**The Analytics Edge**

# Test your knowledge of Linear Regression in R – Solutions

*Note to all.* I have compiled the answers in the following format – for each question, the qualitative or "written" solutions will be provided together with their sub-questions. The R scripts (as well as the console outputs) will be provided *after* each whole question, followed by all the relevant plots. If I have missed anything in the solutions, or if you have any questions, you may email me at benjamin_tanwj@mymail.sutd.edu.sg. Thank you!

1. This question involves the use of simple linear regression on the **Auto** dataset. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset has the following fields:

   - **mpg**: miles per gallon
   - **cylinders** Number of cylinders
   - **displacement**: Engine displacement (cu. inches)
   - **horsepower**: Engine horsepower
   - **acceleration**: Time to accelerate from 0 to 60 mph (sec.)
   - **year**: Model year (modulo 100)
   - **origin**: Origin of car (1. American, 2. European, 3. Japanese)
   - **name**: Vehicle name

   (a) Perform a simple linear regression with **mpg** as the response and **horsepower** as the predictor. Comment on why you need to change the horsepower variable before performing the regression.

   *Solution.* Note that the horsepower variable is read in as a factor variable due to the presence of "?". You need to convert this to numeric as shown in the code, to make it a reasonable model.

   (b) Comment on the output by answering the following questions:

   - Is there a strong relationship between the predictor and the response?
   - Is the relationship between the predictor and the response positive or negative?

   *Solution.* Yes, there is a strong relationship between the predictor and response. p-value is almost zero, implying we can reject the null hypothesis that the corresponding $\beta = 0$.

Since $\beta = -0.1578$ for the relation between **mpg** and **horsepower**, the relation is negative.

(c) What is the predicted **mpg** associated with a **horsepower** of 98? What is the associated 99% confidence interval? Hint: You can check the predict.lm function on how the confidence interval can be computed for predictions with R.

*Solution.* The predicted **mpg** for **horsepower** of 98 is 24.46708 and the 99% confidence interval is [23.816,25.117].

(d) Compute the correlation between the response and the predictor variable. How does this compare with the $R^2$ value?

*Solution.* Code helps compute correlation while dropping observations where even one entry is missing. Squaring the correlation gives the $R^2$ of the model.

(e) Plot the response and the predictor. Also plot the least squares regression line.

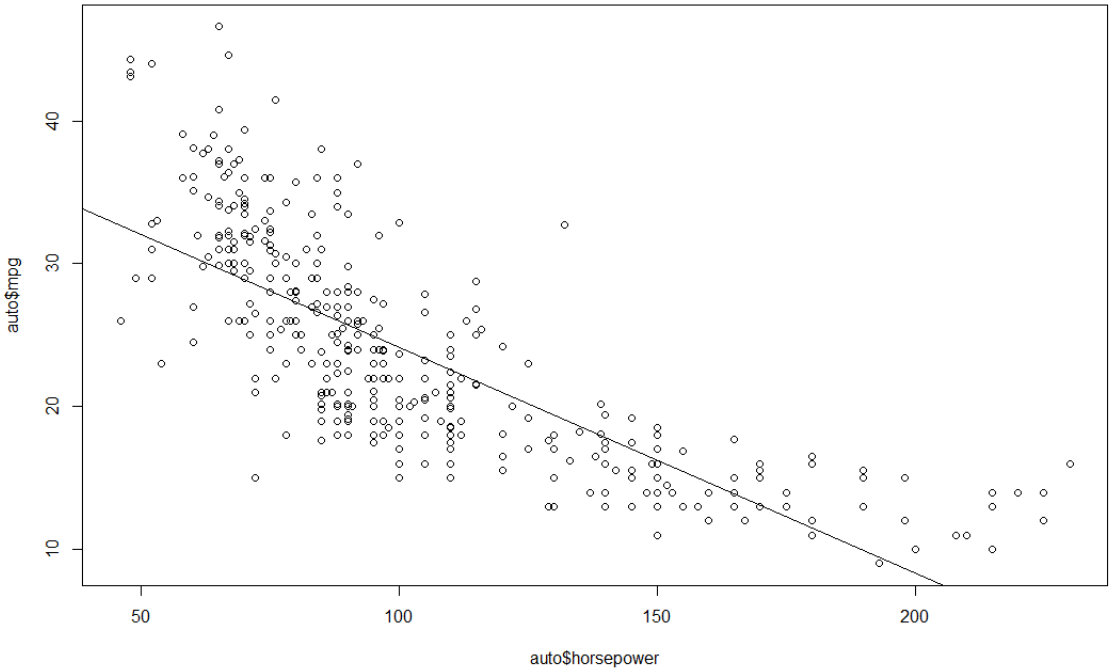
*Solution.* Plot can be found here.

(f) Use the following two commands in R to produce diagnostic plots of the linear regression fit:
> layout(matrix(1:4,2,2))
> plot(*your_model_name*)
Comment on the Residuals versus Fitted plot and the Normal Q-Q plot and on any problems you might see with the fit.

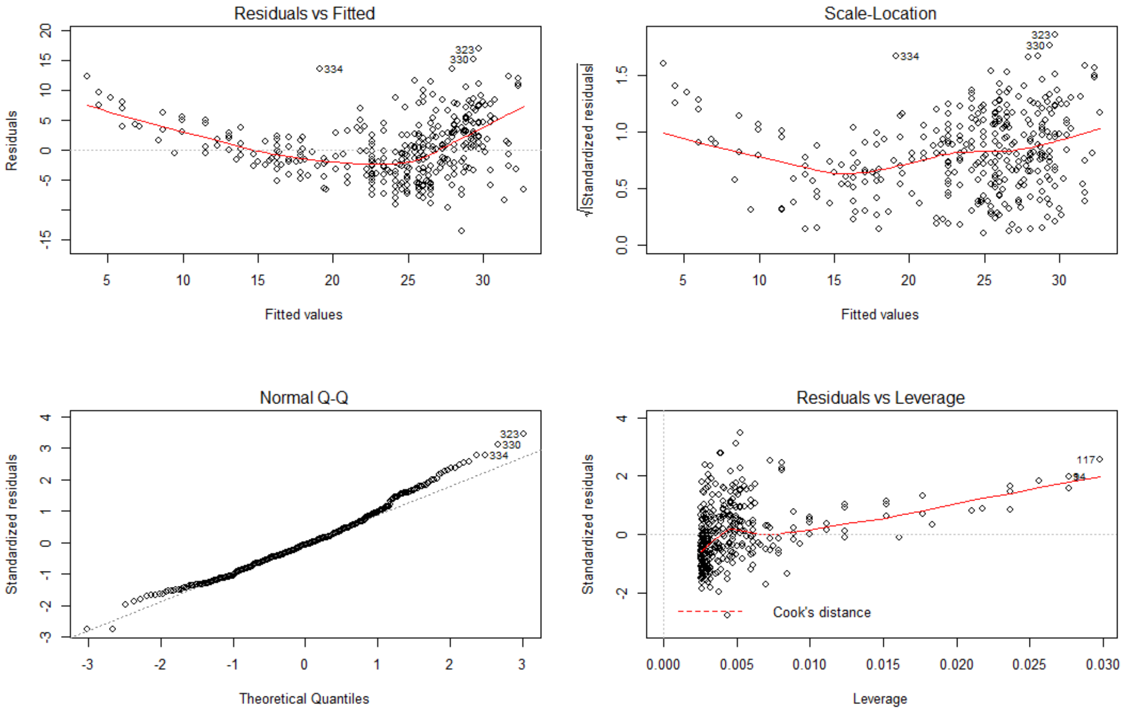
*Solution.* Plots can be found here. A good linear fit should see residuals randomly scattered. In this model we see that the residuals decrease, and then increase as the number of fitted residuals increase. The normal QQ plot also shows that the distribution of the residuals is not normal at the extreme values. This indicates that the data might have evidence of some nonlinearities and outliers.

*R Scripts.*

```
> #1a)
> auto <- read.csv("Auto.csv")
> str(auto)
'data.frame': 397 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : Factor w/ 94 levels "?","100","102",..: 17 35 29 29 24 42 47 46 48 40 ...
 $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2 ...
> auto$horsepower <- as.numeric(as.character(auto$horsepower))

> #1b)
> model1<- lm(mpg~horsepower, data=auto)
> summary(model1)

Call:
lm(formula = mpg ~ horsepower, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.906 on 390 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16

> #1c)
> predict(model1,newdata=data.frame(horsepower=98),interval=c("confidence"),level=.99)
       fit      lwr      upr
1 24.46708 23.81669 25.11747

> #1d)
> cor(auto$mpg,auto$horsepower, use = "pairwise.complete.obs")
[1] -0.7784268
> cor(auto$mpg,auto$horsepower, use = "pairwise.complete.obs")^2
[1] 0.6059483

> #1e)
> plot(auto$horsepower,auto$mpg)
> abline(model1)

> #1f)
> layout(matrix(1:4,2,2))
> plot(model1)
```



Plot for Q1e. Click here to go back to the question.



Plot for Q1f. Click here to go back to the question.

2. This question involves the use of multiple linear regression on the **Auto** dataset building on question 1.

   (a) Produce a scatterplot matrix which includes all the variables in the dataset.

