Data Analytics II: Summary

Yvan Richard

2025-05-26

Contents

Foreword

This script summarizes the course *Data Analytics II* from the University of St. Gallen. This course was taught by Johanna Kutz and Beatrix Eugster. Furthermore, the following literature was relevant:

- Wooldridge, J. M. (2016). Introductory Econometrics: A Modern Approach, 6th ed. Cengage learning.

The purpose of this script is to summarize the theoretical insights from the slides and to cover some **relevant** practical examples to illustrate the theory. The reader should be warned that this script exhaustively covers the content presented in the slides but do not fully address the exercises and assignment.

1. Introduction & Statistics Refresher

Content

- The structure of the course
- Statistics Refresher
- An introductory example of an empirical analysis

Relevant Materials

- Wooldridge Chapter 1
- Slides L01

1.1. Introductory Example

Let us suppose that we have the following data:

```r
# variables
wages <- c(1, 1, 2, 3, 7, 10, 12, 19) # salary in thousands
education <- c(0, 1, 9, 9, 9, 12, 12, 11) # years of schooling

# data frame
df <- data.frame(wage = wages, education = education)
head(df, n = 8)
```

```
##   wage education
## 1    1         0
## 2    1         1
## 3    2         9
## 4    3         9
## 5    7         9
## 6   10        12
## 7   12        12
## 8   19        11
```

And we want to know if more years of schooling *mean*, *imply* a higher salary on average. How should we proceed? Well, this is the typical setting of an **empirical analysis**. Indeed, an empirical analysis uses data to test a theory or to estimate a relationship. The basics of an econometric model is oftentimes constructed this way:

$$y = f(x_1, x_2, \dots, x_n)$$

where we want to study the influence of the **independent variables** x_i (e.g. `education`) on the *dependent variable* y (e.g. wages). To proceed, we will oftentimes (in this course) build a **linear model** (linear since it is linear in its parameters) which we can easily interpret, in our case, we have:

$$E[wage \mid education] = \beta_0 + \beta_1 education$$

where β_0 and β_1 are the parameters of the model. We can estimate them with the built-in function `lm()`:

```r
# we fit the model:
model_11 <- lm(wage ~ education, data = df)

# we look at the coefficient
coefs <- model_11$coefficients
print(coefs)
```

```
## (Intercept)   education
##  -0.5984064   0.9490040
```

Interpretation

Here, we see that with 0 years of education, a worker earns, on average, USD -598 (maybe he is subsidized by the state). Furthermore, we observe that since $\beta_1 > 0$, we could imply that **ceteris paribus** a higher education implies a higher wage (one additional year of schooling increases the monthly wage by roughly USD 949). However, before making **causal** links like those, we have to make a lot of assumptions and cover a lot of other concepts!

Non-observable Factors

When we estimate this kind of relationships, there is always some *noise*, some information we cannot capture and include in the model. This will generate imperfect estimates. Hence, we have this model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \underbrace{u}_{\text{error}}$$

where u is the **error term** and represent what we cannot capture (e.g. skills, IQ, $\ldots$).

1.2. Statistics Refresher

In this section, we cover the key notion of **expected value**, **variance**, **covariance**, and **correlation**.

1.2.1. Expected Value of a Discrete Random Variable

In this course, we will mostly work with discrete random variables (our previous example was realized with discrete random variables).

Definition of $E[X]$ According to Wikipedia, in probability theory, the expected value (also called expectation, expectancy, expectation operator, mathematical expectation, mean, expectation value, or first moment) is a **generalization of the weighted average**. Informally, the expected value is the mean of the possible values a random variable can take, weighted by the probability of those outcomes. We have the following notations:

Expected Value: $E[X]$

$$E[X] = \sum_{i=1}^{n} x_i f(x_i)$$

where $f(x_i)$ is the probability mass function of the random variable X. x_i are the realizations of the random variable X.

Expected Value of a Function: $E[g(X)]$

$$E[X] = \sum_{i=1}^{n} g(x_i) f(x_i)$$

Expected Value of a Linear Construction: $E[Y], \quad Y = a + bX$

$$E[Y] = \sum_{i=1}^{n} (a + b \cdot x_i) f(x_i) = a \sum_{i=1}^{n} f(x_i) + b \cdot \sum_{i=1} x_i f(x_i) = a + bE[X]$$

Conditional Expected Value: $E[Y|X = x]$

$$E[Y \mid X = x] = \sum_{i=1}^{n} y_i f(y_j \mid X = x)$$

The Law of Iterated Expectations The law of iterated expectations is a fundamental result in probability theory and is especially useful in this course since we rely on it to demonstrate numerous theorems. Hence, I will try to provide the best possible explanation of it.

First, the law states that:

$$E[Y] = E[E[Y \mid X]]$$

where X and Y are any random variables (Y must have a defined expected value of course). Basically, this means that:

> To find the overall average of a random quantity, we can first average it within each subgroup defined by another variable, and then average those subgroup-averages, weighted by how likely each subgroup is.

Informal Proof

We suppose X takes values $x_1, \ldots, x_K$ and Y takes values in some set. By definition of conditional expectation for discrete variables,

$$\mathbb{E}[Y \mid X = x_k] \; = \; \sum_{y} y \; P\big(Y = y \mid X = x_k\big).$$

Then the law of total probability gives (see conditional PMF for discrete random variables and definition of a conditional probability):

$$\mathbb{E}[Y] = \sum_{y} \; y \; P(Y = y) \; = \; \sum_{y} \; y \; \sum_{k=1}^{K} P(Y = y, \, X = x_k) \; = \; \sum_{k=1}^{K} \sum_{y} \; y \; P(Y = y \mid X = x_k) \, P(X = x_k).$$

But:

$$\sum_{y} y \, P(Y = y \mid X = x_k) = \mathbb{E}[Y \mid X = x_k]$$

So:

$$\mathbb{E}[Y] = \sum_{k=1}^{K} \mathbb{E}[Y \mid X = x_k] \; P(X = x_k) = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big].$$

Example Now, let us suppose that we throw a dice, we have $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then, we map a random variable:

$$X : w \to x(w) \in \mathbb{R}$$

with the following payoffs:

$$x(w) = \begin{cases} 10 \text{ if } w \in \{1, 2\} \\ 20 \text{ if } w \in \{3, 4\} \\ 40 \text{ if } w \in \{5\} \\ 50 \text{ if } w \in \{6\} \end{cases}$$

What is the expected return? What is the expected return, if we know that at least a 3 has been rolled? We solve:

```
# question 1

# the basic probabilities:
proba_1 <- rep(1/6, 6)
outcomes <- c(10, 10, 20, 20, 40, 80)

# the expected value:
exp_1 <- sum(proba_1 * outcomes)
cat("The expected value is: ", exp_1)
```

```
## The expected value is:  30
```

Then, we have a "new" $\Omega' = \{3, 4, 5, 6\}$, thus $E[Y|X \geq 3]$:

```
# new probabilities:
proba_2 <- rep(0.25, 4)
outcomes_2 <- c(20, 20, 40, 80)

# expected value:
cat("The conditional expected value is: ", sum(proba_2 * outcomes_2))
```

```
## The conditional expected value is:  40
```

1.2.2. Variance of a Random Variable

Definition The variance of a random variable is the expected value of the squared deviation from the mean of a random variable. The standard deviation (SD) is obtained as the square root of the variance. Variance is **a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value**. It is the second central moment of a distribution, and the covariance of the random variable with itself.

$$Var(X) = \sigma_X^2 = E[(X - E[X])^2] = E[X^2] - (E[X])^2 = Cov(X, X)$$

When we are in discrete settings, we have:

$$Var(X) = \sum_{i=1}^{n} x_i^2 \cdot f(x_i) - (\sum_{i=1}^{n} x_i \cdot f(x_i))^2$$

Variance of a Linear Function The variance of a linear function is given by:

$$Var(Y) = Var(a + bX) = b^2 Var(X)$$

Proof:

$$Var(a + bX) = E[(a + bX)^2] - (a + bE[X])^2 = a^2 + 2abE[X] + b^2E[X^2] - a^2 - b^2(E[X])^2 - 2abE[X]$$

Finally:

$$= b^2E[X^2] - b^2(E[X])^2 = \boxed{b^2Var(X)}$$

1.2.3. Standardization

The concept of normalization emerged alongside the study of the normal distribution by Abraham De Moivre, Pierre-Simon Laplace, and Carl Friedrich Gauss from the 18th to the 19th century. As the name "standard" refers to the particular normal distribution with expectation zero and standard deviation one, that is, the standard normal distribution, normalization, in this case, "standardization", was then used to refer to the rescaling of any distribution or data set to have mean zero and standard deviation one. Wikipedia, Normalization.

To obtain the standard score Z (a new random variable), we have:

$$Z := \frac{X - E[X]}{\sigma}$$

where Z has $E[X] = 0$ and $Var[Z] = 0$. It is not normally distributed! In practice, we use **unbiased estimates** to compute the variance and the expected value:

$$\hat{\mu} := \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \hat{\sigma} := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We now illustrate this by generating synthetic income data X. We distribute $X \sim$ uniform(4500, 9000) and we generate 1000 realizations:

```r
set.seed(123)

# synthetic data
sample_size <- 1000
incomes <- runif(n = sample_size, min = 4500, max = 9000)

# we compute the mean and the sd:
my_mean <- function(xs){
  total_sum = 0
  n = 0
  for (x in xs){
    total_sum <- total_sum + x
    n <- n + 1
  }
  return(total_sum/n)
}
```

```r
cat("The mean is : ", my_mean(incomes))
```

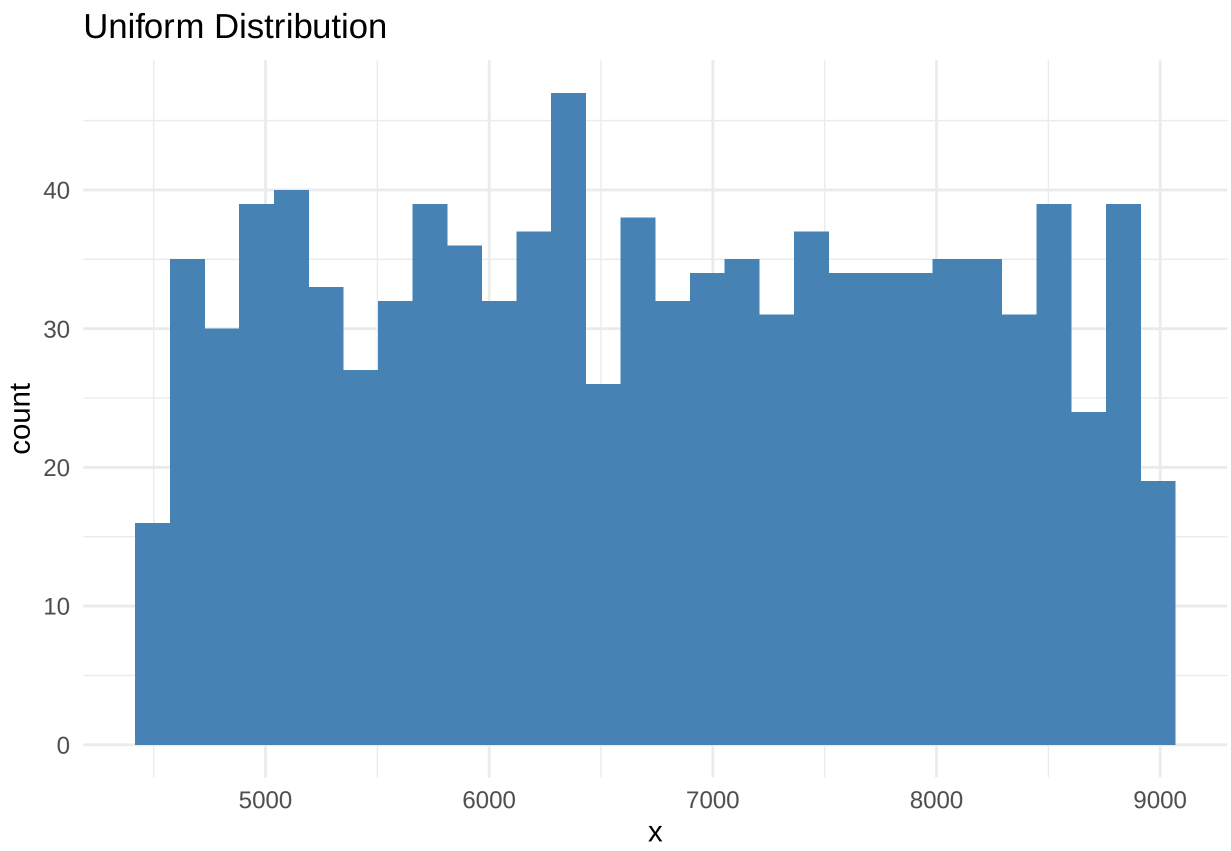
```
## The mean is :  6737.75
```

```r
# the sd
my_sd <- function(xs){
  x_bar <- my_mean(xs)
  square_deviation = 0
  n = 0
  for (x in xs){
    square_deviation <- square_deviation + (x - x_bar)^2
    n <- n + 1
  }

  sd = sqrt(1/(n - 1) * (square_deviation))
  return(sd)
}

cat("The standard deivation is: ", my_sd(incomes))
```

```
## The standard deivation is:  1293.678
```

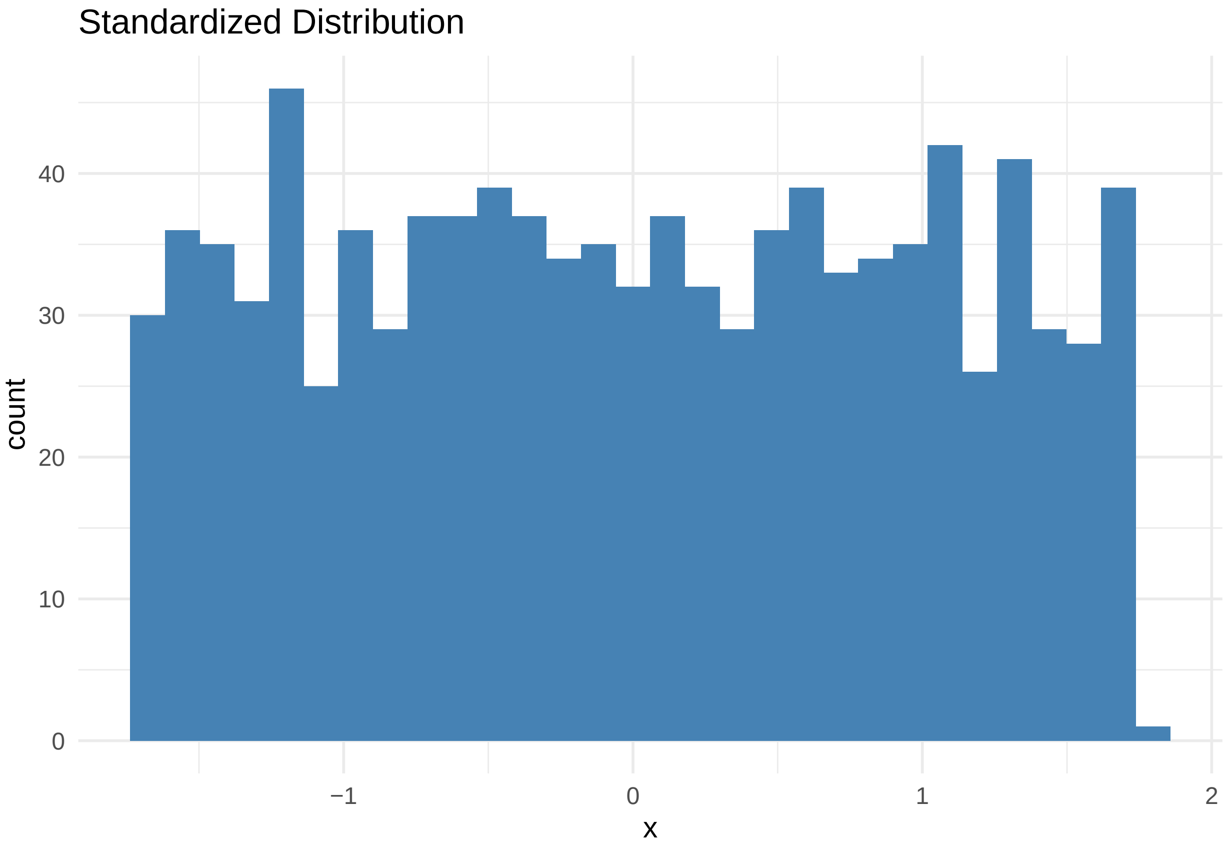
We can plot the original distribution:

```r
ggplot(data = data.frame(x = incomes), aes(x)) +
  geom_histogram(fill = "steelblue") +
  labs(title = "Uniform Distribution") +
  theme_minimal()
```



And the standardized:

```
z = (incomes - my_mean(incomes)) / my_sd(incomes)

ggplot(data = data.frame(x = z), aes(x)) +
  geom_histogram(fill = "steelblue") +
  labs(title = "Standardized Distribution") +
  theme_minimal()
```



1.2.4. Covariance and Correlation

Definition The covariance is a measure of the joint variability of two random variables.

The sign of the covariance, therefore, shows the tendency in the linear relationship between the variables. If greater values of one variable mainly correspond with greater values of the other variable, and the same holds for lesser values (that is, the variables tend to show similar behavior), the covariance is positive.[2] In the opposite case, when greater values of one variable mainly correspond to lesser values of the other (that is, the variables tend to show opposite behavior), the covariance is negative. The magnitude of the covariance is the geometric mean of the variances that are in common for the two random variables. The correlation coefficient normalizes the covariance by dividing by the geometric mean of the total variances for the two random variables.

Wikipedia, Covariance

We have:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X][Y]$$

note: if two random variables are independent, we have:

$$E[XY] = E[X]E[Y] \implies Cov(X,Y) = 0$$

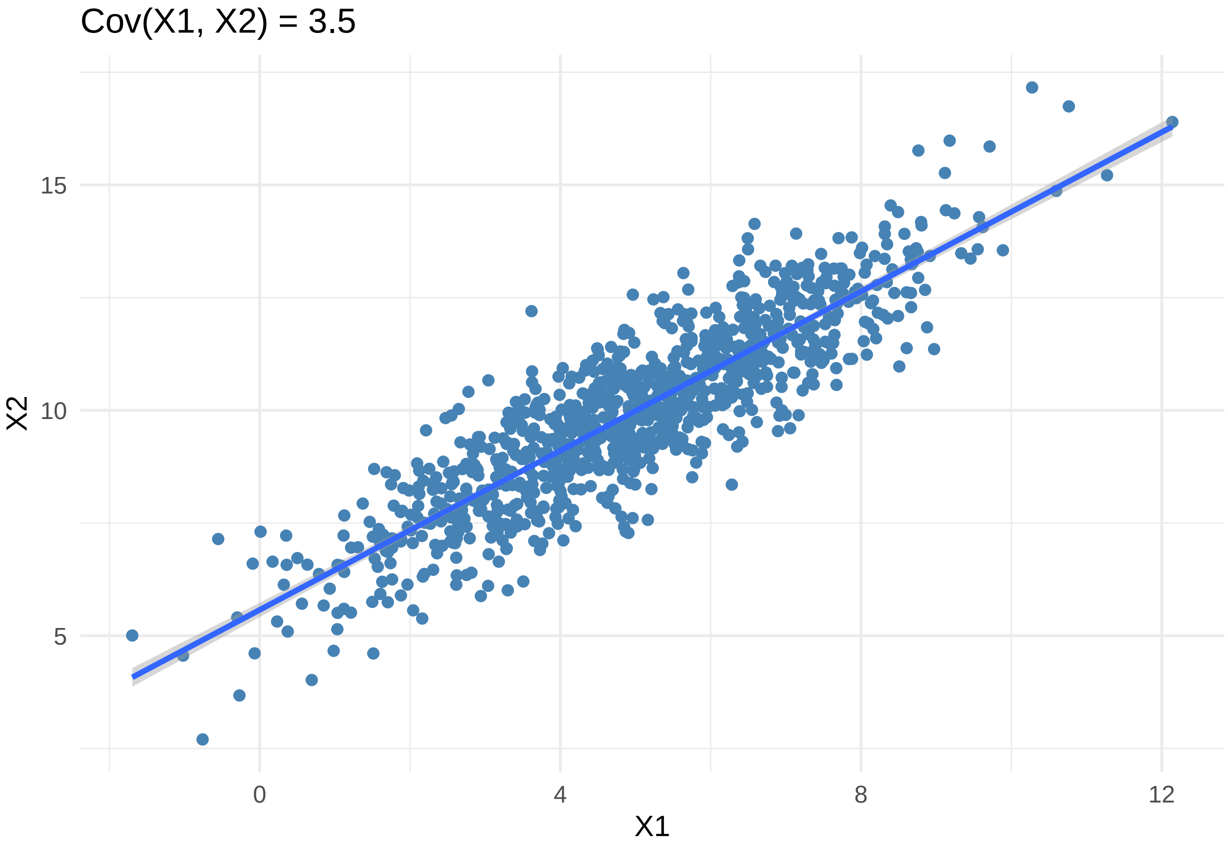
Then, we have the correlation coefficient $\rho_{X,Y} \in [0,1]$:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

In practice, we use the estimation methods displayed above.

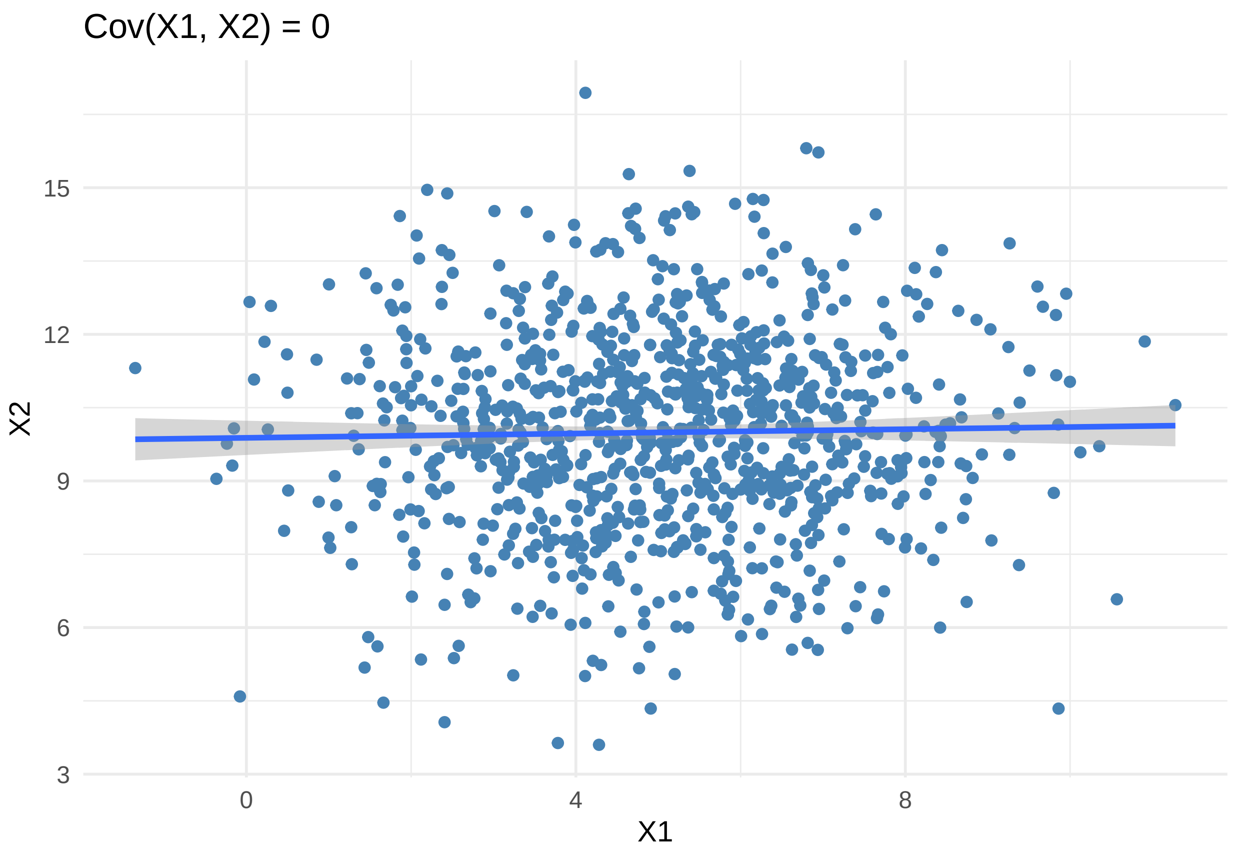
Illustrations Here are two illustrations I generated with the function `mvrnorm()`:

High Covariance



In this case $\rho_{X_1,X_2} > 0$, we speak of a **positive correlation**.

No Covariance



1.3. First Steps Empirical Analysis

1.3.1 Data Types

In econometrics, we encounter four main types of data. Here is a brief presentation of the different one.

Cross-Sectional Data A cross-sectional data set consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, **taken at a given point in time**. If the data on all units do not correspond to the exact same time, we say that the **granularity** is not the same. However, we still view these data as cross-sectional. An important feature of cross-sectional data is that we can often assume that they have been obtained by *random sampling* from the underlying population.

Example:

```
df <- data.frame(
  id          = 1:5,
  age         = c(25, 34, 29, 41, 38),
  gender      = factor(c("M", "F", "F", "M", "F")),
  income      = c(5500, 7200, 4800, 8600, 6300),
  edu_years   = c(16, 14, 12, 18, 16)
)

head(df, n = 5)
```

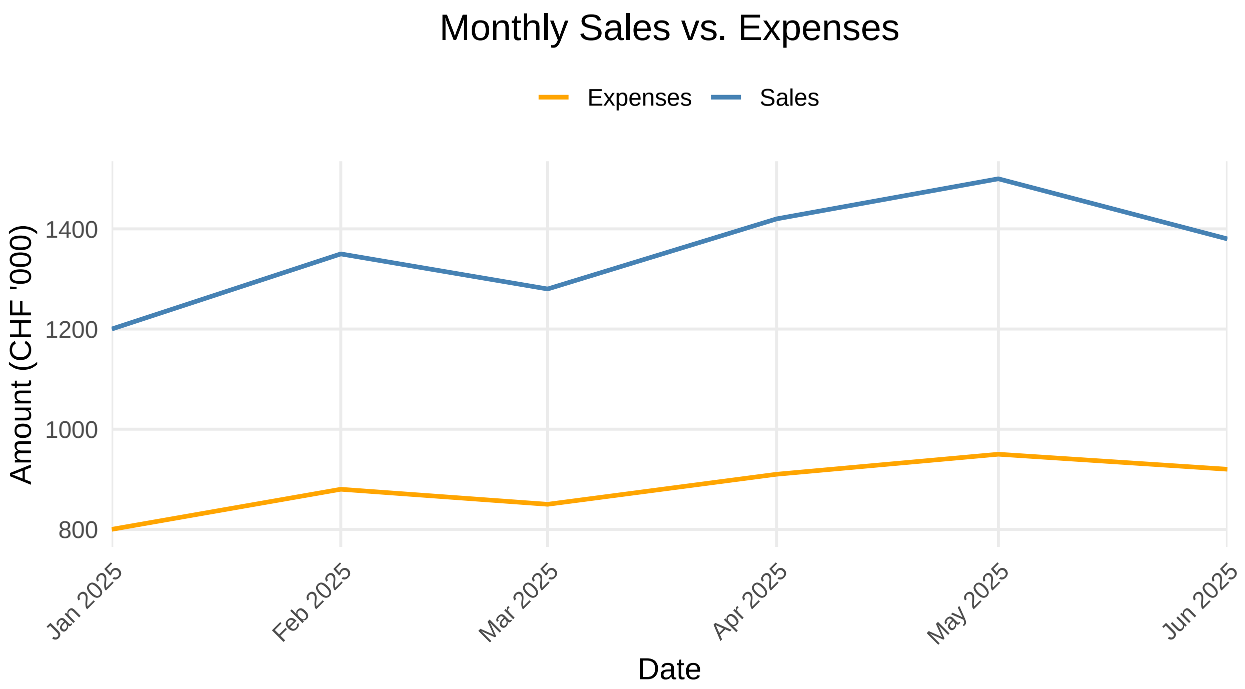
```
##   id age gender income edu_years
## 1  1  25       M   5500         16
## 2  2  34       F   7200         14
## 3  3  29       F   4800         12
## 4  4  41       M   8600         18
## 5  5  38       F   6300         16
```

Time Series Data A time series data set consists of observations on a variable or several variables over time. Examples of time series data include stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, and automobile sales figures. Because past events can influence future events and lags in behavior are prevalent in the social sciences, time is an important dimension in a time series data set. Unlike the arrangement of cross-sectional data, the chronological ordering of observations in a time series conveys potentially important information. Another difficulty of time series data is that each observation is (most of the time) dependent on the precedent ones (i.e. the past influences the future). For this reason, we might use some modelling techniques specially used for time-series (e.g. ARMA model).

Example:

```
# Fake monthly time-series data
# Fake monthly time-series data
df_ts <- data.frame(
  date     = seq.Date(as.Date("2025-01-01"), by = "month", length.out = 6),
  sales    = c(1200, 1350, 1280, 1420, 1500, 1380),  # in CHF ’000
  expenses = c( 800,  880,  850,  910,  950,  920)   # in CHF ’000
)

head(df_ts, n = 6)
```

```
##         date sales expenses
## 1 2025-01-01  1200      800
## 2 2025-02-01  1350      880
## 3 2025-03-01  1280      850
## 4 2025-04-01  1420      910
## 5 2025-05-01  1500      950
## 6 2025-06-01  1380      920
```



Pooled Cross-Sectional Data Some data sets have both cross-sectional and time series features. For example, suppose that two cross-sectional household surveys are taken in the United States, one in 1985 and one in 1990. In 1985, a random sample of households is surveyed for variables such as income, savings, family size, and so on. In 1990, a new random sample of households is taken using the same survey questions. To increase our sample size, we can form a pooled cross section by combining the two years. Therefore, pooling cross sections from different years is often an effective way of analyzing the effects of a new government policy (we observe the data before and after the policy change). We will use this kind of data set in the last assignment, when we discuss Differences in Difference estimates.

