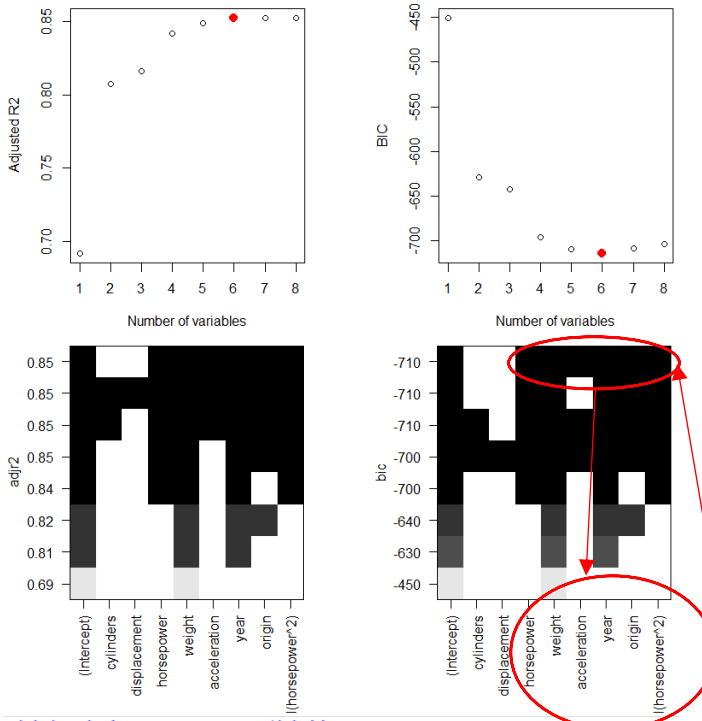
Exercise solutions for exercise class 4 in MMS075, Feb 11, 2020

1. We run best subset selection for predicting miles per gallon values in the Auto dataset. This dataset contains 392 observations on the following variables:
mpg: miles per gallon; **cylinders**: Number of cylinders between 4 and 8; **displacement**: Engine displacement (cu. inches); **horsepower**: Engine horsepower; **weight**: Vehicle weight (lbs.); **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

   We do not want to use vehicle name as a predictor, but we have seen earlier a quadratic relationship between horsepower and mpg, so we add horsepower^2 as a predictor. We get the following outputs in R:

```
          cylinders displacement horsepower weight acceleration year origin I(horsepower^2)
1  ( 1 ) " "       " "          " "        "*"    " "          " "  " "    " "
2  ( 1 ) " "       " "          " "        "*"    " "          "*"  " "    " "
3  ( 1 ) " "       " "          " "        "*"    " "          "*"  "*"    " "
4  ( 1 ) " "       " "          "*"        "*"    " "          "*"  " "    "*"
5  ( 1 ) " "       " "          "*"        "*"    " "          "*"  "*"    "*"
6  ( 1 ) " "       " "          "*"        "*"    "*"          "*"  "*"    "*"
7  ( 1 ) "*"       " "          "*"        "*"    "*"          "*"  "*"    "*"
8  ( 1 ) "*"       "*"          "*"        "*"    "*"          "*"  "*"    "*"
> BestMPGsummary$rsq
[1] 0.6926304 0.8081803 0.8174522 0.8430949 0.8506034 0.8546931 0.8548282 0.8552261
```




```
> coef(BestMPGModel,which.min(BestMPGsummary$bic))
  (Intercept)    horsepower        weight  acceleration          year        origin I(horsepower^2)
1.2151778228 -0.3087574097 -0.0034454734 -0.3122721115 0.7367124307 1.0847454604  0.0009614144
```

   Based on this information, answer the following questions:

   a) What are the predictors in the model selected by the best subsect selection algorithm using BIC for assessing model quality? Write the model equation!
   The predictors are those that are marked with black squares in the top row of the bic plot: horsepower, weight, acceleration, year, origin and (horsepower)^2. For the model equation, we take the coefficient from the output in the line below the plots:

$$\widehat{\text{mpg}} = 1.215 - 0.309 \times \text{horsepower} - 0.003 \times \text{weight} - 0.312 \times \text{acceleration} + 0.737 \times \text{year} + 1.085 \times \text{origin} + 0.001 \times (\text{horsepower})^2$$

b) What is the final model selected by the best subsect selection algorithm using adjusted R^2 for assessing model quality? Write the model equation!