   *Solution.* Plots found here. `pairs(auto)` creates a scatterplot matrix from the *auto* dataframe in question 1.

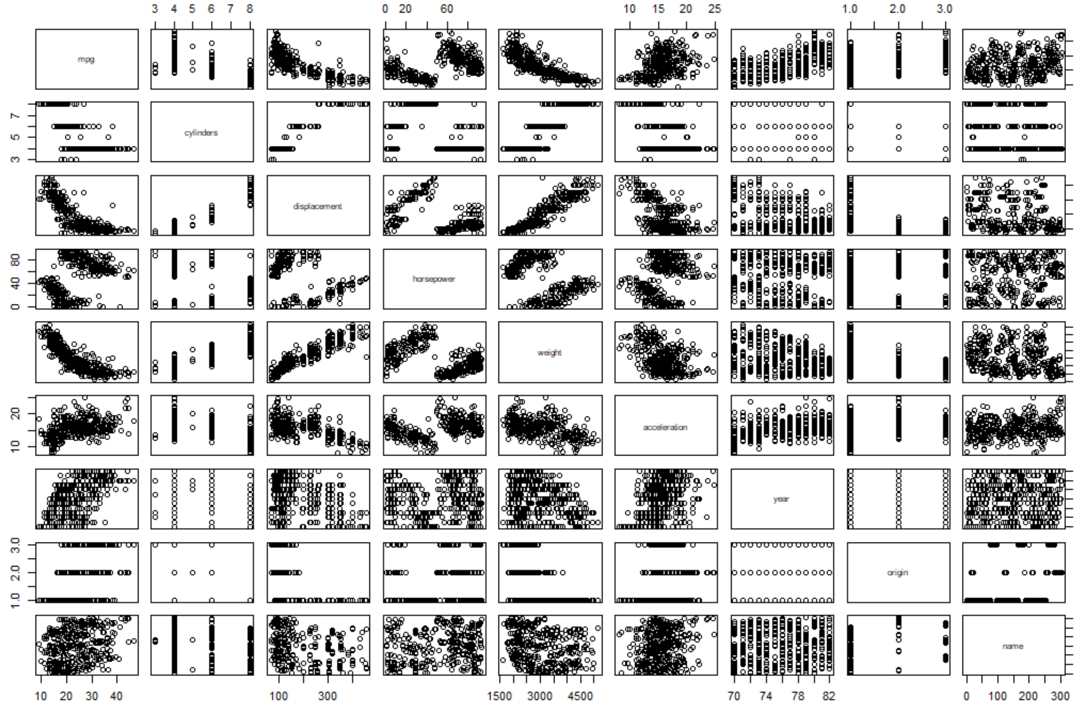
   (b) Compute a matrix of correlations between the variables using the function **cor()**. You need to exclude the **name** variables which is qualitative.

   (c) Perform a multiple linear regression with **mpg** as the response and all other variables except **name** as the predictors. Comment on the output by answering the following questions:

   - Is there a strong relationship between the predictors and the response?
   - Which predictors appear to have a statistically significant relationship to the response?
   - What does the coefficient for the **year** variable suggest?

*R Scripts.*

```
> #2a)
> pairs(auto)
> #2b)
> auto1 <- subset(auto, select= -c(name))
> cor(auto1)
                    mpg  cylinders displacement horsepower     weight acceleration
mpg           1.0000000 -0.7762599   -0.8044430         NA -0.8317389    0.4222974
cylinders    -0.7762599  1.0000000    0.9509199         NA  0.8970169   -0.5040606
displacement -0.8044430  0.9509199    1.0000000         NA  0.9331044   -0.5441618
horsepower           NA         NA           NA          1         NA           NA
weight       -0.8317389  0.8970169    0.9331044         NA  1.0000000   -0.4195023
acceleration  0.4222974 -0.5040606   -0.5441618         NA -0.4195023    1.0000000
year          0.5814695 -0.3467172   -0.3698041         NA -0.3079004    0.2829009
origin        0.5636979 -0.5649716   -0.6106643         NA -0.5812652    0.2100836
                   year     origin
mpg           0.5814695  0.5636979
cylinders    -0.3467172 -0.5649716
displacement -0.3698041 -0.6106643
horsepower           NA         NA
weight       -0.3079004 -0.5812652
acceleration  0.2829009  0.2100836
year          1.0000000  0.1843141
origin        0.1843141  1.0000000
> #2c)
> model2 <-lm(mpg˜., data=auto1)
> summary(model2)

Call:
lm(formula = mpg ˜ ., data = auto1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.328 on 384 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8215,Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```



Plot for Q2a. Click here to go back to the question.

3. This problem focusses on the multicollinearity problem with simulated data.

   (a) Perform the following commands in R:
   > set.seed(1)
   > x1 <− runif(100)
   > x2 <− 0.5*x1 + rnorm(100)/10
   > y <− 2 + 2*x1 + 0.3*x2 + rnorm(100)
   The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

   *Solution.* The form of the linear model is:

   $$y = 2 + 2x_1 + 0.3x_2 + \epsilon$$

   where the coefficients are $\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3$.

   (b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

   *Solution.* Plot found here. Correlation is 0.8351. This clearly shows a strong positive correlation between $x_1, x_2$ as expected from the simulation.

   (c) Using the data, fit a least square regression to predict y using x1 and x2.
   - What are the estimated parameters of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$? How do these relate to the true $\beta_0$, $\beta_1$ and $\beta_2$?
   - Can you reject the null hypothesis $H_0 : \beta_1 = 0$?
   - How about the null hypothesis $H_0 : \beta_2 = 0$?

   *Solution.* From the fit we see that $\hat{\beta}_0 = 2.13, \hat{\beta}_1 = 1.43, \hat{\beta}_2 = 1$. Since the true values are $\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3$, there is a discrepancy in the values $(\beta_1, \hat{\beta}_1)$ and $(\beta_2, \hat{\beta}_2)$, and less so in $(\beta_0, \hat{\beta}_0)$.
   We can reject the null hypothesis that $\beta_1 = 0$ at the 5% level, but we cannot reject the null hypothesis that $\beta_2 = 0$ at the 5% level.

   (d) Now fit a least squares regression to predict y using only x1.
   - How does the estimated $\hat{\beta}_1$ relate to the true $\beta_1$?
   - Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

   *Solution.* The estimated $\hat{\beta}_1 = 1.975$ is close to $\beta_1 = 2$ and we can reject the null hypothesis that $\beta_1 = 0$ as the $p$-value is close to zero.

(e) Now fit a least squares regression to predict y using only x2.

- How does the estimated $\hat{\beta}_2$ relate to the true $\beta_2$?
- Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

*Solution.* We can reject the null hypothesis that $\beta_2 = 0$ as the $p$-value is very close to zero. However, the predicted $\hat{\beta}_2 = 2.89$ is very far off from the actual value 0.3.

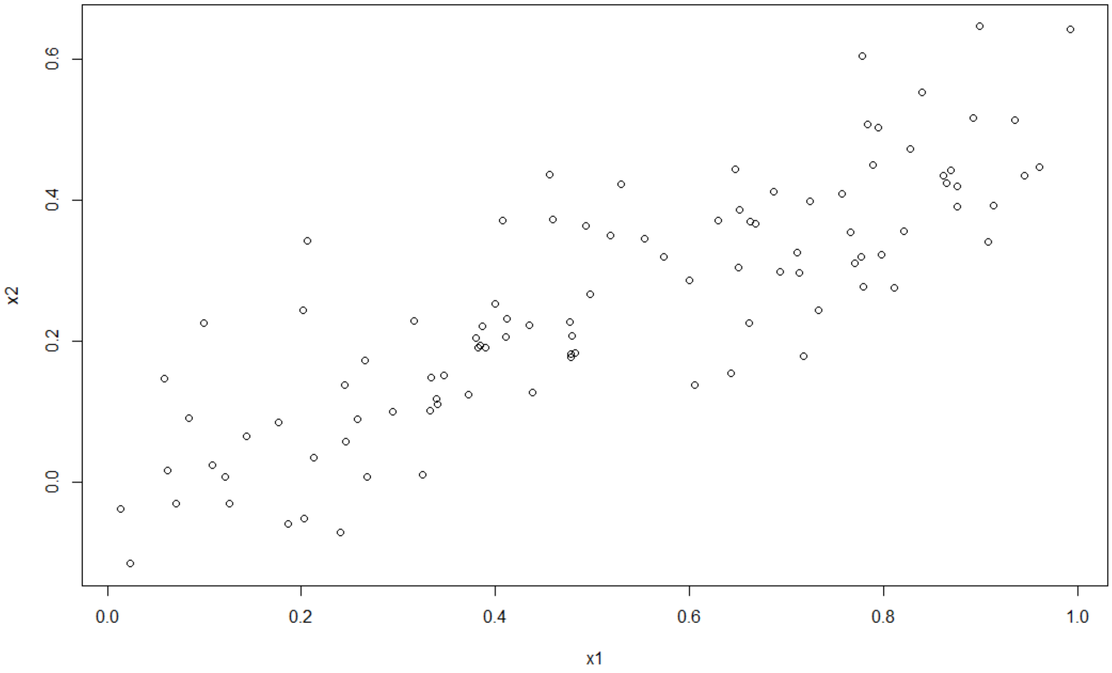
(f) Provide an explanation on the results in parts (c)-(e).

*Solution.* There is multicollinearity in the data between $x_1$ and $x_2$. In doing multiple regression we see this effect where it is difficult to reject $H_0 : \beta_i = 0$ (for one of the coefficients), while we see that with a single regression (with one variable), we can reject $H_0 : \beta_i = 0$. This is caused by multicollinearity.

*R Scripts.*

