Prof. Dr. M. Kateri,
Dr. W. Herff,
L. Kaufmann, M.Sc.

RWTH Aachen, SS 2022
Release date: April 25th, 2022
Solution: Week starting from May 2nd, 2022

# Applied Data Analysis

## R-Laboratory 4

**Model Parsimony – Testing in Linear Models – Analysis of Variance**

**Useful packages and functions:**

- `set.seed()`
- `runif()`
- `lm()`
- `shapiro.test()`
- `rgl`
- `rgl::plot3d()`
- `rgl::planes3d()`
- `summary()`
- `tapply()`
- `car`
- `car::leveneTest()`
- `par()`
- `cooks.distance()`

## Task 13

For values of $x$ in $[0, 100]$, suppose the linear model $(Y \mid X = x) \sim \mathcal{N}(\mu(x), \sigma^2)$ holds with

$$\mathrm{E}(Y \mid X = x) = \mu(x) = 45 + 0.1x + 5 \cdot 10^{-4}x^2 + 5 \cdot 10^{-7}x^3 + 5 \cdot 10^{-11}x^4 + 5 \cdot 10^{-13}x^5$$

and $\sigma = 10$.

(a) Set a seed to 2020 and initialize two numeric vectors `vec.delta.simple` and `vec.delta.correct` of length 100.

(b) Repeat the following procedure 100 times using a loop.

(i) Generate 25 observations from the model with $X$ uniformly distributed on $[0, 100]$.

(ii) Fit a "simple" model with $\mu^{(0)}(x) = \beta_0 + \beta_1 x$ and afterwards the "correct" model with $\mu^{(1)}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$.

(iii) For the first five iterations construct plots showing the data, the true relationship, and both model fits.

(iv) For each model, summarize the quality of the model fit by the mean

$$\Delta_j = \frac{1}{25} \sum_{i=1}^{25} |\hat{\mu}^{(j)}(x_i) - \mu(x_i)|, \qquad j \in \{0, 1\}. \qquad (+)$$

Store the values of $\Delta_j$, $j = 0, 1$, in (+) in the two numeric vectors `vec.delta.simple` and `vec.delta.correct`.

(c) By comparing the values of $\Delta_j$, $j = 0, 1$, in the vectors vectors `vec.delta.simple` and `vec.delta.correct` and the plots of (iii), which model do you prefer? Explain what this Task illustrates about model parsimony.

## Task 14

(a) Load the `.RData` file of the pre-processed data of *Survey1* (Task 7) into the R workspace.

(b) Create a regression model with the approach

$$\texttt{DimSelf} = d + a\ \texttt{DimEmotion} + b\ \texttt{DimBody} \qquad (++)$$

for parameters $a, b, d \in \mathbb{R}$.

(c) Analyze, if the model assumptions are sufficiently satisfied:

(i) Create the following plots (you can directly apply the function `plot` to the fitted model for the first four plots) and interpret them:

i. Residuals versus fitted values

ii. $\sqrt{|\text{Standardized residuals}|}$ versus fitted values

iii. Quantiles of the standardized residuals versus the expected quantiles of the standard normal distribution (called QQ-Plot)

iv. Standardized residuals versus leverage

v. Cook's distance.

(ii) Test on level $\alpha = 0.05$, if there is evidence against the assumption of normally distributed residuals.

*Hint:* Use a Shapiro-Wilk-Test with the R function `shapiro.test`.

(d) Create a 3d scatterplot with `DimEmotion` on the x-axis, `DimBody` on the y-axis and `DimSelf` on the z-axis. Add the regression surface of the model (++) to the plot.

*Hint:* Use the functions `plot3d` and `planes3d` from the package rgl.

(e) In the model (++) test the hypotheses

$$H_0 : b = 0 \quad \text{versus} \quad H_1 : b \neq 0$$

on the significance level $\alpha = 0.05$. Is the null hypothesis rejected?

## Task 15

Load the Scottish hill races data contained in the *Races.dat* data file, which can be found in RWTHmoodle, into your R workspace.

(a) At first step, fit a linear model that predicts women's record time (timeW) using both distance of the course and climb in elevation as explanatory variables and interpret the resulting coefficients.

