Statistics 2

Introduction / Simple Linear Regression I: Estimation

Casper Albers & Jorge Tendeiro

Lecture 1, 2019 – 2020

university of groningen

Overview

- Organization of the course
  - Lecturer, practicals coordinator
  - Important dates
  - Read the course information PDF (Nestor)!
- Browsing through the course's contents
- Simple linear regression: Estimation
  - Regression line
  - Estimation of regression line
  - OLS method
  - Model
  - Estimate σ
  - Regression analysis vs correlation
  - Regression toward the mean

- All course material – lectures, slides, reader, practice exams, etc. – is protected under copyright regulations.
- Recording of course material is prohibited according to the university's studentenstatuut and will be reported to the Examination Committee.

Lecturer

Casper Albers

- **Lecturer PSBA2-07**
- Room H.231; tel. 38239
- c.j.albers@rug.nl

Jorge Tendeiro

- **Lecturer PSBE2-07**
- Room H.206; tel. 36953
- j.n.tendeiro@rug.nl

Practicals coordinator

Karin Siebenga

- **Coordinator**
- Room H.214; tel. 39358
- k.siebenga@rug.nl
- Contact Karin for *all* matters related to practicals management, attendance administration, Nestor enrollments, and so on.

Important dates

Deadline enroll practicals Wednesday 11 September, 17:00

1st partial exam Friday 8 November, 18:45-19:45

2nd partial exam Monday 20 January, 12:15-13:15

Resit exam Monday 6 April, 12:15-14:15

Read the course information PDF (Nestor)!

The course information PDF has all relevant information concerning the course setup, including:

- Literature.
- Software.
- Lectures.
- Practicals (enrollment, requirements, attendance).
- Retaking the course.
- Exams, exam inspection.
- ...

Important

You are expected to be aware of the entire course setup as laid out in the course information PDF in Nestor. You are responsible for not missing some relevant information therein.

Browsing through the course's contents

Statistical methods in a nutshell:

- Regression
  - Simple.
  - Multiple.
- Multivariate relationships.
- Model assumptions: Diagnostics and model validity.
- Code variables.
- ANOVA (Analysis of Variance)
  - One-way ANOVA.
  - Two-way ANOVA.
- Introduction to Bayesian statistics.
- The Replication crisis.

Browsing through the course's contents

- Contents are either new, or build upon material from Statistics I.
- The focus will lie on both
  - Theory: Understanding how and why the methods work.
  - Practice: Understanding how to use the methods.

It is assumed that you have *active knowledge* of the complete contents of Statistics Ia and Ib.

Refresh your knowledge as soon as possible, if necessary.

Overview of the course's contents

Lecture	Week	Literature	Content
0	36	4–7	Refresher Statistics I
1	37	9.1–9.4	Simple linear regression: Estimation
2	38	9.5, A1	Simple linear regression: Inference
3	39	9.6, 10	Model validity. Causality & Association
4	40	11.1–11.3	Multiple regression
5	41	11.4–11.5	Multiple regression: Interaction effects
6	42	11.6–11.7	M.R.: Partial correlation, standardized regression
7	43	-	Assumptions
8	46	12.1	Regression with Categorical Predictors
9	47	12.2	Multiple Comparisons and Contrasts
10	48	12.3	ANOVA, one-way
11	49	12.4	ANOVA, two-way
12	50	A2	Introduction to Bayesian statistics
13	51	A3	Good statistics, bad statistics
14	02	-	Overview

A1: Albers, Inference for Correlations.

A2: Kruschke & Liddell (2018)

A3: Simmons et al. (2011); John et al. (2012)

Please see the reader for detailed information on A1–A3.

Simple linear regression: Estimation

Literature for this lecture:
Chapter 9 (sections 9.1–9.4).

Simple linear regression (SLR)

Type of variables involved in simple linear regression:

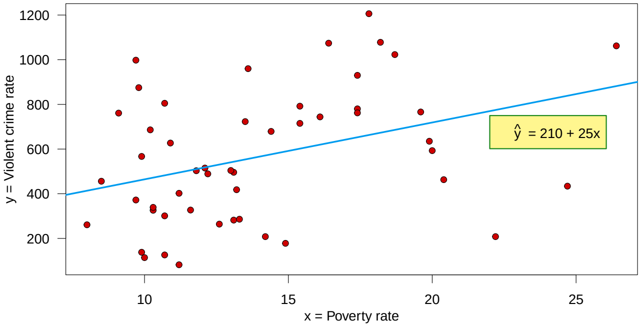
- One continuous predictor (independent, or x, variable).
- One continuous outcome (dependent, or y, variable).