```
> #3a)
> set.seed(1)
> x1 <- runif(100)
> x2 <- .5*x1 + rnorm(100)/10
> y <- 2 + 2*x1 + .3*x2 + rnorm(100)

> #3b)
> cor(x1,x2)
[1] 0.8351212
> plot(x1, x2)

> #3c)
> model3 <- lm(y~x1+x2)
> summary(model3)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8311 -0.7273 -0.0537  0.6338  2.3359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1            1.4396     0.7212   1.996   0.0487 *
x2            1.0097     1.1337   0.891   0.3754
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.056 on 97 degrees of freedom
```

```
Multiple R-squared:  0.2088,Adjusted R-squared:  0.1925
F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05


> #3d)
> model4 <- lm(y˜x1)
> summary(model4)

Call:
lm(formula = y ˜ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89495 -0.66874 -0.07785  0.59221  2.45560

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
x1            1.9759     0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,Adjusted R-squared:  0.1942
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06


> #3e)
> model5 <- lm(y˜x2)
> summary(model5)

Call:
lm(formula = y ˜ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.62687 -0.75156 -0.03598  0.72383  2.44890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,Adjusted R-squared:  0.1679
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```



Plot for Q3b. Click here to go back to the question.

4. This problem involves the **Boston** dataset. This data was part of an important paper in 1978 by Harrison and Rubinfeld titled - "Hedonic housing prices and the demand for clean air" published in the Journal of Environmental Economics and Management 5(1): 81-102. The dataset has the following fields:

   - **crim**: per capita crime rate by town
   - **zn**: proportion of residential land zoned for lots over 25,000 sq.ft
   - **indus**: proportion of non-retail business acres per town
   - **chas**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
   - **nox**: nitrogen oxides concentration (parts per 10 million)
   - **rm**: average number of rooms per dwelling
   - **age**: proportion of owner-occupied units built prior to 1940
   - **dis**: weighted mean of distances to five Boston employment centres
   - **rad**: index of accessibility to radial highways
   - **tax**: full-value property-tax rate per $10,000
   - **ptratio**: pupil-teacher ratio by town
   - **black**: $1000(Bk - 0.63)^2$ where $Bk$ is the proportion of blacks by town
   - **lstat**: lower status of the population (percent)
   - **medv**: median value of owner-occupied homes in $1000s

   We will try to predict the median house value using thirteen predictors

   (a) For each predictor, fit a simple linear regression model using a single variable to predict the response. In which of these models is there a statistically significant relationship between the predictor and the response? Plot the figure of relationship between medv and lstat as an example to validate your finding.

   *Solution.* Plot can be found here. Use `summary(model1)` to `summary(model13)` to verify that the $p$-values for all the predictors is less than 0.001.

   (b) Fit a multiple linear regression models to predict your response using all the predictors. Compare the adjusted $R^2$ from this model with the simple regression model. For which predictors, can we reject the null hypothesis $H_0 : \beta_j = 0$?

   *Solution.* The adjusted $R^2$ is 0.7338 which is larger than the adjusted $R^2$ from the simple regression models. The variables for which we can reject the $H_0 : \beta_i = 0$ are *crim, zn, chas, nox, rm, dis, rad, tax, ptratio, black, lstat* at the 0.05 significance level.

(c) Create a plot displaying the univariate regression coefficients from (a) on the X-axis and the multiple regression coefficients from (b) on the Y-axs. That is each predictor is displayed as a single point in the plot. Comment on this plot.

*Solution.* Plot found here. The figure seems to indicate a fairly positive relationship between the results from the simple and multiple linear regression models. The relationship seems to be linear too.

(d) In this question, we will check if there is evidence of non-linear association between the **lstat** predictor variable and the response? To answer the question, fit a model of the form

$$\mathbf{medv} = \beta_0 + \beta_1\mathbf{lstat} + \beta_2\mathbf{lstat}^2 + \epsilon.$$

You can make use of the poly() function in R. Does this help improve the fit. Add higher degree polynomial fits. What is the degree of the polynomial fit beyond which the terms no longer remain significant?

*Solution.* Plot found here. Yes, adding higher-degree terms helps improve the fit. Beyond degree 5, adding additional terms does not keep the terms significant.

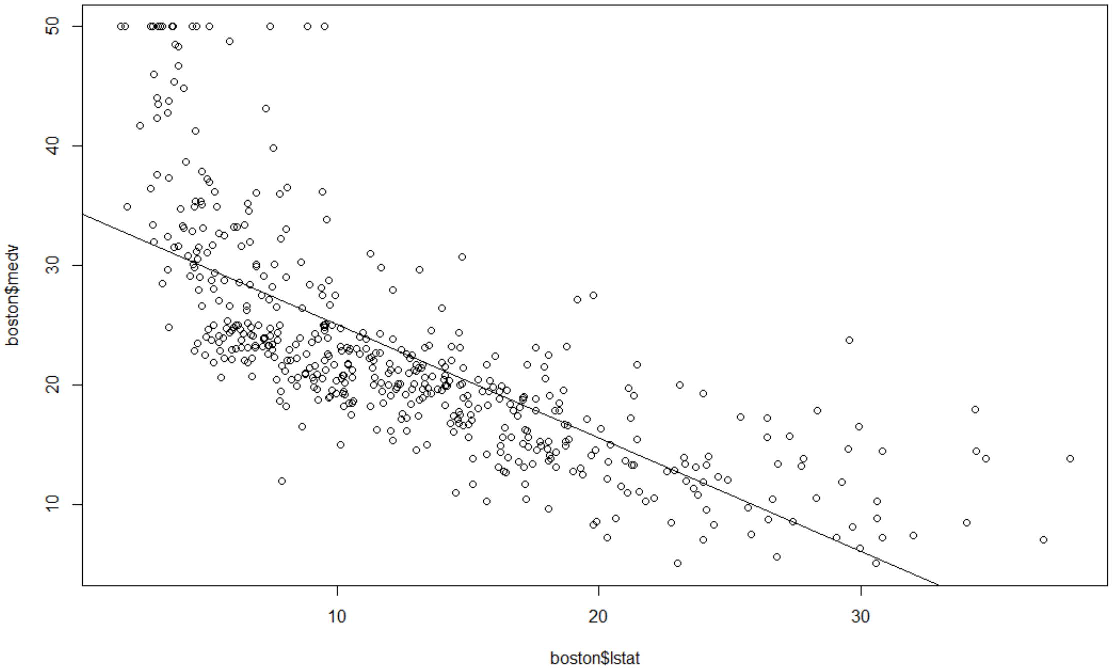
*R Scripts.*

```
> #4a)
> boston <- read.csv("Boston.csv")
> colnames(boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
 [9] "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
> model1 <- lm(medv~crim, data=boston)
> model2 <- lm(medv~zn, data=boston)
> model3 <- lm(medv~indus, data=boston)
> model4 <- lm(medv~chas, data=boston)
> model5 <- lm(medv~nox, data=boston)
> model6 <- lm(medv~rm, data=boston)
> model7 <- lm(medv~age, data=boston)
> model8 <- lm(medv~dis, data=boston)
> model9 <- lm(medv~rad, data=boston)
> model10 <- lm(medv~tax, data=boston)
> model11 <- lm(medv~ptratio, data=boston)
> model12 <- lm(medv~black, data=boston)
> model13 <- lm(medv~lstat, data=boston)
> summary(model13)

Call:
lm(formula = medv ~ lstat, data = boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16

> plot(boston$lstat, boston$medv)
> abline(model13)

> #4b)
> modelall<- lm(medv˜., data=boston)
> summary(modelall)

Call:
lm(formula = medv ˜ ., data = boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

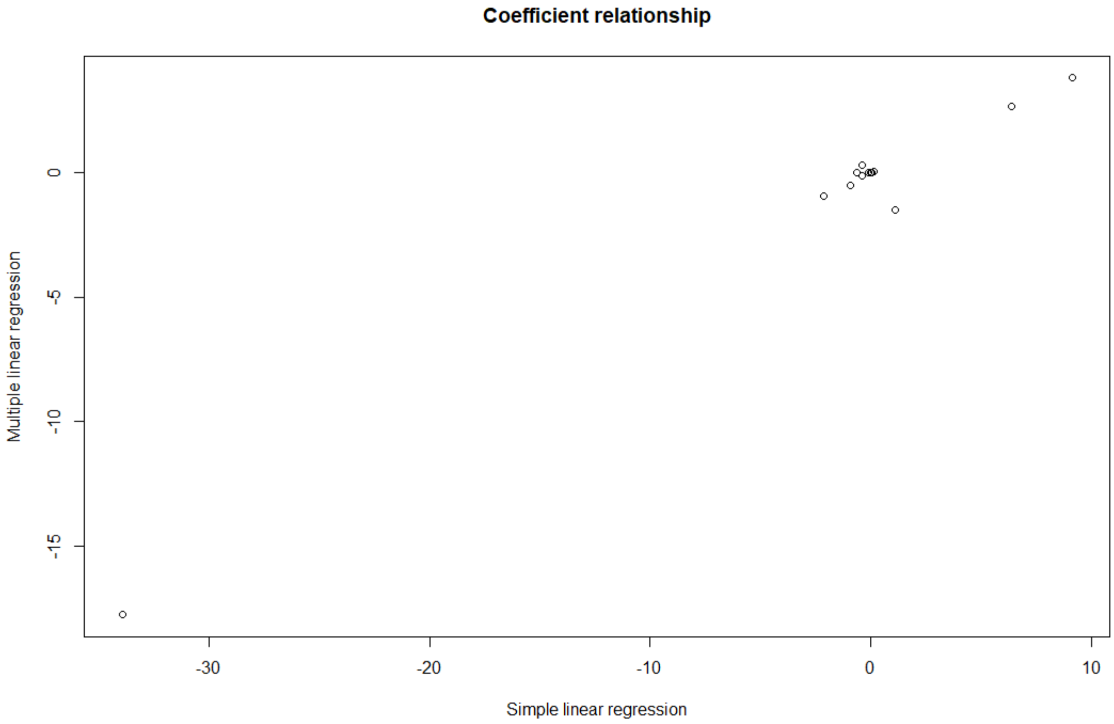
```
> #4c)
> x <- c(model1$coef[2], model2$coef[2], model3$coef[2], model4$coef[2], model5$coef[2],
model6$coef[2], model7$coef[2], model8$coef[2], model9$coef[2], model10$coef[2],
model11$coef[2], model12$coef[2], model13$coef[2])
> y <- modelall$coef[2:14]
> plot(x, y, main = "Coefficient relationship", xlab = "Simple linear regression",
ylab = "Multiple linear regression")