(b) Fit the linear model that predicts timeW using distance as the sole explanatory variable. Compare your results with (a).

Next, a linear model can predict men's record times from women's record times. Fit this linear model.

(c) Show the scatterplot and report the prediction equation. Predict men's record time for the Highland Fling, for which timeW = 490.05 minutes.

(d) Find and interpret the correlation between timeM and timeW.

(e) We could impose the natural constraint that when timeW $= 0$, then timeM $= 0$. Fit the model $E(Y_i) = \beta x_i$. (In R, you can use a command such as `lm(timeM ~ -1+ timeW, data=Races)`.) Interpret the estimated slope.

## Task 16

Load the *Florida.dat* data file, which can be found in RWTHmoodle, into your R workspace.

(a) Estimate the effect of education, marginally and conditional, on urbanization.

(b) Compute and interpret the correlation between crime rate and education

## Task 17

Using the *UN.dat* data file, which can be found in RWTHmoodle, construct a multiple regression model predicting Internet using all other variables. Use the concept of multicollinearity to explain why adjusted $R^2$ is not dramatically greater than when GDP is the sole predictor. Compare the estimated GDP effect in the bivariate model and the multiple regression model and explain why it is so much weaker in the multiple regression model.

```
set.seed(2022)

#vec.delta.simple = runif(100, 0.0, 100)

#vec.delta.correct = runif(100, 0.0, 100)

N=25
# N=1500
vec.delta.simple=rep(0,100)
vec.delta.correct=rep(0,100)
vec.delta.correct.poly=rep(0,100)
for(i in 1:100){
  x=runif(N,0,100)
  #x=runif(N,-50,100)

  mu=45+0.1*x+0.0005*x^2+5e-7*x^3+5e-11*x^4+5e-13*x^5
  y=mu+rnorm(N,sd=10)
  model.correct=lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5))

  # better use poly
  model.correct.poly=lm(y~poly(x,degree=5))
  fitted.vals.poly=predict(model.correct.poly,newdata=data.frame(x=x))

  model.simple=lm(y~x)

  if(i<6){
    # data
    plot(x,y)
    param1=model.correct$coefficients
    param2=model.simple$coefficients
    # true relationship
    curve(45+0.1*x+0.0005*x^2+5e-7*x^3+5e-11*x^4+5e-13*x^5,add=TRUE,col="blue")
    # correct model
    curve(param1[1]+param1[2]*x+param1[3]*x^2+param1[4]*x^3+param1[5]*x^4+param1[6]*x^5,add=TRUE,col="red")
    # simple model
    curve(param2[1]+param2[2]*x,add=TRUE,col="green")
    # correct model fitted with poly
    x.grid=seq(min(x),max(x),length.out = 100)
    y.pred=predict(model.correct.poly,newdata=data.frame(x=x.grid))
    lines(x.grid,y.pred, col="orange")
  }
  vec.delta.correct[i]=mean(abs(model.correct$fitted.values-mu))
  vec.delta.correct.poly[i]=mean(abs(fitted.vals.poly-mu))
  vec.delta.simple[i]=mean(abs(model.simple$fitted.values-mu))
}

mean(vec.delta.simple)
mean(vec.delta.correct)
```

```
#######################
########TASK14########
#######################
#b) create a regression model with the approach
# DimSelf = d + a DimEmotion + b DimBody

DimSe = as.integer(survey_task14$DimSelf)
DimEmo = as.integer(survey_task14$DimEmotion)
DimBo = as.integer(survey_task14$DimBody)

DimSelf.fit = lm(DimSe ~ DimEmo + DimBo)

#c)-1 plot
#i) residual vs fitted
residuals = DimSelf.fit$residuals

fitted_values = DimSelf.fit$fitted.values

plot(residuals, fitted_values)

#ii) standardized residual vs fitted
DimSelf.rstandard = rstandard(DimSelf.fit)
plot(sqrt(abs(DimSelf.rstandard)), fitted_values)

#iii) quantiles of the standardized residuals vs the expected quantiles

qqnorm(DimSelf.rstandard, col = "blue")
qqline(DimSelf.rstandard, col = "red")

#iv)
