Exercise Sheet 1 - Simple and Multiple Linear Regressions - Answers

Fabien Baeriswyl, Jérôme Reboulleau, Tom Ruszkiewicz

Library setup

```
library(car) #for the vif
```

Exercise 1

Setup

You first want to load the dataset. To do so, use the `read.csv()` command. To give you an idea about the dataset itself, use `head()` to display the first 6 rows of the dataset. To directly access the columns of the dataset, use the `attach()` function (careful: run this only ONCE and detach it when you don't want to access the columns anymore, also ONCE only):

```
USComp <- read.csv("USComp.csv")
head(USComp)
```

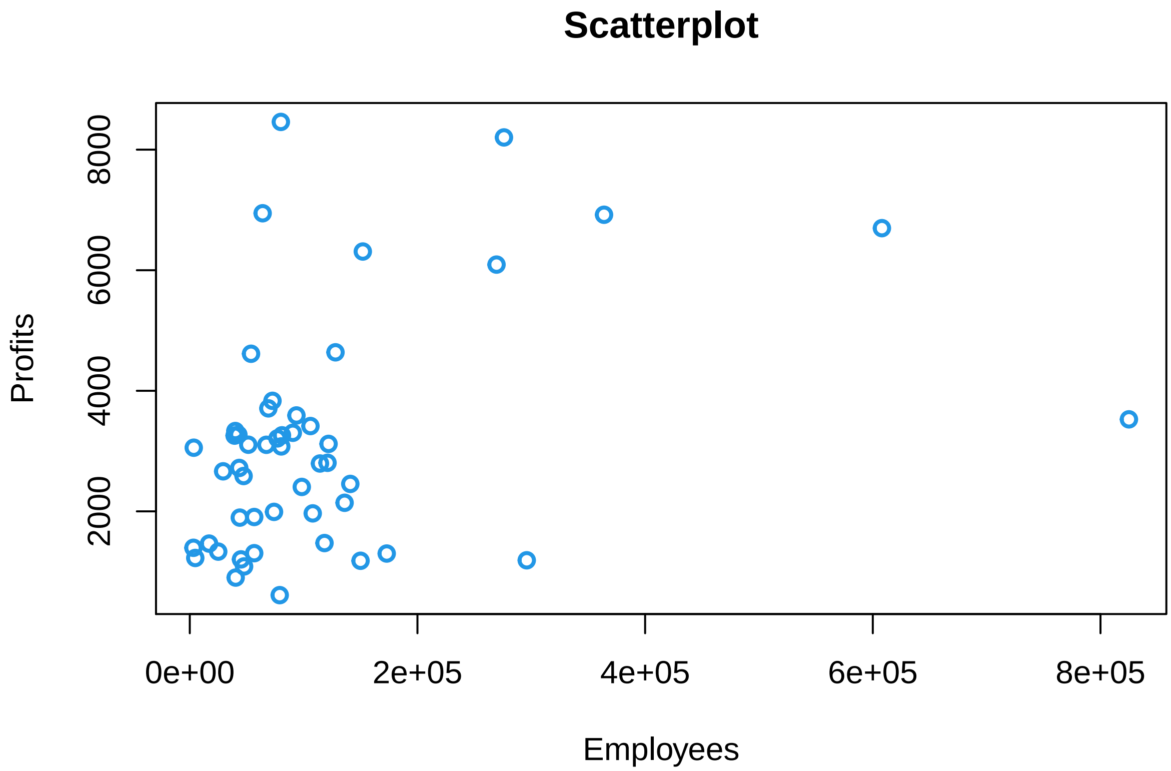
```
##                  Company Assets Earnings_per_share Employees Market_Value
## 1      General_Electric 304012               2.46    276000    260147.17
## 2            Ford_Motor 279097               5.62    363892     73923.00
## 3        General_Motors 228888               8.62    608000     54243.84
## 4                 Exxon  96064               3.37     80000    158783.63
## 5              Citicorp 310897               7.33     93700     66105.35
## 6 Chase_Manhattan_Corp. 365521               8.03     69033     58150.63
##   Profits Revenues Stockholders_Equity Total_Return_to_Investors_1997
## 1    8203    90840               34438                          51.09
## 2    6920   153627               30734                          57.01
## 3    6698   178174               17506                          19.47
## 4    8460   122379               43660                          28.33
## 5    3591    34697               21196                          24.85
## 6    3708    30381               21742                          25.44
```

```
attach(USComp)
```

1a)

We are first asked to plot the number of employees of the company against its profits.

```
plot(Employees, Profits, main="Scatterplot", lwd=2, col=4)
```



The correlation between the number of employees and the profits seems positive, although it is difficult to quantify its strength.

