Chapter 6

Multiple Linear Regression (solutions to exercises)
# Contents

6  **Multiple Linear Regression (solutions to exercises)**  1  
  6.1  Nitrate concentration ........................................... 3  
  6.2  Multiple linear regression model ............................. 6  
  6.3  MLR simulation exercise ........................................ 11
6.1 Nitrate concentration

Exercise 6.1 Nitrate concentration

In order to analyze the effect of reducing nitrate loading in a Danish fjord, it was decided to formulate a linear model that describes the nitrate concentration in the fjord as a function of nitrate loading, it was further decided to correct for fresh water runoff. The resulting model was

\[ Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \]  

(6-1)

where \( Y_i \) is the natural logarithm of nitrate concentration, \( x_{1,i} \) is the natural logarithm of nitrate loading, and \( x_{2,i} \) is the natural logarithm of fresh water runoff.

a) Which of the following statements are assumed fulfilled in the usual multiple linear regression model?

1) \( \epsilon_i = 0 \) for all \( i = 1, ..., n \), and \( \beta_j \) follows a normal distribution
2) \( E[x_1] = E[x_2] = 0 \) and \( V[\epsilon_i] = \beta_1^2 \)
3) \( E[\epsilon_i] = 0 \) and \( V[\epsilon_i] = \beta_1^2 \)
4) \( \epsilon_i \) is normally distributed with constant variance, and \( \epsilon_i \) and \( \epsilon_j \) are independent for \( i \neq j \)
5) \( \epsilon_i = 0 \) for all \( i = 1, ..., n \), and \( x_j \) follows a normal distribution for \( j = \{1, 2\} \)

Solution

1) \( \epsilon_i \) follows a normal distribution with expectation equal zero, but the realizations are not zero, and further \( \beta_j \) is deterministic and hence it does not follow a distribution (\( \hat{\beta}_j \) does), hence 1) is not correct
2)-3) There are no assumptions on the expectation of \( x_j \) and the variance of \( \epsilon \) equal \( \sigma^2 \), not \( \beta_1^2 \) hence 2) and 3) are not correct
4) Is correct, this is the usual assumption about the errors
5) Is incorrect since \( \epsilon_j \) follow a normal distribution, further the are no distributional assumptions on \( x_j \). In fact we assume that \( x_j \) is known
The parameters in the model were estimated in R and the following results are available (slightly modified output from summary):

```R
> summary(lm(y ~ x1 + x2))
```

**Call:**

```R
lm(formula = y ~ x1 + x2)
```

**Coefficients:**

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -2.36500 | 0.22184    | -10.661 | < 2e-16  |
| x1             | 0.47621  | 0.06169    | 7.720   | 3.25e-13 |
| x2             | 0.08269  | 0.06977    | 1.185   | 0.237    |

---

Residual standard error: 0.3064 on 237 degrees of freedom
Multiple R-squared: 0.3438, Adjusted R-squared: 0.3382
F-statistic: 62.07 on 2 and 237 DF, p-value: < 2.2e-16

b) What are the parameter estimates for the model parameters ($\hat{\beta}_i$ and $\hat{\sigma}^2$) and how many observations are included in the estimation?

**Solution**

The number of degrees of freedom is equal $n - (p + 1)$, and since the number of degrees of freedom is 237 and $p = 2$, we get $n = 237 + 2 + 1 = 240$. The parameters are given in the first column of the coefficient matrix, i.e.

\[
\hat{\beta}_0 = -2.365 \quad (6-2)
\]

\[
\hat{\beta}_1 = 0.476 \quad (6-3)
\]

\[
\hat{\beta}_2 = 0.083 \quad (6-4)
\]

and finally the estimated error variance is $\hat{\sigma}^2 = 0.3064^2$.

c) Calculate the usual 95% confidence intervals for the parameters ($\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\beta}_2$).
Solution

