Exam Linear Models – January 2024

Feedback

In this document you find feedback on the exam. I only give general feedback with a short description of one or possible answers. The next items reflect a general impression that I had after marking the exams. This may of course not apply to all students, but I hope it is helpful.

- Some students write **an** answer to **a** question, but not **the** answer to **the** question. I sometimes had the impression that the student did not read the question very well. So I recommend to always read the question carefully, and, when you are finished with answering the question, read the question again and try to reassure yourself that you gave an answer to the question.
- A little bit related to the previous remark: when R output is given, some students start commenting and interpreting the R output, but again without clear reference to the actual question that is asked. Statistical software often gives a lot of output for e.g. a linear regression fit, but for answering a specific exam question you often don't need to use and interpret all R output.
- You should write in well formulated sentences, and make your arguments very clear.
- Do not focus too much on p-values. For some questions, you were even explicitly asked to not use p-values. Don't forget about the effect sizes, their imprecision and relevance. Also report non-significant findings if that is part of the answer to the question.
- Question 1e appeared to be very difficult or unclear for almost all students. I decided to ignore this question (i.e. it did not contribute to your final score).
- Overall the exam scores were low, and I was suprised by this. I thought the exam was not very difficult, particularly because many questions were very similar to examples in the course notes and to problems solved in the homeworks (for which you have received the solutions). For each of the exam questions, I will also indicate why I think you should have known the answer (or at least part of it).

Question 1

In gene expression studies and in microbiome studies, human samples often have to be stored for some time before the samples can be processed in the lab. It is important that during the storage the quality of the DNA does not detoriate too much.

Here we are presented with the data from a study that aims to **study the effects of several storage conditions on the quality of the DNA, and to find the best storage conditions**. In particular, the following storage factors are studied:

- *time*: storage time: one week or one month
- *temp*: storage temparature: -80 degrees Celcius or +20 degrees Celcium (room temperature)
- *medium*: storage medium (i.e. the medium in which the human samples is placed during storage): EN (eNAT medium), G (Glycerol) and ZYM (Zymo medium)

In the dataset there is also the variable *temp.time*, which is the combination (concatenation) of the time and temperature variables.

The outcome variable is *intact*, which is an intact score. It is approximately to be interpreted as the percentage of the DNA present in the human sample that is intact after storage.

In total there are 9 different storage conditions. All human samples are randomly assigned to the storage conditions.

The data is in the *DNA* R data frame. Below you find the results of a data exploration.

```
## Loading required package: carData
```

```r
str(DNA)
```

```
## 'data.frame':    54 obs. of  5 variables:
##  $ temp     : num  -80 -80 20 -80 -80 20 -80 -80 20 -80 ...
##  $ time     : chr  "week" "month" "week" "week" ...
##  $ medium   : chr  "EN" "EN" "EN" "G" ...
##  $ temp.time: chr  "-80 week" "-80 month" "20 week" "-80 week" ...
##  $ intact   : num  91.1 73.1 76.2 74 66.4 60.5 89.5 84.2 69.2 88.1 ...
```

```r
head(DNA)
```

