Exam Linear Models: Feedback

January 24, 2022

Question 1

Introduction

A clinical trial is set up to assess the effect of a vaccin on the blood serum concentration of immunoglobin G (IgG), two weeks after vaccination. In this study 10 children (< 12 year) and 10 adults participated. Two doses of the vaccin were considered: low and high. All adults received the high dose, and all children received the low dose.

The data is in the *IgG* dataset which contains the following variables:

- **dosis**: defined as 0 for low dose, and 1 for high dose
- **adult**: defined as 0 for child and 1 for adult
- **IgG**: concentration of IgG in the blood serum (mg/ml)

The main research question is: is there on average an effect of the dose on the IgG concentration?

For all analyses you may use a significance (or FWER) level of 5% and a 95% confidence level.

```
str(IgG)
```

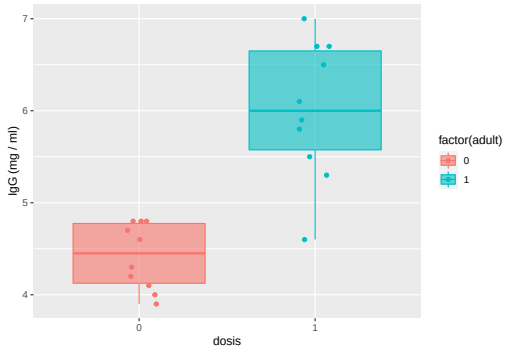
```
## 'data.frame':    20 obs. of  3 variables:
##  $ dosis: num  0 0 0 0 0 1 1 1 1 1 ...
##  $ adult: num  0 0 0 0 0 1 1 1 1 1 ...
##  $ IgG  : num  4.7 4.1 4.8 3.9 4.6 6.7 7 6.7 6.1 5.9 ...
```

```
head(IgG)
```

```
##    dosis adult IgG
## 1      0     0 4.7
## 2      0     0 4.1
## 3      0     0 4.8
## 4      0     0 3.9
## 5      0     0 4.6
## 11     1     1 6.7
```

The data are explored by means of the following figure.

```
library(ggplot2)
ggplot(IgG, aes(x=factor(dosis), y=IgG, fill=factor(adult), colour=factor(adult))) +
  labs(x="dosis", y="IgG (mg / ml)") +
  geom_point(position = position_jitterdodge()) + geom_boxplot(alpha=0.6)
```



Questions

Answer the questions in the order that they are presented to you (i.e. as if you don't know the results of the analyses in the subsequent questions).

a. (2 points) Consider the following R output. What do you conclude about the effect of dose on the average IgG concentration?

```
m1<-lm(IgG~dosis,data=IgG)
summary(m1)
```

```
##
## Call:
## lm(formula = IgG ~ dosis, data = IgG)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.410 -0.345 -0.010  0.380  0.990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4200     0.1848  23.922 4.29e-15 ***
## dosis         1.5900     0.2613   6.085 9.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5843 on 18 degrees of freedom
## Multiple R-squared:  0.6729, Adjusted R-squared:  0.6547
## F-statistic: 37.03 on 1 and 18 DF,  p-value: 9.475e-06
```

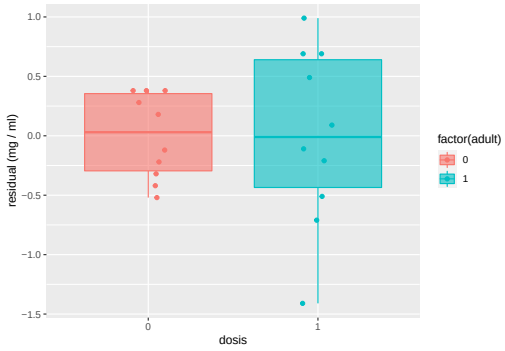
```
m2<-lm(IgG~adult,data=IgG)
summary(m2)
```

```
##
## Call:
## lm(formula = IgG ~ adult, data = IgG)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.410 -0.345 -0.010  0.380  0.990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4200     0.1848  23.922 4.29e-15 ***
## adult         1.5900     0.2613   6.085 9.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5843 on 18 degrees of freedom
```