From Theorem 6.5 we know that the confidence intervals can be calculated by

\[ \hat{\beta}_i \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_i}, \]

where \( t_{1-\alpha/2} \) is based on 237 degrees of freedom, and with \( \alpha = 0.05 \), we get \( t_{0.975} = 1.97 \). The standard errors for the estimates is the second column of the coefficient matrix, and the confidence intervals become

\[
\begin{align*}
\hat{\beta}_0 &= -2.365 \pm 1.97 \cdot 0.222 \quad (6-5) \\
\hat{\beta}_1 &= 0.467 \pm 1.97 \cdot 0.062 \quad (6-6) \\
\hat{\beta}_2 &= 0.083 \pm 1.97 \cdot 0.070 \quad (6-7)
\end{align*}
\]

d) On level \( \alpha = 0.05 \) which of the parameters are significantly different from 0, also find the \( p \)-values for the tests used for each of the parameters?

Solution

We can see directly from the confidence intervals above that \( \beta_0 \) and \( \beta_1 \) are significantly different from zero (the confidence intervals does not cover zero), while we cannot reject that \( \beta_2 = 0 \) (the confidence interval cover zero). The \( p \)-values we can see directly in the R output: for \( \beta_0 \) is less than \( 10^{-16} \) and the \( p \)-value for \( \beta_1 \) is \( 3.25 \cdot 10^{-13} \), i.e. very strong evidence against the null hypothesis in both cases.
6.2 Multiple linear regression model

Exercise 6.2 Multiple linear regression model

The following measurements have been obtained in a study:

<table>
<thead>
<tr>
<th>No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1.45</td>
<td>1.93</td>
<td>0.81</td>
<td>0.61</td>
<td>1.55</td>
<td>0.95</td>
<td>0.45</td>
<td>1.14</td>
<td>0.74</td>
<td>0.98</td>
<td>1.41</td>
<td>0.81</td>
<td>0.89</td>
</tr>
<tr>
<td>x1</td>
<td>0.58</td>
<td>0.86</td>
<td>0.29</td>
<td>0.20</td>
<td>0.56</td>
<td>0.28</td>
<td>0.08</td>
<td>0.41</td>
<td>0.22</td>
<td>0.35</td>
<td>0.59</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>x2</td>
<td>0.71</td>
<td>0.13</td>
<td>0.79</td>
<td>0.20</td>
<td>0.56</td>
<td>0.92</td>
<td>0.01</td>
<td>0.60</td>
<td>0.70</td>
<td>0.73</td>
<td>0.13</td>
<td>0.96</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>0.68</td>
<td>1.39</td>
<td>1.53</td>
<td>0.91</td>
<td>1.49</td>
<td>1.38</td>
<td>1.73</td>
<td>1.11</td>
<td>1.68</td>
<td>0.66</td>
<td>0.69</td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td>x1</td>
<td>0.12</td>
<td>0.65</td>
<td>0.70</td>
<td>0.30</td>
<td>0.70</td>
<td>0.39</td>
<td>0.72</td>
<td>0.45</td>
<td>0.81</td>
<td>0.04</td>
<td>0.20</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>0.21</td>
<td>0.88</td>
<td>0.30</td>
<td>0.15</td>
<td>0.09</td>
<td>0.17</td>
<td>0.25</td>
<td>0.30</td>
<td>0.32</td>
<td>0.82</td>
<td>0.98</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

It is expected that the response variable $y$ can be described by the independent variables $x_1$ and $x_2$. This imply that the parameters of the following model should be estimated and tested

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

a) Calculate the parameter estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$), in addition find the usual 95% confidence intervals for $\beta_0$, $\beta_1$, and $\beta_2$.

You can copy the following lines to R to load the data:

```r
D <- data.frame(
    x1=c(0.58, 0.86, 0.29, 0.20, 0.56, 0.28, 0.08, 0.41, 0.22, 0.35,
         0.59, 0.22, 0.26, 0.12, 0.65, 0.70, 0.30, 0.70, 0.39, 0.72,
         0.45, 0.81, 0.04, 0.20, 0.95),
    x2=c(0.71, 0.13, 0.79, 0.20, 0.56, 0.92, 0.01, 0.60, 0.70, 0.73,
         0.13, 0.96, 0.27, 0.21, 0.88, 0.30, 0.15, 0.09, 0.17, 0.25,
         0.30, 0.32, 0.82, 0.98, 0.00),
    y=c(1.45, 1.93, 0.81, 0.61, 1.55, 0.95, 0.45, 1.14, 0.74, 0.98,
        1.41, 0.81, 0.89, 0.68, 1.39, 1.53, 0.91, 1.49, 1.38, 1.73,
        1.11, 1.68, 0.66, 0.69, 1.98)
)
```
Solution