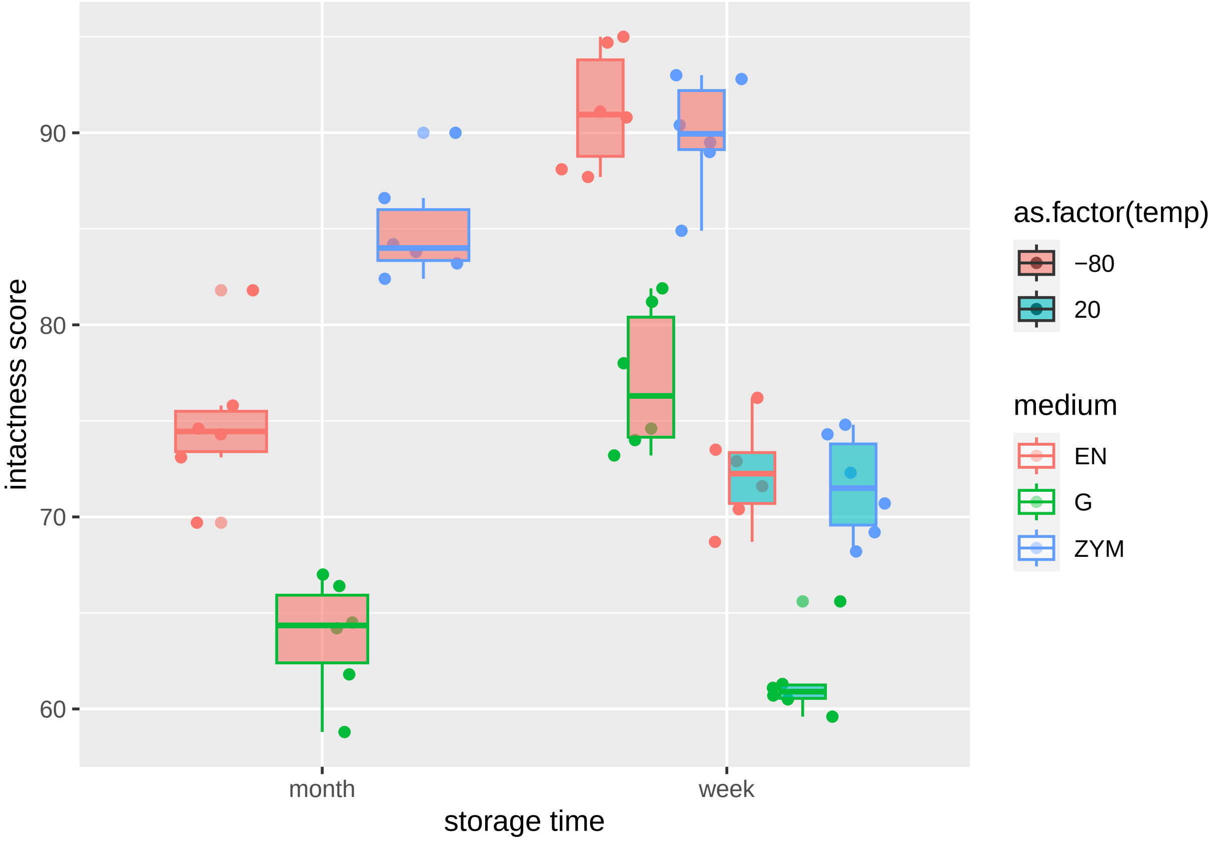
```
##   temp  time medium temp.time intact
## 1  -80  week     EN  -80 week   91.1
## 2  -80 month     EN -80 month   73.1
## 3   20  week     EN   20 week   76.2
## 4  -80  week      G  -80 week   74.0
## 5  -80 month      G -80 month   66.4
## 6   20  week      G   20 week   60.5
```

```r
table(DNA[,1:3])
```

```
## , , medium = EN
##
```

```
##       time
## temp  month week
##   -80     6    6
##   20      0    6
##
## , , medium = G
##
##       time
## temp  month week
##   -80     6    6
##   20      0    6
##
## , , medium = ZYM
##
##       time
## temp  month week
##   -80     6    6
##   20      0    6
```

```r
ggplot(DNA, aes(x=time, y=intact, fill=as.factor(temp), colour=medium)) +
  labs(x="storage time", y="intactness score") +
  geom_point(position = position_jitterdodge()) +
  geom_boxplot(alpha=0.6)
```



a. (2 points). Consider the following R output and explain why not all parameters can be estimated.

```
m1<-lm(intact~(temp+time+medium)^2,data=DNA)
summary(m1)
```

```
##
## Call:
## lm(formula = intact ~ (temp + time + medium)^2, data = DNA)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.183 -1.858 -0.400  2.354  6.917
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       59.670000   1.888905  31.590  < 2e-16 ***
## temp              -0.190167   0.017691 -10.749 5.15e-14 ***
## timeweek          16.350000   1.769121   9.242 5.79e-12 ***
## mediumG           -8.433333   2.671315  -3.157 0.002844 **
## mediumZYM         10.683333   2.671315   3.999 0.000234 ***
## temp:timeweek            NA         NA      NA       NA
## temp:mediumG       0.033333   0.025019   1.332 0.189465
## temp:mediumZYM     0.006667   0.025019   0.266 0.791100
## timeweek:mediumG  -2.983333   2.501915  -1.192 0.239347
## timeweek:mediumZYM -11.450000   2.501915  -4.576 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.064 on 45 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9148
## F-statistic: 72.17 on 8 and 45 DF,  p-value: < 2.2e-16
```

Answer

The answer to this question is similar to questions 1a and 1c of the example-exam that was available on BB.

