Statistical Methods - Assignment 3

Michel Mooiweer (1866761) Thomas Webbers (2560695) Eirik Kultorp (2544992)

December 2016

Theoretical exercises

4.1

```
// todo
```

4.2

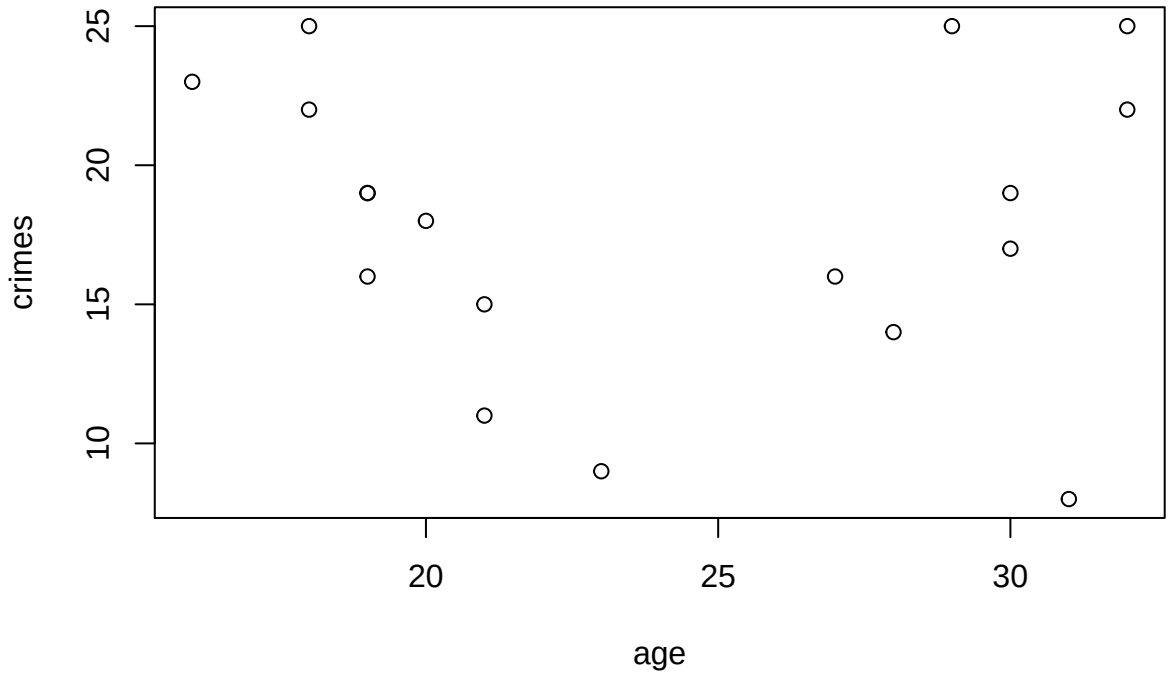
```
// todo
```

R-Exercises

4.3

a)

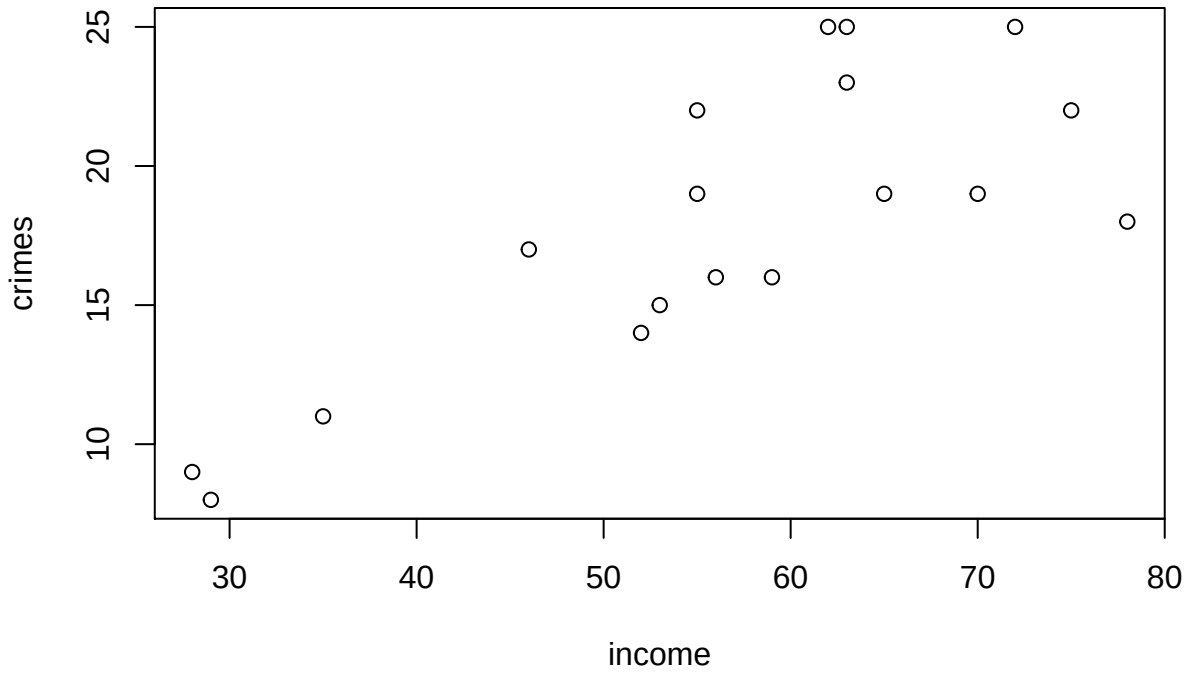
```
##  [1] 16 18 18 19 19 19 20 21 21 23 27 28 29 30 30 31 32 32
##  [1] 23 25 22 16 19 19 18 11 15  9 16 14 25 17 19  8 22 25
```




```
## [1] "Correlation: ( age , crimes ) -0.0709530096415513"
## [1] "Linear correlation seems unlikely"
```