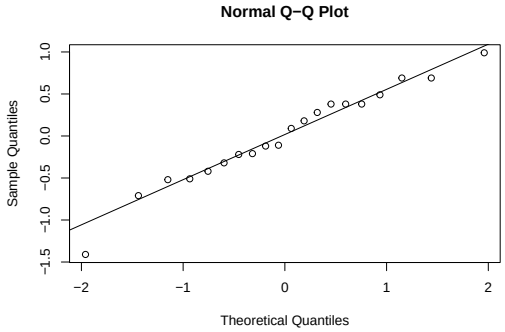
```
## Multiple R-squared:  0.6729, Adjusted R-squared:  0.6547
## F-statistic: 37.03 on 1 and 18 DF,  p-value: 9.475e-06
```

```
IgG$e1<-residuals(m1)
IgG$e2<-residuals(m2)
```

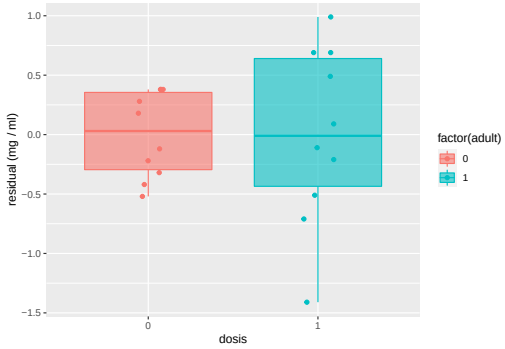
```
ggplot(IgG, aes(x=factor(dosis), y=e1, fill=factor(adult), colour=factor(adult))) +
  labs(x="dosis", y="residual (mg / ml)") +
  geom_point(position = position_jitterdodge()) + geom_boxplot(alpha=0.6)
```



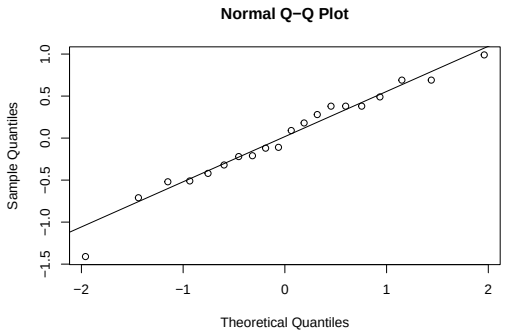

```
qqnorm(IgG$e1)
qqline(IgG$e1)
```




```
ggplot(IgG, aes(x=factor(dosis), y=e2, fill=factor(adult), colour=factor(adult))) +
  labs(x="dosis", y="residual (mg / ml)") +
  geom_point(position = position_jitterdodge()) + geom_boxplot(alpha=0.6)
```




```
qqnorm(IgG$e2)
qqline(IgG$e2)
```



Answer:

From the R output we can see that the parameter estimate for the effect of dose and the parameter estimate for the effect of age (adult/child) coincide. This comes as no surprise. In the design all adults receive the high dose and all children the low dose. There is thus a complete confounding between the effects of age and dose. If you would construct the design matrices for these two models, they would coincide.

Hence, the parameter estimate of 1.59 measures the composite effect of a high dose for adults as compared to a lose dose for children.

The diagnostic residual plots can be interpreted, keeping in mind the restrictions explained in the previous paragraphs.

b. (2 points) The same study is repeated in 5 other hospitals, but in these other hospitals the adults were randomised over the two doses. The hospital of previous question (question 1a) is numbered as hospital 1, and the 5 new hospitals are numbered 2 up to 6. The next R output shows the structure of the dataset.

```
str(IgG2)
```

```
## 'data.frame':    120 obs. of  4 variables:
##  $ dosis   : num  0 0 0 0 0 1 1 1 1 1 ...
##  $ adult   : num  0 0 0 0 0 1 1 1 1 1 ...
##  $ IgG     : num  4.7 4.1 4.8 3.9 4.6 6.7 7 6.7 6.1 5.9 ...
##  $ hospital: Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
table(IgG2$dosis,IgG2$adult,IgG2$hospital)
```

```
## , ,  = 1
##
##
##      0  1
##   0 10  0
##   1  0 10
##
## , ,  = 2
##
##
##      0  1
##   0 10  5
##   1  0  5
##
## , ,  = 3
##
##
##      0  1
##   0 10  5
##   1  0  5
##
## , ,  = 4
##
##
##      0  1
##   0 10  5
##   1  0  5
##
## , ,  = 5
##
##
##      0  1
##   0 10  5
```

```
##   1  0  5
##
## , ,  = 6
##
##
##      0  1
##   0 10  5
##   1  0  5
```

