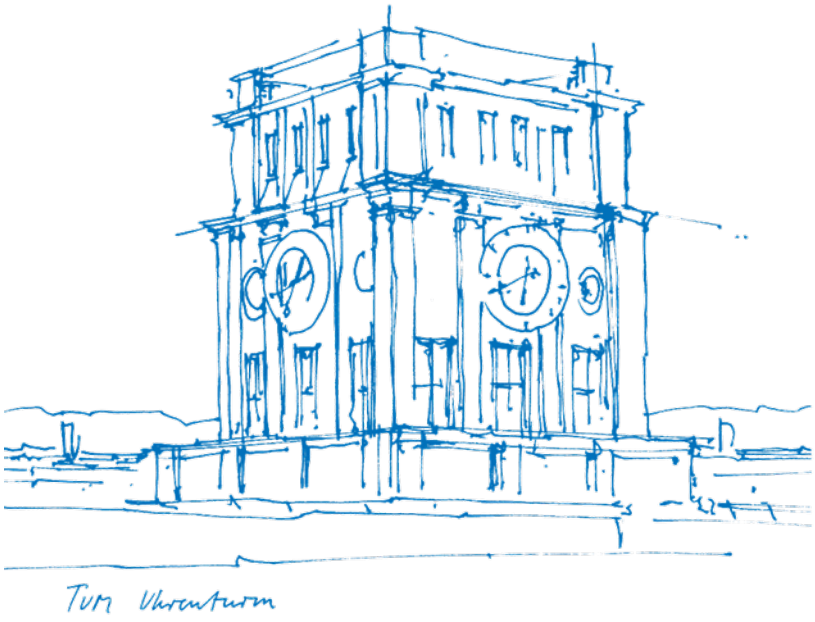
Applied Econometrics
TUM School of Management
Technical University of Munich

Empirical Research Methods - Lecture 3

Prof. Dr. Helmut Farbmacher
Technical University of Munich
TUM School of Management
Munich, May 7, 2024



Outline of today's lecture

Multivariate Random Variables

Introduction

Introduction

In the last lecture we examined random variables separately (univariate random variables).

Many economic applications, however, involve more than one random variable.

Example:

What is the relationship between education and earnings?

What is the relationship between market share and advertising?

How is ownership concentration related to firm performance?

Introduction

Covariance between X and Y describes how variables co-vary or co-move:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \dots \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \end{aligned}$$

Intuition: If X is larger than its mean, than Y tends to be below its mean $\left(\mathrm{Cov}(\cdot) < 0\right)$ or above its mean $\left(\mathrm{Cov}(\cdot) > 0\right)$.

If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$

Note that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)\ \left[= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \right]$.

Introduction

The **correlation** coefficient is defined as

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}_X \cdot \mathrm{SD}_Y}$$

and often denoted by ρ_{XY}.

The correlation is a standardized covariance, i.e. $-1 \leq \rho_{XY} \leq 1$.

Easy to interprete the intensity of the dependence b/w two random variables: the closer a correlation is to 1 (-1), the stronger the positive (negative) dependence b/w X and Y is.

Perfect positive correlation: $\mathrm{Corr}(X, Y) = 1$

Perfect negative correlation: $\mathrm{Corr}(X, Y) = -1$

X and Y are uncorrelated: $\mathrm{Corr}(X, Y) = 0$

Introduction

The expected value of the sum of two random variables is the sum of the expected values:

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

The variance of the sum of two random variables is:

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\,\mathrm{Cov}(X, Y)$$

For **independent** variables, X and Y, we have:

$$\mathrm{Cov}(X, Y) = 0$$

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$$

Multivariate Random Variables

Linear Functions of Random Variables

Linear Functions of Random Variables

Deterministic functions:

- One variable is fully determined by the value of (an)other variable(s).
- A certain input X always gives us the same output Y.

Stochastic functions:

- There is a link between two variables but also some inherent randomness.
- A certain input X tends to result in a lower/larger output Y.

Linear Functions of Random Variables

Sometimes, two variables (X and Y) are related by a (**deterministic**) linear function.

