Homework: Data Analysis

This homework sheet will test your knowledge on analyzing data using R.

30

0

a) Load the data `Boston` from the library `MASS`. All subsequent exercises study potential parameters influencing the housing values in suburbs of Boston.

```
library(MASS)
```

```
data(Boston)
head(as.data.frame(cbind(Boston$medv, Boston$crim, Boston$chas)))
```

```
##     V1      V2 V3
## 1 24.0 0.00632  0
## 2 21.6 0.02731  0
## 3 34.7 0.02729  0
## 4 33.4 0.03237  0
## 5 36.2 0.06905  0
## 6 28.7 0.02985  0
```

```
dim(Boston)
```

```
## [1] 506  14
```

Overall, the dataset `Boston` contains 506 rows and 14 columns.

1

b) The median value of owner-occupied homes (in $ 1000) is given by the column `medv`. Give and discuss its summary statistics!

Solution:

```
summary(Boston$medv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0    17.0    21.2    22.5    25.0    50.0
```

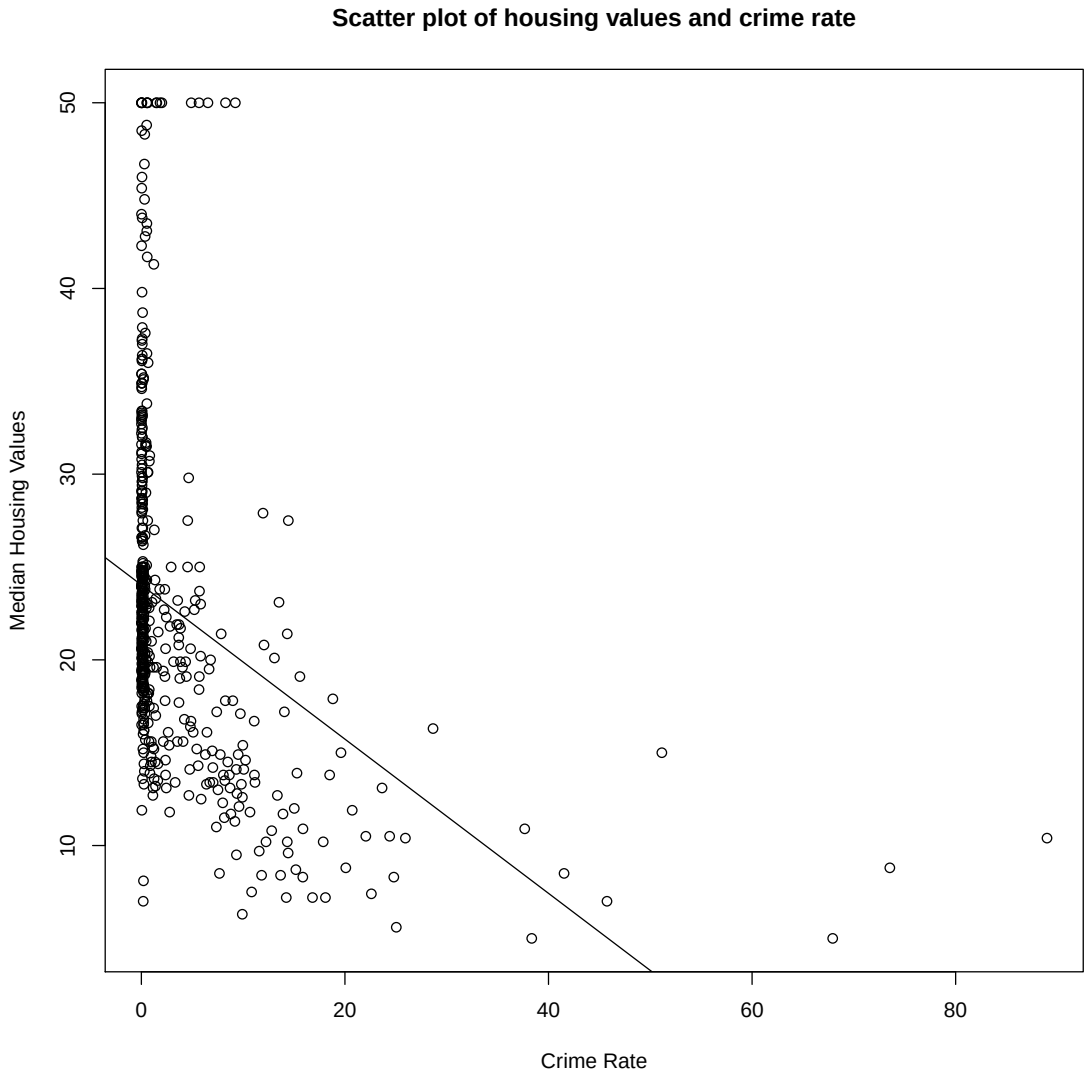
The housing values range from $ 5000 to $ 5×10^4. The average housing value is $ 2.25×10^4, whereas the median is given by $ 2.12×10^4.

