

EXERCISES FOR DAY 4

Exercise 4.1: *Multilinear regression and multicollinearity.*

Multilinear regression refers to the situation where several continuous covariates are used together as explanatory variables in a regression analysis. When doing a multilinear regression you should be aware of the potential pitfalls that may arise if the covariates are *multicollinear*. The purpose of this exercise is to exemplify these pitfalls. This exercise should be done without opening RStudio, but if you want to try the R code yourselves you may find the dataset and the R program in the files `wage.txt` and `solution4_1.R`.

We consider data taken from *The Current Population Survey (CPS)* made in the US in 1985. The dataset contains observations of the following 6 variables for 532 persons:

edu: length of the persons total education in years.

sex: gender of the persons (1=female, 0=male).

exper: length of the persons working experience in years.

wage: wage in US dollars per hour.

age: age of the persons in years.

occup: profession (1=management, 2=trade, 3=office, 4=service, 5=craft, 6=other).

The following R code fits a multilinear regression of wage on length of education, length of working experience, and age among women working with craftsmanship¹:

```
> summary(lm(wage~edu+exper+age,data=subset(wage,(sex==1)&(occup==5))))
```

Residuals:

Min	1Q	Median	3Q	Max
-6.5109	-2.9453	-0.6629	2.0672	14.0105

¹Model validation would reveal that it is better to model the logarithm of the wage. But in order not to give the impression that responses should always be log-transformed, which indeed isn't the case, and also to keep the interpretations of the parameters estimates as simple as possible, we will not transform the response variable. This is ok since the emphasis of this exercise is *multicollinearity* and not *model validity*.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-15.4931	6.6457	-2.331	0.024 *
edu	0.7059	0.8524	0.828	0.412
exper	-0.6247	0.8723	-0.716	0.477
age	0.6775	0.7964	0.851	0.399

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.593 on 48 degrees of freedom

Multiple R-squared: 0.3097, Adjusted R-squared: 0.2666

F-statistic: 7.179 on 3 and 48 DF, p-value: 0.0004461

From the model summary we see that neither of the 3 explanatory variables are close to significance. However, the multilinear regression still explains 30.97 pct (ie. the R^2) of the variation in the wages, and taken together the 3 explanatory variables are highly significant ($p=0.0004461$).

- Compare the **summary**-output to the statements made above and confirm that the reporting of hypothesis tests and R^2 is correct.
- Is there an interpretation of the sign of the estimates for the 3 slopes? E.g. do craftswomen earn more if they have more working experience? Or is it impossible to make such an interpretation in this case?
- An automated backward model reduction would proceed by removing **exper** being the least significant variable. However, what are the arguments for removing **age** instead?

The fit of the multilinear regression after removal of **age** is given on the next page. Please consider the following questions:

- What has happened to the p-values for **edu** and **exper**?
- What has happened to the sign of the slope of **exper**? Do you think that the positive sign makes more sense? Why/why not?

```
> summary(lm(wage~edu+exper,data=subset(wage,(sex==1)&(occup==5))))
```

Call:

```
lm(formula = wage ~ edu + exper, data = subset(wage, (sex ==  
1) & (occup == 5)))
```

Residuals:

Min	1Q	Median	3Q	Max
-5.9828	-3.0854	-0.6495	1.7550	14.1748

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-11.85350	5.07118	-2.337	0.0235 *
edu	1.38007	0.31307	4.408	5.68e-05 ***
exper	0.11552	0.06237	1.852	0.0700 .

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 4.581 on 49 degrees of freedom

Multiple R-squared: 0.2993, Adjusted R-squared: 0.2707

F-statistic: 10.47 on 2 and 49 DF, p-value: 0.0001641

The following output from R shows that `edu` and `exper` may be considered uncorrelated in the subpopulation of craftswomen. Does this have any implication for the interpretation of the slope estimates on `edu` and `exper` given above? Why/why not?

```
> with(subset(wage, (sex==1)&(occup==5)), cor.test(edu, exper))
```

Pearson's product-moment correlation

data: edu and exper

t = -1.0381, df = 50, p-value = 0.3042

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.4022055 0.1329212

sample estimates:

cor

-0.1452478

And what if `edu` and `exper` actually are negatively correlated, i.e. if working experience in general is shorter for craftswomen with a longer education?