The predictors in the final model for best subsect selection algorithm using adjusted R^2 are those that are marked with black squares in the top row of the adjr2 plot (which is to the left of the bic plot). Looking at the plot, we see that in this case, the final model has exactly the same predictors as the other final model (which is not automatic, see an example of different final models in the Lecture 4 presentation).

As for the model equation, we see that both BIC and adjusted R^2 indicate that the best 6-predictor model should be chosen. The coefficients for the best 6-predictor model do not depend on whether BIC or adjusted R^2 was used, because R considers minimum RSS for determining the best model of a given number of variables; therefore, we can again use the above coefficients to write the model equation which will then be the same as in part a) above:

$$\widehat{\text{mpg}} = 1.215 - 0.309 \times \text{horsepower} - 0.003 \times \text{weight} - 0.312 \times \text{acceleration} + 0.737 \times \text{year} + 1.085 \times \text{origin} + 0.001 \times (\text{horsepower})^2$$

c) How well does the best model fit the mpg values?

We take R^2 as a measure of model fit (the other possibility would be RSE, but that is not specified in the outputs given for this exercise). The R^2 values are given in the line above the plots, which is copied here for convenience:

```
> BestMPGsummary$rsq
[1] 0.6926304 0.8081803 0.8174522 0.8430949 0.8506034 0.8546931 0.8548282 0.8552261
```

The best model is the model with 6 predictors; therefore, its R^2 value will be the 6th number in this output, i.e. 0.855. This indicates that about 85.5% of the variation in mpg values can be explained by the best model whose equation was given in parts a) and b).

2. Recall that the prediction model for Advertising example was as follows:

$$\widehat{\text{sales}} = 6.7502 + 0.0191 \times \text{TV} + 0.0289 \times \text{radio} + 0.0011 \times \text{TV} \times \text{radio}$$

Predict the number of sold units using this model for the following advertisement budget distributions:

a) TV budget: \$0, radio budget: \$100 000;

We plug TV=0, radio=100 into the equation above, because the variables were defined in terms of \$1000. This way, we get an estimated value of the sales variable:

$$\begin{aligned}\widehat{\text{sales}} &= 6.7502 + 0.0191 \times 0 + 0.0289 \times 100 + 0.0011 \times 0 \times 100 \\ &= 6.7502 + 0.0289 \times 100 \\ &= 9.6402\end{aligned}$$

This estimated value gives the predicted number of sales in 1000 units. Therefore, with TV budget: \$0, radio budget: \$100 000, we predict 9640 sold units.

b) TV budget: \$100 000, radio budget: \$0;

Plugging TV=100, radio=0 into the above equation gives that

$$\widehat{\text{sales}} = 8.6602$$

and so, with TV budget: $100 000, radio budget: $0, we predict 8660 sold units. This is less than our estimate in part a) which is not surprising, because the coefficient of TV is lower than the coefficient of radio.

c) TV budget: $50 000, radio budget: $50 000;
Plugging TV=50, radio=50 into the above equation gives that

$$\widehat{\text{sales}} = 6.7502 + 0.0191 \times 50 + 0.0289 \times 50 + 0.0011 \times 50 \times 50$$
$$= 11.9002$$

and so, with TV budget: $50 000, radio budget: $50 000, we predict 11900 sold units. Note that this is much higher than the predictions in the previous two parts; according to this model, TV and radio advertisements strengthen each other's effect and therefore, sharing an available budget of $100 000 between them gives better results than investing all of it either in TV or in radio advertisements.

d) TV budget: $50 000, radio budget: $50 000, newspaper budget: $30 000.
As we are using a model that has radio, TV and their interaction as predictors, the budget on newspaper is irrelevant for the prediction of sold units. Therefore, we get the same prediction here as we got in part c), because the TV budget and radio budget are the same as in that part.

