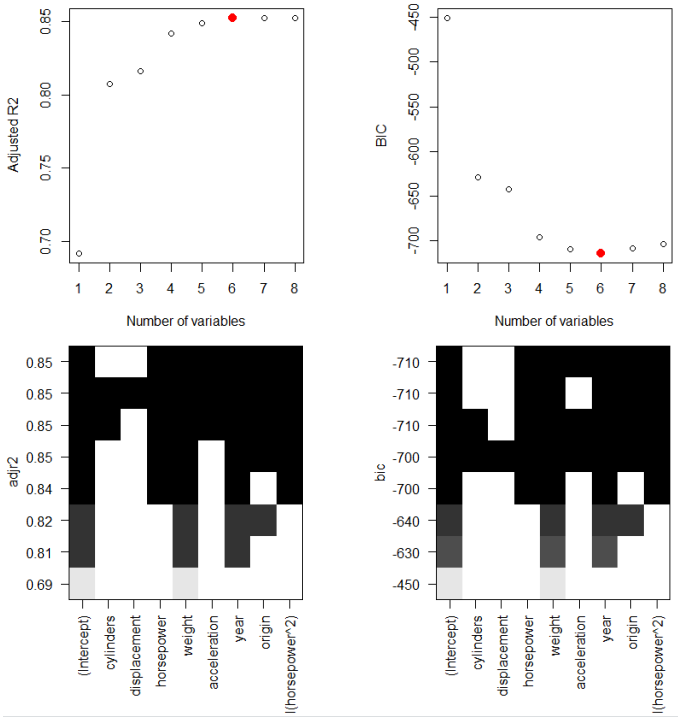
Exercises for exercise class 4 in MMS075, Feb 11, 2020

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

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;
b) TV budget: \$100 000, radio budget: \$0;
c) TV budget: \$50 000, radio budget: \$50 000;
d) TV budget: \$50 000, radio budget: \$50 000, newspaper budget: \$30 000.

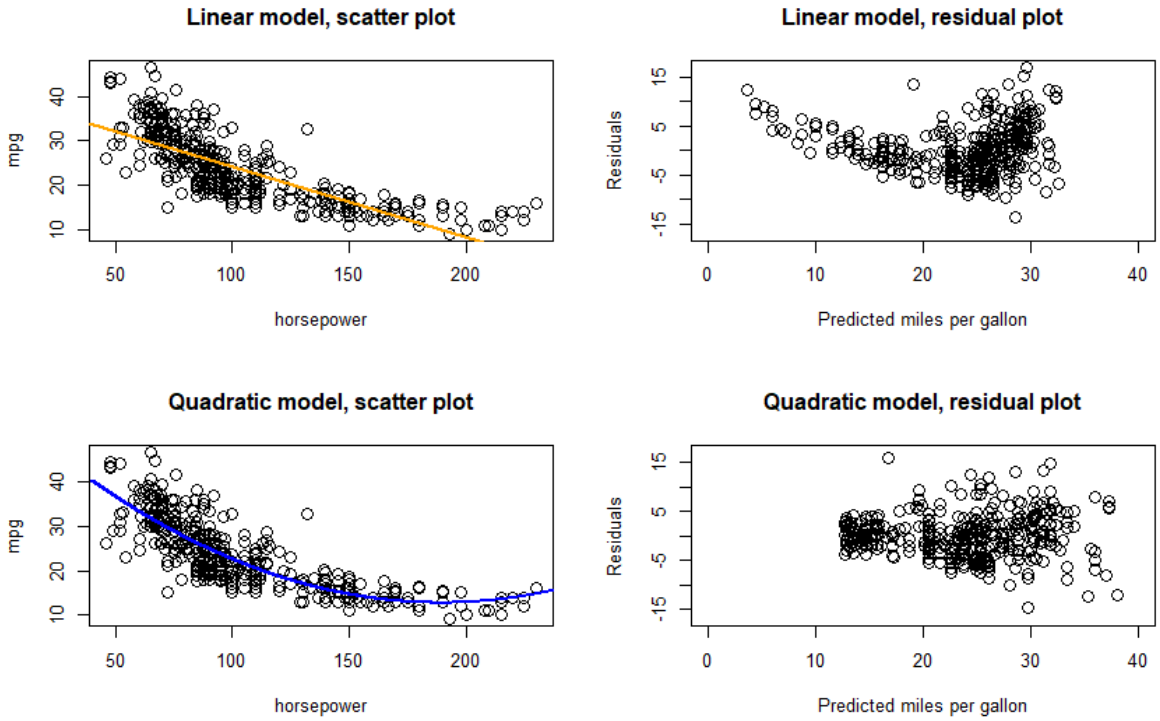
Furthermore, estimate the effect of:

