STAT 1361 - Homework 2

Gordon Lu

2/11/2021

Problem 2)

ISLR Conceptual Exercise 3

3a)

Point iii is correct, as for GPAs above 35/10 = 3.5, males will earn more.

3b)

Y(Gender = 1, IQ = 110, GPA = 4.0) = 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01 (4 * 110) - 10 * 4 = 137.1 thousand dollars.

3c)

False, the IQ score scale is larger than the scale of the other predictors (~100 vs. 1-4 for GPA and 0-1 for gender). Even if all the predictors have the same impact on salary, the will always be ultimately smaller than for IQ predictors.

ISLR Conceptual Exercise 4

4a)

I would expect the cubic regression model to have a lower training RSS than the linear regression model, since it could result in a tighter fit against data that matched with a wider irreducible error. Also, typically having more predictors generally means better (lower) RSS on training data, however issues with Collinearity and confounding variables may be introduced into the model.

4b)

I would expect the linear regression model to have a lower test RSS than the cubic regression model, as the potential overfitting that the cubic model introduces would have more error than a simple linear regression model.

4c)

The cubic regression model would produce a lower RSS on the training set, as it can better account and adjust for non-linearities in the data, additionally, it has a higher flexibility than the linear regression model.

4d)

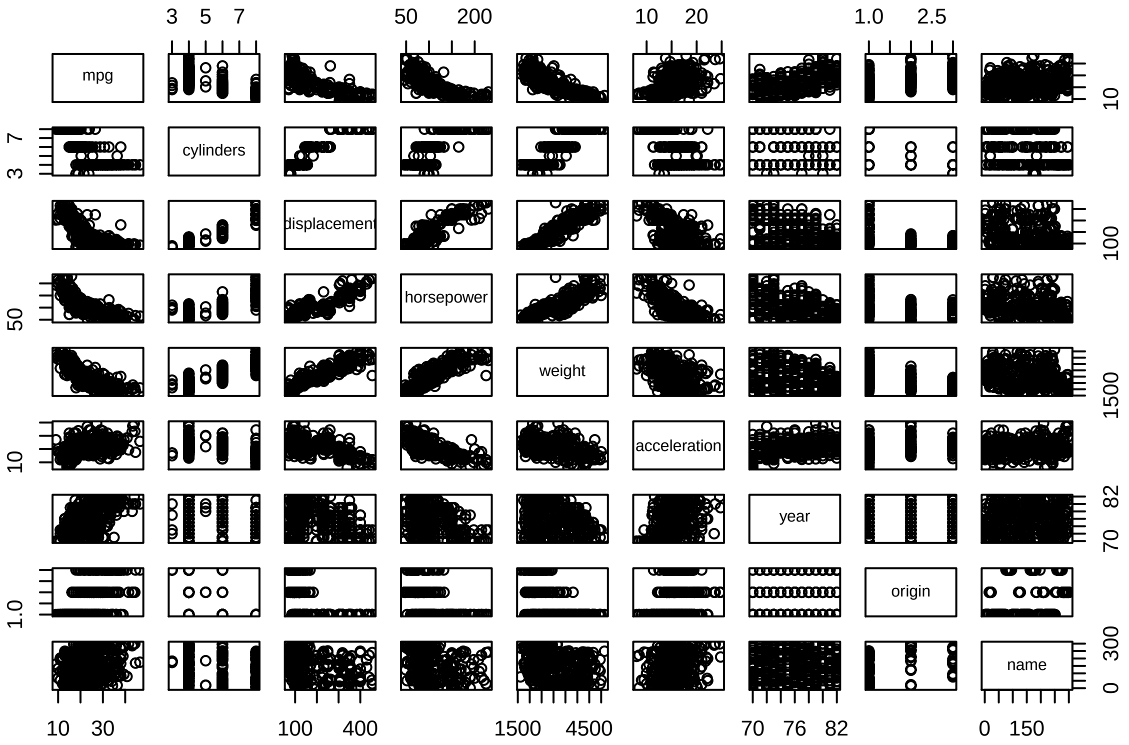
There's not enough information to determine whether test RSS will be lower for linear regression or cubic regression. We simply do not know how "non-linear" the data is. If it's closer to linear than cubic, the linear regression test RSS could be lower than the cubic regression test RSS, and vice versa. This is due to the bias-variance tradeoff: it's not clear what level of flexibility will fit the data better.

