MASTER OF SCIENCE
IN ENGINEERING

REPETITION EXAMINATION FS18

PREDICTIVE MODELING

Date : 12th July 2018 , 13:15-15:15

First Name: ______________________________

Family Name: ______________________________

School: ______________________________

Stick Number: ______________________________

Problem	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
max. points	14	12	34	21	22	17	120
Achieved points							

Please open the file `Lastname_Firstname.R` in the folder `Austausch` on the desktop of the Lernstick environment and save it according to your name.
Good Luck!
Dr. Klaus Frick and Dr. Mirko Birbaumer

General information

1. Write your name on the first page and on supplementary pages you use.
2. The questions may be answered in German or in English.
3. Please answer directly on the question sheet. You may also use the back side.
4. If you need supplementary sheets, please use a separate one for every question. Write your name on every supplementary sheet.
5. Material allowed on the desk during the exam:
   a) Paper, Pen and Ruler
   b) Lecture Notes Predictive Modeling with a summary
   c) R Reference Card (with your comments)
   d) Calculator
   e) Statistical Software **R** within Lernstick environment
6. All solutions to the exam exercises need to be written in a complete and clear manner on paper.
7. You execute all **R** functions that you use for solving the exam problems from an **R** script file that you save according to your last name and first name on the USB stick in the Austausch folder.
8. No question concerning the problems will be answered during the exam. If you don't understand a problem, make an assumption and explain it in your solution. It will be considered by the grader.
9. Communication with others during the exam is forbidden. Mobile phones must be turned off.
10. Don't write in red. This color is reserved for grading.
11. Don't use a pencil for answering the questions.
12. Portions of answers that have been crossed out won't be considered, even if the deleted part is correct.

Problem 1: Multiple Choice .... (14 Points)

Janet and Madlen let fall a ball from different heights (`height`) and measure the time (`time`) until the ball reaches the ground. The following model was fitted to the data

$$\texttt{height}_i = \beta_0 + \beta_1 \texttt{time}_i + \beta_2 \texttt{time}_i^2 + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The model was fitted using the software **R** which yields the following summary output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.987      2.572     ???      ???
time             ???      2.326   -0.82     0.42
time.squared   5.340      0.492   10.86  2.9e-12

