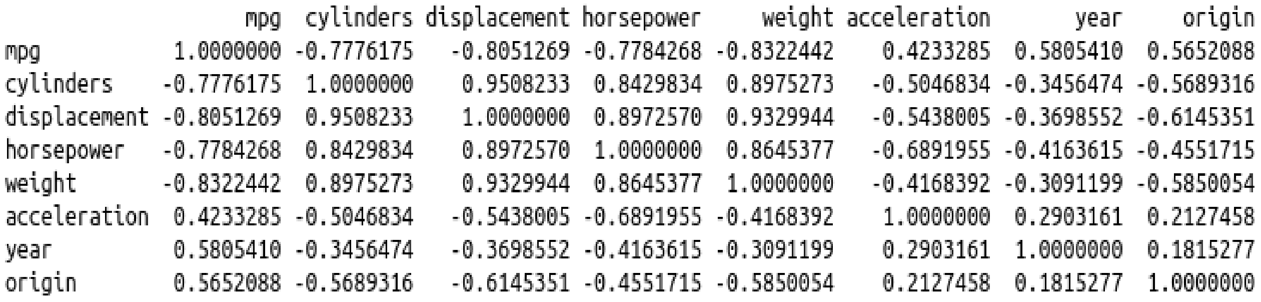
Computational Statistics

Lab 3
Subash Kharel

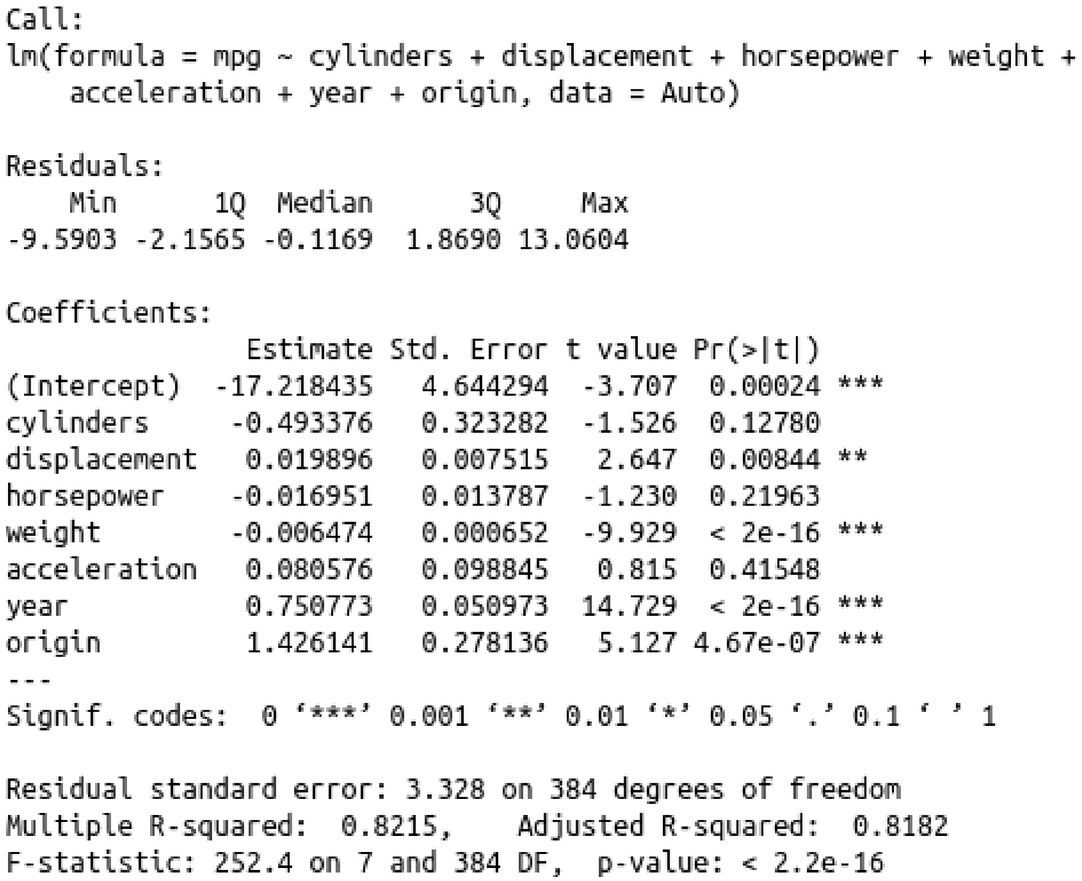
Feb – 17 - 2020

1. **Solution to 9**
   - **a.** The scatterplot is the Figure 1 which is in the PDF file attached.
   - **b.**



```
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

   - **c.** Output of the linear model:

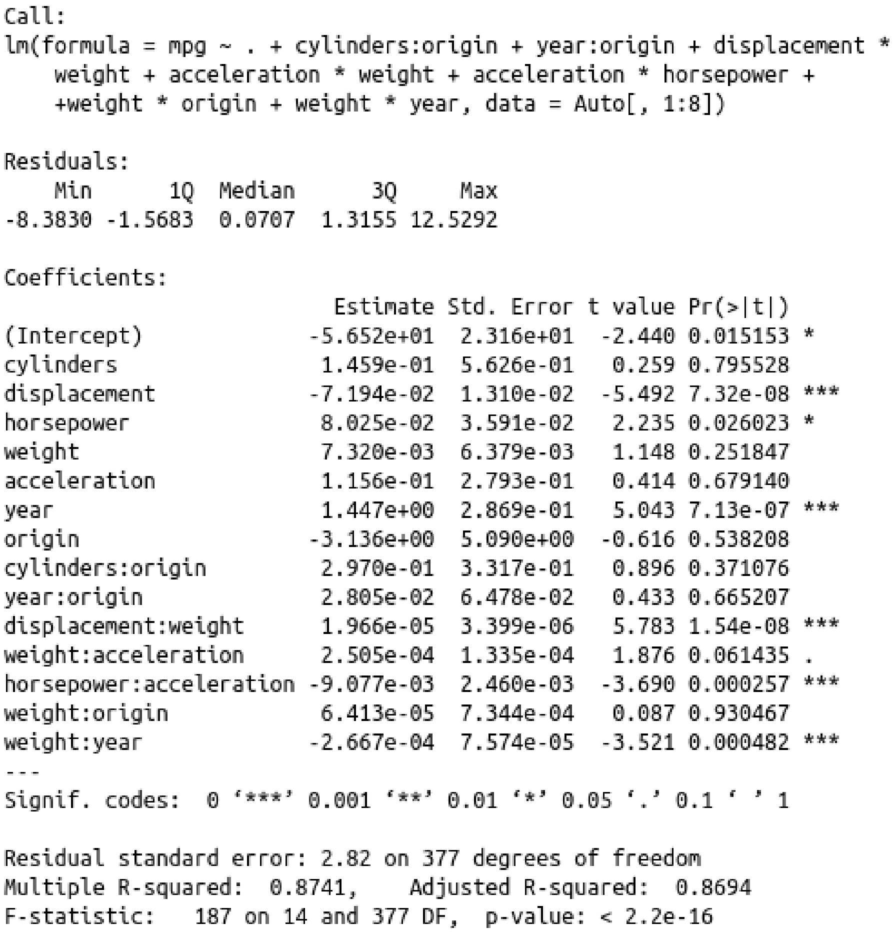


```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight + 
    acceleration + year + origin, data = Auto)

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
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

     - i. **Answer:** Of course, there exists a relation between the predictors and the model. Looking at the p-value of the different predictors, we can say that some of them have significant relation and some of them have negligible relation. But, at the end, there exists a relation among them.
     - ii. **Answer:** The predictor with the least p-value have the most significant relationship to the response. Looking at the output of the code, we can say that weight, year and origin have the most significant relationship to the response, i.e. the value of these predictors contribute to the response more than rest of those.

