MASTER OF SCIENCE
IN ENGINEERING

Examination FS1

Predictive Modeling

Date : 5th July 2017 , 13:15-15:15

First Name: ______________________

Family Name: ______________________

School / Partial School: ______________________

Stick Number: ______________________

Problem	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
max. points	12	30	18	30	18	12	120
Achieved points							

Please open the file `Lastname_Firstname.R` in the folder `/home/user/Vorlagen` 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) Personal handwritten summary of 20 pages
   c) **R** Reference Card (with your comments)
   d) Calculator
   e) Laptop booted from the Lernstick environment with statistical software **R**
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 `/home/user/Vorlagen` 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: Short Comprehension Exercises .... (12 Points)

a) Decide whether the following statements are true or false. Explain your answer in 1-2 sentences.

(i) (2 points) We consider the model $Y = \beta_0 + \beta_1 X + \varepsilon$. Let $[-0.01, 1.5]$ be the 95 %-confidence interval for β_1. In this case, a t-test with significance level 1 % rejects the null hpyothesis $H_0 : \beta_1 = 0$.

(ii) (2 points) Complicated models with a lot of parameters are better for prediction than simple models with just a few parameters.

(iii) (2 points) The following formulas specify all the same model $x \sim x + y + x : y$, $z \sim x * y$ and $z \sim (x + y)^2$.

(iv) (2 points) It can happen that all individual t-tests in a regression do not reject the null hypothesis, although the global F-test is significant.

b) (2 points) Suppose you have a saturated model, i.e. a model containing the same number of parameters as observations. What would the estimate of σ^2 be? Explain your answer in 2 sentences.

c) (2 points) Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2 + \beta_4 X_4 + \varepsilon$$

You want to test the null hypothesis $\beta_2 = \beta_3 = \beta_4 = 0$ against the alternative hypothesis $\beta_2 \neq 0$ and/or $\beta_3 \neq 0$ and/or $\beta_4 \neq 0$. Which class of distribution does the corresponding test statistic have?

Problem 2: Multiple Choice .... (24 Points)

A multiple regression model of the following form is fitted to a data set:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The model is fitted using the software **R** and the following summary output is obtained:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      ???     0.1960   8.438 3.57e-13
x1            5.3036     2.5316     ??? 0.038834
x2            4.0336     2.4796   1.627 0.107111
x3           -9.3153     2.4657  -3.778 0.000276
x4            0.5884     2.2852   0.257 0.797373