Residual standard error: 1.31 on 32 degrees of freedom
Multiple R-squared:  0.994,  Adjusted R-squared:  0.994
```

Only **one** answer is the correct one: mark with a **single cross** the correct answer. If you cross the correct answer, you will get 2 **points** per question. If you cross the wrong answer, **1 point** is subtracted from the total number of points you have achieved. At minimum you will get 0 points for problem 1.

(1.) How many measurements are contained in the data set?

a) 30
b) 32
c) 34
d) 35

(2.) Is the null hypothesis $H_0 : \beta_0 = 0$ rejected at the 5 %-level (the alternative hypothesis is given by $H_A : \beta_0 \neq 0$)?

a) Yes
b) No.
c) It is not possible to draw a conclusion on the basis of the available information.

(3.) What is the value of the estimate $\widehat{\beta}_1$?

a) -2.84
b) -1.91
c) -0.35
d) -0.34

(4.) Which of the following intervals is a precise two-sided 99 % confidence interval for β_2?

a) $[3.99, 6.69]$
b) $[4.34, 6.34]$
c) $[4.36, 6.32]$
d) $[4.59, 5.99]$

(5.) Janet and Madlen let fall a ball from an altitude of 20 m. They measure 2 seconds for the ball to reach the ground. What is the error according to the model ($\equiv$ residual)?

a) 1.4
b) 0.5
c) -1.4
d) -0.5

(6.) Is the regression model (1) a linear model?

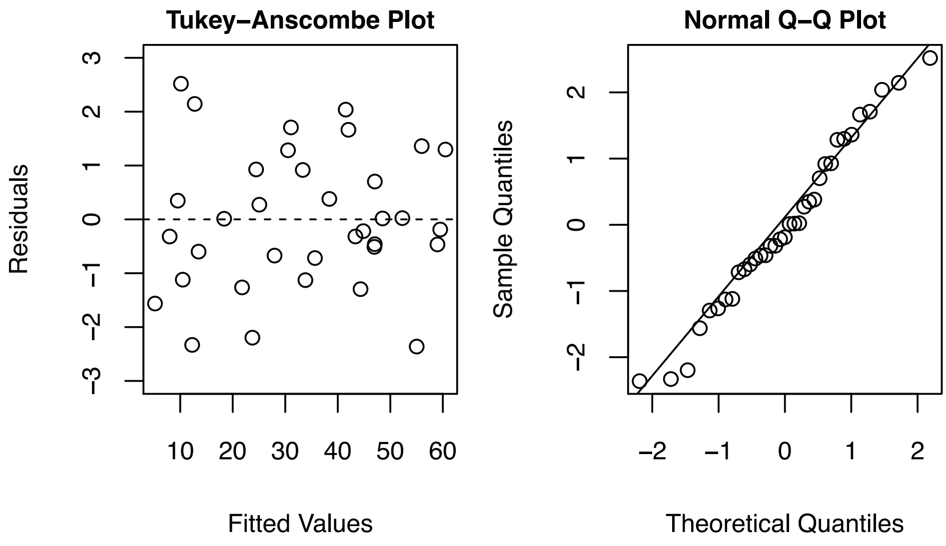
a) No, because the model contains a quadratic predictor term `time.squared`.
b) No, because the predictor variable `time` shows up twice in the regression model.
c) Yes, because the model is linear with respect to the regression coefficients.
d) Yes, because the response variable `height` is not transformed.

(7.) Have a look at the residual plots. Which one of the following statements is true?

a) All model assumptions for the error term ε are plausible.
b) The model assumption concerning the normal distribution of the error term ε is plausible, however the variance of the error term is not constant.
c) The model assumption relying on a constant variance of the error term ε is satisfied, however the error terms do not follow a normal distribution.
d) The model assumptions concerning the normal distribution and the constant variance of the error term ε are violated.

Problem 2: Variable Selection . . . . (12 Points)

In a medical study, the lung function of 25 cystic fibrosis patients was measured. The measured quantities include the maximum expiratory pressure (`pemax`), the age (`age`), the sex (`sex`), the height (`height`), the weight (`weight`), the body mass (%



percent of normal) (**bmp**) and four additional lung function quantities of the patients. A regression model was fitted for the response variable **pemax** and the remaining variables as predictors.

a) (6 points) The analysis of the full regression model is shown in the following **R**-output :

```
> pemax.lm <- lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
> summary(pemax.lm)
Call:
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 +
    rv + frc + tlc)

Residuals:
    Min      1Q  Median      3Q     Max
-37.338 -11.532   1.081  13.386  33.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
```

```
Multiple R-squared:  0.6373,  Adjusted R-squared:  0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

What are the conclusions you can draw from the **summary**-output if you consider on the one hand the p-values of the single regression coefficients and the value of the F statistic? Is this a contradiction?

b) (6 points) Explain (in 3 sentences) the principles of variable selection that is based on the AIC. Which is the next predictor variable that is omitted in this variable selection procedure? Use the following **R**-output to answer this question.

