Linear Models

HW1 – Feedback

November 2024

Note

In this document I give some feedback to HW1. Note that I give more detailed answers than what I expected from you. I hope that you will learn from this feedback.

Question 1

Immunoglobulins in the blood are known to protect against diseases. We consider a pre-clinical study in which a potential new drug, a chemical compound with the name ADDF17, is evaluated for negative side effects. One of these effects, could be its negative effect on the blood serum levels of Immunoglobulin IgG1. Healthy mice have IgG1 blood serum concentrations between 1.2 and 5 mg/l.

A lab animal experiment is set up with 62 mice. For each of the following doses of ADDF17, twelve mice were included in the study: $0.025\mu g$, $0.075\mu g$, $0.1\mu g$, $0.2\mu g$, and $0.5\mu g$. For a dose of $2\mu g$ only 2 mice were included. The data are given in the R data file mice.RData.

Read the data.

```
load(file="~/dropbox/education/LIMO/homeworks/2425/HW1/mice.RData")
```

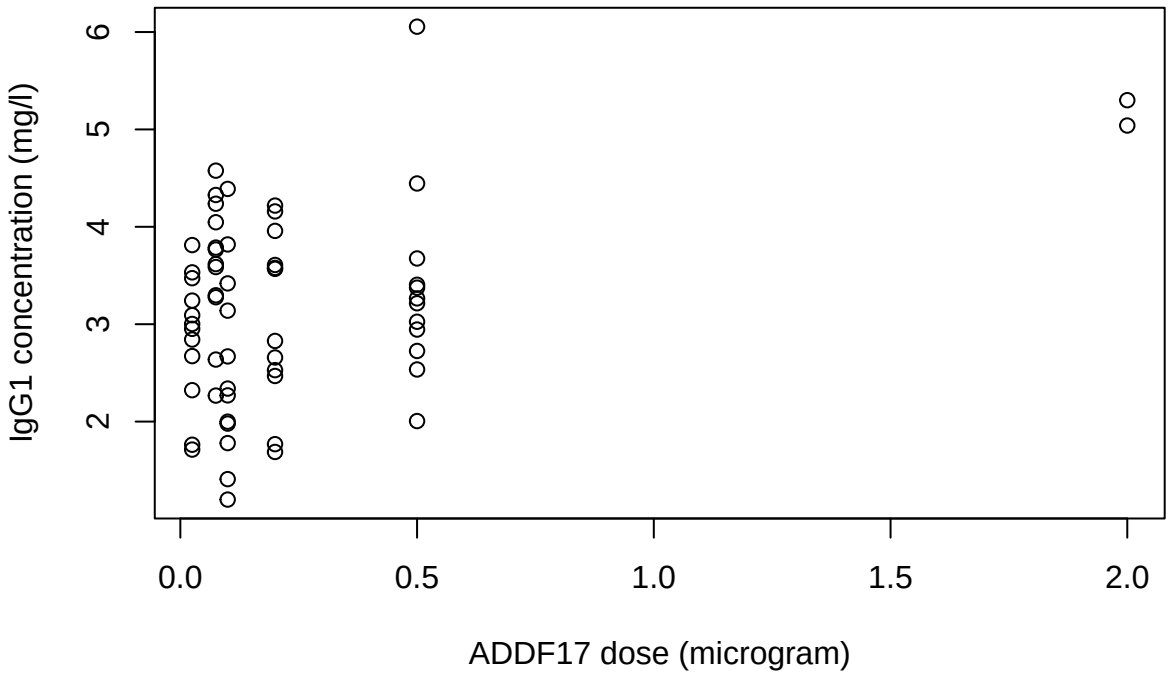
Question 1a

What is the effect of the concentration of ADDF17 on the mean IgG1 blood level concentrations?

This can be answered based on the results of a linear regression analysis. However, it is good statistical practice to always start with a data exploration.

Let's start with a scatter plot of the IgG1 concentrations versus the ADDF17 concentrations.

```
plot(mice$concentration,mice$IgG1,
     xlab="ADDF17 dose (microgram)",
     ylab="IgG1 concentration (mg/l)")
```



The data suggests a tendency of the IgG1 levels to increase with the ADDF17 dose. Also note that most of the observations are for ADDF17 concentrations between 0 and 0.5 microgram, but two observations can be seen at a ADDF17 concentration of 2 microgram. It's good to keep this in the back of our minds.