iii. **Answer**: The coefficient of the year variable suggest that for a single increase in value of year, the value for mpg is increased by 0.750773. It is the positive increment.

d. **Answer**: Looking at the residual vs fitted graph in Figure 2, there is almost horizontal line but the relation is not linear. #323, #326 and #327 seem out of the line and they are the outliers. This implies that these plots might be a potential problem. Normal Q-Q plot is a straight line, it suggests that they are normally distributed except that #323, #326 and #327 are little off the line. Spread-location plot is a horizontal line with residuals equally spreader except #323, #326 and #327 off the mark. At last, the Residual-leverage plot displays most of the points inside the cooks curve but the points #327 and #394 are high leverage points.



```
Call:
lm(formula = mpg ~ . + cylinders:origin + year:origin + displacement *
    weight + acceleration * weight + acceleration * horsepower +
    +weight * origin + weight * year, data = Auto[, 1:8])

Residuals:
    Min      1Q  Median      3Q     Max
-8.3830 -1.5683  0.0707  1.3155 12.5292

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.652e+01  2.316e+01  -2.440 0.015153 *
cylinders                1.459e-01  5.626e-01   0.259 0.795528
displacement            -7.194e-02  1.310e-02  -5.492 7.32e-08 ***
horsepower               8.025e-02  3.591e-02   2.235 0.026023 *
weight                   7.320e-03  6.379e-03   1.148 0.251847
acceleration             1.156e-01  2.793e-01   0.414 0.679140
year                     1.447e+00  2.869e-01   5.043 7.13e-07 ***
origin                  -3.136e+00  5.090e+00  -0.616 0.538208
cylinders:origin         2.970e-01  3.317e-01   0.896 0.371076
year:origin              2.805e-02  6.478e-02   0.433 0.665207
displacement:weight      1.966e-05  3.399e-06   5.783 1.54e-08 ***
weight:acceleration      2.505e-04  1.335e-04   1.876 0.061435 .
horsepower:acceleration -9.077e-03  2.460e-03  -3.690 0.000257 ***
weight:origin            6.413e-05  7.344e-04   0.087 0.930467
weight:year             -2.667e-04  7.574e-05  -3.521 0.000482 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.82 on 377 degrees of freedom
Multiple R-squared:  0.8741,    Adjusted R-squared:  0.8694
F-statistic:   187 on 14 and 377 DF,  p-value: < 2.2e-16
```

e. **Answer**: The interactions suggest that displacement:weight, horsepower:acceleration and weight:year are statistically significant, which can be concluded form the p-value in the output. weight:origin is the most insignificant interaction among above checked interactions.

f. **Answer**: From the residual plots in Figure 3, we can say that the relation between the mpg and weights is non-linear. Using various transformations, we can see that the transformation in figure 4, which is logarithmic transformation gives the best linear output compared to those in figure 5 and figure 6