From the description of the design of the study, we can see that there are no observations for storing conditions given by room temperature and a storage duration of one month: for each of the three mediums, there is only data for three out of the four combinations of temperature and storage time. On the other hand, there are four parameters in the model for modelling the mean outcome at the four combinations of temperature and time (corresponding to lines 1,2,3 and 6 of the summary-output). It is not possible to uniquely estimate the four parameters if there is only data for 3 temperature/time combinations.

Hence, the design matrix X is not of full rank.

b. (2 points) Consider the two following models (m2 and m3) and give the mathematical regression model formulatation for these two models. For example, for a simple linear regression model, this looks like

$$Y_i = \beta_0 + \beta_1 X_i + \cdots$$

This is of course an incomplete model formulation. You are asked to give the full regression formulations for the two models, including the model assumptions.

```
m2<-lm(intact~temp.time,data=DNA, subset=(medium=="G"))
summary(m2)
```

```
##
## Call:
## lm(formula = intact ~ temp.time, data = DNA, subset = (medium ==
##     "G"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9833 -1.9542 -0.2667  2.1750  4.7500
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            63.783      1.250  51.028  < 2e-16 ***
## temp.time-80 week      13.367      1.768   7.562 1.71e-06 ***
## temp.time20 week       -2.317      1.768  -1.311     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.062 on 15 degrees of freedom
## Multiple R-squared:  0.8595, Adjusted R-squared:  0.8407
## F-statistic: 45.87 on 2 and 15 DF,  p-value: 4.057e-07
```

```
m3<-lm(intact~temp+time,data=DNA, subset=(medium=="G"))
summary(m3)
```

```
##
## Call:
## lm(formula = intact ~ temp + time, data = DNA, subset = (medium ==
##     "G"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9833 -1.9542 -0.2667  2.1750  4.7500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.23667    1.88740  27.147 3.63e-14 ***
## temp        -0.15683    0.01768  -8.872 2.35e-07 ***
## timeweek    13.36667    1.76771   7.562 1.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.062 on 15 degrees of freedom
## Multiple R-squared:  0.8595, Adjusted R-squared:  0.8407
## F-statistic: 45.87 on 2 and 15 DF,  p-value: 4.057e-07
```

Answer

This question is about writing down a model in mathematical notation, and correctly defining the notation. The course notes contain many, many model formulations, and at almost all occasions all notation is explictly defined. Throughout the course, we make a clear distinction between regression and ANOVA noation. This should have been clear to you. It is of utmost importance that a statistician can correctly write a statistical model.

Model m2 has *temp.time* as a factor variable. *temp.time* is defined as a combination of temperature (20 and -80) and time (week and month). It is a factor variable with three levels. From the output we learn that there are two dummies, referring to the levels *-80 week* and *20 week*. The missing factor level is *-80 month*, which comes alphabetically first and hence serves (by default) as the reference group. For the regression model we need two dummy variables, defined as

- $X_{i1} = 1$ if DNA sample i comes from a storage temperature of -80 degrees Celcius and a storage time of one week. Otherwise, $X_{i1} = 0$
- $X_{i2} = 1$ if DNA sample i comes from a storage temperature of 20 degrees Celcius and a storage time of one week. Otherwise, $X_{i2} = 0$

The mathematical model can be written as, $i = 1, \ldots, 54$,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

with

- Y_i the intact score of DNA sample i
- X_{i1} and X_{i2} as defined above
- β_0 the intercept. Its interpretation comes from $E(Y \mid$ -80 degrees Celcius, one month$) = \beta_0$
- β_1: the effect of X_1, i.e.

$$E(Y \mid \text{-80 degrees Celcius, one week}) - E(Y \mid \text{-80 degrees Celcius, one month})$$

- β_2: the effect of X_2, i.e.

$$E(Y \mid \text{20 degrees Celcius, one week}) - E(Y \mid \text{-80 degrees Celcius, one month})$$

- ε_i: error term, ε_ii.i.d.$N(0, \sigma^2)$

By formulating the model in this way, the assumptions are already clear and well defined.

