Statistical Methods - Assignment 4

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

December 2016

Theoretical exercises

4.1

```
// todo
```

4.2

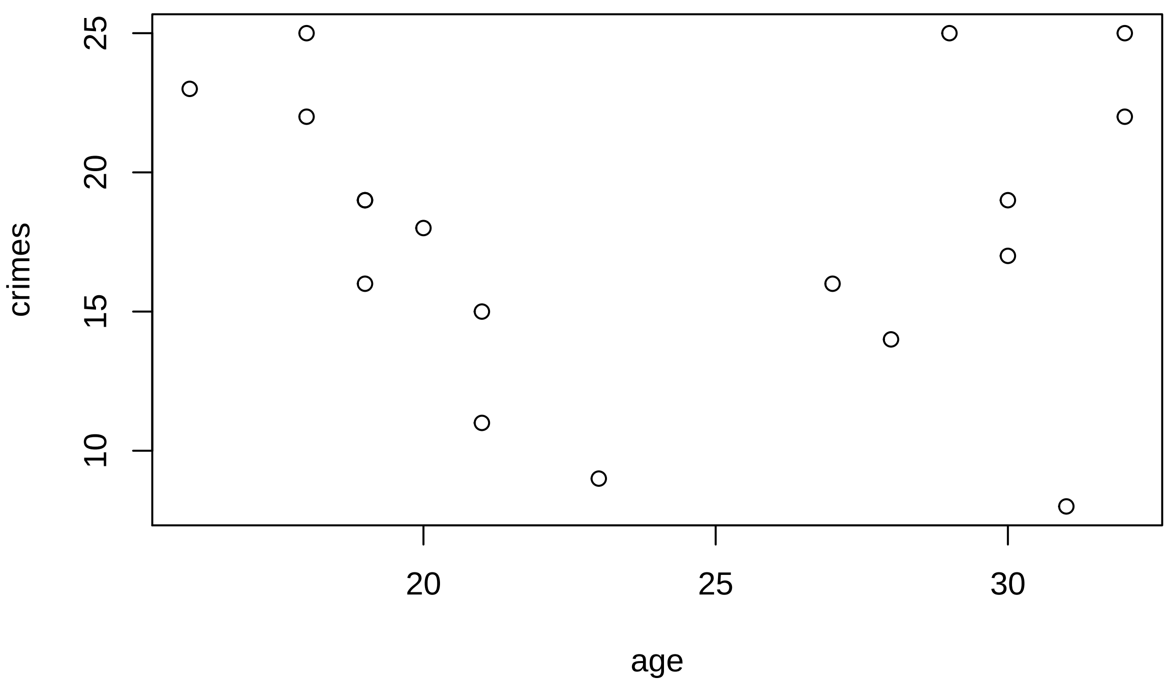
```
// todo
```

R-Exercises

4.3

a)

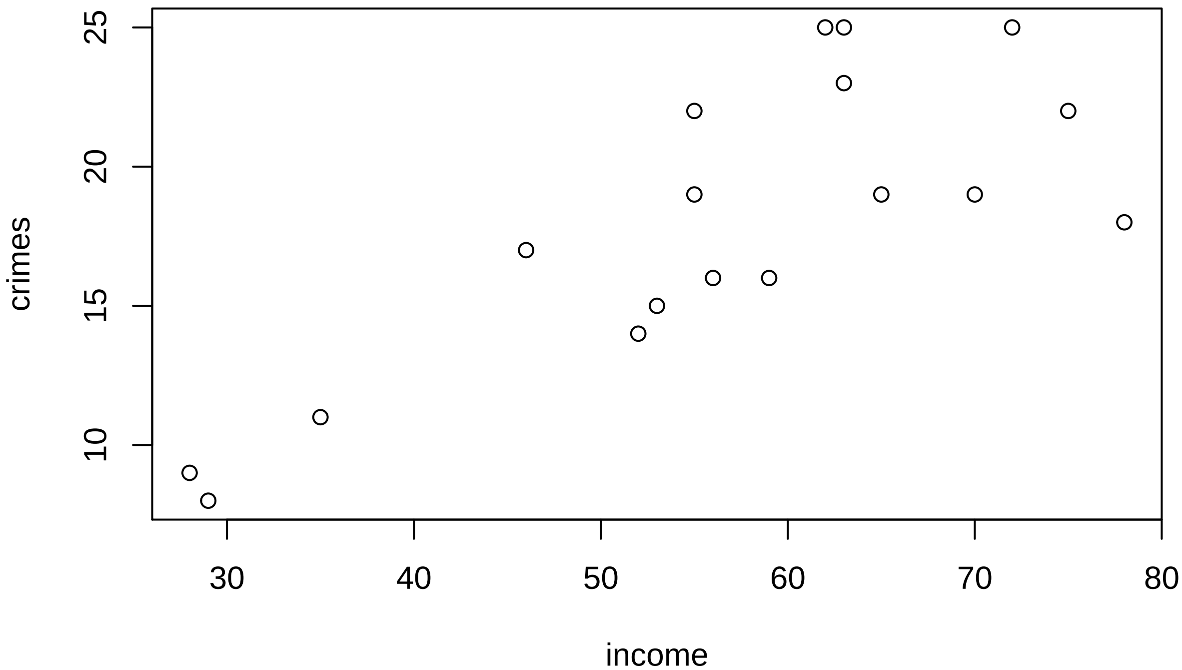
```
##  [1] 16 18 18 19 19 19 20 21 21 23 27 28 29 30 30 31 32 32
##  [1] 23 25 22 16 19 19 18 11 15  9 16 14 25 17 19  8 22 25
```




```
## [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"
```