Problem 3)

ISLR Applied Exercise 9)

9a)

```r
auto <- read.csv(file = 'C:/Users/gordo/Desktop/Auto.csv')
pairs(auto)
```



9b)

```r
cor(subset(auto, select=-name))
```

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

9c)

```
summary(lm(mpg ~ . -name, data = auto))
```

```
##
## Call:
## lm(formula = mpg ~ . - name, 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
```

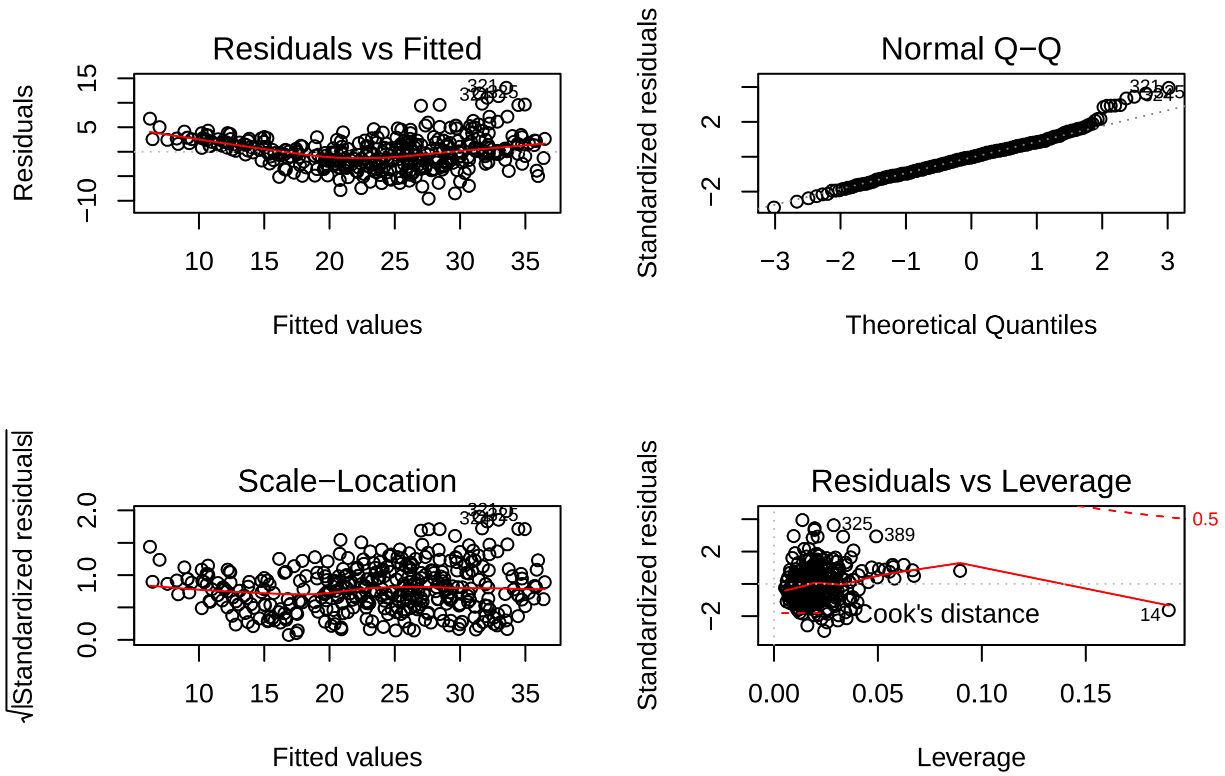
i) Yes, there is a relationship. This is apparent through the test statistics.

ii) It appears that `displacement`, `weight`, `year`, and `origin` all have a statistically significant relationship to the response, `mpg`.

iii) The regression coefficient for year, 0.7508, suggests that for every one year, mpg increases by the coefficient. In other words, cars become more fuel efficient every year by almost 0.75 mpg / year.