Furthermore, estimate the effect of:

e) $1000 increase of TV advertisement on sold units if the radio budget is $10 000;
The increase on the predicted value of the sales variable be the coefficient of TV in the equation, which is

$$0.0191 + 0.0011 \times \text{radio} = 0.0191 + 0.0011 \times 10 = 0.0301$$

Alternatively, one can just make two predictions in both of which the radio budget is $10 000 but the TV budget is increased by $1000 and take the difference of the predictions. For example, the prediction with TV=0, radio=10 is 7.0392 while for TV=1, radio=10, it is 7.0693, and the difference is indeed 0.0301. For your own understanding, check that you get the same difference between the predictions with TV=2, radio=10 and TV=1, radio=10 (or any other value for TV, for example, radio=10, TV=5217 and radio=10, TV=5216) as well!
This estimated value of sales gives the predicted number of sales in 1000 units; therefore, an increase of 0.0301 in this variable corresponds to an increase of 30.1 in the number of sold units.

f) $1000 increase of TV advertisement on sold units if the radio budget is $100 000;
In this case, using either method described in part e) with radio=10 replaced by radio=100, we get an estimated increase of 0.1291 in the sales variable. Therefore, we conclude that the effect of $1000 increase of TV advertisement if the radio budget is $100 000 is estimated to result in an increase of 129 in the number of sold units.

g) $1000 increase of radio advertisement on sold units if the TV budget is $50 000.
The effect on the sales variable can be computed analogously to parts e) and f):

$$0.0289 + 0.0011 \times 50 = 0.0839$$

Therefore, the estimated effect of $1000 increase of radio advertisement assuming a TV budget of $50 000 is an increase of 84 in the number of sold units.

Finally, predict the number of sold units for:

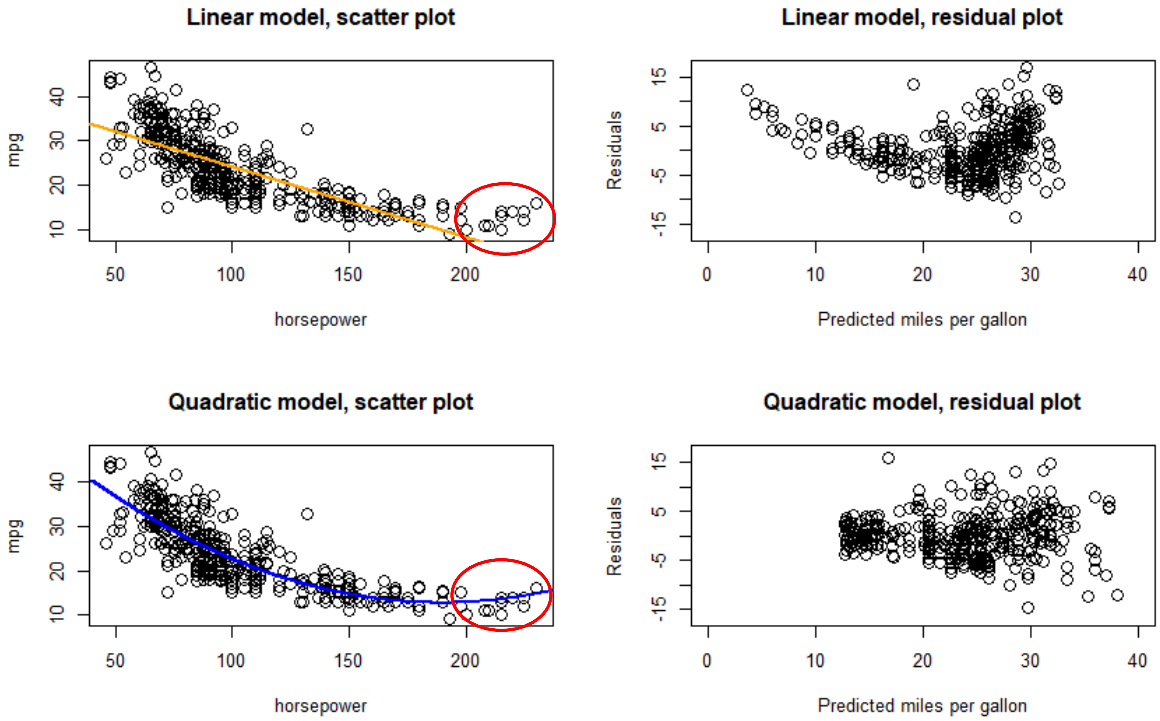
h) TV budget: $50 000, radio budget: $51 000.

