Computer Assignment 2: Multiple Linear Regression

Data files: The assignment uses the data files *'faithful'*, which is already in the R package, *safety.csv*, *welding.csv*, *house.csv* and *election.csv*. The last four may be found in the course directory.

Exercise 1: Simple Linear Regression

A geyser is a warm spring, from which water erupts on a regular basis. One of the more well known geysers is Old Faithful Geyser, in Wyoming. The length of an eruption together with the time until the next eruption has been measured for a large number of eruptions and, based on the data available, the question of interest is to predict the time until the next eruption based on the length of an eruption. The data is found in the file 'faithful'. Type

```
> ?faithful
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
> tail(faithful)
    eruptions waiting
267     4.750      75
268     4.117      81
269     2.150      46
270     4.417      90
271     1.817      46
272     4.467      74
```

We shall analyse the data according to the model:

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

where y denotes the waiting time, which is a response to x, the eruption time and

$$\varepsilon_1, ..., \varepsilon_n \text{ are independent } N(0, \sigma^2).$$

1. Start by plotting y against x. This will establish whether or not a linear regression model is of interest.

   ```
   > plot(faithful$eruptions,faithful$waiting)
   ```

   The plot should indicate that the relationship is linear.

   The correlation coefficient is a measure of the strength of linear association. This is computed as follows:

```
> cor(faithful$eruptions, faithful$waiting)
[1] 0.9008112
```

This also indicates a strong linear relationship, and hence that a regression analysis could be useful. To get confidence intervals on the correlation coefficient, type:

```
> cor.test(faithful$waiting, faithful$eruption)

Pearson's product-moment correlation

data:  faithful$waiting and faithful$eruption
t = 34.089, df = 270, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8756964 0.9210652
sample estimates:
      cor
0.9008112
```

2. Now try a regression analysis. The letters 'lm' stand for *linear model*.

```
 > ?lm
```

will give a description in the bottom right window.

```
 > x<-lm(waiting~eruptions, faithful)
```

The analysis of variance is now obtained by:

```
> anova(x)
Analysis of Variance Table

Response: faithful$waiting
                          Df  Sum Sq Mean Sq  F value    Pr(>F)
faithful$eruptions     1     40644   40644  1162.1    < 2.2e-16 ***
Residuals              270   9443    35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient estimates may be obtained by:

```
> summary(x)

Call:
lm(formula = faithful$waiting ~ faithful$eruptions)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122   3.9246  15.9719

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         33.4744     1.1549   28.98   <2e-16 ***
faithful$eruptions  10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared: 0.8115,Adjusted R-squared: 0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

The estimated regression model is:

$$y = 33.4744 + 10.7296x + \epsilon$$

$$\epsilon \sim N(0, \sigma^2) \qquad s = 5.914$$

Plot the estimated regression line:

```
> plot(faithful$waiting, faithful$eruptions)
> abline(x)
```

Alternatively, there is a nice package in the ggplot2 library, which also gives 95% confidence intervals for the regression line.

```
> install.packages("ggplot2")
> library("ggplot2")
> qplot(waiting, eruptions, data=faithful) + geom_smooth(method=lm)
```

3. The output gives the estimated standard deviations of the coefficients. Let $h = (X^t X)^{-1}$, then $d(\hat{\beta}_0) = s\sqrt{h_{00}}$ and $d(\hat{\beta}_1) = s\sqrt{h_{11}}$. Write out the formula for s^2, the estimate of σ^2. How many degrees of freedom are associated with the variance estimate?

4. Test, at a significance level of $\alpha = 0.01$,

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

5. Now check whether or not the assumption that the errors are normal is reasonable.

```
> z<-x$residuals
```

The column z now contains the residuals.

In the bottom right hand side, check under packages to see if 'car' is there. If not, type

```
> install.packages("car")
> library("car")
```

The standard qqnorm does not give confidence intervals; the one in the 'car' library does. When the package is there, check the box for 'car' under 'packages'.

```
> qqnorm(z)
> qqline(z)
```

gives a normal probability plot, with a straight line indicating the perfect fit. The command

```
 > qqPlot(z)
```

gives, in addition, the Kolmogorov Smirnov confidence bounds.