Model m3 has temperature as a continuous regressor, and time as a factor variable with levels *week* and *month.* By default, R will consider "month' ' as the reference group (alphabetically first). The mathematical formulation of the model. For $i = 1, \ldots, 54$,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

with

- X_{i1} the temperature of observation i, measured in degrees Celcius
- X_{i2}: a dummy variable coding for the storage time. $X_{i2} = 1$ if the storage time is one week, $X_{i2} = 0$ otherwise.
- β_0: the intercept
- β_1 and β_2: the effects of X_1 and X_2, resp.
- ε_i: error term, ε_ii.i.d.$N(0, \sigma^2)$

c. (2 points) Consider again the two models (m2 and m3) from the previous question. Do the parameter estimates of the two models result in the same conclusions, or in different conclusions. Motivate your answer. Note: the question is about parameter estimates and *not* about hypothesis tests.

Answer

Many students referred to the p-values in their answers, whereas in the formulation of the question it is clearly stated that you need to use parameter estimates and no p-values.

In terms of the parameter estimates, the two models give the same conclusions. This can be seen as follows.

- From m2 we learn that the average difference in intact score between storage at -80 degrees for one week and storage at -80 degrees for one month, is estimated as 13.367. This is thus the effect of time, given a temperature of -80 degrees. From m3, keeping temperature fixed, the effect of time is also estimated to be 13.367.

- From m2 we learn that the average difference in intact score between storage at 20 degrees for one week and storage at -80 degrees for one month, is given by -2.317. The same estimated effect can also be deduced from m3:

$$(51.237 + 20 \times (-0.15683) + 13.367) - (51.237 + (-80) \times (-0.15683)) = -2.317$$

Note that also the R^2 of the two models coincide, telling us that the two models explain the same percentage of the total variability of the intact score.

You could also have remarked that with model m3 we can have a more detailed conclusion for the effect of temperature (because it enters the model as continuous regressor). The parameter estimate for the effect of temperature is -0.15, telling us that, while keeping storage time constant, the intact score decreases on average with 0.15 when temperature increases with one degree Celsius.

d. (2 points) Based on the following model fit (m4), answer the original research question as complete as possible. You may use the parameter estimates and the p-values; there is no need to report standard errors or confidence intervals for this question. You may assume that all model assumptions hold.

```
m4<-lm(intact~temp.time*medium,data=DNA)
summary(m4)
```

```
## 
## Call:
## lm(formula = intact ~ temp.time * medium, data = DNA)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -5.183 -1.858 -0.400  2.354  6.917 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                   74.883      1.251  59.861  < 2e-16 ***
## temp.time-80 week             16.350      1.769   9.242 5.79e-12 ***
## temp.time20 week              -2.667      1.769  -1.507    0.139    
## mediumG                      -11.100      1.769  -6.274 1.22e-07 ***
## mediumZYM                     10.150      1.769   5.737 7.65e-07 ***
## temp.time-80 week:mediumG     -2.983      2.502  -1.192    0.239    
## temp.time20 week:mediumG       0.350      2.502   0.140    0.889    
## temp.time-80 week:mediumZYM  -11.450      2.502  -4.576 3.71e-05 ***
## temp.time20 week:mediumZYM   -10.783      2.502  -4.310 8.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.064 on 45 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9148 
## F-statistic: 72.17 on 8 and 45 DF,  p-value: < 2.2e-16
```

```
Anova(m4,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: intact
##                    Sum Sq Df   F value    Pr(>F)
## (Intercept)         33645  1 3583.3161 < 2.2e-16 ***
## temp.time            1272  2   67.7433 2.675e-14 ***
## medium               1356  2   72.1875 9.068e-15 ***
## temp.time:medium      314  4    8.3553 3.957e-05 ***
## Residuals             423 45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m4,type="II")
```