The linear regression model that we will consider is given by, $i = 1, \ldots, n = 62$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with

- x_i the ADDF17 dose for mouse i (in microgram)
- Y_i the blood serum immunoglobulin IgG1 concentration (mg/l) for mouse i
- ε_i the error term, for which we assume ε_i i.i.d. $N(0, \sigma^2)$.

The sample size is $n = 62$ and so perhaps the normality assumption will not be very important, but still it is good statistical practice to have a look at the distribution of the residuals. This will be done later.

Now the regression analysis.

```
m1<-lm(IgG1~concentration, data=mice)
summary(m1)
```

```
##
## Call:
## lm(formula = IgG1 ~ concentration, data = mice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81822 -0.64593  0.00123  0.59608  2.60921
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.9101     0.1367  21.287   <2e-16 ***
## concentration   1.0714     0.3150   3.401   0.0012 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8989 on 60 degrees of freedom
## Multiple R-squared:  0.1616, Adjusted R-squared:  0.1476
## F-statistic: 11.57 on 1 and 60 DF,  p-value: 0.0012
```

From the output we find that the estimated effect of ADDF17 dose on the average IgG1 concentration equals 1.07 (mg/l)/(μg). This can be interpreted as follows: when the ADDF17 dose increases with 1 μg, we estimate that the IgG1 blood serum level increases with 1.1 mg/l on average.

Question 1b

Give an appreciation/interpretation of the (im)precision of the previous estimate (standard error and a 95% confidence interval).

So we need to report the standard error and the 95% CI of the regression coefficient.

The standard error is part of the output of summary(m1); see earlier for the output. We read that the standard error (i.e. the estimated standard deviation) of the parameter estimate equals 0.32 (mg/l)/μg. This value is about a factor 3 smaller than the parameter estimate itself. However, a more convenient interpretation comes from the 95% CI.

```
confint(m1)
```

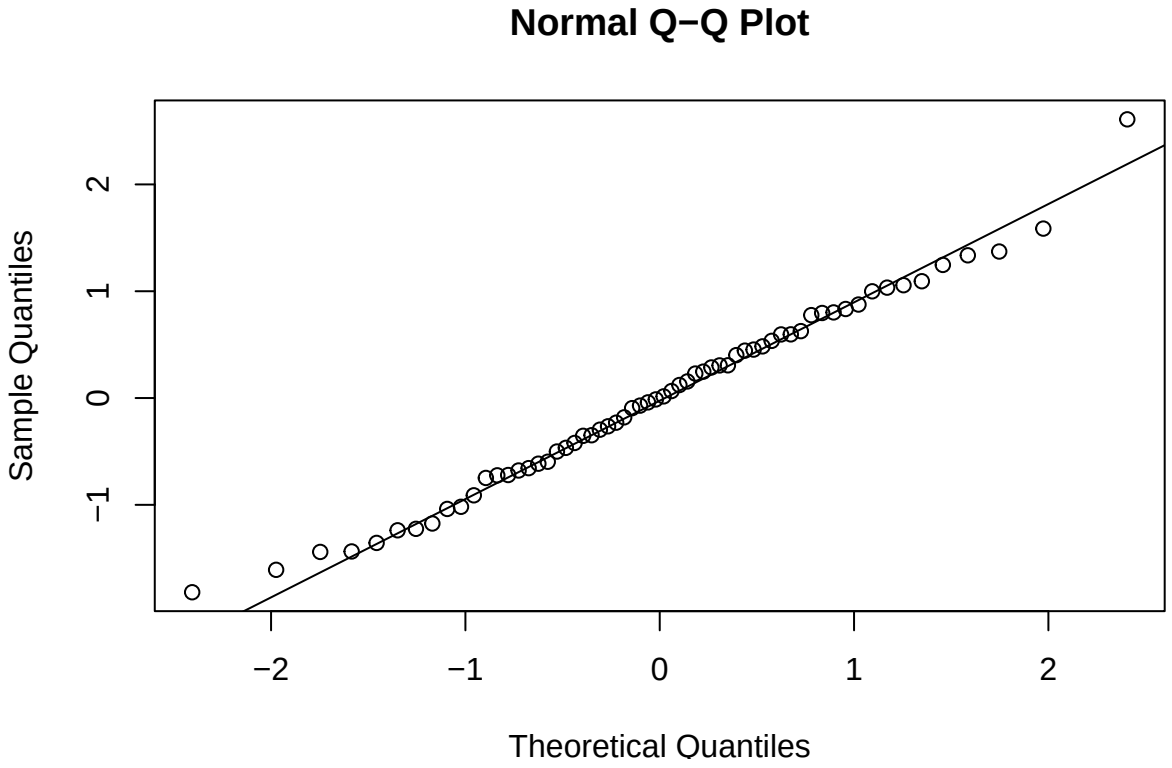
```
##                   2.5 %   97.5 %
## (Intercept)   2.6366273 3.183535
## concentration 0.4412574 1.701578
```

