Statistical Theory II

Lab 5 — Linear Regression

Vilnius University

Use any resources available to you — lecture slides, textbooks, online documentation, R's built-in `help()` function, and GenAI with caution – more for debugging, explanation, and learning purposes.

Exercise 1 — Real Estate Valuation

The dataset **Real Estate Valuation Data Set** contains transactions from the Taiwan real estate market. The variables of interest are:

X2 House age (years)
X3 Distance to the nearest MRT station (metres)
X4 Number of convenience stores nearby
Y House price per unit area

1. Data Properties

i. Read in the Real Estate Valuation dataset. Subset the relevant X2–X4 and Y variables. Display a summary of the subsetted data.

> **Hint:** *The file is an* `.xlsx` *— look into the* `readxl` *package.*

ii. Summarise the data and produce a scatter plot and a box plot investigating how house age is related to house price.

iii. Briefly describe the relationships you observe in the two graphs. Is there a clear trend? Does the spread look consistent across age groups?

2. Linear Regression

i. Run a linear regression predicting house price (Y) as a function of house age (X_2), distance to the nearest MRT station (X_3), and number of convenience stores (X_4).

ii. Describe the output, interpreting the key results. For each predictor, state the estimated effect and whether it is statistically significant. Interpret the coefficient of determination.

iii. Describe and write down each of the following terms in your own words (under 50 words per concept): **Residual**, **Coefficient**, **Intercept**, **Multiple R^2 vs. Adjusted R^2**, and the **F-statistic and its p-value**.

iv. Plot a scatter plot visualising how house age predicts the *fitted* house prices. Obtain the fitted prices using `predict(model_name)` and include a regression line in the plot.

3. Model Diagnostics

i. Produce the following two diagnostic plots for your model:

```
plot(model_name, which = 1)   # Residuals vs Fitted
plot(model_name, which = 2)   # Normal Q-Q
```

For each plot, write a short paragraph covering: what the plot is checking, what a well-behaved result looks like, and what you actually observe in your output.

> **Hint:** *These plots need only the fitted model object — no additional packages required.*

Exercise 2 — Motor Trend Car Data

The `mtcars` dataset is built into base R — no loading required. It contains performance data on 32 car models extracted from a 1974 Motor Trend magazine. Type `?mtcars` in R for the full variable descriptions. You will work with:

mpg	Miles per gallon (fuel efficiency)
wt	Weight of the car (1000 lbs)
hp	Gross horsepower
cyl	Number of cylinders (4, 6, or 8)

i. Run `summary(mtcars)` and `head(mtcars)`. How many observations are there? Describe any notable features of the variables you will be working with.

ii. Produce scatter plots of `wt` vs. `mpg` and `hp` vs. `mpg`. Describe the direction and approximate strength of each relationship.

iii. Run a simple linear regression predicting `mpg` from `wt` alone. Interpret the slope coefficient in plain language: what does it mean for a car to be 1000 lbs heavier?

iv. Extend the model to include `hp` as a second predictor. Compare the two models: does the adjusted R^2 improve? Does the coefficient on `wt` change noticeably when `hp` is added? What does that tell you?

v. Plot the residuals of your two-predictor model against the fitted values. Is there any systematic pattern? What might explain it?

Exercise 3 — Weather Conditions in World War Two

The file **Summary of Weather.csv** provided with this lab contains daily weather observations recorded at stations across multiple WW2 theatres of operation, 1940–1945. Run `names()` and `head()` after loading to orient yourself. The key variables are:

MeanTemp	Mean daily temperature
MaxTemp	Maximum daily temperature
MinTemp	Minimum daily temperature
Precip	Precipitation
WindGustSpd	Wind gust speed
STA	Station identifier
Date	Date of observation

i. Load the CSV into R. How many observations and stations are there? What date range does the data cover? Check for missing values in the temperature columns — how might you handle them before modelling?

ii. Plot the distribution of `MeanTemp`. Then produce a box plot comparing `MeanTemp` across months (you will need to extract the month from `Date`). Does temperature vary seasonally as you would expect?

iii. Run a simple linear regression predicting `MaxTemp` from `MinTemp` alone. Interpret the slope: if the minimum temperature rises by one degree, what does the model predict for the maximum? How much of the variation in `MaxTemp` does `MinTemp` explain?

iv. Extend the model by adding `Precip` and `WindGustSpd` as additional predictors. Does the adjusted R^2 improve? Are the new coefficients statistically significant, and do their signs make intuitive sense?

v. Plot the residuals of your extended model against the fitted values. Given that this is time-series data with daily observations, what potential problem might you be worried about that a standard linear regression does not account for?

Bonus 1 — Patient Medical Costs

The file **patient_data.txt** provided with this lab contains health insurance records for 1,338 individuals in the United States. Load it with `read.csv("patient_data.txt")`. The variables are:

age	Age of the primary beneficiary
sex	Gender
bmi	Body mass index
children	Number of dependants covered
smoker	Smoking status (`yes` / `no`)
region	Residential region in the US
charges	Individual medical costs billed by insurance

i. Load the data and run `summary()`. Plot the distribution of `charges`. Describe its shape — is it symmetric, or skewed? What does that tell you about medical costs in this sample?

ii. Produce a scatter plot of `bmi` vs. `charges`. Describe the relationship. Is it strong? Is it what you would expect?

iii. Run a simple linear regression predicting `charges` from `bmi`. Interpret the intercept and slope. Is BMI a statistically significant predictor of medical costs?

iv. Extend the model by adding `age` as a second predictor. Does the adjusted R^2 improve? How does the coefficient on `bmi` change, if at all? Interpret the coefficient on `age` in plain language.

v. *Extension.* Add `smoker` to your model as a third predictor. Note that this is a categorical variable — R will handle it automatically, creating a dummy variable. How does the model change? What does the coefficient on `smoker` tell you, and does it surprise you?

Bonus 2 — Exploratory Data Analysis (R for Data Science, Ch. 10)

These problems are open-ended and draw on Chapter 10 of *R for Data Science* (2e): https://r4ds.hadley.nz/EDA.html. Use the `diamonds` dataset from `ggplot2` (`library(ggplot2)`).

B.1. Variation. Plot the distribution of `price` using a histogram. Experiment with different bin widths. What do you notice about the shape? Are there unusual spikes or gaps?

B.2. Typical and unusual values. Plot the distribution of `carat` for diamonds below 3 carats using a narrow bin width. Are there more diamonds just above round-number carats (0.5, 1.0, 1.5, ...) or just below? Why might this be?

B.3. Covariation — categorical and continuous. Use a box plot to compare `price` across levels of `cut`. Does a better cut mean a higher price? Is the result what you expected? How might `carat` be confounding the relationship?

B.4. Covariation — two continuous variables. Scatter plot of `carat` vs. `price`. The overplotting will be severe — try `alpha` transparency or `geom_bin2d()` to make the pattern clearer. Does the relationship look linear?

B.5. Patterns and models. Run a regression predicting `price` from `carat`, then extend it by adding `cut`. Compare the two models. Does your regression confirm or complicate what the box plot in B.3 suggested? Explain in 2–3 sentences.

Good luck. Clear, well-commented code is always appreciated.