```
## Anova Table (Type II tests)
##
## Response: intact
##                     Sum Sq Df  F value    Pr(>F)
## temp.time          2901.60  2 154.5153 < 2.2e-16 ***
## medium             2205.29  2 117.4357 < 2.2e-16 ***
## temp.time:medium    313.81  4   8.3553 3.957e-05 ***
## Residuals           422.52 45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer

This is a question that requires a typical interpretation of a model fit. Many students seemed to have difficulties with this question and made a severe error: many students (tried to) interpret all parameter estimates and p-values that are listed in the output of the summary function. However, the anova table shows that there is a significant interaction effect, and hence main effects cannot be interpreted! We have seen this several times in the course.

Thus, from the ANOVA table with type III sum of squares, we conclude at the 5% level of significance that there is a signifcant interaction effect of temperature-time and medium. This is of course already an important conclusion! So, the effect of temperature-time is not the same for all 3 mediums, or, equivalently, for each temperature-time combination, the effect of medium is not the same.

Because of the interaction effect, we cannot interpret the main effects, but we can interpret the parameter estimates of the interaction effect. First we note that the reference group is formed by -80 degrees Celsius and a storage time of one month (for the temp.time factor), and the EN medium (for the factor medium). From the output of the summary function we can now conclude (note that I will only use parameter estimates and p-values, as in the formulation of question 1d):

- DNA stored at 20 degrees Celsius for one week in the ZYM medium is estimated to have on average an intact score of 10.783 lower than DNA stored at -80 degrees Celsius for one month in the EN medium. This difference is significant, with $p < 0.0001$.

- DNA stored at -80 degrees Celsius for one week in the ZYM medium is estimated to have on average an intact score of 11.45 lower than DNA stored at -80 degrees Celsius for one month in the EN medium. This difference is significant, with $p < 0.0001$.
- DNA stored at 20 degrees Celsius for one week in the G medium is estimated to have on average an intact score of 0.35 higher than DNA stored at -80 degrees Celsius for one month in the EN medium. This difference is not significant ($p = 0.889$).
- DNA stored at -80 degrees Celsius for one week in the G medium is estimated to have on average an intact score of 1.192 lower than DNA stored at -80 degrees Celsius for one month in the EN medium. This difference is not significant ($p = 0.239$).

This is not a complete answer to the original research question, but based on the available output there is not much more you can do. It would also be possible to compare other storage conditions with one another (by adding up parameter estimates), but based on the available output you cannot calculate the corresponding p-values.

Finally, the original research question referred to looking for the best storage conditions (highest expected intact score). Based on the parameter estimates, you can find that these conditions are: -80 degrees Celsius, one week and the ZYM medium. The mean intact score is then estimated as

$$74.883 + 16.35 + 10.15 - 11.45 = 89.933$$

e. (2 points) As a follow up to the previous question: The R output from the previous question is perhaps insufficient for comparing all storage conditions with one another in a pairwise-comparison fashion. What contrasts are missing? Construct the contrast matrix that can be used to construct tests for the missing comparisons. For each line of the matrix, state the null and alternative hypotheses.

Answer

This question has been removed. Almost noone answered the question. So for the calculation of your final score, this question is ignored.

f. (2 points) Consider the following output, formulate a conclusion, and give an interpretation to the interval. You may assume that all model assumptions hold.