Consider the linear function $Y = a + bX$

For example, a worker earns 2000 € plus 20 € sales commission for every sold item. Overall earnings Y can then be written as a function of sales X:

$$Y = 2000 + 20X$$

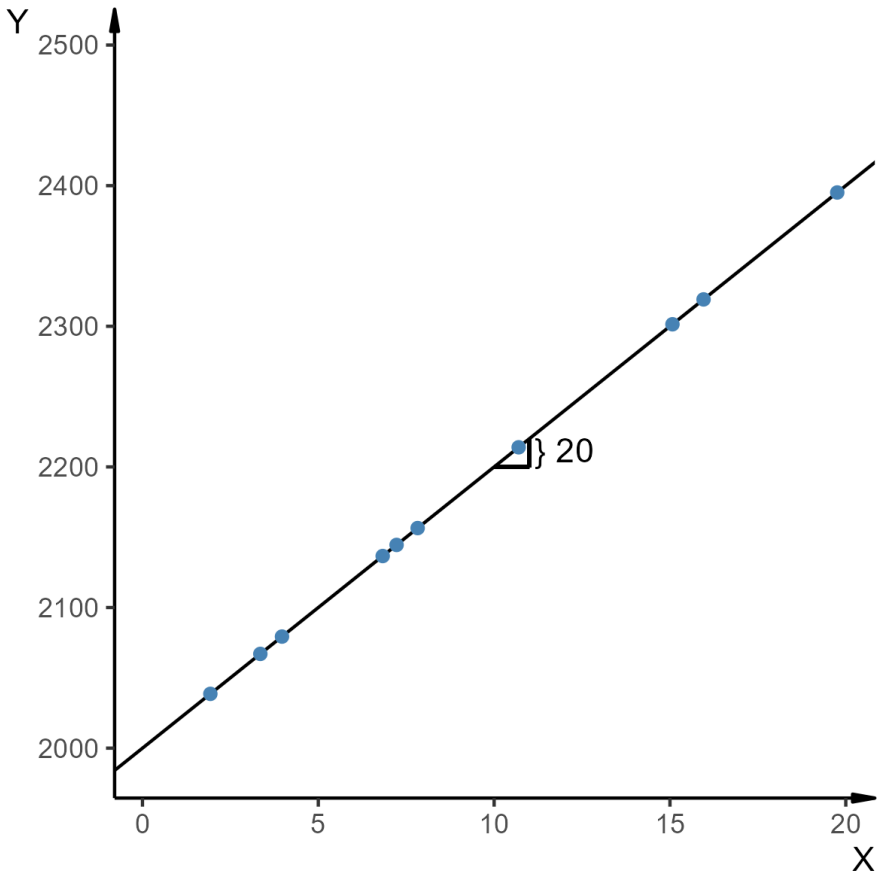
Let X be a random variable with $\mathrm{E}(X) = \mu_X$, and $\mathrm{Var}(X) = \sigma_X^2$.

What is $E(Y)$ and $\mathrm{Var}(Y)$?

$$\begin{aligned} \mathrm{E}(Y) &= a + b\mu_X \\ \mathrm{Var}(Y) &= b^2\sigma_X^2 \end{aligned}$$

Linear Functions of Random Variables

$$Y = 2000 + 20X$$



Linear Functions of Random Variables

Far more interesting from an econometric point of view are **stochastic** relationships between (two) random variables.

For example, education & wages or advertisement & sales.
Clearly, these variables are related. However, the relation is not exact.