From this output we read that the 95% CI of the regression coefficient ranges between 0.44 (mg/l)/μg to 1.7 (mg/l)/μg.

Hence, with a probability of 95% we expect that an increase of the ADDF17 dose with 1 μg, results in an average increase of the IgG1 blood serum level of somewhere between 0.44 mg/l and 1.7 mg/l.

Although it was not part of your assignment, and although we have argued earlier that the normality assumption is not very important here, we will have a look at the normal QQ-plot of the residuals.

```
e<-m1$residuals
qqnorm(e)
qqline(e)
```



This normal QQ-plot does not suggest any serious deviation from the normality assumption. Hence, confidence intervals and hypothesis tests will be valid and allow for a correct interpretation.

Question 1c

Is there a significant effect of the concentration of ADDF17 on the mean IgG1 blood level concentrations at the 5% level of significance? What is the null and alternative hypothesis considered, and give a motivation for you choice.

Again the answer can be found in the output of the summary function. However, we need to know exactly the null and alternative hypotheses. From the introduction to question 1a (particularly the sentence: "One of these effects, could be its negative effect on the blood serum levels of Immunoglobulin IgG1.' '). This suggests a one-sided alternative hypothesis. The hypotheses are:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 < 0.$$

From the output of the summary function, we read the p-value 0.0012. However, this p-value corresponds to the two-sided alternative hypothesis. To convert it to the correct one-sided p-value, we first look at the sign of the parameter estimate $\hat{\beta}_1$. This sign is positive, which does not agree with the alternative hypothesis. Hence, the correct one sided p-value is equal to one minus half of the two-sided p-value, i.e.

$$p = 0.9994.$$

Since this is larger than the nominal signifance level of $\alpha = 0.05$, we conclude at this significance level that we cannot reject the null hypothesis of no-effect of the ADDF17 level on the mean IgG1 blood serum concentration is negative. There is obviously insufficient evidence.

NOTE: here we have tranlated the research question into a one-sided alterantive $H_1 : \beta_1 < 0$. However, it would also been OK to stick to a two-sided alternative, $H_1 : \beta_1 \neq 0$. This can be used for detecting negative side effects, but it also allows to conclude the opposite.

Question 1d

Repeat the data analysis, but without the data of the mice that received a dose of $2\mu g$. No need to give a detailed conclusion, but only report on the major differences in the conclusion and explain this difference.

```
mice2<-mice[mice$concentration<2,]
dim(mice2)
```

```
## [1] 60  2
```

```
m2<-lm(IgG1~concentration, data=mice2)
summary(m2)
```

```
##
## Call:
## lm(formula = IgG1 ~ concentration, data = mice2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83402 -0.64787  0.02151  0.58186  2.69190
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.9505     0.1717   17.18   <2e-16 ***
## concentration   0.8252     0.6937    1.19    0.239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9127 on 58 degrees of freedom
## Multiple R-squared:  0.02382,    Adjusted R-squared:  0.006985
## F-statistic: 1.415 on 1 and 58 DF,  p-value: 0.2391
```

```
confint(m2)
```