This can be computed by plugging TV=50, radio=51 in the equation:

$$\widehat{\text{sales}} = 6.7502 + 0.0191 \times 50 + 0.0289 \times 51 + 0.0011 \times 50 \times 51$$
$$= 11.9841$$

Alternatively, we computed in part c) that the predicted number of sold units with TV budget: $50 000, radio budget: $50 000 is 11900, and in part g) that in case of a TV budget of $50 000, having an increase of $1000 in the radio budget results in an increase of 84 in the number of sold units; therefore, the number of sold units at TV budget: $50 000, radio budget: $51 000 is estimated as 11900 + 84 = 11984.

3. Check the mpg vs horsepower graphs presented in the lecture:



Based on these graphs, address the following points:

a) Why are there several points in the residual plot of the linear model with x-coordinates between 0 and 10 and no such points at all in the residual plot of the quadratic model?

Because the x-axis of the residual plot shows the predicted values by the model. In the linear model, these are represented by the orange line which goes under 10 for all horsepower values greater than (approximately) 200. However, in the quadratic model, the predicted values are represented by the blue curve, and that curve does not go below 10 for any horsepower values.

b) Find the corresponding points on the scatter plots!

By the argument presented in part a), those are the values corresponding to horsepower ≥ 200, which are now highlighted in the plots.

4. Consider a simple linear model predicting body weight with height as predictor, based on 20 observations. The R summary for the model is given below:

```
Call:
lm(formula = BodyWeight ~ Height)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9108  -4.2229   0.4685   8.1552  11.2213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -95.1240    25.3518  -3.752  0.00146 **
Height        0.9736     0.1427   6.821 2.19e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.093 on 18 degrees of freedom
Multiple R-squared:  0.7211,    Adjusted R-squared:  0.7056
F-statistic: 46.53 on 1 and 18 DF,  p-value: 2.192e-06
```

Now let us assume that I accidentally doubled the data before the analysis, i.e. copy the set of observations twice into a table before importing it in R without noticing this. This gives the following output in R:

```
Call:
lm(formula = BodyWeight ~ Height)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9108  -4.2229   0.4685   8.1552  11.2213

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -95.12397   17.44833  -5.452 3.21e-06 ***
Height        0.97359    0.09823   9.911 4.37e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.851 on 38 degrees of freedom
Multiple R-squared:  0.7211,    Adjusted R-squared:  0.7137
F-statistic: 98.23 on 1 and 38 DF,  p-value: 4.374e-12
```

Compare the two sets of results:

a) Which values in the summary remain the same?
   The coefficient estimates and the R^2 values remain the same. Doubling the points will not change which line fits the points best and what percentage of variability in body weight is explained by this model.

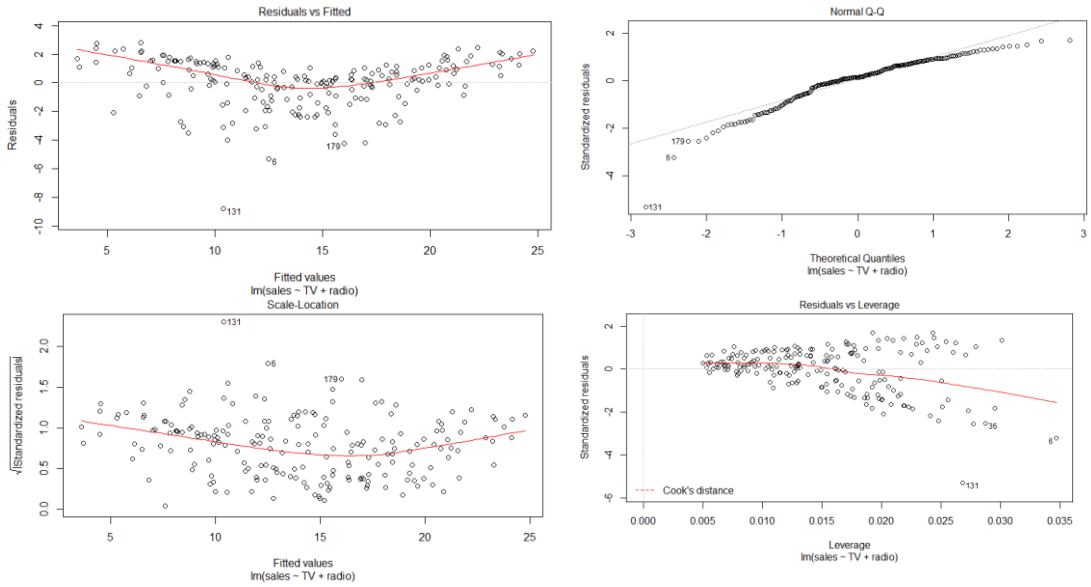
b) Which values change?
   All error-related values (i.e. standard errors of coefficients and RSE) decrease in the doubled model, because the same patterns are seen in a model that has twice as large sample size. Consequently, the t-value and F-statistic take on more extreme values in the doubled model, resulting in smaller p-values for variable significance and model significance.

c) Which model looks better based on the R output?
   Since we like small errors and small p-values, the doubled model looks better judged by the R output alone. This, however, is misleading, see part d) below.

d) Is there any assumption of linear regression violated in the second model based on doubled data?