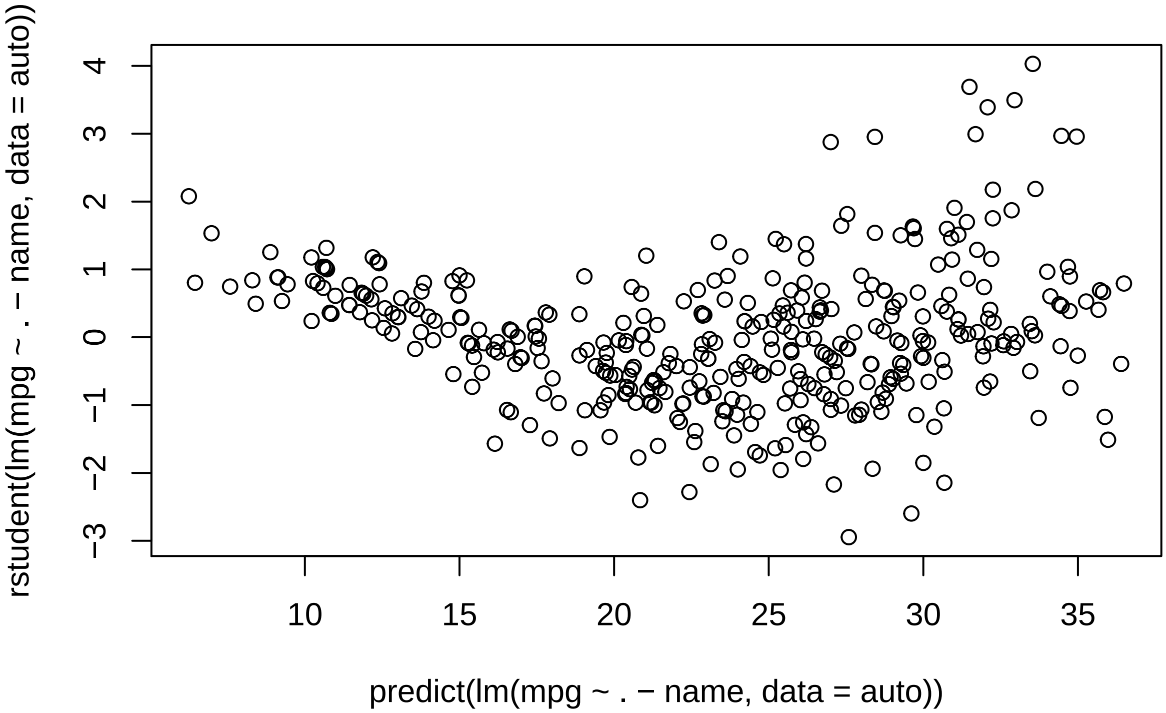
9d)

```r
par(mfrow=c(2,2))
plot(lm(mpg ~ . -name, data = auto))
```



The residual plot seems to indicate a weak fit, likely indicating evidence of non-linearity. From the leverage plot, it appears that observation 14 has high leverage, though not a very large residual to be regarded as an outlier.

```r
plot(predict(lm(mpg ~ . -name, data = auto)), rstudent(lm(mpg ~ . -name, data = auto)))
```



By honing in on the residual plot, it is apparent that points with studentized residuals larger than 3 are potential outliers.