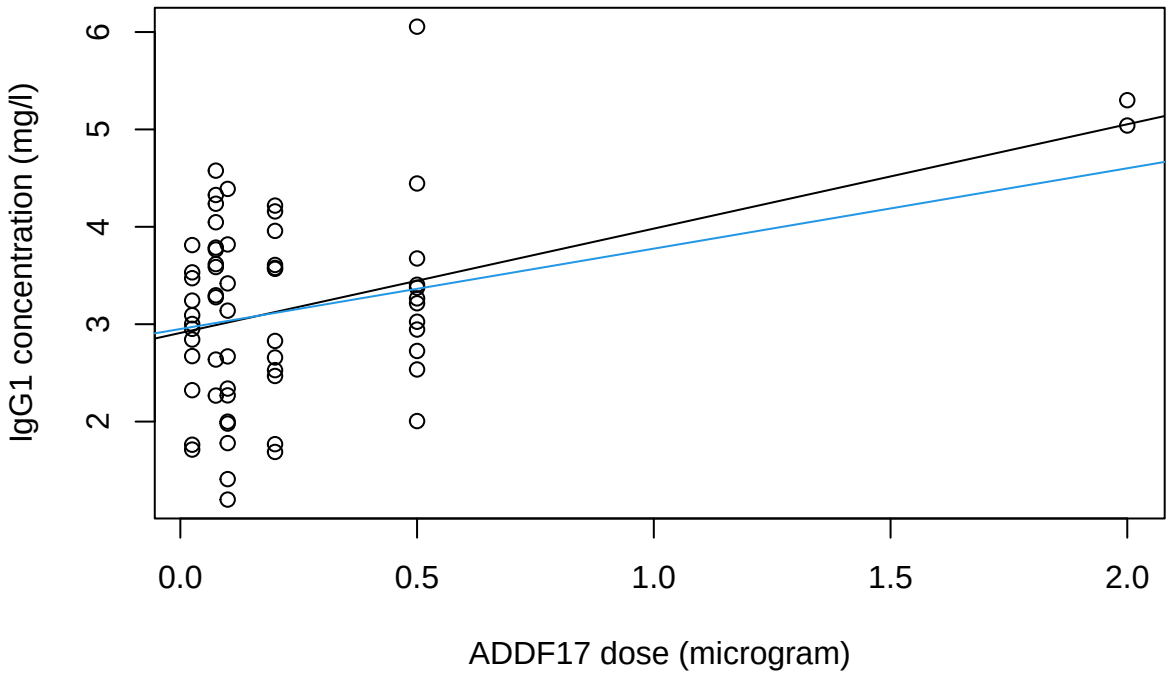
```
##                    2.5 %   97.5 %
## (Intercept)    2.6068235 3.294167
## concentration -0.5634331 2.213855
```

The major differences are

- the effect size is estimated to be 0.83, which is smaller than what we estimated from the full dataset, but, just like before, the effect size is estimated to be positive. This time the 95% CI goes from -0.56 to 2.21, which contains zero. So the data support both a positive and negative effect of ADDF17 on the mean IgG1 concentration.
- the effect is still not significant at the 5% level of significance. The one-sided p-value is $1 - p/2$, with p the two-sided p-value as reported in the output: 0.8805. If the two-sided alternative would have been preferred, it would also result in failing to reject the null hypothesis.

The following plot shows the data and the fits from the two regression models: with the concentrations of 2 microgram (black) and without these concentrations (blue). Would you consider the two observations at the large concentration to be influential observations (try to answer the question yourself).

```
plot(mice$concentration,mice$IgG1,
     xlab="ADDF17 dose (microgram)",
     ylab="IgG1 concentration (mg/l)")
abline(coef(m1), col=1)
abline(coef(m2), col=4)
```



Final notes

The assignment mentions that healthy mice have IgG1 blood serum levels between 1.2 and 5 mg/l. Within the range of ADDF17 concentrations studied here, the IgG1 concentration stayed within the healthy interval. We could further investigate this by e.g. computing 95% prediction intervals for a IgG1 blood serum level for a mice that receives the extreme doses of 0.5 and 2 mg/l of ADDF17.

```
predict(m1,newdata = data.frame(concentration=c(0.5,2)),
        interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.445790 1.625761 5.265818
## 2 5.052916 2.927532 7.178301
```

This interval shows that there is a chance that a mouse receiving an ADDF17 dose of 0.5 or 2 μg could have IgG1 blood serum levels outside the healthy range. Hence, even though ADDF17 has a postive effect on the IgG1 concentration, it may affect the IgG1 concentration to go outside of the healthy region.