> #4d)
> modelpoly2 <- lm(medv~poly(lstat,2,raw=TRUE), data = boston)
> summary(modelpoly2)

Call:
lm(formula = medv ~ poly(lstat, 2, raw = TRUE), data = boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2834  -3.8313  -0.5295   2.3095  25.4148 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 42.862007   0.872084   49.15   <2e-16 ***
poly(lstat, 2, raw = TRUE)1 -2.332821   0.123803  -18.84   <2e-16 ***
poly(lstat, 2, raw = TRUE)2  0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,Adjusted R-squared:  0.6393 
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16 

> modelpoly3 <- lm(medv~poly(lstat,3,raw=TRUE), data = boston)
> summary(modelpoly3)

Call:
lm(formula = medv ~ poly(lstat, 3, raw = TRUE), data = boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.5441  -3.7122  -0.5145   2.4846  26.4153 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 48.6496253  1.4347240  33.909  < 2e-16 ***
poly(lstat, 3, raw = TRUE)1 -3.8655928  0.3287861 -11.757  < 2e-16 ***
poly(lstat, 3, raw = TRUE)2  0.1487385  0.0212987   6.983 9.18e-12 ***
poly(lstat, 3, raw = TRUE)3 -0.0020039  0.0003997  -5.013 7.43e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 
```

```
Residual standard error: 5.396 on 502 degrees of freedom
Multiple R-squared:  0.6578,Adjusted R-squared:  0.6558
F-statistic: 321.7 on 3 and 502 DF,  p-value: < 2.2e-16

> modelpoly4 <- lm(medv~poly(lstat,4,raw=TRUE), data = boston)
> summary(modelpoly4)

Call:
lm(formula = medv ~ poly(lstat, 4, raw = TRUE), data = boston)

Residuals:
    Min      1Q  Median      3Q     Max
-13.563  -3.180  -0.632   2.283  27.181

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.731e+01  2.280e+00  25.134  < 2e-16 ***
poly(lstat, 4, raw = TRUE)1 -7.028e+00  7.308e-01  -9.618  < 2e-16 ***
poly(lstat, 4, raw = TRUE)2  4.955e-01  7.489e-02   6.616 9.50e-11 ***
poly(lstat, 4, raw = TRUE)3 -1.631e-02  2.994e-03  -5.448 7.98e-08 ***
poly(lstat, 4, raw = TRUE)4  1.949e-04  4.043e-05   4.820 1.90e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.28 on 501 degrees of freedom
Multiple R-squared:  0.673,Adjusted R-squared:  0.6704
F-statistic: 257.8 on 4 and 501 DF,  p-value: < 2.2e-16

> modelpoly5 <- lm(medv~poly(lstat,5,raw=TRUE), data = boston)
> summary(modelpoly5)

Call:
lm(formula = medv ~ poly(lstat, 5, raw = TRUE), data = boston)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5433  -3.1039  -0.7052   2.0844  27.1153

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.770e+01  3.604e+00  18.783  < 2e-16 ***
poly(lstat, 5, raw = TRUE)1 -1.199e+01  1.526e+00  -7.859 2.39e-14 ***
poly(lstat, 5, raw = TRUE)2  1.273e+00  2.232e-01   5.703 2.01e-08 ***
poly(lstat, 5, raw = TRUE)3 -6.827e-02  1.438e-02  -4.747 2.70e-06 ***
poly(lstat, 5, raw = TRUE)4  1.726e-03  4.167e-04   4.143 4.03e-05 ***
poly(lstat, 5, raw = TRUE)5 -1.632e-05  4.420e-06  -3.692 0.000247 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.215 on 500 degrees of freedom
Multiple R-squared:  0.6817,Adjusted R-squared:  0.6785
```

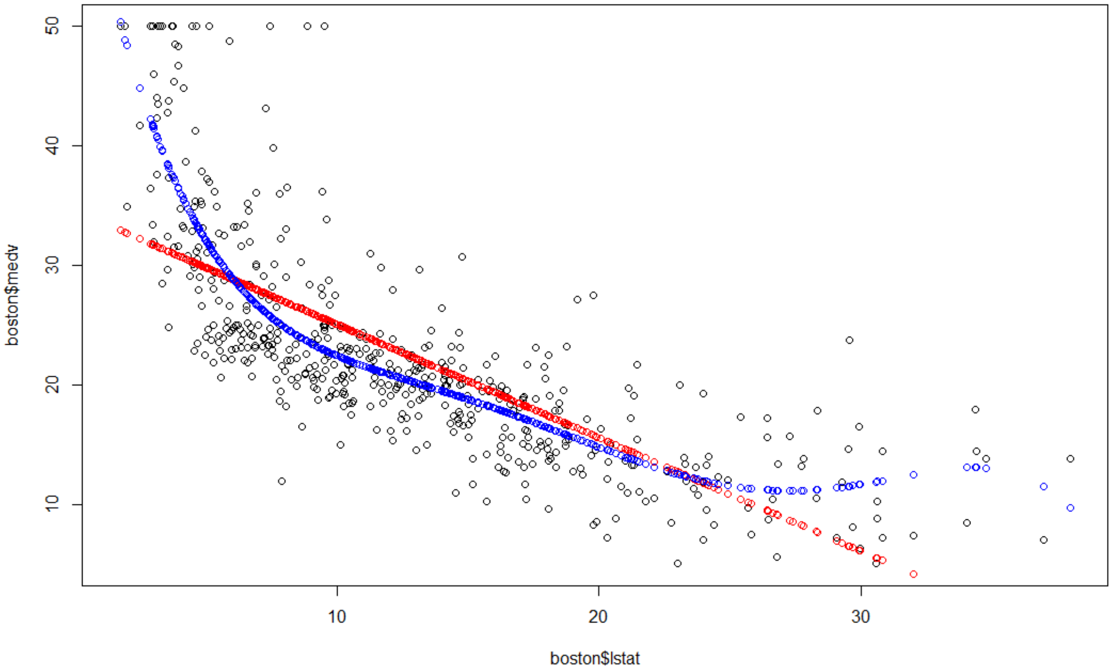
```
F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16

> modelpoly6 <- lm(medv~poly(lstat,6,raw=TRUE), data = boston)
> summary(modelpoly6)

Call:
lm(formula = medv ~ poly(lstat, 6, raw = TRUE), data = boston)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7317  -3.1571  -0.6941   2.0756  26.8994

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  7.304e+01  5.593e+00  13.059  < 2e-16 ***
poly(lstat, 6, raw = TRUE)1 -1.517e+01  2.965e+00  -5.115 4.49e-07 ***
poly(lstat, 6, raw = TRUE)2  1.930e+00  5.713e-01   3.378 0.000788 ***
poly(lstat, 6, raw = TRUE)3 -1.307e-01  5.202e-02  -2.513 0.012295 *
poly(lstat, 6, raw = TRUE)4  4.686e-03  2.407e-03   1.947 0.052066 .
poly(lstat, 6, raw = TRUE)5 -8.416e-05  5.450e-05  -1.544 0.123186
poly(lstat, 6, raw = TRUE)6  5.974e-07  4.783e-07   1.249 0.212313
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.212 on 499 degrees of freedom
Multiple R-squared:  0.6827,Adjusted R-squared:  0.6789
F-statistic: 178.9 on 6 and 499 DF,  p-value: < 2.2e-16

