in mind that *an estimator is a random variable*.

## A Monte-Carlo Simulation of $\hat{\beta}$ in R

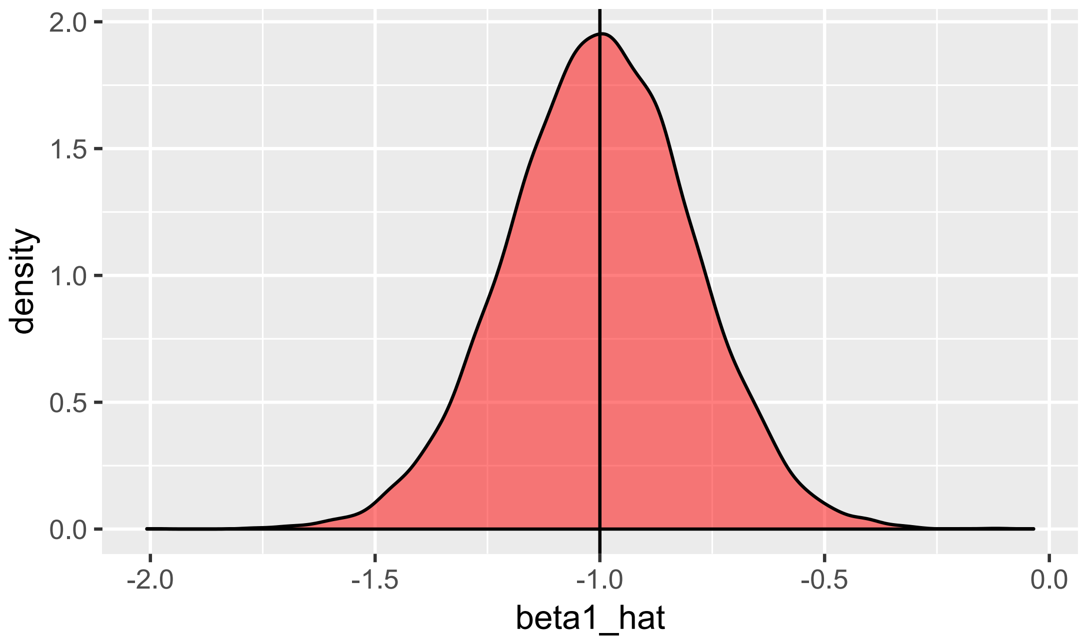
*Analysis in R* (R analysises would be shown life in the lecture and I made videos for some. You will do similar analysis yourself in an RTutor problem set)

- Start with the simulation of our ice cream model from Chapter 1a with random prices and estimate the demand function.
- Write an R function with name `sim.and.est` that performs this simulation and returns the estimated coefficients $\hat{\beta}$ of the demand function. Your function shall have the sample size $T$ as an argument.
- Repeat the simulation and estimation several times (draw new $\varepsilon$ each time) and compute and store the OLS estimates. Use the function `simulation.study` in the package `sktools` to conduct a systematic simulation study. Plot the distribution of the resulting estimates $\hat{\beta}_1$ and compare it with the true value of $\beta_1$. How does the distribution change if we change the sample size $T$?



## Distribution of our estimator $\hat{\beta}_1$

We found in our Monte-Carlo simulation study that the estimator $\hat{\beta}_1$ has a distribution, that depends on the sample size $T$. Here it is shown for $T = 20$ and a true $\beta_1 = -1$



## Standard Error of OLS estimator



- In a simple linear regression (one explanatory variable)

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

where the $\varepsilon$ are independently, identically normal distributed, the standard deviation of the OLS estimator $\hat{\beta}_1$ can be *estimated* by

$$se(\hat{\beta}_1) = \hat{sd}(\hat{\beta}_1) = \frac{1}{\sqrt{T}} \frac{sd(\hat{\varepsilon})}{sd(x)}$$

- We call this estimate of the standard deviation the *standard error* of $\hat{\beta}_1$.
- Observations: We can estimate $\beta_1$ more precisely if we have...
  - a larger the sample size $T$
  - more variation in $x$ (higher standard deviation).

*Analysis in R:* We run a linear regression with `lm` and call `summary` on the result to see besides the estimated coefficients also the standard errors.



## Robust Standard Errors

- There is also a matrix formula to compute the standard errors for all $\hat{\beta}$ that can also be used for multiple linear regressions with more than one explanatory variable.
- If the $\varepsilon$ are not identically, independently normal distributed, one should use approbriate *robust* standard errors. Most empirical papers in economics use some robust standard errors.
- We don't explain robust standard errors further in this course. Just note that in R a convenient way to use robust standard errors is the function `lm_robust` in the package `estimatr` or the function `felm` in the package `lfe` .