A histogram can also be used as a diagnostic; make a histogram of the residuals and superimpose a normal density function.

```
> m = mean(z)
> s = sd(z)
> hist(z,probability=TRUE)
> pt <- seq(min(z),max(z),length = 40)
> lines(pt,dnorm(pt,mean=m,sd=s),col = "red")
```

Alternatively, try:

```
> h<-hist(z, xlab="residual", main="Histogram with Normal Curve")
> xfit<-seq(min(z),max(z),length=40)
> yfit<-dnorm(xfit,mean=mean(z),sd=sd(z))
> yfit <- yfit*diff(h$mids[1:2])*length(z)
> lines(xfit, yfit, col="red", lwd=2)
```

6. An erruption of 4 minutes has just taken place. Compute a 95% prediction interval for the time to the next eruption.

```
> attach(faithful)
> x<-lm(waiting~eruptions, data=faithful)
> newdata = data.frame(eruptions = 4)
> predict(x, newdata, interval = "predict")
       fit      lwr     upr
1 76.39296 64.72382 88.0621
> detach(faithful)
```

Check the syntax for 'predict'; 95% is the default.

The estimated waiting time is 76.39 minutes; the 95% prediction interval is $(64.72, 88.06)$ minutes.

Fill in the control sheet and add any interesting plots.

Exercise 2: Using dummy variables and computing prediction intervals

The data is found in the course directory in the file 'safety.csv'. The aim of the exercise is to determine whether an active safety programme at the work place is associated with the number of working hours lost through accidents at the work place. 40 companies have been chosen at random, 20 with an active safety programme and 20 without. For each company, the three pieces of data are:

$$
\begin{aligned}
Y &= \text{number of working hours lost in a year,} \\
x_1 &= \text{number of employees,} \\
x_2 &= \begin{cases} 1, & \text{if there is an active safety programme;} \\ 0, & \text{otherrwise.} \end{cases}
\end{aligned}
$$

number employed (x_1)	safety programme (x_2)	number of working hours lost (y) (yearly, thousands of hours)
6490	0	121
7244	0	169
⋮	⋮	⋮
5526	0	123
3077	1	44
6600	1	73
⋮	⋮	⋮
1701	1	6

Firstly, import the data set. This is best done by clicking on 'Import data set', finding the data set in the appropriate directory, and clicking on it.

Obs: Under 'headers', click on 'yes'. This means that the first row, containing y, $x1$, $x2$ will be treated as headers. Otherwise, $y, x1, x2$ will be treated as the first row of data and there will be problems because it is not numerical.

To decide which models are suitable for analysing the data, one starts by plotting y against x_1 with different symbols for each value of x_2.

```
> install.packages("ggplot2")
> library("ggplot2")
```

Now try

```
> qplot(x1,y, colour = x2, data = safety)
```

Now consider two different models:

$$\textbf{Model 1:} \quad Y = \gamma_0 + \gamma_1 z_1 + \gamma_2 x_2 + \tilde{\varepsilon},$$

where $z_1 = x_1/1000$ and $x_2 = 1$ with a safety programme and $x_2 = 0$ without, and

$$\textbf{Model 2:} \quad Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon,$$

where $z_2 = x_2 z_1$.

For Model 1, the expected values are:

$$\mathbb{E}[Y] = \begin{cases} \gamma_0 + \gamma_1 z_1 + \gamma_2, & \text{with safety programme,} \\ \gamma_0 + \gamma_1 z_1, & \text{without safety programme.} \end{cases}$$

The expected value for Model 2 satisfies:

$$\mathbb{E}[Y] = \begin{cases} \beta_0 + (\beta_1 + \beta_2) z_1, & \text{with safety programme,} \\ \beta_0 + \beta_1 z_1, & \text{without safety programme.} \end{cases}$$

For both models, sketch the lines corresponding to the expected values both with and without a safety programme, as a function of the number of employees. How do they compare? Using the plot made above, explain why model 1 is not as good as model 2.