The question is answered by R. Start by loading data into R and estimate the parameters in R:

```r
fit <- lm(y ~ x1 + x2, data=D)
summary(fit)
```

Call:
`lm(formula = y ~ x1 + x2, data = D)`

Residuals:
```
       Min       1Q   Median       3Q      Max
-0.15500 -0.07800 -0.02000  0.05000  0.30100
```

Coefficients:
```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.433551  0.065983  6.5728 1.312e-06 ***
x1           1.65299  0.09525    17.359 2.462e-14 ***
x2           0.00394  0.07485     0.053   0.9597
```

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.113 on 22 degrees of freedom
Multiple R-squared: 0.9403, Adjusted R-squared: 0.9342
F-statistic: 172 on 2 and 22 DF, p-value: 3.729e-14

Solution

The parameter estimates are given in the first column of the coefficient matrix, i.e.

\[ \hat{\beta}_0 = 0.434, \]
\[ \hat{\beta}_1 = 1.653, \]
\[ \hat{\beta}_2 = 0.0039, \]

and the error variance estimate is \( \hat{\sigma}^2 = 0.11^2 \). The confidence intervals can either be calculated using the second column of the coefficient matrix, and the value of \( t_{0.975} \) (with degrees of freedom equal 22), or directly in R:
confint(fit)

2.5 % 97.5 %
(Intercept) 0.2967 0.5704
x1 1.4555 1.8505
x2 -0.1513 0.1592

b) Still using confidence level \( \alpha = 0.05 \) reduce the model if appropriate.

\[\text{Solution}\]

Since the confidence interval for \( \beta_2 \) cover zero (and the p-value is much larger than 0.05), the parameter should be removed from the model to get the simpler model

\[y_i = \beta_0 + \beta_1 x_1 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),\]

the parameter estimates in the simpler model are

```r
fit <- lm(y ~ x1, data=D)
summary(fit)
```

Call:
```
lm(formula = y ~ x1, data = D)
```

Residuals:
```
       Min        1Q  Median        3Q       Max
-0.15630 -0.07629 -0.02145  0.05159  0.29989
```

Coefficients:  
```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.436109  0.044034   9.910 9.07e-10 ***
x1          1.651206  0.086886  18.960 1.46e-15 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.11 on 23 degrees of freedom  
Multiple R-squared: 0.94, Adjusted R-squared: 0.937  
F-statistic: 360 on 1 and 23 DF, p-value: 1.54e-15

and both parameters are now significant.
c) Carry out a residual analysis to check that the model assumptions are fulfilled.

**Solution**

We are interested in inspecting a q-q plot of the residuals and a plot of the residuals as a function of the fitted values.

```r
par(mfrow=c(1,2))
qqnorm(fit$residuals, pch=19)
qqline(fit$residuals)
plot(fit$fitted.values, fit$residuals, pch=19,
xlab="Fitted.values", ylab="Residuals")
```

![Normal Q-Q Plot](image)

there are no strong evidence against the assumptions, the qq-plot is a straight line and there are no obvious dependence between the residuals and the fitted values, and we conclude that the assumptions are fulfilled.

d) Make a plot of the fitted line and 95% confidence and prediction intervals of the line for $x_1 \in [0, 1]$ (it is assumed that the model was reduced above).
Solution