9e)

```
interaction1 <- lm(mpg ~ displacement + weight + year * origin, data = auto)
interaction2 <- lm(mpg ~ displacement+ origin + year* weight, data = auto)
base <- lm(mpg ~ displacement + weight + year + origin, data = auto)
interaction3 <- lm(mpg ~ year + origin + displacement * weight, data = auto)
interaction4 <- lm(mpg ~ cylinders * displacement+displacement * weight, data = auto)

summary(base)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + weight + year + origin, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8102 -2.1129 -0.0388  1.7725 13.2085
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.861e+01  4.028e+00  -4.620 5.25e-06 ***
## displacement  5.588e-03  4.768e-03   1.172    0.242
## weight       -6.575e-03  5.571e-04 -11.802  < 2e-16 ***
## year          7.714e-01  4.981e-02  15.486  < 2e-16 ***
## origin        1.226e+00  2.670e-01   4.593 5.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.346 on 387 degrees of freedom
## Multiple R-squared:  0.8181, Adjusted R-squared:  0.8162
## F-statistic: 435.1 on 4 and 387 DF,  p-value: < 2.2e-16
```

```r
summary(interaction1)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + weight + year * origin, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7541 -1.8722 -0.0936  1.6900 12.4650
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.927e+00  8.873e+00   0.893 0.372229
## displacement  1.551e-03  4.859e-03   0.319 0.749735
## weight       -6.394e-03  5.526e-04 -11.571  < 2e-16 ***
## year          4.313e-01  1.130e-01   3.818 0.000157 ***
## origin       -1.449e+01  4.707e+00  -3.079 0.002225 **
## year:origin   2.023e-01  6.047e-02   3.345 0.000904 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.303 on 386 degrees of freedom
## Multiple R-squared:  0.8232, Adjusted R-squared:  0.8209
## F-statistic: 359.5 on 5 and 386 DF,  p-value: < 2.2e-16
```

```r
summary(interaction2)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + origin + year * weight, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9402 -1.8736 -0.0966  1.5924 12.2125
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.076e+02  1.290e+01  -8.339 1.34e-15 ***
## displacement -4.020e-04  4.558e-03  -0.088 0.929767
## origin        9.116e-01  2.547e-01   3.579 0.000388 ***
## year          1.962e+00  1.716e-01  11.436  < 2e-16 ***
## weight        2.605e-02  4.552e-03   5.722 2.12e-08 ***
## year:weight  -4.305e-04  5.967e-05  -7.214 2.89e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.145 on 386 degrees of freedom
## Multiple R-squared:  0.8397, Adjusted R-squared:  0.8376
## F-statistic: 404.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
summary(interaction3)
```

```
##
## Call:
## lm(formula = mpg ~ year + origin + displacement * weight, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6119  -1.7290  -0.0115   1.5609  12.5584
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -8.007e+00  3.798e+00  -2.108   0.0357 *
## year                 8.194e-01  4.518e-02  18.136  < 2e-16 ***
## origin               3.567e-01  2.574e-01   1.386   0.1666
## displacement        -7.148e-02  9.176e-03  -7.790 6.27e-14 ***
## weight              -1.054e-02  6.530e-04 -16.146  < 2e-16 ***
## displacement:weight  2.104e-05  2.214e-06   9.506  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.016 on 386 degrees of freedom
## Multiple R-squared:  0.8526, Adjusted R-squared:  0.8507
## F-statistic: 446.5 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
summary(interaction4)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders               7.606e-01  7.669e-01   0.992    0.322
## displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

It appears interactions 1-3 all have statistically significant effects. As for interaction 4, after observing the p-values, we can see the interaction between cylinders and displacement is not statistically significant, while the interaction between displacement and weight is statistically significant.

9f)

```
nonlineart1 <- lm(mpg ~ poly(displacement,3) + weight + year + origin, data = auto)
nonlineart2 <- lm(mpg ~ displacement + I(log(weight)) + year + origin, data = auto)
nonlineart3 <- lm(mpg ~ displacement + I(weight^2) + year + origin, data = auto)
summary(nonlineart1)
```

```
##
## Call:
## lm(formula = mpg ~ poly(displacement, 3) + weight + year + origin,
##     data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8131  -1.8012   0.0788   1.5566  12.3181
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2.342e+01  3.802e+00  -6.160 1.84e-09 ***
## poly(displacement, 3)1 -1.701e+01  9.820e+00  -1.732   0.0840 .
## poly(displacement, 3)2  2.840e+01  3.610e+00   7.866 3.74e-14 ***
## poly(displacement, 3)3 -7.996e+00  3.164e+00  -2.527   0.0119 *
## weight                 -5.285e-03  5.419e-04  -9.753  < 2e-16 ***
## year                    8.189e-01  4.660e-02  17.572  < 2e-16 ***
## origin                  2.422e-01  2.761e-01   0.877   0.3810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.102 on 385 degrees of freedom
## Multiple R-squared:  0.8445, Adjusted R-squared:  0.842
## F-statistic: 348.4 on 6 and 385 DF,  p-value: < 2.2e-16
```