The assumption of independent error terms is violated here, because the error terms in the doubled set are perfectly correlated with those in the original set of points. As mentioned in the presentation, having correlated error terms often gives a false feeling of certainty about the model as a result of underestimated errors.

5. See the default R plots below for the Advertisement model of sales with TV and radio budgets as predictors.



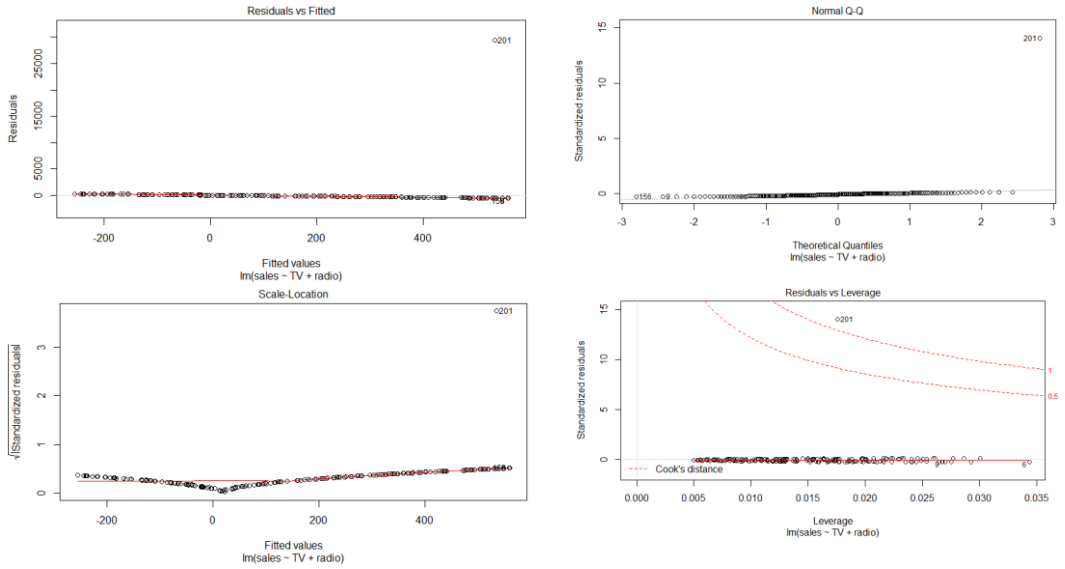
Based on these graphs, address the following points:

a) Do these graphs suggest any non-linear relationship between the response and the predictors?
Yes, the residual plot (i.e. the top-left plot) shows a curvy pattern which may be a sign of a non-linear relationship.

b) Are there any outliers in the model?
Based on the bottom-right graph, points 131 and maybe point 6 are outliers, because their standardized residuals are -3 or lower. Generally, we would like to check whether the *studentized* residuals are below -3 or above 3, so we would need a new graph to be certain, but the standardized residuals can also give a strong indication.

c) Are there any high leverage points in the model?
Based on the bottom-right graph, point 6 has highest leverage at 0.035. This is quite a bit higher than (p+1)/n = 3/200 = 0.015; e.g. in Bruce, P. & Bruce, A. (2017), Practical Statistics for Data Scientists, points with leverage above 2(p+1)/n are called high leverage points, which this point satisfies.