Panel Data A panel data (or longitudinal data) set consists of a time series for each cross-sectional member in the data set. For instance, we might collect information, such as investment and financial data, about the same set of firms over a five-year time period. The key feature of panel data that distinguishes them from a pooled cross section is that the same cross-sectional units (individuals, firms, or counties in the preceding examples) are followed over a given time period. We should bear in mind that the analysis of panel data is hard in comparison to the previous data structures.

```r
# 1) "long" panel structure
ids   <- 1:3                # three firms
years <- 2020:2023          # four years

df_panel <- expand.grid(
  firm = ids,
  year = years
)

set.seed(2025)

# some fake outcomes
df_panel$revenue <- round(runif(nrow(df_panel), 100, 200), 1)
```

```
df_panel$cost <- round(runif(nrow(df_panel),  50, 150), 1)
df_panel$profit <- df_panel$revenue - df_panel$cost

# head
head(df_panel)
```

```
##   firm year revenue  cost profit
## 1    1 2020   173.3 115.4   57.9
## 2    2 2020   147.6  52.4   95.2
## 3    3 2020   151.4  96.8   54.6
## 4    1 2021   149.8 135.7   14.1
## 5    2 2021   178.0  86.5   91.5
## 6    3 2021   150.4  86.6   63.8
```

2. Univariate Linear Regression Model (1/3)

Content

- Definition of the univariate linear regression model
- The estimation method: Ordinary Least Squares (OLS)

Materials

- Wooldridge Chapter 2.1. and 2.2.
- Slides L02

2.1. Assumptions & Definitions

In this section, our purpose is to find a relationship that allows us to explain the behavior of y, the target variable, in function of x, the independent variable. A simple equation could capture this relationship as explicited in the following assumption.

2.1.1. Assumption SLR. 1: Linear Regression

> We make the assumption that the random variables X and Y are related by the following relationship:
> $$Y = \beta_0 + \beta_1 X + u$$

This is the so-called **simple linear regression model** (SLR) or more commonly, the bivariate regression model. where:

- β_0 is the **intercept**
- β_1 is the **slope**
- Y is the **dependent variable**
- X is the **independent variable**
- u is the error term and captures all the *unobserved* (all except X) factors that influence Y.

Hence, the first assumption we formulate about the simple linear regression is the functional form it should take. We can directly move on to the second assumption.

2.1.2. Assumption SLR. 2: Random Sampling

> We have a sample of size n, such that $\{(x_i, y_i) : i = 1, \dots, n\}$ the x_i and y_i are independent and identically distributed realizations of the random variables (X, Y).

Oftentimes, this assumption can be violated in many different ways (even if we do not specifically treat it with all the focus we should). Here are some examples:

1. **Self-selection / Volunteer bias**. What happens? Individuals choose to participate (e.g. an online survey), instead of being drawn at random. In consequence, our sample over-represents people with strong opinions or high motivation, so (x_i, y_i) are not independent draws from the target population.
2. **Convenience sampling**. We grab whoever's easily accessible—say, students in our lecture hall or customers at one store. Hence, observations are correlated by location or context, and they don't reflect the broader population.

3. **Cluster sampling with ignored within-cluster correlation**. We sample clusters (e.g. schools) and then everybody in each cluster, but we treat all observations as independent. Pupils in the same school share unobserved factors (teaching quality, neighborhood), so errors are correlated within clusters (we discuss this in another assumption).
4. **Panel / time-series data treated as cross-sectional** We pool repeated observations on the same units over time, but still assume each (x_i, y_i) is a fresh draw. Serial correlation (e.g. GDP this year predicts GDP next year) violates independence.

Importantly, we should note that if this assumption holds, this means that the residuals u_i should also be i.i.d. realizations.

2.1.3. Assumption SLR. 3: Sample Variation in the Explanatory Variable

> The sample outcomes on X, namely, $\{x_i : i = 1, ..., n\}$, are not all equal to the same value with probability one.

This assumption is extremelly easy to check (usually) and is most of the time satisfied. For instance, this X would not work:

```r
income <- c(1000, 1000, 1000, 1000, 1000)
cat("The variance of the random variable income is: ", var(income))
```

```
## The variance of the random variable income is:  0
```

But the probability that this kind of situation happens in a large sample of size n tends toward 0. Furthermore, even if this assumption is not satisfied, we can simply discard the variable that violates it since it will not help us to make any useful predictions (e.g. if I say that all the individuals in the survey are human beings and create a dummy variable `humans`, this will not help me to explain the variance in the dependent variable).

The random variable `size` is valid:

```r
size <- c(1.84, 1.62, 1.91, 1.57, 1.47)
cat("The variance of size is: ", var(size))
```

```
## The variance of size is:  0.03457
```

2.1.4. Assumption SLR. 4: Exogeneity (Zero Conditional Mean)

> The error U (also considered as a random variable) has an expected value of zero conditional on any value of the explanatory variabe.
>
> $$E[u \mid X] = E[u] = 0$$

What does this mean in words? This implies that the error term u does not, in itself, capture a systematic pattern, conditional on X. For instance, if we run the model:

$$wages = \beta_0 + \beta_1 \cdot experience + u$$

We might encounter a problem because it is highly likely that we do not capture all the relevant variables in our model. For instance, we do not account for intellectual dexterity (e.g., IQ), which might be correlated to *experience*.

Why is this assumption foundational for causal inference?

In causal analysis, we aim to identify the effect of a single independent variable on the target variable. However, in such a case where $E[u \mid X] \neq 0$, we will have **biased** estimates for β_0 and β_1 because they will try to "make up for" the unobserved variables (this will become clearer later) and incorporate some of their influence in them. This means that we will likely overestimate or underestimate the real effect of *experience* on *wages* because β_1 will also reflects the effects of other omitted / unobserved variables, such as IQ.

Important Consequences

Once we have established this assumption, important consequences logically follow (law of iterated expectations):

$$E[u] = E[E[u \mid X]] \implies E[u] = 0E[Y \mid X] = E[\beta_0 + \beta_1 X + u \mid X] = \beta_0 + \beta_1 X$$

Then:

$$E[Y] = E[E[Y \mid X]] = \beta_0 + \beta_1 E[X]Cov(u, X) = E[Xu] - E[X]E[u] = E[Xu] = 0$$

Logically, if one of these is violated, this likely means that the assumption of exogeneity does not hold.

Avoiding a Classic Trap

When we use the OLS method, it is an algebraic consequence that we will obtain:

1. $\sum \hat{\varepsilon}_i = 0$ (so $\bar{\hat{\varepsilon}} = 0$).
2. $\sum x_i, \hat{\varepsilon}_i = 0$ (so $\text{Cov}(x, \hat{\varepsilon}) = 0$).

Thus, this would not be a good idea to measure empirically the residuals or to plot them against X because we will think that SLR. 4 is satisfied, whereas it is not!

Monte-Carlo Simulation In this section, we run a quick Monte Carlo Simulation to show that when a variable is omitted from the model (and correlated to the independent variable, by consequence $E[u \mid X] \neq 0$), we will **consistently** obtain biased estimates:

```r
# parameters
set.seed(2025)
n      <- 1000      # sample size
beta0  <- 5         # true intercept
beta1  <- 2         # true slope on experience
gamma  <- 3         # effect of IQ (the one we will omit)
rho     <- 0.3      # strength of correlation between experience and IQ

# we simulate the data once:
experience <- rnorm(n, mean = 10, sd = 2)
IQ <- rho * experience + sqrt(1 - rho^2) * rnorm(n)  # corr(IQ,exp)=rho
noise <- rnorm(n, 0, 1)

# our true data-generating process:
wages <- beta0 + beta1 * experience + gamma * IQ + noise

# we use Monte-Carlo replication to see bias:
```

```r
B <- 500
est_b1 <- replicate(B, {
  x  <- rnorm(n,10,2)
  iq <- rho * x + sqrt(1-rho^2) * rnorm(n)
  y <- beta0 + beta1 * x + gamma * iq + rnorm(n)
  coef(lm(y ~ x))[2] # we retrieve the slope coef
})

cat("True slope =", beta1, "\n")
```

```
## True slope = 2
```

```r
cat("Average estimated slope over", B, "simulations:", round(mean(est_b1), 4), "\n")
```

```
## Average estimated slope over 500 simulations: 2.9015
```

In the true model, IQ indeed influences wages, but since we fail to capture it, we obtain biased estimates, i.e., estimates that do not isolate the true effect that their associated variable has on the response variable!

2.1.5. Assumption SLR. 5: Homoskedasticity

We assume that the variance of the error term u conditional on given X is constant.

$$Var[u \mid X] = Var[u] = \sigma^2$$

Why does this assumption even matter?

It is a bit more technical and less intuitive to understand why this assumption matters since it will be explained only in subsequent lectures. However, as for now, we can still provide the following explanation:

- If errors are heteroskedastic ($Var[u \mid X]$ not fixed), OLS remains unbiased for β, but it is no longer the most efficient linear estimator—we could get smaller variance (and hence more precise estimates) by giving less weight to high-variance observations.
- In hypothesis testing, we will need to compute **OLS standard errors** (how much the OLS coefficient estimates $\hat{\beta}$ would vary if we re-drew your sample many times, remember the previous Monte-Carlo simulation) and in the classic variation of the formula, we assume that the variance of the error term is homoskedastic. If this assumption fails, the statistical significance of the coefficients needs to be determined with more sophisticated methods!

2.2. Methods of Estimation

In this section, we formally derive how to obtain the OLS estimates and we also cover another estimation method, the method of moments (MoM).

2.2.1. Ordinary Least Squares (OLS)

In this section, we formally derive the OLS formula for the simple linear regression. Our purpose is to reduce the squared error between the prediction we would make with our model and the true value. Mathematically speaking, this means that we want to reduce for each x_i, y_i (error term neglected since we focus on minimizing the expected value):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

And we want to find the optimal $hat\beta_0$ and $\hat{\beta}_1$. We take the two first order conditions and we have (L stands for loss function):

$$\frac{\partial L}{\partial \hat{\beta}_1} = 0 \implies -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Then, we divide on each side by n and use the same notations as in the course to obtain:

$$\overline{xy} = \hat{\beta}_0 \overline{x} + \hat{\beta}_1 \overline{x^2} \quad [1]$$

Then, we take the second first-order condition:

$$\frac{\partial L}{\partial \hat{\beta}_0} = 0 \implies -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0 \implies \overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \quad [2]$$

We can now isolate $\hat{\beta}_1$ and we find:

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Alternatively, we could easily demonstrate that $\hat{\beta}_1$ is also given by:

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

By convention, we call:

- $\hat{y}_i$ the predicted values
- $\hat{u}_i = y_i - \hat{y}_i$ the residuals, we should not confound them with the error term!

Thus, we can also write that:

$$y_i = \hat{y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

2.2.2. Method of Moments

As usual, the goal with the method of moment is to equate the **theoretical** and **empirical** moments to find the desired parameters. In our case, we made several assumption on the conditional expected value of the error term in our SLR. 4. Thus, we can first use:

$$E[u] = 0 \implies \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0 \quad [A]$$

Then, we also know that $E[X \mid u] = 0$. This also means that, $Cov(X,u) = 0$, hence we can also say that:

$$E[uX] = 0 \implies \frac{1}{n}\sum_{i=1}^{n}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right] x_i = 0 \quad [B]$$

We could solve with $[A]$ and $[B]$ and find the exact same results as before! In the following lecture, we will discuss extensively these results and some of their implications.

2.2.3. Remarks on Notebook 3

If one does not feel comfortable with functions in R, it might be interesting to code a function, without any built-in functionals, named `my_univariate_ols()`, which outputs a vector `coefficients` with the intercept and the slope.

Example Usage

```r
# parameters
beta_0 = -3
beta_1 = 2

# synthetic data
x <- rnorm(n = 1000, mean = 3, sd = 2)
error <- rnorm(n = 1000, mean = 0, sd = 1)
y <- beta_0 + beta_1 * x + error

# example usage
# betas <- my_ols(x = x, y = y)
# print(betas)

# Output
# -3.1445, 2.1678
```

3. Univariate Linear Regression Model (2/3)

Content

- Fitted Values and Residuals
- Algebraic Properties

Materials

- Wooldridge Chapter 2.3.
- Slides L03

3.1. Residuals of the OLS

By definition, we saw in the previous lecture that:

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

And we agreed on the fact that the residual $\hat{u}_i$ captures the deviation of our prediction $\hat{y}_i$ from the true value y_i. Hence, we compute the residuals as:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

where:

- $\hat{y}_i$ underestimated y_i if $\hat{u}_i > 0$
- $\hat{y}_i$ overestimated y_i if $\hat{u}_i < 0$

Now, from this point, we can derive other algebraic properties for the residuals. (We do this by relying on the previously derived first-order conditions.)

3.1.1. Algebraic Properties

First, we know that, on average, residuals have to cancel out because $\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$. Hence, we can write:

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

Then, we can rely of the second first-order condition to derive a useful result:

$$\overline{xy} = \hat{\beta}_0 \overline{x} + \hat{\beta}_1 \overline{x^2}$$

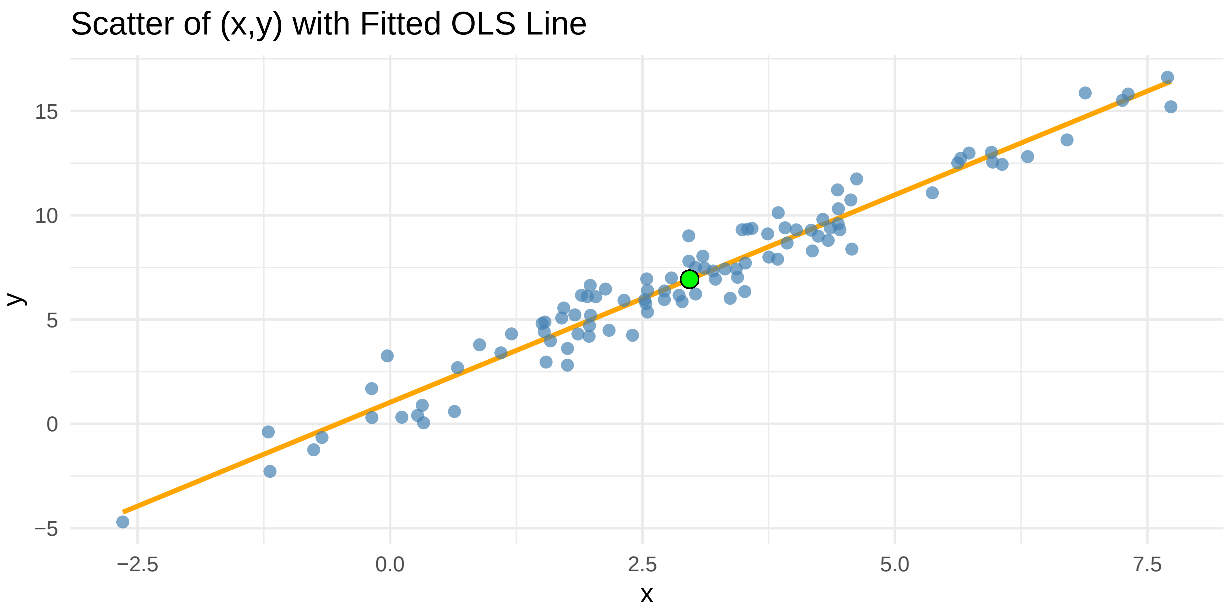
Thus, when we develop:

$$\sum_{i=1}^{n} \hat{u}_i x_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$= n \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\hat{\beta}_0 \sum_{i=1}^{n} x_i - \frac{1}{n}\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \right) = n \cdot 0 = 0$$

Finally, the last property states that $(\overline{x}, \overline{y})$ is on the regression line. Indeed, we can do a quick simulation to prove this:

```r
# settings
beta_0 <- 1
beta_1 <- 2
x <- rnorm(n = 100, mean = 3, sd = 2)
u <- rnorm(n = 100, mean = 0, sd = 1)

# model
y <- beta_0 + beta_1 * x + u

# data frame
df <- data.frame(
  x = x,
  y = y,
  u = u
)

# regression
model_31 <- lm(y ~ x, data = df)

# coefficients
betas <- model_31$coefficient
beta_0 <- betas[1]
beta_1 <- betas[2]

# regression line
df$xs <- seq(min(df$x), max(df$x), length.out = 100)
df$y_predicted <- beta_0 + beta_1 * df$xs

# point
mean_x = mean(df$x)
mean_y = mean(df$y)

# plot
ggplot(data = df, aes(x = xs, y = y_predicted)) +

  # fitted line
  geom_line(color = "orange", size = 1.2) +

  # raw points
  geom_point(aes(x = x, y = y),
             color = "steelblue",
             size = 2.5,
             alpha = 0.7) +

  # mean point
  geom_point(aes(x = mean_x, y = mean_y),
             shape = 21,
             fill = "green",
             color = "black",
             size = 4) +
  # titles and labels
  labs(
    title     = "Scatter of (x,y) with Fitted OLS Line",
    x         = "x",
    y         = "y"
  ) +

  # a cleaner minimal theme
  theme_minimal(base_size = 14)
```



3.2. Variance in the OLS

By convention, we denote three ways of looking at the variance in the OLS model. This also indicates the extent to which we can explain the variance in the dependent variable thanks to the explanatory variables (in this case X).

3.2.1. Total Sum of Squares

The total sum of squares is a proxy of the variance of y. It measures the sum of the squared deviation of each y_i from their mean $\overline{y}$. We have:

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

Example

```
y1 <- rnorm(100, mean = 0, sd = 1) # low variance
y2 <- rnorm(100, mean = 0, sd = 4) # higher variance

# total sum of squares
```

```r
sst_1 <- sum((y1 - mean(y1))^2)
sst_2 <- sum((y2 - mean(y2))^2)

# print
cat("SST y1: ", sst_1, "\nSST y2: ", sst_2)
```

```
## SST y1:  96.23893
## SST y2:  1806.407
```

Furthermore, we can show that this is nothing more than the sample variance multiplied by $(n-1)$:

```r
n = 100
sst = (n - 1) * var(y1)
print(sst)
```

```
## [1] 96.23893
```

3.2.2. Explained Sum of Squares

The explained sum of squares (SSE) provides an information on how well our model explain the variance in the dependent variable:

$$\text{SSE} = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2$$

For instance, if we have a very simplistic model and we make predictions as: $\hat{y}_i = \overline{y}$, we will have $SSE = 0$. This means that we cannot explain the variance in y with our model! Hopefully, the OLS regression can do better!

3.2.3. Residual Sum of Squares

By now, we should be familiar with this measure since we attempted to minimize it for deriving our OLS parameters. We have:

$$SSR = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

The lower the SSR, the better the quality of our fit. If we have $SSR = 0$, this means that we can explain 100% of the variance in y with our model. In other words the sum of squared residuals capture the **unexplained variance**. (In machine learning, this kind of situation is a warning for overfitting or data leakage.) Hence, this measure helps us to estimate the goodness of our fit. (Its mean is called the Mean Squared Error, and it is widely used in ML).

Finally, quite logically, the total sum of squares (SST) is equal to the sum of the explained variance (SSE) and the unexplained variance (SSR):

$$SST = SSR + SSE$$

The proof can be realized if we recombine the development of $SSR + SSE$.

Now that we have uncovered all of those measures, we can look at a widely used statistical measure used to evaluate the goodness of our fit.

3.2.4. The Coefficient of Determination R^2

The R score is the ratio of the explained variation compared to the total variation; thus, it is interpreted as the fraction of the sample variation in y that is explained by x. If we have $R^2 = 1$, this means that we reached a perfect fit (i.e. we explain all the variation in the data). In the social sciences, low R-score in regression equations are not uncommon, especially for cross-sectional analysis.

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

We have to be aware of the fact that in our case, when we use the simple linear regression model, the R^2 measure the quality of the **linear relationship**.

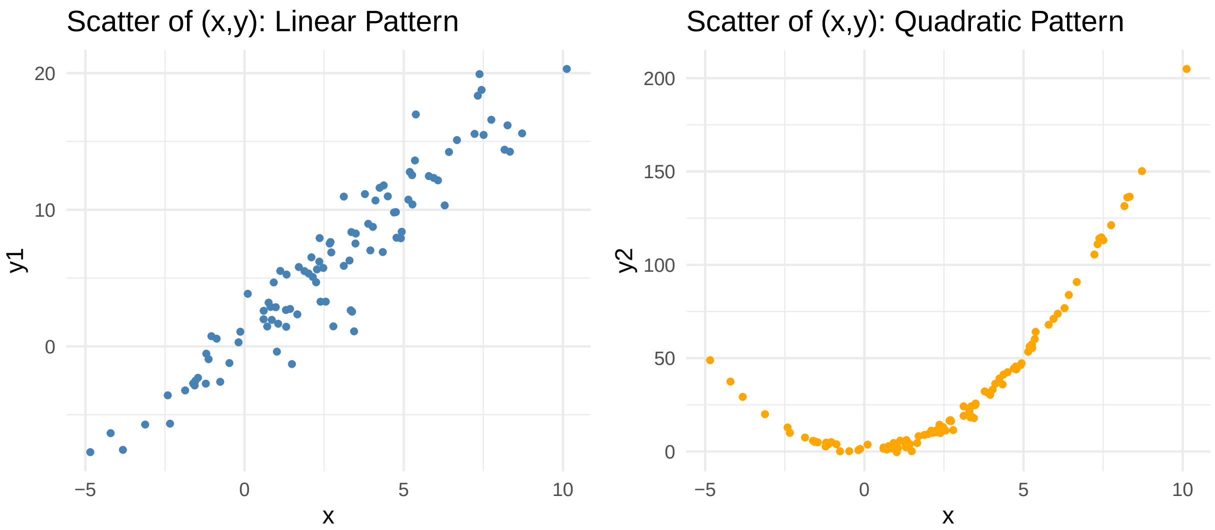
Example We build two model to illustrate two kinds of relation:

$$y_1 = \beta_0 + \beta_1 x_1 + u_1 y_2 = \beta_0 + \beta_1 x_1^2 + u_2$$

And we show that, since R^2 measures the strength of the linear relationship, it should be low when we use the simple linear regression.

```
library(patchwork)

# settings
beta_0 <- 1
beta_1 <- 2
x <- rnorm(n = 100, mean = 3, sd = 3)
u <- rnorm(n = 100, mean = 0, sd = 2)

# model
y1 <- beta_0 + beta_1 * x + u
y2 <- beta_0 + beta_1 * x^2 + u

# data frame
df <- data.frame(
  x = x,
  y1 = y1,
  y2 = y2,
  u = u
)

# plots
p1 <- ggplot(data = df, aes(x = x, y = y1)) +
  geom_point(color = "steelblue") +
  # titles and labels
  labs(
    title    = "Scatter of (x,y): Linear Pattern",
    x        = "x",
    y        = "y1"
  ) +

  # a cleaner minimal theme
  theme_minimal(base_size = 14)
```

```
p2 <- ggplot(data = df, aes(x = x, y = y2)) +
  geom_point(color = "orange") +

  # titles and labels
  labs(
    title    = "Scatter of (x,y): Quadratic Pattern",
    x        = "x",
    y        = "y2"
  ) +

  # a cleaner minimal theme
  theme_minimal(base_size = 14)

p1 + p2
```



Now, we can have a look at two different R-scores (we focus only on y_2):

```
# This is the correct quadratic model on y2: (we did not cover that yet)
lm_quad <- lm(y2 ~ x + I(x^2), data = df)

# We extract R_squared for both
r2_lin  <- summary(lm(y2 ~ x, data = df))$r.squared
r2_quad <- summary(lm_quad)$r.squared

cat("Linear fit r2 on y2:     ", round(r2_lin, 3), "\n",
    "Quadratic fit r2 on y2:  ", round(r2_quad, 3), "\n")
```

```
## Linear fit r2 on y2:      0.633
##  Quadratic fit r2 on y2:   0.998
```

In the case, where we do not capture the quadratic relationship, we fail to have a good r2 score!

Additional Remark There is no relation between predictive power and unbiasedness!

4. Univariate Linear Regression Model (3/3)

Content

- The expected values of the OLS estimator (i.e., under which conditions the OLS estimator is unbiased)
- The variance of the OLS estimator

Materials

- Wooldridge Chapter 2.5.
- Slides L04

4.1. Unbiasedness of the OLS Estimator

In this first section, we discuss in greater details what unbiasedness means and why the OLS might or might not be a biased estimator.

4.1.1. Definitions

An estimator is unbiased if the expected value of this estimator is equal to the theoretical value of the estimated parameter. That is:

$$E[\hat{\theta}] = \theta$$

On the other hand, if the estimator is biased, we compute the bias as:

$$\text{bias} = E[\hat{\theta}] - \theta$$

To visualize better what is at play here, let us do an example.

4.1.2. Visualizing Bias with a Monte-Carlo Simulation

In this brief example, we will create two models. One ill-specified model, which violates the exogeneity assumption ($E[u \mid X \neq 0]$), and one well-specified multivariate model, which will not suffer from the omitted variable bias (OVB).

Our real model will be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where (X_1, X_2) are distributed according to a bivariate normal distribution and display some correlation:

$$\rho_{(X_1, X_2)} = 0.5$$

Then, we will estimates the model's parameters by running an ill specified regression:

$$y^{(1)} = \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x_1 + \hat{u}^{(1)}$$

and a well specified model:

$$y^{(2)} = \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x_1 + +\hat{\beta}_2^{(2)} x_2 + \hat{u}^{(2)}$$

and the Monte-Carlo will estimate $E[\hat{\beta}_1^{(1)}]$ and $E[\hat{\beta}_1^{(2)}]$. If we believe our theoretical insights, we should observe that $E[\hat{\beta}_1^{(1)}] \neq \beta_1$ and $E[\hat{\beta}_1^{(2)}] \approx \beta_1$ (because in the first case the exogeneity assumption is violated, if $\beta_2 \neq 0$ of course).