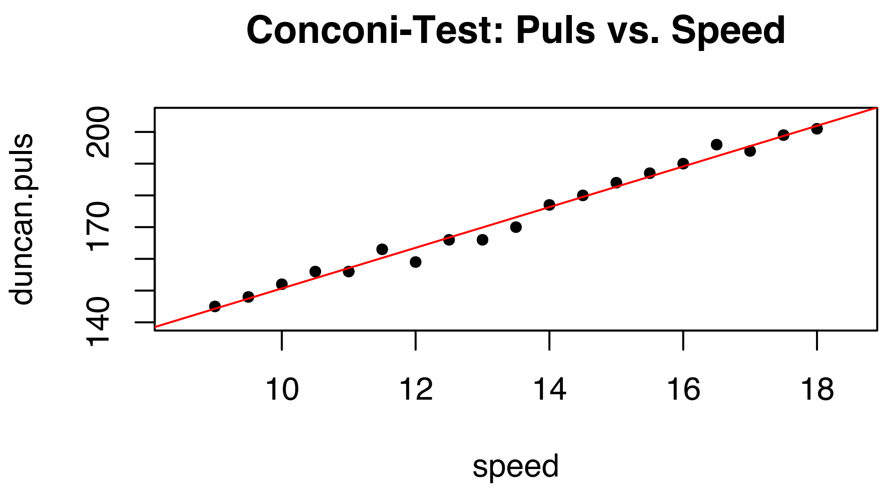
```
> step(pemax.lm)
Start:  AIC=169.11
pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc +
    tlc

         Df Sum of Sq     RSS    AIC
- sex     1     37.90  9769.2 167.20
- tlc     1     92.40  9823.7 167.34
- height  1    158.32  9889.6 167.51
- age     1    181.81  9913.1 167.57
- frc     1    254.55  9985.8 167.75
- fev1    1    648.45 10379.7 168.72
- rv      1    653.78 10385.0 168.73
<none>                 9731.2 169.11
- weight  1   1441.21 11172.5 170.56
- bmp     1   1480.12 11211.4 170.65
```

Problem 3: Conconi Test ....(34 Points)

The Conconi test measures the endurance performance of a person. It takes place on a 400 m-track where one starts running slowly (9 km h^{-1}). Every 200 meters the speed is increased by 0.5 km h^{-1}. At the end of every 200 m section the pulse is measured. The test continues until the speed can no longer be increased. Duncan's and Macduff's data are contained in the file **conconi.rda**:

```r
load("./Austausch/conconi.rda")
```

The scatter plot for Duncan's data with the least squares regression line looks as follows:



a) Visualize the data in a scatter plot as shown above. Fit the regression line with the command `lm()` and generate the summary output with **R** and answer the following questions:

   (i) (3 points) To what extent can we explain the variance in Duncan's pulse by the increase in speed?

   (ii) (4 points) By what amount does Duncan's pulse increase on average when the speed is increased by 1 km h^{-1}? What other values are also plausible?

(iii) (4 points) How large is Duncan's resting heart rate (i.e. when there is no movement)? What is the interval you expect this value to fall into? Does this seem plausible?

b) (5 points) We now consider Macduff's values for the Conconi test. If you fit a least squares model, then you obtain:

```
summary(fit.macduff)
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)        ???        ???     ???      ???
Speed          4.09323    0.09972   41.05   <2e-16 ***
```

Whose pulse is increasing more slowly when the speed is increased, Duncan's or Macduff's? Can we say whether there is a significant difference between the two increases of the pulse? Answer the question by means of confidence intervals for the regression coefficients.

c) (3 points) Generate the residual plots for Duncan's regression model. Decide which of the following three assumptions are fulfilled:

- The regression line captures the relation correctly, i.e. $\mathrm{E}(\varepsilon) = 0$
- The variance of the error is constant, i.e. $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$
- The errors follow a Normal distribution, i.e. $\mathcal{N}(0, \sigma_\varepsilon^2)$

What statements made in the previous sub-problems are still valid, which ones aren't?

d) (3 points) We create a data frame that contains all observations of the variables **puls**, **speed** and **runner** which should be a categorical variable with the levels **duncan** and **macduff**, indicating what person the corresponding observation belongs to.

```
## load data
load("Daten/conconi.rda")
## preprocess
speed <- conconi$speed[c(1:19, 7:26)]
puls <- c(conconi$macduff.puls[1:19], conconi$duncan.puls[7:26])
runner <- factor(c(rep("macduff", 19), rep("duncan", 20)))
conconi2 <- data.frame(puls, speed, runner)
```

Fit a least squares regression model for the main effects:
puls ~ speed + runner . What does this model assume with respect to the initial pulse and the increase in pulse of the two runners?

e) (12 points) Formulate a model (`lm(...)`) that assumes different initial pulses as well as different slopes, i.e. such that two distinct regression lines are fitted. Compute the estimates for the initial pulse (i.e. when `speed=0`) of Duncan and Macduff as well as the estimates for the amount the pulse increases with every additional $\mathrm{km\,h^{-1}}$ in speed. Is the difference significant?