e) \$1000 increase of TV advertisement on sold units if the radio budget is \$10 000;
f) \$1000 increase of TV advertisement on sold units if the radio budget is \$100 000;
g) \$1000 increase of radio advertisement on sold units if the TV budget is \$50 000.

Finally, predict the number of sold units for:

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

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?
b) Find the corresponding points on the scatter 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
```

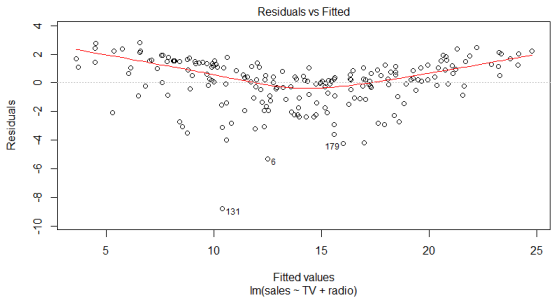
Now let us assume that I accidentally doubled the data before the analysis, i.e. copy the set of observations twice into a table before importing it in R without noticing this. This gives the following output in R:

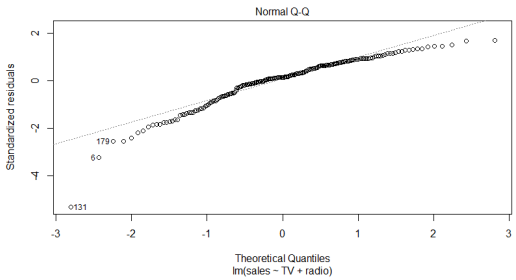
```
Call:
lm(formula = BodyWeight ~ Height)