Multicollinearity means that some of the covariates explains the same property in the experimental units. E.g. if you have a long education as well as long working experience, then you necessarily also will have a comparably high age. So we will only need two of the three variables `edu`, `exper`, `age` in order to characterize these properties of a person. Right? To decide which two of these variables provides the “correct” explanation can not be

done based on statistics, but relies on the interpretation of the variables. When there is multicollinearity among the explanatory variables, the p-values may change from non-significant to highly significant and the estimates may change sign after model reduction. That there indeed is multicollinearity in the present dataset may be seen from the following analysis²

```
> summary(lm(age~edu+exper,data=wage))
```

```
Call: lm(formula = age ~ edu + exper, data = wage)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
	-3.8507	-0.3801	-0.0122	0.4081	2.1230

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.09160	0.19182	31.76	<2e-16 ***
edu	0.98494	0.01281	76.91	<2e-16 ***
exper	1.05558	0.00271	389.51	<2e-16 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.7235 on 529 degrees of freedom
```

```
Multiple R-squared:  0.9966, Adjusted R-squared:  0.9966
```

```
F-statistic: 7.793e+04 on 2 and 529 DF,  p-value: < 2.2e-16
```

Please do the following:

- Comment on the R^2 -value as well as the significance tests.
- What is the interpretation of the estimate of the intercept?
- What is the interpretation of the null hypothesis that the slopes on `edu` and `exper` both equal 1?
- What is the interpretation of the error term, and the RMSE=0.7235?

²A few more suggestions for the identification of multicollinearity may be found in `solution4_1.R`.

Exercise 4.2: *ANCOVA, statistical modeling, and more.*

In this exercise we investigate the world records for outdoor running distances. The records were taken from the website <http://www.iaaf.org> of the International Association of Athletics Federation on May 7, 2011. We want to examine the dependence of the record (*time*) on the *distance*, and to examine the difference between men and women. The purpose of this exercise is to give a non-trivial example of the choices needed in making simple statistical models with good interpretations. Reference: Based on exercise 8.2 from Anders Tolver & Helle Sørensen: *Lecture notes for Applied Statistics*.

The following items guide you through such an analysis step-by-step:

- Read the dataset available in the text file `WR2011.txt` into R (in a data frame called *wr*), and have a look at the variables:
 - Please note, that the distances are more or less doubled between consecutive running disciplines. Thus, the running distances are almost equidistant on a logarithmic scale.
 - The variable *DOB* contains the data-of-birth of the record holder. The variables *Place* and *Date* contain the place and date of the record. These variables will not be used in this exercise.
 - The variable *bend* I made myself, and it will be used later. This variable quantifies how many times longer than 1500 meters the running distance in question is, and it is set to 1 if the distance is shorter than 1500 meters.
 - Make sure that the variables *time*, *distance* and *bend* are numerical, and that *sex* is a categorical factor.
- Make a plot of *time* against *distance* using the code:

```
library(ggplot2)
ggplot(wr) + geom_point(aes(x=distance,y=time,col=sex))
```

This plot corresponds to the relationship:

$$\text{time} = \alpha + \beta * \text{distance}$$

The parameter β describes the running velocity. Thus, in this model the running velocity is the same no matter the distance. This is clearly unrealistic. Do you agree?

- Make a plot of $\log(\text{time})$ against $\log(\text{distance})$. This plot corresponds to the relationship:

$$\log(\text{time}) = \alpha + \beta * \log(\text{distance})$$

Taking the exponential function on both sides we find:

$$\text{time} = \exp(\alpha) * \text{distance}^\beta$$

It is not obvious that this is a good model. But what do you think looking at the plot?

- Make the following linear regression:

```
m1 <- lm(log(time)~sex+log(distance)+sex:log(distance),data=wr)
```

In this model both the alpha and the beta parameter depend on the gender:

$$\log(\text{time}_i) = \alpha(\text{sex}_i) + \beta(\text{sex}_i) * \log(\text{distance}_i) + \text{error}_i$$

Such a model is called an ANCOVA (ANalysis of COVariance). The ANCOVA will allow us to compare the records for men and women.