Carry out regression analyses corresponding to the models 1 and 2. This may be done using the command

```
> fit1 <- lm(y~I(x1/1000)+ x2, safety)
> fit2 <- lm(y~I(x1/1000) + I(x1*x2/1000), safety)
```

In each case decide which explanatory variables are useful. In each case, compute a prediction inteval for a company with safety programme and 6600 employees.

```
> attach(safety)
> new <- data.frame(x1 = 6600, x2 = 1)
> predict(fit1,new,interval="prediction", level = 0.95)
       fit      lwr      upr
1 67.64132 26.14405 109.1386
> predict(fit1,new,interval="confidence",level=0.95)
       fit      lwr     upr
1 67.64132 58.41485 76.8678
> predict(fit2,new,interval="prediction", level = 0.95)
       fit      lwr      upr
1 63.59839 34.77465 92.42213
> detach(safety)
```

1. Write down the estimated regression line in each case.

2. Compare the error sum of squares for each model.

3. Write down both prediction intervals.

4. What is the difference between a confidence interval and a prediction interval?

5. Does the second model indicate that the safety programme leads to fewer lost working hours on average? Answer the question by considering an appropriate parameter in the model.

```
> confint(fit1, level=0.95)
                2.5 %    97.5 %
(Intercept)  14.32461  49.01473
I(x1/1000)   11.79527  16.74854
x2          -71.01977 -45.42604
> confint(fit2, level=0.95)
                     2.5 %    97.5 %
(Intercept)      -9.096787 13.113284
I(x1/1000)       17.465864 21.149302
I(x1 * x2/1000) -11.350460 -8.601027
```

Fill in the summary sheet and add any interesting plots.

Exercise 3: Linearity

With electrical welding, the strength of the current influences the strength of the weld. In an experiment, six different current strengths were used on three trials each. The results (in suitable units of strength) were as follows:

current	strength		
4000	600	750	550
5000	1850	2000	1750
6000	2700	2650	2650
7000	3650	3800	3600
8000	4400	4550	4350
9000	4900	4800	4850

The data is found in 'welding.csv' in the course directory.

Plot 'strength' y against 'current' x. y mot x. The plot should indicate that there is a linear association, but that it is not entirely apprpriate:

$$\textbf{Model 1:} \quad Y = \beta_0 + \beta_1 x + \varepsilon \text{ where } \varepsilon \sim N(0, \sigma_1^2)$$

Now make a regression analysis, plot y against x and add the best fitting regression line.

```
> fit <- lm(strength~current, welding)
> plot(welding$current, welding$strength)
> abline(fit)
```

Does the straight line seem appropriate?
Now look at the residuals:

```
> z <- residuals(fit)
> plot(welding$current, z)
```

What is s_1, the estimate of σ_1? You'll find it by typing

```
> aov(fit)
Call:
   aov(formula = fit)

Terms:
                 current Residuals
Sum of Squares  38058857    857254
Deg. of Freedom        1        16

Residual standard error: 231.47
Estimated effects may be unbalanced
```

In the residual plot, there is a clear pattern, which suggests that the linear model is not satisfactory. An additional quadratic term may provide a better fit.

Try

```
> fit2 <- lm(strength~current + I(current^2), welding)
> plot(welding$current, welding$strength)
> points(welding$current,predict(fit2),type="l",col="red")
```

This is the model:

$$\textbf{Modell 2:} \quad Y = \beta_0' + \beta_1' x + \beta_2' x^2 + \varepsilon'' \text{ where } \varepsilon'' \sim N(0, \sigma_2^2).$$

Denote the estimate of σ_2 by $\hat{\sigma}_2 = s_2$. What is it?

Plot the residuals against 'current'. Do the residuals for Model 2 appear better than those for Model 1? If so, why?

Would it be appropriate to use this regression analysis to predict the weld strength for a current of 12000?

Fill in the control sheet and print out interesting plots.