Residual standard error: 1.892 on 95 degrees of freedom
Multiple R-squared:  0.1984,  Adjusted R-squared:  ???
F-statistic: 5.745 on 4 and 95 DF,  p-value: 0.0003483
```

Only **one** answer is the correct one: mark by means of a **single cross** the correct answer. If you cross the correct answer, you will get 3 points per question. If you cross the wrong answer, one point is subtracted from the total number of points you have achieved. At minimum you will get 0 points for problem 2.

(1.) What is the value of the t-statistic of $\widehat{\beta}_1$?

a) 0.099

b) 13.43

c) 2.095

d) 0.015

(2.) How many observations are in the data set?

a) 100
b) 99
c) 96
d) 95

(3.) Has the null hypothesis $H_0 : \beta_3 = 0$ to be rejected at the 5 % level?

a) Yes
b) No
c) No answer possible.

(4.) What is the estimate of the intercept $\widehat{\beta}_0$?

a) 1.654
b) 0.324
c) 43.051
d) 1.591

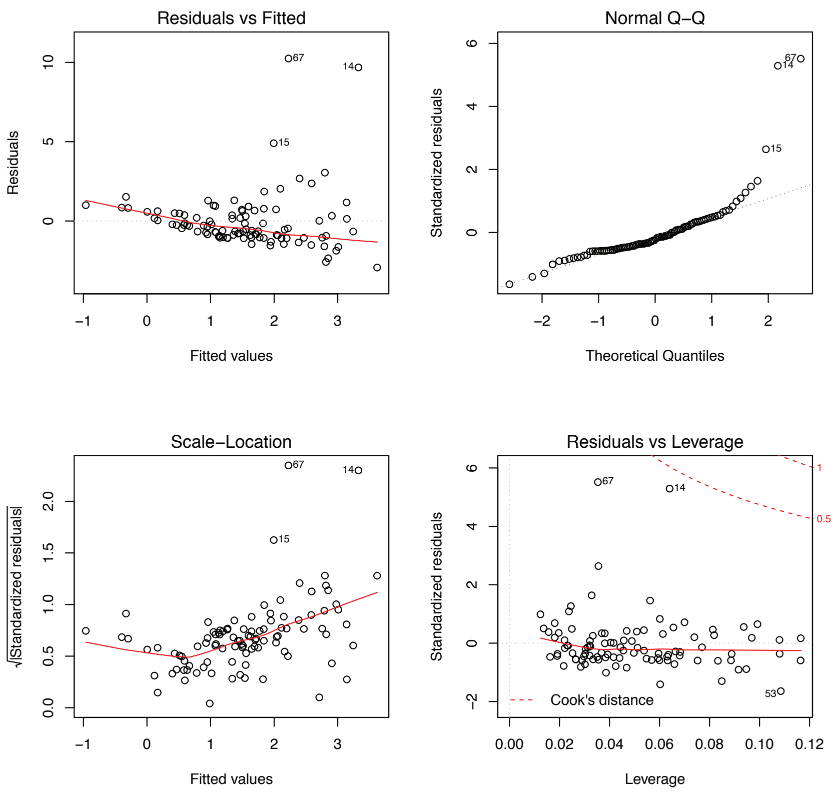
(5.) What is the estimate of $\text{Var}(\varepsilon)$?

a) 1.892
b) 3.579
c) 1.375
d) 9.46

(6.) Which of the following intervals is a two-sided 95 % confidence interval for β_3?

a) $-9.315 \pm 1.99 \cdot 0.00028$
b) $-9.315 \pm 1.99 \cdot \frac{2.466}{\sqrt{95}}$
c) $-9.315 \pm 1.99 \cdot \frac{0.00028}{\sqrt{95}}$
d) $-9.315 \pm 1.99 \cdot 2.466$

(7.) Have a look at the residual plots. Are the model assumptions on the ε fullfilled and if not, what is the main problem?



a) Yes.

b) No, since leverage points exist.

c) No, since the assumption of constant variance of the ε is violated.

d) No, since the ε are dependent.

(8.) You want to repeat the regression, but with a better model and/or adapted data basis. What action do you take?

a) Leave out all non significant variables.

b) Investigate the data without leverage points and outliers.

c) Add a quadratic term.

d) Apply a transformation to the response variable.

Problem 3: Boston Crime Per Capita .... (24 Points)

This problem involves the **`Boston`** data set, which records among other variables the per capita crime in suburbs of Boston. We will now predict **`crim`**, which refers to the per capita crime rate, using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

Load the data as follows:

```r
load("../Boston.RData")
```

a) (4 points) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

b) (5 points) Perform a hybrid stepwise selection (**`method='both'`**). Propose a model that seems to perform well on this data set. Explain which criteria you used to select this particular model.

c) (5 points) Perform a residual analysis of your selected model and comment on potentially problematic aspects of your model.

d) (6 points) Consider now the model **`crim ~ lstat * age`**, where **`age`** refers to the proportion of owner-occupied units built prior to 1940 and **`lstat`** refers to the percentage of households with low socio-economic status. Is the interaction effect significant? How do you interpret this interaction term and the signs of the coefficient estimates?

e) (4 points) Determine a 95 % prediction interval and a 95 % confidence interval for **`crim`** given **`age=80`**, **`lstat=40`** and the model **`crim ~ lstat * age`**. If a mayor wants a prognosis for the per capita crime rate in his town, will he consider the prediction interval or rather the confidence interval?

Problem 4: Pima Indians Diabetes Database .... (12 Points)

The data set contains 8 diagnostic parameters $X_1, \ldots, X_8$ of 767 female individuals of Pima heritage. The population lives near Phoenix, AZ. The following parameters have been determined

1. `Number of pregnancies`
2. `Plasma glucose concentration`
3. `Diastolic blood pressure`
4. `Triceps skin fold thickness`
5. `2-hours serum insulin`
6. `Body mass index`
7. `Diabetes pedigree function`
8. `Age`
9. `Diagnosis`

The diagnosis parameter is binary (0 or 1) and indicates whether the individual has diabetes or not.

a) Load the data with