```
summary(nonlineart2)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + I(log(weight)) + year + origin,
##     data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7136 -1.9214  0.0447  1.5790 12.9864
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    131.274483  11.082986  11.845  < 2e-16 ***
## displacement     0.007711   0.004052   1.903 0.057810 .
## I(log(weight)) -21.584745   1.451851 -14.867  < 2e-16 ***
## year             0.804835   0.046532  17.296  < 2e-16 ***
## origin           0.836143   0.250485   3.338 0.000925 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.113 on 387 degrees of freedom
## Multiple R-squared:  0.8425, Adjusted R-squared:  0.8409
## F-statistic: 517.7 on 4 and 387 DF,  p-value: < 2.2e-16
```

```r
summary(nonlineart3)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + I(weight^2) + year + origin,
##     data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0988  -2.2549  -0.1057   1.8704  13.4702
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.609e+01  4.349e+00  -5.999 4.56e-09 ***
## displacement -9.114e-03  5.118e-03  -1.781   0.0757 .
## I(weight^2)  -7.068e-07  9.075e-08  -7.789 6.28e-14 ***
## year          7.336e-01  5.380e-02  13.635  < 2e-16 ***
## origin        1.488e+00  2.900e-01   5.132 4.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.628 on 387 degrees of freedom
## Multiple R-squared:  0.7861, Adjusted R-squared:  0.7839
## F-statistic: 355.7 on 4 and 387 DF,  p-value: < 2.2e-16
```

Along all of the tested models, `displacement^2` has a larger effect than any other `displacement` polynomials tested.

ISLR Applied Exercise 10)

10a)

```r
carseats <- read.csv(file = 'C:/Users/gordo/Desktop/Carseats.csv')
clm <- lm(Sales ~ Price + Urban + US, data = carseats)
```

10b)

```r
summary(clm)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.9206 -1.6220 -0.0564  1.5786  7.0581 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936    
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335 
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

Price: The result of the multiple regression model implies that there is a significant relationship between Price and Sales. The coefficient implies a negative relationship between Price and Sales. As Price increases, Sales decrease, vice versa.

UrbanYes: The result of the multiple regression model implies that there is not s significant relationship between Sales and Urban. However the coefficient from the summary tells us that Sales are 2.2% lower for Urban locations.

USYes: The result of the multiple regression model implies that there is a significant relationship between US and Sales. The coefficient implies a positive relationship between US and Sales. As more people are in the US, Sales increase, vice versa.

10c)

Sales = 13.043 - 0.054 x Price - 0.022 x UrbanYes + 1.201 x USYes

10d)

We can reject the null hypothesis for Price and USYes (coefficients have low p-values < 0.01).