> pr1 <- predict(model13,newdata=boston)
> pr5 <- predict(modelpoly5,newdata=boston)
> plot(boston$lstat,boston$medv)
> points(boston$lstat,pr1,col="red")
> points(boston$lstat,pr5,col="blue")
```



Plot for Q4a. Click here to go back to the question.



Plot for Q4c. Click here to go back to the question.



Plot for Q4d. Click here to go back to the question.

5. There have been many studies documenting that the average global temperature has been increasing over the last century. The consequences of a continued rise in global temperature will be dire. Rising sea levels and an increased frequency of extreme weather events will affect billions of people. In this problem, we will attempt to study the relationship between average global temperature and several other factors. The file **climate_change.csv** contains climate data from May 1983 to December 2008. The available variables include:

   - **Year**: Observation year
   - **Month**: Observation month
   - **Temp**: Difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.
   - **CO2, N2O, CH4, CFC.11, CFC.12**: Atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division.
     CO2, N2O and CH4 are expressed in ppmv (parts per million by volume – i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere) CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).
   - **Aerosols**: Mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.
   - **TSI**: Total solar irradiance (TSI) in W/m2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.
   - **MEI**: Multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

   (a) We are interested in how changes in these variables affect future temperatures, as well as how well these variables explain temperature changes so far. Read the dataset **climate_change.csv** into R. Then, split the data into a training set, consisting of all the observations up to and including 2006, and a testing set consisting of the remaining years. A training set refers to the data that will be used to build the model, and a testing set refers to the data we will use to test our predictive ability. Build a linear regression model to predict the dependent variable Temp, using MEI, CO2, CH4, N2O, CFC.11, CFC.12, TSI and Aerosols as independent variables (Year and Month should not be used in the model). Use the training set to build the model. What is the model $R^2$?

*Solution.* $R^2$ for the model is 0.7509.

(b) Which variables are significant in the model? We will consider a variable significant in this example only if the p-value is below 0.05.

*Solution.* The variables significant in this model with $p$-values below 0.05 are *MEI, CO2, CFC.11, CFC.12, TSI, Aerosols.*

(c) Current scientific opinion is that nitrous oxide and CFC-11 are greenhouse gases: gases that are able to trap heat from the sun and contribute to the heating of the Earth. However, the regression coefficients of both the N2O and CFC-11 variables are negative, indicating that increasing atmospheric concentrations of either of these two compounds is associated with lower global temperatures. Compute the correlations in the training set. Which of the following is the simplest correct explanation for this contradiction?

- Climate scientists are wrong that N2O and CFC-11 are greenhouse gases - this regression analysis constitutes part of a disproof.
- There is not enough data, so the regression coefficients being estimated are not accurate.
- All of the gas concentration variables reflect human development - N2O and CFC.11 are correlated with other variables in the data set.

*Solution.* *N2O* is highly correlated with *CO2, CH4, CFC.12, Temp* and quite correlated with *CFC.11*. *CFC.11* is fairly positively correlated with *CO2, N2O* and strongly correlated with *CH4, CFC.12.* The results seem to indicate that all the gas concentration variables reflect human development and the variables are correlated in the dataset.

Note that the correlation between *Temp* and *N2O* is 0.778 and that between *Temp* and *CFC.11* is 0.40 (fairly high).

(d) Given that the correlations are so high, let us focus on the N2O variable and build a model with only MEI, TSI, Aerosols and N2O as independent variables. Remember to use the training set to build the model. What is the coefficient of N2O in this reduced model? How does this compare to the coefficient in the previous model with all of the variables? What is the model $R^2$?

*Solution.* The coefficient for *N2O* in this reduced model is 0.0253. The variable is also very significant as compared to the model with all variables in it. By comparing the $R^2$ and adjusted $R^2$, we also see the model does not lose a lot of explanatory power while variables are reduced. This is typical of models where the independent variables are highly correlated with each other.

(e) We have many variables in this problem, and as we have seen above, dropping some from the model does not decrease model quality. R provides a function **step()**, that will automate the procedure of trying different combinations of variables to find a good compromise of model simplicity and $R^2$. This trade-off is formalized by the Akaike information criterion (AIC) - it can be informally thought of as the quality of the model with a penalty for the number of variables in the model. The step function has one argument - the name of the initial model. It returns a simplified model. Use the step function in R to derive a new model, with the full model as the initial model. What is the $R^2$ value of the model produced by the step function? Which of the variable(s) are eliminated from the full model by the step function? It is interesting to note that the step function does not address the collinearity of the variables, except that adding highly correlated variables will not improve the $R^2$ significantly. The consequence of this is that the step function will not necessarily produce a very interpretable model - just a model that has balanced quality and simplicity for a particular weighting of quality and simplicity (AIC).

*Solution.* $R^2$ is 0.7508 which is slightly lower than the full model, but adjusted $R^2$ is 0.7445 and higher. *CH4* is eliminated in this model to get a better fit but with fewer predictors.

(f) We have developed an understanding of how well we can fit a linear regression to the training data, but does the model quality hold when applied to unseen data? Using the model produced from the step function, calculate temperature predictions for the testing data set, using the predict function. What is the test $R^2$?

*Solution.* The test $R^2 valueis 0.6286051$. Note that with the full model, you get test $R^2 = 0.6274$.

*R Scripts.*

```
> #5a)
> climate <- read.csv("climate_change.csv")
> training <- subset(climate, climate$Year <= 2006)
> test <- subset(climate, Year > 2006)
> model1 <- lm(Temp~MEI+CO2+CH4+N2O+CFC.11+CFC.12+TSI+Aerosols, data = training)
> summary(model1)

Call:
lm(formula = Temp ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 +
    TSI + Aerosols, data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25888 -0.05913 -0.00082  0.05649  0.32433

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.246e+02  1.989e+01  -6.265 1.43e-09 ***
MEI          6.421e-02  6.470e-03   9.923  < 2e-16 ***
CO2          6.457e-03  2.285e-03   2.826  0.00505 **
CH4          1.240e-04  5.158e-04   0.240  0.81015
N2O         -1.653e-02  8.565e-03  -1.930  0.05467 .
CFC.11      -6.631e-03  1.626e-03  -4.078 5.96e-05 ***
CFC.12       3.808e-03  1.014e-03   3.757  0.00021 ***
TSI          9.314e-02  1.475e-02   6.313 1.10e-09 ***
Aerosols    -1.538e+00  2.133e-01  -7.210 5.41e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09171 on 275 degrees of freedom
Multiple R-squared:  0.7509,Adjusted R-squared:  0.7436
F-statistic: 103.6 on 8 and 275 DF,  p-value: < 2.2e-16



> #5c)
> cor(training)
                Year         Month           MEI         CO2         CH4         N2O
Year      1.00000000 -0.0279419602 -0.0369876842  0.98274939  0.91565945  0.99384523
Month    -0.02794196  1.0000000000  0.0008846905 -0.10673246  0.01856866  0.01363153
MEI      -0.03698768  0.0008846905  1.0000000000 -0.04114717 -0.03341930 -0.05081978
CO2       0.98274939 -0.1067324607 -0.0411471651  1.00000000  0.87727963  0.97671982
CH4       0.91565945  0.0185686624 -0.0334193014  0.87727963  1.00000000  0.89983864
N2O       0.99384523  0.0136315303 -0.0508197755  0.97671982  0.89983864  1.00000000
CFC.11    0.56910643 -0.0131112236  0.0690004387  0.51405975  0.77990402  0.52247732
CFC.12    0.89701166  0.0006751102  0.0082855443  0.85268963  0.96361625  0.86793078
TSI       0.17030201 -0.0346061935 -0.1544919227  0.17742893  0.24552844  0.19975668
Aerosols -0.34524670  0.0148895406  0.3402377871 -0.35615480 -0.26780919 -0.33705457
Temp      0.78679714 -0.0998567411  0.1724707512  0.78852921  0.70325502  0.77863893
              CFC.11        CFC.12         TSI    Aerosols        Temp
Year      0.56910643  0.8970116635  0.17030201 -0.34524670  0.78679714
Month    -0.01311122  0.0006751102 -0.03460619  0.01488954 -0.09985674
MEI       0.06900044  0.0082855443 -0.15449192  0.34023779  0.17247075
CO2       0.51405975  0.8526896272  0.17742893 -0.35615480  0.78852921
CH4       0.77990402  0.9636162478  0.24552844 -0.26780919  0.70325502
N2O       0.52247732  0.8679307757  0.19975668 -0.33705457  0.77863893
CFC.11    1.00000000  0.8689851828  0.27204596 -0.04392120  0.40771029
CFC.12    0.86898518  1.0000000000  0.25530281 -0.22513124  0.68755755
TSI       0.27204596  0.2553028138  1.00000000  0.05211651  0.24338269
Aerosols -0.04392120 -0.2251312440  0.05211651  1.00000000 -0.38491375
```

```
Temp      0.40771029  0.6875575483  0.24338269 -0.38491375  1.00000000



> #5d)
> model2 <- lm(Temp˜MEI+TSI+Aerosols+N2O, data = training)
> summary(model2)