1. We generate the synthetic data:

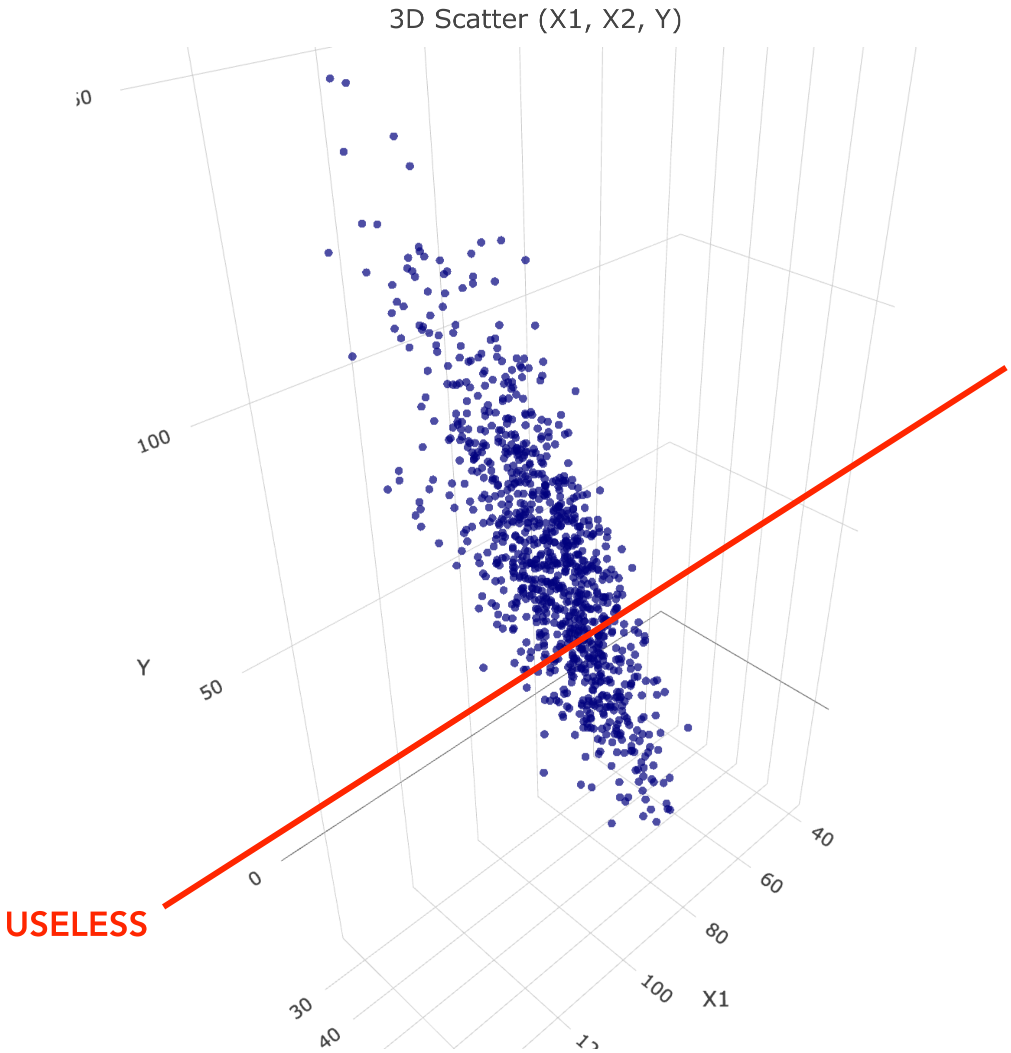
```
library(MASS)
# true parameters
beta_0 <- 10
beta_1 <- 2
beta_2 <- -3

# X1, X2
############
data_size <- 1000

# means
mu_vec <- c(100, 50)

# standard deviations
sd1 <- 16    # sd(X1) = sqrt(40)
sd2 <- 8              # sd(X2)

# target correlation
rho  <- 0.5

# covariance matrix
Sigma <- matrix(c(sd1^2,      rho * sd1 * sd2,
                  rho * sd1 * sd2, sd2^2),
                nrow = 2, byrow = TRUE)

# bivariate sample
sim_data <- mvrnorm(n = data_size,
                    mu = mu_vec,
                    Sigma = Sigma)

# X1, X2
X1 <- sim_data[, 1]
X2 <- sim_data[, 2]

#####
#u
u <- runif(n = data_size,
           min = -10,
           max = 10) # E[u] = 0, this is respected

# the real values of y
Y <- beta_0 + beta_1 * X1 + beta_2 * X2 + u

# we store everything in a data frame
```

```r
complete_df <- data.frame(
  X1 = X1,
  X2 = X2,
  u = u,
  Y = Y
)

# head
head(complete_df, n = 3)
```

```
##          X1       X2          u         Y
## 1 113.00108 44.12482  2.0255974 105.65331
## 2  94.76472 30.31088  0.9989391 109.59572
## 3 111.64746 54.56318 -6.4271944  63.17818
```

2. We can also plot our data generating process (DGP):

```r
library(plotly)

plot_ly(
  data = complete_df,
  x    = ~X1,
  y    = ~X2,
  z    = ~Y,
  type = "scatter3d",
  mode = "markers",
  marker = list(size = 3, color = "navy", opacity = 0.7)
) %>%
  layout(
    title = "3D Scatter (X1, X2, Y)",
    scene = list(
      xaxis = list(title = "X1"),
      yaxis = list(title = "X2"),
      zaxis = list(title = "Y")
    )
  )
```



3. And now, we run the Monte-Carlo simulation for estimating the true parameter β_1 in the ill-specified model:

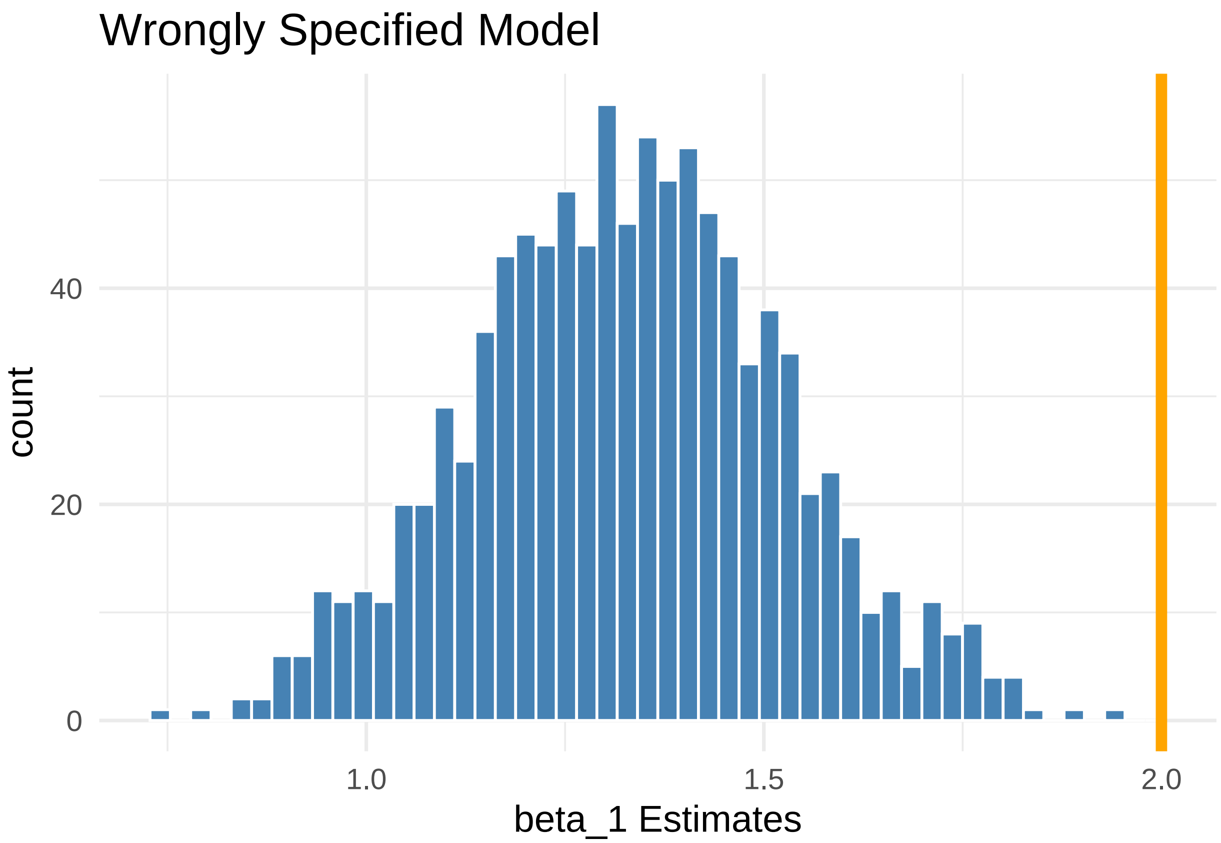
```r
# simulation parameters
sample_size = 50
M = 1000

# storage
betas_1 = numeric(M)

for (i in 1:M){

  # we draw a sample
  samp_idx   <- sample(nrow(complete_df), size = sample_size, replace = FALSE)
  sample_df  <- complete_df[samp_idx, ]

  # we run the regression on the sample
  model <- lm(Y ~ X1, data = sample_df)
  betas_1[i] <- coef(model)["X1"]

}

# summarize the Monte Carlo results
mean_beta_1 <- mean(betas_1)
sd_beta_1   <- sd(betas_1)

cat("Expected value of beta_1 (MC):", round(mean_beta_1, 4), "| True value: ", beta_1, "\n")
```

```
## Expected value of beta_1 (MC): 1.3299 | True value:  2
```

```r
cat("Std. dev. of beta_1   :", round(sd_beta_1,   4), "\n")
```

```
## Std. dev. of beta_1   : 0.1941
```

```r
# plot
ggplot(data = data.frame(betas_1 = betas_1), aes(x = betas_1)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  geom_vline(xintercept = beta_1, color = "orange", linewidth = 2) +
  labs(title = "Wrongly Specified Model",
       x = "beta_1 Estimates") +
  theme_minimal(base_size = 14)
```



4. and in the well-specified model:

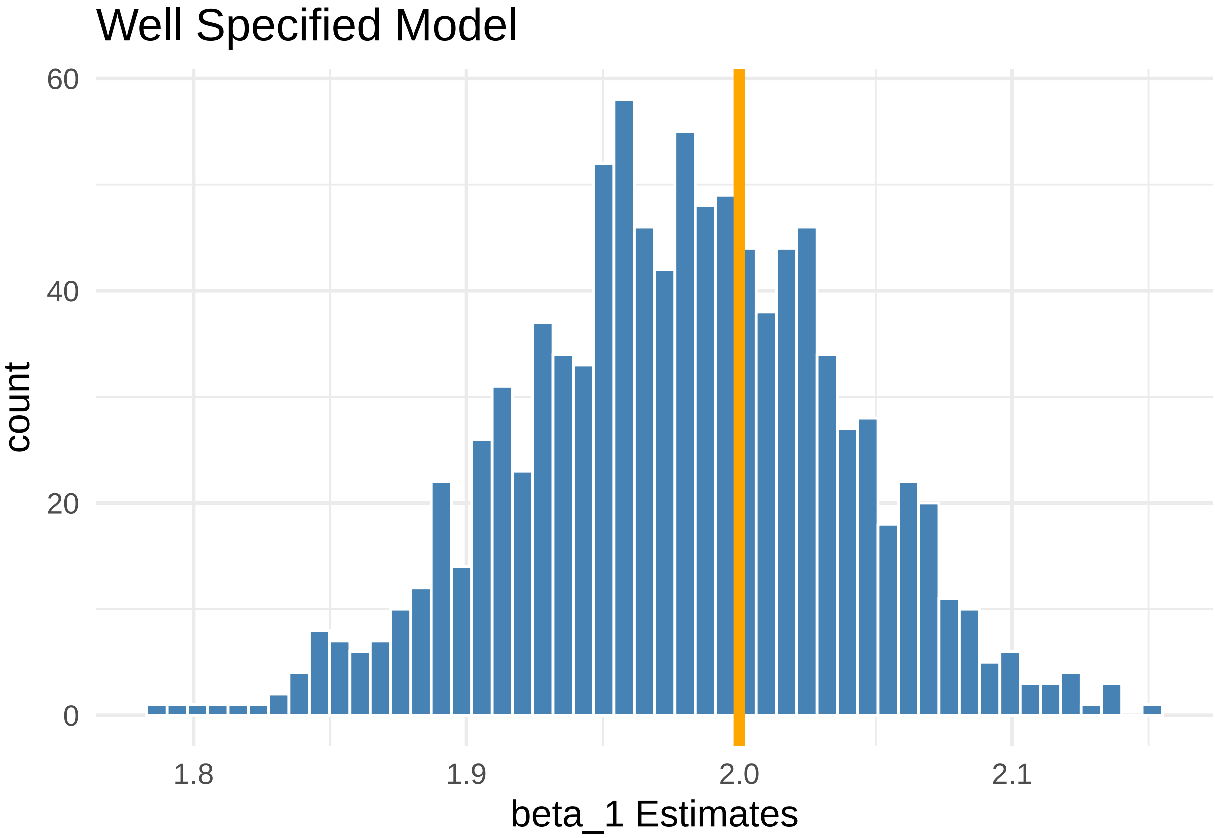
```
# simulation parameters
sample_size = 50
M = 1000

# storage
betas_12 = numeric(M)

for (i in 1:M){

  # we draw a sample
  samp_idx   <- sample(nrow(complete_df), size = sample_size, replace = FALSE)
  sample_df  <- complete_df[samp_idx, ]

  # we run the regression on the sample
  model2 <- lm(Y ~ X1 + X2, data = sample_df)
  betas_12[i] <- coef(model2)["X1"]

}

# summarize the Monte Carlo results
mean_beta_12 <- mean(betas_12)
sd_beta_12   <- sd(betas_12)

cat("Expected value of beta_1 (MC):", round(mean_beta_12, 4), "| True value: ", beta_1, "\n")
```

```
## Expected value of beta_1 (MC): 1.9781 | True value:  2
```

```r
cat("Std. dev. of beta_1   :", round(sd_beta_12,   4), "\n")
```

```
## Std. dev. of beta_1   : 0.0592
```

```r
# plot
ggplot(data = data.frame(betas_1 = betas_12), aes(x = betas_1)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  geom_vline(xintercept = beta_1, color = "orange", linewidth = 2) +
  labs(title = "Well Specified Model",
       x = "beta_1 Estimates") +
  theme_minimal(base_size = 14)
```



Conclusion

We see that when the exogeneity assumption (SLR. 4) is violated, we obtain biased estimates! On the other hand, our Monte-Carlo simulation also revealed that when the model is well specified, it seems to be unbiased. We will formally derive this in the following section.

4.1.3. Formal Proof of Unbiasedness

Theorem If we assume that SLR.1 to SLR. 4 hold, we can demonstrate that:

$$E[\hat{\beta}_1] = \beta_1 E[\hat{\beta}_0] = \beta_0$$

Proof Here is my proof for unbiasedness under SLR. 1 to SLR. 4

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \quad \text{we want to show:} \quad \mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \text{[Unbiasedness Theorem]}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\sum x_i(\beta_0 + \beta_1 x_i + u_i) - \frac{1}{n}\sum x_i \sum(\beta_0 + \beta_1 x_i + u_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$\hat{\beta}_1 = \frac{\sum x_i(\beta_0 + \beta_1 x_i + u_i) - \frac{1}{n}\sum x_i \left[n\beta_0 + \beta_1 \sum x_i + \sum u_i\right]}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \quad \text{(use that } \sum \hat{u}_i = 0 \text{ under SLR.4)}$$

$$(1)\ \mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\beta_0 \sum x_i + \beta_1 \sum x_i^2 + \sum x_i u_i - \frac{1}{n}\sum x_i \left(n\beta_0 + \beta_1 \sum x_i + \sum u_i\right)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}\right]$$

$$(2)\ \mathbb{E}[\hat{\beta}_1] = \frac{\beta_0 \sum x_i + \beta_1 \sum x_i^2 - \beta_0 \sum x_i - \frac{1}{n}\beta_1 (\sum x_i)^2}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \beta_1 \left[\frac{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}\right] = \beta_1$$

We have demonstrated unbiased for $\hat{\beta}_1$. The proof is quite simple since we just need to develop with respect to the theoretical coefficients and use SLR. 4. Secondly, we use this result to prove the unbiasdness of $\hat{\beta}_0$:

Evidently, we know that (FOC):

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

Hence,

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}\left[\frac{1}{n}\sum(\beta_0 + \beta_1 x_i + u_i)\right] - \mathbb{E}[\hat{\beta}_1] \cdot \frac{1}{n}\sum x_i$$

$$= \mathbb{E}\left[\frac{1}{n} \cdot n \cdot \beta_0 + \frac{1}{n}\beta_1 \sum x_i + \bar{u}\right] - \beta_1 \cdot \frac{1}{n}\sum x_i$$

$$\text{(since } \sum_{i=1}^{n} [u_i] = 0\text{), we have:} \quad \mathbb{E}[\hat{\beta}_0] = \beta_0$$

4.2. Efficiency of the OLS Estimator

The OLS estimator is efficient if it is the one with the smallest variance among all other estimators. First, we establish the theorem and look at some comparative statistics (this time we also assume that SLR. 5 holds):

$$Var(\hat{\beta}_1 \mid x_1, \ldots, x_k) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2}{SST_x} Var(\hat{\beta}_0 \mid x_1, \ldots, x_k) = \frac{\frac{\sigma^2}{n}\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

We notice two important facts: first, the variance of both estimators increases as the variance σ^2 of the error term u increases. Secondly, another less expected result is that as the total sum of squares (SST) increases, both estimators become more precise. This means that variation in the independent variable is good. We will now formally derive the variance.

4.2.1. Derivation of the Variance

We have the following development, closely inspired from the lecture slides:

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

We continue with the same development:

$$\frac{\sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i + \bar{x}\bar{y}n - n\bar{y}\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + u_i)(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n x_i(x_i - \bar{x}) + \sum_{i=1}^n u_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0, \quad \sum_{i=1}^n x_i(x_i - \bar{x}) = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

Thus:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Once we have established this, we see that the constant term β_1 should not be included in the variance. Furthermore it is important to note that since we are working with the conditional variance, we treat $(x_1, \ldots, x_n)$ as non-random. This means that when we compute the variance we consider them like a constant, where, for a quick refresh, we know that:

$$Var(bX) = E[(bX)^2] - (E[bX])^2 = b^2E[X^2] - b^2(E[X])^2 = b^2Var(X)$$

Thus, in our case, we have:

$$\text{Var}(\hat{\beta}_1 \mid x_1, x_2, \ldots, x_n) = \text{Var}\left(\frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \,\middle|\, x_1, \ldots, x_n \right)$$

and we know that $Var(u_i \mid x_i) = \sigma^2$, thus, we have:

$$\text{Var}(\hat{\beta}_1 \mid x_1, \ldots, x_n) = \frac{\sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Again , variance in the independent variable is good! The only issue we face here is that in practice we do not have access to the theoretical variance σ of the error term u. Hence, we must find an estimator for it!

4.2.2. Derivation of Standard Errors

The standard error of the parameter β_1 is derived as:

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

and now, we should find an **unbiased estimator** for $\text{Var}(u) = \sigma^2$.

Unbiased Estimator of the Variance σ^2 The unbiased estimaor is given by:

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

and we divide by $(n-2)$ since we have two degrees of freedom in the SLR. Then, we simply use this estimator in the formula of the standard error:

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

How can we prove this? We recall first that the total variance is the sum of the explained and unexplained variance, we have:

$$SST = SSE + SSR \Leftrightarrow \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 + \sum_{i=1}^{n}\hat{u}_i^2$$

and we search $E[\sum_{i=1}^{n}\hat{u}_i^2]$. We will solve it as:

$$E[SSR] = E[SST] - E[SSE]$$

where:

$$E[SST] = (n-1)\sigma_y^2 \quad \text{(unbiased sample variance)}$$

and we can also write that:

$$\sigma_y^2 = \text{Var}(y_i) = \text{Var}(\beta_0 + \beta_1 x_i + u_i) = \beta_1^2\sigma_x^2 + \sigma^2 + 2\beta_1\text{Cov}(x_i,u_i) = \beta_1^2\sigma_x^2 + \sigma^2.$$

Then, we focus on developing the SSE:

$$SSE = \sum_{i=1}^{N}(\hat{y}_i-\overline{y})^2$$

$$= \sum_{i=1}^{N}\left(\hat{y}_i - \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_0+\hat{\beta}_1 x_i+\hat{u}_i)\right)^2 \quad \text{where } \sum_{i=1}^{N}\hat{u}_i = 0$$

$$= \sum_{i=1}^{N}\left(\hat{y}_i - \hat{\beta}_0 - \hat{\beta}_1\bar{x}\right)^2$$

$$= \sum_{i=1}^{N}\left(\hat{\beta}_1(x_i-\bar{x})\right)^2$$

$$= \hat{\beta}_1^2\sum_{i=1}^{N}(x_i-\bar{x})^2$$

and from previous calculations, we know that this is equivalent to:

$$SSE = (\beta_1 + \frac{\sum_{i=1}^{n} u_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})^2 \sum_{i=1}^{N}(x_i - \bar{x})^2$$

and we compute:

$$E[SSE \mid x_1, x_2, \dots x_n]$$

First, we define:

$$A = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{(a constant given the } x_i\text{'s)}$$

We let:

$$Z := \frac{\sum_{i=1}^{n} u_i(x_i - \bar{x})}{A} \quad \Rightarrow \quad SSE = (\beta_1 + Z)^2 \cdot A$$

Now we compute the conditional expectation:

$$\mathbb{E}[SSE \mid x_1, \dots, x_n] = \mathbb{E}[(\beta_1 + Z)^2 \cdot A \mid x_1, \dots, x_n] = A \cdot \mathbb{E}[(\beta_1 + Z)^2 \mid x_1, \dots, x_n]$$

We expand the square:

$$\mathbb{E}[(\beta_1 + Z)^2 \mid x] = \beta_1^2 + 2\beta_1 \mathbb{E}[Z \mid x] + \mathbb{E}[Z^2 \mid x]$$

Recall:

$$Z = \frac{\sum_{i=1}^{n} u_i(x_i - \bar{x})}{A} \quad \text{with} \quad \mathbb{E}[u_i \mid x_i] = 0 \Rightarrow \mathbb{E}[Z \mid x] = 0$$

Also, since u_i are uncorrelated with equal variance σ^2, we get:

$$\text{Var}(Z \mid x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sigma^2}{A^2} = \frac{\sigma^2 A}{A^2} = \frac{\sigma^2}{A} \Rightarrow \mathbb{E}[Z^2 \mid x] = \frac{\sigma^2}{A}$$

Finally:

$$\mathbb{E}[SSE \mid x_1, \dots, x_n] = A \cdot \left(\beta_1^2 + 0 + \frac{\sigma^2}{A}\right) = \beta_1^2 A + \sigma^2 = \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sigma^2 = \beta_1^2(n-1)\sigma_x^2 + \sigma^2$$

Finally:

$$E[SSR] = E[SST] - E[SSE] = (n-1)(\beta_1^2 \sigma_x^2 + \sigma^2) - (\beta_1^2(n-1)\sigma_x^2 + \sigma^2) = (n-2)\sigma^2$$

This finally justifies the division by $(n-2)$.

---

5. Multiple Linear Regression Model (1/2)

Content

- Multiple Linear Regression Model (MLR)
- Assumptions, Estimator and Algebraic Properties of MLR

Materials

- Wooldridge Chapters 3.1. to 3.3.
- Slides L05

5.1. Introduction

In this new lecture, we extend the linear regression to a case where we have k explanatory variables. Therefore, the underlying model we now face is defined by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k + u$$

We usually write this model in matrix form as:

$$Y = X\beta + u$$

where:

$$X = \begin{bmatrix} 1 & X_{1(1)} & X_{2(1)} & \dots & X_{K(1)} \\ 1 & X_{1(2)} & X_{2(2)} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1(n)} & \dots & \dots & X_{K(n)} \end{bmatrix}$$

Thus, the matrix of the K independent variables is of dimension $(n \times k)$. We defined $X_0 = 1$.

And we also denote:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad Y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

In such a model, the partial effects are known as:

$$\frac{\partial y}{\partial x_k} = \beta_k$$

5.2. Assumptions and Definitions

As before, we formulate 5 key different assumptions for having unbiased estimators (β).

5.2.1. MLR. 1: Functional Form

The random variables $X_1, \ldots, X_k$ and Y are related via the equation:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + U = X\beta + u$$

The model should respect linearity in parameters.

5.2.2. MLR. 2: Random Sampling

We have a sample of size $N, \{(x_{i1}, \ldots, x_{ik}, y_i) : i = 1, \ldots, N\}$, such that the $(x_{i1}, \ldots, x_{ik}, y_i)$ are independent and identically distributed (i.i.d.) realizations of the random variables $(X_1, \ldots, X_k, Y)$.

Without this assumption, we cannot formulate causal links.

5.2.3. MLR. 3: No Perfect Multicollinearity

There exists sufficient variation in all regressors $(1, x_1, ..., x_k)$ in that they are linearly independent of each other, i.e., there is no linear combination between the variables that equals 0:

$$\lambda_0 + \lambda_1 x_1 + \ldots + \lambda_k x_k = 0$$

only if $\forall j \in K, \lambda_j = 0$

This assumption needs to hold otherwise we would not be able to invert the X matrix and when regressors are perfectly collinear, their individual effects on y become non-identifiable. We can't interpret the coefficient of one regressor while holding the others constant — because holding one constant means the others must move too.

5.2.4. MLR. 4: Exogeneity

The error u has an expected value of zero given any value of the explanatory variables $X_1, \ldots, X_k$. In other words,

$$E[u \mid X_1, \ldots, X_k] = 0$$

This assumption makes reference to the same exogeneity assumption we formulated for the SLR. It is essential in causal analysis; otherwise, we will make wrong interpretations about the direction and magnitudes of causal effects.

5.2.5. MLR. 5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}[u \mid X_1, \ldots, X_k] = Var[u] = \sigma^2$$

This assumption can also be harder to check and is important for ensuring the validity of the different test (t-test for hypothesis testing) we will be conducting. Indeed, we need the standard error of the estimators ($\text{se}(\hat{\beta}_j)$) for making these tests and the standard error computation relies on this assumption. If homoskedasticity is violated, our tests become flawed.

5.3. Method of Estimation: Ordinary Least Squares

This idea is the same here, we want to minimize a loss function, namely the squared loss function to find the vector $\hat{\beta}$ which minimize it:

$$\min_{\hat{\beta}} (y - X\hat{\beta})^2 = \text{MSE}$$

This is an optimization problem with a convex quadratic function:

$$\nabla_{\hat{\beta}} \text{MSE}(\hat{\beta}) = 0 \Rightarrow \nabla_{\hat{\beta}} \left[(y - X\hat{\beta})^\top (y - X\hat{\beta}) \right] = 0$$

$$= \nabla_{\hat{\beta}} \left[y^\top y + \hat{\beta}^\top X^\top X \hat{\beta} - 2y^\top X \hat{\beta} \right] = 0$$

Take the gradient:

$$\nabla_{\hat{\beta}} \left[y^\top y + \hat{\beta}^\top X^\top X \hat{\beta} - 2y^\top X \hat{\beta} \right] = 0$$

$$= 0 + 2X^\top X \hat{\beta} - 2X^\top y = 0$$

$$\Rightarrow X^\top X \hat{\beta} = X^\top y$$

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

This is the closed-form formula for $\hat{\beta}$. Finally, we note the following properties.

Properties The mean of the residuals in the random sample is zero:

$$\frac{1}{N} \sum_{i=1}^{N} \hat{u}_i = \overline{\hat{u}} = 0$$

The covariance between the regressors and residuals in the sample is zero:

$$\frac{1}{N} \sum_{i=1}^{N} x_i j \hat{u}_i = 0$$

The regression hyper plane runs through the mean values of the data:

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x_1} + \hat{\beta}_2 \overline{x_2} + \ldots + \hat{\beta}_k \overline{x_k}$$

We can visualize this last property:

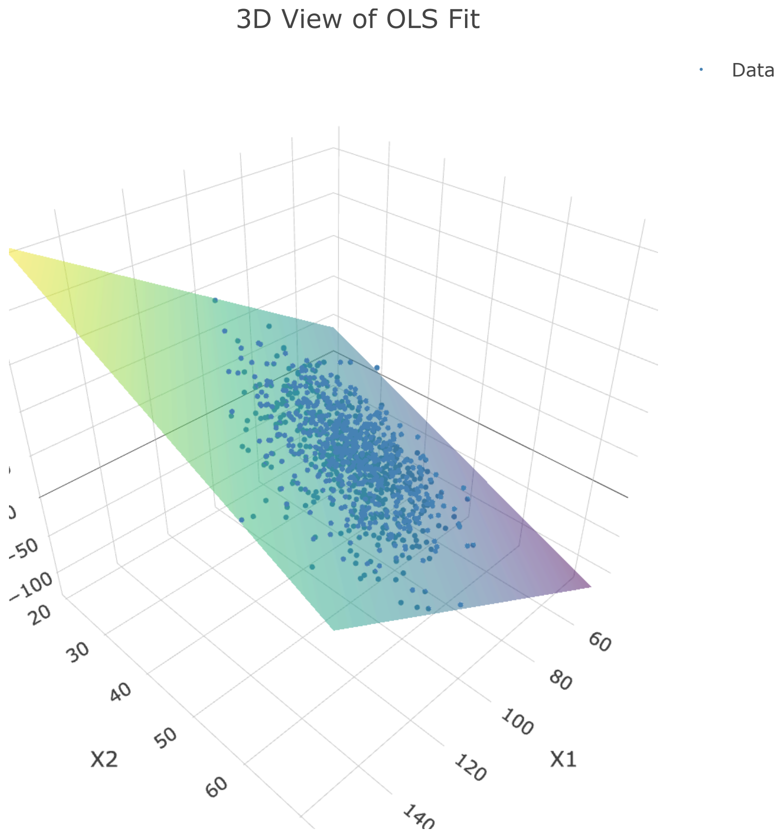
```r
library(MASS)
# true parameters
beta_0 <- 10
beta_1 <- 2
beta_2 <- -3

# X1, X2
############
data_size <- 1000

# means
mu_vec <- c(100, 50)

# standard deviations
sd1 <- 16    # sd(X1) = sqrt(40)
sd2 <- 8              # sd(X2)

# target correlation
rho  <- 0.5

# covariance matrix
Sigma <- matrix(c(sd1^2,      rho * sd1 * sd2,
                  rho * sd1 * sd2, sd2^2),
                nrow = 2, byrow = TRUE)

# bivariate sample
sim_data <- mvrnorm(n = data_size,
                    mu = mu_vec,
                    Sigma = Sigma)

# X1, X2
X1 <- sim_data[, 1]
X2 <- sim_data[, 2]

#####
#u
u <- runif(n = data_size,
           min = -10,
           max = 10) # E[u] = 0, this is respected

# the real values of y
Y <- beta_0 + beta_1 * X1 + beta_2 * X2 + u

# we store everything in a data frame
complete_df <- data.frame(
  X1 = X1,
  X2 = X2,
  u = u,
  Y = Y
)

# head
head(complete_df, n = 3)
```

```
##         X1       X2         u        Y
## 1 98.99714 46.05144  2.317690 72.15764
## 2 93.24444 59.33336 -3.724929 14.76387
## 3 94.47048 46.79627 -7.184351 51.36778
```

Then, we build the plan equation:

```
# the model
formula <- Y ~ X1 + X2

# fit with lm()
model <- lm(formula, data = complete_df)

# We build the plan
n.grid <- 30

# sequences over the observed range
x1.seq <- seq(min(complete_df$X1), max(complete_df$X1), length.out = n.grid)
x2.seq <- seq(min(complete_df$X2), max(complete_df$X2), length.out = n.grid)

# all combinations
grid <- expand.grid(X1 = x1.seq, X2 = x2.seq)

# 2.4 Predicted Y^ on that grid
grid$Yhat <- predict(model, newdata = grid)

library(plotly)

plot_ly() |>
  add_markers(
    data = complete_df,
    x    = ~X1, y = ~X2, z = ~Y,
    marker = list(size = 2, color = 'steelblue'),
    name = "Data"
  ) |>
  add_surface(
    x = x1.seq, y = x2.seq, z = t(matrix(grid$Yhat, n.grid, n.grid)),
    opacity = 0.5,
    showscale = FALSE,
    name = "Fitted plane"
  ) |>
  layout(
    scene = list(
      xaxis = list(title = "X1"),
      yaxis = list(title = "X2"),
      zaxis = list(title = "Y")
    ),
    title = "3D View of OLS Fit"
  )
```

```
## file:////private/var/folders/7v/_v_y1jpx0rl056gg5rkjsw4r0000gn/T/Rtmp6flSpo/filef5a53734a479/widgetf5
```


3D View of OLS Fit
Data
0
−50
−100
20
30
40
50
60
X2
60
80
100
120
140
X1

5.4. Insights from Notebook 5

5.4.1. Avoiding Perfect Multicollinearity: The Dummy Trap

As we previously discussed, one key assumption of the MLR model is that we do not have perfect multicollinearity (MLR. 3). One thing we must be careful about in this context is the so-called **dummy-variable** trap. Indeed, let us imagine that we have the following data set:

```r
df_pay <- data.frame(
  gender = rep(c("M", "F"), 3),
  hourlyPay = rnorm(6, mean = 100, sd = 10),
  expYear = runif(6, min = 2, max = 15)
)

head(df_pay, n = 6)
```

```
##   gender hourlyPay   expYear
## 1      M  95.92282  5.119811
## 2      F 108.77277  7.825001
## 3      M  95.36259  4.945302
## 4      F 106.54533 11.384035
## 5      M 100.85228  6.203236
## 6      F  93.50458  5.398124
```

and we want to use the following model:

$$hourlyPay_{(i)} = \beta_0 + \beta_1 gender_{(i)} + \beta_2 expYear_{(i)} + u$$

In such a case, we have to be careful and create only 1 dummy variable (e.g. `gender`) to avoid perfect multicollinearity! For instance, this would be wrong:

$$hourlyPay_{(i)} = \beta_0 + \beta_1 male_{(i)} + \beta_2 female_{(i)} + \beta_3 expYear_{(i)} + u$$

because the variables `male` and `female` would be perfectly correlated! The proper way to encode the information is:

$$gender_{(i)} = \begin{cases} 1, & \text{if individual } i \text{ is a male.} \\ 0, & \text{else} \end{cases}$$

In practice:

```r
df_pay <- df_pay |>
  mutate(gender = ifelse(gender == "M", 1, 0))

head(df_pay, n = 6)
```

```
##   gender hourlyPay   expYear
## 1      1  95.92282  5.119811
## 2      0 108.77277  7.825001
## 3      1  95.36259  4.945302
## 4      0 106.54533 11.384035
## 5      1 100.85228  6.203236
## 6      0  93.50458  5.398124
```

This way we can properly run our model! (We will discuss later how to interpret the coefficient on gender, and how to choose the "base group").