Main aspects of regression analyses:

- Explore the existence of a linear relationship between predictor and outcome variables.
- Study this relationship (e.g., strength, direction).
- Predict values of the outcome variable from values of the predictor.

SLR: Crime data

- $x =$ Poverty rate; % population with income below the poverty level.
- $y =$ Violent crime rate $=$ number serious crimes per 100,000 people.
- $n = 50$ American states.



SLR: Equation

$$y = 210 + 25x = \alpha + \beta x$$

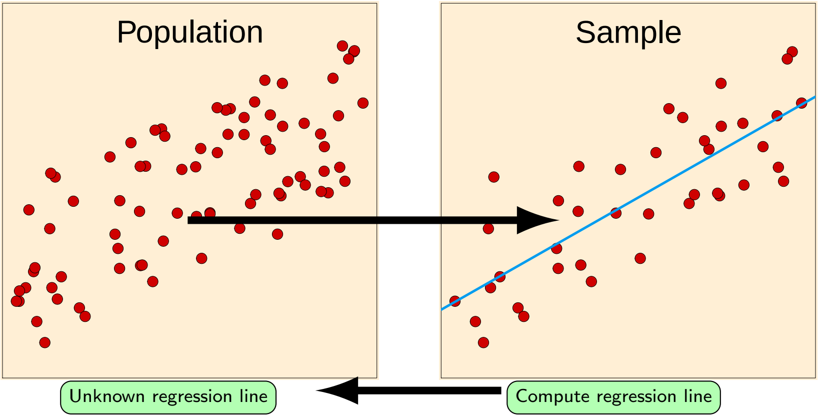
Interpreting the equation coefficients:

- α is the intercept:
  $\alpha = 210$ is the number of serious crime rates per 100,000 when x, the poverty rate, is 0.
- β is the slope:
  The number of serious crime rates per 100,000 increases by $\beta = 25$ when x, the poverty rate, increases by one unit (percent).

The sign of the slope β determines the direction of the regression line:

- $\beta > 0 \longrightarrow$ increasing line, i.e., positive relation between x and y.
- $\beta = 0 \longrightarrow$ horizontal line, i.e., no relation between x and y.
- $\beta < 0 \longrightarrow$ decreasing line, i.e., negative relation between x and y.

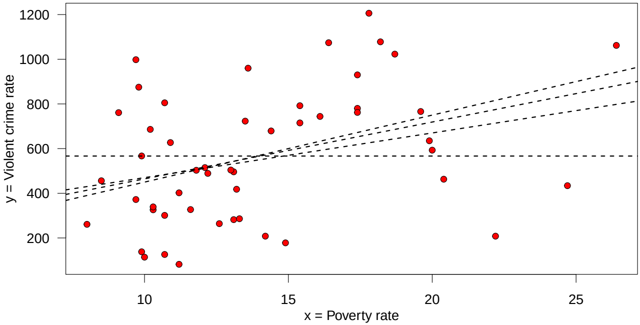
Use a fact about a sample to estimate the truth about the whole population.



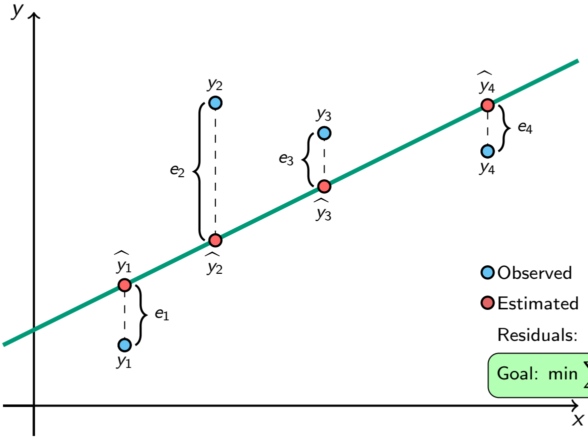
$$\underbrace{y = \alpha + \beta x}_{\text{Population}} \quad \longrightarrow \quad \underbrace{\widehat{y} = a + bx}_{\text{Sample}}$$

- a: Sample estimate of α.
- b: Sample estimate of β.

But how to compute a and b from the sample?