Exercise 4: Stepwise Regression

The data for this exercise is found in the file 'house.csv' in the course directory. The six variables are:

1. property taxes (dollars)
2. House size (square feet)
3. Plot size (acres)
4. Plot size squared
5. Attractiveness index

6. Selling price.

The aim is to find a suitable regression model to explain the median house price. Do this using stepwise regression.
First, find the explanatory variable with the strongest correlation to the selling price:

```
 > cor(house)
                       tax  housesize    plotsize  plotsizesq attractiveindex sellingprice
tax             1.00000000  0.2430790  0.81446708  0.80378262     -0.02144253    0.4396969
housesize       0.24307896  1.0000000 -0.28982630 -0.26804966     -0.12844139    0.6258176
plotsize        0.81446708 -0.2898263  1.00000000  0.98477889     -0.01889162    0.1206351
plotsizesq      0.80378262 -0.2680497  0.98477889  1.00000000     -0.06571432    0.1181183
attractiveindex -0.02144253 -0.1284414 -0.01889162 -0.06571432      1.00000000    0.3610034
sellingprice     0.43969687  0.6258176  0.12063505  0.11811831      0.36100336    1.0000000
```

This would appear to be the house size. You can use cor.test (as in the first exercise) to get confidence intervals on the correlation coefficients. Make a regression based on this variable

```
> fit <- lm(sellingprice~housesize, house)
```

Is the coefficient β_1 connected with 'house size' significant?

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 219.51726444 411.9688846
housesize     0.03500754   0.1002785
```

(yes)

What is the residual sum of squares? the associated degrees of freedom?

Having established that the β coefficient is significant, now try to find the best model, containing 'house size' plus one other variable.

```
> fit21 <- lm(sellingprice~housesize + plotsize, house)
> anova(fit21)
Analysis of Variance Table

Response: sellingprice
          Df Sum Sq Mean Sq F value    Pr(>F)
housesize  1  36283   36283 20.7842 9.962e-05 ***
plotsize   1   9225    9225  5.2844   0.02949 *
Residuals 27  47134    1746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(fit22)
Analysis of Variance Table
```

```
Response: sellingprice
                Df Sum Sq Mean Sq F value    Pr(>F)
housesize        1  36283   36283  25.775 2.477e-05 ***
attractiveindex  1  18351   18351  13.036  0.001228 **
Residuals       27  38007    1408
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Clearly ‘attractiveness index’ is more significant than ‘plot size’. Both are significant, but according to the forward principle, choose ‘attractiveness index’ rather than ‘plot size’ for the second variable. Try the others and take the one with the largest sum of squares.

While the p value of the additional variable is less than 0.05, the additional variable is significant. Continue until there are no more significant variables.

Which model does the forward principle produce?

Now make two regression analyses; one with all the variables and the other with only the variables that the forward model produced.

1. Which variables appear to be significant according to the full model?
2. By comparing the full model with the ‘forward principle’ model and using an appropriate F test, determine whether or not the omitted parameters are significant (at a 5% significance level).

The explanation is that ‘plot size’ and ‘plot size squared’ are strongly correlated. This leads to ill conditioning of the X^tX matrix and hence to *variance inflation*, as described in the lectures.

What is the correlation between ‘plot size’ and ‘plot size squared’?

Stepwise regression is (of course) automated in R.

```
> step(lm(sellingprice~1, data = house), scope = list(upper = ~tax + housesize
+ plotsize + plotsizesq + attractiveindex, lower = ~1), direction = "forward")
```

Which model is returned? What happens if you change ‘forward’ to ‘backward’? Which model is returned if you use ‘both’?
Fill in the control sheet.

Exercise 5: Outliers

The data for this exercise is from the USA presidential elections of 2000. There was much discussion over the result from Florida; the proportion of spoiled ballot papers in Palm Beach seemed substantially higher than the proportion in other places, which seemed to favour Pat Buchanan.

The table below gives the results for some of the electoral districts in Florida.

County	Bush	Gore	Buchanan	Total
Alachua	34124	47365	263	85729
Baker	5610	2392	73	8154
⋮	⋮	⋮	⋮	⋮
Palm Beach	152951	269732	3411	433186
⋮	⋮	⋮	⋮	⋮
Washington	4994	2798	88	8025
Absentee	1575	836	5	2490

The data is found in the file 'election.csv'. Palm Beach is not included in this data set; all the other districts are included.