2. Answer to 13

a. Check code

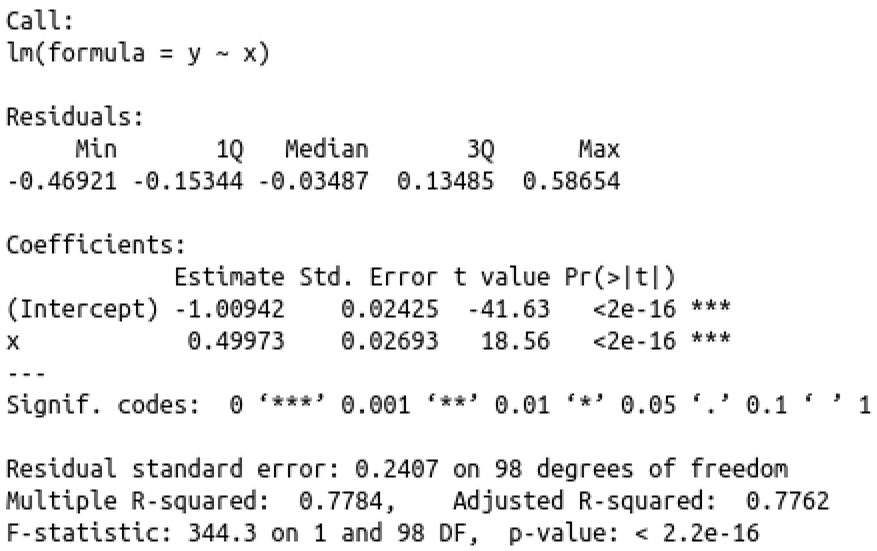
b. Check code

c. **Answer:** The length of vector y is 100. The values are:
β0 = -1
β1 = 0.5

d. The scatterplot (Figure 7) shows somewhat linear relation between x and y. Though there are few outliers, they are not causing the linear relation to deviate much.

e. Observation: Using the Least Square linear model, we get the following the values:
β0 = -1.00942