```r
m5<-lm(intact~temp+time*medium,data=DNA)
predict(m5, newdata = data.frame(temp=-18,time="week",medium="ZYM"),
        interval="confidence")
```

```
##        fit      lwr      upr
## 1 78.63633 76.84005 80.43261
```

Answer

This was supposed to be an easy question. A very similar example is in the course notes.

From the output we read that we estimate that for DNA stored at a temperature of -18 degrees Celcius for one week in the ZYM medium, the average intact score is 78.64. The corresponding 95% confidence interval goes from 76.84 to 80.43. In other words, with a probability of 95% we expect that the mean intact score to be somewhere between 76.84 and 80.43.

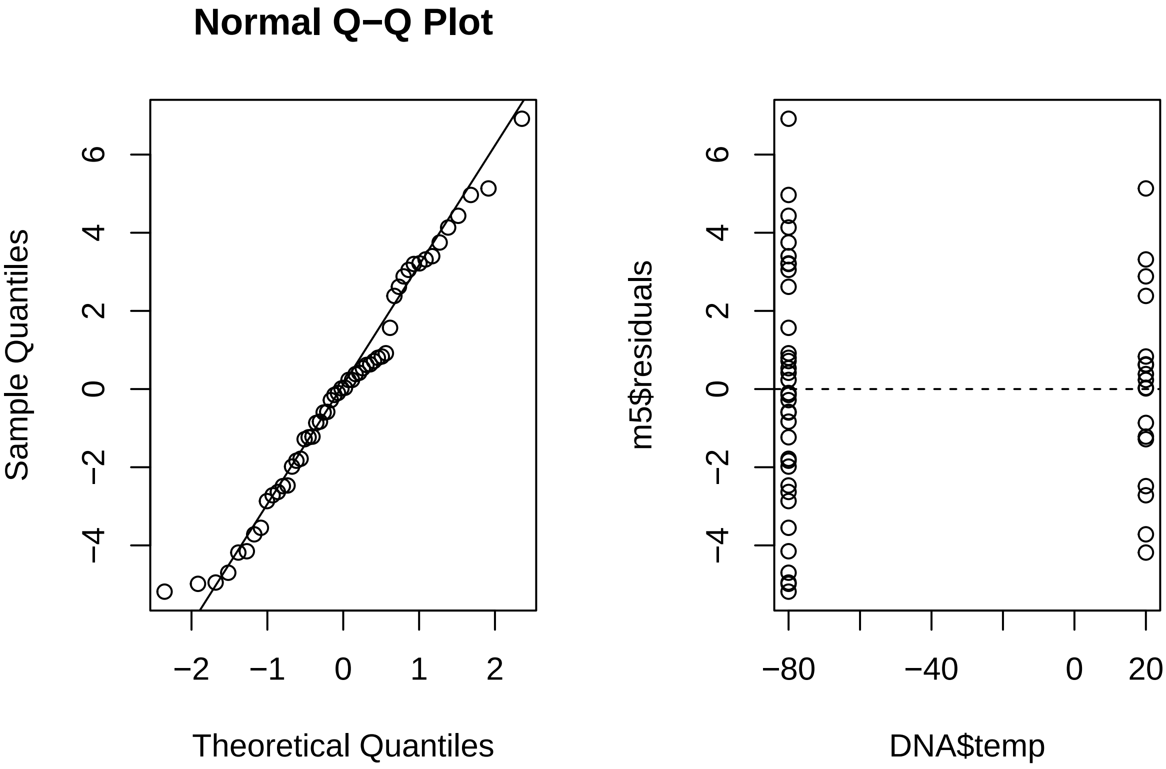
Note that this is all about the **average** intact score, because the R code includes *interval="confidence"*. So it is **not** about prediction.

You could also have included something about the usual way we interpret confidence intervals: if the experiment would be repeated many times, and each time the 95% confidence interval is calculated, then 95% of these intervals will contain the true mean intact score of DNA stored at a temperature of -18 degrees Celcius for one week in the ZYM medium.

g. (2 points) Below you see a normal QQ-plot and a residual plot. What can you conclude about the normality of the residuals and about the linearity of the effect of temperature?

```r
par(mfrow=c(1,2))
qqnorm(m5$residuals)
qqline(m5$residuals)

plot(DNA$temp,m5$residuals)
abline(h=0,lty=2)
```



```r
par(mfrow=c(1,1))
```

Answer

QQ-plots have been interpreted several times in the course notes. Almost all students interpreted the QQ-plot correctly. The residual plot, on the other hand, was not interpreted to the full extent by many students.

The QQ-plot shows only a minor deviation from normality (in the middle of the plot), but since the sample size is 54, this deviation will not cause problems. The central limit theorem will make the parameter estimators approximately normally distributed.

A residual plot can typically be used for assessing the linearity of the model (correctness of the model for the conditional mean outcome). In this respect the plot does not show a systematic deviation, **but** there is only data for two temperatures and through two points you can always draw a straight line. As a consequence, based on the plot we cannot assess the linearity of the model. There should have been observations at intermediate temparatures.

h. (2 points) Does this study allow causal conclusions to be drawn from the data analysis? Motivate your answer.

Answer

The answer is actually very simple. The description of the design of the study mentions: *All human samples are randomly assigned to the storage conditions.* . In the course notes (Chapter 4), it is written that randomised studies allow for causal conclusions.

Question 2

Consider again a study on the storage conditions (as in Question 1), but this time only storage time is considered (7, 14, 21 and 28 days). All experiments make use of the glycerol medium and a storage temperature of -80 degrees Celcius. The data is now in the data frame *DNA2*. In the R code you will see how three dummy regressors are constructed (T1, T2 and T3). The dummy coding system, however, is different from what we have done in the course.

Give an interpretation to the parameters estimates of the model fit of model *m6* (no need to report on the standard errors and hypothesis tests). Based on your analysis, do you think that a linear regression model with storage time as a continuous regressor would give a good model fit?