1

c) Depict the relationship between housing values (column: `medv`) and crime rate (column: `crim`) in a scatter plot. Also draw the line of best fit!

Solution:

```
plot(Boston$crim, Boston$medv, main = "Scatter plot of housing values and crime rate",
    xlab = "Crime Rate", ylab = "Median Housing Values")
m <- lm(Boston$medv ~ Boston$crim)
abline(m)
```



One can see a clear relationship.

1

d) Analyze this relationship using the Pearson correlation coefficient (and its corresponding hypothesis test)!

Solution:

```
cor(Boston$crim, Boston$medv)

## [1] -0.3883

cor.test(Boston$crim, Boston$medv)

##
##  Pearson's product-moment correlation
```

```
##
## data:  Boston$crim and Boston$medv
## t = -9.46, df = 504, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4599 -0.3117
## sample estimates:
##     cor
## -0.3883
```

Although the correlation is relatively small, the P-value smaller than 0.01 indicates a significant linear dependence at the $1\,\%$-significance level.

2

e) Estimate the median housing values in Boston through a linear OLS regression model M_1 specified by

$$medv(x) = \alpha + \beta_1\, crim(x) + \beta_2\, chas(x) + \varepsilon\,(x).$$

Column `crim` gives the per capita crime rate by town, `chas` a Charles River dummy variable ($= 1$ if house tract bounds river; 0 otherwise) and $\varepsilon(x)$ is the error term. Give a short interpretation of your result.

Solution:

```
m1 <- lm(medv ~ crim + chas, data = Boston)
summary(m1)
```

```
##
## Call:
## lm(formula = medv ~ crim + chas, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.54  -5.42  -1.88   2.58  30.13
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.6140     0.4186   56.41  < 2e-16 ***
## crim         -0.4060     0.0434   -9.36  < 2e-16 ***
## chas          5.5777     1.4693    3.80  0.00016 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.37 on 503 degrees of freedom
## Multiple R-squared:  0.174,Adjusted R-squared:  0.171
## F-statistic: 53.1 on 2 and 503 DF,  p-value: <2e-16
```

From the P-values, we see that both regressors reveal an effect on housing value within the model that is statistically significant at a $0.1\,\%$ significance level. From the coefficient signs, we observe that more crimes relate to lower housing values, whereas a tract bounding the Charles river related to higher housing values. With a P-value below $0.1\,\%$, we can reject the null hypothesis that all coefficients are equal to zero at all common significance levels. Overall, an adjusted R^2 of 0.1712 shows that the model has limited explanatory power. However, the previous statements are subject to the preconditions of the OLS estimator that needs

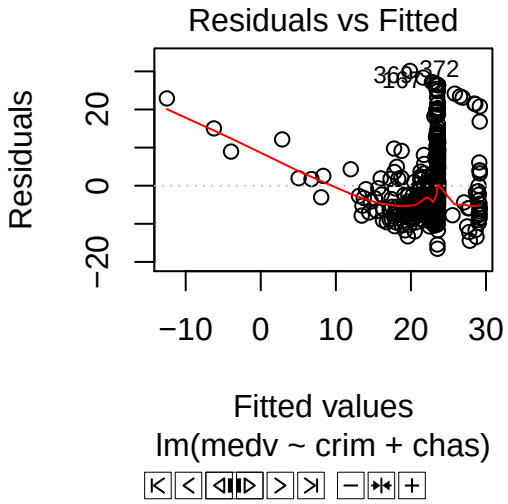
to be fulfilled.

4

f) Test the model for heteroskedasticity using visual regression diagnostics and generate the following plots: residuals vs. fitted values, residuals across observations and a Q-Q plot to check normal distribution of residuals.