Call:
lm(formula = Temp ˜ MEI + TSI + Aerosols + N2O, data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27916 -0.05975 -0.00595  0.05672  0.34195 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.162e+02  2.022e+01  -5.747 2.37e-08 ***
MEI          6.419e-02  6.652e-03   9.649  < 2e-16 ***
TSI          7.949e-02  1.487e-02   5.344 1.89e-07 ***
Aerosols    -1.702e+00  2.180e-01  -7.806 1.19e-13 ***
N2O          2.532e-02  1.311e-03  19.307  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09547 on 279 degrees of freedom
Multiple R-squared:  0.7261,Adjusted R-squared:  0.7222 
F-statistic: 184.9 on 4 and 279 DF,  p-value: < 2.2e-16 



> #5e)
> model3 <- step(model1)
Start:  AIC=-1348.16
Temp ˜ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + TSI + Aerosols

           Df Sum of Sq    RSS     AIC
- CH4       1   0.00049 2.3135 -1350.1
<none>                  2.3130 -1348.2
- N2O       1   0.03132 2.3443 -1346.3
- CO2       1   0.06719 2.3802 -1342.0
- CFC.12    1   0.11874 2.4318 -1335.9
- CFC.11    1   0.13986 2.4529 -1333.5
- TSI       1   0.33516 2.6482 -1311.7
- Aerosols  1   0.43727 2.7503 -1301.0
- MEI       1   0.82823 3.1412 -1263.2

Step:  AIC=-1350.1
Temp ˜ MEI + CO2 + N2O + CFC.11 + CFC.12 + TSI + Aerosols

           Df Sum of Sq    RSS     AIC
```

```
<none>                  2.3135 -1350.1
- N2O       1   0.03133 2.3448 -1348.3
- CO2       1   0.06672 2.3802 -1344.0
- CFC.12    1   0.13023 2.4437 -1336.5
- CFC.11    1   0.13938 2.4529 -1335.5
- TSI       1   0.33500 2.6485 -1313.7
- Aerosols  1   0.43987 2.7534 -1302.7
- MEI       1   0.83118 3.1447 -1264.9
> summary(model3)

Call:
lm(formula = Temp ~ MEI + CO2 + N2O + CFC.11 + CFC.12 + TSI + 
    Aerosols, data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25770 -0.05994 -0.00104  0.05588  0.32203 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.245e+02  1.985e+01  -6.273 1.37e-09 ***
MEI          6.407e-02  6.434e-03   9.958  < 2e-16 ***
CO2          6.402e-03  2.269e-03   2.821 0.005129 ** 
N2O         -1.602e-02  8.287e-03  -1.933 0.054234 .  
CFC.11      -6.609e-03  1.621e-03  -4.078 5.95e-05 ***
CFC.12       3.868e-03  9.812e-04   3.942 0.000103 ***
TSI          9.312e-02  1.473e-02   6.322 1.04e-09 ***
Aerosols    -1.540e+00  2.126e-01  -7.244 4.36e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09155 on 276 degrees of freedom
Multiple R-squared:  0.7508,Adjusted R-squared:  0.7445 
F-statistic: 118.8 on 7 and 276 DF,  p-value: < 2.2e-16


> #5f)
> pred <- predict(model3, newdata = test)
> sse <- sum((test$Temp-pred)^2)
> sst <- sum((test$Temp-mean(training$Temp))^2)
> testR2 <- 1-sse/sst
> testR2
[1] 0.6286051
```

6. Orley Ashenfelter in his paper "Predicting the Quality and Price of Bordeaux Wines" published in The Economic Journal showed that the variability in the prices of Bordeaux wines is predicted well by the weather that created the grapes. In this question, you will validate how these results translate to a dataset for wines produced in Australia. The data is provided in the file **winedata.csv**. The dataset contains the following variables:

   - **vintage**: Year the wine was made
   - **price91**: 1991 auction prices for the wine in dollars
   - **price92**: 1992 auction prices for the wine in dollars
   - **temp**: Average temperature during the growing season in degree Celsius
   - **hrain**: Total harvest rain in mm
   - **wrain**: Total winter rain in mm
   - **tempdiff**: Sum of the difference between the maximum and minimum temperatures during the growing season in degree Celsius

   (a) Define two new variables **age91** and **age92** that captures the age of the wine (in years) at the time of the auctions. For example, a 1961 wine would have an age of 30 at the auction in 1991. What is the average price of wines that were 15 years or older at the time of the 1991 auction?

   *Solution.* The average price of wine that were 15 years or older at the 1991 auction is \$96.435.

   (b) What is the average price of the wines in the 1991 auction that were produced in years when both the harvest rain was below average and the temperature difference was below average?

   *Solution.* The average price in 1991 when harvest rain and temperature difference were below average is \$72.867.

   (c) In this question, you will develop a simple linear regression model to fit the **log** of the price at which the wine was auctioned in 1991 with the age of the wine. To fit the model, use a training set with data for the wines up to (and including) the year 1981. What is the R-squared for this model?

   *Solution.* $R^2$ for this model is 0.6675.

   (d) Find the 99% confidence interval for the estimated coefficients from the regression.

   *Solution.* For intercept ($\beta_0$): $[3.159, 3.98]$.

For *age91* ($\beta_1$): $[0.022, 0.062]$.

(e) Use the model to predict the **log** of prices for wines made from 1982 onwards and auctioned in 1991. What is the test R-squared?

*Solution.* Test $R^2 = 0.9213742$.

(f) Which among the following options describes best the quality of fit of the model for this dataset in comparison with the Bordeaux wine dataset that was analyzed by Orley Ashenfelter?

- The result indicates that the variation of the prices of the wines in this dataset is explained much less by the age of the wine in comparison to Bordeaux wines.
- The result indicates that the variation of the prices of the wines in this dataset is explained much more by the age of the wine in comparison to Bordeaux wines.
- The age of the wine has no predictive power on the wine prices in both the datasets.

*Solution.* In comparison to the results for the Bordeaux wine data, the trainint (model) $R^2$ and test $R^2$ is higher for this new dataset. This seems to indicate that the variation in the prices of the wine in the dataset is explained much more by the age of the wines in comparison to the Bordeaux dataset.

(g) Construct a multiple regression model to fit the log of the price at which the wine was auctioned in 1991 with all the possible predictors (**age91, temp, hrain, wrain, tempdiff**) in the training dataset. To fit your model, use the data for wines made up to (and including) the year 1981. What is the R-squared for the model?

*Solution.* $R^2$ for this model $= 0.7938$.

(h) Is this model preferred to the model with only the age variable as a predictor (use the adjusted R-squared for the model to decide on this)?

*Solution.* With only the age variable, adjusted $R^2 = 0.65$. With all the variables, adjusted $R^2 = 0.7145$. This seems to indicate that the latter model (with more variables included) is preferred.

(i) Which among the following best describes the output from the fitted model?

- The result indicates that less the temperature, the better is the price and quality of the wine

- The result indicates that greater the temperature difference, the better is the price and quality of wine.
- The result indicates that lesser the harvest rain, the better is the price and quality of the wine.
- The result indicates that winter rain is a very important variable in the fit of the data.

*Solution.* The result indicates that the lesser the harvest rain, the better the price and quality of the wine will be. This is because the corresponding $\beta = 0.002$ and is significcant at the 0.1 level. All other statements appear to be false.

(j) Of the five variables (**age91, temp, hrain, wrain, tempdiff**), drop the two variables that are the least significant from the results in (g). Rerun the linear regression and write down your fitted model.

*Solution.* The least significant variables are *wrain* and *tempdiff* with $p$-values 0.53 and 0.416 respectively.

(k) Is this model preferred to the model with all variables as predictors (use the adjusted R-squared in the training set to decide on this)?

*Solution.* In the training set, adjusted $R^2$ for this model is 0.73 while for model2, adjsuted $R^2 = 0.7145$. In this case, the new model3 is preferred to model2.

(l) Using the variables identified in (j), construct a multiple regression model to fit the log of the price at which the wine was auctioned in 1992 (remember to use **age92** instead of **age91**). To fit your model, use the data for wines made up to (and including) the year 1981. What is the R-squared for the model?

*Solution.* $R^2$ for this model is 0.5834.

(m) Suppose in this application, we assume that a variable is statistically significant at the 0.2 level. Would you would reject the hypothesis that the coefficient for the variable **hrain** is nonzero?

*Solution.* The $p$-value for *hrain* is 0.32. Hence we canot reject the null hypothesis that the $\beta$ coefficient for *hrain*is zero.

(n) By separately estimating the equations for the wine prices for each auction, we can better establish the credibility of the explanatory variables because:

- We have more data to fit our models with.
- The effect of the weather variables and age of the wine (sign of the estimated coefficients) can be checked for consistency across years.
- 1991 and 1992 are the markets when the Australian wines were traded heavily.

Select the best option.

*Solution.* The best explanation seems to be that we can check for consistency of the effect of weather variables and age by looking at the sign of the estimated coefficients.

(o) The current fit of the linear regression using the weather variables drops all observations where any of the entries are missing. Provide a short explanation on when this might not be a reasonable approach to use.

*Solution.* Clearly, dropping missing entries is reliable. However, if there are many missing entries, then this implies we can lose a lot of data.

*R Scripts.*