```
load("./Daten/PimaIndians.Rda")
```

Your workspace now contains a variable `Pima`. Genereate training and test set by the command

```
set.seed(18)
idx.train = sample.int(nrow(X), 500)
train.set = X[idx.train, ]
test.set = X[-idx.train, ]
```

b) Fit a logistic regression model `Pima.fit` to the training set using the `glm` function. Which are the significant predictors?

c) Make another fit `Pima.fit.sig` invoking only the significant predictors in the previous question and write down the logistic regression model for this case.

d) Given a woman that has never been pregnant and additionally has the parameters `PlasmaGlucose=101`, `Diastolic=71`, `BMI = 28.1` and `DiabPedigree=0.621`, what is the probability of her suffering from diabetes?

e) The command

```
train.prob = predict(Pima.fit.sig, type = "response")
train.class = as.integer(train.prob > 0.5)
table(train.class, X$Diagnosis)
```

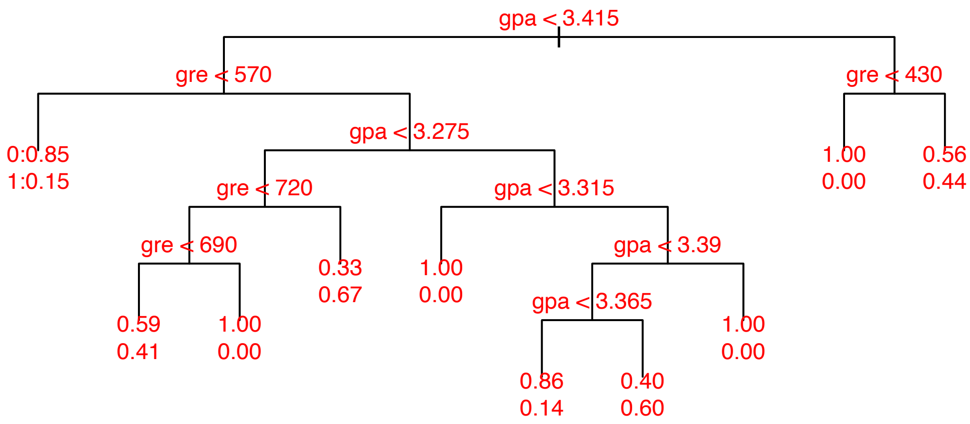
computes the confusion matrix on the training set. What is the classification error on the training set?

f) Provide the confusion matrix and classification error on the test set.

Problem 5: School data . . . . (12 Points)

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores) and GPA (grade point average) in an undergraduate institution effect admission into graduate school. The response variable, admit/don???t admit, is a binary variable. There are $n = 400$ observations of **gre**, **gpa** and the response **admit**.

The researcher models the response **admit** by means of a classification tree. Binary splitting is performed an the following tree is generated. The terminal nodes show the class probabilites for **admit=yes** (bottom) and **admit=no** (top).



a) Write the predicted class at each terminal node.

b) Compute the Gini index for the left most terminal node.

c) Draw by hand the predictor space and the partition implied by the tree.

d) If a student has **gre**$= 700$ as well as **gpa**$= 3.6$, will he be admitted to graduate school? What is the probability for that?

Problem 6: Discrete stochastic process .... (12 Points)

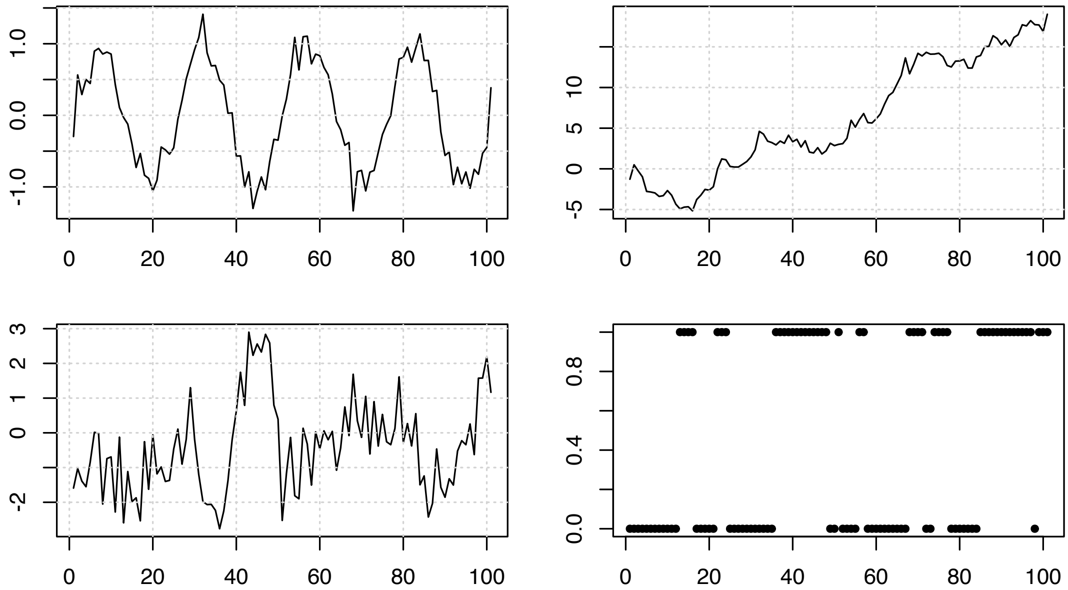
We are given a discrete Stochasic process