b)

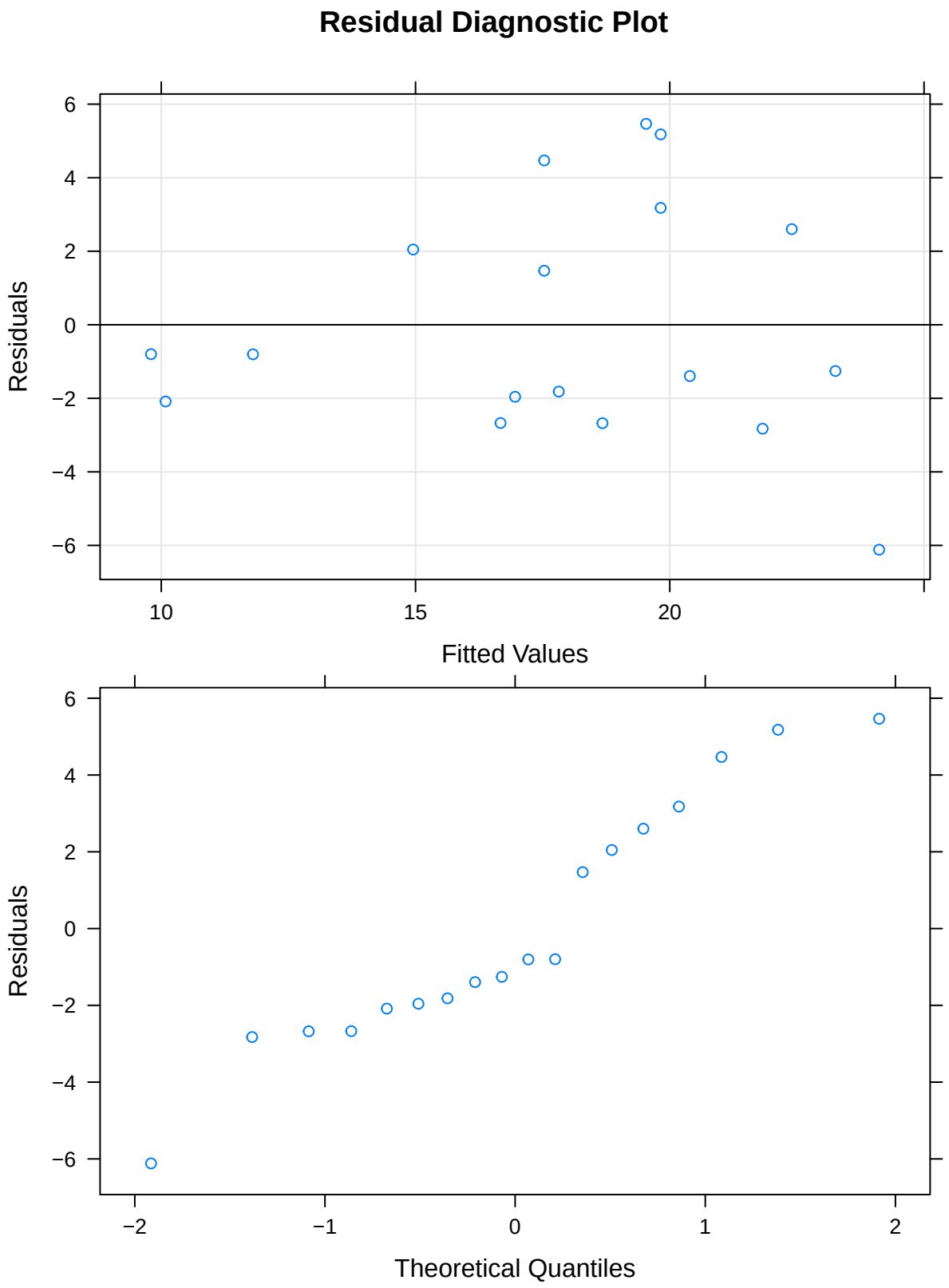
```
##  [1] 63 72 75 59 65 70 78 35 53 28 56 52 63 46 55 29 55 62
##  [1] 23 25 22 16 19 19 18 11 15  9 16 14 25 17 19  8 22 25
```




```
## [1] "Correlation: ( income , crimes ) 0.791557270082001"
## [1] "Linear correlation seems plausible"
```