Ordinary least squares (OLS) method



Residuals: $e_i = y_i - \widehat{y_i}$

Goal: $\min \sum_i e_i^2$

SLR: OLS method

Find a, b that minimize the sum of squared distances between the observations and the regression line:

$$\min \sum_i e_i^2 = \min \sum_i (y_i - \widehat{y}_i)^2 = \min \sum_i [y_i - (a + bx_i)]^2.$$

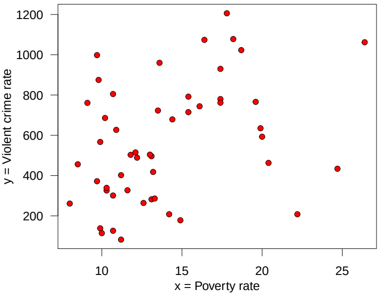
Mathematical solution:

$$\boxed{b = r_{xy}\frac{s_y}{s_x}} \qquad \boxed{a = \overline{y} - b\overline{x}}$$

where

- r_{xy} = sample correlation between x and y.
- s_x, s_y = sample standard deviation of x, y.
- $\overline{x}, \overline{y}$ = sample mean of x, y.

SLR: Crime data



Descriptive Statistics

	PovertyRate	ViolentCrime
Valid	50	50
Missing	0	0
Mean	14.016	566.660
Std. Deviation	4.287	295.877

Pearson Correlations

		PovertyRate	ViolentCrime
PovertyRate	Pearson's r	—	
	p-value	—	
ViolentCrime	Pearson's r	0.369	—
	p-value	0.008	—

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	2.5%	97.5%
1	(Intercept)	209.920	135.613		1.548	0.128	−62.748	482.588
	PovertyRate	25.452	9.260	0.369	2.749	0.008	6.833	44.072

$$b = r_{xy}\frac{s_y}{s_x} = .369 \times \frac{295.877}{4.287} = 25.5$$

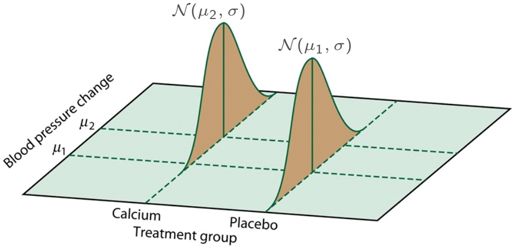
$$a = \overline{y} - b\overline{x} = 566.660 - 25.452 \times 14.016 = 209.9$$

SLR model – Motivation from t-test

Recall the two-sample t test:

- Two populations: $y_1 \sim \mathcal{N}(\mu_1, \sigma)$, $y_2 \sim \mathcal{N}(\mu_2, \sigma)$.
  Parameters μ_1, μ_2, and σ unknown. Same σ assumed.
- Take one sample from each population; sample sizes n_1 and n_2.

$$\boxed{\mathcal{H}_0 : \mu_1 = \mu_2 \text{ versus } \mathcal{H}_a : \mu_1 \neq \mu_2}$$

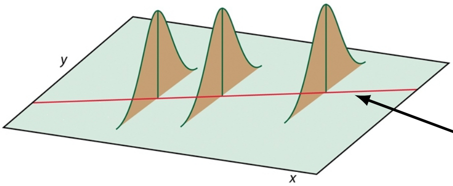


SLR model

The population regression equation is:

$$\boxed{E(y) = \alpha + \beta x}$$

- $E(y)$: Population mean y-score conditional on x.
- α: Population intercept, i.e., the mean value of y when $x = 0$.
- β: Population slope, i.e., the change rate of $E(y)$ when x increases 1 unit.



$E(y) = \alpha + \beta x$:

The SLR assumes a linear relationship between x and $E(y)$ in the population.

SLR model

Population regression equation:

$$\boxed{E(y) = \alpha + \beta x}$$

Assumptions:

- Given x, the y values are normally distributed.
- The spread of the y values is the same for conditional distributions (i.e., same σ).

So, individual y scores spread around the mean $E(y)$ according to the value of σ:

$$\boxed{y_i = \underbrace{\alpha + \beta x_i}_{E(y_i)} + \varepsilon_i} \longrightarrow \varepsilon_i \sim \mathcal{N}(0, \sigma)$$

(unrelated to x)

SLR model

Statistical model:

$$y_i = \underbrace{\alpha + \beta x_i}_{E(y_i)} + \varepsilon_i$$

Data = Model + Error

Model parameters:

- The intercept α.
- The slope β.
- The standard deviation of the residuals ε_i, σ.

We already know how to estimate α and β.
What about σ?

SLR: Estimate σ

Recall the formulas to estimate the population intercept α and the slope β:

$$\boxed{a = \overline{y} - b\overline{x}} \qquad \boxed{b = r_{xy}\frac{s_y}{s_x}}$$

Having a and b computed, the following estimate for σ^2 can be computed:

$$\boxed{s^2 = \frac{\sum_i e_i^2}{n-2} = \frac{\sum_i (y_i - \widehat{y_i})^2}{n-2}}$$

where $\widehat{y_i} = a + bx_i$.

