

Ordinary Least-Squares Estimation

Contents

Pre-Requisite Knowledge	2
Lab Assignment Settings	2
Specific Learning Objectives	2
Lab Assignment	2
Setup	2
Rubric	3
The Facebook Dataset	3
1. Main Statistical Inquiry	3
2. Loading Data	4
3. Exploratory Data Analysis (EDA)	4
4. Data Modelling	6
5. Estimation	6
5.1	6
5.2	7
5.3	8
5.4	8
(Optional) 5.5	9
6. Conclusion	10
Submission	11
Reference	11

```
BEGIN ASSIGNMENT
requirements: requirements.R
files:
  - data
generate:
  show_stdout: true
  show_hidden: true
environment: environment.yml
export_cell: false
```

Pre-Requisite Knowledge

This lab assignment includes practical questions involving estimation in simple linear regression via ordinary least-squares. All its content is in a frequentist framework. Furthermore, this assignment would be the introductory practicum in the course **DSCI 561 (Regression I)**.

Besides the sample lecture, students specifically need to be familiar with the following courses and topics:

- **DSCI 551 (Descriptive Statistics and Probability for Data Science)**. Random variables, expected values (and their properties), and normality.
- **DSCI 552 (Statistical Inference and Computation I)**. Estimators, sampling distributions, hypothesis testing, and confidence intervals.
- **DSCI 531 (Data Visualization I)**. Data visualization via the package `ggplot2`.

Lab Assignment Settings

This assignment has the following characteristics:

- It is expected to be submitted as an **R** markdown along with its corresponding PDF file.
- The handout incorporates auto-graded items for instantaneous feedback. These items are built using Otter Grader.
- Moreover, the practicum is designed to be submitted to the online grading platform Gradescope.

Specific Learning Objectives

By the end of this lab assignment, students are expected to attain the following:

- Define linear regression models.
- Estimate their terms using **R** via a sample and interpret them.

Lab Assignment

This lab assignment will allow you to define and conceptualize the simple regression model via a practical case. Moreover, you will get familiar with the process of a typical statistical model involving (but not limited to!) a main statistical inquiry, exploratory data analysis (EDA), mathematical modelling, estimation, and data storytelling.

Setup

To solve this assignment, you need to load the packages below. If you fail to load any of them, you can install them and rerun the cell.

```
library(tidyverse)
library(broom)
library(tree)
library(digest)
library(testthat)
```

Rubric

This assignment is worth **17 points in total** plus 2 bonus points if you solve the optional question. Most of the questions are **auto-graded**. **Thus, you will need to pass all the corresponding auto-grading tests to get their full marks.** The rest of the points belong to **reasoning**.

The Facebook Dataset

It is time to explore another engaging dataset, such as the Spotify data we covered in our lecture. This time we will work with Facebook data. In their work related to data mining for predicting performance metrics of posts on Facebook pages linked to brands, Moro et al. (2016) provide a dataset related to post metrics on Facebook user engagement. This engagement data comes from 2014 on a Facebook page of a famous cosmetics brand. The original dataset has 500 observations, each belonging to specific classes of page posts. You can find the raw dataset in data.world. The CSV file to be used in this assignment is a modified version of this dataset with only 491 observations.

Moreover, it is essential to clarify that the raw dataset has 17 different continuous and discrete variables. Nevertheless, for this assignment, let us narrow them down to the following:

1. The continuous variable `total_engagement_percentage` is a **key variable** for any brand with a Facebook page. It tells us how engaged the Facebook users are with the company's posts, **regardless of whether they previously liked their Facebook page or not**. *The larger the percentage, the better the total engagement.* We compute it as follows:

$$\text{total_engagement_percentage} = \frac{\text{Lifetime Engaged Users}}{\text{Lifetime Post Total Reach}} \times 100\%$$

- **Lifetime Post Total Reach:** The number of overall *Facebook unique users* who *saw* the post.
 - **Lifetime Engaged Users:** The number of overall *Facebook unique users* who *saw and clicked* on the post. This count is a subset of **Lifetime Post Total Reach**.
2. The continuous `share_percentage` is the percentage that the number of *shares* represents from the sum of *likes*, *comments*, and *shares* in each post. It is computed as follows:

$$\text{share_percentage} = \frac{\text{Number of Shares}}{\text{Total Post Interactions}} \times 100\%$$

- **Total Post Interactions:** The sum of *likes*, *comments*, and *shares* in a given post.
- **Number of Shares:** The number of *shares* in a given post. This count is a subset of *Total Post Interactions*.