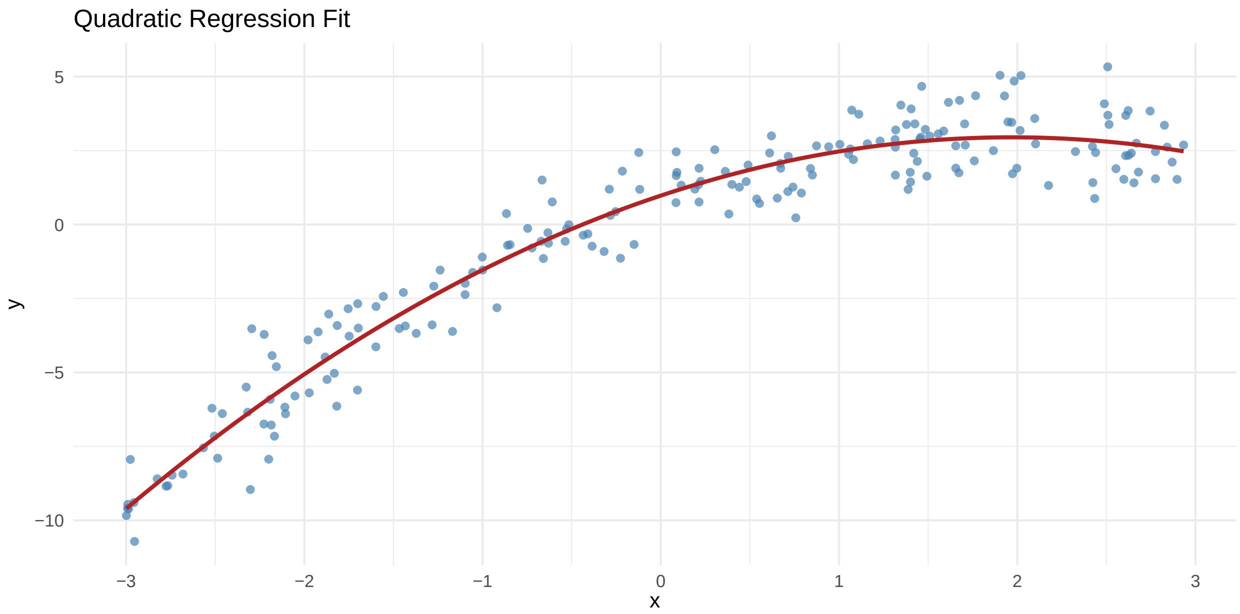
5.4.2. Polynomial Features

The linear regression is called linear because it is linear in its parameter **BUT NOT NECESSARILY IN ITS INDEPENDENT PARAMETERS**! Indeed, we can imagine a model as:

$$y_{(i)} = \delta_0 + \delta_1 x_1 + \delta_2 x_1^2 + u$$

Here is a brief example to demonstrate it:

```r
# 1. our simulated data and the DGP
set.seed(42)
n      <- 200
x      <- runif(n, -3, 3)
u      <- rnorm(n, 0, 1)
delta0 <- 1
delta1 <- 2
delta2 <- -0.5
y      <- delta0 + delta1 * x + delta2 * x^2 + u

df <- data.frame(x = x, y = y)

# 2. We fit the model
model <- lm(y ~ x + I(x^2), data = df)

# 3. We build a prediction grid (here a vector)
x.seq <- seq(min(df$x), max(df$x), length = 300)
pred_df <- data.frame(
  x    = x.seq,
  yhat = predict(model, newdata = data.frame(x = x.seq))
)

# 4. we plot with ggplot2
ggplot(df, aes(x = x, y = y)) +
  geom_point(color = "steelblue", alpha = 0.7) +
  geom_line(data = pred_df, aes(x = x, y = yhat),
            color = "firebrick", size = 1) +
  labs(
    title = "Quadratic Regression Fit",
    x     = expression(x),
    y     = expression(y)
  ) +
  theme_minimal()
```



When we proceed that way, with one variable being a higher-order of another independent variable, we call this a **polynomial feature**. On the other hand, if we multiply two independent variables with each other, we call this an **interaction term**. We will also talk about how to interpret the coefficient on these.

6. Multiple Linear Regression Model (2/2)

Content

- Expected value and Variance of MLR Estimates
- Gauss-Markov Theorem and Gauss-Markov Theorem and Goodness of Fit

Materials

- Wooldridge Chapters 3.2. - 3.4.
- Slides L06

6.1. Unbiasedness of the OLS Estimator

In this new variation of the OLS method, it is quite straightforward to show that the estimators contained in the vector β are unbiased. We rely on the law of iterated expectations:

$$E[\hat{\beta}] = E[E[\hat{\beta} \mid X]] = E[E[(X^\top X)^{-1}X^\top y \mid X]] = E[X^\top X)^{-1}X^\top \cdot E[y \mid X]] = E[X^\top X)^{-1}X^\top X\beta] = \beta$$

In the following sections, we will study more in depth the issues we might encounter in the multivariate linear regression.

6.2. Overspecification

First, what is overspecification? Let us imagine that we have written down:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

but in fact $\beta_2 = 0$ — that means X_2 has no real partial effect on Y. Yet we go ahead and estimate:

$$\hat{Y} \;=\; \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2.$$

Because all the Gauss-Markow assumptions hold, we do not alter the unbiasedness of $\hat{\beta}_1$. The only issue here is that by increasing the degrees of freedom we are in fact inflating the variance of the estimator $\hat{\beta}_1$:

$$\mathrm{Var}(\hat{\beta}_1 \mid X_1, X_2) \;=\; \frac{\sigma^2}{\sum (X_{1i} - \bar{X}_1)^2 \, (1 - R^2_{X_1|X_2})},$$

where $R^2_{X_1|X_2}$ is the R^2 from regressing X_1 on X_2. Because $0 \leq R^2 < 1$, the denominator is smaller than $\sum (X_{1i} - \bar{X}_1)^2$ alone. Hence we obtain a larger variance than in the model without X_2.

6.3. Underspecification: The Omitted Variable Bias

In this section, we talk about a much more frequent issue in causal analysis, the **omitted variable bias**. Let us suppose that the true model or data generating process is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

but because we cannot observe X_2, we estimate β_0 and β_1 with this model:

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{u}$$

In this situation, if $cov(X_1, X_2) \neq 0$, we are indeed violating the exogeneity assumption. This means that we no longer have an unbiased estimate of β_1 (we actually demonstrated this empirically in section 4.1.).

Indeed, the OVB emerges and one can show that we can compute it as:

$$\mathbb{E}[\tilde{\beta}_1] = \beta_1 \; + \; \underbrace{\beta_2 \, \delta_1}_{\text{omitted-variable bias}},$$

where:

$$\delta_1 = \mathbb{E}\left[\frac{\sum i = 1^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^N (x_{1i} - \bar{x}_1)^2}\right] \approx \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)}.$$

In essence, the omitted-variable bias arises whenever (i) we leave out a variable that really belongs in the model and (ii) that omitted variable is correlated with one of our included regressors. In practice, it means our estimated slopes can be systematically too large or too small, and no amount of data will "wash it out" unless we include the missing regressor (or find a valid instrument, more on this later on). Finally, we can notice that we usually have two kinds of bias:

- **upward**: $E[\tilde{\beta}_1] > 0$
- **downward**: $E[\tilde{\beta}_1] < 0$

6.4. Variance

6.4.1. Definitions

In this case, the variance of the different estimators is a little bit more complex to understand than in the simple linear regression. We have:

$$\mathrm{Var}(\hat{\beta}_j \mid X) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad \forall j \in \{1, \ldots, K\} \mathrm{se}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sqrt{SST_j(1 - R_j^2)}}$$

The size of this variance is practically important. A larger variance means a less precise estimator, and this translates into larger confidence intervals and less accurate hypotheses tests.

6.4.2. Components of the Variance

We have three different important components for the variance.

1. The Error Variance σ^2. a larger σ^2 means larger variances for the OLS estimators. This is not at all surprising: more "noise" in the equation makes it more difficult to estimate the partial effect of any of the independent variables on y, and this is reflected in higher variances for the OLS slope estimators. This time, we use the same estimator as before but with the change degrees of freedom where $df = n - (k+1)$:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i}{n - (k+1)}$$

2. The Total Variation in X_k. As we can see, the larger the total sample variation in the variable x the lower the variance of the estimators. Therefore, this is good if a feature has a high variance. Since it is the total variation and not the mean variation we are looking at, increasing the sample size n could be a good way to proceed.

The Mutual Coefficient of Determination R_k^2. In the general case, this r-score is the proportion of the total variation in x_j that can be explained by the other independent variables appearing in the equation. Therefore, the more the variables are correlated (linearly) to each other, the higher the variance of the estimators. We note that $R_k^2 = 1$ is forbidden since we assume no perfect collinearity. High (but not perfect) correlation between two or more independent variables is called multicollinearity. Since multicollinearity is not explicitly forbidden in our assumptions, this is an issue we must approach with carefulness. In practice we obtain this score by regressing X_j with all the other independent variables. If the independent variable for which we measure the coefficient's variance, we would have $R_j^2 = 0$, thus the smallest possible variance.

6.5. Gauss-Markov Theorem & Goodness of the Fit

6.5.1. BLUE

Under the Gauss–Markov assumptions (MLR.1–MLR.5), the OLS estimator enjoys the celebrated BLUE property:

1. Linear

Each OLS coefficient $\hat{\beta}_j$ can be written as a fixed (i.e. non–random) linear combination of the observed outcomes y_i:

$$\hat{\beta}_j = \sum_{i=1}^{N} w_{ij} \, y_i,$$

where each weight $w_{ij} = f_j(x_{i1}, \ldots, x_{ik}$ depends only on the regressor values $\{x_{i1}, \ldots, x_{ik}\}$, not on any y.

2. Unbiased

Because: 1. $\mathbb{E}[u_i \mid X] = 0$ (no systematic error) and 2. the weights w_{ij} sum correctly,

we have:

$$\mathbb{E}\big[\hat{\beta}_j \mid X\big] = \mathbb{E}\Big[\sum_i w_{ij} y_i\Big] = \sum_i w_{ij} \, \mathbb{E}[y_i] = \sum_i w_{ij} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) = \beta_j.$$

In other words, on average OLS hits the true β_j.

3. Best (Minimum Variance)

Among all linear and unbiased estimators of β_j, OLS has the smallest conditional variance:

$$\mathrm{Var}\big(\hat{\beta}_j \mid X\big) \;\le\; \mathrm{Var}(\tilde{\beta}_j \mid X)$$

For every other linear unbiased $\tilde{\beta}_j$

That "best" property means we cannot find another estimator—still a fixed linear combination of the y_i and still unbiased—that is strictly more precise than OLS.

6.5.2. Goodness of the Fit

With the same notation, one can show that

$$SST \;=\; SSE + SSR,$$

where
- $SST = \sum_{i=1}^{N}(y_i - \bar{y})^2$ is the total sum of squares,
- $SSE = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$ is the explained (regression) sum of squares,
- $SSR = \sum_{i=1}^{N} \hat{u}_i^2$ is the residual (error) sum of squares.

Definition:

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

- R^2 measures the fraction of the total variation in y explained by the regressors.
- If the model includes an intercept, $0 \le R^2 \le 1$.

> **Note:** In a regression *without* a constant term, R^2 can be negative, and all $\hat{\beta}_j$ may be biased if the true intercept $\beta_0 \neq 0$. We've introduced an omitted-intercept bias into every slope coefficient.

Because OLS minimizes SSR, adding extra regressors (even irrelevant ones) can never decrease R^2.

Adjusted R^2 penalizes the inclusion of additional regressors:

$$\bar{R}^2 \;=\; 1 \;-\; \frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} \;=\; 1 \;-\; \frac{SSR/(N-k-1)}{SST/(N-1)},$$

where
- k is the number of regressors (excluding the intercept),
- $SSR/(N-k-1)$ is the mean squared error, and
- $SST/(N-1)$ is the sample variance of y.

7. Hypothesis Testing

Content

- Hypothesis Testing

Materials

- Woolridge chapter 4.1. to 4.5.
- Slides L07

7.1. Motivation

In this chapter, we want to start with hypothesis testing to start with some inference analysis. We need to understand how an estimator $\hat{\beta}_j$ is distributed around its mean and how likely it is to deviate from that mean. This will allow us to determine the **statistical significance** of the coefficients $\hat{\beta}_j$ that we compute.

7.2. Structure of a Hypothesis Test

1. First, we formulate the research and determine the parameters for the study. Based on this the null hypothesis H_0 and the alternative hypothesis H_1 can be determined.

2. We set the significance level (i.e. the probability of a false positive or the probability to reject a correct null hypothesis-this what we call the type one error.)

3. We get the data and compute the relevant test statistics (e.g., the standard error).

4. We reject the null hypothesis if the value of the t-test is large enough (depending on the confidence level we set).

   Note: To compute the standard errors and make some approximation about how the estimators are distributed, we will need to make an additional assumption.

7.2.1. Assumption MLR. 6: Sampling Distribution

> u is independent of $(X_1, X_2, ..., X_k)$ and:
>
> $$u \sim \mathcal{N}(0, \sigma^2)$$

This assumption is stronger than what we saw before, and it actually encompasses two important assumptions: MLR. 4 (the exogeneity assumption) and MLR. 5 (the homoskedasticity assumption).

Consequences of Assumption MLR. 6

If assumption MLR. 6 holds, this entails that:

1. **$\hat{\beta}$ is a linear combination of normals.**

The OLS estimator

$$\hat{\beta} \;=\; (X^\top X)^{-1}X^\top Y \;=\; (X^\top X)^{-1}X^\top\bigl(X\beta+\varepsilon\bigr) \;=\; \beta \;+\; (X^\top X)^{-1}X^\top\varepsilon.$$

Since $\varepsilon \sim N(0,\sigma^2 I)$ and $(X^\top X)^{-1}X^\top$ is a fixed matrix (once we condition on X), it follows from the "reproductive property" of the normal distribution that:

$$(X^\top X)^{-1}X^\top\varepsilon \;\sim\; N\bigl(0,\; \sigma^2\,(X^\top X)^{-1}X^\top X\,(X^\top X)^{-1}\bigr) \;=\; N\Bigl(0,\; \sigma^2\,(X^\top X)^{-1}\Bigr).$$

In other words, each component $\hat{\beta}_j$ is itself a (fixed) linear combination of the entries of ε. Hence:

$$\hat{\beta} \;\big|\; X \;\sim\; N\bigl(\beta,\; \sigma^2\,(X^\top X)^{-1}\bigr).$$

Taking the j-th component, we get

$$\hat{\beta}_j \;\big|\; X \;\sim\; N\bigl(\beta_j,\; \mathrm{Var}(\hat{\beta}_j)\bigr), \quad \mathrm{Var}(\hat{\beta}_j) \;=\; \sigma^2\,\bigl[(X^\top X)^{-1}\bigr]jj.$$

That justifies this result:

$$\hat{\beta}_j \mid X \;\sim\; N\bigl(\beta_j,\, \mathrm{Var}(\hat{\beta}_j)\bigr)$$

2. Standardization

$$\frac{\hat{\beta}_j-\beta_j}{\mathrm{sd}(\hat{\beta}_j)} \sim N(0,1)$$

Once we know:

$$\hat{\beta}_j \;\big|\; X \sim\; N\bigl(\beta_j,\; \mathrm{Var}(\hat{\beta}_j)\bigr)$$

it is elementary that if a random variable Z follows $N(\mu,\;\tau^2)$, then:

$$\frac{Z-\mu}{\tau} \;\sim\; N(0,1).$$

In our case,

$$Z=\hat{\beta}_j, \quad \mu=\beta_j, \quad \tau^2=\mathrm{Var}(\hat{\beta}_j), \quad \text{so} \quad \tau=\mathrm{sd}(\hat{\beta}_j) \;=\; \sqrt{\mathrm{Var}(\hat{\beta}_j)}.$$

Therefore

$$\frac{\hat{\beta}_j-\beta_j}{\mathrm{sd}(\hat{\beta}_j)}\;\Bigg|\; X \;\sim\; N(0,1).$$

Because this holds for every fixed X, if we "uncondition" on X (i.e. treat X as nonrandom or consider it fixed in repeated samples), we still have:

$$\frac{\hat{\beta}_j-\beta_j}{\mathrm{sd}(\hat{\beta}_j)} \sim N(0,1).$$

Finally and most importantly, replacing $\mathrm{sd}(\hat{\beta}_j)$ by $\mathrm{se}(\hat{\beta}_j)$ is then a $t_{\,n-k-1}$-statistic.

In practice, we do not know σ^2, so we cannot compute the true:

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \,[\,(X^\top X)^{-1}\,]jj$$

exactly. Instead we estimate σ^2 by the usual unbiased estimator:

$$\hat{\sigma}^2 \;=\; \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}, \quad \hat{\varepsilon}_i = Y_i - X_i^\top \,\hat{\beta}.$$

We then define the estimated standard error of $\hat{\beta}_j$ as:

$$\mathrm{se}(\hat{\beta}_j) \;=\; \sqrt{\hat{\sigma}^2 \left[(X^\top X)^{-1}\right] jj} \;=\; \hat{\sigma}\,\sqrt{\left[(X^\top X)^{-1}\right] jj}.$$

Because $\hat{\sigma}^2$ itself is a random variable (and in fact is proportional to a χ^2 under the normal-errors assumption), the ratio:

$$T_j \;=\; \frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \;=\; \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\,\sqrt{[(X^\top X)^{-1}]jj}}$$

is no longer exactly N(0,1). Instead, one can show (via the fact that $\hat{\beta}_j$ is normal and $\hat{\sigma}^2$ is independent and scaled $\chi^2\, n-k-1)$ that:

$$T_j \;\sim\; t_{\,n-k-1}$$

a Student's t-distribution with $n-k-1$ degrees of freedom. In other words, whenever we replace the true standard deviation $\mathrm{sd}(\hat{\beta}_j)$ by its estimator $\mathrm{se}(\hat{\beta}_j)$, the sampling distribution of the resulting standardized coefficient is:

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \;\sim\; t_{\,n-k-1}.$$

The t score when then compute represents how many "standard deviation away" we are from the true or hypothesized coeffcient β_j. The higher t_{n-k-1} is, the higher the probability that β_j is not the "correct" true value.

7.3. t-Test: Single Parameter Test

Typically, in a single parameter test, we have the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

and we will look at a single parameter β_j. The classic **two-sided hypotheses** we formulate are:

$$\begin{cases} H_0 : \beta_j = 0, & \text{no effect} \\ H_1 : \beta_j \neq 0, & \text{effect} \end{cases}$$

7.3.1. Two-Sided t-Test

Let us make a quick example:

$$children_{(i)} = \beta_0 + \beta_1 \cdot income_{(i)} + \beta_2 \cdot educ_{(i)} + u$$

where we try to look at the influence of the `income` ($income_{(i)}$) of the parents and the accumulated number of years of `education` ($educ_{(i)}$) of the parents on the number of children per couple.

```r
library(MASS)

set.seed(13)

# parameters
beta_0 <- 3 # average number of children when other para are 0
beta_1 <- -0.5 # effect of more income
beta_2 <- 0.3 # effect of more years of educ

# DGP
N = 3000
mu <- c(4.5, 9)
Sigma <- matrix(c(2, 1,
                  1, 4),
                ncol = 2, nrow = 2)

# N joint observations:
draws <- mvrnorm(n = N, mu = mu, Sigma = Sigma)

# extract variables:
income <- log(draws[,1] + 0.001)
educ <- draws[,2]
u <- rnorm(N, mean = 0, sd = 4) # MLR. 6 is respected, sigma = 2

# model
children = beta_0 + beta_1 * income + beta_2 * educ + u

# data fram
df_families <- data.frame(
  income = income,
  educ = educ,
  children = children
)

head(df_families, n = 3)
```

```
##     income      educ   children
## 1 1.354362 10.520193  2.8273740
## 2 1.554988  8.265692 -0.7073612
## 3 1.737417 12.547461  2.8682208
```

We ignore the fact that some couples have “digit” children. This is just a toy example. In our case, we want to investigate the relationship between $income_{(i)}$ and $children_{(i)}$. We want to test the relationship for a 99% confidence level. How should we proceed?

1. **We state the null and alternative hypothesis**

$$\begin{cases} H_0 : \beta_j = 0, & \text{no effect} \\ H_1 : \beta_j \neq 0, & \text{effect} \end{cases}$$

2. We regress our model and compute the standard error

```r
# fit the model
toy_model <- lm(children ~ income + educ, data = df_families)

# summary
summary(toy_model)
```

```
##
## Call:
## lm(formula = children ~ income + educ, data = df_families)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5989  -2.7175  -0.0181   2.5903  13.4335
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.33164    0.39148   8.510  < 2e-16 ***
## income      -0.46870    0.20884  -2.244   0.0249 *
## educ         0.25079    0.03783   6.629 3.99e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.963 on 2993 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.01451,    Adjusted R-squared:  0.01385
## F-statistic: 22.03 on 2 and 2993 DF,  p-value: 3.185e-10
```

As we can see, our model estimates the coefficient on income as $\hat{\beta}_1 = -0.46870$ and compute:

$$\text{se}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sqrt{SST_1(1 - R_1^2)}} \approx 0.20884$$

Then, what should we do? Compute the t-statistics for $df = 2993$ (degrees of freedom). Furthermore, since we look at a two-tailed test for $\alpha = 1\%$, we should compute the critical value (we have $\alpha/2$ in each tail of the student t distribution):

```r
# critical value
df = 2993
alpha = 0.01
crit_1pct <- qt(1 - alpha/2, df)
cat("This is the critical t-value: ", crit_1pct)
```

```
## This is the critical t-value:  2.577473
```

Then, we compute OUR t-statistic:

$$t_{n-k-1} = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} = \frac{-0.46870}{0.20884} \approx -2.2443$$

We take the absolute value of this to compare it against the critical value. Since our t-statistic absolute value is lower than the critical value we computed, we **fail to reject the null hypothesis**. This means that we CANNOT be sure at 99% that income has indeed an impact on the number of children.

How Confidently Can We Reject the Null?

An alternative is now to compute the p-value. This value is is the probability of observing a test-statistic at least as "extreme" as what we actually saw, assuming the null is true.

Concretely, under $H_0 : \beta_1 = 0$ we know:

$$T \ = \ \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} \ \sim \ t_\nu \quad \big(\nu = 2993\big).$$

We computed:

$$t_{\text{obs}} \ = \ \frac{-0.46870}{0.20884} \ \approx \ -2.2443.$$

A two-tailed p-value is

$$p \ = \ P\big(|T| \ \geq \ |\, t_{\text{obs}}\,| \ \big| \ H_0\big) \ = \ 2\left[1 - F_{t_\nu}\big(|t_{\text{obs}}|\big)\right],$$

Numerically, we have:

```
df       <- 2993
t_obs    <- -0.46870/0.20884   # = -2.2443
p_value  <- 2*(1 - pt(abs(t_obs), df))

cat("The p-value is: ", round(100*(p_value), 2), "%")
```

```
## The p-value is:  2.49 %
```

This means that if we had been slightly more tolerant of type I error, we would have rejected the null! The p-value is the "smallest significance level" for which we could reject the null hypothesis, observing t_{n-k-1}

Of course, we can generalize this kind of two-sided test for one sided test, typically:

$$\begin{cases} H_0 : \beta_j = 0, & \text{no effect} \\ H_1 : \beta_j < 0, & \text{effect} \end{cases}$$

What changes is that we reject the null if $t_{n-k-1} < c$, where c is the critical value we compute for:

$$P_{t_v}[T \leq c] = \alpha$$

and we compute the c that satisfies this equation. We can do it quickly here:

```
df    <- 2993
alpha <- 0.01

# critical value for lower-tail alpha
c_lower <- qt(alpha, df)
cat("\nCritical value for one-sided test: ", c_lower)
```

```
##
## Critical value for one-sided test:  -2.327594
```

```
# compute observed t
t_obs <- (-0.46870) / 0.20884   # = -2.2443
cat("\nOur test statistic: ", t_obs)
```

```
##
## Our test statistic:  -2.244302
```

```
if (t_obs <= c_lower) {
  cat("\nReject H0 in favor of beta < 0.\n")
} else {
  cat("\nFail to reject H0.\n")
}
```

```
##
## Fail to reject H0.
```

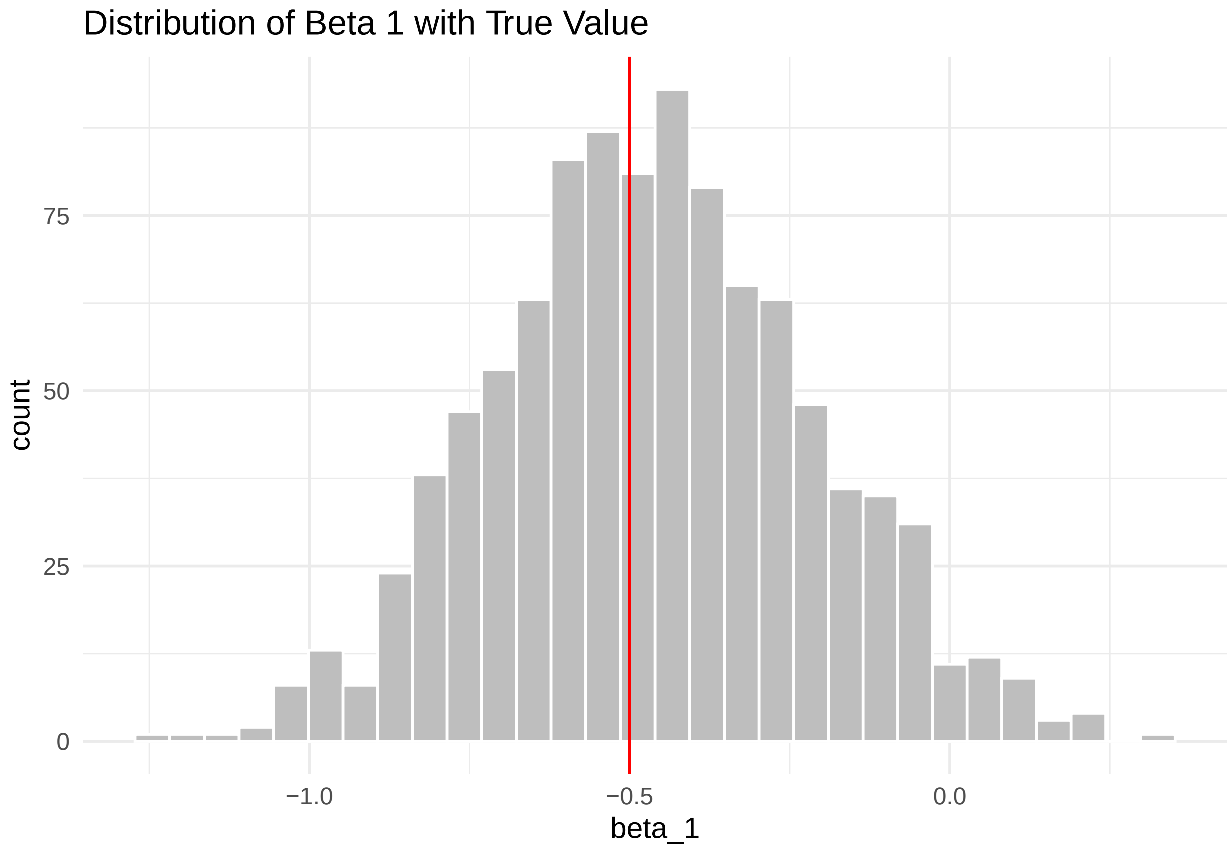
Remarks: Type I Error Under MLR. 1 to MLR. 6, it holds that:

> the Type 1 error of the test, i.e., the probability to reject under the null hypothesis, equals α.