Based on the following model fit, what do you conclude now about the effect of dose on the mean IgG concentration? Would you prefer to remove *hospital* from the model? Motivate your answer. (You do not have to worry about the model assumptions in answering this question; the model assumptions are the topic of the next question.)

```
library(car)
```

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

```
m<-lm(IgG~dosis+adult+hospital,data=IgG2)
Anova(m, type="III")
```

```
## Anova Table (Type III tests)
##
## Response: IgG
##              Sum Sq  Df F value    Pr(>F)
## (Intercept) 309.997   1 449.046 < 2.2e-16 ***
## dosis       139.281   1 201.755 < 2.2e-16 ***
## adult        39.129   1  56.680 1.401e-11 ***
## hospital      1.370   5   0.397      0.85
## Residuals    77.319 112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m)
```

```
##
## Call:
## lm(formula = IgG ~ dosis + adult + hospital, data = IgG2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26946 -0.47417  0.03323  0.55554  2.03300
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.367692   0.206114  21.191  < 2e-16 ***
## dosis        3.207077   0.225786  14.204  < 2e-16 ***
## adult       -1.512462   0.200895  -7.529  1.4e-11 ***
## hospital2    0.121769   0.268739   0.453    0.651
```

```
## hospital3   -0.188231   0.268739  -0.700    0.485
## hospital4    0.001769   0.268739   0.007    0.995
## hospital5    0.106769   0.268739   0.397    0.692
## hospital6    0.101769   0.268739   0.379    0.706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8309 on 112 degrees of freedom
## Multiple R-squared:  0.6664, Adjusted R-squared:  0.6455
## F-statistic: 31.96 on 7 and 112 DF,  p-value: < 2.2e-16
```

Answer:

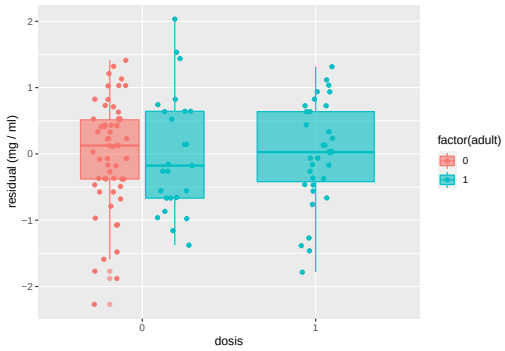
According to the description of the desing of the study, hospital is obviously a stratification factor (randomisation restriction, because subjects are randomized within hospitals) and hence it is required to be in the model.

This time we have adults in the low and the high dose groups, but we still only have children within low dose. With this data it will be possible to assess the main effect of dose, but only under the assumption of no-interaction effect of dose and age. An interaction effect cannot be assessed with this design because for children the effect of dose cannot be estimated (they only received the low dose) and hence it is not possible to estimate any difference in the effects of dose between children and adults.

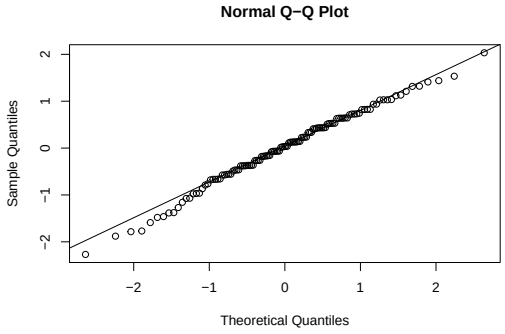
Under the assumption of no interaction, we estimate that for adults and for children, on average the Ig concentration is 3.2 mg/ml higher in the high dose group as compared to the low dose group. This comes with a standard error of 0.23 mg/ml. This effect is significant at the 5% level of significance, with a two-sided p-value < 0.0001.

c. (2 points) Next you see some model diagnostics. What are the model assumptions and based on this R output, do these model assumptions hold true? Do you need other graphs/statistics or extra model fits in order to assess some of the model assumptions?

```
IgG2$e<-residuals(m)
ggplot(IgG2, aes(x=factor(dosis), y=e, fill=factor(adult), colour=factor(adult))) +
  labs(x="dosis", y="residual (mg / ml)") +
  geom_point(position = position_jitterdodge()) + geom_boxplot(alpha=0.6)
```




```
qqnorm(IgG2$e)
qqline(IgG2$e)
```




```
vif(m)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## dosis    1.830769  1        1.353059
## adult    1.753846  1        1.324329
## hospital 1.076923  5        1.007438
```

Answer:

The model assumptions (and their assessment):

- normality of the error term. Based on the QQ-plot this assumption seems reasonable. Moreover, the total sample size is large enough to rely on the central limit theorem and so the normality assumption is not very important.