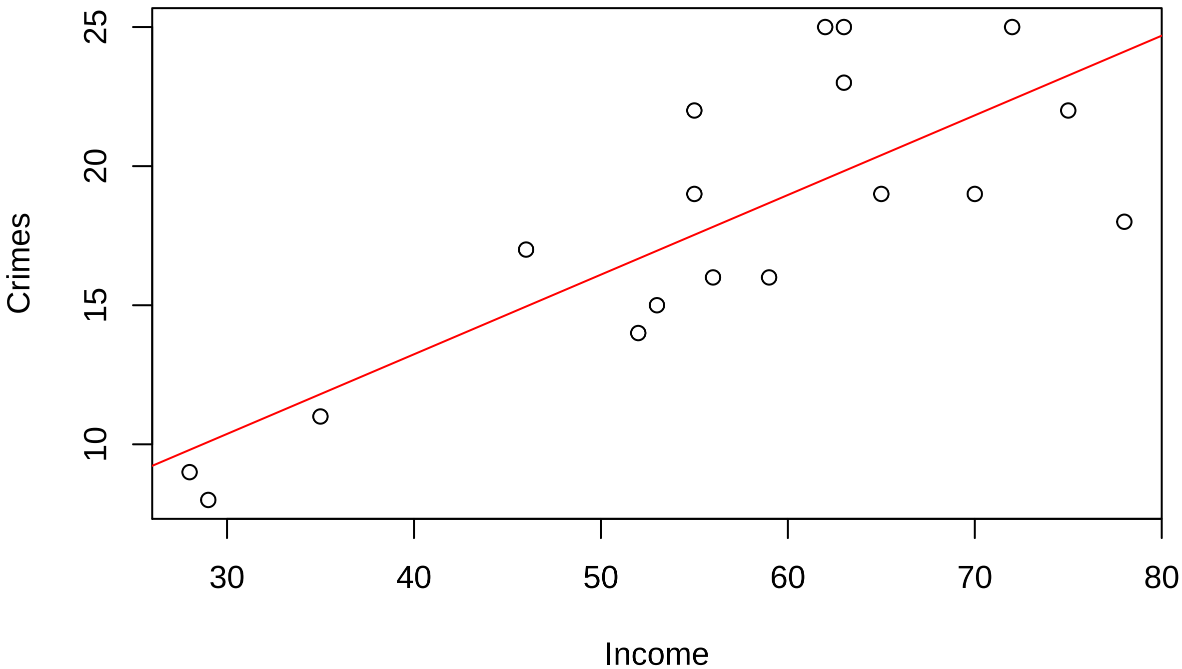
c)

```
##
## 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
```

```
## [1] "Intercept ="
```

```
## (Intercept)
##    1.781109
```

```
## [1] "Slope ="
## income
## 0.2863583
```



d)