7.4. Confidence Interval

```
library(ggplot2)

set.seed(1)
batch_size <- 2000
M <- 1000
betas_1 <- numeric(M)

for (i in 1:M) {
  # draw 'batch_size' row-indices with replacement:
  idx <- sample(seq_len(nrow(df_families)),
                size = batch_size,
                replace = TRUE)
  # subset by those rows:
  df_sample <- df_families[idx, ]

  # fit the model on the bootstrap sample
  model <- lm(children ~ income + educ, data = df_sample)
  betas_1[i] <- coef(model)[ "income" ]
}

# distribution
ggplot(data = data.frame(beta_1 = betas_1), aes(x = beta_1)) +
  geom_histogram(fill = "gray", color = "white") +
  geom_vline(xintercept = beta_1, color = "red") +
  labs(title = "Distribution of Beta 1 with True Value") +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The confidence interval will allow us to determine an interval:

$$CI = [\hat{\beta}_j - c \cdot \text{se}(\hat{\beta}_j); \hat{\beta}_j + c \cdot \text{se}(\hat{\beta}_j)]$$

Then, under MLR. 1 to MLR. 6 it holds that:

$$P(\beta_j \in CI) = 1 - \alpha$$

In our previous example:

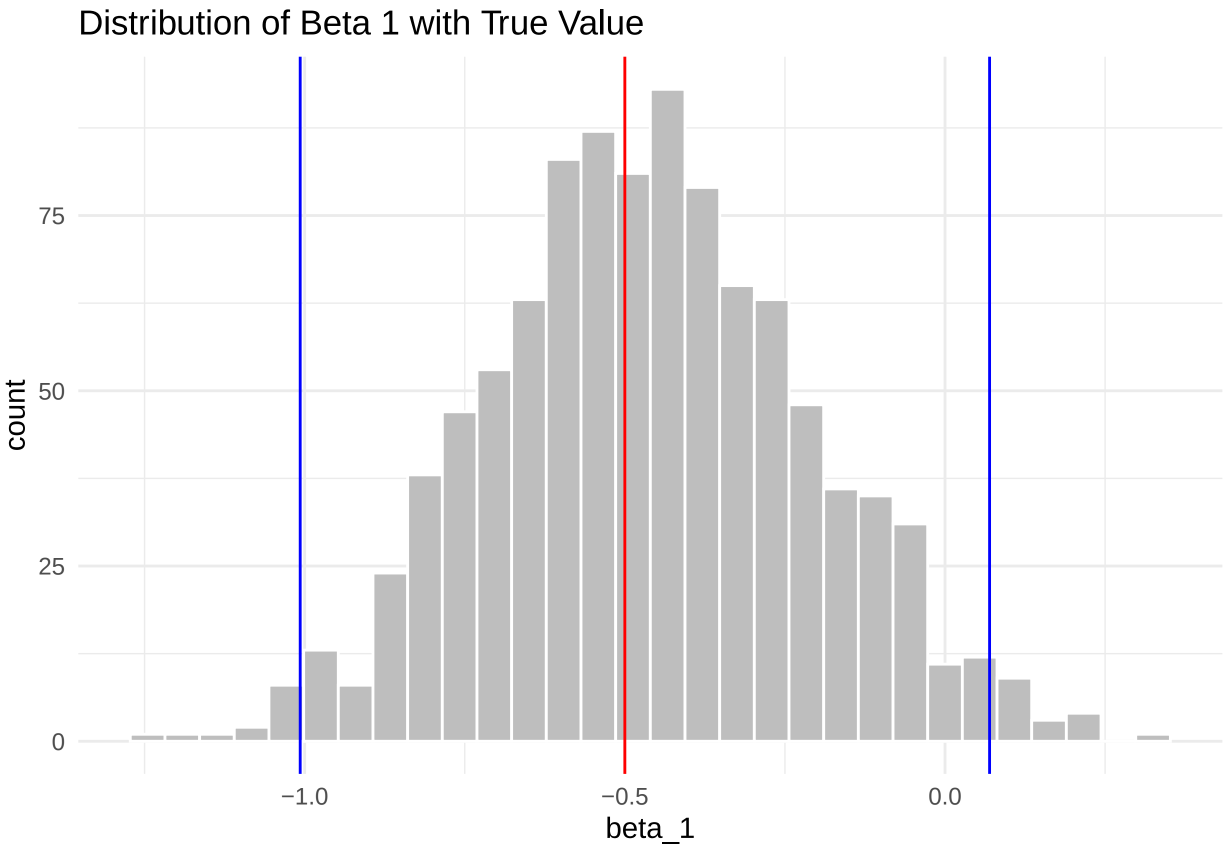
```
# known values
beta_1_hat <- -0.46870
```

```
df = 2993
alpha = 0.01
crit_1pct <- qt(1 - alpha/2, df)
se <- 0.20884

# CI
beta_low = beta_1_hat - crit_1pct * se
beta_high = beta_1_hat + crit_1pct * se

print(c(beta_low, beta_high))
```

```
## [1] -1.00697946  0.06957946
```

This is a quite broad confidence interval! But we can be sure at 95% that the true value of β_1 is in this interval!

```
ggplot(data = data.frame(beta_1 = betas_1), aes(x = beta_1)) +
  geom_histogram(fill = "gray", color = "white") +
  geom_vline(xintercept = beta_1, color = "red") +
  geom_vline(xintercept = -1.00697946, color = "blue") +
  geom_vline(xintercept = 0.06957946, color = "blue") +
  labs(title = "Distribution of Beta 1 with True Value") +
  theme_minimal()
```



WARNING: Since we used `batch_size = 2000`, this is not entirely correct (only for having a visual).

7.5. t-Test: Single Linear Combination

In hypothesis testing, we do not need to test the parameters β_j in isolation. Indeed, we can also run hypothesis test for some linear combinations of the parameters. For instance, if we take our previous toy model:

$$children_{(i)} = \beta_0 + \beta_1 \cdot income_{(i)} + \beta_2 \cdot educ_{(i)} + u$$

One possible hypothesis to test is the fact that `income` has the same magnitude of influence on `children` than `education`. In this case, we would test the hypotheses as:

$$\begin{cases} H_0 : \beta_1 - \beta_2 = 0 \\ H_1 : \beta_1 - \beta_2 \neq 0 \end{cases}$$

In such a case, we compute our t-statistic as:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\mathrm{se}(\hat{\beta}_1 - \hat{\beta}_2)}$$

and here we have to be careful since:

$$sd(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{var(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2cov(\hat{\beta}_1, \hat{\beta}_2)}$$

This means that we would compute our t-value as:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\mathrm{se}(\hat{\beta}_1)^2 + \mathrm{se}(\hat{\beta}_1)^2 - 2cov(\hat{\beta}_1, \hat{\beta}_2)}}$$

7.6. The F-Test: Multiple Linear Restrictions

7.6.1. The Classic F-test

We begin with the leading case of testing whether a set of independent variables has no partial effect on a dependent variable.

Imagine that we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_3 + \beta_2 x_4 + u$$

And we want to test if the subset of independent variables (x_3, x_4) has no influence on y, i.e.:

$$H_0 : \beta_3 = 0 \text{ and } \beta_4 = 0$$

This null hypothesis contains **two exclusion restrictions**. If this hypothesis is not rejected, this means that the subgroup of features (x_3, x_4) has no statistical significance. This is what we call a joint hypotheses test. We should also not that using separate t statistics to test a multiple hypothesis like this can be very misleading. **We need a way to test the exclusion restrictions jointly**.

To proceed, a good way is to observe how the SSR (residual sum of squares) evolve when we drop x_3 and x_4 from the regression. We remember that, because the OLS estimates are chosen to minimize the sum of squared residuals, the SSR always increases when variables are dropped from the model; this is an algebraic fact. The question is whether this increase is large enough, relative to the SSR in the model with all the variables, to warrant rejecting the null hypothesis. Therefore, we now have the restricted model (as opposed to the unrestricted model):

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

The statistic charged to compute whether this relative change is significant is the so-called F statistic. It is defined by:

$$F = \frac{\left(SSR_r - SSR_{ur}\right)/q}{SSR_{ur}/\left(n-k-1\right)}$$

where: SSR_r: Sum of squared residuals for the restricted model SSR_{ur}: Sum of squared residuals for the unrestricted model q: Number of restrictions being tested n: Number of observations k: Number of regressors in the unrestricted model (excluding intercept)

The rejection rule is simple. Once c has been obtained, we reject H_0 in favor of H_1 at the chosen significance level if:

$$F \ > \ c$$

Basically, this means that the subset of predictors (x_3, x_4) does contribute to making accurate predictions– i.e. is statistically significant. To obtain the c value, we need to compute it from the F distribution (which is distributed like a random variable). Again, it is outside the scope of this course to derive this distribution. We should note that the F statistic can also be computed from the R^2 scores:

$$F = \frac{\left(R^2_{ur} - R^2_r\right)/q}{\left(1 - R^2_{ur}\right)/\left(N-k-1\right)}$$

7.6.2. The Overall Significance Test

Furthermore, we can use this formula for computing the overall significance test, where the null is:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$$

Then, we have:

$$F = \frac{R^2/k}{(1-R^2)(n-k-1)}$$

7.6.3. Example

We have the following data set:

```r
library(MASS)
library(dplyr)

df_car <- mtcars |>
  select(mpg, cyl, hp, drat, wt)

head(df_car, n = 3)
```

```
##                mpg cyl  hp drat    wt
## Mazda RX4     21.0   6 110 3.90 2.620
## Mazda RX4 Wag 21.0   6 110 3.90 2.875
## Datsun 710    22.8   4  93 3.85 2.320
```

We attempt to predict the miles per gallon based on the other features:

```r
car_model <- lm(mpg ~ ., data = df_car) # unrestricted model
summary(car_model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df_car)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6171 -1.5663 -0.6058  1.2612  5.8161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.49588    7.44101   4.636  8.1e-05 ***
## cyl         -0.76229    0.63502  -1.200  0.24040
## hp          -0.02089    0.01295  -1.613  0.11845
## drat         0.81771    1.38684   0.590  0.56034
## wt          -2.97331    0.81818  -3.634  0.00116 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.541 on 27 degrees of freedom
## Multiple R-squared:  0.8451, Adjusted R-squared:  0.8222
## F-statistic: 36.84 on 4 and 27 DF,  p-value: 1.438e-10
```

We see that we can compute the F-statistic for the overall significance as:

$$F = \frac{0.8451/4}{(1-0.8451)/(27)} \approx 36.83$$

And now, we want to see if the regressors `hp` and `wt` are relevant. We run a model without them and look at the new R^2 score and then compute the F-statistics. Formally:

$$\begin{cases} H_0 : \beta_3 = 0 \text{ and } \beta_4 = 0 \\ \alpha = 0.05 \end{cases}$$

We run the new model:

```r
car_model2 <- lm(mpg ~ cyl + drat, data = df_car)
r_r2 <- summary(car_model2)$r.squared
ur_r2 <- summary(car_model)$r.squared
cat("The r-squared in the restricted model is: ", r_r2,
    "\nIn the unrestricted model we had r2 = ", ur_r2)
```

```
## The r-squared in the restricted model is:  0.7402482
## In the unrestricted model we had r2 =  0.8451439
```

We can now compute the F-statistics:

```
F = ((ur_r2 - r_r2)/2)/((1-ur_r2)/(27))
cat("The F-statistic is: ", F)
```

```
## The F-statistic is:  9.144572
```

Then, the associated p-value is:

```
F_p <- 1 - pf(9.144572, df1 = 2, df2 = 27)
print(F_p)
```

```
## [1] 0.0009278768
```

This means that we have strong evidence against the null!

8. Consistency & Convergence in Probability

Content

- consistency of estimators
- convergence in probability and some of its properties

Materials

- Woolridge C.2. and C.3.
- Slides L08

8.1. Introduction

Overview
- We have already studied **small-sample** properties of estimators (e.g., unbiasedness under MLR assumptions).
- In Lectures 8 & 9, we shift focus to **large-sample (asymptotic) properties** as $N \to \infty$.

8.1.1. Asymptotic Concepts

- **Consistency**
  - An estimator $\widehat{\theta}_N$ is *consistent* if $\widehat{\theta}_N \xrightarrow{p} \theta_0$ (i.e., it converges in probability to the true parameter θ_0 as sample size N grows).
  - Consistency relies on the **Law of Large Numbers (LLN)**: sample averages converge to population expectations.
- **Asymptotic Normality**
  - Once consistency is established, the **Central Limit Theorem (CLT)** shows $\sqrt{N}\,(\widehat{\theta}_N - \theta_0)$ often converges in distribution to a Normal.

  - This allows us to build approximate confidence intervals and hypothesis tests *without* requiring exact normality of the errors (MLR.6).
- **Convergence in Probability and in Distribution**
  - We recall basic rules (Slutsky's theorem, continuous mapping theorem) for manipulating convergences.
  - Distinguish between:
    - $X_N \xrightarrow{p} c$ (convergence in probability)
    - $X_N \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ (convergence in distribution)

8.1.2. Why Study Asymptotics?

1. **Assessing Large-Sample Behavior**
   - We want estimators that remain "close" to the true parameter with high probability as N grows.

   - If an estimator fails to be consistent, we must (i) recognize its failure and (ii) switch to a different estimator.
2. **Distributional Approximations**

- Even when we cannot assume normal errors (MLR.6), the CLT provides an *approximate* sampling distribution for $\sqrt{N}(\widehat{\beta} - \beta)$.

- This underpins the construction of asymptotic z-tests and confidence intervals in large samples.

8.2. Consistency

8.2.1. The Idea

We already knwo that an estimator W_n of θ is unbiased if:

$$E[W_n] = \theta$$

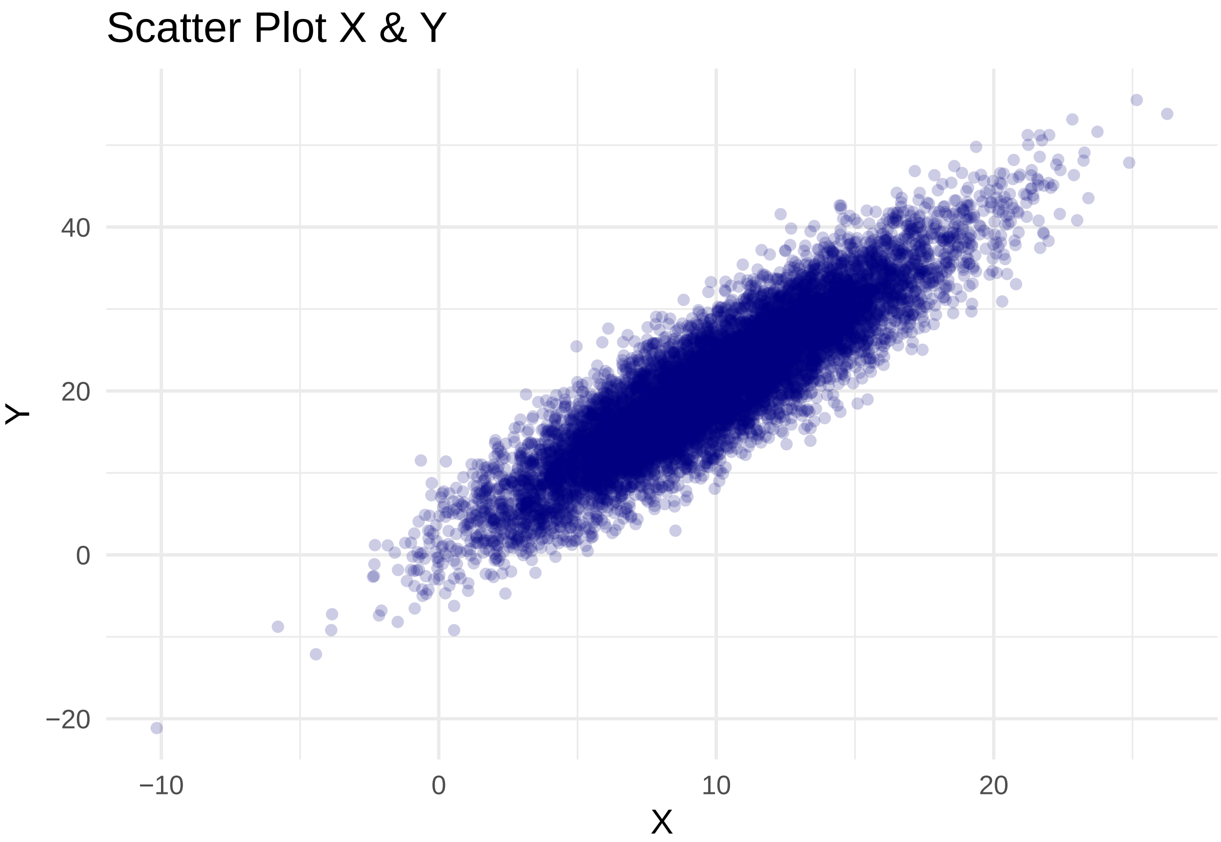
W_n is also a random variable in itself and is distributed according to the pdf f_{W_n}. Now, what consistency tells us verbally is:

> Consistency means that the distribution f_{W_n} gets more and more concentrated around the true parameter θ with larger samples, and when $n = \infty$, all its mass is concentrated at θ.

We can do a quick experiment to show this (univariate model):

$$Y = \beta_0 + \beta_1 X + u$$

```r
# parameters
n = 10000
beta_0 = 1
beta_1 = 2

# DGP
X = rnorm(n, mean = 10, sd = 4)
Y = beta_0 + beta_1 * X + rnorm(n, mean = 0, sd = 4)

# data frame and plot
df <- data.frame(X = X,
                 Y = Y)
ggplot(data = df, aes(x = X, y = Y)) +
  geom_point(color = "navy", alpha = 0.2) +
  labs(title = "Scatter Plot X & Y") +
  theme_minimal(base_size = 13)
```



Now, we will see if, as the sample size we select grow, the parameter $\hat{\beta}_1$ is more concentrated around its mean. To proceed, we use some bootstrap estimates of the distribution and choose three sub-sample size (we resample without replacement):

$$N_1 = 100, \quad N_2 = 1,000, \quad N_3 = 9,000$$

Then, we will plot the distribution of the estimator $\hat{\beta}_1$ for the different sample sizes. We have:

```
# sub sample sizes
n1 <- 100
n2 <- 1000
n3 <- 9000
M <- 1000 # number of bootstraps

# storage for the result
B <- matrix(0, ncol = 3, nrow = M)
colnames(B) <- c("beta_N1", "beta_N2", "beta_N3")

# the experiment
for (j in 1:3){

  if (j == 1){
    n = n1
  } else if (j == 2){
    n = n2
  } else {
    n = n3
  }

  for (i in 1:M){

    # 1. select the sub sample
    sample_idx = sample(nrow(df), size = n, replace = FALSE)
    df_sample = df[sample_idx, ]

    # 2. model
    model <- lm(Y ~ X, data = df_sample)
    B[i, j] <- model$coefficients["X"]

  }
}

# data frame
df_B <- data.frame(B)
head(df_B, n = 3)
```

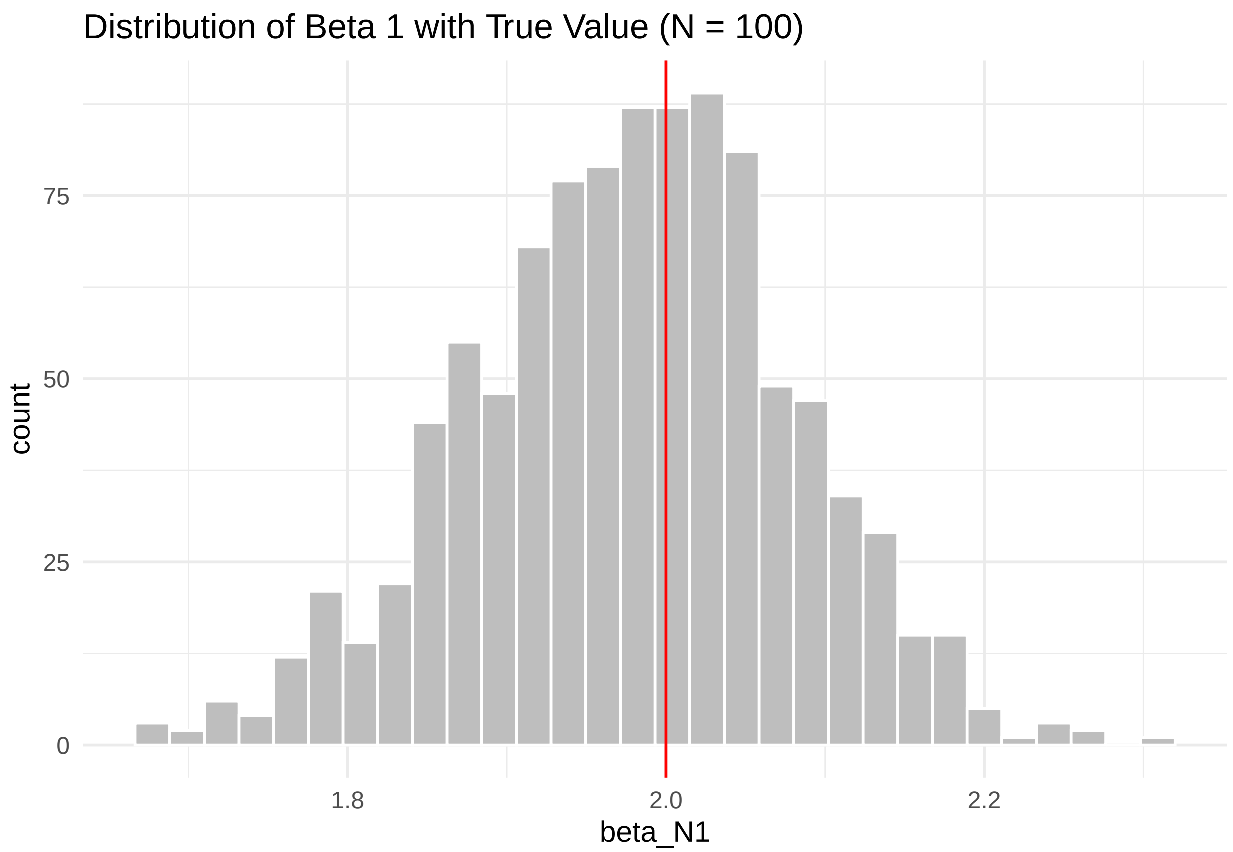
```
##    beta_N1  beta_N2  beta_N3
## 1 1.915013 2.024535 1.974586
## 2 2.053306 1.982442 1.965832
## 3 1.995646 1.943560 1.974472
```

Then, we plot the three different distribution:

For $n = 100$:

```
ggplot(data = df_B, aes(x = beta_N1)) +
  geom_histogram(fill = "gray", color = "white") +
  geom_vline(xintercept = beta_1, color = "red") +
  labs(title = "Distribution of Beta 1 with True Value (N = 100)") +
  theme_minimal()
```

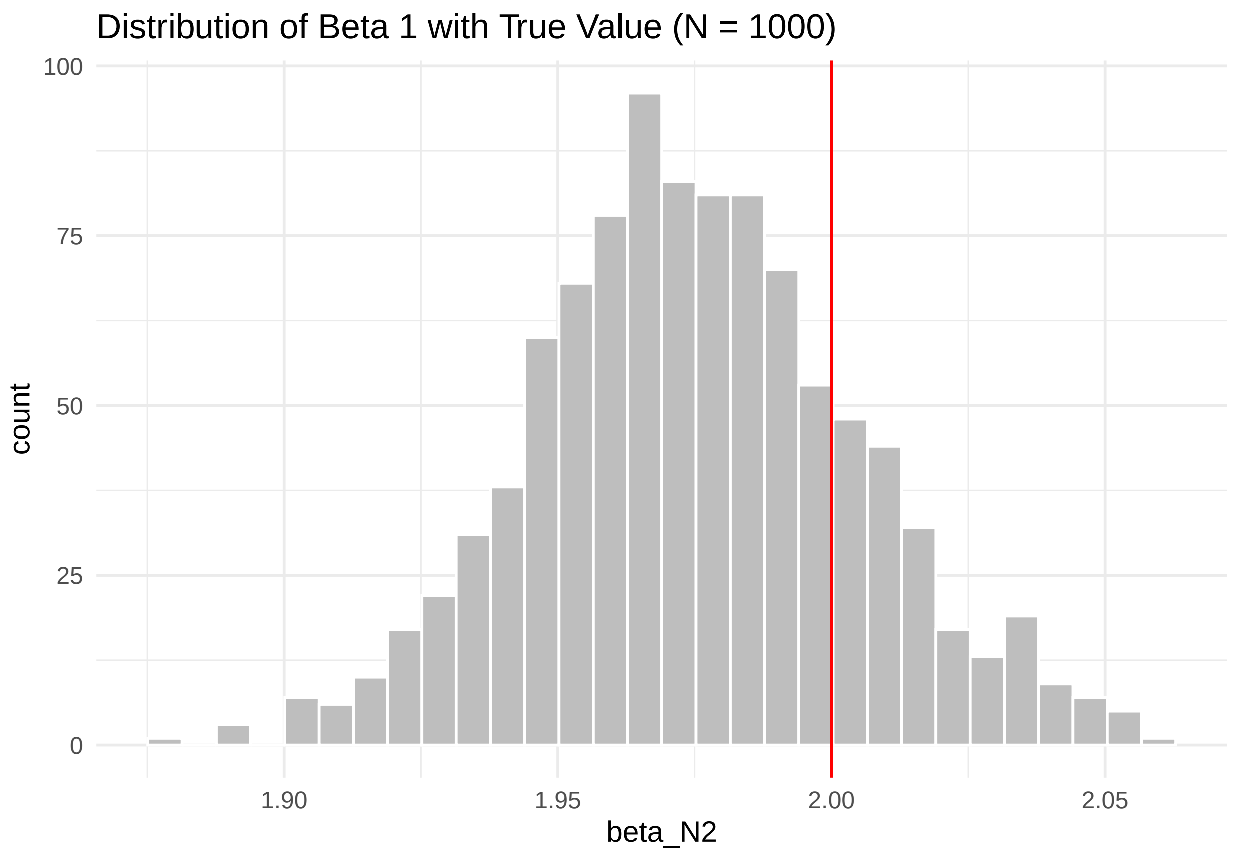
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



For $n = 1000$:

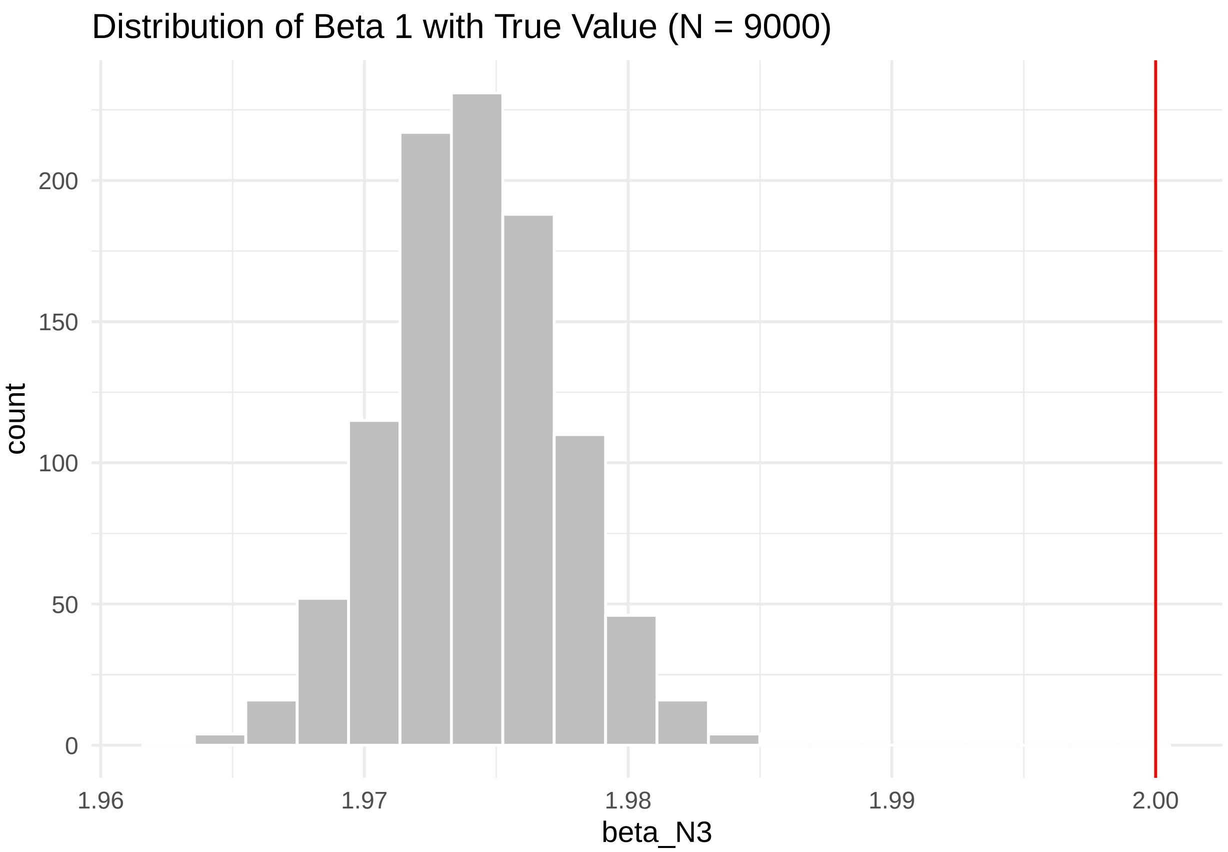
```
ggplot(data = df_B, aes(x = beta_N2)) +
  geom_histogram(fill = "gray", color = "white") +
  geom_vline(xintercept = beta_1, color = "red") +
  labs(title = "Distribution of Beta 1 with True Value (N = 1000)") +
  theme_minimal()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



For $n = 10000$:

```
ggplot(data = df_B, aes(x = beta_N3)) +
  geom_histogram(fill = "gray", color = "white", bins = 20) +
  geom_vline(xintercept = beta_1, color = "red") +
  labs(title = "Distribution of Beta 1 with True Value (N = 9000)") +
  theme_minimal()
```



```
mean(B[, "beta_N1"])  # average of the 1,000 estimates at N = 100
```

```
## [1] 1.978305
```

```
mean(B[, "beta_N2"])  # average of the 1,000 estimates at N = 1000
```

```
## [1] 1.974998
```

```
mean(B[, "beta_N3"])  # average of the 1,000 estimates at N = 9000
```

```
## [1] 1.974153
```

We see that the distribution concentrates around 2.