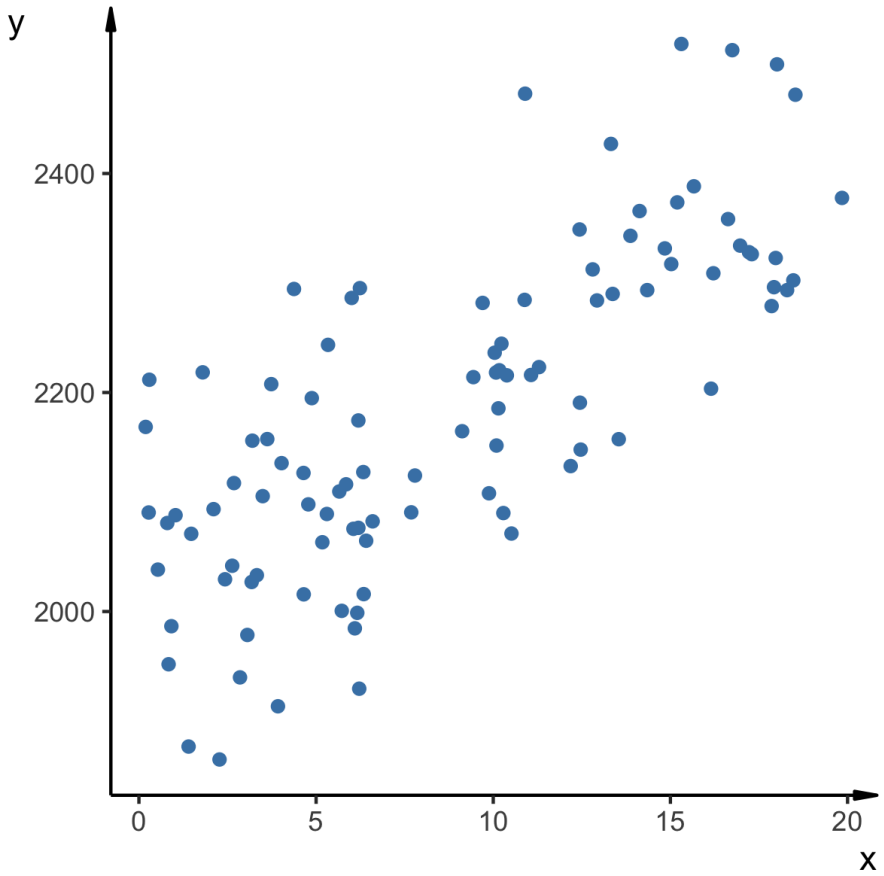
To account for the stochastic nature, we add an (unknown) error term u

$$y = 2000 + 20x + u$$

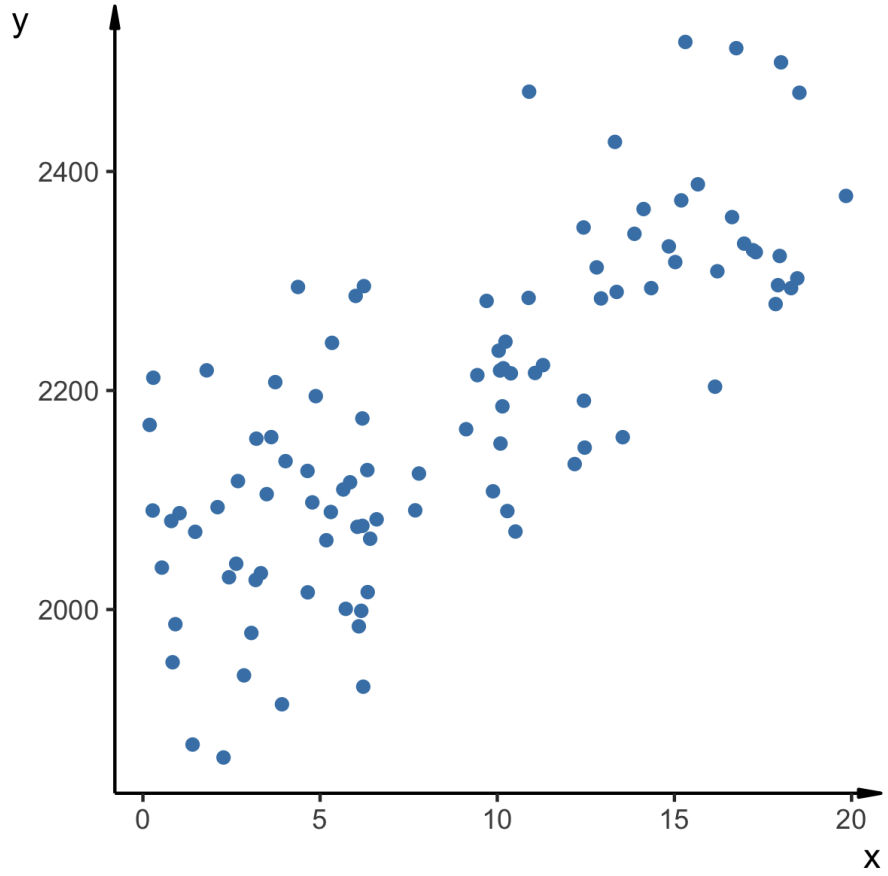
The error term contains several other factors that are relevant for y. For example, weather conditions if y measures sales of beer (& x measures e.g. advertisement).