```
> #6a)
> wine<-read.csv("winedata.csv")
> str(wine)
'data.frame': 25 obs. of  7 variables:
 $ vintage : int  1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 ...
 $ price91 : num  131.5 156 118 92.8 119.6 ...
 $ price92 : num  132 181 180 115 157 ...
 $ temp    : num  18.8 18.8 18 NA 18.6 ...
 $ hrain   : num  20.3 41 42.1 41.1 0.5 ...
 $ wrain   : num  328 411 453 372 374 ...
 $ tempdiff: num  7.36 7.21 6.84 7.76 8.38 7.26 6.88 7.2 6.58 7.47 ...
> wine$age91<-1991-wine$vintage
> wine$age92<-1992-wine$vintage
> mean(subset(wine$price91,wine$age91>=15))
[1] 96.43563


> #6b)
> mean(subset(wine$price91,wine$hrain<mean(wine$hrain)&wine$tempdiff<mean(wine$tempdiff)))
[1] 72.86714

> #6c)
> train<-subset(wine,vintage<=1981)
> model1<-lm(log(price91)˜age91,data=train)
> summary(model1)

Call:
```

```
lm(formula = log(price91) ~ age91, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26897 -0.13328  0.01939  0.10452  0.41913 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.571442   0.144163  24.774 6.30e-16 ***
age91       0.042610   0.006899   6.176 6.19e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1914 on 19 degrees of freedom
Multiple R-squared:  0.6675,Adjusted R-squared:   0.65 
F-statistic: 38.15 on 1 and 19 DF,  p-value: 6.188e-06


> #6d)
> confint(model1,  level = 0.99)
                 0.5 %     99.5 %
(Intercept) 3.15900093 3.98388288
age91       0.02287254 0.06234705


> #6e)
> test<-subset(wine,vintage>=1982)
> predtest<-predict(model1,newdata=test)
> sse<-sum((log(test$price91)-predtest)^2)
> sst<-sum((log(test$price91)-mean(log(train$price91)))^2)
> testR2<- 1-sse/sst
> testR2
[1] 0.9213742


> #6g)
> model2<-lm(log(price91)~temp+hrain+wrain+tempdiff+age91,data=train)
> summary(model2)

Call:
lm(formula = log(price91) ~ temp + hrain + wrain + tempdiff + 
    age91, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32298 -0.07019 -0.01634  0.11051  0.24455 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2626138  1.4914578   1.517   0.1532
temp         0.1135015  0.0734344   1.546   0.1462
```

```
hrain       -0.0028825  0.0016205  -1.779   0.0987 .
wrain        0.0002520  0.0003918   0.643   0.5313
tempdiff    -0.1213280  0.1445947  -0.839   0.4166
age91        0.0482331  0.0075346   6.402 2.34e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1803 on 13 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7938,Adjusted R-squared:  0.7145
F-statistic: 10.01 on 5 and 13 DF,  p-value: 0.000421


> #6j)
> model3<-lm(log(price91)~temp+hrain+age91,data=train)
> summary(model3)

Call:
lm(formula = log(price91) ~ temp + hrain + age91, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32198 -0.09905  0.00491  0.14536  0.29828

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.801013   1.297697   1.388    0.185
temp         0.097500   0.068750   1.418    0.177
hrain       -0.001983   0.001236  -1.604    0.130
age91        0.045670   0.006702   6.814 5.85e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1752 on 15 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7753,Adjusted R-squared:  0.7304
F-statistic: 17.26 on 3 and 15 DF,  p-value: 3.973e-05


> #6l)
> model4<-lm(log(price92)~temp+hrain+age92,data=train)
> summary(model4)

Call:
lm(formula = log(price92) ~ temp + hrain + age92, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29394 -0.15545  0.03238  0.10221  0.37661

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.723393   1.574961   1.729 0.104296
temp         0.072494   0.083333   0.870 0.398043
hrain       -0.001539   0.001498  -1.027 0.320713
age92        0.035237   0.008124   4.338 0.000586 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2123 on 15 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.5834,Adjusted R-squared:  0.5001
F-statistic: 7.002 on 3 and 15 DF,  p-value: 0.003621
```

7. In his 2003 bestseller Moneyball, the author Michael Lewis makes an interesting claim that the Oakland Athletics team manager Billy Beane felt that the valuation of skills in the market for baseball players at that time was inefficient and undervalued important batting skills such as on-base percentage while overvalued batting skills such as slugging percentage. In this question, we will verify with data if these claims made by Lewis are indeed true by looking at data on player performances and salaries for the 1996 and 2006 seasons. The data is provided in the file **batters.csv** and contains the following variables:

   - **playerID**: Player identity code
   - **yearID**: Year
   - **teamID**: Team identity code
   - **G**: Number of games in which the player played during the season
   - **AB**: At Bats
   - **R**: Runs
   - **H**: Hits (Times reached base because of a batted, fair ball without error by the defense)
   - **X2B**: Doubles (Hits on which the batter reached second base safely)
   - **X3B**: Triples (Hits on which the batter reached third base safely)
   - **HR**: Homeruns
   - **BB**: Base on balls
   - **HBP**: Hit by pitch
   - **SF**: Sacrifice flies
   - **salary**: Salary for players at the start of the next season

   (a) Read the data into the dataframe **batters**. Which player made the most salary in the 2006 season?

   *Solution.* the `which.max` function indicates the batter at row 1159 made the maximum salary – whose name is coded as *giambja01*.

   (b) What is the ratio of the maximum salary to the minimum salary among all players in the 2006 season?

   *Solution.* Ratio is 61.65 – indicating the top paid player made 61 times what the least paid player made.

   (c) At the end of the 1996 season, which teams had the set of batters with the minimum and maximum total sum of salaries respectively?

*Solution.* OAK (Oakland Athletics) had the smallest payroll while NYA (New York Yankees) had the highest payroll.

(d) Write down the R command(s) that you used to answer question (c).

(e) Plot the histogram of the **salary** variable. What best describes the distribution of player salaries?

- Most of the salaries are large, with a relatively small number of much smaller salaries (this is referred to as "left-skewed").
- The salaries are balanced, with equal numbers of unusually large and unusually small salaries.
- Most of the salaries are small, with a relatively small number of much larger salaries (this is referred to as "right-skewed").

*Solution.* Plot found here. Most of the salaries are small with a relatively small number of much larger salaries (right-skewed).

(f) When handling a skewed dependent variable, it is often useful to predict the logarithm of the dependent variable instead of the dependent variable itself - this prevents the small number of unusually large or small observations from having an undue influence on the predictive model. In this problem, you will predict the natural logarithm of the salary variable at the end of a season with the number of runs scored in the season and a constant (intercept). Use the entire dataset to build your model with linear regression. What does your model predict to be the logarithm of the salary of a batter who scores 0 runs in a season?

*Solution.* The model fit gives:

$$log\,(salary) = 13.41 + 0.0149R$$

When $R = 0,\ log\,(salary) = 13.41$.

(g) What is the actual average of the logarithm of the salary of batters who score 0 runs in a season in the dataset? Remember to drop missing entries in computing this number.

*Solution.* The actual average in the dataset with $R = 0$ or the logarithm of salary is 13.60.

(h) Comment on whether the results in questions (f) and (g) are close to each other. If yes provide a brief explanation.

*Solution.* The values are close. Note that if you only regress $log\,(salary)$ with a constant, the best fit would be the mean, which is the reason for this result.

(i) Assume that the number of runs scored by a player increases by 1. Suppose $\beta_1$ is the coefficient of the number of runs scored in question (f). What best describes how your model would predict the change in the salary?

- New salary = Old salary + $e^{\beta_1}$
- New salary = Old salary × $e^{\beta_1}$
- New salary = Old salary + $\beta_1$
- New salary = Old salary × $\beta_1$

*Solution.*

$$log\,(salary) = \beta_0 + \beta_1 R \implies salary = e^{\beta_0 + \beta_1 R}$$

If runs increases by 1, we have $salary_{new} = salary_{old} \cdot e^{\beta_1}$.

(j) We will now compare the effect of two baseball statistics on the salaries of the players. To do this, we need to define two new variables **OBP** (on-base percentage) and **SLG** (slugging percentage) as follows:

$$\text{OBP} = \frac{\text{Hits} + \text{Base on balls} + \text{Hit by pitch}}{\text{At Bats} + \text{Base on balls} + \text{Hit by pitch} + \text{Sacrifice flies}}$$

$$\text{SLG} = \frac{\text{Hits} + \text{Doubles} + 2 \times \text{Triples} + 3 \times \text{Homeruns}}{\text{At Bats}}$$

What is the average on-base percentage in the 2006 season? Drop observations with missing entries.

*Solution.* The average on-base percentage in the 2006 season was 0.2707.

(k) Perform a two sided t-test to check if the average slugging percentage in the 1996 and 2006 seasons are different. What is the p-value of the test and your conclusion?

*Solution.* $p$-value of test is 0.4045, which seems to indicate that there is not enough evidence to reject the null hypothesis that the average slugging percentage of the 1996 and 2006 seasons are the same.

(l) We will now use linear regression to predict the logarithm of the salary using the **OBP** and **SLG** variables and the constant (intercept). To build the model, we will consider only batters with at least 130 at-bats, since this is required to qualify as honors for rookie of the year and helps provide an objective cutoff to check the effect of performance on players with relatively large sample of at-bats. Using only data for the year 1996, what is adjusted R-squared for your model?

*Solution.* Adjusted $R^2 = 0.2589$.

(m) Is there evidence that you can reject the three null hypothesis H0 : $\beta_j = 0$ for the **OBP**, **SLG** and constant variables? Use a p-value of 0.05 to make your conclusion.

*Solution.* Since all the $p$-values are very close to zero, there is enough evidence to indicate that we can reject each of the null hypotheses that $\beta_j = 0$.

(n) Redo the linear regression from question (l) using only data for 2006. What is the adjusted R-squared for your model?

*Solution.* Adjusted $R^2 = 0.1164$.

(o) Billy Beane, the Oakland Athletics coach believed that on-base percentage was much more important than the slugging percentage to help win a game. By looking at the coefficients of the **OBP** and **SLG** variables from the regressions in questions (l) and (n), we can conclude that:

- The market undervalued the **SLG** statistic relative to the **OBP** statistic in 1996 before Moneyball was published. This inefficiency still remained in 2006.
- The market undervalued the **SLG** statistic relative to the **OBP** statistic in 1996 before Moneyball was published. This has been corrected in 2006.
- The market undervalued the **OBP** statistic relative to the **SLG** statistic in 1996 before Moneyball was published. This inefficiency still remained in 2006.
- The market undervalued the **OBP** statistic relative to the **SLG** statistic in 1996 before Moneyball was published. This has been corrected in 2006.

Select the best option.

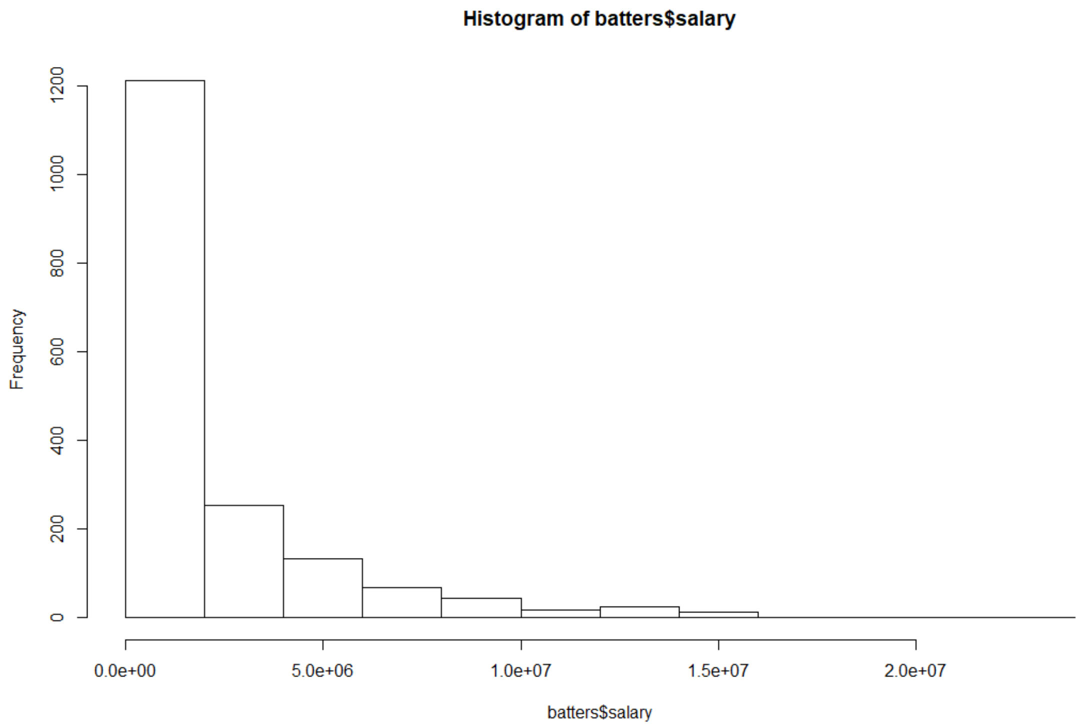
*Solution.* In model 2 (1996), we have $\beta_{OBP} = 4.87$, $\beta_{SLG} = 5.46$. In model 3 (2006), we have $\beta_{OBP} = 6.64$, $\beta_{SLG} = 2.90$. In both models, both of the variables are statistically significant.

The market undervalued OBP compared to SLG in 1996 ($4.87 < 5.46$) before Moneyball was published. This has been corrected since in 2006 ($6.64 > 2.90$).

*R Scripts.*