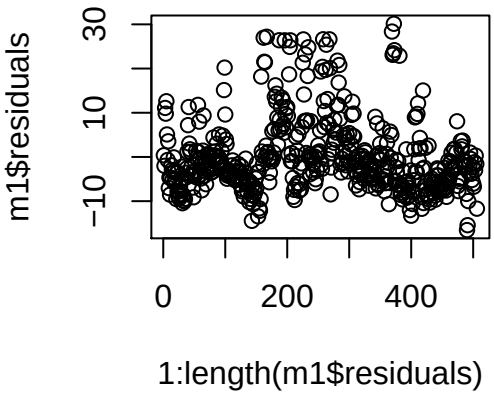
Solution:

```
# residuals vs. fitted and Q-Q plot
plot(m1)
```



Apparently, the plot comparing residuals vs. fitted values shows a non-uniformly distributed pattern. This is a strong indicator of heteroskedasticity (and/or autocorrelation). In addition, looking at the Q-Q plot, we can clearly see that the residuals do not follow a normal distribution since there is a large gap on the right side.

```
# residuals across observations
plot(1:length(m1$residuals), m1$residuals)
```



Since many observations in the middle of the plot seem to be shifted upwards, this is a possible indication of heteroskedasticity.

2

g) Test the model for heteroskedasticity with the Breusch-Pagan test. Interpret the result!

Solution:

```
bptest(m1)

##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 11.95, df = 2, p-value = 0.002539
```

As the P-value is below 0.05, the null hypothesis of homoscedasticity can be rejected at the $5\,\%$ significance level and we instead assume heteroskedasticity.

2

h) Test the model for autocorrelation with the Durbin-Watson test and interpret the result.

Solution:

```
dwtest(m1)

##
##  Durbin-Watson test
##
## data:  m1
```

```
## DW = 0.7595, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

As the P-value is below 0.05, the null hypothesis of no autocorrelation can be rejected at the $5\,\%$ significance level and we assume autocorrelation instead.

1

i) Explain why it does not make sense to test autocorrelation by plotting a correlogram?

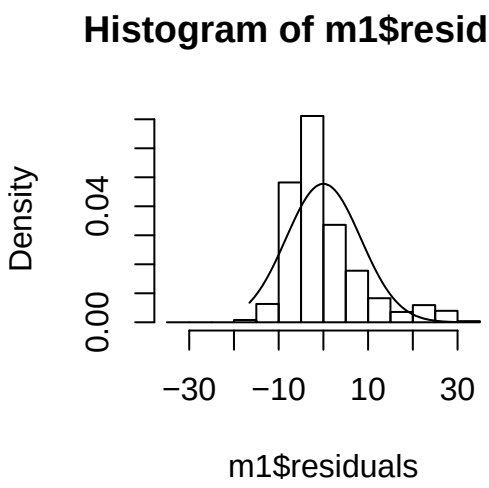
Solution:
Correlograms (i. e. the autocorrelation function) only aim at time series with a time lag.

2

j) Plot a histogram of the residuals of the model and the residuals versus fitted and interpret both.

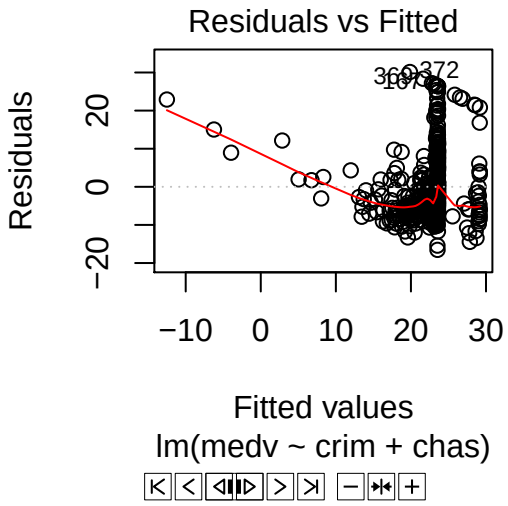
Solution:

```
hist(m1$residuals, freq = FALSE, breaks = seq(-35, 35, 5))
xx <- seq(min(m1$residuals), max(m1$residuals), 0.01)
lines(xx, dnorm(xx, mean = mean(m1$residuals), sd = sd(m1$residuals)))
```



The histogram reveals, though badly visible, that the residuals do not follow a normal distribution.