Question 2

Consider the R code provided to you in the R Markdown file. What do you conclude from this simulation study. Your answer should be formulated in less than half a page (hence no need to describe the R code, only formulate your conclusion).

```
set.seed(2678)

p1<-p2<-vector(length=1000)
for(i in 1:1000) {
  conc2<-mice$concentration
  mice2<-data.frame(
    concentration=mice$concentration,
    IgG1=3+0.83*conc2+rnorm(62,sd=0.9)
  )
```

```
  conc1<-mice$concentration[mice$concentration<2]
  mice1<-data.frame(
    concentration=conc1,
    IgG1=3+0.83*conc1+rnorm(60,sd=0.9)
  )

  m2<-lm(IgG1~concentration,data=mice2)
  p2[i]<-summary(m2)$coef[2,4]

  m1<-lm(IgG1~concentration,data=mice1)
  p1[i]<-summary(m1)$coef[2,4]
}

mean(p2<0.05)
```

```
## [1] 0.707
```

```
mean(p1<0.05)
```

```
## [1] 0.21
```

The final output shows mean(p2<0.05) and mean(p1<0.05). These show the powers of the tests for the hypotheses

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0,$$

based on the 62 concentrations in the original mice dataset (including those at a ADDF17 concentration of 2 microgram) and on the 60 concentrations in the reduced mice2 dataset (excluding those at a ADDF17 concentration of 2 microgram), respectively (both tests at the 5% level of significance).

From the results we learn that the power of 0.707 for the 62 observations is much larger than the power of 0.21 for the 60 observations .

I did not expect that you give a detailed explanation for this result, but I will tell you: it is not dominated by the difference in sample size (62 versus 60), it is rather the effect of a much smaller SE when the two extreme observations are included (see also the results of models m1 and m2 for Question 1). It is as if these two extreme points confirm the linear relationship that could also be established with only the data with concentrations < 2 microgram.

General feedback to the reports

- Follow the instructions precisely and make use of the R Markdown template. For this HW the instructions (as in the markdown template) told you to give a textual, concise, to-the-point answer to the questions, and provide the R output in the Appendix.
- Please make sure that you submit your reports using the correct file names: LastName_FirstName_HW1 ... Follow the instructions precisely
- You must submit both the R Markdown and the rendered pdf files (if you produced an html file, then print it to a pdf file).
- Do not submit a zip file, but submit both files simultaneously.
- Do not forget to explicitly submit the files (otherwise they remain a "saved draft" and it will not be considered as a submission).
- Use neat English and write full sentences. Use the correct units (e.g. concentration in mg/l).
- Provide sufficient motivation for your choice of data analysis.

- Stick to the questions. Some students have e.g. assessed the model assumptions. This was not needed. It's good statistical practice to so, but particularly for the final exam it is best to read the questions very well, and stick to these questions.
- Do not just report estimates, standard errors, CIs and p-values, but include them in full sentences to give a meaning to these values.
- When reporting confidence intervals, try to avoid simply reporting $[-2, 0.08]$, but rather express it in a sentence (and provide the correct units). This will also make it easier for you to report that the data supports both a descrease of the IgG1 concentration and an increase.
- Mind the number of decimals when reporting results. Usually just a few decimals are more than enough.
- When you use mathematical symbols and give equations, please make sure that they are nicely formatted. For example, do not write y_i, but rather y_i.
- Some students translated $H_0 : \beta_1 \neq 0$ as "the slope is significantly different from zero". We reserve the term "significant" for the result of a statistical test, not for the formulation of the hypotheses.
- Try to see the relevance of the data analysis results. For example, here it was given that healthy mice have blood serum levels of IgG1 between 1.2 and 5 mg/l), and the statistical analyses demonstrated that the doses of ADDF17 administered to the mice can possibly result in the IgG1 level outside the healthy interval.
- Note that upon submission your reports are automatically tested for plagiarism. Please do avoid copying from one another for the next homeworks! I do not mind that you discuss problems and issues together, but the final result (including the R code) should be your own work.
- I've seen a simple copy-and-paste from the feedback to HW1 from last year. Note that it was a different assignment this year, and the code of last year will not produce the correct results.