1. Main Statistical Inquiry

rubric={reasoning:2}

Suppose you are the sales manager of the cosmetics brand; you are interested in the following:

Is the **mean** total engagement percentage dependent on the share percentage on our Facebook page? If so, by how much?

Suppose you want to use simple linear regression (SLR) to answer this inquiry. Hence, answer the following:

1. **In one sentence**, what would be the model's response?

The model's response would be total engagement percentage.

Reasoning rubric for feedback:

- Grant one mark for the correct variable.
2. **In one sentence**, what would be the model's regressor?

The model's regressor would be share percentage.

Reasoning rubric for feedback:

- Grant one mark for the correct variable.

2. Loading Data

Now, let us load the data.

```
facebook_data <- read_csv("data/facebook_data.csv")
head(facebook_data)
```

```
## # A tibble: 6 x 2
##   total_engagement_percentage share_percentage
##           <dbl>           <dbl>
## 1             6.47             17
## 2            13.9             17.7
## 3             7.34             17.5
## 4             4.41             8.27
## 5             9.26             12.5
## 6            11.4             17.7
```

3. Exploratory Data Analysis (EDA)

```
rubric={autograde:3,reasoning:2}
```

Using the variables of interest stored in `facebook_data`, create the **proper visualization** with the regressor and response on the x and y -axes, respectively. Add appropriate axes labels and titles. Store the plot in the variable `facebook_plot`.

Furthermore, **comment on your findings in one to two sentences**.

From our scatterplot, we see a positive relationship between page engagement and share percentages. However, graphically speaking, data points are pretty scattered.

Reasoning rubric for feedback:

- Grant full marks if they state the positive relationship along with the scattered points.
- Penalize one mark if they do not state the scattered points.
- Penalize one mark if they do not state the positive relationship.

BEGIN QUESTION

name: Q3

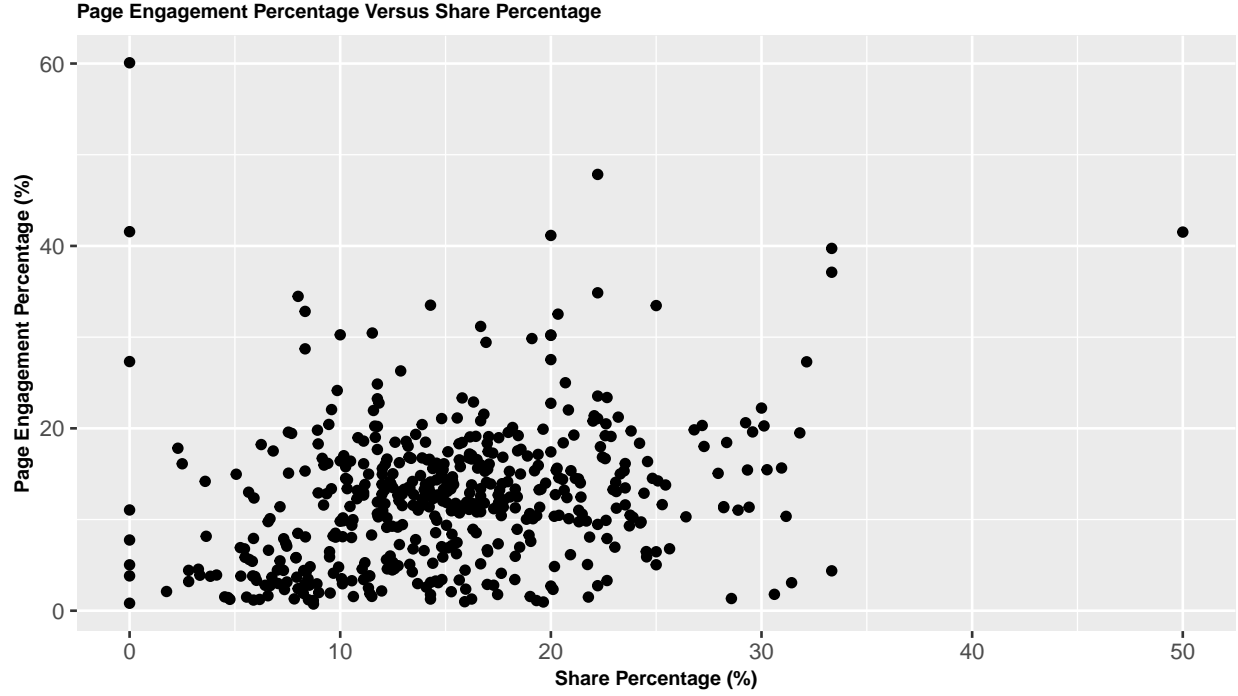
points:

- 0
- 0.5
- 0.5
- 0.5
- 0.5
- 0.5
- 0.5

```
facebook_plot <- NULL

# BEGIN SOLUTION
facebook_plot <- ggplot(
  facebook_data,
  aes(
    x = share_percentage,
    y = total_engagement_percentage,
  )
) +
  geom_point() +
  labs(
    title = "Page Engagement Percentage Versus Share Percentage",
    x = "Share Percentage (%)",
    y = "Page Engagement Percentage (%)"
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 8),
    axis.title = element_text(face = "bold", size = 8)
  )
# END SOLUTION

facebook_plot
```



4. Data Modelling

Once we have our EDA, let us proceed to the SLR modelling. Our training set (i.e., `facebook_data`) has a size of $n = 491$. For the i th observation in our training set ($i = 1, \dots, 491$), the regression equation is the following:

$$\underbrace{Y_i}_{\text{Response}} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Systematic Component}} + \underbrace{\varepsilon_i}_{\text{Random Component}}$$

Recall that X_i is the regressor, whereas β_0 and β_1 are the unknown regression intercept and coefficient, respectively.

5. Estimation

Now, let us start with the model estimation. We will break down this stage into four questions.

5.1 rubric={autograde:2}

As seen during lecture time, we will use ordinary least-squares (OLS) estimation to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$:

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

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

Using the response and regressor from the training set `facebook_data`, compute $\hat{\beta}_0$ and $\hat{\beta}_1$ via the formulas above **by hand**. Bind your results to the **numeric vector-type** variable `beta_0_hat` for $\hat{\beta}_0$ and `beta_1_hat` for $\hat{\beta}_1$.

BEGIN QUESTION

name: Q5.1

points:

- 0
- 0
- 1
- 1

```
beta_0_hat <- NULL
```

```
beta_1_hat <- NULL
```

```
# BEGIN SOLUTION
```

```
beta_1_hat <- sum((facebook_data$share_percentage -  
  mean(facebook_data$share_percentage)) *  
  (facebook_data$total_engagement_percentage -  
    mean(facebook_data$total_engagement_percentage))) /  
  sum((facebook_data$share_percentage -  
    mean(facebook_data$share_percentage))^2)
```

```
beta_0_hat <- mean(facebook_data$total_engagement_percentage) -  
  beta_1_hat * mean(facebook_data$share_percentage)
```

```
# END SOLUTION
```

```
beta_0_hat
```

```
## [1] 7.735435
```

```
beta_1_hat
```

```
## [1] 0.3098828
```

5.2 rubric={autograde:2}

Now, using `lm()` with `facebook_data`, estimate a SLR called `facebook_SLR` to help determine the association of share percentage and page engagement percentage.

BEGIN QUESTION

name: Q5.2

points:

- 1
- 1

```
facebook_SLR <- NULL
```

```
# BEGIN SOLUTION
```

```
facebook_SLR <- lm(total_engagement_percentage ~ share_percentage,  
  data = facebook_data  
)
```

```
# END SOLUTION
```

```
facebook_SLR
```

```
##
## Call:
## lm(formula = total_engagement_percentage ~ share_percentage,
##     data = facebook_data)
##
## Coefficients:
##      (Intercept)  share_percentage
##           7.7354           0.3099
```

5.3 rubric={autograde:1}

Use `tidy()` from the `broom` package to obtain the estimated coefficients of `facebook_SLR`. Bind your results to the variable `tidy_SLR`. Your model estimates have to be equal to the manual computations `beta_0_hat` and `beta_1_hat`.

BEGIN QUESTION

name: Q5.3

points:

- 0.5

- 0.5

```
tidy_SLR <- NULL

# BEGIN SOLUTION
tidy_SLR <- tidy(facebook_SLR)
# END SOLUTION

tidy_SLR
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)    7.74     0.847     9.14 1.73e-18
## 2 share_percentage 0.310    0.0512     6.05 2.92e- 9
```

5.4 rubric={autograde:1}

Compute the corresponding sum of squared residuals (SSR). You will need to use the training set `facebook_data` along with `beta_0_hat` and `beta_1_hat`.

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

Bind your results to the **numeric vector-type** variable `SSR_facebook`.