β1 = 0.49973

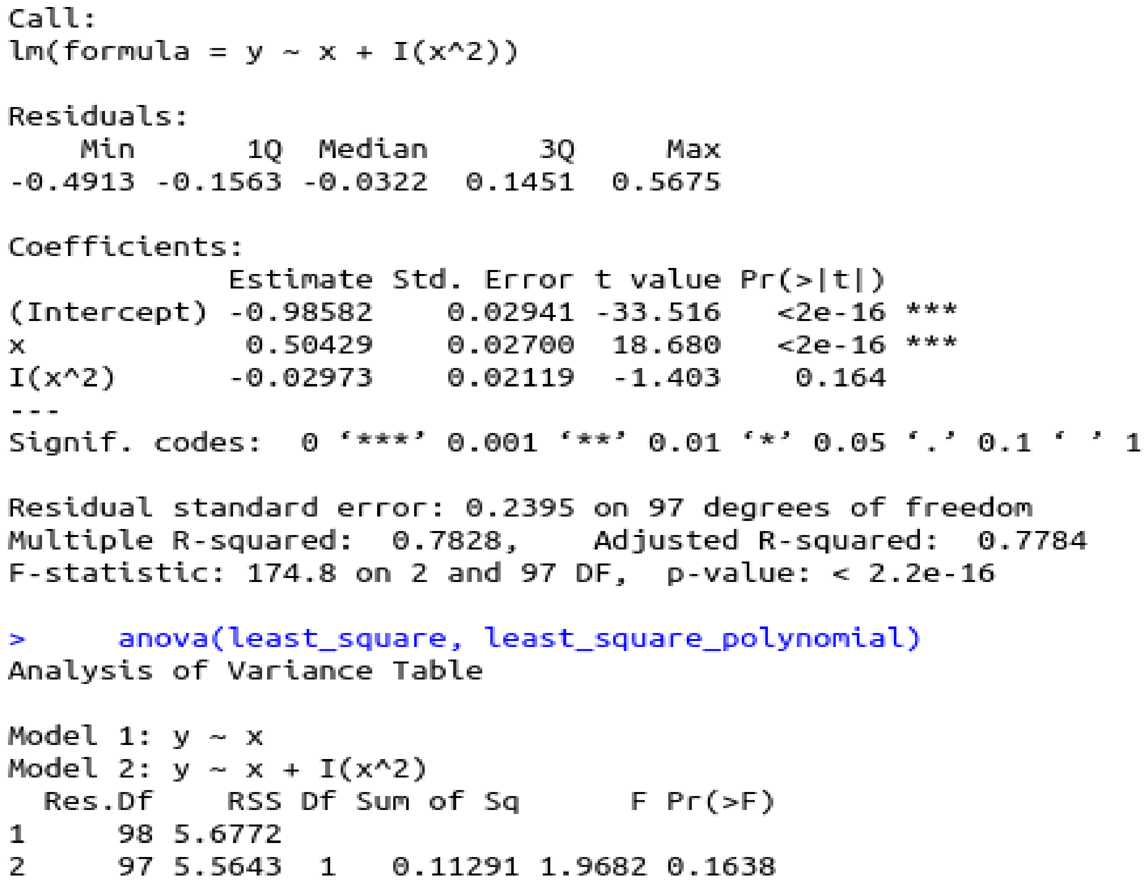


```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.46921 -0.15344 -0.03487  0.13485  0.58654 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.00942    0.02425  -41.63   <2e-16 ***
x            0.49973    0.02693   18.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2407 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762 
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Comparing the β0 and β1 used to generate the linear model with the values obtained from the modeling, we get the same result. The values closely resemble to each other and it is not a surprise.

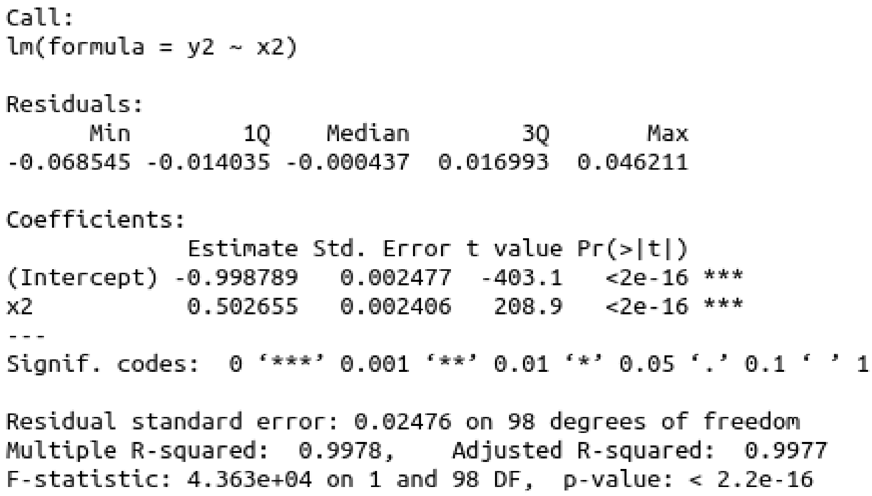
f. Figure 7 contains all the plot required by the question.

g. **Observation**: The significant p-value associated with the quadratic term suggests that the model is not improved. Also, the anova shows that the F-statistic is 1.9682 and associated p-value is not zero. Thus, we can conclude that introducing the quadratic term is not helping to fit the data well.



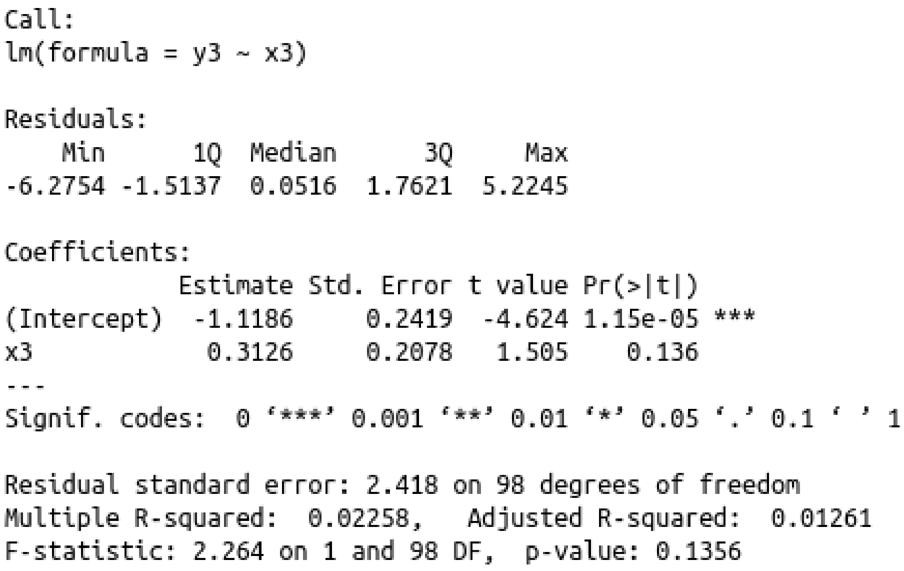
```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4913 -0.1563 -0.0322  0.1451  0.5675 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.98582    0.02941 -33.516   <2e-16 ***
x            0.50429    0.02700  18.680   <2e-16 ***
I(x^2)      -0.02973    0.02119  -1.403    0.164    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2395 on 97 degrees of freedom
Multiple R-squared:  0.7828,    Adjusted R-squared:  0.7784 
F-statistic: 174.8 on 2 and 97 DF,  p-value: < 2.2e-16

