Exercises on chapter 8: Multiple linear regression

Lieven Clement, Jeroen Gilis and Milan Malfait

statOmics, Ghent University (https://statomics.github.io)

Contents

On the fourth day of the "Practical Statistics for the Life Sciences" course, we will have 6 tutorials on multiple linear regression, based on different datasets:

- The poison dataset
- The poison dataset (2)
- The puromycin dataset
- The KPNA2 dataset
- Blocking: the rats dataset (1)
- Blocking: the rats dataset (2)

1 The poison dataset

In this experiment, 96 fish (dojofish, goldfish and zebrafish) were placed separately in a tank with two liters of water and a certain dose (in mg) of the poison EI-43,064. The resistance of the fish against the poison was measured as the amount of minutes the fish survived after being exposed to the poison (`Surv_time`, in minutes). Additionally, the weight of each fish was measured.

In this tutorial session we will focus on Dojofish, and we will model the survival time in function of the poison dose while correcting for the weight of the fish.

1. We will first analyse the survival data by only considering the dose as an explanatory variable for survival time
2. Next we will model the survival data with and **additive model** for dose and weight

- Exercise: Exercise1

- Data path: “https://raw.githubusercontent.com/statOmics/PSLSData/main/poison.csv”
- Solution1: Solution1

---

2 The poison dataset (2)

Again, we will work with the poison dataset. In contrast to the exercise above, we here we will model the survival time in function of the dose and the weight of the fish, **and including an interaction between dose and weight.**

- Exercise: Exercise2
- Data path:

  `https://raw.githubusercontent.com/statOmics/PSLSData/main/poison.csv`
- Solution2: Solution2

---

3 The puromycin dataset

Data on the velocity of an enzymatic reaction were obtained by Treloar (1974). The number of counts per minute of radioactive product from the reaction was measured as a function of substrate concentration in parts per million (ppm) and from these counts the initial rate (or velocity) of the reaction was calculated (counts/min/min). The experiment was conducted once with the enzyme treated with Puromycin, and once with the enzyme untreated.

Assess if there is an association between the substrate concentration and rate **for both the treated and untreated enzymes.** To do this, fit a model that includes a main effect for concentration, a main effect for enzyme state, and an interaction term between these two variables.

- Exercise: Exercise3
- Data path: Not required
- Solution: Solution3
- Wrapup summary

---

4 The KPNA2 dataset

Histologic grade in breast cancer provides clinically important prognostic information. Researchers examined whether histologic grade was associated with gene expression profiles of breast cancers and whether such profiles could be used to improve histologic grading.

In this tutorial we will assess the association between histologic grade and the expression of the KPNA2 gene that is known to be associated with poor breast cancer prognosis. The patients, however, do not only differ in the histologic grade, but also on their lymph node status. The lymph nodes were not affected (0) or surgically removed (1).

- Exercise: Exercise4
- Data path:

  `https://raw.githubusercontent.com/statOmics/SGA21/master/data/kpna2.txt`

- Solution: Solution4

---

5 Blocking: the rats dataset (1)

Researchers are studying the impact of protein sources and protein levels in the diet on the weight of rats. They feed the rats with diets of beef, cereal and pork and use a low and high protein level for each diet type. The researchers can include 60 rats in the experiment. Prior to the experiment, the rats were divided in 10 homogeneous groups of 6 rats based on characteristics such as initial weight, appetite, etc.

Within each group a rat is randomly assigned to a diet. The rats are fed during a month and the weight gain in grams is recorded for each rat.

The researchers want to assess the effect of the type of diet and the protein level on the weight of the rats.

In this exercise we will perform the data exploration using all diets, but, to keep the data analysis simple we will only assess the beef and cereal diets.

- Exercise: Exercise5
- Data path:

  `https://raw.githubusercontent.com/statOmics/PSLSData/main/dietRats.txt`
- Exercise: Solution5

---

6 Blocking: the rats dataset (2)

Again, we make use of the *rats* dataset defined above. In contrast to the previous exercise, we perform the analysis for all three diets in the dataset.

- Exercise: Exercise6
- Data path:

  `https://raw.githubusercontent.com/statOmics/PSLSData/main/dietRats.txt`
- Exercise: Solution6

---

7 Power analysis: the rodents dataset

A biologist examined the effect of a fungal infection on the eating behavior of rodents. Infected apples were offered to a group of eight rodents, and sterile apples were offered to a group of 4 rodents. The amount of grams of apples consumed per kg body weight are given in the dataset.

We will answer four research questions:

- What is the power of the experiment if the effect size and standard deviation in the population would be equal to the ones you observed in the experiment?
- What would the power by if number of rodents would balanced in both groups?
- How many observations would you need to pick up the treatment effect with a power of 90%?
- Suppose that we would like to pick up an effect size of $\beta_1 = 60g/kg$. How many samples would be required in each group to obtain a power of 90%?

- Exercise: Exercise7
- Data path: Not required

---

8 Power analysis: the Puromycin dataset

Data on the velocity of an enzymatic reaction were obtained by Treloar (1974). The number of counts per minute of radioactive product from the reaction was measured as a function of substrate concentration in parts per million (ppm) and from these counts the initial rate (or velocity) of the reaction was c alculated (counts/min/min). The experiment was conducted once with the enzyme treated with Puromycin, and once with the enzyme untreated.

We will answer four research questions:

- Use the data to calculate the power to pick up an association that is as least as strong as the association you observed in the dataset when using an experiment with the same design.
- Use the data to calculate the power to pick up an association where the reaction rate increases on average with 10 counts/min when the substrate concentration is 10 times higher ($\beta_1 = 10$).
- Use the data to calculate the number of repeats you need for each concentration to pick up an association where the reaction rate increases on average with 10 counts/min when the substrate concentration is 10 times higher with a power of at least 90%. ($\beta_1 = 10$)
- Suppose that you would setup an experiment with a design similar with the same concentrations as in the puromycin dataset and you have the following restriction: you need to use each concentration at least once and can setup at most 12 reactions, how would you choose your design points? Calculate the power for this design when the effect size is 10 counts/min per 10 times increase in the substrate concentration ($\beta_1 = 10$).
- Exercise: Exercise8
- Data path: Not required

---

9 Power analysis: the KPNA2 dataset

Histologic grade in breast cancer provides clinically important prognostic information. Researchers examined whether histologic grade was associated with gene expression profiles of breast cancers and whether such profiles could be used to improve histologic grading.

In this tutorial we will assess the association between histologic grade and the expression of the KPNA2 gene that is known to be associated with poor breast cancer prognosis. The patients, however, do not only differ in the histologic grade, but also on their lymph node status. The lymph nodes were not affected (0) or surgically removed (1).

We will answer three research questions:

- What is the power to pick up each of the contrasts when their real effect sizes would be equal to the effect sizes we observed in the study?
- How does the power evolves if we have 2 upto 10 repeats for each factor combination of grade and node when their real effect sizes would be equal to the ones we observed in the study?
- What is the power to pick up each of the contrasts when the FC for grade for patients with unaffected lymph nodes equals 1.5 ($\beta_g = log2(1.5)$)?

- Data path:

  `https://raw.githubusercontent.com/statOmics/SGA21/master/data/kpna2.txt`