1b)

In R, one can find the sample Pearson correlation coefficient r by using the cor() function:

```
cor(Employees, Profits)
```

```
## [1] 0.3637726
```

One can then compute the test statistic t for the correlation test, by using the formula

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

where n is the number of observations, here $n = 50$. We find

$$t \approx 2.7057.$$

Knowing that the test statistic follows a Student's t distribution under the null with $n - 2 = 48$ degrees of freedom, one can find the p-value corresponding to t with the pt() function in R (and multiply by two, since we are looking at a two-sided test):

```
pt(2.7057, df=48, lower.tail=FALSE)*2
```

```
## [1] 0.00940766
```

The p-value being (much) smaller than 0.05, we reject the null hypothesis at this level and conclude that the correlation is statistically different from 0.

1c)

Since no company has 0 employee (or less), we can, in principle, expect this intercept to be positive. Note that the intercept is interpreted as the average profits when the company has 0 employee. But in this case, one could also say that a company is in this case expected to make 0 profit… Sometimes, the intercept does not make sense, per se.

We fit the linear model using the `lm()` function (for linear model) and report its summary:

```r
mod1 <- lm(Profits~Employees)
summary(mod1)
```

```
##
## Call:
## lm(formula = Profits ~ Employees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2957.1 -1176.0   -38.0   566.8  5522.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.557e+03  3.283e+02   7.788 4.61e-10 ***
## Employees   4.759e-03  1.759e-03   2.706  0.00941 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1794 on 48 degrees of freedom
## Multiple R-squared:  0.1323, Adjusted R-squared:  0.1143
## F-statistic: 7.321 on 1 and 48 DF,  p-value: 0.009409
```

Both the intercept and the slope are statistically different from 0, as suggested by the respective t-tests in the summary outputs. b_1 is interpreted as the increase in the average profit per unit increase in the number of employees. Since the profits are in millions, it means that, as we increase the number of employees by 1 unit (1 employee), the average profits increase by 0.004759M$ (M stands for million), that is 4759$ per additional employee.

The regression equation is given by:

$$p_i = 2557 + 4.759\text{e-03} \cdot e_i.$$

As mentioned, the slope is significantly different from 0 even at $\alpha = 0.01$, since the associated p-value is 0.0094 and $0.0094 < 0.01$ and we therefore reject the hypothesis that $H_0 : b_1 = 0$. Note that checking whether the slope is statistically different from 0 in this case is equivalent to test whether the coefficient of the independent variable is different from 0, since we have only one regressor. These tests use the same test statistic, which under the null follows a Student's t distribution with $n - 2$ degrees of freedom.

An additional point which has to be addressed when dealing with linear regression is the fact that we never explicitly say that increasing the number of employees by 1 unit (here 1 employee) causes the profits to increase by 4579$, there is no causation. We only say that an increase of 1 unit in the number of employees is associated with an increase of 4579$ in the profits, on average.

1d)

To obtain the SSR and the SSE, one can use the `anova()` function in `R`:

```r
anova(mod1)
```

```
## Analysis of Variance Table
```

```
##
## Response: Profits
##           Df    Sum Sq  Mean Sq F value   Pr(>F)
## Employees  1  23548019 23548019  7.3206 0.009409 **
## Residuals 48 154400512  3216677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have that

$$SSR = 23548019, \quad SSE = 154400512, \quad SST = 23548019 + 154400512 = 177948531$$

and hence that

$$R^2 = \frac{23548019}{177948531} \approx 0.1323$$

meaning that only approximately 13% of the variation of the response is explained by our simple linear model. Note however that, as covered in the lecture, this is not directly related to the prediction power of the model. One would have to look at the estimate s_ϵ to talk about the prediction power of the model. However, in this situation, since the model does not even explain the variation in the response, one can expect this model to perform poorly in terms of predictions.

Note that one can also compute these figures "manually", by using:

```r
SST <- var(Profits) * (50 - 1) # R uses the unbiased variance estimator.
SSE <- sum(mod1$resid^2)
SSR <- SST-SSE
```

1e)

The mean number of employees can be found by using the simple `mean()` function:

```r
mean(Employees)
```

```
## [1] 118514.6
```

Then, recall that the $1-\alpha\%$ confidence interval for the average response, given a specific dependent value x_p is given by

$$CI_{1-\alpha} = \bar{y} \pm t_{n-2,\alpha/2} \cdot s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

where

$$s_\epsilon = \sqrt{\frac{SSE}{n-2}}.$$

In this particular case, we have that

$$\bar{y} = b_0 + b_1 \cdot x_p = 2557 + 0.004759 \cdot 118514.6 \approx 3121.011$$

and, from the question,

$$x_p = \bar{x}$$

. We find that

$$t_{48,0.025} \approx 2.010635$$

and that

$$s_\epsilon = \sqrt{\frac{SSE}{48}} = 1793.51,$$

meaning that, eventually,

$$CI_{0.95} \approx [2611.032; 3630.99].$$

This confidence interval is quite large and therefore, we see that the prediction for the average value, at the average value of the number of employees, is of poor quality.

1f)

We saw earlier that it was not the case. What we could do is to try to find another independent variable correlated with the profits with a higher strength. One can use the correlation matrix for that purpose:

```r
cor(USComp[,-1]) #removing the first column which is character (or factor)
```

```
##                                     Assets Earnings_per_share   Employees
## Assets                          1.00000000         0.18192621 -0.07386325
## Earnings_per_share              0.18192621         1.00000000 -0.02956875
## Employees                      -0.07386325        -0.02956875  1.00000000
## Market_Value                   -0.05436295        -0.07917917  0.28273623
## Profits                         0.06694198         0.10290828  0.36377260
## Revenues                        0.09074903         0.08630084  0.75877759
## Stockholders_Equity             0.24516474        -0.01632086  0.17008323
## Total_Return_to_Investors_1997  0.31474373         0.11489213  0.15358346
##                                Market_Value    Profits   Revenues
## Assets                          -0.05436295 0.06694198 0.09074903
## Earnings_per_share              -0.07917917 0.10290828 0.08630084
## Employees                        0.28273623 0.36377260 0.75877759
## Market_Value                     1.00000000 0.75896990 0.36887398
## Profits                          0.75896990 1.00000000 0.69148321
## Revenues                         0.36887398 0.69148321 1.00000000
## Stockholders_Equity              0.49137904 0.71837127 0.53414150
## Total_Return_to_Investors_1997   0.35366694 0.16544678 0.08824224
##                                Stockholders_Equity
## Assets                                  0.24516474
## Earnings_per_share                     -0.01632086
## Employees                               0.17008323
## Market_Value                            0.49137904
## Profits                                 0.71837127
## Revenues                                0.53414150
## Stockholders_Equity                     1.00000000
## Total_Return_to_Investors_1997          0.06341139
##                                Total_Return_to_Investors_1997
## Assets                                             0.31474373
## Earnings_per_share                                 0.11489213
## Employees                                          0.15358346
## Market_Value                                       0.35366694
## Profits                                            0.16544678
## Revenues                                           0.08824224
## Stockholders_Equity                                0.06341139
## Total_Return_to_Investors_1997                     1.00000000
```

In this correlation matrix, one sees that the higher (linear) correlation is attained between profits and market value. We therefore want to use market value as an independent variable to try to predict the response, i.e. the profits.

```r
mod2 <- lm(Profits ~ Market_Value)
summary(mod2)
```

```
##
## Call:
## lm(formula = Profits ~ Market_Value)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2042.7  -814.7  -230.1   331.8  3822.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+03  2.947e+02   4.138 0.000141 ***
## Market_Value 3.054e-02  3.781e-03   8.076 1.69e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1254 on 48 degrees of freedom
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.5672
## F-statistic: 65.22 on 1 and 48 DF,  p-value: 1.693e-10
```

The coefficient of determination R^2 is now much higher, at 0.576, emphasising the fact that the variation in the response is now better explained by our model. The slope coefficient is also statistically different from 0, at any level α considered. This better fit was expected, since profits and market value have a higher linear correlation coefficient than profits and employees.

Exercise 2

2a)

From our correlation matrix in 1f), we expect the market value, the revenues, the stockholders equity and (possibly although it is unlikely) the number of employees to be statistically different from 0. For the other variables, namely assets, earnings per share and the total return to investors, we expect them to matter less in predicting the average value of the response.

2b)

We fit the complete model using the same function as before:

```r
mod3 <- lm(Profits ~ Employees + Assets + Market_Value + Revenues
           + Stockholders_Equity + Total_Return_to_Investors_1997)
summary(mod3)
```