```
plot(m1)[1]
```




```
## NULL
```

Apparently, the plot comparing residuals vs. fitted values shows a non-uniformly distributed pattern. This is a strong indicator of heteroskedasticity and/or autocorrelation.

2

k) Estimate a second model M_2 specified by

$$medv(x) = \beta_1\, crim(x) + \beta_2\, chas(x) + \beta_3\, rad(x) + \beta_4\, tax(x) + \beta_5\, dis(x) + \varepsilon\,(x),$$

with `dis` being the weighted mean of distances to five Boston employment centres, `tax` being the full-value property-tax rate per $10,000 and `rad` being the index of accessibility to radial highways. Give a short explanation of the result!

Solution:

```
m2 <- lm(medv ~ crim + chas + tax + dis + rad, data = Boston)
summary(m2)


##
## Call:
## lm(formula = medv ~ crim + chas + tax + dis + rad, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.69  -4.81  -1.76   2.66  33.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.02075    1.84454   18.44  < 2e-16 ***
## crim        -0.23353    0.05232   -4.46  1.0e-05 ***
## chas         5.16953    1.39167    3.71  0.00023 ***
## tax         -0.03615    0.00516   -7.01  7.8e-12 ***
```

```
## dis          0.02719    0.19871    0.14  0.89122
## rad          0.38243    0.10095    3.79  0.00017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.82 on 500 degrees of freedom
## Multiple R-squared:  0.284,Adjusted R-squared:  0.276
## F-statistic: 39.6 on 5 and 500 DF,  p-value: <2e-16
```

All variables (except $dis(x)$) reveal a P-value below 0.001 and are thus statistically significant at the $0.1\,\%$-significance level. While $crim(x)$ and $tax(x)$ have a negative relationship with housing values, all others are positive. Compared to M_1, the adjusted R^2 increases to 0.1712, showing that the variance of the dependent variable can be explained to a larger degree.

2

l) Calculate and interpret the correlation matrix of model M_2.

Solution:

```
cor(as.data.frame(cbind(Boston$crim, Boston$chas, Boston$tax, Boston$dis, Boston$rad)))
```

```
##          V1        V2       V3       V4        V5
## V1  1.00000 -0.055892  0.58276 -0.37967  0.625505
## V2 -0.05589  1.000000 -0.03559 -0.09918 -0.007368
## V3  0.58276 -0.035587  1.00000 -0.53443  0.910228
## V4 -0.37967 -0.099176 -0.53443  1.00000 -0.494588
## V5  0.62551 -0.007368  0.91023 -0.49459  1.000000
```

Some correlation values (off-diagonal elements) are above 0.4. So this gives an indication that there might be a multicollinearity problem.

2

m) Is there a multicollineartity problem in one of the models? Check using the variance inflation factors.

Solution:

```
library(car)
```

```
vif(m1)
```

```
##  crim  chas
## 1.003 1.003
```

```
vif(m2)
```

```
##  crim  chas   tax   dis   rad
## 1.671 1.031 6.234 1.445 6.375
```

There seems to be no indication of multicollinearity in model M_1 as all variance inflation factors are below 4. However, the VIF of the parameter $tax(x)$ exceeds 4, which might indicate multicollinearity in the second model M_2.

2

n) Calculate and interpret the condition number for model M_2.

Solution:

```
kappa(as.data.frame(cbind(Boston$crim, Boston$chas, Boston$tax, Boston$dis,
    Boston$rad)))

## [1] 1879
```

The condition number is above 30. So this gives an indication that there might be a multicollinearity problem.

2

o) Why does it not make sense to use an OLS estimator in a setting with the first model M_1?

Solution:
The OLS estimator requires three assumptions: no multicollinearity, homoskedasticity and non-autocorrelation. However, the latter two prerequisites are violated. As a result, both t-values and P-values can be erroneous, while coefficients will be correct.

2

p) Compare your models, which one would you choose and why? Take both AIC and BIC into account.

Solution:

```
c(AIC(m1), AIC(m2))

## [1] 3591 3526

c(BIC(m1), BIC(m2))

## [1] 3608 3555
```

According to both information criteria, the second model M_2 has a lower value and, therefore, second model M_2 is preferable. This model M_2 provides a better trade-off between complexity and goodness of fit.

1

q) What do you think about the model specification?

Solution:
One of the most important influences is missing; the size (i. e. the number of rooms). Additionally, reverse

effects are not covered, such as higher housing values attracting home owners that themselves lead to lower crime rates. More precisely, causality is not guaranteed.

1

r) Use model M_1 to make a prediction. What would be your expected median housing value assuming the crime rate is 20 and the Charles river dummy is 1?

Solution:

```
nd <- data.frame(crim = 20, chas = 1)
predict(m1, newdata = nd)
```

```
##     1
## 21.07
```