# leverage: a measure of how far away the independent variable values
# of an observation are from of the other observations

# get the leverage
leverages = hatvalues(DimSelf.fit)
plot(DimSelf.rstandard, leverages)

plot(cooks.distance(DimSelf.fit))

# another option
par(mfrow =c(2,2))
plot(DimSelf.fit)
par(mfrow=c(1,1))
plot(cooks.distance(DimSelf.fit))

#d) Null hypothesis b = 0 => DimBody

hypo.fit = lm(DimSe ~ DimEmo)

Y = DimSe

ones2 = rep(1, nrow(survey_task14))

B0 = cbind(ones2, DimEmo)

B = cbind(B0, DimBo)

Q = B %*% solve(t(B)%*%B, t(B))
Q0 = B0 %*% solve(t(B0)%*%B0, t(B0))
n = nrow(survey_task14)
r = 3
r0 = 2
numer2 = Y %*% (Q-Q0) %*% Y / (r-r0)
denom2 = Y %*% (diag( nrow(survey_task14))) %*% Y / (n-r)
F_statistic = numer2 / denom2

reject = F_statistic > 0.05
```

```
#######################
#########TASK15########
#######################

#a)
race_task15 = read.table("R-Lab-Datasets/Races.dat", header = TRUE)

timeW.fit = lm(timeW ~ climb + distance, data = race_task15)

#b)

timeW.fit2 = lm(timeW ~ distance, data = race_task15)
#
# Analysis of Variance Table
#
# Model 1: timeW ~ distance
# Model 2: timeW ~ climb + distance
# Res.Df   RSS Df Sum of Sq      F    Pr(>F)
# 1     66 30686
# 2     65 12675  1     18011 92.36 4.223e-14 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
# reject the null hypothesis that climb is not sigfinicant

distance = race_task15$distance
climb = race_task15$climb
ones_task15 = rep(1, nrow(race_task15))

mens_dat = cbind(ones_task15, climb, distance)

predicted_timeM = mens_dat %*% timeW.fit$coefficients
# which(race_task15["timeW"] == 490.05)
# [1] 41
# > predicted_timeM[41]
# [1] 456.1487
#

# #d)
# > cor(race_task15$timeM, race_task15$timeW)
# [1] 0.9958732

cor(Races$timeW,Races$timeM)
# we have strong positive correlation
# this indicates that in a race where the males need more time, the females will also need more time

#e)
fit.timeMW = lm(timeM ~ -1 + timeW, data = race_task15)
```

```
#####################
########TASK16########
#####################

urban.fit = lm(Crime ~ HS, data = florida_task16)
urban.fit2 = lm(Crime ~ Urban + HS, data = florida_task16)

# Adjusting for urbanization, the effect of education changes sign:
#the crime rate tends to decrease as education increases.
# Hence, we have a positive marginal correlation when we ignore urbanization
# This phenomenon of association reversal between
#marginal and conditional associations is called "simpsons paradox"

#####################
########TASK17########
#####################

Internet.fit = lm(Internet ~
                  GDP+
                  HDI+
                  GII+
                  Fertility+
                  CO2+
                  Homicide+
                  Prison, data = UN_task17)

Internet.GDP.fit = lm(Internet ~ GDP,data = UN_task17)

# the p-value of GDP is essentially 0 when it is the sole explanatory variable
# when we add the other variables, the SE of the GDP effect increases
#from 0.1217 to 0.290680
# the p-value increases to 0.13856  ( > 0.05) when we add the other variables
#to the model
# the dramatic change in the SE for GDP and the
#lack of statistical significance for the conditional effects is
#due to the high correlation

# because of the multicollinearity,
#we can attain nearly as large and R^2 value in predicting
#the response with a reduced set of explanatory variables
# the fact that the effect of GDP is so different
#in the multiple regression model compared with the bivariate model is
#caused by the multicollinearity,
# meaning that GDP "overlaps" considerably with other explanatory variables.
#Hence, the effects on the multiple regression model are not significant even
# if the effect is highly significant marginally
```