```
str(DNA2)
```

```
## 'data.frame':    12 obs. of  2 variables:
##  $ time  : num  7 14 21 28 7 14 21 28 7 14 ...
##  $ intact: num  95 85 81.3 71 90.1 89.8 81.6 83 87.1 87.1 ...
```

```
head(DNA2)
```

```
##   time intact
## 1    7   95.0
## 2   14   85.0
## 3   21   81.3
## 4   28   71.0
## 5    7   90.1
## 6   14   89.8
```

```
DNA2$T1<-DNA2$T2<-DNA2$T3<-0

DNA2$T1[DNA2$time==7]<--1
```

```r
DNA2$T1[DNA2$time==14]<-1

DNA2$T2[DNA2$time==7]<--1
DNA2$T2[DNA2$time==21]<-1

DNA2$T3[DNA2$time==7]<--1
DNA2$T3[DNA2$time==28]<-1
```

```r
m6<-lm(intact~T1+T2+T3,data=DNA2)
summary(m6)
```

```
##
## Call:
## lm(formula = intact ~ T1 + T2 + T3, data = DNA2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5667 -1.4250 -0.4167  2.1000  5.4333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   84.508      1.133  74.611 1.16e-12 ***
## T1             2.792      1.962   1.423  0.19254
## T2            -2.075      1.962  -1.058  0.32109
## T3            -6.942      1.962  -3.538  0.00764 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.924 on 8 degrees of freedom
## Multiple R-squared:  0.7069, Adjusted R-squared:  0.5971
## F-statistic: 6.433 on 3 and 8 DF,  p-value: 0.01587
```

Answer

Note that the dummy coding is different from what we used in the course notes, and, as a consequence the interpretation of the model parameters will be different. You are expected to know how to interpret any dummy coding system. First, for the 0/1 dummy coding used throughout the lectures, section 2.9 of the course notes explains how we go from dummies to the interpretation of the parameters. Second, in HW2 you were asked to work with another (new) dummy coding system, and again the same method was used to get to the interpretation of the parameters. We will use this method again for solving this question.

First, we write the model that corresponds to m6:

$$E(Y \mid T_1, T_2, T_3) = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3$$

with T_1, T_2 and T_3 the dummies as defined in the R code. You were asked to interpret the β-parameters.

For DNA stored for a period of 7 days, the model becomes

$$E(Y \mid 7 \text{ days}) = E(Y \mid T_1 = -1, T_2 = -1, T_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3$$

For DNA stored for a period of 14 days, the model becomes

$$E(Y \mid 14 \text{ days}) = E(Y \mid T_1 = 1, T_2 = 0, T_3 = 0) = \beta_0 + \beta_1$$

For DNA stored for a period of 21 days, the model becomes

$$E(Y \mid 21 \text{ days}) = E(Y \mid T_1 = 0, T_2 = 1, T_3 = 0) = \beta_0 + \beta_2$$

For DNA stored for a period of 28 days, the model becomes

$$E(Y \mid 28 \text{ days}) = E(Y \mid T_1 = 0, T_2 = 0, T_3 = 1) = \beta_0 + \beta_3$$

First, note that average of the four conditional means equals β_0. Hence, β_0 has the interpretation of the average intact score (averaged over the four storage times). Let's call this the *overall* mean intact score. We then find

- $\beta_1 = E(Y \mid 14 \text{ days}) - \beta_0$, i.e. the difference between the mean intact score for a storage time of 14 days and overall mean intact score.
- $\beta_2 = E(Y \mid 21 \text{ days}) - \beta_0$, i.e. the difference between the mean intact score for a storage time of 21 days and overall mean intact score.
- $\beta_3 = E(Y \mid 28 \text{ days}) - \beta_0$, i.e. the difference between the mean intact score for a storage time of 28 days and overall mean intact score.

Finally, you were asked whether a model with time as a continuous regressor would give a good model fit. From the R output we find the following parameter estimates: $\hat{\beta}_0 = 84.5$, $\hat{\beta}_1 = 2.8$, $\hat{\beta}_2 = -2.1$ and $\hat{\beta}_3 = -6.9$. Thus, the estimated mean intact scores for storage times of 7, 14, 21 and 28 days are 90.7, 87.3, 82.4 and 77.6. If you would plot these against the regressor values 7, 14, 21 and 28, you get to see a nice linear relationship. So the answer to the question is: yes, a model with time as a continuous regressor would give a good model fit.

Question 3

The Holm-Bonferroni method is known to be better than the Bonferroni method. In what sense is the former better than the latter?

answer

The answer to this question comes straight from the course notes.

Question 4

The following R code is for a simulation study. It contains three different scenarios. The simulation results are stored in the matrices *results1*, *results2* and *results3*. At the end of the R code, the column means of these matrices are shown. What do you learn from the 1st column, and what do you learn from the second column?

For your information:

- If $U \sim U[a,b]$, then $\text{Var}(U) = \frac{1}{12}(b-a)^2$ and $E(U) = b-a$
- If $W \sim \exp(\lambda)$ (λ is the rate parameter), then $\text{Var}(W) = \frac{1}{\lambda^2}$ and $E(W) = \frac{1}{\lambda}$