Now we add an extra point to the data where we forget that sales are measured in 1000 units, fit the model and create the default plots again:



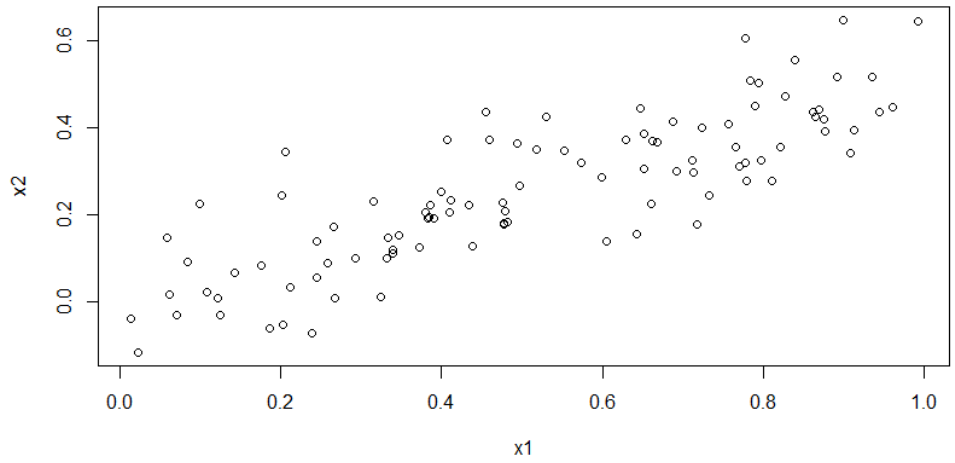
What are your answers to the questions a)-c) based on the updated graphs?

Non-linear effects are difficult to judge, because the new point has changed the y-scale on each graph so much that it is difficult to see any pattern. The new point is clearly an outlier based on its standardized residual and a point of influential value (see the bottom-right graph); however, it does not have an unusual predictor value and does not have a high leverage. Therefore, the only high leverage point here is point 6, as above. Note: the concepts of influential values and high leverage points will be discussed again at the next lecture to clarify potential confusion.

6. Do exercise 14 on page 125 of ISL. The R outputs required for the exercise are as follows:

The solution to this exercise will be provided in Computer lab 4 - R code example.pdf, to be uploaded in the Computer labs folder in Canvas.

Part b)



Part c)

```
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: 0.00001164
```

Part d)

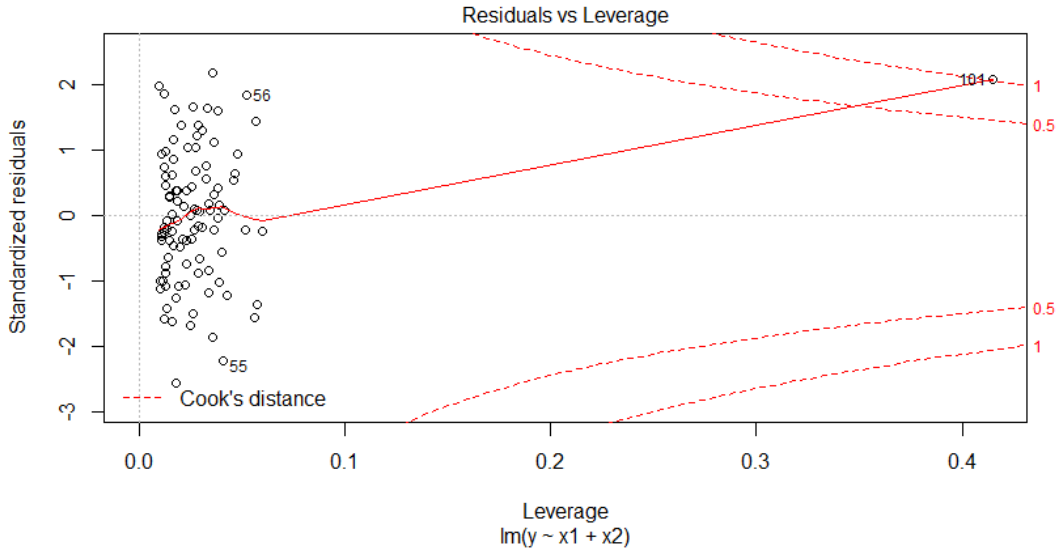
```
Call:
lm(formula = y ~ x1)

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
```

