

TMA4267: Assignment 4

Assignment 4 requires the use of R. Use AI chat-bots or other resources to obtain relevant code outside of what is given in the text. Start by installing recommended packages:

```
install.packages(c("ggplot2", "readr", "GGally"))
```

Task 1: Application

In task 1 you will work with linear regression in R, focusing on getting to know a data set, visualizing observations, fitting linear models, reading and interpreting model output.

Our task is to aid a (fictitious) grocery store chain in decision making: Where should the chain build a new store? We will model weekly turnover using data from $n = 130$ grocery stores. For each store we have the following data:

- **turnover**: Weekly turnover (measured in 1000 NOK).
- **population**: Number of people living within 1 km of the store.
- **distance**: Distance (in km) to the nearest competing grocery store.
- **central**: Indicator variable equal to 1 if the store is located in a city center, 0 otherwise.
- **mall**: Indicator variable equal to 1 if the store lies within a mall (shopping center), 0 otherwise.

a) The data set can be loaded directly into R using the readr package:

```
library(readr)
groceries = read_csv("https://www.math.ntnu.no/emner/TMA4267/2026v/assignments/groceries.csv")
```

Print the first 5 rows using `head()`. Which variables are continuous and which are discrete?

Visualize the empirical distributions of **turnover**, **population** and **distance** in a ‘correlogram’, use e.g. the GGally package and the function `ggpairs(groceries[,1:3])`, see <https://r-graph-gallery.com/correlogram.html>.

Visualize the distribution of weekly turnover stratified on levels (0/1) of the discrete variables using box plots. Example code for the **central** variable:

```
library(ggplot2)
ggplot(groceries, aes(y = turnover, x = as.factor(central))) + geom_boxplot()
```

Discuss what you learn about the data from the figures that you have created.

b) Create a new column in the data frame:

```
groceries$log_distance = log(groceries$distance)
```

Fit a multiple linear regression model with **turnover** as the response and **population** and **log_distance** as covariates:

```
model1 = lm(turnover ~ population + log_distance, data = groceries)
summary(model1)
```

Read the model summary, what are the estimated coefficients of the model? And what is the interpretation of these? In particular, give a detailed interpretation of the estimated regression coefficient for **population**. Also, give an interpretation of the intercept, β_0 .

Discuss how the model reflects what you have observed in the figures created in a). Can you argue for the log-transformation of distance?

c) Extend the model to include the covariates **central**, **mall** and their interaction (product);

```
model2 = lm(turnover ~ population + log_distance + central*mall, data = groceries)
```

How do you interpret the estimated model coefficients? Recall that **central** and **mall** are binary (0/1) and so will the product of them be. Does your interpretation of β_0 change?

Task 2: Distribution of estimators

We will study the distribution of estimators by repeatedly re-doing a (stochastic) experiment. Our experiment is defined by the following design matrix:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

You can think of the first covariate (column 2 of the design matrix) to be an indicator of light (0 = dim light, 1 = bright light), and the second covariate (column 3) to be noise (0 = silent, 1 = noisy). The response variable y that we ‘measure’ in our experiment is reaction time. We see from the design matrix that two experiments will be performed in darkness and silence ($x_1 = 0$, $x_2 = 0$) and so on.

Reaction time, in milliseconds, can be described by the model

$$Y = 225 - 30x_1 + 45x_2 + \varepsilon, \quad \varepsilon \sim N(0, 10^2).$$

a) Write R code that generates observations from this model, for the experiment defined by X. One experimental run gives 8 (independent) observations of reaction times. Fit a linear regression model to the data you have generated and write down the estimated regression coefficients ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$), as well as the estimated variance (s^2). Compare your estimates to the true model parameters.

b) Perform the experiment from a) 1000 times. Store regression coefficient and variance estimates in each repetition. Plot histograms of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $5s^2/10^2$. What do you observe? Where do your observations from a) lie within these histograms? Where do the true model parameters lie?

Discuss how your findings correspond to the theory of chapter 7.6 The normal model. Use your simulation study to describe the difference between an estimator and an estimate.