- $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$
- Significance level $\alpha = .05$
- Test statistic:

$$T_\beta = b_1 / s_{b_1}$$

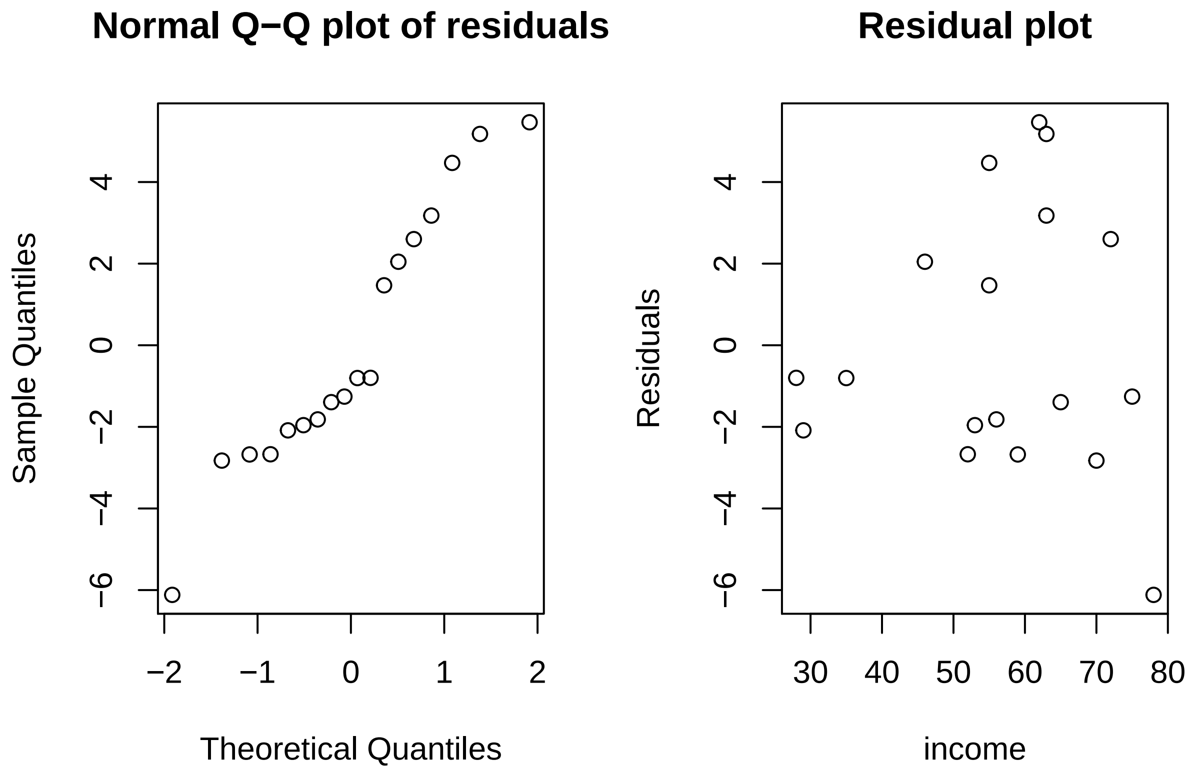
  has a t-distribution with df = n-2 = 16 under H_0
- Observed value:

```
## [1] 5.18108
```

- Critical values: two-tailed test and $\alpha = 0.05$ so $-t_{16,0.025}$ and $t_{16,0.025}$ i.e. -2.120 and 2.120
- Since $5.181 > 2.120$ we reject H_0
- There is sufficient evidence to warrant a rejection of the claim that there is no linear relationship between income and crime

e)

Requirements for testing linearity: - Independence: (difficult to check) - Normal distribution of residuals - Fixed standard deviation



(Visusally) The residual plot shows no obvious pattern
The Q-Q plot seems to approach normal distribution

So the requirements are met

4.4

a)

$E_9 = 0.046 *$ 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 69 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]
```

```r
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

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

4.3.c

```r
lmres = lm(crimes ~ income)
summary(lmres)
"Intercept ="
lmres$coefficients[1]
"Slope ="
lmres$coefficients[2]
plot(income,crimes,xlab="Income",ylab="Crimes")
abline(lmres$coefficients, col='red')
```

4.3.d

```r
unname(lmres$coefficients[2]/0.05527)
```

4.3.e

```r
par(mfrow=c(1,2))
qqnorm(lmres$res, main = "Normal Q-Q plot of residuals")
plot(income, lmres$res, ylab="Residuals", main="Residual plot")
```

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)
```