This assumption only makes sense if the model for the mean outcome is correctly specfied (see next point).

- correctness of the model for the mean outcome. As explained before, we assume absence of the interaction effect. However, this assumption cannot be verified based on the data,

because we do not have data for children with high dose. Thus we cannot compare the dose effect for adults (which can be estimated) with the dose effect of children.

- independence of the observations: this assumption can in general not be assessed based on the data. Based on the description of the study design we cannot assess this assumption.
- homoskedasticity: we can use the boxplots of the residuals to assess this assumption, but only under the assumption that the model for the mean outcome is correct (see earlier). The boxplots do not indicate a serious deviation from the constant-variance assumption.

d. (2 points) For the model fit of question b, give the mathematical model formulation (factor effects model notation).

Answer:

for $i = 1, 2$, $j = 1, 2$, $k = 1, \ldots, 6$, $l = 1, \ldots, n_{ijk}$,

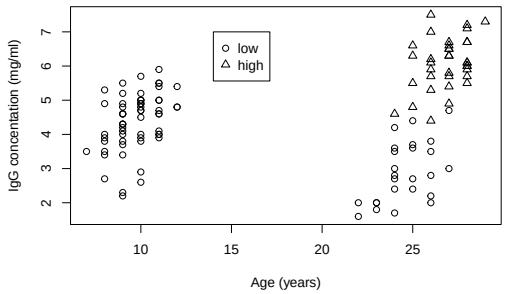
$$Y_{ijk} = \mu + \tau_i + \alpha_j + \gamma_k + \varepsilon_{ijkl}$$

with

- μ: the intercept
- τ_i: the main effect of the factor *adult*. The reference group is *child* ($i = 1$), for which $\tau_i = 0$.
- α_j: the main effect of the factor *dose*. The reference group is *low* ($j = 1$), for which $\alpha_j = 0$.
- γ_k: the main effect of the factor *hostpital*. The reference group is *hospital1* ($k = 1$), for which $\gamma_k = 0$.
- ε_{ijkl}: the error term, which is assumed to be i.i.d. $N(0, \sigma^2)$.