8.2.2. Formal Definition

> A sequence of estimators W_n is called consistent if the sequence of random variables W_n converges in probability to the true value θ; that is, if for every $\varepsilon > 0$:
>
> $$P(|W_n - \theta| > \varepsilon) \Rightarrow 0, \text{ as } N \to \infty$$

The probability that the absolute difference between the estimator and the true value is greater than any positive number approaches zero asymptotically. If W_n converges to θ in probability we write:

$$\text{plim}(W_n) = \theta$$

8.2.3. Properties of Convergence in Probability

Property 1. If $f : \mathbb{R} \to \mathbb{R}$ is continuous at w, then

$$plim_{N\to\infty} f\big(W_N\big) \;=\; f\big(plim W_N\big) \;=\; f(w)$$

.

Why this is true:

- By assumption $\text{plim} W_N = w$. Equivalently, $W_N \xrightarrow{p} w$.
- The Continuous Mapping Theorem says that whenever a sequence of random variables W_N converges in probability to a constant w, then applying any function f that is continuous at w preserves the convergence. Symbolically,

$$W_N \;\xrightarrow{p}\; w \;\;\Longrightarrow\;\; f\big(W_N\big) \;\xrightarrow{p}\; f(w).$$

Intuitively, if W_N is very near w (with high probability) for all large N, and f does not "blow up" (i.e. f is continuous at w), then $f(W_N)$ will be very near $f(w)$.

Property 2.

$$plim\big(W_N + Z_N\big) \;=\; w + z.$$

Why this is true?

- We know $W_N \xrightarrow{p} w$ and $Z_N \xrightarrow{p} z$. A basic property of convergence in probability is that sums of convergent-in-probability sequences converge to the sum of their limits.

Formally, for any $\varepsilon > 0$:

$$\Pr\big(|(W_N + Z_N) - (w+z)| > \varepsilon\big) \;\le\; \Pr\big(|W_N - w| > \varepsilon/2\big) \;+\; \Pr\big(|Z_N - z| > \varepsilon/2\big),$$

and each of those two pieces goes to zero as $N \to \infty$. Therefore $(W_N + Z_N) \xrightarrow{p} (w+z)$.

Example:

If $W_N \xrightarrow{p} 5$ and $Z_N \xrightarrow{p} -2$, then $W_N + Z_N \xrightarrow{p} 5 + (-2) = 3$.

Property 3.

$$plim\big(W_N\, Z_N\big) \;=\; w\, z.$$

Why this is true?

- We already know $W_N \xrightarrow{p} w$ and $Z_N \xrightarrow{p} z$. Multiplication is a continuous function on $\mathbb{R}^2$. In particular $(x, y) \mapsto x \cdot y$ is continuous at (w, z).

Property 4.

$$\text{plim}\big(W_N/Z_N\big) \;=\; w/z, \quad \text{provided } z \neq 0.$$

Why this is true?

Division is also a continuous operation on $\mathbb{R}^2$ as long as the denominator is not zero. The mapping $(x, y) \mapsto x/y$ is continuous at (w, z) whenever $z \neq 0$.

Since $Z_N \xrightarrow{p} zandz \neq 0$, with high probability Z_N will stay away from zero for large N. Hence $W_N/Z_N \xrightarrow{p} w/z$ by the Continuous Mapping Theorem (or by a decomposition argument analogous to the product case).

8.3. law of Large Numbers

> Let Y_i be a sequence of i.i.d. random variables with expected value $E[Y_i] = \mu$. Then, the sample average is a consistent estimator of μ, that is:
>
> $$plim(\frac{1}{n}\sum_{i=1}^{n} Y_i = \mu)$$

Proof for unbiasedness and consistency:

We assume that each Y_i has finite variance $Var(Y)$. We recall that

$$E\!\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) \;=\; \frac{1}{N}\sum_{i=1}^{N} E(Y_i) \;=\; \frac{1}{N}\left(N\mu\right) \;=\; \mu,$$

and that:

$$Var\!\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) \;=\; \frac{1}{N^2}\sum_{i=1}^{N} Var(Y_i) \;=\; \frac{1}{N^2}\big(N\,Var(Y)\big) \;=\; \frac{Var(Y)}{N} \;\longrightarrow\; 0 \quad \text{as } N \to \infty.$$

WARNING An estimator can be unbiased but not consistent or biased but consistent!

8.4. The Central Limit Theorem

This fundamental theorem allows us to understand how the estimator is distributed for large N. Indeed, many estimators are asymptotically distributed in the sense that:

$$Z_N := \frac{W_N - E(W_N)}{\sqrt{Var(W_N)}} \quad \text{satisfies} \quad P(Z_N \leq z) \to \Phi(z) \text{ for } N \to \infty$$

Formally, the **central limit theorem** tells us that:

If we let the random variables Y_i be i.i.d. with expectation μ and variance σ^2, and denote:

$$\bar{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i s$$

Then, we have:

$$Z_N := \frac{\bar{Y_n} - E(\bar{Y_n})}{\sqrt{Var(\bar{Y_n})}} = \frac{\bar{Y_N} - \mu}{\sigma/\sqrt{n}}$$

that satisfies $P(Z_N \leq z) \to \Phi(z)$ for $N \to \infty$.

9. OLS Asymptotics

Content

- consistency of the OLS estimator
- asymptotic normality of the OLS estimator

Materials

- Woolridge chapters 5.1-5.2 and C.3.
- Slides L09

9.1. Consistency of the OLS Estimator

In the multivariate model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

under MLR. 1 to MLR. 4, the OLS estimator is **consistent** and (adding MLR. 5) asymptotically normal. In this section, we will prove this statement.

1: Repetition: Derivation of $\hat{\beta}_1$ and $\hat{\beta}_0$ Model:
$y = \beta_0 + \beta_1 x + u$

OLS estimation minimizes:

$$\min_{(\hat{\beta}_0, \hat{\beta}_1)} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

(a) FOC w.r.t. $\hat{\beta}_1$:
Let $L(\hat{\beta}_0, \hat{\beta}_1)$ be the loss function:

$$\frac{\partial L}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i \cdot (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Rewriting:

$$\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0 \quad \text{[A]}$$

(b) FOC w.r.t. $\hat{\beta}_0$:

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Rewriting:

$$\sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0 \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{[B]}$$

Using (A) and (B):

$$xy - x\bar{y} = \hat{\beta}_1 (x^2 - x^2) \Rightarrow \hat{\beta}_1 = \frac{xy - x\bar{y}}{x^2 - \bar{x}^2}$$

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

2: Transformation of β_1 We know:

$$\bar{x} = \frac{1}{n}\sum x_i \quad ; \quad \bar{y} = \frac{1}{n}\sum y_i$$

We transform as:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\sum x_i(\beta_0 + \beta_1 x_i + u_i) - \frac{1}{n}\sum x_i \sum(\beta_0 + \beta_1 x_i + u_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

Then:

$$= \frac{\sum x_i(\beta_0 + \beta_1 x_i + u_i) - \frac{1}{n}\sum x_i\left[n\beta_0 + \beta_1 \sum x_i + \sum u_i\right]}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \beta_1 + \frac{\sum u_i(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}$$

This is an **important result** we already discussed.

3: Plim of $\hat{\beta}_1$ According to probability theory and Law of Large Numbers:

$$\text{plim}(\hat{\beta}_1) = \beta_1 + \text{plim}(\frac{\frac{1}{n}\sum u_i(x_i - \overline{x})}{\frac{1}{n}\sum(x_i - \overline{x})^2}) = \beta_1 + \frac{Cov(X,u)}{Var(X)}$$

If $Cov(x,u) \neq 0$, violated $\Rightarrow$ $\hat{\beta}_1$ is biased (not consistent)!

9.2. Asymptotic Normality

Under assumptions MLR. 1-5, we have:

$$\frac{\hat{\beta}_j - \beta_j}{\text{sd}(\hat{\beta}_j)} \sim N(0,1), \forall j \in K$$

and this statement holds if we replace $\text{sd}(\hat{\beta}_j)$ with $\text{se}(\hat{\beta}_j)$. This means that t-test and F-test are valid even without MLR. 6. We can also build confidence intervals. Now, all our tests are **asymptotically valid**.

10. Heteroskedasticity

Content

- consequences of violations of Assumption MLR.6 (homoskedasticity)
- heteroskedasticity robust inference despite MLR.6 being violated

Materials

- Wooldridge Chapters 8.1. to 8.2.
- Slides L10

10.1. Why do we need the homoskedasticity assumption in our models?

There is mostly two reasons that jutifies the necessity of the homoskedasticity assumption, namely:

10.1.1. Gauss–Markov efficiency (BLUE)

Under the standard OLS assumptions (linearity, exogeneity, no perfect multicollinearity, and homoskedasticity), the Gauss–Markov theorem tells us that the OLS estimator $\hat{\beta}$ is not only unbiased but also has the smallest variance among all linear unbiased estimators. In other words, homoskedasticity is what guarantees that OLS is the best (i.e. minimum-variance) linear unbiased estimator. If the errors actually have nonconstant variance (heteroskedasticity), OLS remains unbiased, but there exist other (linear or nonlinear) estimators with strictly smaller variance.

10.1.2. Validity of the usual standard errors and tests

The textbook formula for the OLS variance is defined as:

$$\mathrm{Var}(\hat{\beta}) \;=\; \sigma^2 \big(X^\top X\big)^{-1},$$

this formula implicitly assumes that:

$$\mathrm{Var}(u_i \mid X_i) = \sigma^2 \quad \forall i.$$

If instead $\mathrm{Var}(u_i \mid X_i)$ depends on X_i, then the usual $\hat{\sigma}^2(X^\top X)^{-1}$ formula is incorrect. In practice, that means:

- Our reported standard errors will be wrong (often too small or too large), so t-tests and confidence intervals based on them are invalid.
- p-values and confidence bounds we compute will be misleading.

But what should we do if the homoskedasticity assumption fail?

10.2. Heteroskedasticity

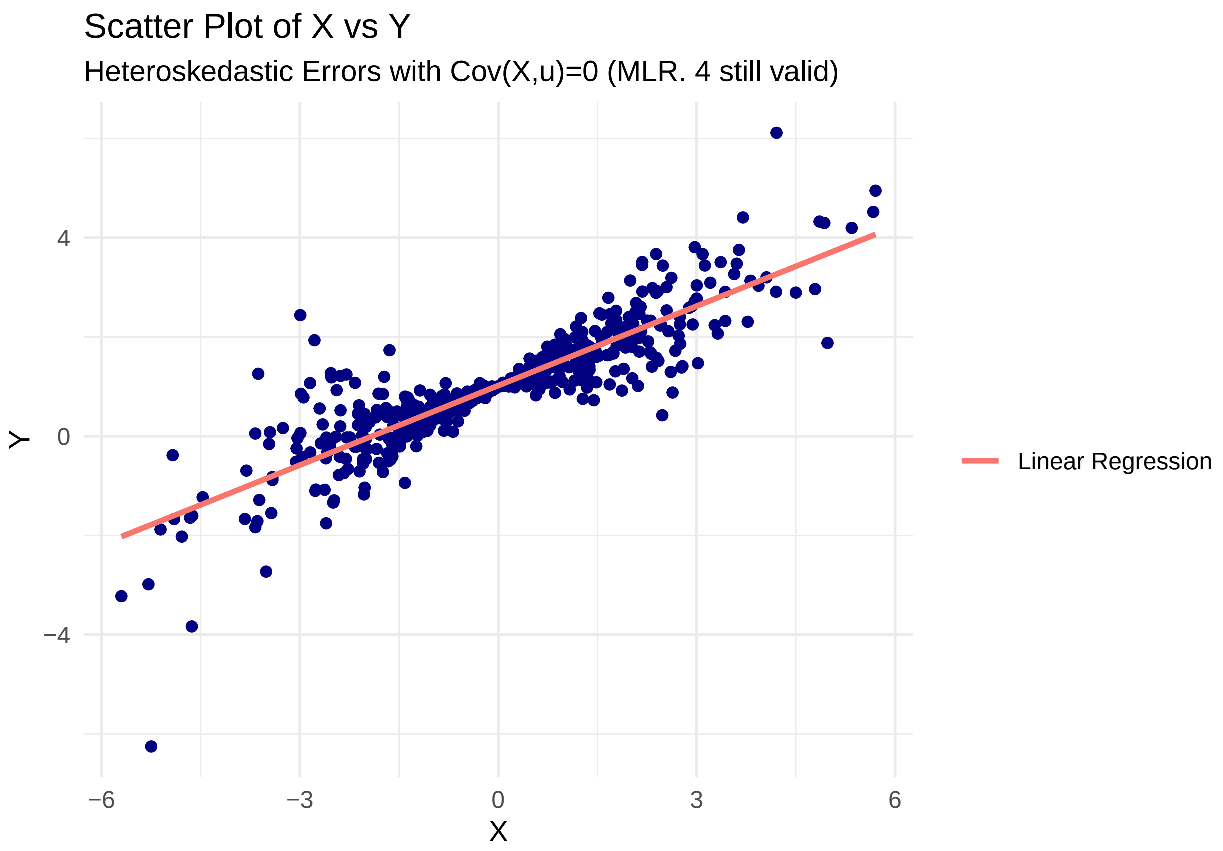
10.2.1 Definition & Example

We have heteroskedasticity in our model as soon as:

$$\mathrm{Var}(u_i \mid x_1, x_2, \ldots, x_k) = \sigma_i^2 \neq \sigma^2$$

Graphically, we could illustrate this as:

```r
# parameters
beta_0     <- 1
beta_1     <- 0.5
scale_effect <- 0.3

# DGP with var(u) growing in |X| but E[u|X]=0
set.seed(2025)
N  <- 500
X  <- rnorm(N, mean = 0, sd = 2)

# eps ~ N(0,1), independent of X
eps <- rnorm(N, mean = 0, sd = 1)

# now u has mean zero (conditional on X) and Var(u|X)=scale_effect^2 * X^2
u  <- eps * abs(X) * scale_effect

# outcome
Y  <- beta_0 + beta_1 * X + u

df <- data.frame(X = X, Y = Y)

ggplot(df, aes(x = X, y = Y)) +
  geom_point(color = "navy") +
  geom_smooth(method = "lm", aes(color = "Linear Regression"), se = FALSE) +
  labs(
    title    = "Scatter Plot of X vs Y",
    subtitle = "Heteroskedastic Errors with Cov(X,u)=0 (MLR. 4 still valid)",
    color    = ""
  ) +
  theme_minimal()
```



We implemented a scale effect and we clearly see tha the variance increases as we deviate from $\mu = 0$.

10.2.2. Consequences

Our assumptions MLR. 1 to MLR. 4 are still valid, but this time MLR. 5 fails! As we said before, our estimator is no longer BLUE, and the standard errors are no longer correctly computed!

10.2.3. Solutions

We basically have **three distinct solutions**:

1. Improve the regression specification
2. Find a new, more efficient unbiased estimator for β_1
3. Live with OLS's inefficiency and search new unbiased estimators for:

- standard errors
- confidence intervals
- hypothesis test

The Weighted Least Squares The WLS is an alternative which we can use to scale with a lower factor observations displaying high variance. A Weighted Least Squares (WLS) model is an extension of ordinary least squares (OLS) that accommodates nonconstant error variances (heteroskedasticity) by giving each observation a weight inversely proportional to its error variance. In effect, observations with "more reliable" (lower-variance) errors receive larger weight in fitting the regression line, while those with "noisier" (higher-variance) errors count for less.

Idea

Suppose we know (or can estimate) each observation's error variance σ_i^2. Then we define weights:

$$w_i \;=\; \frac{1}{\sigma_i^2} \quad \text{(or any constant multiple of } 1/\sigma_i^2\text{)}$$

In WLS, we give observation i weight w_i in the regression criterion. Intuitively, if an observation's error variance is large, we "trust" that point less (small weight), and if its error variance is small, we "trust" it more (large weight).

The WLS objective function

Concretely, we let X be the $n \times k$ matrix of regressors (including a column of 1s for the intercept, if any), Y the $n \times 1$ outcome vector, and β the $k \times 1$ coefficient vector. We denote:

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$$

Then if we know Σ, the WLS estimator solves:

$$\min_{\beta} \sum_{i=1}^{n} w_i \left(y_i - x_i^\top \beta\right)^2, \quad \text{where } w_i \;=\; \frac{1}{\sigma_i^2}$$

Robust Standard Errors The last solution we could implement is to recompute heteroskedasticity robust standard errors for our model. In practice, we can then always use these and not spend extensive amount of time analyzing whether our errors are heteroskedastic or not.

Therefore, we use (in the MLR model):

$$\text{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

where $\hat{r}_{ij}^2$ is the i-th residual of a regression of x_j on all other independent variables:

$$\hat{r}_{ij} = x_{ij} - \left[\hat{\gamma}_0 + \sum_{\forall k \neq j} \hat{\gamma}_k x_{ik}\right]$$

This estimator is not unbiased but consistent. If we do not use this one (see the notebook on the coverage experiment), we will over- or under-estimate our standard errors! We are always on the safe side if we use the robust standard errors!

USEFUL: Notebook 10 Summary

The α-coverage of an estimator measures how often a $(1-\alpha)$-confidence interval overcoats the true value of an estimator.

The heteroskedastic robust variance estimator shows undercoverage or overcoverage in small samples even when the error term is heteroskedastic.

The homoscedastic variance estimator may be substantially biased when the error term is heteroscedastic. Hypothesis tests based on these standard errors are not valid, even asymptotically.

---

11. Model Specification

Content

- Data scaling
- Dummy variables
- Ordinal and nominal variables

Materials

- Wooldridge Chapters 6.1.-6.2. to 7.1.-7.3.
- Slides L11

11.1. Linear Scaling

In the context of the multiple linear regression (MLR) model, "linear scaling" refers to the fact that we can multiply one of our variables (either the outcome or a regressor) by a constant, and the regression coefficients simply scale in the obvious, "inverse" way so that predicted values remain unchanged. In other words, MLR is linear in parameters, so changing the units or "scale" of a variable just rescales its coefficient.

11.1.1. Scaling the Dependent Variable

We suppose our original model is:

$$Y_i \;=\; \beta_0 \;+\; \beta_1 X_{1i} \;+\; \cdots + \;\beta_k X_{ki} \;+\; u_i.$$

If we define a new variable $Y_i' = c \cdot Y_i$ (for some nonzero constant c), then:

$$Y_i' \;=\; c\,\beta_0 \;+\; (c\,\beta_1)\,X_{1i} \;+\; \cdots + \;(c\,\beta_k)\,X_{ki} \;+\; (c\,u_i).$$

In other words: • Every coefficient β_j is multiplied by c. • The residual u_i is also multiplied by c.

So if we regress $Y' = cY$ on the same X-variables, we'll get:

$$\beta_0' = c\,\beta_0, \quad \beta_1' = c\,\beta_1, \;\ldots,\; \beta_k' = c\,\beta_k.$$

That is the hallmark of linear scaling in the dependent variable: multiplying Y by a constant simply multiplies each estimated coefficient (and each residual) by that same constant. The standard error is also scaled by c but the t-statistics will remain the same!

11.1.2. Scaling on of the Regressor X_j

Likewise, if we keep Y and all other X's the same, but define a new regressor:

$$X_{1i}' \;=\; d \cdot X_{1i} \quad \text{(for some nonzero } d\text{)},$$

then our model becomes:

$$Y_i \;=\; \beta_0 \;+\; \beta_1\bigl(X'_{1i}/d\bigr) \;+\; \sum_{j=2}^{k}\beta_j\,X_{ji} \;+\; u_i.$$

Re-writing it as a regression in terms of X_1':

$$Y_i \;=\; \beta_0 \;+\; \left(\tfrac{\beta_1}{d}\right)X'_{1i} \;+\; \sum_{j=2}^{k}\beta_j\,X_{ji} \;+\; u_i.$$

Hence:

- The coefficient on X_1' becomes $\beta_1' = \frac{\beta_1}{d}$.
- All other β_j stay the same (for $j \neq 1$), as does β_0.
- Residuals u_i are unchanged.

Again, predictions line up exactly:

$$\hat{Y}_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \sum_{j=2}^{k}\hat{\beta}_j X_{ji} \;=\; \hat{\beta}_0 + \hat{\beta}_1'\,X'_{1i} + \sum_{j=2}^{k}\hat{\beta}_j X_{ji}.$$

So multiplying X_1 by d forces its slope estimate to shrink by a factor of d, and nothing else changes in the fitted line.

11.2. Logarithmic Scaling

Logarithmic scaling is useful because, as we saw in the previous lecture, it can sometimes help us remove heteroskedastic errors for having homoskedastic errors. Furthermore, logarithmic scaling is also extremely useful for "facilitating" the interpretation of some coefficients.

11.2.1. Log-Log

For instance, let us suppose that we have the following model:

$$\log(Y) = \beta_0 + \beta_1\log(X) + u$$

How should we interpret the estimated parameter $\hat{\beta}_1$? Here is the explanation:

$$e^{\log(Y)} = e^{\beta_0+\beta_1\log(X)+u} \Leftrightarrow Y = cX^{\beta_1} \implies \frac{\partial Y}{\partial X} = c\beta_1 X^{\beta_1 - 1} = \beta_1\frac{cX^{\beta_1}}{X} = \beta_1\frac{Y}{X}$$

This is quite an important result since it implies that:

$$\beta_1 = \frac{\partial Y}{\partial X}\cdot\frac{X}{Y} = \varepsilon_{y,x}$$

This means that β_1 is equal to the elasticity of Y with respect to X! Therefore, we can interpret β_1 in our model as the measure of elasticity. For instance, if we have:

$$\ln(income_{(i)}) = 2.14_{(SE=0.67)} + 2.1_{(SE=0.01)}\cdot\ln(experience_{(i)}) + u_{(i)}$$

This means that for a 1% increase in the years of experience, we have a 2.1% increase in income!

11.2.2. Level-Log

Alternatively, we could have the following model:

$$y = \beta_0 + \beta_1 \log(x) + u$$

How should we interpret the coefficient β_1? Well, if we take the derivative we observe:

$$\frac{\partial y}{\partial x} = \frac{\beta_1}{x} \implies \beta_1 = \frac{\partial y}{\partial x} x$$

This new β_1 could be interpreted as a *semi-elasticity*, in other words, when x changes by 1%, y changes by β_1 units (not in percentage anymore since it is a semi-elasticity, i.e., we do not divide by y).

11.2.3. Log-Level

Finally, we can also imagine the model:

$$\log(y) = \beta_0 + \beta_1 x + u$$

In this case, how should we interpret β_1? Again, we can apply the following mathematical transformation:

$$y = c \cdot e^{\beta_1 x}, \qquad c = e^{\beta_0 + u}$$

This means that:

$$\frac{\partial y}{\partial x} = c\beta_1 e^{\beta_1 x} = \beta_1 y \implies \beta_1 = \frac{\partial y}{\partial x} \cdot \frac{1}{y}$$

This could also be considered as a semi-elasticity, only this time the interpretation is inverted. Indeed, β_1 indicates by how much y increase in *percentage* if x increase by 1 *unit*.

More precisely, for a discrete change of 1 unit, we see that:

$$\frac{\partial \ln(y)}{\partial x} = \beta_1 \implies \Delta \ln(y) = \beta_1 \Delta x \Leftrightarrow \ln(y(x + \Delta x)) - \ln(y(x)) = \beta_1 \Delta x$$

Now, if we set $\Delta x = 1$, we have:

$$\ln(y(x+1)) - \ln(y(x)) = \beta_1 \implies \ln(y(x+1)) = \ln(y(x)) + \beta_1 \implies y(x+1) = y(x) \cdot e^{\beta_1}$$

With this result, we have:

$$\frac{y(x+1) - y(x)}{y(x)} = \frac{y(x)e^{\beta_1} - y(x)}{y(x)} = e^{\beta_1} - 1 \implies \Delta y = e^{\beta_1} - 1$$

Therefore, if x changes by one unit, y changes by this amount.

11.2.4. Practical Considerations

1. We should use logarithmic scaling only when $x \in (0, \infty)$!
2. The t-values change with the log-scale, hence statistical significance can also change!

11.3. Polynomials

As we have seen before, we could specify this kind of model:

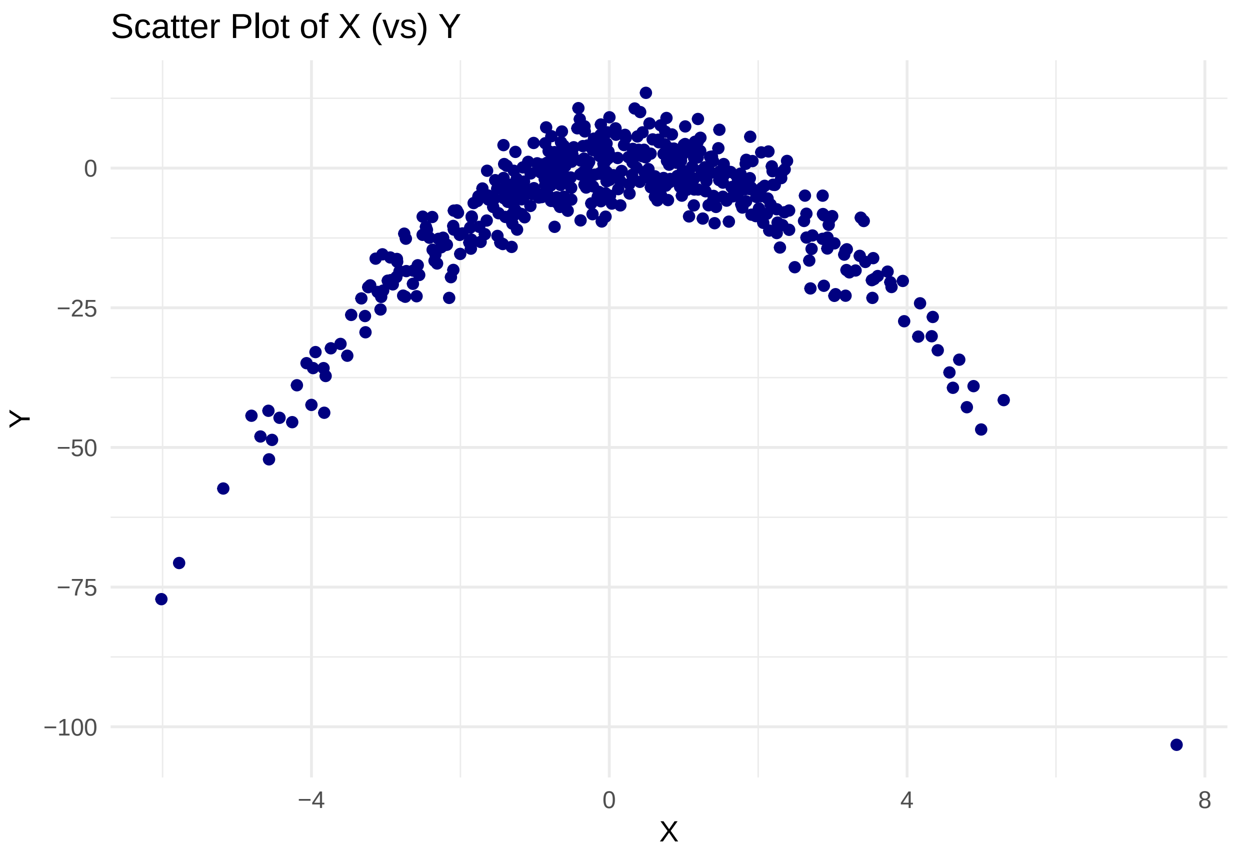
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

In this case, we can no longer interpret β_1 in isolation since the ceteris-paribus does not hold anymore. Therefore, we have to compute the marginal effect as:

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

An example of the marginal effect evolution is given here:

```r
set.seed(1)
# parameters

beta_0 = 1
beta_1 = 1.5
beta_2 = -2

# DGP
N <- 500
X <- rnorm(N, mean = 0, sd = 2)
u <- rnorm(N, mean = 0, sd = 4)
Y <- beta_0 + beta_1 * X + beta_2 * X^2 + u

# data frame
df <- data.frame(
  X = X,
  Y = Y
)

# plot
ggplot(data = df, aes(x = X, y = Y)) +
  geom_point(color = "navy") +
  labs(title = "Scatter Plot of X (vs) Y") +
  theme_minimal()
```



Now, we run the regression for identifying marginal effect of X on Y:

```r
model_quadra <- lm(Y ~ X + I(X^2), data = df)
summary(model_quadra)
```

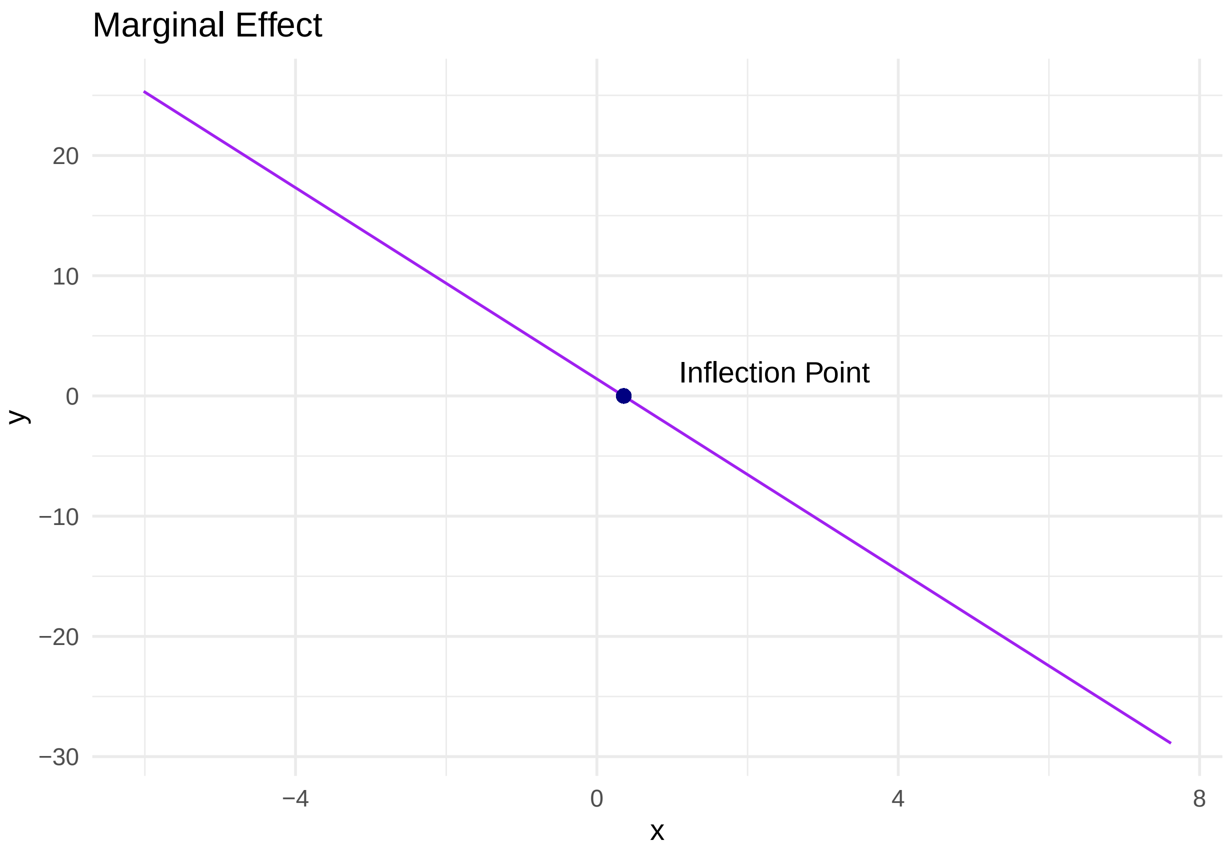
```
##
## Call:
## lm(formula = Y ~ X + I(X^2), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7751  -2.9959   0.0273   2.9094  12.4909
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.77268    0.22860    3.38 0.000782 ***
## X             1.41184    0.09380   15.05  < 2e-16 ***
## I(X^2)       -1.98837    0.03137  -63.39  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.234 on 497 degrees of freedom
## Multiple R-squared:  0.893,  Adjusted R-squared:  0.8926
## F-statistic:  2074 on 2 and 497 DF,  p-value: < 2.2e-16
```