c)

```
// todo figure how to interpret this output and the plots
// based on https://www.r-bloggers.com/simple-linear-regression-2/
##
## Call:
## lm(formula = crimes ~ income)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.117 -2.054 -1.031  2.462  5.465
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.78111    3.21597   0.554    0.587
## income       0.28636    0.05527   5.181  9.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.315 on 16 degrees of freedom
## Multiple R-squared:  0.6266, Adjusted R-squared:  0.6032
## F-statistic: 26.85 on 1 and 16 DF,  p-value: 9.097e-05
```


Residual Diagnostic Plot
Residuals
Fitted Values
Residuals
Theoretical Quantiles

d)

e)

4.4

a)

$E_9 = 0.046 * \text{total number of files} =$

```
## [1] 6.9
```

b)

- H_0 = the observed leading digits follow Benford's law
  H_a = the digits do not follow Benford's law
- Significance level: $\alpha = 5\%$
- Test statistic:
$$X^2 = \sum_{i=1}^{k} \frac{(o_i - E_i)^2}{E_i}$$
which has a X^2 distribution with $k-1$ degrees of freedom under the null hypothesis
- Observed value(X^2) and P-value:

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 10.299, df = 8, p-value = 0.2447
```

With a significance level of 0.05 we do not have sufficient evidence to reject the null hypothesis and say the digits do not follow Benford's law

4.5

a)

Because we are interested in if each 'subgroup' (which contain the result against one person) follow the same distribution of win, loss, draw we do a test of homeogenity.

- H_0 = andy performs equally well against all opponents
- H_a = andy does not perform equally well against all opponents

b)

- For hypotheses see 4.5a)
- Significance level: $\alpha = 5\%$
- Test statistic:
$$X^2 = \sum_{i,j} \frac{(o_{ij} - E_{ij})^2}{E_{ij}}$$
which has a X^2 distribution with $(rows-1)(columns-1)$ degrees of freedom under the null hypothesis

- Observed value(X^2) and P-value:

```
##
##  Pearson's Chi-squared test
##
## data:  result
## X-squared = 4.7235, df = 6, p-value = 0.5797
```

-Because the p value is not below our significance level we can not reject the null hypothesis. Andy might perform equally well against all opponents

c)

If they play 96 games he is expected to win:

```
## [1] 43.243
```

Appendix

4.3.a

```
dat=matrix(as.numeric(as.matrix(read.table("crimemale.txt"))[2:19,]),ncol=3)
age=dat[,1]
income=dat[,2]
crimes=dat[,3]

investigate_linear_correlation <- function(v1,v2,xlab,ylab){
  print(v1)
  print(v2)
  plot(v1,v2,xlab=xlab,ylab=ylab)
  corr=cor(v1,v2)
  print(paste("Correlation: (",xlab,",",ylab,")",corr))
  corr=abs(corr)

  # TODO adjust these thresholds based on statistical standards (if they exist)

  if (corr<0.7) w ="unlikely"
  else if (corr<0.8) w = "plausible"
  else w="likely"
  print(paste("Linear correlation seems",w))
}

investigate_linear_correlation(age,crimes,"age","crimes")
```

4.3.b

```
investigate_linear_correlation(income,crimes,"income","crimes")
```

4.3.c

```
lmres = lm(crimes ~ income)
summary(lmres)
library("lattice")
xyplot(resid(lmres) ~ fitted(lmres),
  xlab = "Fitted Values",
  ylab = "Residuals",
  main = "Residual Diagnostic Plot",
  panel = function(x, y, ...)
  {
    panel.grid(h = -1, v = -1)
    panel.abline(h = 0)
    panel.xyplot(x, y, ...)
  }
)
qqmath( ~ resid(lmres),
  xlab = "Theoretical Quantiles",
  ylab = "Residuals"
)
```

4.3.d

4.3.e

4.4.a

```
sum(c(45,32,18,12,9,3,13,9,9))*0.046
```

4.4.b

```
expected <- c(0.301,0.176,0.125,0.097,0.079,0.067,0.058,0.051,0.046)
observed <-c(45,32,18,12,9,3,13,9,9)
chisq.test(observed, p=expected)
```

4.5.a

4.5.b

```
result <- matrix(c(179,96,52,39,47,17,13,15,57,36,18,15), ncol = 3)
colnames(result) <-  c('won', 'lost', 'draw')
rownames(result) <- c('Bob', 'Cecilia', 'David', 'Emma')
chisq.test(result)
```

4.5.c

```
round(chisq.test(result)$exp['Emma', 'won'],3)
```