Problem 4: Time Series .... (21 Points)

In this exercise we study the `JohnsonJohnson` time series that is built-in into `R`. It contains the earnings per Johnson & Johnson share over a time period of about 20 years.

a) (2 points) Use the output of the following code to answer the questions below:

```
start(JohnsonJohnson)
end(JohnsonJohnson)
frequency(JohnsonJohnson)
```

- What are the time increments in the time series?
- In which month and year does the time series start and end?

b) (2 points) Plot the time series with the command

```
plot(JohnsonJohnson)
```

Give at least two reasons why the given times series is not weakly stationary.

c) (5 points) The log-return of a time series $x_1, \ldots, x_n$ is a time series $y_1, \ldots y_{n-1}$ computed as

$$y_i = \log(x_{i+1}) - \log(x_i).$$

Compute the log-return for the `JohnsonJohnson` data using the `diff` and `log` functions in `R`. What is the mean and variance of the log-return series?

d) (3 points) Compute the partial autocorrelation function of the log-returns and determine the largest lag for which the partial autocorrelation is significantly different from zero.

e) (4 points) From your findings in d) compute an autoregressive model for the log-returns using the following code (replace **your_logret_sequence** with the name of your log-return time series):

```
mod = ar(your_logret_sequence, aic=FALSE, order.max = ???)
```

Provide the coefficients of the model.

f) (4 points) Predict the Johnson & Johnson share for the first quarter of the year 1981 using the **predict**-function and your model for the log-returns in e).

Problem 5: Logistic Regression ....(22 Points)

The file **PimaIndians.Rda** contains measurements of 8 clinical indicators of 768 individuals of the Pima Indian population near Phoenix, Arizona. Additionally, each individual has been tested on diabetes by means of an oral glucose tolerance test. The nine paramters are summarized in the following table

No_Pregnant	Number of times pregnant	**Insuline**	Insulin concentration
Plasma	Plasma glucose concentration	**BMI**	Body mass index
Blood_Pressure	Blood pressure	**DBF**	Diabetes pedigree function
Skin	Triceps skin fold thickness	**Age**	Age of the probant
		Diagnosis	Diabetes diagnosis

The aim of this exercise is to model **Diagnoses** as a function of the remaining 8 parameters. Load the data and generate a training and test set by the following code

```
load("./Austausch/PimaIndians.Rda")  # load a data frame called 'X'
set.seed(983)
idx.train = sample(nrow(X), 600, replace = F)
X.train = X[idx.train, ]
X.test = X[-idx.train, ]
```

a) (3 Points) Compute a logistic regression model on the training set for **Diagnosis** that incorporates all remaining 8 parameters as predictors (Hint: use the **glm** function). List all predictors that are significant for the model with $p < 0.01$

b) (5 Points) Recompute the logistic regression model but this time only use the significant predictors in a). Write down the logistic regression equation explicitely for this case.

c) (6 Points) Assume that the model in b) estimates a probability of 0.61 of an individual for having diabetes. How does this probability change, if the BMI of this individual increases by 5 points (while the other parameters stay unchanged)?

d) (4 Points) The following code predicts **`Diagnosis`** on the test set by thresholding the probability at 0.5. Also a confusion matrix is computed. Replace **`myModel`** with the **full model** from a).

```
pred.prob = predict(myModel, newdata = X.test, type = "response")
pred.class = as.integer(pred.prob > 0.5)
tb = table(X.test$Diagnosis, pred.class)
addmargins(tb)
```

Compute the classification error, as well as the false positive and false negative rate.