Linear Functions of Random Variables

$$y = 2000 + 20x + u$$



Linear Functions of Random Variables



Scatterplots are a nice way to illustrate data, but not easy to *quantify* relationships.

Linear Functions of Random Variables

Let us first theoretically think about the population model that generated our data:

$$\begin{aligned} \mathrm{E}(y|x) &= \mathrm{E}(2000 + 20x + u|x) \\ &= 2000 + 20\underbrace{\mathrm{E}(x|x)}_{=x} + E(u|x) \end{aligned}$$

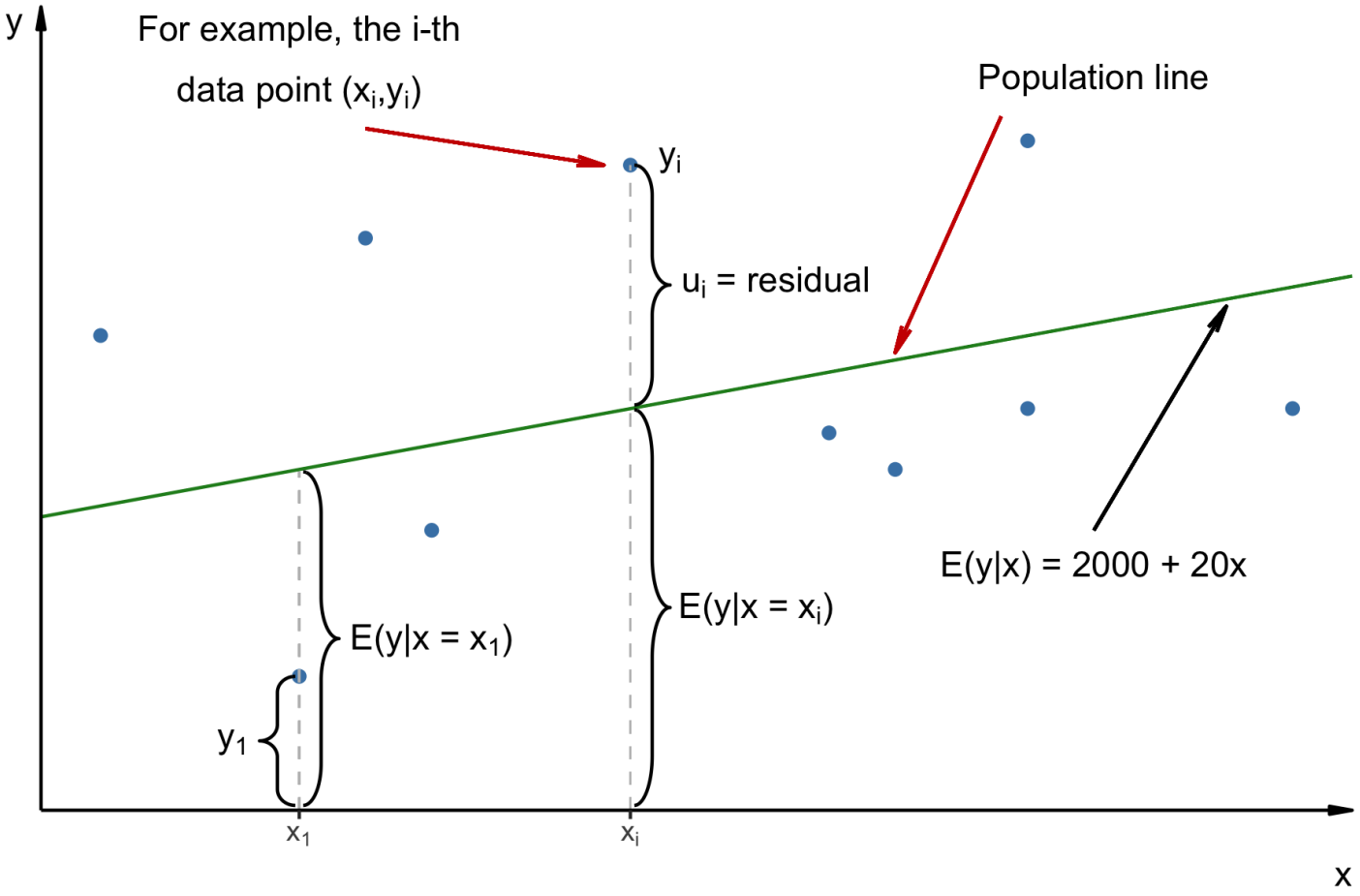
If we are willing to assume $\mathrm{E}(u|x) = 0$, then

$$\begin{aligned} \mathrm{E}(y|x) &= 2000 + 20x + \underbrace{E(u|x)}_{=0} \\ \mathrm{E}(y|x) &= 2000 + 20x \end{aligned}$$