## Criteria for estimators: Bias

- **Bias:** Recall that an estimator $\hat{\beta}$ is a random variable since it depends on the realizations of $\varepsilon$. Let $E\hat{\beta}$ be the expected value of $\hat{\beta}$. The bias of $\hat{\beta}$ measures a systematic over- or underestimation of $\hat{\beta}$ compared to $\beta$:

$$Bias(\widehat{\beta}) = E\hat{\beta} - \beta.$$

- **Unbiasedness:** An estimator $\hat{\beta}$ is unbiased if its Bias is 0, i.e.

$$E\hat{\beta} = \beta$$



## Criteria for estimators: Standard Deviation

- For two unbiased estimators of $\beta_i$, one would typically prefer an estimator with a lower standard deviation $sd(\hat{\beta}_i)$ (or equivalently the one with the lower variance $Var(\hat{\beta}_i)$)

## Criteria for estimators: Mean Squared Error



- **Mean squared error**: The mean squared error of $\hat{\beta}_i$ is given by

$$
\begin{aligned}
MSE(\hat{\beta}_i) &= E(\hat{\beta}_i - \beta_i)^2 \\
&= Bias(\hat{\beta}_i)^2 + Var(\hat{\beta}_i)
\end{aligned}
$$

- *Analysis in R:* We go back to the previous simulation study in R and compute an approximation to the bias and mean squared error of $\hat{\beta}_0$ by analysing the simulated distribution of $\hat{\beta}_0$. How does the MSE change when you increase the number of observations $T$ in the simulated data set?

## Criteria for estimators: Consistency



- An estimator $\hat{\beta}$ is (strongly) **consistent** if its MSE converges to 0 as the sample size $T$ grows large

$$\lim_{T\to\infty} MSE(\hat{\beta}) = 0.$$

- A note for the mathematic students: Strong consistency implies weak consistency, which means that the estimated parameters $\widehat{\beta}$ converges (in probability) to the true parameters $\beta$

$$\operatorname*{plim}_{T\to\infty} \widehat{\beta} = \beta$$

- **Consistency is often seen the most important requirement for an estimator.**
- If an estimator is inconsistent that is typically because it is biased and the bias does not go away as $T \to \infty$.

## Criteria for estimators: Efficiency



- An estimator $\hat{\beta}$ is **efficient** (within a specified class of estimators) if there is no other estimator that has a lower mean squared error.
- We won't discuss efficiency deeper in this course.

## Assumptions of the simple linear regression model



- We now state a series of assumptions for the simple linear regression model (one explanatory variable).
  - A1: $E(\varepsilon_t|x) = 0$
  - A2: The $\varepsilon_t$ are identically and independently distributed.
  - A3: The $\varepsilon_t$ are normally distributed
  - A1-A3 are often compactly written as $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
  - A4: The explanatory variable $x$ must have positive variance and be deterministic or a stationary random variable. (We don't discuss what stationary means in this course, but you can look it up on Wikipedia.)
- If all assumptions are satisfied, the OLS estimator $\hat{\beta}$ will be consistent, unbiassed and efficient.
- Good intuitive overview: “An Introduction to Econometrics” by Peter Kennedy

## Assumption A1



- **A1** No matter which values of $x$ we observe, the conditional expected value of $\varepsilon_t$ is always zero:

$$E(\varepsilon_t|x) = 0$$

- The important thing is not the 0 on the right. If it were a positive or negative value, we could always redefine the constant $\beta_0$ to make it 0.
- The important thing is that the expected value of $\varepsilon_t$ does not depend on $x$. This means knowing $x$ shall give us no information about the expected value of $\varepsilon_t$.
- In our ice cream example with profit maximizing prices this condition is violated. Higher demand shocks lead to higher prices. This means if we observe a high price, we expect that there was a positive demand shock $\varepsilon_t$.

## Exogenous and Endogenous Variables



- We say the explanatory variable $x$ is **exogenous**, if it is uncorrelated with $\varepsilon$

$$\mathrm{cor}(x_t, \varepsilon_t) = 0$$

- We say $x$ is **endogenous** if $\mathrm{cor}(x_t, \varepsilon_t) \neq 0$
- Condition A1 $E(\varepsilon|x) = 0$ can only be satisfied if $x$ is exogenous.
- We will typically just check whether $x$ is exogenous, even though A1 is a stronger condition. A1 is sometimes called *strong exogeniety*. In all examples studied in this course, exogeniety of $x$ implies that also A1 holds.