e. (2 points) Since the actual age of the participants was also recorded, we will now replace the 0/1 dummy *adult* with the age as a continuous covariate. Based on the following model fits, what do you conclude now about the effect of dose on the mean IgG concentration? Motivate your choice of model. You may assume that that all model assumptions hold true.

```
plot(IgG2$age,IgG2$IgG,pch=IgG2$dosis+1,
     xlab="Age (years)",
     ylab="IgG concentation (mg/ml)")
legend(14,7,legend=c("low","high"),pch=c(1,2),col=1)
```




```
m1<-lm(IgG~dosis+age+hospital,data=IgG2)
Anova(m1,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: IgG
##              Sum Sq  Df  F value    Pr(>F)
## (Intercept) 280.891   1 358.0608 < 2.2e-16 ***
## dosis       119.742   1 152.6394 < 2.2e-16 ***
## age          28.586   1  36.4398 2.088e-08 ***
## hospital      1.602   5   0.4084    0.8421
## Residuals    87.862 112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = IgG ~ dosis + age + hospital, data = IgG2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29002 -0.46757  0.01351  0.57958  2.09886
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.156511   0.272507  18.922  < 2e-16 ***
## dosis        3.220935   0.260705  12.355  < 2e-16 ***
## age         -0.085274   0.014126  -6.037 2.09e-08 ***
```

```
## hospital2    0.057015   0.285950   0.199    0.842
## hospital3   -0.252985   0.285950  -0.885    0.378
## hospital4   -0.007557   0.287232  -0.026    0.979
## hospital5    0.054806   0.286220   0.191    0.848
## hospital6    0.100970   0.287455   0.351    0.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8857 on 112 degrees of freedom
## Multiple R-squared:  0.6209, Adjusted R-squared:  0.5972
## F-statistic: 26.21 on 7 and 112 DF,  p-value: < 2.2e-16
```

```
m2<-lm(IgG~dosis*age+hospital,data=IgG2)
Anova(m2,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: IgG
##              Sum Sq  Df  F value    Pr(>F)
## (Intercept) 286.106   1 385.0092 < 2.2e-16 ***
## dosis         2.265   1   3.0474  0.083632 .
## age          30.585   1  41.1572 3.553e-09 ***
## hospital      1.611   5   0.4334  0.824416
## dosis:age     5.376   1   7.2340  0.008258 **
## Residuals    82.486 111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = IgG ~ dosis * age + hospital, data = IgG2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2447 -0.4792  0.1077  0.5814  2.1366
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.238051   0.266953  19.622  < 2e-16 ***
## dosis       -6.003407   3.438994  -1.746  0.08363 .
## age         -0.088549   0.013803  -6.415 3.55e-09 ***
## hospital2    0.060688   0.278313   0.218  0.82778
## hospital3   -0.283825   0.278545  -1.019  0.31044
## hospital4   -0.070781   0.280544  -0.252  0.80128
## hospital5    0.041714   0.278615   0.150  0.88126
## hospital6    0.003561   0.282108   0.013  0.98995
```

```
## dosis:age    0.345131   0.128320   2.690  0.00826 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.862 on 111 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6185
## F-statistic: 25.11 on 8 and 111 DF,  p-value: < 2.2e-16
```

Answer:

With the actual age as a continuous regressor the interaction effect of age and dose can be included (the design matrix is now of full column rank). Two models are fitted: one without and one with the interaction effect of age and dose. Adapting the hierarchical modelling approach, we first look at the output of the latter model. This tells us that the interaction effect is significant at the 5% level of significance ($p = 0.0083$) and so we need to use this model for formulating conlcusions. From the estimated parameters, we conclude that the effect of dose (high dose versus low dose) is increased with 0.35 mg/ml (se 0.13mg/ml) for every increase of age with one year. For example, for a child of the age of 10 years, the effect of dose on the average IgG concentration is $-6 + 0.35 \times 10 = -2.5$ mg/ml and for an adult of the age of 25 the effect of dose is $-6 + 0.35 \times 25 = +2.75$ mg/ml. However, when looking at the graph of IgG concentration versus age, we see that for children of the age of about 10 years old, we do not have data for the high dose. Thus the interpretation based on the model, for people of age 10 and of oge 25, is an extrapolation and thus relies heavily on the correctness of the model. There is no data to assess the model assumption beyond the scope of the data.

Yet another way of interpreting the interaction effect is to look at the effect of age for low and high dose separately. For low dose the effect of age is negative (-0.089 mg/ml, se 0.014 mg/ml), and for high dose the effect of age is estimated as $-0.089 + 0.35 = 0.211$ mg/ml.

However, looking at the scatter plot, we can see that the model assumptions probably do not hold true (although you were asked to answer the question as if all model assumptions hold true). Within the low dose group of children we see a positive age effect, and also within the adult group we see a positive effect of age. The data thus rather suggests that it is not only the age that matters, but the data also suggests that the immune system of adults and chidren are fundamentally different.

...

f. (2 points) Consider the data analyses from questions a and e. How do you now interpret the results of the model fit of question a?

The main message is that we need data from children treated with the high dose to say something sensible about the effect of dose for children. As long as there is no data, we can only form conclusions based on model assumptions, but that is dangerous (against scientific integrity).

g. (2 points) For this question only the data of the adults are used, and the participants are grouped into age classes. This can be seen from the next R code.

```
IgG3<-IgG2[IgG2$adult==1,]
IgG3$age.group<-cut(x=IgG3$age,breaks=c(20,22,24,26,28,30))
table(IgG3$age.group)
```

```
##
## (20,22] (22,24] (24,26] (26,28] (28,30]
##       2      12      22      23       1
```

Consider the following R code and output. What do you conclude from the output of *m.mcp*? Be complete and precise.

```
library(multcomp)
```