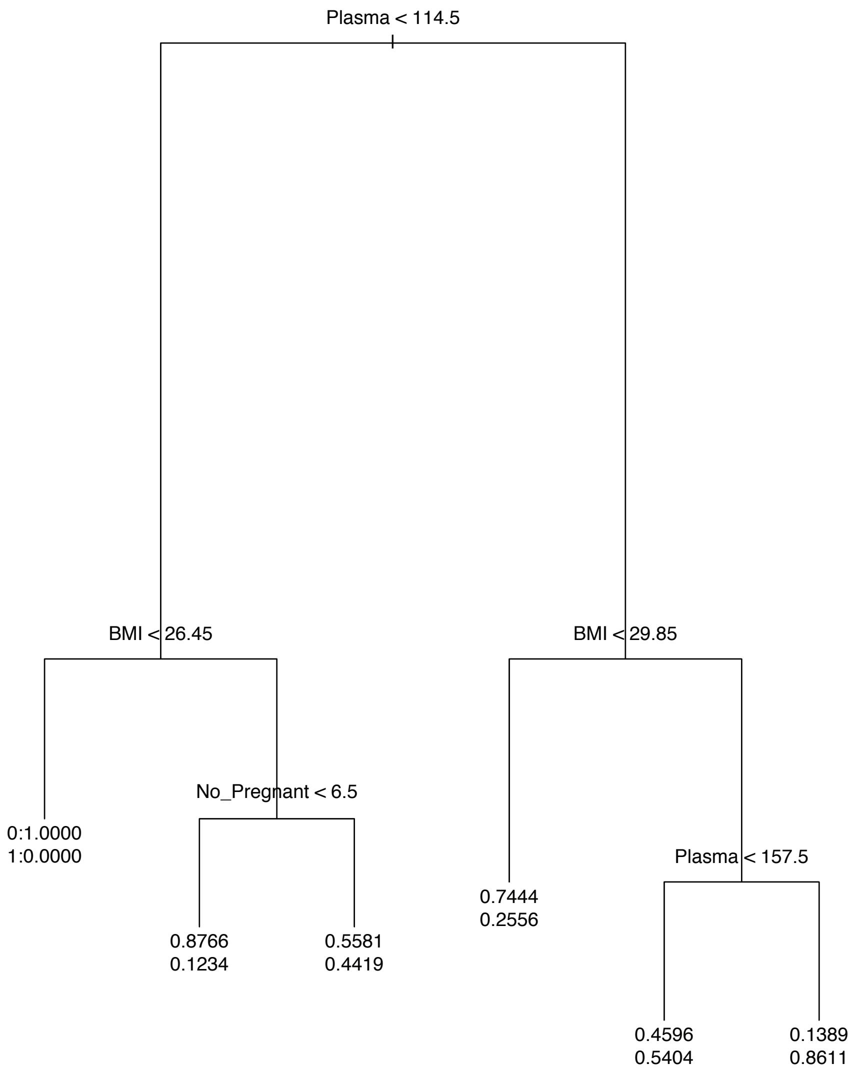
e) (4 Points) Is there are threshold other than 0.5 that gives a lower classification error? If yes, provide such a threshold.

Problem 6: Conceptual Questions .... (17 Points)

a) (7 points) Assume we are given a white noise process $W_1, W_2, \dots$ with variance $\sigma^2 = 1$. Compute the autocorrelation at lag 1 of the process