```r
set.seed(2698)

### simulation scenario 1

db<-data.frame(x=rep(c(0,0,0,0,0,0,0,0,0,10,10,10,10,10,10,10,10,10),5))

N<-1000
results1<-matrix(nrow=N,ncol=2)
for(i in 1:N) {
  db$y<-100+2*db$x+runif(90,min=-30,max=30)
  m<-lm(y~x,data=db)
  sm<-summary(m)

  results1[i,]<-c(sm$coefficients[2,1],
                  sm$coefficients[2,2]^2)
}

### simulation scenario 2

db<-data.frame(x=rep(c(0,0,0,0,0,0,2,2,2,8,8,8,10,10,10,10,10,10),5))

N<-1000
results2<-matrix(nrow=N,ncol=2)
for(i in 1:N) {
  db$y<-10+2*db$x+rnorm(90,sd=sqrt(300))
  m<-lm(y~x,data=db)
  sm<-summary(m)
  results2[i,]<-c(sm$coefficients[2,1],
                  sm$coefficients[2,2]^2)
}

### simulation scenario 3

db<-data.frame(x=rep(c(0,0,0,4,4,4,2,2,2,8,8,8,6,6,6,10,10,10),5))

N<-1000
results3<-matrix(nrow=N,ncol=2)
for(i in 1:N) {
  db$y<-10+2*db$x+(rexp(90,rate=sqrt(1/300))-sqrt(300))
  m<-lm(y~x,data=db)
  sm<-summary(m)

  results3[i,]<-c(sm$coefficients[2,1],
                  sm$coefficients[2,2]^2)
}


colMeans(results1)
```

```
## [1] 2.0032895 0.1332293
```

```r
colMeans(results2)
```

```
## [1] 1.9859135 0.1702882
```

```r
colMeans(results3)
```

```
## [1] 1.9767906 0.2833595
```

Answer

We've seen several simulation studies in the course notes. The overall structure of this simulation study is not much different from what we have seen many times: repeated sampling from a linear regression model, and each time fitting a linear regression model, and storing parameter estimates in a vector After the for-loop the averages of the elements in the vectors are calculated. We have seen that such averages are approximately equal to the expectations of the corresponding estimator.

Here we are looking at three scenarios. The scenarios differ in two aspects: (1) the distribution for the error term, and (2) the values of the regressor x. The error distributions are: normal, uniform and exponential. You were also given expressions for the mean and variance of these distributions. These will tell you that all error distributions have mean equal to zero and variance equal to 300. From the theory in the course notes we know that least squares parameter estimators of the β-parameters in a linear model are unbiased, irrespective of the distribution of the error term. The same holds true for MSE as an estimator of the variance of the error term. We only assumed that the error terms are i.i.d.

The column means of the vectors results1, results2 and results3, are approximations of $E(\hat{\beta}_1)$ and $Var(\hat{\beta}_1)$. From the theory (see previous paragraph), we know that we have used unbiased estimators. For $E(\hat{\beta}_1)$ this is easy to see, because $\beta_1 = 2$ is given in the R-code. On the other hand, $Var(\hat{\beta}_1)$, is not known to us, but the column means tell us that for scenarios 1, 2 and 3, $Var(\hat{\beta}_1)$ is (approximately) equal to 0.133, 0.170 and 0.283. As argued earlier, these differences cannot be attributed to the different error distributions, and so there is only one aspect remaining: the values of the regressor!

Going from scenario 1, over scenario 2, to scenario 3, the values of x show more and more a uniform distribution, i.e. for scenario 1 only the extreme values 0 and 10 are considerd, and for the other scenarios also intermediate values are included. So we can conclude that the variance of $\hat{\beta}_1$ increases when the distribution of the regressor becomes more uniform.