10e)

```r
better_fit <- lm(Sales ~ Price + US, data = carseats)
summary(better_fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

10f)

Based on the RSE and R^2 of the linear regressions, they both fit the data similarly, with linear regression from (e) fitting the data slightly better.

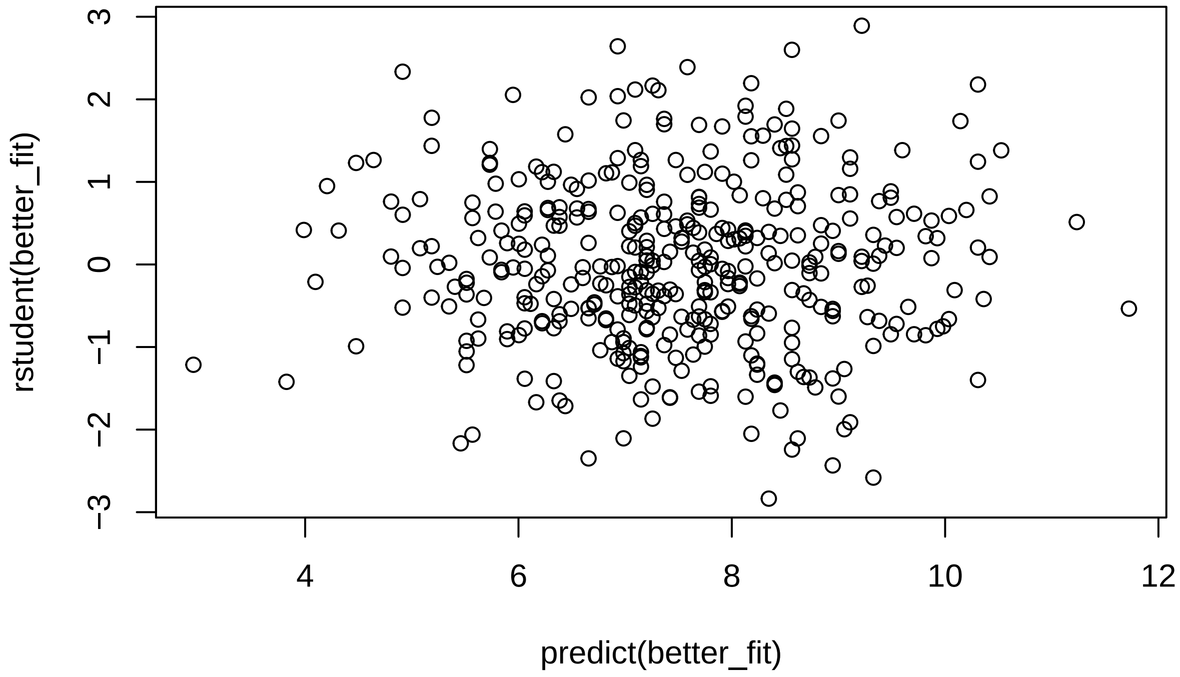
10g)

```r
confint(better_fit)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

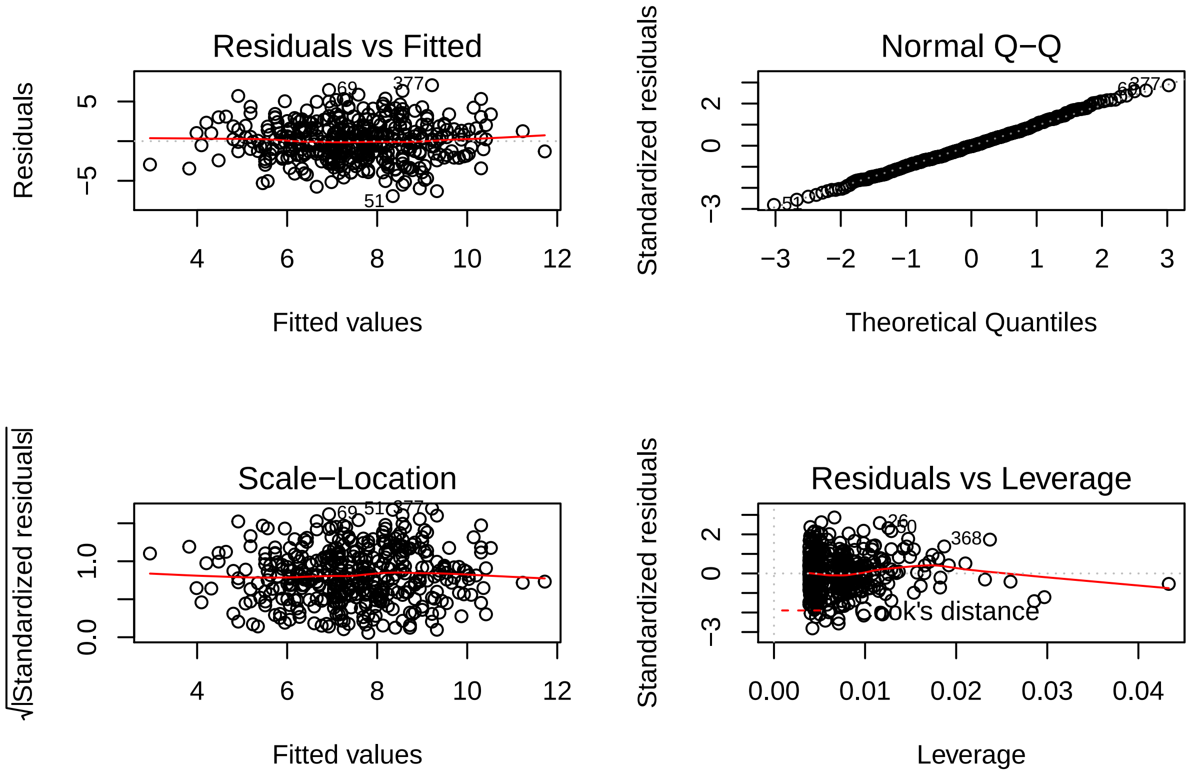
10h)

```r
plot(predict(better_fit), rstudent(better_fit))
```