$$X_n = \frac{1}{2}X_{n-1} + \frac{1}{3}X_{n-2} + W_n,$$

where W_n is a white noise process.

a) Prove that the process is stationary.

b) Which of the following time series is a realization of the process?



c) Let σ_X^2 be the variance of X_n. Show that the autocovariance at lag 1 is given by

$$\gamma(1) = \mathrm{Cov}(X_n, X_{n+1}) = \frac{3}{4}\sigma_X^2.$$

(**Hint:** Use the definition of the process and the fact that $\mathrm{Cov}(X_n, X_{n+1}) = \mathrm{Cov}(X_{n-1}, X_n)$ which follows from stationarity.)

d) Answer the following questions

Question	**True**	**False**
A discrete stochastic process with constant mean and constant variance is weakly stationary.		
If the characteristic polynomial of an AR(p) has only real zeros, then the process is stationary.		
If all zeros of the characteristic polynomial of an AR(p) are larger than 1 in absolute value, then the process is stationary.		
Let X_n be weakly stationary. Then $\mathrm{Cov}(X_3, X_8) = \mathrm{Cov}(X_5, X_{10})$.		

Solutions

Solution 1

a) (i) False. The corresponding 5 % hypothesis test accepts the null hypothesis β_1, since 0 lies in the 95 % confidence interval. The test decision thus remains even more true in the case of a 99 % confidence interval.

(ii) False. When several predictors need to be estimated, the variance of the prediction, that is the variance of the response variable, may increase. In the case of a simpler model the bias is likely to be larger. Hence, we need to find a good balance between the bias and variance.

(iii) True. $x * y = x + y + x : y$. Since $x^2 = x : x$, we have $(x + y)^2 = x + x : y + y$

(iv) True. This may occur when the predictor variables exhibit a strong correlation among each other.

b) In a saturated model all residual are zero. The residual standard error thus will be estimated to 0.

c) This is a partial F-test. The corresponding test statistic follows an F-distribution.

Solution 2

(1.) (c)

(2.) (a)

(3.) (a)

(4.) (a)

(5.) (b)

(6.) (d)

(7.) (c)

(8.) (d)

Solution 3

a)
```r
library(MASS)
lm.all = lm(crim ~ ., data = Boston, method = "forward")
summary(lm.all)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston, method = "forward")
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