```
> #7a)
> batters <- read.csv("batters.csv")
> str(batters)
'data.frame':	1770 obs. of  14 variables:
 $ playerID: Factor w/ 1474 levels "aardsda01","abbotku01",..: 2 3 5 6 7 8 10 11 11 12 ...
 $ yearID  : int  1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 ...
 $ teamID  : Factor w/ 33 levels "ARI","ATL","BAL",..: 12 13 23 12 7 23 18 11 18 3 ...
 $ G       : int  109 15 22 9 69 12 19 11 25 54 ...
 $ AB      : int  320 22 NA 0 6 NA NA NA NA 68 ...
 $ R       : int  37 1 NA 0 0 NA NA NA NA 6 ...
 $ H       : int  81 5 NA 0 0 NA NA NA NA 7 ...
 $ X2B     : int  18 1 NA 0 0 NA NA NA NA 0 ...
 $ X3B     : int  7 0 NA 0 0 NA NA NA NA 0 ...
 $ HR      : int  8 0 NA 0 0 NA NA NA NA 0 ...
 $ BB      : int  22 2 NA 0 1 NA NA NA NA 3 ...
 $ HBP     : int  3 0 NA 0 0 NA NA NA NA 0 ...
 $ SF      : int  0 0 NA 0 0 NA NA NA NA 0 ...
 $ salary  : int  650000 150000 165000 151000 215000 157500 3075000 225000 225000 205000 ...
> which.max(batters$salary)
[1] 1159
> batters[1159,]
      playerID yearID teamID   G  AB  R   H X2B X3B HR  BB HBP SF   salary
1159 giambja01   2006    NYA 139 446 92 113  25   0 37 110  16  7 23428571


> #7b)
> max(batters$salary[batters$yearID==2006])/min(batters$salary[batters$yearID==2006])
[1] 61.65413

> #7cd)
> sort(tapply(batters$salary[batters$yearID==1996],batters$teamID[batters$yearID==1996],sum))
     OAK      MIN      KCA      PIT      CAL      ML4      MON      CHN      DET      FLO
23380333 23632500 24285000 28411667 30474500 30574338 33087500 35139004 35181000 36982000
     NYN      TOR      SFN      PHI      HOU      CIN      SDN      COL      SLN      SEA
37561150 37857333 38197380 38682500 39580000 40065000 41141672 41884667 43622667 45319494
     LAN      TEX      BOS      CHA      CLE      ATL      BAL      NYA
45983304 51743005 51776167 53703500 63646793 64987500 65094734 73661183


> #7e)
> hist(batters$salary)

> #7f)
> model1 <- lm(log(salary)~R,data=batters)
> summary(model1)

Call:
lm(formula = log(salary) ~ R, data = batters)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8332 -0.8636 -0.1999  0.8610  3.1706
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.417505   0.037924  353.80   <2e-16 ***
R            0.014991   0.000872   17.19   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.192 on 1591 degrees of freedom
  (177 observations deleted due to missingness)
Multiple R-squared:  0.1567,Adjusted R-squared:  0.1561
F-statistic: 295.5 on 1 and 1591 DF,  p-value: < 2.2e-16


> #7g)
> mean(log(batters$salary[batters$R==0]),na.rm=TRUE)
[1] 13.60544

> #7j)
> batters$OBP <- (batters$H+batters$BB+batters$HBP)/(batters$AB+batters$BB+batters$HBP+batters$SF)
> batters$SLG <- (batters$H+batters$X2B+2*batters$X3B+3*batters$HR)/(batters$AB)
> mean(batters$OBP[batters$yearID==2006],na.rm=TRUE)
[1] 0.2707986

> #7k)
> t.test(batters$SLG[batters$yearID==1996],batters$SLG[batters$yearID==2006])

Welch Two Sample t-test

data:  batters$SLG[batters$yearID == 1996] and batters$SLG[batters$yearID == 2006]
t = 0.83382, df = 1340.7, p-value = 0.4045
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01199352  0.02972624
sample estimates:
mean of x mean of y
0.3397733 0.3309069


> #7l)
> model2 <- lm(log(salary)~OBP+SLG,data=subset(batters,batters$yearID==1996&batters$AB>=130))


> #7n)
> model3 <- lm(log(salary)~OBP+SLG,data=subset(batters,batters$yearID==2006&batters$AB>=130))
> summary(model3)

Call:
lm(formula = log(salary) ~ OBP + SLG, data = subset(batters,
    batters$yearID == 2006 & batters$AB >= 130))

Residuals:
    Min      1Q  Median      3Q     Max
-2.7655 -1.0813  0.1921  0.9722  2.3209

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.9822     0.5647  19.448  < 2e-16 ***
OBP           6.6440     2.2181   2.995  0.00294 **
SLG           2.9093     1.1125   2.615  0.00930 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.152 on 350 degrees of freedom
Multiple R-squared:  0.1214,Adjusted R-squared:  0.1164
F-statistic: 24.19 on 2 and 350 DF,  p-value: 1.443e-10
```



Plot for Q7e. Click here to go back to the question.