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
```

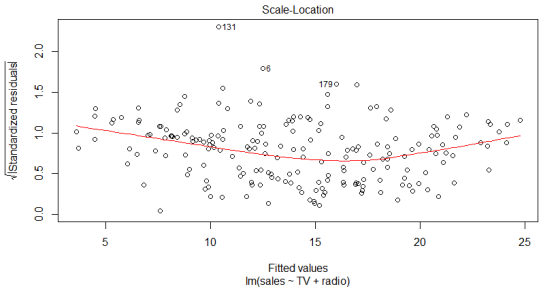
Compare the two sets of results:

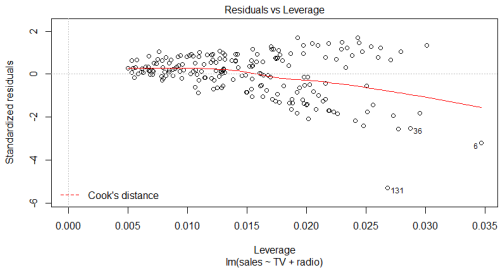
a) Which values in the summary remain the same?
b) Which values change?
c) Which model looks better based on the R output?
d) Is there any assumption of linear regression violated in the second model based on doubled data?

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









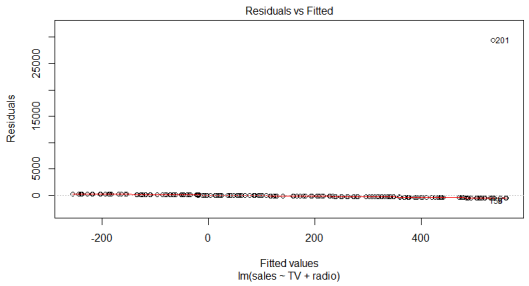
Based on these graphs, address the following points:

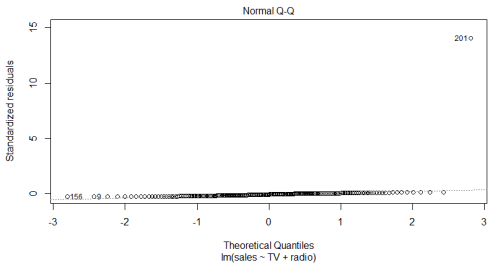
a) Do these graphs suggest any non-linear relationship between the response and the predictors?

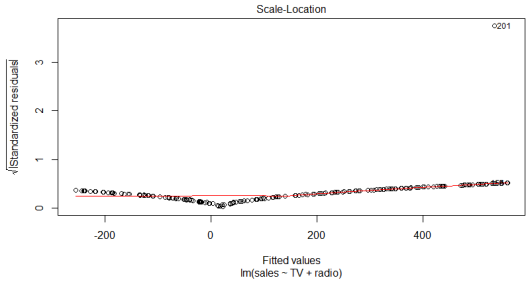
b) Are there any outliers in the model?

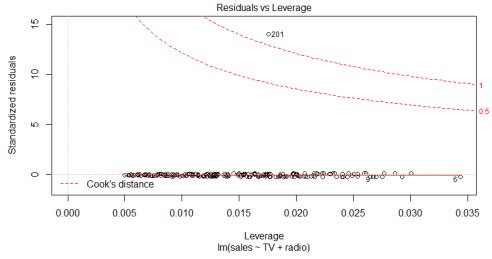
c) Are there any high leverage points in the model?

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:





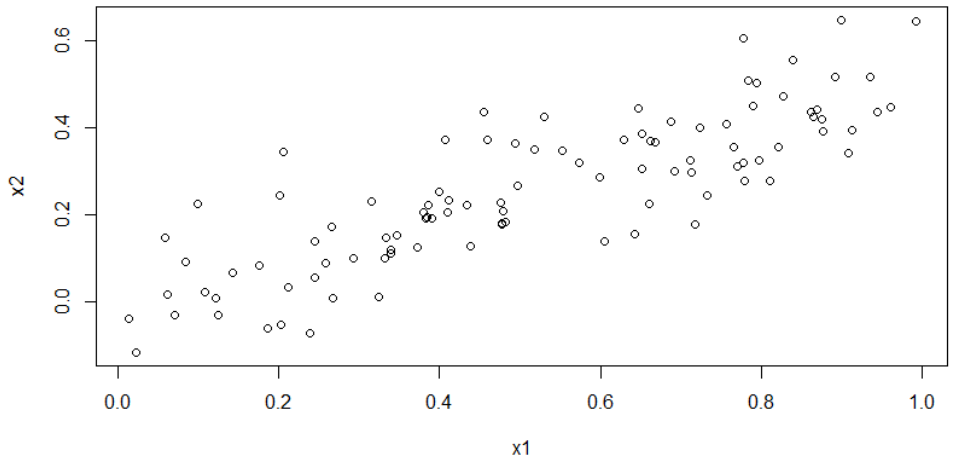




What are your answers to the questions a)-c) based on the updated graphs?

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

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)

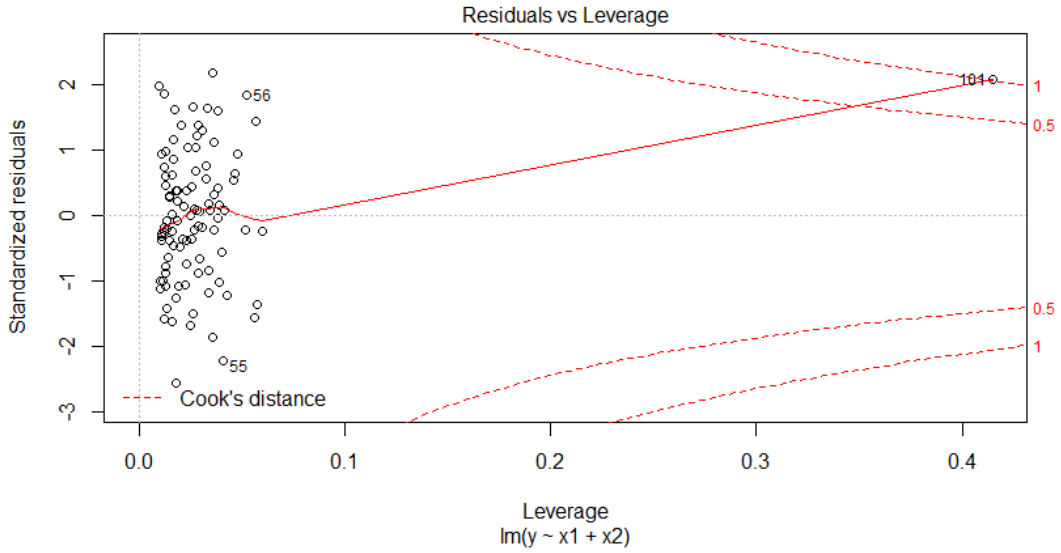
```
Call:
lm(formula = y ~ x1)