For the following predictors we can reject the null hypothesis $H_0 : \beta_j = 0$:

(i) zn

(ii) dis

(iii) rad

(iv) black

(v) medv

b) Hybrid stepwise selection:

```
library(leaps)
f_full <- lm(crim ~ ., data = Boston)
f_empty <- lm(crim ~ NULL, data = Boston)
```

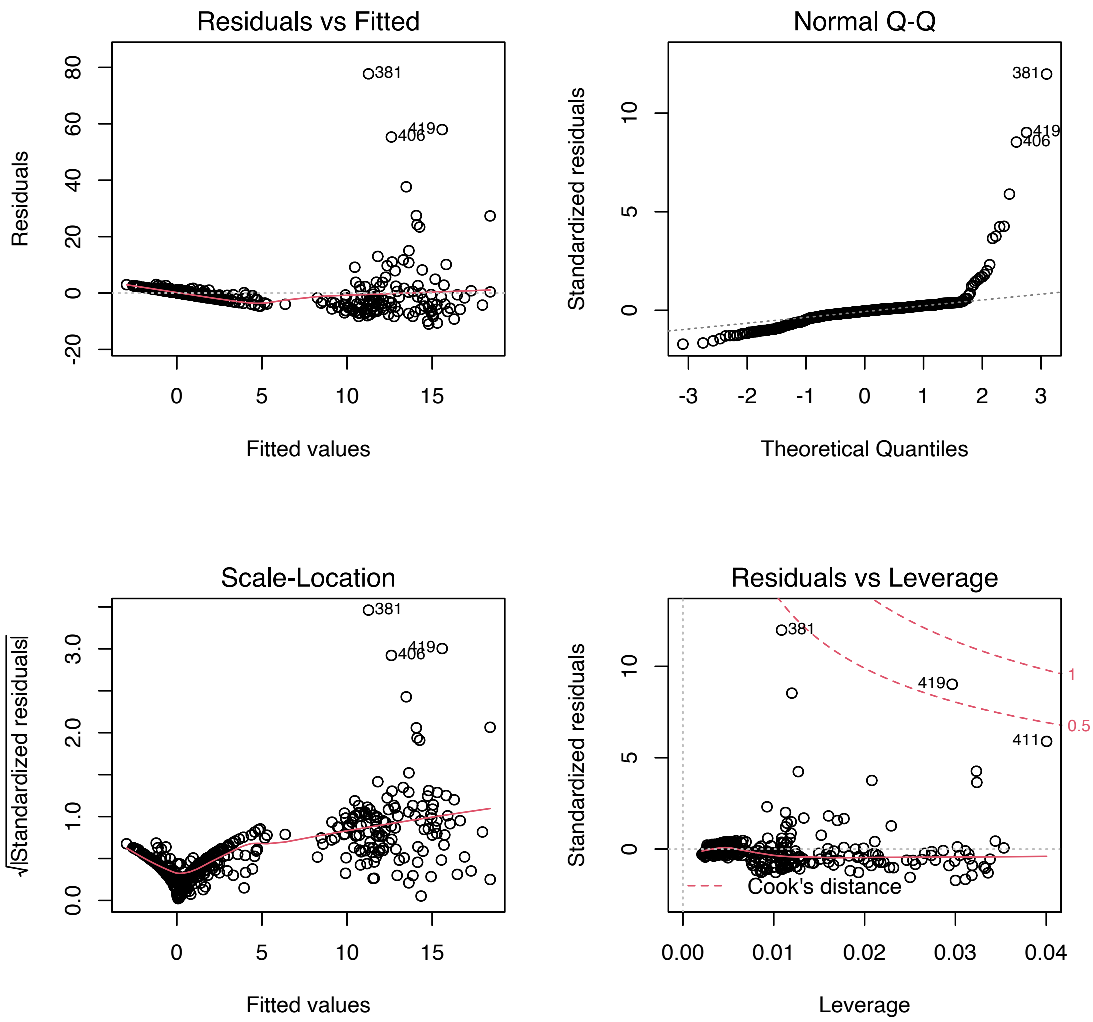
```r
regfit <- step(f_empty, direction = "both", scope = list(lower = f_em
    upper = f_full), trace = 0, k = log(nrow(Boston)))
regfit
```

```
##
## Call:
## lm(formula = crim ~ rad + lstat + black, data = Boston)
##
## Coefficients:
## (Intercept)          rad        lstat        black
##   -0.372585     0.488172     0.213596    -0.009472
```