>     anova(least_square, least_square_polynomial)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     98 5.6772                           
2     97 5.5643  1   0.11291 1.9682 0.1638
```

h. **Answer**: Adding less noise to the data decreases the number of outliers (Figure 8). This implies that the prediction line appears to be more close to the data. Also, the dependency of the response on the intercept increases drastically which can be interpreted from the p-value. The sum of residuals is too.



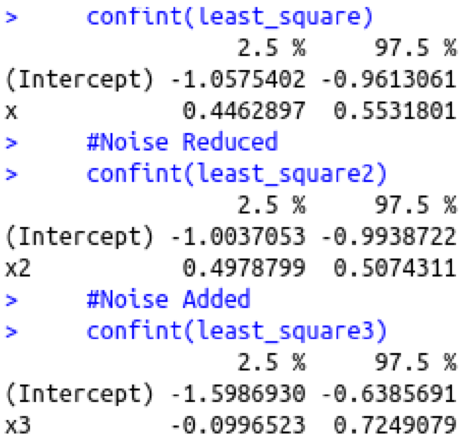
```
Call:
lm(formula = y2 ~ x2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.068545 -0.014035 -0.000437  0.016993  0.046211 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.998789   0.002477  -403.1   <2e-16 ***
x2           0.502655   0.002406   208.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02476 on 98 degrees of freedom
Multiple R-squared:  0.9978,	Adjusted R-squared:  0.9977 
F-statistic: 4.363e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

i. **Answer:** Adding more noise to the data increases the number of outliers (Figure 9). This means that the prediction line appears to be separating the bunch of highly scattered data. Also, the dependency of the response on the intercept decreases drastically which can be interpreted from the p-value.



```
Call:
lm(formula = y3 ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.2754 -1.5137  0.0516  1.7621  5.2245 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.1186     0.2419  -4.624 1.15e-05 ***
x3            0.3126     0.2078   1.505    0.136    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.418 on 98 degrees of freedom
Multiple R-squared:  0.02258,	Adjusted R-squared:  0.01261 
F-statistic: 2.264 on 1 and 98 DF,  p-value: 0.1356
```

j. **Answer:** The confidence interval shows that the range of confidence interval is increased for the predictors as well as the intercepts from the original data to the more noisy data(i). Whereas, the less noisy data(h) displays a drastic decrease in confidence interval. We can conclude that the more noise is added to the data, the more the confidence interval increases.



```
>     confint(least_square)
                 2.5 %     97.5 %
(Intercept) -1.0575402 -0.9613061
x            0.4462897  0.5531801
>     #Noise Reduced
>     confint(least_square2)
                 2.5 %     97.5 %
(Intercept) -1.0037053 -0.9938722
x2           0.4978799  0.5074311
>     #Noise Added
>     confint(least_square3)
                 2.5 %     97.5 %
(Intercept) -1.5986930 -0.6385691
x3          -0.0996523  0.7249079
```