- Is this ANCOVA model valid?
- Look in particular at the *residual plot* (called “Residuals vs Fitted” if you use `plot(m1)`). In my opinion there is a bend at observations number 6 and 22. One way to identify these observation numbers is to use the `identify()` function. To use this the residual plot should fill the entire graphics window, so we will make it again. Try the R code³

```
par(mfrow=c(1,1))  
plot(predict(m1),residuals(m1))  
identify(predict(m1),residuals(m1))
```

and use the mouse to click on the points where you think the bend is positioned. After you are done finish the identifier as signified in the graphics window (in Windows you should press the *Esc*-key).

- Check that observations number 6 and 22 correspond to the 1500 meter distance for men and women, respectively. Thus, there appears to be a difference between short running distances (less than 1500 meters) and long running distances (more than 1500 meters).

³The line `par(mfrow=c(1,1))` is only necessary if you did `par(mfrow=c(2,2))` before.

- In order to allow the regression lines to bend at 1500 meters we include the variable *bend* in the ANCOVA model:

$$\log(\text{time}_i) = \alpha(\text{sex}_i) + \beta(\text{sex}_i) * \log(\text{distance}_i) + \gamma(\text{sex}_i) * \log(\text{bend}_i) + \text{error}_i$$

Fit this model!

- Is the extended ANCOVA model valid? No not really...
- Refit the extended ANCOVA without using observations number 1, 11, 12, 14, 15, 17, 27, 28, 30, and 31. This may be done using the option

```
data=wr[-c(1,11,12,14,15,17,27,28,30,31),]
```

in the call to `lm()`. Does this improve the model validity?

- Which distances does the removed observations represent? Do you think it is fair to remove the world records for these running distances from the present analysis? Why/why not?

Hint: This may require some knowledge about athletics.

Remark: In any case, in the remaining of this exercise you should remove these observations from the analysis!

- Use the function `step()` to do model selection based on the Akaike Information Criterion.
- The AIC based selection should result in the model, where the interactions between *sex* and $\log(\text{distance})$, $\log(\text{bend})$ are removed:

$$\log(\text{time}_i) = \alpha(\text{sex}_i) + \beta * \log(\text{distance}_i) + \gamma * \log(\text{bend}_i) + \text{error}_i$$

Thus, the difference between men and women is only via the alpha parameter. Realize that the number

$$\exp(\alpha(\text{woman}) - \alpha(\text{man}))$$

quantifies how much slower the women run than the men.

- Give a 95% confidence interval for the relative running speed of women compared to men.

Hint: The contrast $\alpha(\text{woman}) - \alpha(\text{man})$ is called *sexwoman* in the parametrization used by R.

Exercise 4.3: *Incomplete block experiment.*

The following experiment was carried out by H. Wolffhechel, KVL, in 1986. The purpose was to compare 12 sphagnum lots with respect to water and air content. Each of these was applied to four pots with small cucumber plants. The pots were placed in one of six watering troughs, each containing eight pots. The experimental design and the volume (water and air content), in percent, is given in the following table (dataset available in file `sphagnum.txt`) for each pot.

Sphagnum lot	Watering trough						Mean
	1	2	3	4	5	6	
1	37.0		44.6		42.5	47.1	42.80
2			49.0	50.5	51.0	44.8	48.83
3		34.6	42.7	41.8	37.8		39.23
4	45.3	42.7	47.7	42.8			44.63
5	32.1	38.5		32.0		31.6	33.55
6	34.3	33.3		34.0	22.6		31.05
7	32.3		28.1	28.1	32.3		30.20
8	38.9	36.5		39.7		34.8	37.48
9	33.9		31.4	32.1		23.0	30.10
10		39.7	41.8		43.5	33.8	39.70
11		41.1	38.1		31.1	37.9	37.05
12	35.9	7.5			36.2	25.5	26.28

Make and validate the additive model for *volume*, i.e. the model including only the main effects of the two explanatory variables *lot* and *trough*:

1. Remember that the variables *lot* and *trough* should be used as factors. You can achieve this by using the `factor()` function in the call to `lm()`. However, in this exercise I recommend that you change the type of the variables in the data frame at the beginning of your R code:

```
sphagnum      <- read.delim("sphagnum.txt")
sphagnum$lot  <- factor(sphagnum$lot)
sphagnum$trough <- factor(sphagnum$trough)
```

2. You probably would want to remove the “extreme” observation `volume=7.5` for `(lot,trough)=(12,2)`. Can you use the validation plots in R to identify the number of this observation?