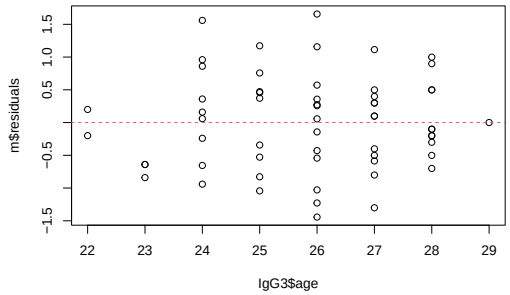
```
## Loading required package: mvtnorm
## Loading required package: survival
## Warning: package 'survival' was built under R version 4.1.1
## Loading required package: TH.data
## Warning: package 'TH.data' was built under R version 4.1.1
## Loading required package: MASS
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
m<-lm(IgG~dosis+age.group,data=IgG3)
summary(m)
```

```
##
## Call:
## lm(formula = IgG ~ dosis + age.group, data = IgG3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44313 -0.53197  0.02844  0.46059  1.65687
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.8000     0.5263   3.420  0.00120 **
## dosis              2.6149     0.2600  10.058 5.57e-14 ***
## age.group(22,24]   0.8404     0.5689   1.477  0.14542
## age.group(24,26]   1.4282     0.5677   2.516  0.01488 *
## age.group(26,28]   1.7864     0.5979   2.988  0.00422 **
```

```
## age.group(28,30]   2.8851     0.9480   3.043  0.00361 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7443 on 54 degrees of freedom
## Multiple R-squared:  0.8404, Adjusted R-squared:  0.8256
## F-statistic: 56.87 on 5 and 54 DF,  p-value: < 2.2e-16
```

```
plot(IgG3$age,m$residuals)
abline(h=0,lty=2,col=2)
```




```
m.mcp<-glht(m,linfct=mcp(age.group="Tukey"))
m.mcp$linfct
```

```
##                    (Intercept) dosis age.group(22,24] age.group(24,26]
## (22,24] - (20,22]            0     0                 1                 0
## (24,26] - (20,22]            0     0                 0                 1
## (26,28] - (20,22]            0     0                 0                 0
## (28,30] - (20,22]            0     0                 0                 0
## (24,26] - (22,24]            0     0                -1                 1
## (26,28] - (22,24]            0     0                -1                 0
## (28,30] - (22,24]            0     0                -1                 0
## (26,28] - (24,26]            0     0                 0                -1
## (28,30] - (24,26]            0     0                 0                -1
## (28,30] - (26,28]            0     0                 0                 0
##                    age.group(26,28] age.group(28,30]
## (22,24] - (20,22]                 0                0
## (24,26] - (20,22]                 0                0
## (26,28] - (20,22]                 1                0
## (28,30] - (20,22]                 0                1
## (24,26] - (22,24]                 0                0
## (26,28] - (22,24]                 1                0
## (28,30] - (22,24]                 0                1
## (26,28] - (24,26]                 1                0
## (28,30] - (24,26]                 0                1
## (28,30] - (26,28]                -1                1
## attr(,"type")
## [1] "Tukey"
```

```r
summary(m.mcp,test=adjusted("bonferroni"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = IgG ~ dosis + age.group, data = IgG3)
##
## Linear Hypotheses:
##                          Estimate Std. Error t value Pr(>|t|)
## (22,24] - (20,22] == 0   0.8404     0.5689   1.477   1.0000
## (24,26] - (20,22] == 0   1.4282     0.5677   2.516   0.1488
## (26,28] - (20,22] == 0   1.7864     0.5979   2.988   0.0422 *
## (28,30] - (20,22] == 0   2.8851     0.9480   3.043   0.0361 *
## (24,26] - (22,24] == 0   0.5878     0.2929   2.007   0.4977
## (26,28] - (22,24] == 0   0.9460     0.3417   2.768   0.0771 .
## (28,30] - (22,24] == 0   2.0447     0.8106   2.523   0.1463
## (26,28] - (24,26] == 0   0.3582     0.2417   1.482   1.0000
## (28,30] - (24,26] == 0   1.4569     0.7702   1.892   0.6392
## (28,30] - (26,28] == 0   1.0987     0.7607   1.444   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- bonferroni method)
```

Answer:

An additive two-way ANOVA is fitted with main effects of the factors *dose* and *age.group*. The summary of the *m.mcp* object shows the results of a multiple (or pairwise) comparisons of means between all age groups. In particular, for each pair of age groups, the null hypothesis is tested that the mean IgG concentration in the two age groups is the same, while controlling for the dose (two-sided hypothesis tests).