Compare to the deterministic relation we discussed before:

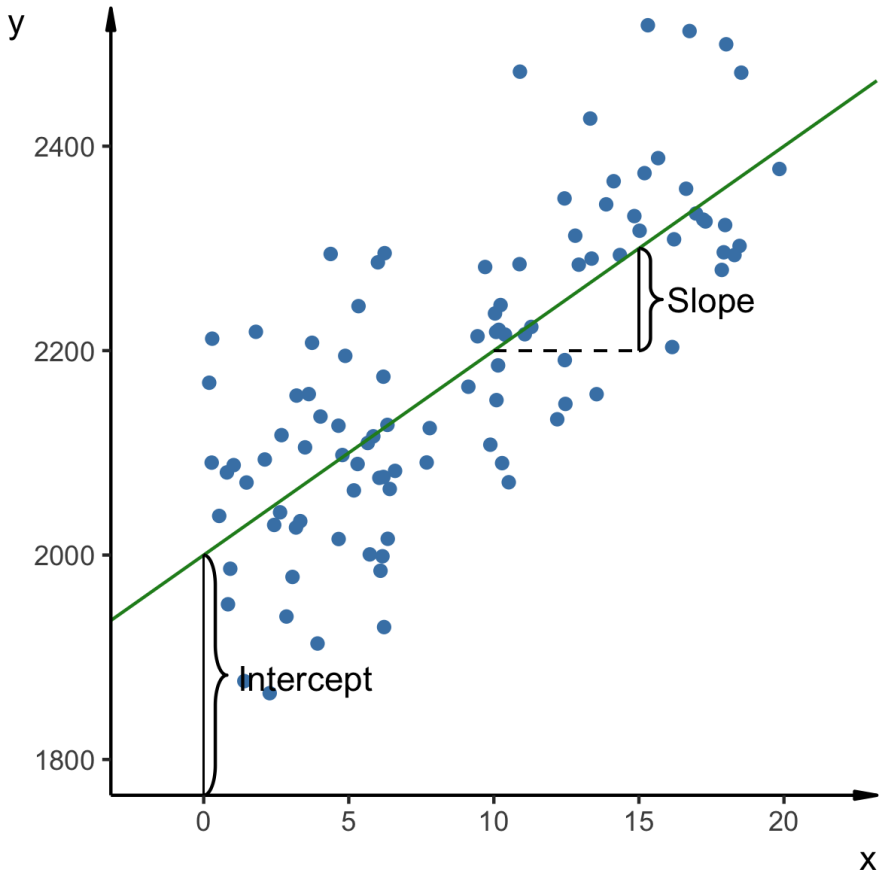
$$Y = 2000 + 20X$$

Linear Functions of Random Variables



Linear Functions of Random Variables

Green line: $\mathrm{E}(y|x) = 2000 + 20x$



Linear Functions of Random Variables

A linear line can be uniquely characterized by intercept and slope.

Note that we do **not** know the population parameters:

$$\mathrm{E}(y|x) = 2000 + 20x$$

where $\alpha = 2000$ (intercept) and $\beta = 20$ (slope).

Therefore, we have to estimate them using data:

$$\widehat{\mathrm{E}}(y|x) = \widehat{\alpha} + \widehat{\beta}x$$

Linear Bivariate Regression Model

Introduction

Introduction

We would like to summarize the information contained in the scatterplot.

This may be done by finding

- the best line in terms of fitting the data
- and then report the estimated **intercept** ($\widehat{\alpha}$) and **slope** ($\widehat{\beta}$) of this line.

Most common way to accomplish this is known as the **linear regression model**.

The **Ordinary Least Squares** (OLS) estimator is a common way to estimate linear regression models.