## The problem of endogeniety



- **A1 is the most important assumption:**
  - **If $x$ is endogenous, the OLS estimator $\hat{\beta}$ will (typically) be inconsistent and biased.**
- We typically check whether there is endogeniety as follows:
  - Think of a true model for the data generating process; specify which factors are part of the random shock $\varepsilon$
  - See if in that model the explanatory variable $x$ is a function of $\varepsilon$ or of some unobserved variable that is part of $\varepsilon$. If that is the case $x$ is endogenous.

## Exploring Endogeniety in R



*Analysis in R*: We consider the ice cream model from slides 1a.

- We compare the two cases that prices are randomly drawn and prices are chosen optimally. In which model is A1 satisfied / violated and the OLS estimator consistent / inconsistent? We compute the correlation between $\varepsilon$ and $p$ in your simulated data.
- For real world data, we never observe $\varepsilon$, i.e. we cannot simply see in the data that there is an endogeniety problem. If we compute the correlation between $p$ and $\hat{\varepsilon}$ (the OLS residual), we see that it is always 0 (or due to rounding errors maybe slightly away) even if $\varepsilon$ and $p$ are correlated.
- We now assume prices $p_t$ are simply always set 10% above the cost $c_t$ (and costs shall be uncorrelated with $\varepsilon$). Are prices $p_t$ then endogenous or exogenous? Is the OLS estimator consistent?
- We now consider the model in which prices are chosen optimal. Consider in your simulation the limit case that $\sigma_\varepsilon \rightarrow 0$. What is then the main source of variation in prices? To which value will $cor(\varepsilon, p)$ converge? We study in the simulation whether the OLS estimator of prices seems to be consistent and unbiased in this limit case.

## A2, No auto-correlation and no heteroskedasticity



- **A2** The $\varepsilon_t$ are identically and independently distributed.
- Typical violations of A2:
  - auto-correlation: demand shocks may be persistent across periods
  - heteroskedasticity: the variance of $\varepsilon_t$ can depend on the explanatory variable (this alone does not yet mean that A1 is violated)
- A2 is moderately important. If violated, the OLS estimator $\hat{\beta}$ is still consistent but not efficient. One must calculate standard errors using an appropriate formula for robust standard errors.
- We don't study violations of A2 in this course.

## A3: Normally distributed disturbances



- **A3** $\varepsilon_t$ is normally distributed
- It is nice if A3 holds, but it is not crucial. Even if A3 is violated, the OLS estimate $\hat{\beta}$ is the best unbiased linear estimators of $\beta$ (Gauss-Markov Theorem). Significance tests would only be asymptotically correct. If A1-A3 (and the other assumptions) holds, $\hat{\beta}$ coincides with Maximum Likelihood estimator and is efficient.

# 95% Confidence Intervals



- If assumptions A1 holds (no endogeniety problem) then with approximately 95% probability we find an estimate $\hat{\beta}_i$ such that the interval of plus-minus 2 standard errors around $\hat{\beta}_i$ contains the true parameter $\beta_i$. We call this interval

$$[\hat{\beta}_i - 2 \cdot se(\hat{\beta}_i) \ ; \ \hat{\beta}_i + 2 \cdot se(\hat{\beta}_i)]$$

  the approximate **95% confidence interval**.

*Analysis in R*

1. Run a linear regression on some simulated data without endogenitey problem and apply the R function `summary` on the result to see the estimated standard errors. Compute in your head the approximate 95% confidence interval and also use the function `confint` for the exact confidence interval.
2. Now simulate data with an endogeniety problem. Does the true parameter $\beta_1$ still typically lie within the 95% confidence interval around $\hat{\beta}_1$ or can it be that it is almost always very far away from it?

# Bias Formula



- Consider a simple linear regression

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

  and assume we would observe $\varepsilon$.
- One can show that

$$\hat{\beta}_1 - \beta_1 = cor(x, \varepsilon) \frac{sd(\varepsilon)}{sd(x)}$$

  using the sample correlations and sample standard deviations.
- This expression is an estimator of the bias of $\hat{\beta}_1$. (The actual bias is the expected value of it.)
- Thus essentially the bias has the same sign as the correlation between $x$ and $\varepsilon$.

*Analysis in R*

- Simulate a demand model with endogenous prices and check for your simulation that the "bias formula" above holds.