The p-values are adjusted with the Bonferroni method, and the nominal FWER is set to 5%. At this 5% level of the FWER, we conclude that, while controlling for dose, the following age groups show signicant differences in the mean IgG concentration:

- IgG concentration among subjects of age (26-28] is on average 1.79 mg/ml (se 0.60 mg/ml) larger then among subjects of age (20,22] that recieved the some dose of the vaccin (adj. p = 0.0422)
- IgG concentration among subjects of age (28-30] is on average 2.89 mg/ml (se 0.95 mg/ml) larger then among subjects of age (20,22] that recieved the some dose of the vaccin (adj. p = 0.0361)

Also the non-significant results should be reported, including effect size estimates with standard error (or CI) and the adjusted p-values.

Overall this analysis suggests that among adults that receive the same dose, the effect of age

on the mean IgG concentration is positive.

...

h. (2 points) Is age (as continuous regressor) a confounder in the context of this study?

Answer:

In the description of the design of the study, we read that this is randomised study, but with the complication that children only received the low dose.

Among adults the situation is clear: the randomisation breaks the association between *age* and *dose* and hence age cannot be a confounder.

Among children, there is no variability in the dose, and so the dose can also not be affected by age. So here the age is also not a confounder.

However, being child or adult depends on the age, and so when the age is small (< 12 years, which is used as the threshold), the dose is always low. So from this perspective there is causal relation between age and dose. However, this causal relation is *by-design*, i.e. it is part of the study design. We typically only use the term *confounder* in observational studies. We sometimes say that effects are *confounded*, as in question a, where the effects of age (as a binary factor) and dose were completely confounded, i.e. their effects cannot be distinguished from one another.

(note that for the exam, you can earn points from any of these perspectives, as long as your arguments are OK)

i. (2 points) In the output ANOVA type III sum of squares were used. What is/are the advantage(s) of type III sum of squares as compared to type I or II?

Answer:

Some advantages:

- they do not depend on the order in which the terms were added to the model (in contrast to type I)
- they measure the variability in the outcome that can be attributed to a factor or regressor, given that all the other factors/regressors are in the model. This corresponds to the interpretation we give to parameters in the model.

Question 2

(4 points) The following R code is for a simulation study. You are asked to explain what is illustrated. Do these empirical results agree with what the theory tells you (explain).

```
beta.hat<-c()
var.beta.hat<-c()

x<-2:11
db<-data.frame(x=x,y=NA)

set.seed(62547)
for(i in 1:10000) {
  db$y<-exp(4+0.3*x)+rnorm(10,sd=25)
  db$y2<-log(db$y)
  m<-lm(y2~x,data=db)

  beta.hat<-c(beta.hat,coef(m)[2])
  var.beta.hat<-c(var.beta.hat,summary(m)$coef[2,2]^2)
}

mean(beta.hat)
```

```
## [1] 0.3030301
```

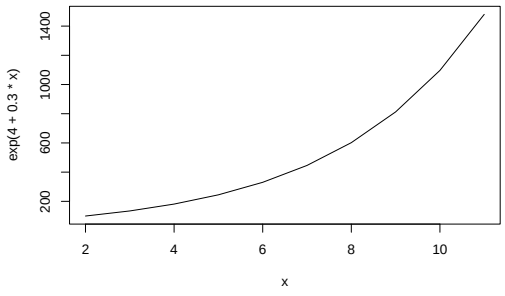
```
mean(var.beta.hat)
```

```
## [1] 0.000175066
```

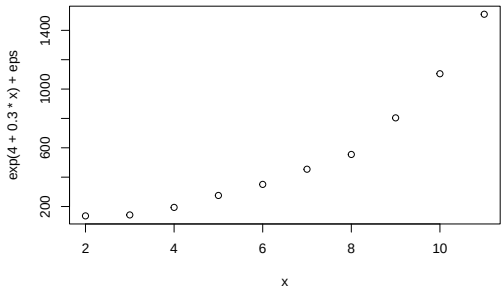
```
var(beta.hat)
```

```
## [1] 0.0003322823
```