SLR: Crime data

Model Summary

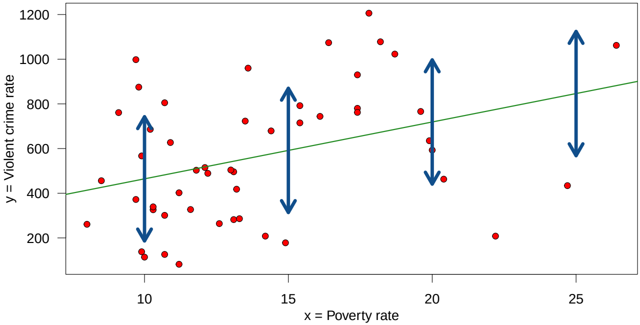
Model	R	R²	Adjusted R²	RMSE
1	$\boxed{0.369}$	$\boxed{0.136}$	$\boxed{0.118}$	$\boxed{277.876}$

$r(\text{Poverty}, \text{Violence})$; better, it is $r(y, \widehat{y})$

proportion of explained variance $\rightarrow$ measure of model fit

$s^2 = \frac{\sum_i e_i^2}{n-2} = 277.88^2$ $\rightarrow$ estimate of σ^2

SLR: Crime data



$s \simeq 280 \quad \longrightarrow \quad \text{length arrows} = 2s \simeq 560$

s is an estimate of the variability about the population regression line.

Regression analysis vs correlation

- SLR tries to model a linear relationship between x and y.
- Therefore, there is a strong connection between regression and correlation.
- Recall the formula of the simple regression slope, b:

$$b = r_{xy}\frac{s_y}{s_x} \iff r_{xy} = b\frac{s_x}{s_y}.$$

The correlation is a standardized slope:
When $s_x = s_y$ (e.g., when x and y are standardized) then $r_{xy} = b$.

The correlation is given by:

$$\boxed{r = \frac{cov(x,y)}{sd(x)sd(y)} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left[\sum_i (x_i - \overline{x})^2\right]\left[\sum_i (y_i - \overline{y})^2\right]}}}$$

Regression analysis vs correlation

$$r = \frac{cov(x,y)}{sd(x)sd(y)} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left[\sum_i (x_i - \overline{x})^2\right]\left[\sum_i (y_i - \overline{y})^2\right]}}$$

Properties of r:

- r is standardized: $-1 \leq r \leq 1$.
- r indicates the direction (sign of r) and strength (magnitude of r) of the linear x-y relationship:
  - ✓ $r = 1$: Perfect positive linear relationship;
  - ✓ $r = 0$: No linear relationship;
  - ✓ $r = -1$: Perfect negative linear relationship.
- $\text{sign}(r) = \text{sign}(b)$.
- Careful: r is sensitive to outliers.

Regression toward the mean

What happens when x increases by one SD?

- y at x:
$$\widehat{y}_x = a + bx.$$
- y at $(x + s_x)$:
$$\widehat{y}_{x+s_x} = a + b(x + s_x) = \widehat{y}_x + bs_x.$$

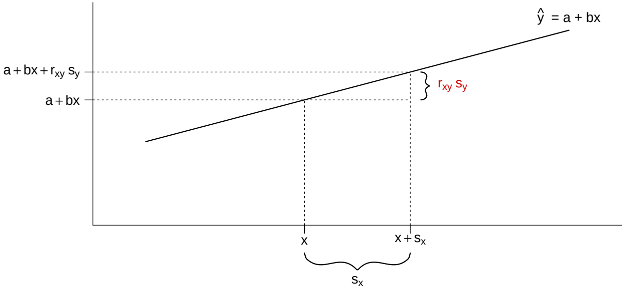
So, when x increases by one SD, y increases bs_x units.
But $b = r_{xy}\frac{s_y}{s_x}$, so $bs_x = r_{xy}s_y$.

Conclusion:
When x increases by one SD, y increases only by $r_{xy}s_y$, that is, less than one SD (recall that $|r_{xy}| \leq 1$).

The closer r_{xy} is from 0:

- The closer the slope b is from 0.
- The closer the regression line is from being horizontal.
- The closer the y values are to $\overline{y}$.

This is known as regression toward the mean.



Next week

Contents:

- Simple linear regression/correlation: Inference

Read:

- Section 9.5.
- Additional text in the reader (see Nestor) in ‘Inference for Correlations’.