Introduction

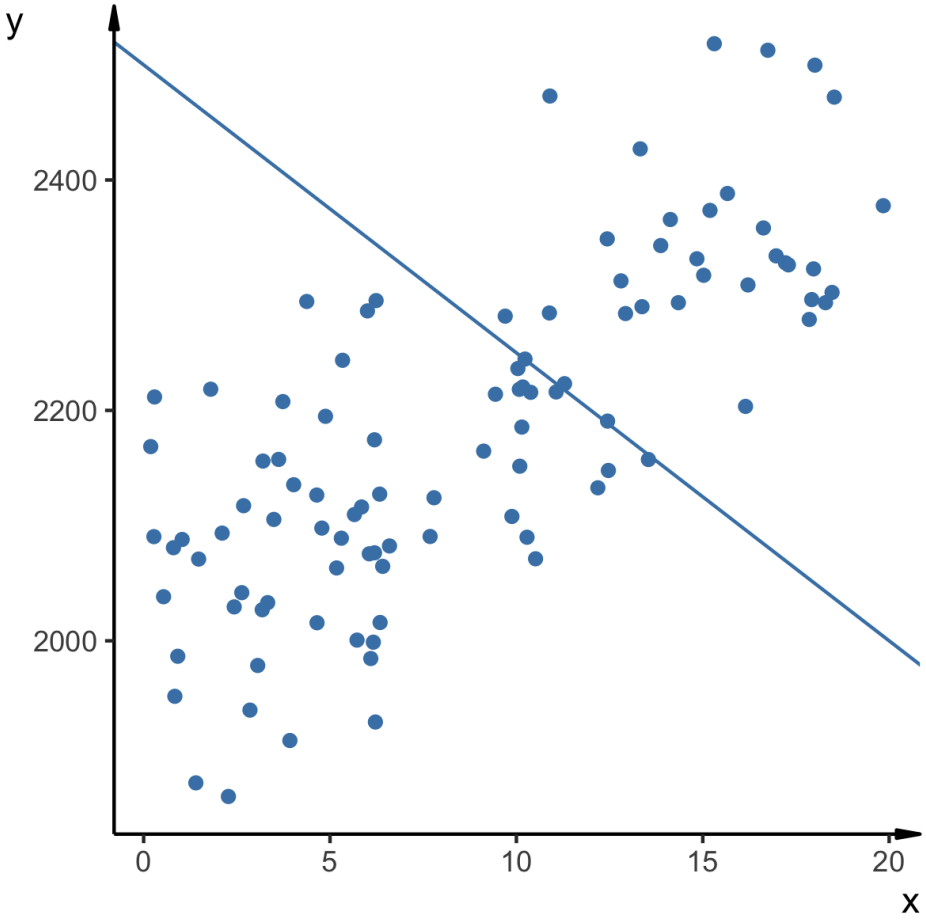
An **estimator** is a (clever) function of our data (here, y and x).

An **estimate** is a number computed from data using an estimator.

Although the estimator (e.g., OLS) will not change with the sample drawn, the estimates (e.g., of the intercept and the slope) will change everytime we draw a new sample from the population.