On the basis of the BIC we select the following model:

crim ∼ rad + black + lstat

c)
```r
par(mfrow = c(2, 2))
plot(regfit)
```



There are several outliers (e.g. points 381, 406, 411 and 419) that however do not seem to be dangereous in terms of Cook's distance. However, the residuals show a rather long-tailed distribution which is to some extent due to these outliers. The normality assumption seems to be violated. The most problematic issue indicated by these residual plots is certainly the shape of the smoother curve in the scale-location plot. This may indicate that there is a non-linear relationship between predictors and response variable.

d)
```
lm_interaction <- lm(crim ~ age * lstat, data = Boston)
summary(lm_interaction)

##
## Call:
## lm(formula = crim ~ age * lstat, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -15.695  -2.162  -0.283   0.424  82.389
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.018871   1.800236   0.566 0.571670
## age         -0.033778   0.024348  -1.387 0.165958
## lstat       -0.268568   0.205097  -1.309 0.190975
## age:lstat    0.008406   0.002268   3.706 0.000234 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.531 on 502 degrees of freedom
## Multiple R-squared:  0.2381,Adjusted R-squared:  0.2335
## F-statistic: 52.28 on 3 and 502 DF,  p-value: < 2.2e-16
```

The higher the percentage of older (prior to 1940) owner-occupied houses of a suburb, the smaller is the effect on the per capita crime rate. The higher the percentage of people having a low socio-economic status in a suburb, the smaller the effect on the per capita crime rate. However, these two quantities are on the one hand not significant and on the other hand, they do not have an additive effect on the per capita crime rate, since their interaction effect is significant. That is, the combination of a suburb having a high popultion percentage of low socio-economic status and a high percentage of old owner-occupied houses leads to higher crime per capita rate.

e)
```
predict(lm_interaction, data.frame(lstat = 40, age = 80),
    interval = "confidence")

##        fit     lwr     upr
## 1 14.4721 11.2263 17.7179
```

The 95 % confidence interval thus is given by : $[11.23, 17.72]$.

```
predict(lm_interaction, data.frame(lstat = 40, age = 80),
    interval = "prediction")

##        fit       lwr      upr
## 1 14.4721 -0.675162 29.61937
```

The 95 % prediction interval thus is given by : $[0, 29.62]$.

Since the mayor is interested to predict the per capita crime rate for a single town (observation), he would consider a prediction interval.

Solution 4

a) -

b) We compute

```
Pima.fit.complete = glm(Diagnosis ~ ., data = train.set,
    family = "binomial")
summary(Pima.fit.complete)
```

```
##
## Call:
## glm(formula = Diagnosis ~ ., family = "binomial", data = train.set
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9413  -0.6989  -0.3741   0.6989   2.7739
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.5369531  0.9062783  -9.420  < 2e-16
## No_Pregnant     0.1543596  0.0397410   3.884 0.000103
## Plasma          0.0355823  0.0047113   7.552 4.27e-14
## Blood_Pressure -0.0147201  0.0065037  -2.263 0.023615
## Skin           -0.0045355  0.0085574  -0.530 0.596104
## Insuline       -0.0007086  0.0010696  -0.662 0.507674
## BMI             0.0904864  0.0191801   4.718 2.38e-06
## DBF             1.4302113  0.3712967   3.852 0.000117
## Age             0.0098361  0.0113939   0.863 0.387984
##
## (Intercept)    ***
## No_Pregnant    ***
## Plasma         ***
## Blood_Pressure *
## Skin
## Insuline
## BMI            ***
## DBF            ***
## Age
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 651.08  on 499  degrees of freedom
## Residual deviance: 456.59  on 491  degrees of freedom
## AIC: 474.59
##
```

```
## Number of Fisher Scoring iterations: 5
```