We retrieve the two coefficients and plot the marginal effect as:

$$\Delta y = \beta_1 + 2\beta_2 x$$

```r
beta_1 <- model_quadra$coefficients["X"]
beta_2 <- model_quadra$coefficients["I(X^2)"]

# we build the marginal effect:
x <- seq(min(X), max(X), length.out = 200)
y = beta_1 + 2 * beta_2 * x

df_marginal <- data.frame(x = x,
                          y = y)

# plot
ggplot(data = df_marginal, aes(x = x, y = y)) +
  geom_line(color = "purple") +
  geom_point(aes(x = x[y <= 0.01 & y >= -0.01], 0), color = "navy",
             size = 2) +
  annotate("text", label = "Inflection Point",
           x = x[y <= 0.01 & y >= -0.01] + 2, y = 2) +
  labs(title = "Marginal Effect") +
  theme_minimal()
```

```
## Warning in geom_point(aes(x = x[y <= 0.01 & y >= -0.01], 0), color = "navy", : All aesthetics have le
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```



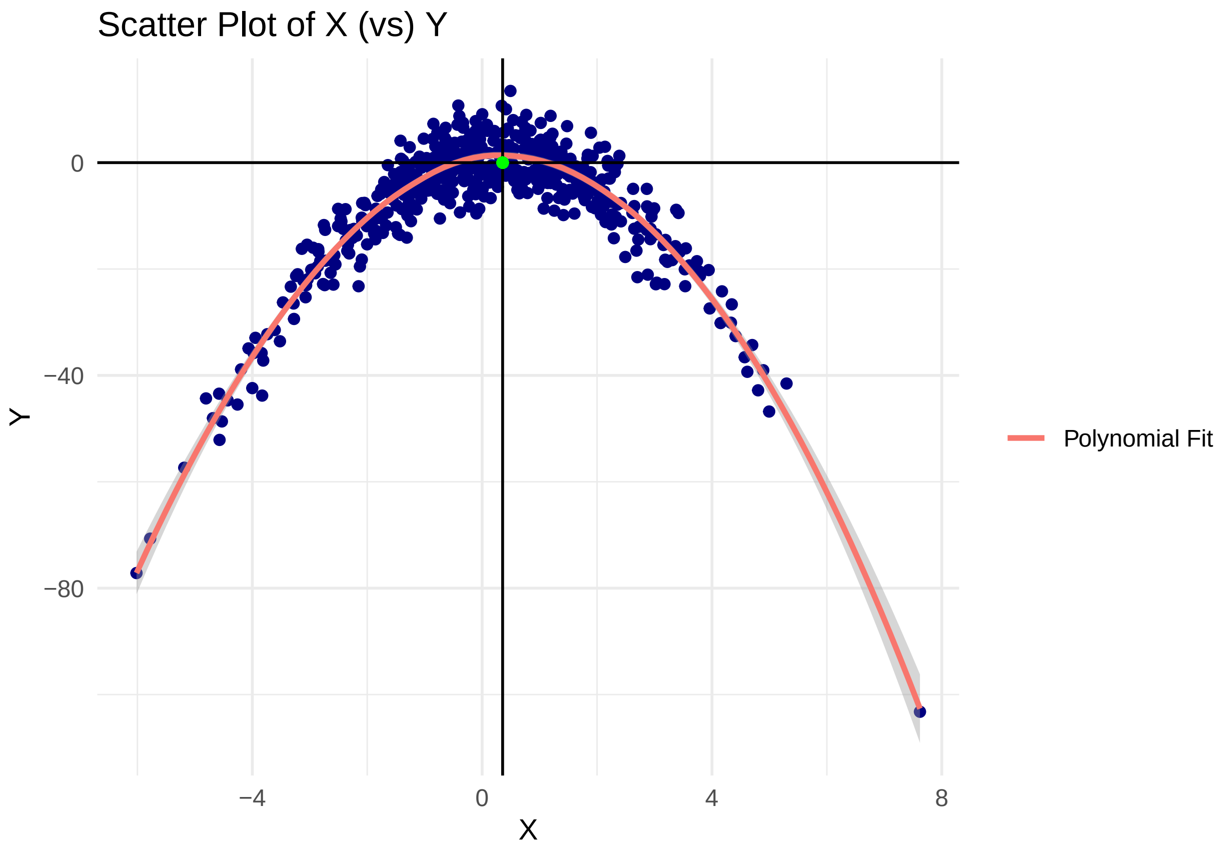
We can identify this point as:

```
coordinates <- c(df_marginal$x[y < 0.01 & y > -0.01], 0)
names(coordinates) <- c("X", "Y")
coordinates
```

```
##         X         Y
## 0.3568098 0.0000000
```

and plot this in our scatter plot:

```
# plot
ggplot(data = df, aes(x = X, y = Y)) +
  geom_point(color = "navy") +
  geom_smooth(method = "loess", aes(color = "Polynomial Fit"), se = F) +
  geom_vline(xintercept = coordinates[1], color = "black") +
  geom_hline(yintercept = coordinates[2], color = "black") +
  geom_point(aes(x = coordinates[1], y = coordinates[2]), color = "green") +
  labs(title = "Scatter Plot of X (vs) Y",
       color = "") +
  theme_minimal()
```



This was a synthetic example to display how to handle marginal effects when we have polynomial features.

11.4. Dummy Variables

In regression (and more broadly in statistical modeling), a dummy variable (also called an indicator variable) is a numeric variable that takes on exactly two values—usually 0 and 1—to encode a categorical feature. By turning categories into numbers, we can include qualitative information (gender, treatment vs. control, region, etc.) in a linear-model framework.

11.4.1. Binary Dummy Variable: Example

We suppose you have a variable female that is "F" or "M." We define

$$D_i = \begin{cases} 1, & \text{if individual } i \text{ is female,} \\ 0, & \text{if individual } i \text{ is male.} \end{cases}$$

Now we can include D in your regression:

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i.$$

- β_0 = average y for the "base" group (male, $D = 0$).
- β_1 = difference in average y between females and males (i.e. "being female" effect).

This example is very important because we introduced the notion of *reference group.* The reference group is the group for whom the intercept is only β_0, here the male, since $D_{(i)} = 0$ if individual i is a male!

11.4.2. Avoid the Dummy Variable Trap

Here, we must avoid that kind of scenario:

```
##   income educ male female
## 1    1.3    9    1      0
## 2    2.0   10    0      1
## 3    3.0    8    0      1
```

Here, we have two variables, `male` and `female`, for encoding only one piece of information! Therefore, we are creating a situation with perfect multicollinearity, and this means that our regressor matrix is no longer invertible!

11.4.3. Multiple Categorical Variable

Let us imagine that we face a data set with data on country A, B, and C. How should we encode properly this information? We start from this kind of data:

```
##   country   pop  gdp
## 1       A   9.2  100
## 2       B  63.1  123
## 3       C 320.0 1000
```

and now we use a technique named "one hot code encoding". What does that mean? One hot means that when a particular value is true, we assign to value 1 to the indicator of this country (to avoid the dummy variable trap, we only have two indicators). This is done as:

```
##   country_A country_B   pop  gdp
## 1         1         0   9.2  100
## 2         0         1  63.1  123
## 3         0         0 320.0 1000
```

In this case, the *reference group* is now country C! **Importantly, unlike a single dummy variable, the choice of reference group for ordinal or nominal variables can have an impact on the significance of the coefficients.**

---

12. Effect Heterogeneity

Content

- Effect Heterogeneity

Materials

- Wooldridge Chapters 6.2. and 7.4.
- Slides L12

12.1. Effect Heterogeneity

12.1.1. The Idea

Our good old MLR assumes **homogeneous effect** in the sense that when we look at parameter β_j it does not discriminate between two groups but average the effect. In the previous lecture, we demonstrated that the intercept can be different for two subgroups. Well, here it is exactly the same scenario, we propose that the partial effects of some variables might differ between two groups!

12.1.2. Mathematical Intuition

We start from the usual MLR:

$$y_i \;=\; \beta_0 \;+\; \sum_{j=1}^{K} \beta_j \, x_{ij} \;+\; u_i,$$

or in vector form

$$y = X\beta + u$$

Here each coefficient β_j is constant across all observations i, so we are assuming, for example, that the "treatment effect" (if one of the x's is a treatment dummy) is the same for everyone.

However, this might be misleading. If subgroups exist (by age, gender, genetic type, prior skill, etc.) for which the effect is systematically different—say positive for some, negative for others—then a single β_j will average those out. In the extreme, one subgroup's +5 and another's –5 average to zero, and we mistakenly conclude "no effect," even though each subgroup experienced a large effect in opposite directions.

12.2. Interaction Terms

The simplest way in MLR to let effects differ across a binary subgroup $D_i \in \{0, 1\}$ is to add an interaction term. We suppose x_{i1} is our "treatment" and D_i is a dummy for Group 1 vs. Group 0. Then:

$$y_i = \;\beta_0 \;+\; \beta_1 \, x_{i1} \;+\; \gamma \, D_i \;+\; \delta \, (D_i \times x_{i1}) \;+\; \sum_{j=2}^{K} \beta_j \, x_{ij} \;+\; u_i.$$

When $D_i = 0$ (Group 0):

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \sum_{j=2}^{K} \beta_j \, x_{ij} + u_i.$$

So the slope on x_{i1} is β_1.

When $D_i = 1$ (Group 1):

$$y_i = (\beta_0 + \gamma) + (\beta_1 + \delta) \, x_{i1} + \sum_{j=2}^{K} \beta_j \, x_{ij} + u_i.$$

So the slope on x_{i1} is now $\beta_1 + \delta$. Thus δ measures the difference in the effect of x_{i1} between Group 1 and Group 0. If $\delta \neq 0$, we have effect heterogeneity.

---

13. Binary Dependent Variables

Content

- Why models for binary dependent variables?
- Specifying Logit and Probit Models
- Estimation of Logit and Probit Models
- Interpretation of Logit and Probit Models

Materials

- Wooldridge Chapter 17.1.
- Slides L13

13.1. Binary Dependent Variables

Sometimes, interesting outcomes are binary. An example could be this dataset `df`:

```
## # A tibble: 5 x 4
##   default loan_val   inc   age
##     <dbl>    <dbl> <dbl> <dbl>
## 1       0      103    80    64
## 2       0      139    73    18
## 3       0      124    71    59
## 4       0       50    70    37
## 5       1       85    67    64
```

And we are interested to know what factor influence the probability that an individual will default on its loan. Therefore, the dependent variable `default` is binary (e.g. $default_{(i)} = 1$ if individual i will default on its loan). Now, if we fit an OLS model to this, we will have (these are fake data) this kind of results:

```r
# we fit the model
lin_model <- lm(default ~ ., data = df_loan)

# we show the results
summary(lin_model)
```

```
##
## Call:
## lm(formula = default ~ ., data = df_loan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3730 -0.2870 -0.2348  0.6350  0.8200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.140752   0.169145   0.832    0.406
## loan_val     0.001194   0.001024   1.166    0.245
## inc          0.001234   0.001844   0.669    0.504
## age         -0.001322   0.002070  -0.639    0.524
```

```
##
## Residual standard error: 0.4431 on 196 degrees of freedom
## Multiple R-squared:  0.01231,    Adjusted R-squared:  -0.002804
## F-statistic: 0.8145 on 3 and 196 DF,  p-value: 0.4872
```

Now, the question arises as to how we should interpret the coefficients at our disposal. Furthermore, if we use them to make a prediction, how should we interpret the output of the prediction, which will be continuous? What happens when the predicted $\hat{default}_{(i)}$ exceeds 1? All those new questions emerge because now we have an independent **binary** variable. Well, we have two models which enable us to avoid predictions outside the range $(0, 1)$:

- The Logit Model: the outcome of the linear regression is modeled as logistic function.
- The Probit Model: the outcome of the linear regression is modeled as a standard normal cdf $\Phi(X\beta)$.

13.2. Specification of the Models

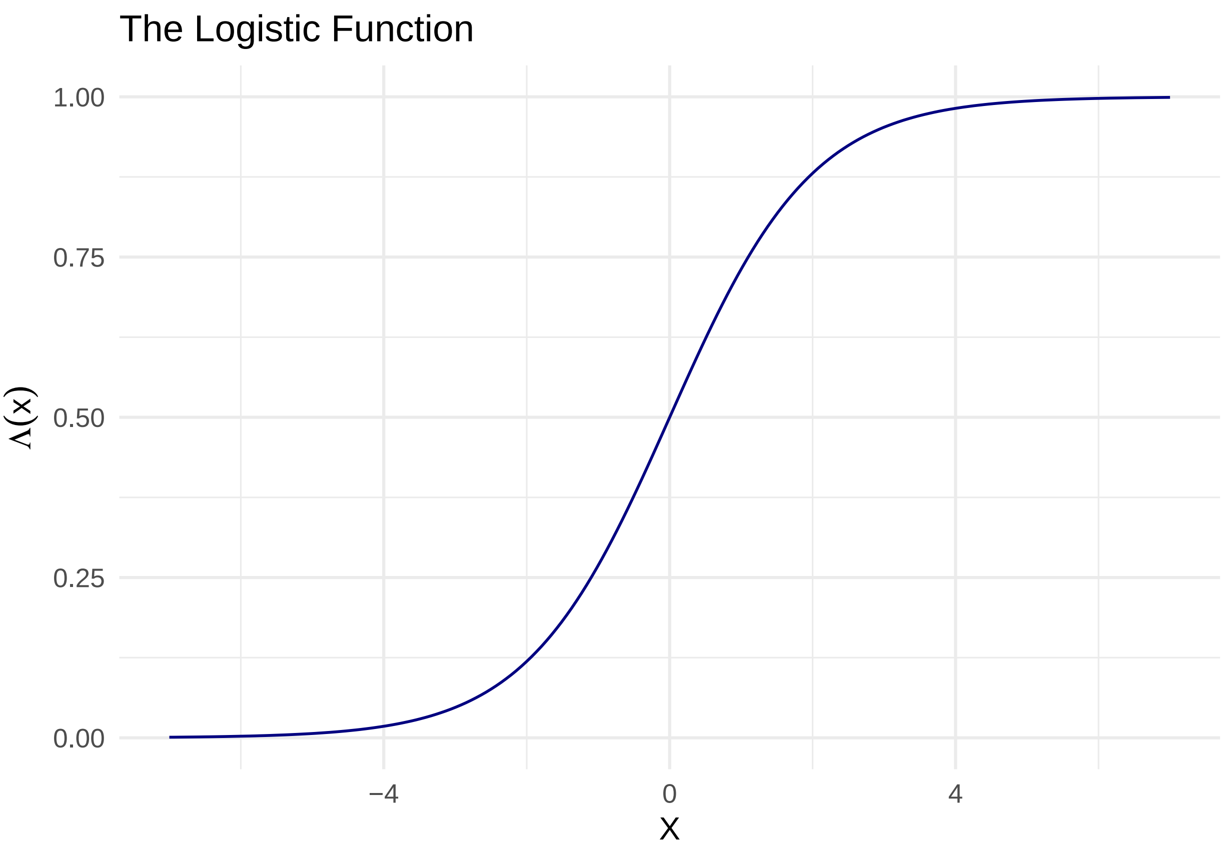
13.2.1. The Logit Model

The logit model will use the ouput of our linear regression and model it as a logistic function for generating continuous values in the range $(0, 1)$. The logistic function is given by:

$$G(z) = \frac{e^z}{1 + e^z} := \Lambda(z)$$

We can map this function as:

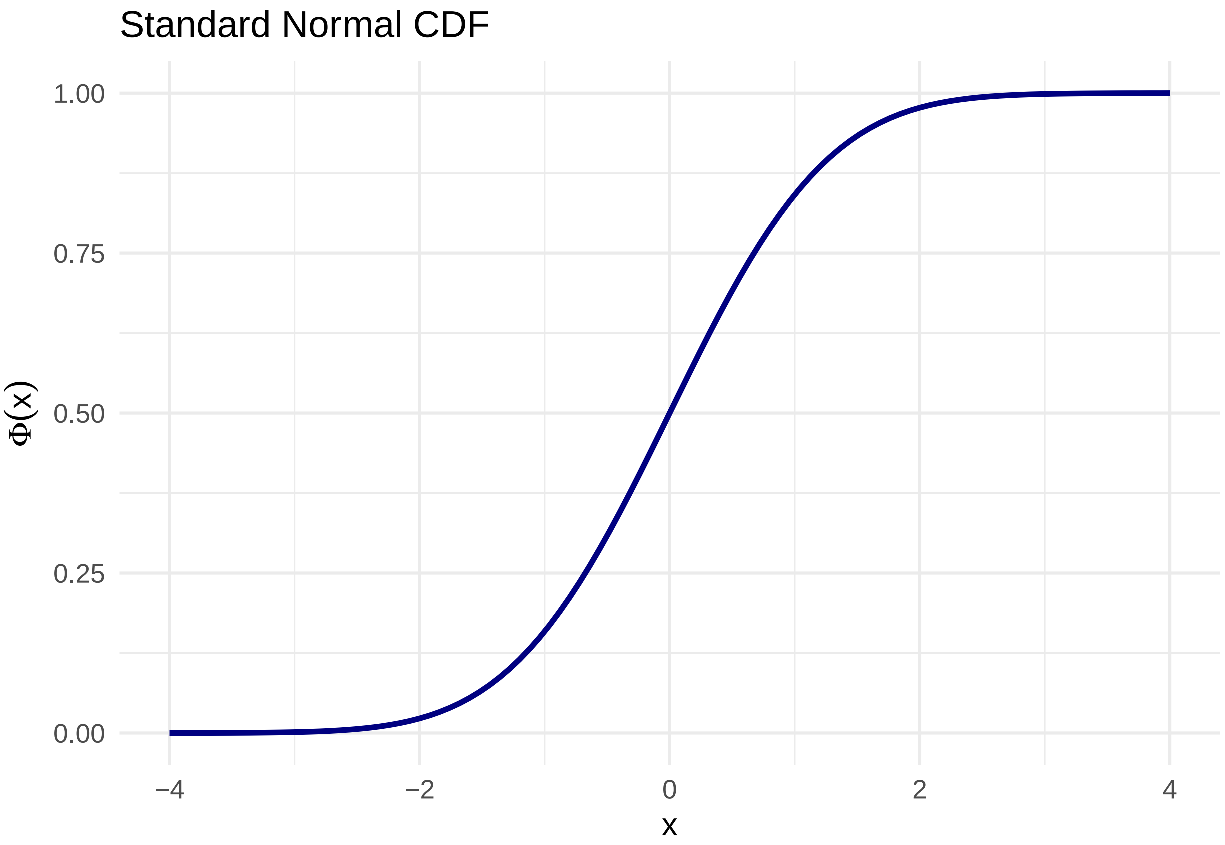
```r
logistic <- function(x){
  a <- exp(x)
  b <- 1 + a
  return (a / b)
}

# the variables
x <- seq(from = -7, to = 7, length.out = 200)
y <- logistic(x)

# the data frame
df <- data.frame(
  x = x,
  y = y
)

# the plot
ggplot(data = df, aes(x = x, y = y)) +
  geom_line(color = "navy") +
  labs(title = "The Logistic Function",
       x = "X",
       y = expression(Lambda(x))) +
  theme_minimal(base_size = 12)
```



13.2.2. The Probit Model

Alternatively, we have the probit model which relies on the very familiar standard normal cdf:

```
df <- data.frame(x = seq(-4, 4, length.out = 300))

ggplot(df, aes(x = x)) +
  stat_function(fun = pnorm,
                args = list(mean = 0, sd = 1),
                linewidth = 1,
                color = "navy") +
  labs(
    title = "Standard Normal CDF",
    x     = "x",
    y     = expression(Phi(x))
  ) +
  theme_minimal(base_size = 12)
```


Standard Normal CDF

13.3. Derivation of the Logit and Probit Models

13.3.1. Derivation

The underlying framework is a **latent variable model** (latent since we transform the target variable we obtain with the first linear regression). We have:

$$y^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + e, \qquad \text{whith: } y = 1[y^* > 0]$$

- y^* is the latent variable
- $1[\cdot]$ is the indicator function ($= 1$ if condition if fulfilled)
- e is independent of x and has either a standard logistic or standard normal distribution (symmetric around 0).

Once we have defined this, we can look at the **response probability** for y:

$$P(y = 1 \mid x_1, \dots, x_k) = P(y^* > 0 \mid x_1, \dots, x_k) = P(e > -(\beta_0 + \sum_{j=1}^{k} \beta_j x_j) \mid x_1, \dots, x_k)$$

Now, since we assumed that e is distributed according to a standard normal or logistic CDF. This means that we have:

$$c = \beta_0 + \sum_{j=1}^{k} \beta_j x_j P(e > -c) = 1 - P(e \leq -c)$$

and since both of our CDF are symmetric:

$$P(e \leq -c) = 1 - (1 - P(e \leq c)) = G(\beta_0 + \sum_{j=1}^{k} \beta_j x_j) = \begin{cases} \Lambda(\beta_0 + \sum_{j=1}^{k} \beta_j x_j) & \text{Logit Model} \\ \Phi(\beta_0 + \sum_{j=1}^{k} \beta_j x_j) & \text{Probit Model} \end{cases}$$

13.3.2. Estimation

Now, we ask ourselves the question of how we should estimate this model, namely, how do we choose the best value for the weight vector β? The value of the parameter β will determine what is the accuracy of our model (the relative number of correct predictions).

To derive the optimal β, we will use the **Maximum Likelihood Estimation**. This means that we want to maximize the joint density function at the observed data sample $y = (y_1, y_2, \ldots, y_n)$.

$$\mathcal{L}((y \mid x_j); \beta) = f((y \mid x_j); \beta)$$

and now, since we computed before $f(y = 1 \mid x_j)$ as $G(X\beta)$, where $X = (x_1, x_2, \ldots, x_k)^\top$, we have:

$$\mathcal{L}_n(\beta; (y \mid x_i)) = \prod_{i=1}^{n} G(x_i\beta)^y (1 - G(x_i\beta))^{1-y}$$

This is the case since y exactly indicates the number of realized “true” realizations. Then, we can take the log-MLE for facilitating the derivation of the optimal β. However, teh precise derivation is outside the scope of this course.

13.4. Interpretation of Coefficients

The coeffcients β_j show the effect of the independent variables on the latent variable y^*. This is often not interesting since we do not have defined units for those (standardization of all the variables could contribute to feature importance…). We are interested in kowing how the independent variables x_j influence:

$$P(y = 1)$$

First, we should be aware of the fact that the direction is the same for both the latent and independent variable (this means that a negative β_j implies that x_j is decreasing the probability $P(y = 1)$). What we must compute now is the magnitude of the influence. This can be computed as:

$$\frac{\partial p(x_1, \ldots, x_k)}{\partial x_j} = g(x_1, \ldots, x_k)\beta_j, \qquad \text{with } g(z) = \frac{\partial G(z)}{\partial z}(z)$$

In general, the marginal effect $\partial p / \partial x_j$ is dependent on the other variables, but the relative effect of x_j and x_h is not dependent on $(x_1, \ldots, x_k)$. Therefore, we have:

$$\Delta \hat{P}(y = 1 \mid x_1, \ldots, x_k) = [g(\hat{\beta}_0 + \hat{\beta}_1 x_1 + (\cdots) + \hat{\beta}_k x_k)\hat{\beta}_j]\Delta x_j$$

And here we see again that the partial effect is dependent on the values of the other regressors. Therefore, to generalize, we define two common summaries of marginal effects:

13.4.1. Partial Effect at the Average (PEA)

We plug in the mean of each regressor $\bar{x}_j$ into the density g to get a single "typical" slope:

$$\text{PEA}j = \frac{\partial P(y = 1 \mid x)}{\partial x_j}\Bigg| x = \bar{x} = g\big(\hat{\beta}_0 + \sum_{m=1}^{k} \hat{\beta}_m \, \bar{x}_m\big)\, \hat{\beta}_j.$$

In practice, we compute:

- $\bar{x}_m = \frac{1}{n} \sum_{i=1}^{n} x_{im}$ for each m.
- $\widehat{z} = \hat{\beta}_0 + \sum_{m=1}^{k} \hat{\beta}_m \, \bar{x}_m$.
- evaluate $g(\widehat{z})$, where $g(z) = G'(z)$ is the logistic (or normal) pdf.
- multiply by $\hat{\beta}_j$.

This gives us an easy-to-interpret "slope" in probability terms, at the "average" individual.

13.4.2. Average Partial Effect (APE)

Rather than one point, we can average the individual marginal effects over our sample:

$$\text{APE}j = \frac{1}{n} \sum_{i=1}^{n} \Big[g\big(\hat{\beta}_0 + \textstyle\sum_{m=1}^{k} \hat{\beta}_m \, x_{im}\big)\, \hat{\beta}_j \Big].$$

This accounts for the fact that $g\big(x_i^\top \hat{\beta}\big)$ varies across observations.

13.4.3. Example

We will take our simulated data set `df_loan` to make a short example:

```
## # A tibble: 6 x 4
##   default loan_val   inc   age
##     <dbl>    <dbl> <dbl> <dbl>
## 1       0      103    80    64
## 2       0      139    73    18
## 3       0      124    71    59
## 4       0       50    70    37
## 5       1       85    67    64
## 6       0      133    40    42
```

We run a linear regression to compute the coefficients β_j:

```
# OLS model
mod_1 <- lm(default ~ ., data = df_loan)

# summary
summary(mod_1)
```

```
##
## Call:
## lm(formula = default ~ ., data = df_loan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3730 -0.2870 -0.2348  0.6350  0.8200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.140752   0.169145   0.832    0.406
## loan_val     0.001194   0.001024   1.166    0.245
## inc          0.001234   0.001844   0.669    0.504
## age         -0.001322   0.002070  -0.639    0.524
##
## Residual standard error: 0.4431 on 196 degrees of freedom
## Multiple R-squared:  0.01231,    Adjusted R-squared:  -0.002804
## F-statistic: 0.8145 on 3 and 196 DF,  p-value: 0.4872
```

Now that we have established the different coefficients, we can use what we learned to compute both the partial effect at the average (PEA) and the average partial effect (APE). We have:

```
PEA_lm <- function(df, target, model_selection) {
  # target: the name of the response column in df, as a string
  var_names <- colnames(df)
  reg_names <- setdiff(var_names, target)

  # 1) compute the column means of the regressors
  avgs <- colMeans(df[, reg_names, drop = TRUE])

  # 2) fit the (glm) model
  fam <- if (model_selection == "logit") binomial(link = "logit")
         else                             binomial(link = "probit")
  f   <- as.formula(paste(target, "~", paste(reg_names, collapse = " + ")))
  model <- glm(f, data = df, family = fam)

  # 3) extract coefficients
  beta <- coef(model)

  # 4) compute the linear predictor at average x
  zbar <- beta["(Intercept)"] + sum(beta[reg_names] * avgs)

  # 5) get the density g(zbar)
  g <- if (model_selection == "logit") {
         dlogis(zbar, location = 0, scale = 1)
       } else {
         dnorm(zbar, mean = 0, sd = 1)
       }

  # 6) compute PEA for each regressor
  PEAs <- g * beta[reg_names]
  return(PEAs)

}
```

Once we coded our function, we use it as:

```
pea_1 <- PEA_lm(df = df_loan,
                target = "default",
                model_selection = "logit")