```
##
## Call:
## lm(formula = Profits ~ Employees + Assets + Market_Value + Revenues +
##     Stockholders_Equity + Total_Return_to_Investors_1997)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1415.92  -375.68   -93.44   287.44  2549.04
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.727e+02  2.932e+02   0.930   0.3577
## Employees                      -3.636e-03  1.351e-03  -2.692   0.0101 *
## Assets                         -1.594e-04  1.157e-03  -0.138   0.8911
## Market_Value                   2.169e-02  3.165e-03   6.853 2.10e-08 ***
## Revenues                       3.359e-02  6.430e-03   5.224 4.84e-06 ***
## Stockholders_Equity            4.243e-02  2.091e-02   2.029   0.0486 *
## Total_Return_to_Investors_1997 -3.704e+00  5.831e+00  -0.635   0.5286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 782.4 on 43 degrees of freedom
## Multiple R-squared:  0.8521, Adjusted R-squared:  0.8315
## F-statistic: 41.29 on 6 and 43 DF,  p-value: 2.785e-16
```

We see that the estimated coefficients for the number of employees and the stockholders equity are slightly different from 0, at least at $\alpha = 0.05$; the estimated coefficients for the market value and the revenues are different from 0, at any α usually considered (we rarely go below 0.01); for the other variables, we cannot reject the hypothesis that they are equal to 0, at any usual α (usually, we do not go above 0.1). The coefficient of determination is 0.852, signalling a good linear fit. The adjusted R^2 is also quite high, at 0.8315.

(Side note: when dealing with multiple tests, note that it is usual to use p-values adjustement, such as the Bonferroni adjustement. This is, again, beyond the scope of this course, yet I thought it would be interesting to mention that when we are simultaneously testing multiple assumptions, we are piling up possibilities for errors…)

2c)

The estimated coefficient for the number of employees is negative, signalling that an additional unit of employees decreases the profits, which seems counter-intuitive. This signals potential multicollinearity problems. To compute the VIF, simply use the `vif()` function on your model from part b):

```r
vif(mod3)
```

```
##                      Employees                         Assets
##                       3.100497                       1.391465
##                   Market_Value                       Revenues
##                       1.798897                       4.002848
##            Stockholders_Equity Total_Return_to_Investors_1997
##                       2.249017                       1.454815
```

This indicates the degree of multicollinearity. As seen in the class, a VIF higher than 5 signals severe multicollinearity. In our case, note that some VIF values are close to that threshold, signalling potential multicollinearity. But nothing too severe.

2d)

Even though we did not find evidence for multicollinearity in the last point, we might still want to reduce the dimensionality of our linear model. Remember: the smaller the better! Since we started this exercise with a complete model, we will proceed to backward selection, using AIC as a criteria. The `step()` function does it automatically for us in `R`:

```r
nullmod <- lm(Profits ~ 1) # a model with just an intercept (constant model)
selmod <- step(mod3, scope=list(lower=nullmod, upper=mod3), direction="backward")
```

```
## Start:  AIC=672.69
## Profits ~ Employees + Assets + Market_Value + Revenues + Stockholders_Equity +
##     Total_Return_to_Investors_1997
##
##                                  Df Sum of Sq      RSS    AIC
## - Assets                          1     11610 26331741 670.71
## - Total_Return_to_Investors_1997  1    247021 26567152 671.16
## <none>                                         26320130 672.69
## - Stockholders_Equity             1   2520708 28840838 675.26
## - Employees                       1   4434329 30754460 678.48
## - Revenues                        1  16706357 43026487 695.27
## - Market_Value                    1  28749284 55069414 707.60
##
## Step:  AIC=670.71
## Profits ~ Employees + Market_Value + Revenues + Stockholders_Equity +
##     Total_Return_to_Investors_1997
##
##                                  Df Sum of Sq      RSS    AIC
## - Total_Return_to_Investors_1997  1    366917 26698658 669.41
## <none>                                         26331741 670.71
## - Stockholders_Equity             1   2623957 28955698 673.46
## - Employees                       1   4472668 30804409 676.56
## - Revenues                        1  16809946 43141687 693.40
## - Market_Value                    1  32939801 59271541 709.28
##
## Step:  AIC=669.41
## Profits ~ Employees + Market_Value + Revenues + Stockholders_Equity
##
##                       Df Sum of Sq      RSS    AIC
## <none>                              26698658 669.41
## - Stockholders_Equity  1   2778596 29477254 672.36
## - Employees            1   4755279 31453937 675.60
## - Revenues             1  17223756 43922414 692.30
## - Market_Value         1  34745830 61444488 709.08
```

The selected model is the following:

$$p_i = b_0 + b_1 e_i + b_2 mv_i + b_3 r_i + b4 stoeq_i + \epsilon_i.$$