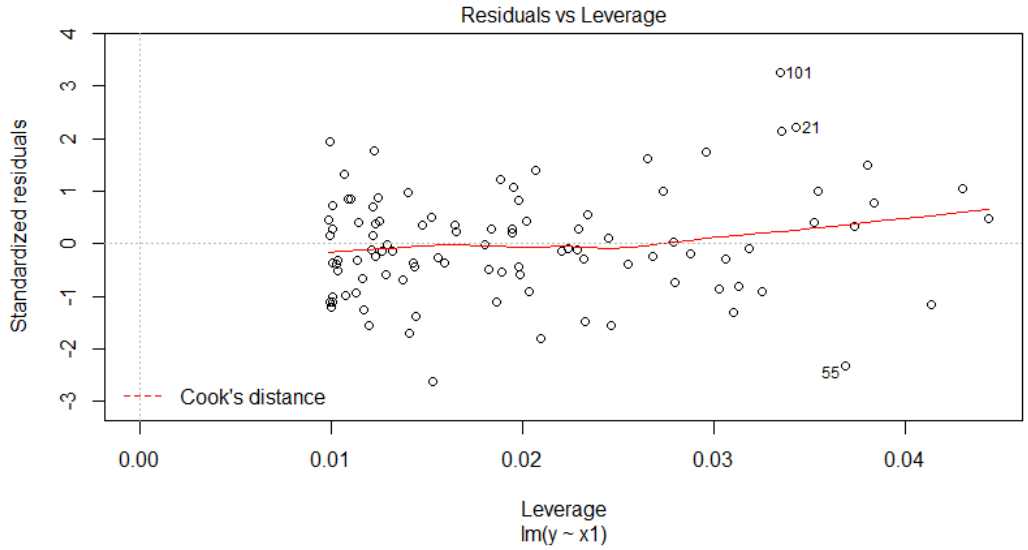
Part e)

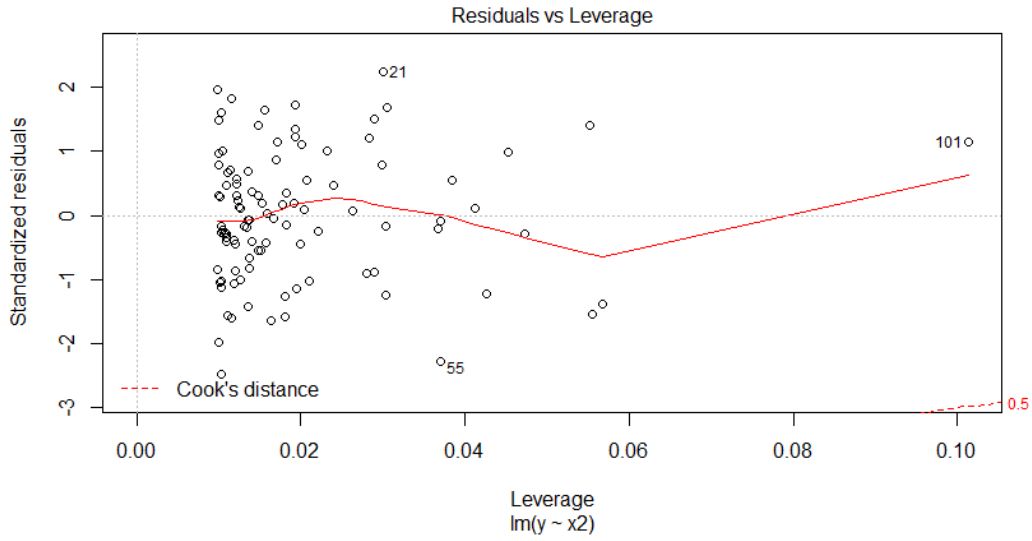
```
Call:
lm(formula = y ~ x2)

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 0.0000137 ***
---
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: 0.00001366
```

Part g) – NOTE: THESE ARE THE OUTPUTS WITH THE MISMEASURED OBSERVATION INCLUDED!

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.73348 -0.69318 -0.05263  0.66385  2.30619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
x1            0.5394     0.5922   0.911  0.36458
x2            2.5146     0.8977   2.801  0.00614 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.075 on 98 degrees of freedom
Multiple R-squared:  0.2188,    Adjusted R-squared:  0.2029
F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06


Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8897 -0.6556 -0.0909  0.5682  3.5665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
x1            1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared:  0.1562,    Adjusted R-squared:  0.1477
F-statistic: 18.33 on 1 and 99 DF,  p-value: 0.00004295


Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64729 -0.71021 -0.06899  0.72699  2.38074

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
x2            3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2042
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```







7. Feedback quiz (optional): Go to www.menti.com and use the code 30 81 24.