Simple Linear Regression - Tasks

M Douch based on Florian Oswald et al.

October 14, 2025

Task 1

Task 1: Getting to know the data

- Load the data from [here](https://www.dropbox.com/s/wwp2cs9f0dubmhr/grade5.dta?dl=1) as grades.
- Need a function to import *.dta* files (Stata).

R Code

```
library(haven)
link <- "https://www.dropbox.com/s/wwp2cs9f0dubmhr/grade5.dta?dl=1"
grades <- read_dta(link)
```

Task 1: Getting to know the data

Alternative: load from the directory

R Code

1. Clear the environment (remove all objects)

```
rm(list = ls())
```

2. Check your current working directory

```
getwd()
```

3. Set your working directory (optional, adjust path as needed)

```
setwd("C:/Users/YourName/Path/To/Data")
```

4. Confirm the directory has the data file

```
list.files()
```

5. Load the data

```
grades <- read_dta("grade5.dta")
```

Task 1: Getting to know the data (cont.)

- **Unit of observation:** What does each row correspond to?

Task 1: Getting to know the data (cont.)

- **Unit of observation:** What does each row correspond to?
- How many observations are there?

R Code: Number of Observations

```
nrow(grades)
```

Task 1: Getting to know the data (cont.)

- **Unit of observation:** What does each row correspond to?
- How many observations are there?

R Code: Number of Observations

```
nrow(grades)
```

- What variables do we have?

R Code: Variable Names

```
names(grades)
```

*Note that if you view the data, under each column name you have the variable's label, which is very convenient.

Task 1: Getting to know the data (cont.)

- What do `avgmath` and `avgverb` correspond to?

Task 1: Getting to know the data (cont.)

- What do `avgmath` and `avgverb` correspond to?
- Use `skim()` for summary statistics of `classize`, `avgmath` and `avgverb`.

- Use the ‘skim‘ function from the ‘skimr‘ package to obtain common summary statistics for the variables ‘classize‘, ‘avgmath‘ and ‘avgverb‘. (*Hint: use ‘dplyr‘ to ‘select‘ the variables and then simply pipe (‘%>%‘) ‘skim()‘.*)

R Code: Summary Statistics with `skimr`

```
library(skimr)
library(tidyverse)
grades %>%
select(classize, avgmath, avgverb) %>%
skim()
```

Task 1: Getting to know the data (cont.)

- What do `avgmath` and `avgverb` correspond to?
- Use `skim()` for summary statistics of `classize`, `avgmath` and `avgverb`.

- Use the ‘skim‘ function from the ‘skimr‘ package to obtain common summary statistics for the variables ‘classize‘, ‘avgmath‘ and ‘avgverb‘. (*Hint: use ‘dplyr‘ to ‘select‘ the variables and then simply pipe (‘%>%‘) ‘skim()‘.*)

R Code: Summary Statistics with `skimr`

```
library(skimr)
library(tidyverse)
grades %>%
select(classize, avgmath, avgverb) %>%
skim()
```

Note: Class sizes range from 5 to 44 (average $\approx$ 30). Average math scores were slightly lower and more dispersed than average verb scores. No missing values.

Task 1: Getting to know the data (cont.)

- Do you have any priors about the actual (linear) relationship?

Task 1: Getting to know the data (cont.)

- Do you have any priors about the actual (linear) relationship?
- What would you do to get a first insight?

Task 1: Getting to know the data (cont.)

- Do you have any priors about the actual (linear) relationship?
- What would you do to get a first insight? **First Insight:** A **scatter plot** would provide a first insight.

R Code: Scatter Plots

```
library(cowplot)
scatter_verb <- grades %>%
ggplot() + aes(x = classize, y = avgverb) + geom_point() + scale_x_continuous(limits = c(0, 45), breaks =
seq(0,45,5)) + scale_y_continuous(limits = c(0, 100), breaks = seq(0, 100, 20)) + labs(x = "Class size", y =
"Average reading score")

scatter_math <- grades %>%
ggplot() + aes(x = classize, y = avgmath) + geom_point() + scale_x_continuous(limits = c(0, 45), breaks =
seq(0,45,5)) + scale_y_continuous(limits = c(0, 100), breaks = seq(0, 100, 20)) + labs(x = "Class size", y =
"Average math score")
plot_grid(scatter_verb, scatter_math, labels = c("Reading", "Mathematics"))
```

Task 1: Getting to know the data (cont.)

- Compute the correlation between class size and math/verbal scores.
- Is the relationship positive/negative, strong/weak?