$$X_k = W_k + \frac{1}{2}W_{k-1} - \frac{2}{3}W_{k-2}.$$

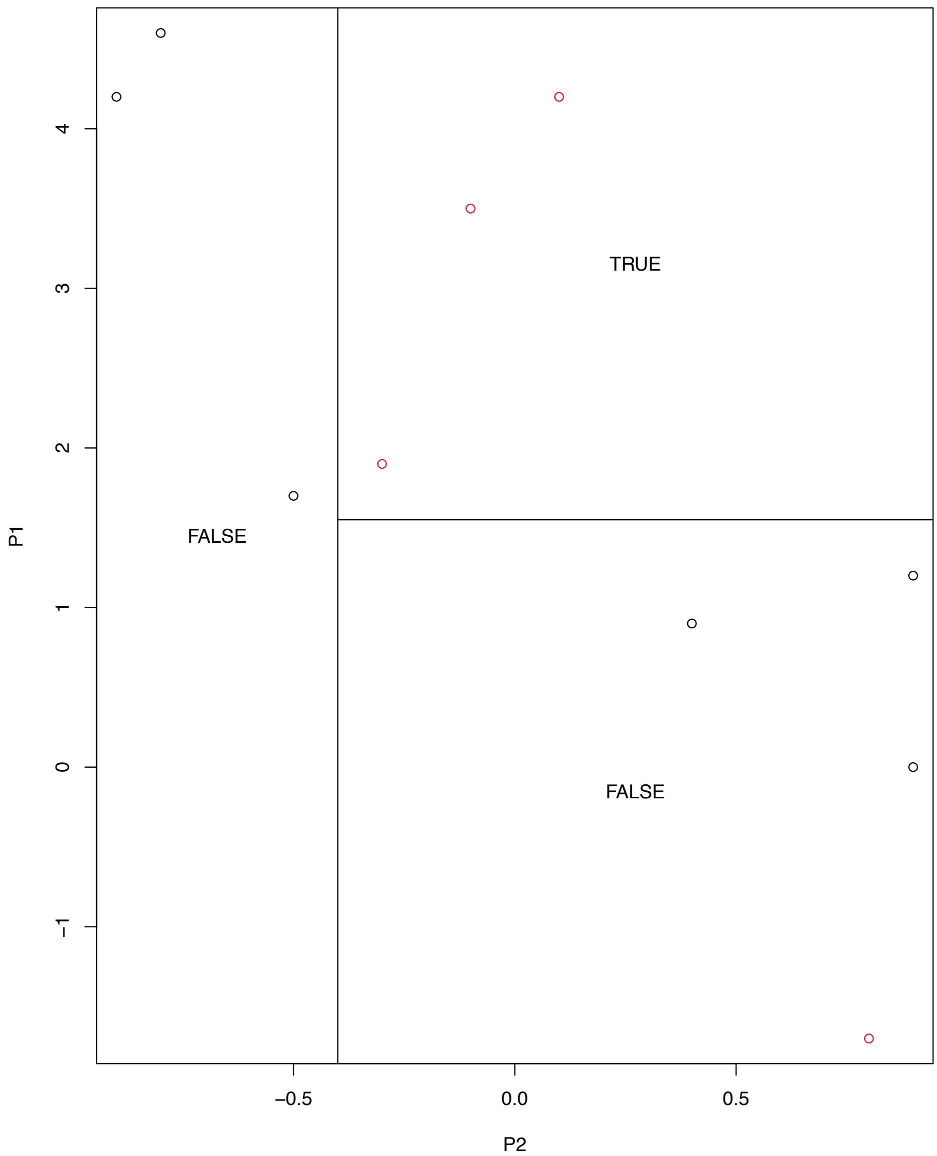
b) (3 points) Below a decision tree for the **PimaIndians** data set is shown. Compute the cross-entropy for each terminal node.



(b) (3 points) Which of the following assertions about random forests (RF) are true?

- o RF is an ensemble method that performs bootstrap aggregation.
- o The out-of-bag (OOB) error estimate is always computed from a test set.
- o The OOB error is strictly decreasing with the number of trees.
- o RF usually have high variance compared to simple tree models.
- o For each tree and each split, RF only consider a random subset of predictors.

c) (4 points) Below the partition of a 2 dimensional predictor space is shown. The data points are coloured acccording to a class variable (binary classification). The partition is due to a decision tree model



- Draw the resulting tree and annotate all splits and the terminal nodes.
- Compute the predicted class of a new observation $(P_1, P_2) = (-0.5, 2)$.
- Compute the estimated classification error for this class and interpret the result.