BEGIN QUESTION

name: Q5.4

points:

- 0

- 1


```
SSR_facebook <- NULL

# BEGIN SOLUTION

SSR_facebook <- sum((facebook_data$total_engagement_percentage -
  beta_0_hat -
  beta_1_hat * facebook_data$share_percentage)^2)

# END SOLUTION

SSR_facebook
```

```
## [1] 28083.82
```

(Optional) 5.5 rubric={autograde:2}

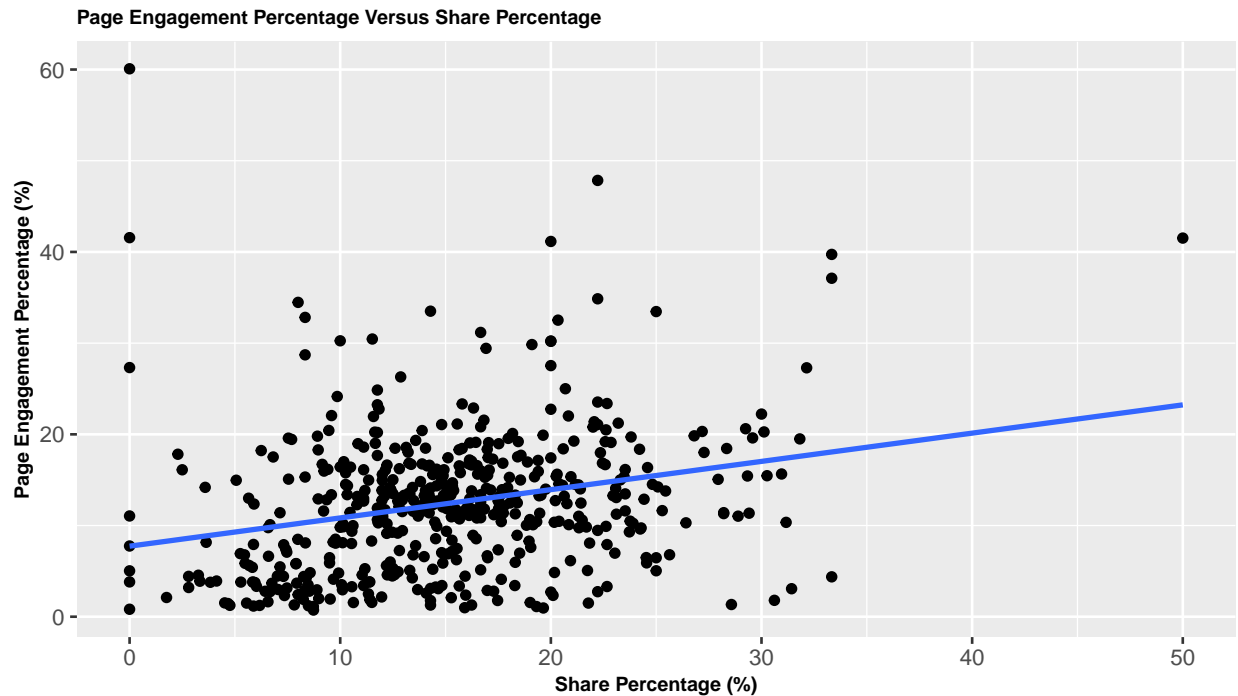
There is a way to automatically plot the estimated OLS regression line via `ggplot2`. Do it on top of `facebook_scatterplot`.

```
BEGIN QUESTION
name: Q5.5
points:
  - 1
  - 1
```

```
# BEGIN SOLUTION
facebook_plot <- facebook_plot +
  geom_smooth(method = lm, formula = y ~ x, se = FALSE)

# END SOLUTION

facebook_plot
```



6. Conclusion

rubric={reasoning:2}

Run the cell below before continuing.

```
tidy_SLR
```

```
## # A tibble: 2 x 5
##   term                estimate std.error statistic  p.value
##   <chr>                <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)          7.74      0.847      9.14 1.73e-18
## 2 share_percentage     0.310    0.0512     6.05 2.92e- 9
```

Use this output to answer the main statistical inquiry **in one or two sentences**. Recall the following:

Is the **mean** total engagement percentage dependent on the share percentage on our Facebook page? If so, by how much?

Yes, we estimate that the mean total engagement percentage is linearly dependent on the share percentage via our regression model. Specifically, we estimate that for each 1% increase in share percentage, the mean total engagement percentage increases by 0.31% on our Facebook page.

Reasoning rubric for feedback:

- Grant full marks if they answer both questions correctly.
- Penalize one mark if they do not mention **estimate**.
- Penalize one mark if they do not mention **mean**.

Submission

You are done with the assignment. Follow these final instructions:

- Knit the assignment to generate the PDF file.
- Submit both the Rmd AND the PDF files to Gradescope.

Reference

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 69(9), 3341-3351.