R Code: Correlation

```
grades %>%
summarise(cor_verb = cor(classize, avgverb), cor_math = cor(classize,
avgmath))
```

Task 2

Task 2: OLS Regression

Run the following code to aggregate the data at the class size level:

R Code: Data Aggregation

```
grades_avg_cs <- grades %>%
group_by(classize) %>%
summarise(avgmath_cs = mean(avgmath), avgverb_cs = mean(avgverb))
```

Task 2: OLS Regression (cont.)

- Compute the OLS coefficients b_0 and b_1 of the regression of `avgmath_cs` on `classize` using the formulas:

$$b_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

(*Hint:* you need to use the ‘cov‘, ‘var‘, and ‘mean‘ functions.)

Task 2: OLS Regression (cont.)

- Compute the OLS coefficients b_0 and b_1 of the regression of `avgmath_cs` on `classize` using the formulas:

$$b_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

(*Hint:* you need to use the ‘cov‘, ‘var‘, and ‘mean‘ functions.)

R Code: Manual OLS Calculation

```
cov_x_y = grades_avg_cs %>%
summarise(cov(classize, avgmath_cs))
var_x = var(grades_avg_cs$classize)
b_1 = cov_x_y / var_x
b_1

y_bar = mean(grades_avg_cs$avgmath_cs)
x_bar = mean(grades_avg_cs$classize)
b_0 = y_bar - b_1 * x_bar
b_0
```

Task 2: OLS Regression (cont.)

- Regress average verbal score (`avgverb_cs`) on class size (`classize`).
- Interpret the coefficients b_0 and b_1.

R Code: OLS Regression

```
lm(avgverb_cs ~classize, grades_avg_cs)
```

Task 2: OLS Regression (cont.)

- Is the slope coefficient for verbal score similar to the one for average math score? Was this expected?
- What is the predicted average verbal score when class size is equal to 0?

Task 2: OLS Regression (cont.)

- Is the slope coefficient for verbal score similar to the one for average math score? Was this expected?
- What is the predicted average verbal score when class size is equal to 0? **Answer:** $b_0 \approx 69.19$. This makes **no sense** in this context.
- What is the predicted average verbal score when the class size is equal to 30 students?

Task 2: OLS Regression (cont.)

- Is the slope coefficient for verbal score similar to the one for average math score? Was this expected?
- What is the predicted average verbal score when class size is equal to 0? **Answer:** $b_0 \approx 69.19$. This makes **no sense** in this context.
- What is the predicted average verbal score when the class size is equal to 30 students?

Task 3

Task 3: R^2 and Goodness of Fit

1. Regress `avgmath_cs` on `classize`. Assign to `math_reg`.

R Code: Math Regression

```
math_reg <- lm(avgmath_cs ~classize, grades_avg_cs)
```

2. Pass 'math_reg' in the 'summary()' function. What is the (multiple) R^2 for this regression? How can you interpret it?

R Code: Summary of Regression

```
summary(math_reg)
```

Task 3: R^2 and Goodness of Fit (cont.)

3. Compute the squared correlation between `classize` and `avgmath_cs`.
   - What does this tell you about the relationship between R^2 and correlation?

R Code: Squared Correlation

```
grades_avg_cs %>%
summarise(cor_sq = cor(classize, avgmath_cs)^2)
```

Task 3: R^2 and Goodness of Fit (cont.)

- Use $R^2 = \frac{\text{SSE}}{\text{SST}}$, where $\text{SST} = \text{Var}(y)$ and $\text{SSE} = \text{Var}(\hat{y})$ (predicted values).

4. Install and load the 'broom' package. Pass 'math_reg' in the 'augment()' function and assign it to a new object. Use the variance in 'avgmath_cs' (SST) and the variance in '.fitted' (predicted values; SSE) to find the R^2 using the formula on the previous slide.

R Code: R^2 from Variance

```
library(broom)
math_reg_aug <- augment(math_reg)
SST = var(grades_avg_cs$avgmath_cs)
SSE = var(math_reg_aug$.fitted)
SSE/SST
```

Task 3: R^2 and Goodness of Fit (cont.)

- Repeat steps 1 and 2 for `avgverb_cs`.
- For which exam does class size explain more variance in students' scores?

R Code: Verbal Regression Summary

```
verb_reg <- lm(avgverb_cs ~classize, grades_avg_cs)
summary(verb_reg)
```

Task 3: R^2 and Goodness of Fit (cont.)

- Repeat steps 1 and 2 for `avgverb_cs`.
- For which exam does class size explain more variance in students' scores?

R Code: Verbal Regression Summary

```
verb_reg <- lm(avgverb_cs ~classize, grades_avg_cs)
summary(verb_reg)
```

Comparison: The R^2 is **greater for maths** (≈ 0.28) than for reading (≈ 0.046). Therefore, class size explains **more of the variation** in math scores than in reading scores.