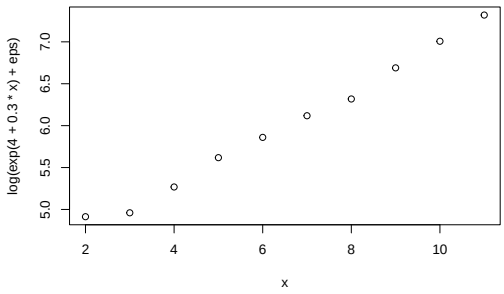
```
x<-2:11
plot(x,exp(4+0.3*x),type="l")
```




```
eps<-rnorm(10,sd=25)
plot(x,exp(4+0.3*x)+eps)
```




```
plot(x,log(exp(4+0.3*x)+eps))
```



Answer:

The regressor takes the values $x = 2, 3, \ldots, 11$, and the outcomes Y_i^* are simultated with the model

$$Y_i^* = \exp(\alpha_0 + \alpha_1 x_i) + \varepsilon_i^*$$

with $\alpha_0 = 4$, $\alpha_1 = 0.3$ and ε_i^* i.i.d. $N(0, 0.25^2)$.

Next the outcomes are log-transformed

$$Y_i = \log(Y_i^*)$$

and with these transformed outcomes the following regression model is fitted:

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

with ε_i i.i.d. $N(0, \sigma^2)$.

Before discussing the rest of the simulation study, it is instructional to think about the meaning of this setting. The outcomes are generated with a non-linear model (the mean outcome shows an exponential relation with the regressor – as can also be seen from the graphs). The fit of a linear model to this data would result in a residual plot that shows a deviation that indicates a non-linear effect of the regressor. One of the classical "solutions'' is to apply a transformation. Here the outcomes are transformed with a log-transformation. The graphs show that after the log-transformation there appears to be a linear relationship between the mean (transformed) outcome and the regressor. However, we are not primarily interested in the parameter estimates of this linear model (the β parameters), but rather in the parameters of the nonlinear model (the α parameters). It is tempting to consider the estimates of the β parameters as the estimates of the α parameters, because if $\log(y) = \beta_0 + \beta_1 x$, then $y = \exp(\beta_0 + \beta_1 x)$ and hence $\alpha_0 = \beta_0$ and $\alpha_1 = \beta_1$. However, these equalities are NOT implied by the statistical models. The model for the log-transformed outcome gives an expression for the mean log-transformed outcome, i.e. $E(\log(Y)) = \beta_0 + \beta_1 x$. If we not exponentiate both

sides of the equality we find $\exp(E(\log(Y))) = \exp(\beta_0 + \beta_1 x)$, but $\exp(E(\log(Y))) \neq E(Y)$, and therefore we cannot easily backtransform and we cannot consider the estimates of the β parameters as the estimates of the α parameters. This is further illustrated in the simulation study.

In each simulation run the β_1 parameter is estimated, as well as its variance. Let $\hat{\beta}_{1(j)}$ and $S^2_{\beta j}$ denote these two estimates in simulation run $j = 1, \ldots, 10000$.

When the simulations are finished (after 10000 runs), the average of the 10000 $\hat{\beta}_{1(j)}$s is computed. This is an approximation of $E(\hat{\beta}_1)$. Thus we have $E(\hat{\beta}_1) \approx 0.303$. This is very close to the true value of α_1 ($\alpha_1 = 0.3$). So the simulation study suggests that $\hat{\beta}_1$ is an unbiased estimator of α_1. However, this is in general not true.

Based on the 10000 $\hat{\beta}_{1(j)}$s we can also approximate the variance of the estimator $\hat{\beta}_1$. From the R output we read $\mathrm{Var}(\hat{\beta}_1) \approx 0.00033$. The estimator S^2_β taken from the results of the *lm* function is supposed to be a good estimator of this variance. However, when we compute the average of the 10000 $S^2_{\beta j}$s we conclude $E(S^2_\beta) \approx 0.00018$, which is almost a factor 2 smaller than the true variance. Hence, the variance estimator S^2_β is a biased estimator of the true variance. As a consequence, all statistical inference on the β_1 (or α_1) parameter will be invalid.

...