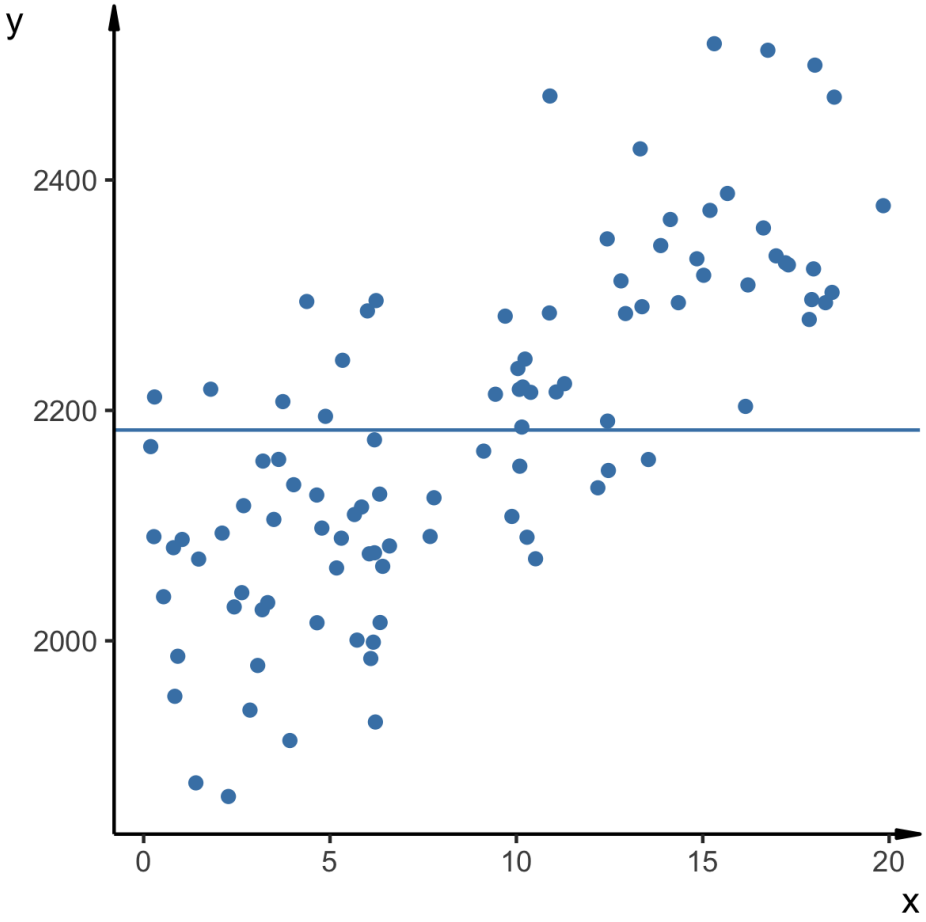
Therefore, estimates are random variables, which have a distribution (remember the previous lecture on the sample mean).

Introduction



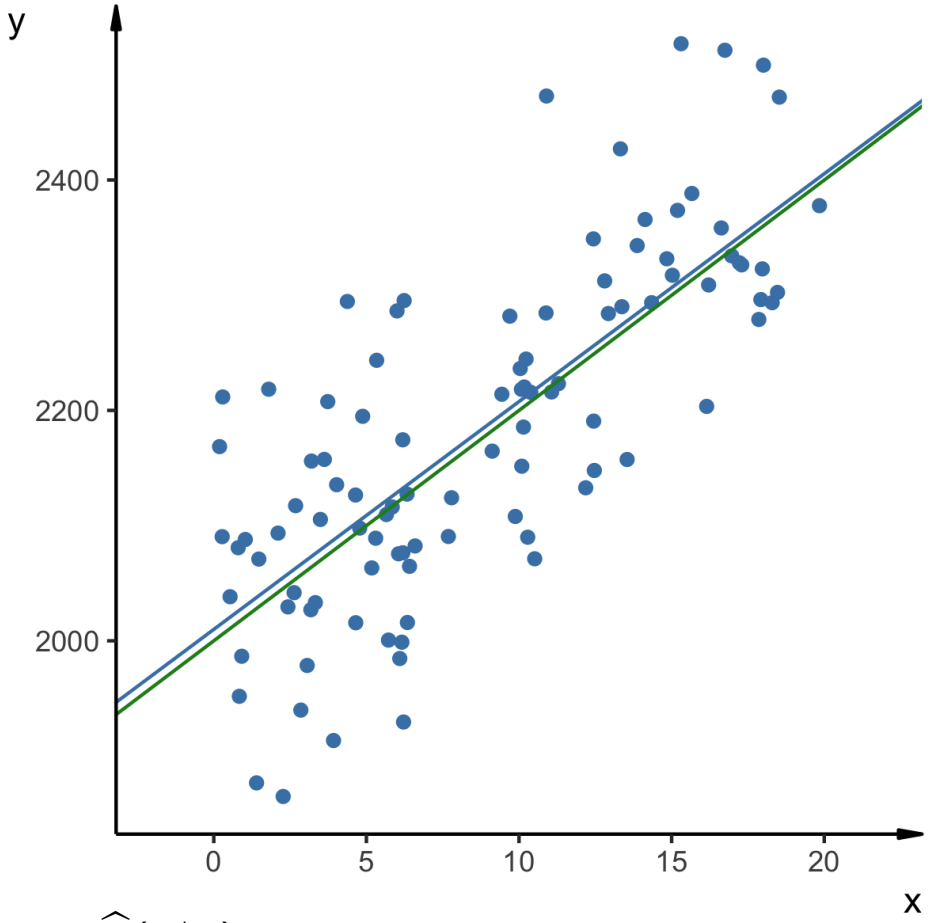
Not clever!