# percentage
100 * pea_1
```

```
##    loan_val       inc        age
##   0.1202613  0.1223652 -0.1329316
```

Interpretation

loan_val ß0.0012026 A one-unit increase in loan value (e.g. USD 1, if our units are dollars) is associated with a +0.0012026 increase in the probability of $y = 1$ (about 0.12 percentage points).

*inc*ß0.0012237 A one-unit increase in income (again in whatever units we used, say USD 1k) raises $P(y = 1)$ by roughly 0.122 percentage points.

*age*ß˘0.0013293 A one-year increase in age lowers the probability of $y = 1$ by about 0.133 percentage points.

Odds Ratio In a logistic regression:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{\text{female}} \cdot (\text{female indicator}) + \dots$$

the coefficient β_{female} is itself a log odds ratio. Exponentiating it gives

$$\exp\left(\beta_{\text{female}}\right) = \frac{\text{odds if female}}{\text{odds if male}},$$

which you we directly interpret!

For instance, if the odds ratio is 0.98, this means that women are 2% less likely to be employed.

14. Introduction to Causal Inference

Content

- Correlation vs. causation
- Selection bias.

Materials

- Slides L14

14.1. Treatment Effects and Potential Outcomes

In causal inference, we want to estimate **Treatment Effects**. The questions we are asking are:

- What is the effect of a hospital visit on health?
- What is the effect of a voluntary math course on the econometrics grade?
- What is the effect of an elite college on wages?
- What is the effect of playing basketball on height?

To do so, the ideal would be to observe how a person reacts with and without treatment. However, in real life, this ideal scenario is, most of the time, not fulfilled. Therefore, we have to estimate the unobserved potential outcome.

14.2. Selection Bias: Endogeneity

One issue we encounter quite often in economics is the so-called selection bias. This term refers to the fact that people benefiting from a treatment self-selected themselves for being treated. Self-selection bias arises in any situation in which individuals select themselves into a group, causing a biased sample with **nonprobability sampling** (this violates MLR. 2 and, as a consequence, this potentially induces a violation of MLR. 4). It is commonly used to describe situations where the characteristics of the people which cause them to select themselves in the group create abnormal or undesirable conditions in the group.

In causal inference, we must find clever ways to avoid selection bias!

14.3. Randomized Controlled Trial (RCT): Experiments in Labs & In Nature

When the researchers are testing the significance of a regressor (e.g., a treatment where $D_{(i)} = 1$ if the individual took the treatment), they oftentimes use RCT. Participants who enroll in RCTs differ from one another in known and unknown ways that can influence study outcomes (violating endogeneity assumption), and yet cannot be directly controlled. However, by randomly allocating participants among compared treatments, an RCT enables statistical control over these influences. Provided it is designed well, conducted properly, and enrolls enough participants, an RCT may achieve sufficient control over these confounding factors to deliver a useful comparison of the treatments studied.

Mathematically, we “push” all the unkown features $(x_2, \dots, x_k)$ in the error term u and thanks to our controlled random process, we can say that:

$$E(u \mid x_1) = 0$$

However, economists cannot really use RCT because in real life the government would typically not attribute social coverage completely randomly… Still, in the so-called ”credibility revolution” in economics, natural experiments have been increasingly used to estimate causal effects over the past 30 years or so. Researchers use clever methods to counter the self-selection bias such as:

- Instrument Variables (L15)
- Proxy Variables (L17)
- Regression Discontinuity

14.4. Formal Representation

14.4.1. Settings

Notation:

- Treatment indicator: $D_i \in \{0, 1\}$
- Potential outcome: Y_{D_i}
- observed outcome: $Y_i = D_i \cdot Y_{1i} + (1 - D_i)Y_{0i}$

Treatment Effect

- Individual causal effect: $Y_{1i} - Y_{0i}$ (typically not observable in economics)
- Average Treatment Effect: $E(Y_{1i} - Y_{0i})$.
- Average Treatment Effect: $E(Y_{1i} - Y_{0i} \mid D_i = 1)$

14.4.2. The Prima Facie Contrast

We compare mean outcomes in the treated and control groups:

$$\underbrace{E\big(Y_i \mid D_i = 1\big) \; - \; E\big(Y_i \mid D_i = 0\big)}_{\text{Prima-facie contrast}}$$

Using $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$:

If $D_i = 1$ then $Y_i = Y_{1i}$, so:

$$E(Y_i \mid D_i = 1) = E(Y_{1i} \mid D_i = 1)$$

If $D_i = 0$ then $Y_i = Y_{0i}$, so:

$$E(Y_i \mid D_i = 0) = E(Y_{0i} \mid D_i = 0).$$

Thus the contrast is:

$$E(Y_{1i} \mid D_i = 1) \; - \; E(Y_{0i} \mid D_i = 0)$$

Then, we can add and subtract the same term to expose the treatment-effect term: add and subtract $E(Y_{0i} \mid D_i = 1)$:

$$
\begin{aligned}
E(Y_{1i} \mid D_i = 1) \; - \; E(Y_{0i} \mid D_i = 0) = {} & \big[E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 1)\big] \\
& + \; \big[E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)\big].
\end{aligned}
$$

We recognize:

First bracket

$$E(Y_{1i} - Y_{0i} \mid D_i = 1)$$

is the **Average Treatment Effect on the Treated (ATET)**.

Second bracket

$$E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)$$

is the selection bias: it measures how the treated group's untreated potential outcome differs from the control group's.

Hence

$$\underbrace{E(Y_i \mid D_i = 1) \; - \; E(Y_i \mid D_i = 0)}_{\text{Prima-facie contrast}} = \underbrace{E(Y_{1i} - Y_{0i} \mid D_i = 1)}_{\text{ATET}} \; + \; \underbrace{E(Y_{0i} \mid D_i = 1) \; - \; E(Y_{0i} \mid D_i = 0)}_{\text{Selection bias}}.$$

The Role of Randomization

If treatment is randomly assigned, then D_i is independent of (Y_{0i}, Y_{1i}). In particular

$$E(Y_{0i} \mid D_i = 1) \; = \; E(Y_{0i} \mid D_i = 0) \quad \Longrightarrow \quad \text{Selection bias} = 0$$

Therefore under randomization:

$$E(Y_i \mid D_i = 1) \; - \; E(Y_i \mid D_i = 0) \; = \; E(Y_{1i} - Y_{0i} \mid D_i = 1) \; = \; \text{ATET}$$

And furthermore, because of $D_i \perp (Y_{0i}, Y_{1i})$, one usually also has

$$E(Y_{1i} - Y_{0i} \mid D_i = 1) = E(Y_{1i} - Y_{0i})$$

so the prima-facie contrast recovers the overall ATE as well.

14.5. Challenges of Social Experiments

Randomization bias:

- Samples individuals are different from overall population because of randomization
- Not everybody is willing to participate in experiment (e.g. to test new drug)

Attrition bias:

- Some people drop out of experiment/data collection
- Often selective attrition, due to e.g. unfavorable treatment

External validity:

- Do participants behave differently, because they are aware that they are in an experiment?

Substitution bias:

- Can control group get treatment elsewhere?

15. Instrument Variables

Contents

15.1. Motivation, Omitted Variable

For illustration, let us consider a simple model:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e$$

where e is the error term. Unfortunately, it now becomes impossible to observe the variable ability for an adult population, and we cannot use the IQ as a proxy variable. Hence, we are forced to include abil in the error term:

$$u = e + \beta_2 abil$$

We thus have the newly defined model:

$$log(wage) = \beta_0 + \beta_1 \cdot educ + u$$

and this model would give us biased estimates, because $E\left(u|educ\right) \neq 0$. It turns out that we can still use this model as the basis for estimation, provided we can find an instrumental variable for educ. Let us suppose that we have:

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

Where we think that x and u are uncorrelated, i.e. $Cov\left(u, x\right) = 0$. The method of instrumental variables works whether x and u are correlated or not, but for efficiency reasons we will see later, OLS should be used if x is uncorrelated with u. Now, let us suppose that we have a new observable variable z, where:

$$Cov\left(u, z\right) = 0 Cov\left(z, x\right) \neq 0$$

Then, we call z an instrumental variable for x, or sometimes simply an instrument for x. In the context of omitted variables, instrument exogeneity means that z should have no partial effect on y (after x and omitted variables have been controlled for), and z should be uncorrelated with the omitted variables. If $Cov\left(x, z\right) \neq 0$, we speak of instrumental relevance, and contrary to the other measure, we can test it in the population we have. The easiest way to do this is to estimate a simple regression between x and z. In the population, we have:

$$x = \pi_0 + \pi_1 z + v$$

If x and z are correlated, this means that we should be able to reject the null hypothesis $H_0 : \pi_1 = 0$ against the two-sided alternative at a sufficiently small α(5. In wage equations, labor economists have used family background variables as IVs for education. For example, mother's education (*motheduc*) is positively correlated with child's education, as can be seen by collecting a sample of data on working people and running a simple regression of educ on motheduc. The problem here is that *mothereduc* might be correlated with the abil, thus $Cov(motheredu\ (z),\ u)\ \neq 0)$.

There is a final point worth emphasizing before we turn to the mechanics of IV estimation. It is important to take note of the sign (and even magnitude) of $\hat{\pi}_1$ and not just its statistical significance. We should be able to interpret meaningfully the magnitude and direction of $\hat{\pi}_1$ in our argument.

15.2. Reverse Causality

Reverse causality occurs when, instead of X causing Y, it is really Y that causes changes in X. In observational data this can masquerade as a spurious "effect" of X on Y when in fact the true arrow of causation runs the other way.

Why it's a problem?

1. Biased Estimates

If we regress Y on X but really Y drives X, our estimated coefficient will capture that backward link. We might wrongly conclude that increasing X would change Y, when in fact it's changes in Y that lead people (or firms, or markets) to adjust X.

2. Policy Mistakes

Policies targeted at shifting X (the presumed "cause") will fail if the true causal lever is actually Y. We'll spend resources pushing on the wrong variable.

Classic Examples

Health and Income

We observe that richer people live longer. Is it that higher income causes better health, or that healthier people are more likely to earn higher income? If reverse causality holds, promoting income won't improve health as much as we expect.

Advertising and Sales

Companies see that high sales volumes coincide with large ad budgets. But is more advertising driving sales, or are firms ramping up ads in response to high-demand periods? Misreading the direction can lead to ineffective marketing strategies.

Instrument variables (IV) can help us to counter this issue!

Indeed, when Y partly causes X, this necessarily implies $E[u \mid X] \neq 0$!!!

15.3. IV Estimation

15.3.1. Assumptions

The validity of the derivation we derive in this sub-section hinges upon two key assumptions:

Instrumental Relevance $Cov(z, x_1) \neq 0$ This assumption is testable because if we run the following model:

$$x_1 = \pi_0 + \pi_1 x_1 + \varepsilon$$

We can also run the following hypothesis test:

$$\begin{cases} H_0 : \pi_1 = 0 \\ H_1 : \pi_1 \neq 0 \end{cases}$$

For our OLS-IV enhanced model to be valid, this needs to hold.

Exogeneity of the IV The second assumption the IV must fulfill is $E[u \mid z] = 0$. This assumption, as for the OVB, cannot be tested. Validity must be argued. We may assume that exogeneity is only valid if certain other exogenous variables are controlled for! It is also important to avoid $E(z \mid \varepsilon) \neq 0$, otherwise we carry the bias with us!

15.3.2. Derivation of the Estimates

The Two Stage Least Squares (2SLS) We consider the structural model

$$Y_i = \beta\, X_i + u_i,$$

where X_i is endogenous (i.e. $\mathrm{Cov}(X_i, u_i) \neq 0$). We have an instrument Z_i satisfying:

1. **Relevance**: $\mathrm{Cov}(Z_i, X_i) \neq 0$
2. **Exclusion**: $\mathrm{Cov}(Z_i, u_i) = 0$

1. First Stage

We project the endogenous regressor X_i onto the instrument(s) Z_i:

$$X_i = \pi_0 + \pi_1\, Z_i + \varepsilon_i, \qquad E[\varepsilon_i \mid Z_i] = 0.$$

1. Estimate the first-stage regression by OLS to obtain $\widehat{\pi}_0, \widehat{\pi}_1$.
2. Compute the fitted (predicted) values

   $$\widehat{X}_i = \widehat{\pi}_0 + \widehat{\pi}_1\, Z_i.$$

   By construction, $\widehat{X}_i$ is uncorrelated with the original error u_i.

---

2. Second Stage

We replace X_i by its predicted value $\widehat{X}_i$ in the structural equation:

$$Y_i = \beta\, \widehat{X}_i + v_i.$$

1. Run OLS of Y_i on $\widehat{X}_i$.
2. The resulting slope coefficient $\widehat{\beta}_{\mathrm{2SLS}}$ is

   $$\widehat{\beta}_{\mathrm{2SLS}} = \frac{\sum_i \widehat{X}_i\, Y_i}{\sum_i \widehat{X}_i^2} \xrightarrow{p} \beta.$$

3. Intuition

- The first stage "filters out" the part of X that is correlated with u by using only variation in X driven by Z.
- The second stage regresses Y on this "clean" variation $\widehat{X}$, yielding a consistent estimator of the causal effect β. (We cannot use the SE since they negelec the fact that X is indeed estimated by $\hat{X}$).

4. Summary of Assumptions

Requirement	Mathematical statement	Testable?
Relevance	$\pi_1 \neq 0 \iff \text{Cov}(Z, X) \neq 0$	Yes (first-stage F-test)
Exclusion	$\text{Cov}(Z, u) = 0$	No (must be argued)

When both hold, 2SLS consistently recovers the structural parameter β even in the presence of endogeneity.

15.4. Weak Instrument

There is also another method to derive the slope parameter β_1 when we use an IV.

Model:

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

We want:

$$\text{Cov}(z, u) = 0$$

So we consider:

$$\text{Cov}\left(z, y - \beta_0 - \beta_1 x_1\right) = 0$$

This means:

$$\text{Cov}(z, y - \beta_0 - \beta_1 x_1) = \text{Cov}(z, y) - \text{Cov}(z, \beta_0) - \text{Cov}(z, \beta_1 x_1)$$

Expanding:

$$\mathbb{E}\left[zy - \beta_1 z x_1 - \beta_2 z x_2\right] = \beta_1 \mathbb{E}[x_1]\mathbb{E}[z] + \beta_2 \mathbb{E}[x_2]\mathbb{E}[z] + \mathbb{E}[z]\mathbb{E}[y] \text{Cov}(z, y - \beta_0 - \beta_1 x_1) = \text{Cov}(z, y) - \beta_1 \text{Cov}(z, x_1)$$

Then:

$$\implies \hat{\beta}_1 = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x_1)}$$

Then, we have:

$$\text{plim}(\hat{\beta}_1) = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x_1)} = \frac{\text{Cov}(z, \beta_0 + \beta_1 x_1 + v)}{\text{Cov}(z, x_1)} = \frac{\beta_1 \text{Cov}(z, x_1) + \text{Cov}(z, \nu)}{\text{Cov}(z, x_1)} = \beta_1 + \frac{\text{Cov}(z, \nu)}{\text{Cov}(z, x_1)}$$

We see that the weaker the IV relevance, the more severe a violation of the exogeneity assumption becomes!

15.5. Efficiency of the IV-Estimator

The IV estimator is always at best as good as the OLS but no better because:

$$\text{Var}(\hat{\beta}_1^{IV}) = \frac{\sigma^2}{SST_{x_1} R^2_{(x_1,z)}}$$

where $R^2_{(x_1,z)} \in (0,1)$ is the R^2 score of the first stage OLS!

16. Proxy Variables

Content

- Proxy Variables

Materials

- Slides L16

Proxy Variables

In econometrics, a **proxy variable** is used when a relevant explanatory variable is unobservable or unavailable in the dataset. A proxy is a substitute variable that is correlated with the unobserved factor and helps reduce omitted variable bias.

16.1. Definition

We let the true model be:

$$y_i = \beta_0 + \beta_1 x_i^* + u_i$$

where x_i^* is unobserved. If we observe a proxy variable z_i that is linearly related to x_i^*, such as:

$$x_i^* = \delta_0 + \delta_1 z_i + v_i$$

and if z_i is correlated with x_i^* but uncorrelated with the error term u_i, then we can substitute x_i^* with z_i in the regression.

16.2. Conditions for a Valid Proxy

A valid proxy z_i must satisfy:

1. **Relevance**: $\text{Cov}(z_i, x_i^*) \neq 0$
2. **Exogeneity**: $\text{Cov}(z_i, u_i) = 0$

16.3. Example

We have a regression where ability affects wage, but ability is unobserved. Instead, we observe a proxy: IQ.

So the true model (not directly estimable) is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^* + u \quad \text{where } x_2^* = \text{ability (unobserved)}$$

But since we can't observe x_2ˆ*, we use its proxy:

$$x_2^* = \delta_0 + \delta_2 x_2 + \nu_2 \quad \text{where } x_2 = \text{IQ}$$

Substitute into the regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_2 x_2 + \nu_2) + u$$

Group terms to get an estimable model:

$$y = \underbrace{(\beta_0 + \beta_2\delta_0)}_{\alpha_0} + \underbrace{\beta_1}_{\alpha_1} x_1 + \underbrace{\beta_2\delta_2}_{\alpha_2} x_2 + \underbrace{(u + \beta_2\nu_2)}_{e}$$

So the error term in the estimable model is:

$$e = u + \beta_2\nu_2$$

In this scenario, we can only estimate β_1 consistently, but that's all we care about! Furthermore, we see that if $Cov(u, x_1) \neq 0$ (the variable we are interested in is correlated to the source of the OVB) we will have:

$$x_2^* = \delta_0 + \underbrace{\delta_1 x_1}_{\neq 0} + \delta_2 x_2 + \nu_2$$

and we can no longer reliably estimate β_1. The notebook n°17 displays this result in an elegant way with a simulation. The more x_1 is correlated to the missing variable, the more biased β_1 becomes:

$$(\beta_1 + \delta_1) \uparrow$$

as the correlation increases...

16.4. Proxy Versus IV

	Proxy Variable	**Instrumental Variable (IV)**
Purpose	Replace a **missing or unobserved regressor**	Correct for **endogeneity** of an observed regressor
Target variable	Proxy is a substitute for an **unobserved variable**	Instrument helps identify an **observed but endogenous variable**
Estimation	Proxy is **directly included** in the regression	IV is used to construct a predicted regressor (**2SLS**)
Relevance condition	$\text{Cov}(z, x^*) \neq 0$	$\text{Cov}(z, x) \neq 0$
Exogeneity condition	$\text{Cov}(z, u) = 0$	$\text{Cov}(z, u) = 0$
Example use case	Unobserved **ability** → use **IQ score** as proxy	Endogenous **education** → use **distance to college** as IV

17. Causal Inference: Difference-in-Differences

Contents

- Natural Experiments
- Difference-in-Differences

Materials

- Slides L17

17.1. Natural Experiments

Difference-in-Differences (DiD) is a cornerstone methodology in causal inference, particularly well-suited to the analysis of natural experiments—situations in which external factors or policy changes affect some units (the treatment group) but not others (the control group), and crucially, this allocation is plausibly exogenous to the outcome of interest. In such settings, DiD exploits temporal variation in treatment exposure, comparing the evolution of outcomes in treated versus untreated units before and after the intervention. This approach leverages the key identifying assumption of parallel trends—that in the absence of treatment, the difference between the two groups would have remained constant over time. When this assumption holds, DiD can provide a credible estimate of the treatment effect despite the lack of randomization, making it particularly powerful in real-world policy evaluations where controlled experiments are infeasible. Thus, natural experiments provide the quasi-experimental variation that DiD capitalizes on to uncover causal relationships.

17.2. Difference-in-Differences

17.2.1. Derivation

As mentioned in the previous paragraph, the difference in difference approach exploits the temporality of a data set. Therefore, we typically look at **pooled cross-sectional data**. Then, we compute the average treatment effect on the treated (ATET) as:

$$\text{ATET} = [E(Y_i^a \mid D_i = 1) - E(Y_i^a \mid D_i = 0)] - \left[E(Y_i^b \mid D_i = 1) - E(Y_i^b \mid D_i = 0)\right]$$

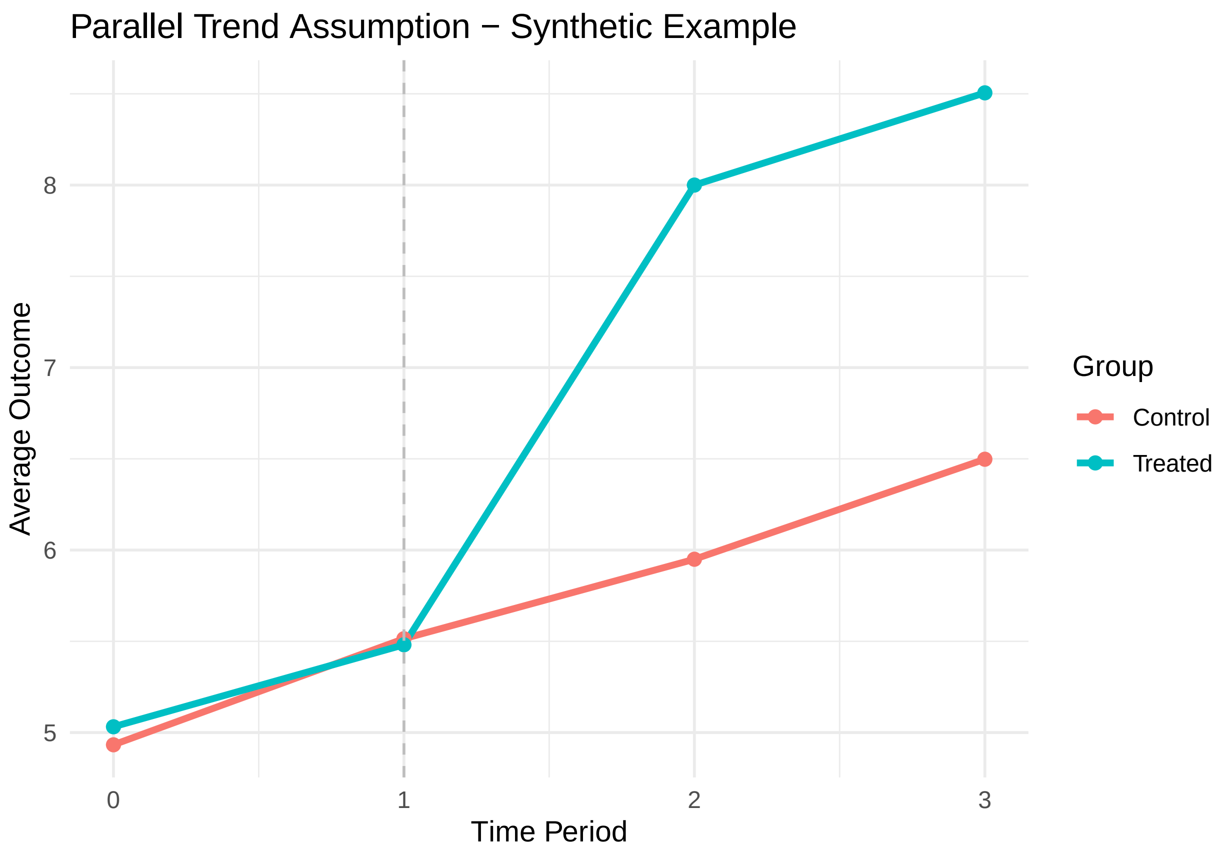
where:

- Y_i^b is the outcome **before** the realization of the event, shock.
- Y_i^a is the outcome **after** the realization of the event.

17.2.2. Assumption for the DiD

For the DiD to be accepted as a valid measure of the ATET, we must respect one key assumption, **the parallel trend assumption**. In the absence of the event (e.g. treatment), the outcomes of the affected and control group would have evolved in a parallel fashion.

We can generate some synthetic data for illustration purpose.



In this case, we see that before the treatment, both groups were evolving according to the same trend. Then, once the treatment started in period time $t = 1$, we see that the treated group diverged from the control group over the period $t \in (1, 2)$.

17.2.3. Specification in the OLS with DiD

When we use the DiD technique, we would typically specify our linear model as (where $A_i = 1$ if the observation is taken after the event, shock, or treatment and $D_i = 1$, if i received treatment):

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 A_i + \beta_3 D_i A_i + u_i$$

And then ask ourselves, which parameter measures the effect of the treatment? In this specification, we have:

- β_0: **baseline outcome** in both groups before the treatment.
- β_1: **difference in average outcome between the groups before treatment**. This parameter typically capture the selection bias we might face in our data…
- β_2: **temporal change** displays how the average outcome evolved (in both groups) after the event/treatment. We note that this is the complete effect for the control group but not for the treated group.
- β_3: **DiD estimate**, this coefficient actually measures the difference in average output, which is due to the treatment. This is the additional change in outcome for the treated group after the treatment, above and beyond any baseline group differences and common time trends.

17.3. Case Study

Now, let us make a quick study inspired from assignment 6. We have the following cleaned data set `df_players`:

```
## # A tibble: 3 x 5
##      ID  year  post treated n_players
##   <dbl> <dbl> <dbl>   <dbl>     <dbl>
## 1     1  1991     0       0      6.04
## 2     2  1991     0       1      8.81
## 3     3  1991     0       0      8.03
```

where `ID` identifies a unique town, the variable `post` indicates the observations made after the treatment, the variable `treated` indicates which town received a financial subsidy for building a tennis court (the treatment), and `n_players` measures the average number of tennis players in a town.

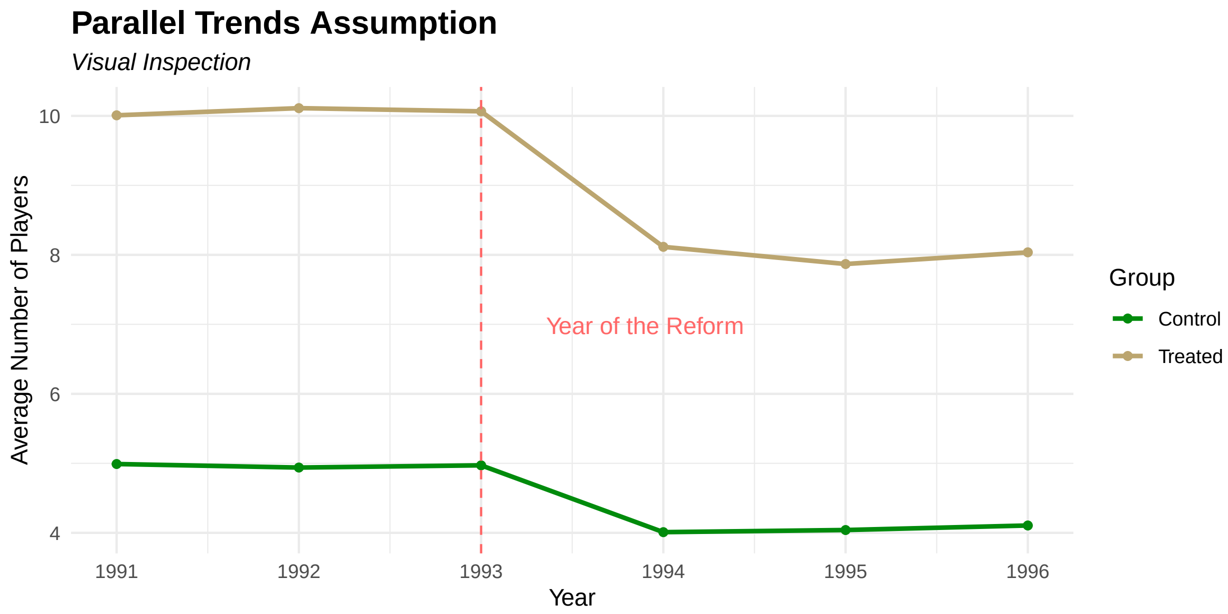
First, we can check the parallel trend assumption:

```r
# first, we should aggregate the data to plot them:
df_avg <- df_2 |>
  group_by(year, treated) |>
  summarise(mean_players = mean(n_players), .groups = "drop")

head(df_avg, n = 3)
```

```
## # A tibble: 3 x 3
##    year treated mean_players
##   <dbl>   <dbl>        <dbl>
## 1  1991       0         4.99
## 2  1991       1        10.0
## 3  1992       0         4.94
```

Then, we plot the data:

```r
ggplot(df_avg, aes(x = year, y = mean_players, color = factor(treated))) +
  geom_line(linewidth = 1) +
  geom_point() +
  geom_vline(xintercept = 1993, linetype = "dashed", color = "#FF6767") +
  annotate("text", x = 1993.9, y = max(df_avg$mean_players)/1.5,
           label = "Year of the Reform", angle = 0, vjust = -0.5, size = 4,
           color = "#FF6767") +
  scale_color_manual(values = c("#008B0C", "#BBA56F"),
                     labels = c("Control", "Treated")) +
  labs(title = "Parallel Trends Assumption",
       subtitle = "Visual Inspection",
       x = "Year", y = "Average Number of Players",
       color = "Group") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 15, face = "bold"),
    plot.subtitle = element_text(face = "italic")
  )
```



and we conclude that there is a reasonable chance for the parallel trend assumption to hold. Once this is done, we estimate our model as specified above:

```
# the model
model_did <- lm(n_players ~ post * treated, data = df_2)
summary(model_did)
```

```
##
## Call:
## lm(formula = n_players ~ post * treated, data = df_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1929 -1.0402  0.0214  1.0239  7.9924
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.96487    0.04520  109.85   <2e-16 ***
## post         -0.68303    0.05535  -12.34   <2e-16 ***
## treated       5.09500    0.06388   79.76   <2e-16 ***
## post:treated -0.85530    0.07823  -10.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.79 on 9416 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.6291
## F-statistic:  5327 on 3 and 9416 DF,  p-value: < 2.2e-16
```

Interpretation `(Intercept) = 4.96487`

This is β_0, the mean outcome for the control group before the treatment (i.e., when post $= 0$ and treated $= 0$). On average, individuals in the control group had a baseline outcome of 4.96 before the treatment period.

`post = -0.68303`

This is β_1: Change over time in the control group, from before to after the treatment. In the control group, the outcome decreased by 0.68 after the treatment period — this captures time trends not caused by treatment. Therefore, we must have faced an adverse shock!

```
treated = 5.09500
```

This is β_2: Difference between treated and control groups before treatment. Before the treatment, the treated group had outcomes that were 5.10 units higher than the control group on average. This reflects pre-existing group differences (**selection effect**).

```
post:treated = -0.85530
```

This is β_3: **Difference-in-Differences (DiD) estimate** — the causal effect of the treatment. After adjusting for baseline differences and time trends, the treatment caused a decrease of 0.86 units in the outcome for the treated group. (The reform was not successful. . .).

END

Yvan Richard
University of St. Gallen
June 2025

Phone: +41 79 318 72 85
Email: yvan.richard2004@gmail.com