Firstly, plot Buchanan's vote against the total.

```
> plot(election$total,election$buchanan)
```

There is clearly an association between Y, buchanan's vote and x the total vote, but the variance seems larger for larger values of x. The residuals are not identically distributed. Therefore, it seems reasonable to plot the logarithms (`log10` gives log base 10). Make sure `ggplot2` is operating and type

```
> qplot(log10(total),log10(buchanan),data=election)
```

Now do a regression with the logarithmic values:

```
> fit <- lm(I(log10(buchanan))~I(log10(total)), data=election)
```

Plot the logarithmic values against each other and add a regression line:

```
> plot(log10(election$total),log10(election$buchanan))
> abline(fit)
```

Now make a 99.5% prediction interval for Buchanan's vote in Palm Beach (where the total number voting was 433186.

Using the prediction interval for $\log_{10}$(Buchanan), compute a prediction interval for the number of voters for Buchanan in Palm Beach and compare with the actual number of 3411.

```
> attach(election)
> newdata <- data.frame(total=433186)
> ?predict
> a<-predict(fit,newdata,interval="prediction",level=0.995)
> a
       fit      lwr      upr
1 2.901225 2.296942 3.505508
> 10^a
       fit      lwr     upr
1 796.5723 198.1264 3202.64
> detach(election)
```

Fill in the control sheet; print off any interesting plots.

Note We have taken the most deviant value and shown that it is unlikely. However, in 68 observations, the Buonferoni upper bound on the significance is $68 \cdot 0.005 = 0.34$, which is rather large. It is therefore inappropriate to read too much into the result.

Control Sheet

Names:

1)
Erik Ekelund, Niklas Ericson Hans-Filip Elo

2)

3)
S^2= 1/(n+m-1) * ((m-1)s1^2 +(n-1)s2^2)

Exercise 1

a) Estimated correlation between y and x: 0,85

b) Estimated regression line 33,82 + 10,7x

c) $d(\hat{\beta}_0) =$ 2,2618 $d(\hat{\beta}_1) =$ 0,6263

Formula: $s^2 =$ —— Degrees of freedom for s^2: 105

d) Test statistic $T =$ 14,96 ; critical value for $|T|$: 2,58 ; Reject H_0? Ja ty 14,96>2,58

e) Prediction interval for time to the next eruption: 70, 83,5

$(\mathbf{X}^T\mathbf{X})^{-1} =$

e) Is the residual plot OK? Nej, den måste vara lika bred kring x-axeln

Exercise 2

a) Write out the regression lines for the two models.

Model 1: 31,67 + 14,27z1 - 58,22x2

Model 2: -2,01+ 19,31z1-9,98z2

Give a short explanation of the differences between the two models and why model 2 is better.
Ur datat ser vi att det har skett en ändring i lutning mellan de två urvalsgrupperna. modell1 ger direkt lodrät translatering medan model 2 ändrar lutning därför är model 2 bättre.

	Modell 1	Modell 2
b) R^2-value:	0,8592	0,9322
c) Prediction interval:	26,3, 109	34.9, 92.3

d) Confidence interval: [-11,33 ; -8,62]

Exercise 3

a) $s_1 = \ldots\ldots\ldots$ s1=231.47

b) Is there a pattern in the residual plot? Sketch it: Ser ut som en andragradare

c) $s_2 = \ldots\ldots\ldots$ s2=116,05

d) Is the residual plot OK? Nej Is model 2 OK? Ja

e) Answer: Ja, Why? Eftersom vi hittills bara haft fel <200 är det rimligt att anta att vi kan fortsätta använda modellen för att skatta med strömstyrkor upp mot 12 000.

Exercise 4

Best single explanatory variable: x3

a) Model according to the forward principle (write in the theoretical model, not the estimated parameters)
y=beta0 + x2*beta1 + x5*beta2 +*x4*beta3

b) Test statistic $v =$ -2.5065 Critical boundary: .....
Conclusion:

c) $\hat{\beta}_3 =$ 31,6 $T_3 =$ 0 $\hat{\beta}_4 =$ 2,4 $T_4 =$ 0
Critical boundary for $|T_i|$: 2,05
Correlation between x_3 och x_4: 0,98

Exercise 5

a) Does the estimated regression line appear reasonable? Yes, it seems linear
b) Prediction interval for Buchanan's vote: 10^2.33, 10^3.47