We see that all predictors save **Skin**, **Insulin** and **Age** are significant.

c)
```
Pima.fit.sig = glm(Diagnosis ~ No_Pregnant + Plasma + Blood_Pressure
    BMI + DBF, data = train.set, family = "binomial")
coefficients(Pima.fit.sig)

##     (Intercept)     No_Pregnant          Plasma
##     -8.16652464      0.17615418      0.03547311
## Blood_Pressure             BMI             DBF
##     -0.01445621      0.08276500      1.37145769
```

If X_1, X_2, X_3, X_4, X_5 are the predictors corresponding to **NoPreg, PlasmaGlucose, Diastolic, BMI, DiabPedigree**, then we have for

$$p(x_1, \ldots, x_5) = P(\texttt{Class = 1} | X_1 = x_1, \ldots, X_5 = x_5)$$

the following model

$$\log\left(\frac{p(x_1,\ldots,x_5)}{1-p(x_1,\ldots,x_5)}\right) = -7.503 + 0.145x_1 + 0.033x_2 - 0.019x_3 + 0.095x_4 + 0.679x_5.$$

d)
```
library(evaluate)
Pima.fit.sig <- glm(Diagnosis ~ No_Pregnant + Plasma + Blood_Pressure
    BMI + DBF, data = train.set, family = "binomial")
prob <- predict(Pima.fit.sig, newdata = data.frame(No_Pregnant = 0,
    Plasma = 101, Blood_Pressure = 71, BMI = 28.1, DBF = 0.621),
    type = "response")
log_odds <- log(prob/(1 - prob))
```

With the given values the log-odds for suffering from diabetes are $u = -2.433$. Thus, we find that

$$p(0, 101, 71, 28.1, 0.621) = e^{u/(1+u)} = 0.0807.$$

e) We find

```
train.prob = predict(Pima.fit.sig, type = "response")
train.class = as.integer(train.prob > 0.5)
table(train.class, train.set$Diagnosis)

##
## train.class   0   1
##           0 286  70
##           1  36 108
```

Thus we find for the classification error

$$\text{Err} = \frac{34 + 68}{500} = 0.204.$$

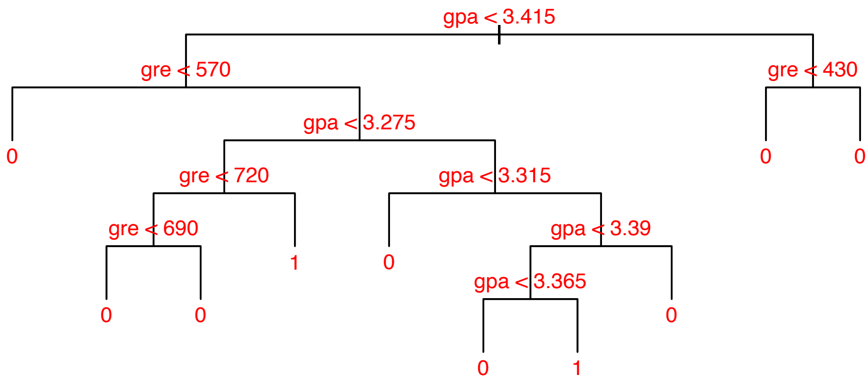
f)
```
test.prob = predict(Pima.fit.sig, type = "response", newdata = test.s
test.class = as.integer(test.prob > 0.5)
table(test.class, test.set$Diagnosis)

##
## test.class   0   1
##          0 155  46
##          1  23  44
```