Find the estimated marginal means of volume for the 12 different sphagnum lots using the additive model. Why are these em-means different from

the raw means listed in the above table? Do you prefer the raw means or the em-means? And why?

Remark: The **emmeans**-package may be used to compute and compare the em-means. The statistical computations done in the **emmeans**-package are based on standard errors extracted from the model objects. Suppose e.g. that your model is available in an **lm**-object called *m2*, and try the following R code (and think about what the code does):

```
# load library
library(emmeans)

# make and plot em-means
emmeans(m2,~lot)
plot(emmeans(m2,~lot))

# Tukey grouping of em-means
# Note: the p-values are adjusted for multiple testing,
#       but the confidence intervals are not adjusted!
CLD(emmeans(m2,~lot))

# Remark: The author of the emmeans-package, Russell Lenth,
# does not like the CLD, and intend to remove this
# functionality from the package. I disagree on this.
# Luckily, in the future you may use multcomp::cld() in
# place of CLD().

# As an replacement for the CLD() functionality Russell
# proposes the following plot. Personally, I find this
# display to be too busy. But what do you think?
pwpp(emmeans(m2,~lot))

# An alternative to pwpp() is to find and plot
# simultaneous confidence intervals.
# Note, however, that replacing hypothesis tests by
# looking for overlap between confidence intervals
# may be misleading.
confint(emmeans(m2,~lot),adjust="tukey")
plot(confint(emmeans(m2,~lot),adjust="tukey"))
```

Remark: An alternative is to use the **multcomp** package. However, in my opinion the syntax for that package is much more difficult to learn. So I

clearly recommend the `emmeans` package. However, there are some pitfalls. For instance try the code

```
plot(confint(emmeans(m2,~lot),adjust="tukey"))  
plot(confint(emmeans(m2,~lot)),adjust="tukey")
```

You might think that these two lines would produce the same graphical visualization, i.e. simultaneous confidence intervals. But only the first line does that! The difference is, that `adjust="tukey"` is an option to `confint.emmGrid()` in the first line, but to `plot.emmGrid()` in the second line. And this option does two different things in the two functions. For more information see

```
?confint.emmGrid  
?plot.emmGrid
```

Exercise 4.4: *Linear regression.*

In a field experiment the concentration of phosphor available for plant growth was measured for each of 18 plants. Furthermore, the concentration of inorganic phosphor was chemically determined and the concentration of an organic phosphor component was measured for each plant. The primary interest is to describe the concentration of phosphor available as a function of the concentrations of inorganic and organic phosphor. Thus, we have the following *Table of Variables*:

Variable	Type	Usage
inorganic	continuous	fixed effect
organic	continuous	fixed effect
available	continuous	response

The dataset is shown below, and it is also available in the text file `phosphor.txt`:

inorganic	organic	available
0.4	53	64
0.4	23	60
3.1	19	71
0.6	34	61
4.7	24	54
1.7	65	77
9.4	44	81
10.1	31	93
11.6	29	93
12.6	58	51
10.9	37	76
23.1	46	96
23.1	50	77
21.6	44	93
23.1	56	95
1.9	36	54
26.8	58	168
29.9	51	99

Analyse the data, i.e. answer the generic questions:

- Is there an association?
- What is the association?
- Can the conclusions be trusted?

Hints and suggestions: If you do a multilinear regression of **available** on **inorganic** and **organic**, then one of the observations is not well-modelled. You may either decide to remove this observation (what is the most easy way to do this in R?). Alternatively, you may e.g. try a logarithmic transformation of the response variable.

(Reference: Exercise 8.4 from Anders Tolver & Helle Sørensen: *Lecture notes for Applied Statistics*.)

End of exercises.