```r
x1new <- seq(0,1,by=0.01)
pred <- predict(fit, newdata=data.frame(x1=x1new),
                interval="prediction")
conf <- predict(fit, newdata=data.frame(x1=x1new),
                interval="confidence")
plot(x1new, pred[, "fit"], type="l", ylim=c(0.1,2.4),
     xlab="x1", ylab="Prediction")
lines(x1new, conf[, "lwr"], col="green", lty=2)
lines(x1new, conf[, "upr"], col="green", lty=2)
lines(x1new, pred[, "lwr"], col="red", lty=2)
lines(x1new, pred[, "upr"], col="red", lty=2)
legend("topleft", c("Prediction","Confidence band","Prediction band"),
        lty=c(1,2,2), col=c(1,3,2), cex=0.7)
```

![Graph showing prediction and confidence bands](image-url)
6.3 MLR simulation exercise

Exercise 6.3 MLR simulation exercise

The following measurements have been obtained in a study:

<table>
<thead>
<tr>
<th>Nr.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>9.29</td>
<td>12.67</td>
<td>12.42</td>
<td>0.38</td>
<td>20.77</td>
<td>9.52</td>
<td>2.38</td>
<td>7.46</td>
</tr>
<tr>
<td>x1</td>
<td>1.00</td>
<td>2.00</td>
<td>3.00</td>
<td>4.00</td>
<td>5.00</td>
<td>6.00</td>
<td>7.00</td>
<td>8.00</td>
</tr>
<tr>
<td>x2</td>
<td>4.00</td>
<td>12.00</td>
<td>16.00</td>
<td>8.00</td>
<td>32.00</td>
<td>24.00</td>
<td>20.00</td>
<td>28.00</td>
</tr>
</tbody>
</table>

a) Plot the observed values of $y$ as a function of $x_1$ and $x_2$. Does it seem reasonable that either $x_1$ or $x_2$ can describe the variation in $y$?

You may copy the following lines into R to load the data:

```r
D <- data.frame(
  y=c(9.29, 12.67, 12.42, 0.38, 20.77, 9.52, 2.38, 7.46),
  x1=c(1.00, 2.00, 3.00, 4.00, 5.00, 6.00, 7.00, 8.00),
  x2=c(4.00, 12.00, 16.00, 8.00, 32.00, 24.00, 20.00, 28.00)
)
```

Solution

The data is plotted with

```r
par(mfrow=c(1,2))
plot(D$x1, D$y, xlab="x1", ylab="y")
plot(D$x2, D$y, xlab="x1", ylab="y")
```
There does not seem to be a strong relation between $y$ and $x_1$ or $x_2$.

b) Estimate the parameters for the two models

\[ Y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \]

and

\[ Y_i = \beta_0 + \beta_1 x_{2,i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \]

and report the 95% confidence intervals for the parameters. Are any of the parameters significantly different from zero on a 5% confidence level?
Solution

The models are fitted with

```r
fit1 <- lm(y ~ x1, data=D)
fit2 <- lm(y ~ x2, data=D)
confint(fit1)

2.5 % 97.5 %
(Intercept) -0.5426 24.898
x1 -3.1448 1.893

confint(fit2)

2.5 % 97.5 %
(Intercept) -7.5581 15.9659
x2 -0.2958 0.8688
```

since all confidence intervals cover zero we cannot reject that the parameters are in fact zero, and we would conclude neither $x_1$ nor $x_2$ explain the variations in $y$.

c) Estimate the parameters for the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (6-8)$$

and go through the steps of Method 6.16 (use confidence level 0.05 in all tests).
### Solution

The model is fitted with

```r
fit <- lm(y ~ x1 + x2, data=D)
summary(fit)
```

Call:
```
  lm(formula = y ~ x1 + x2, data = D)
```

Residuals:
```
   1     2     3     4     5     6     7     8
0.9622 0.1783 -0.3670 -1.0963 -0.3448 -0.2842 0.0178 0.9339
```

Coefficients:
```
                Estimate Std. Error   t value Pr(>|t|)     
(Intercept)     8.0325     0.6728   11.9063    0.0000727 ***
x1            -3.5734     0.1955   -18.2852    0.0000090 ***
x2             0.9672     0.0489    19.8361    0.0000061 ***
```

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.821 on 5 degrees of freedom
Multiple R-squared: 0.988, Adjusted R-squared: 0.983
F-statistic: 208 on 2 and 5 DF, p-value: 0.0000154

### Solution

Before discussing the parameter let’s have a look at the residuals:

```r
par(mfrow=c(1,2))
qqnorm(fit$residuals)
qqline(fit$residuals)
plot(fit$fitted.values, fit$residuals, 
     xlab="Fitted values", ylab="Residuals")
```
The are no obvious structures in the residuals as a function of the fitted values and also there does not seem be be serious departure from normality, but lets try to look at the residuals as a function of the independent variables anyway.

### Solution

```r
par(mfrow=c(1,2))
plot(D$x1, fit$residuals, xlab="x1", ylab="Residuals")
plot(D$x2, fit$residuals, xlab="x1", ylab="Residuals")
```

the plot of the residuals as a function of $x_1$ suggest that there could be a quadratic dependence.
Solution

Now include the quadratic dependence of $x_1$

```r
D$x3 <- D$x1^2
fit3 <- lm(y ~ x1 + x2 + x3, data=D)
summary(fit3)
```

```
Call:
  lm(formula = y ~ x1 + x2 + x3, data = D)

Residuals:
  1       2       3       4       5       6       7       8
0.0417  -0.0233  -0.0107  -0.0754  -0.0252  0.1104  0.0585  -0.0758

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.1007    0.1212  83.3 1.2e-07 ***
x1       -5.0024    0.0709 -70.5 2.4e-07 ***
x2        1.0006    0.0054 185.2 5.1e-09 ***
x3       0.1474    0.0070  21.1 3.0e-05 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0867 on 4 degrees of freedom
Multiple R-squared: 1.000, Adjusted R-squared: 1.000
F-statistic: 1.26e+04 on 3 and 4 DF, p-value: 2.11e-08
```

we can see that all parameters are still significant, and we can do the residual analysis of the resulting model.

```
par(mfrow=c(2,2))
qqnorm(fit3$residuals)
qqline(fit3$residuals)
plot(fitted.values(fit3), fit3$residuals, xlab="Fitted values", ylab="Residuals")
plot(D$x1, fit3$residuals, xlab="x1", ylab="Residuals")
plot(D$x2, fit3$residuals, xlab="x2", ylab="Residuals")
```
There are no obvious structures left and there is no departure from normality, and we can report the finally selected model as

\[ Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i}^2 + \varepsilon_i, \quad \varepsilon_i \sim (N(0, \sigma^2)), \]

with the parameters estimates given above.

d) Find the standard error for the line, and the confidence and prediction intervals for the line for the points \( (\min(x_1), \min(x_2)), (\bar{x}_1, \bar{x}_2), (\max(x_1), \max(x_2)) \).
Solution

The question is solved by

```r
## New data
Dnew <- data.frame(x1=c(min(D$x1), mean(D$x1), max(D$x1)),
                   x2=c(min(D$x2), mean(D$x2), max(D$x2)),
                   x3=c(min(D$x1), mean(D$x1), max(D$x1))^2)

## standard error for the line
predict(fit3, newdata=Dnew, se=TRUE)$se

  1 2 3
0.07306 0.04785 0.07985

## Confidence interval
predict(fit3, newdata=Dnew, interval="confidence")

fit lwr  upr
2  8.587  8.454  8.720
3 11.538 11.317 11.760

## Prediction interval
predict(fit3, newdata=Dnew, interval="prediction")

fit lwr  upr
1  9.248  8.934  9.563
2  8.587  8.312  8.862
3 11.538 11.211 11.866
```

e) Plot the observed values together with the fitted values (e.g. as a function of $x_1$).
Solution

The question is solved by

```r
plot(D$x1, D$y, pch=19, col=2, xlab="x1", ylab="y")
points(D$x1, fit3$fitted.values, pch="+", cex=2)
legend("topright", c("y1","fitted.values"), pch=c(19,3), col=c(2,1))
```

Notice that we have an almost perfect fit when including $x_1$, $x_2$ and $x_1^2$ in the model, while neither $x_1$ nor $x_2$ alone could predict the outcomes.