Thus we find for the classification error

$$\text{Err} = \frac{20 + 55}{268} = 0.280.$$

Solution 5



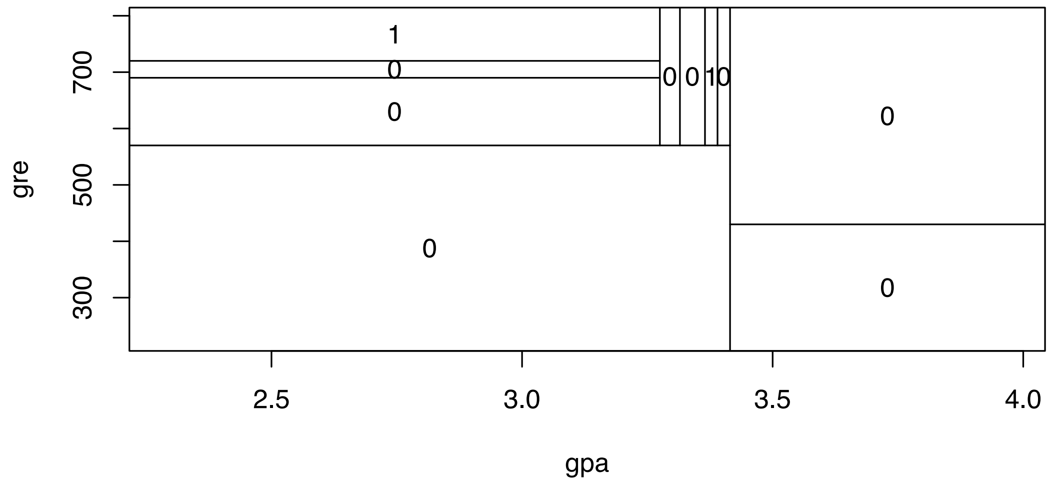
a)

b) Compute the Gini index for every terminal node. Which is the purest and which the impurest node? The Gini index for a two class problem for node m reads as

$$G_m = \hat{p}_m(1 - \hat{p}_m)$$

where $\hat{p}_m$ is the frequence of **admit = yes** in training data at node m. So we find that

$$G_1 = 0.85 \cdot 0.15 = 0.1275.$$



c)

d) Pushing down the given data along the tree we find that we end up in the right most node. The prediction amounts to say that the student will not be admitted. The estimated probability for that is 0.56.

Solution 6

a) The process is of AR(2) type. Hence stationarity is can be assessed by means of the characteristic polynomial. We compute the zeros of $\phi(x) = 1 - \frac{1}{2}x - \frac{1}{3}x^2$ and find that

$$x_1 = \quad 1.137 \text{ and } \quad x_2 = -2.637.$$

Hence, both have an absolute value larger than 1 and hence the process is stationary.

b) According to a) the process is stationary. Hence the lower left process is the only plausible candidate for a time series that is generated from the process. The first image exhibits seasonality, the second a clear trend. The last image is a descrete provesss.

c) Let σ_X^2 be the variance of X_n. The autocovariance at lag 1 is given by

$$\begin{aligned}\mathrm{Cov}(X_n, X_{n+1}) &= \mathrm{Cov}(X_n, a_1 X_n + a_2 X_{n-1} + W_{n+1}) \\ &= a_1 \mathrm{Cov}(X_n, X_n) + a_2 \mathrm{Cov}(X_n, X_{n-1}) + \mathrm{Cov}(X_n, W_{n+1}) \\ &= a_1 \sigma_X^2 + a_2 \mathrm{Cov}(X_n, X_{n+1}).\end{aligned}$$

Rearranging gives

$$(1 - a_2)\mathrm{Cov}(X_n, X_{n+1}) = a_1 \sigma_X^2$$

and hence

$$\mathrm{Cov}(X_n, X_{n+1}) = \frac{a_1}{1 - a_x}\sigma_X^2 = \frac{3}{4}\sigma_X^2.$$

d) Answer the following questions

Question	**True**	**False**
A discrete stochastic process with constant mean and constant variance is weakly stationary.		x
If the characteristic polynomial of an AR(p) has only real zeros, then the process is stationary.		x
If all zeros of the characteristic polynomial of an AR(p) are larger than 1 in absolute value, then the process is stationary.	x	
Let X_n be weakly stationary. Then $\mathrm{Cov}(X_3, X_8) = \mathrm{Cov}(X_5, X_{10})$.	x	