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
```

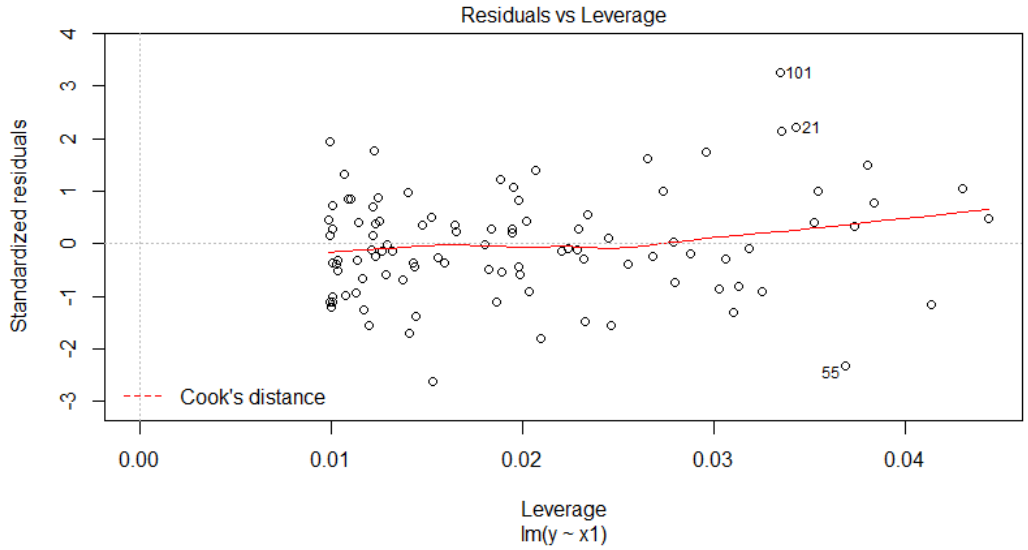
Part e)

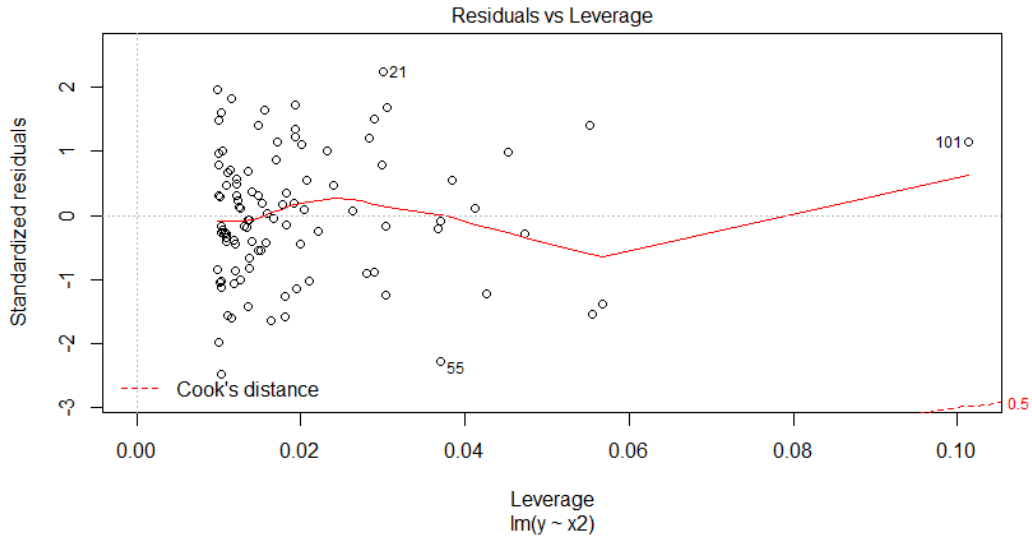
```
Call:
lm(formula = y ~ x2)

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.