Introduction



Not clever!

Introduction



Blue line: $\widehat{\mathrm{E}}(y|x) = 2009.92 + 19.78x$ (OLS regression)

Green line: $\mathrm{E}(y|x) = 2000 + 20x$ (population model)

Linear Bivariate Regression Model

The Population Model

The Population Model

The linear **bivariate** regression model investigates the relationship between two variables y and x.

Assume, for every individual i our model defines the following relation:

$$y_i = \alpha + \beta x_i + u_i \, .$$

Outcome variable y (also called the dependent variable, regressand or left hand side variable).

A single **explanatory variable** x (also called independent variable, regressors or right hand side variable).

An **error term** u (that allows for other variables to affect y).

The Population Model

$$y_i = \alpha + \beta x_i + u_i$$

α (the intercept/constant) and β (the slope) are parameters (aka coefficients). Population parameters are generally **unknown**.

In most applications the focus is on β which quantifies the relation between y and x.

We can easily transform this model to get rid of the constant

$$\begin{aligned} y_i &= \alpha + \beta x_i + u_i && (1) \\ \bar{y} &= \alpha + \beta \bar{x} + \bar{u} && (2) \\ y_i - \bar{y} &= \beta \left(x_i - \bar{x}\right) + \left(u_i - \bar{u}\right) && (3) \end{aligned}$$

We (effectively) partialled out the constant: α and β are *numerically* the same in equation (1), (2) and (3).

The Population Model

$$y_i = \alpha + \beta x_i + u_i \tag{1}$$
$$\bar{y} = \alpha + \beta \bar{x} + \bar{u} \tag{2}$$
$$y_i - \bar{y} = \beta \left(x_i - \bar{x}\right) + \left(u_i - \bar{u}\right) \tag{3}$$

In the following we want to focus on β and therefore use equation (3).

Although there seems to be no constant in equation (3), we implicitly have one in this model as well.

To simplify notation, we still use y_i, x_i and u_i to denote $y_i - \bar{y}$, $x_i - \bar{x}$ and $u_i - \bar{u}$.

If you feel confused now, simply think about $\alpha = 0$ in equation (1).

Linear Bivariate Regression Model

Example: Private Returns to Education

Example: Private Returns to Education

$$y_i = \beta x_i + u_i$$

Private returns to education: y **income** and x **years of schooling**

Research question: How much more can you expect to earn with one additional year of schooling?

Of course, there may be other factors that influence the outcome of y (e.g. income also depends on job experience or soft skills).

In the above linear regression these other factors are summarized in the error term u.

Example: Private Returns to Education

Data info:

The Current Population Survey (*cps09mar*) contains information on employment, earnings, educational attainment, income etc. for 57.000 U.S. households (March 2009).

Variables:

earnings	total annual wage and salary earnings
education	years of education (based on highest degree)

Data can be downloaded here: `https://www.ssc.wisc.edu/~bhansen/econometrics/`.

Example: Private Returns to Education

$$lm(formula = earnings \sim education, data = cps09mar)$$

MODEL INFO:
Observations: 50742
Dependent Variable: earnings
Type: OLS linear regression

MODEL FIT:
$F(1,50740) = 8796.52$, $p = 0.00$
$R^2 = 0.15$
Adj. $R^2 = 0.15$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	-46755.01	1106.79	-42.24	0.00
education	7314.13	77.98	93.79	0.00

Example: Private Returns to Education

$lm(formula = earnings\_demeaned \sim education\_demeaned - 1, data = cps09mar)$

MODEL INFO:
Observations: 50742
Dependent Variable: earnings_demeaned
Type: OLS linear regression

MODEL FIT:
$F(1,50741) = 8796.69$, $p = 0.00$
$R^2 = 0.15$
Adj. $R^2 = 0.15$

Standard errors: OLS

	Est.	S.E.	t val.	p
education_demeaned	7314.13	77.98	93.79	0.00

Recommended reading

For next week please read chapter:

2.13 Ordinary least squares

in `https://mixtape.scunning.com/index.html`

and

2.3 Conditional Expectation

in `https://www.ssc.wisc.edu/~bhansen/econometrics/`

Contact

Helmut Farbmacher
office.econometrics@mgt.tum.de