The studentized residuals all seem to be between |3|, so there does not appear to be any apparent outliers.

```r
par(mfrow=c(2,2))
plot(better_fit)
```



There are a few observations that greatly exceed (p+1)/n (0.0076) on the leverage plot that suggest that the corresponding points have high leverage.

ISLR Applied Exercise 13)

13a)

```r
set.seed(1)
x <- rnorm(100, 0, 1)
```

13b)

```r
eps <- rnorm(100, 0, 0.25)
```

13c)

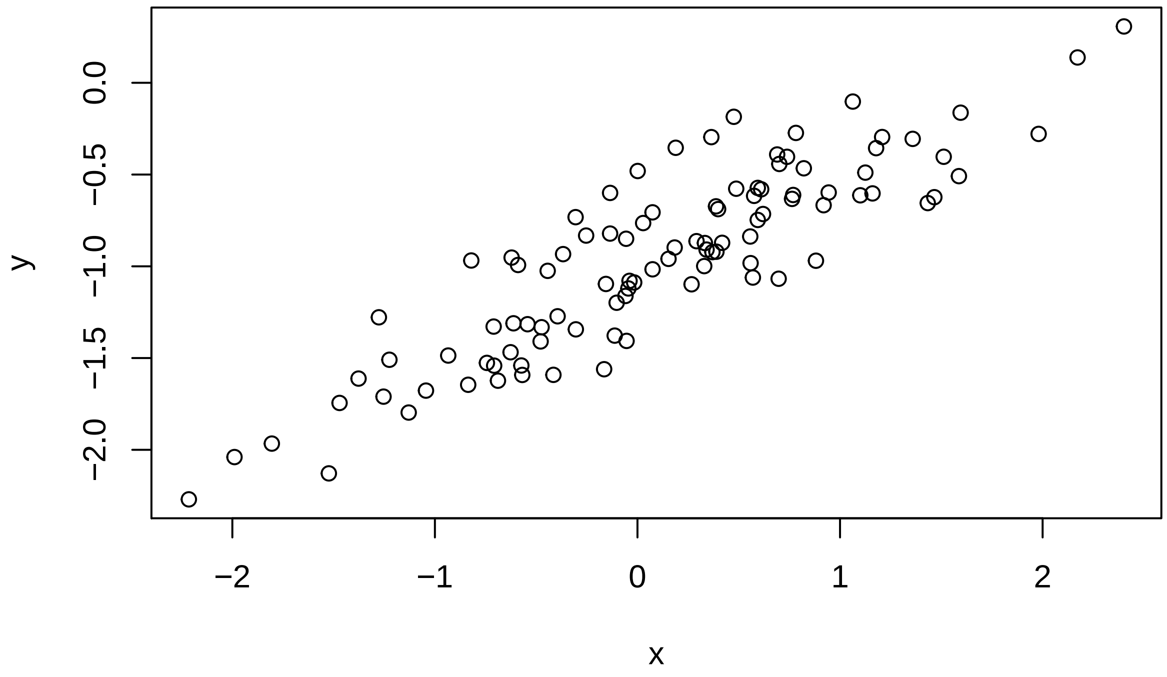
```r
y <- -1 + (0.5*x) + eps
length(y)
```

```
## [1] 100
```

The length of y is 100, β_0 is -1 and β_1 is 0.5

13d)

```
plot(x,y)
```



There is a strong, positive linear relationship between y and x, as to be expected.

13e)

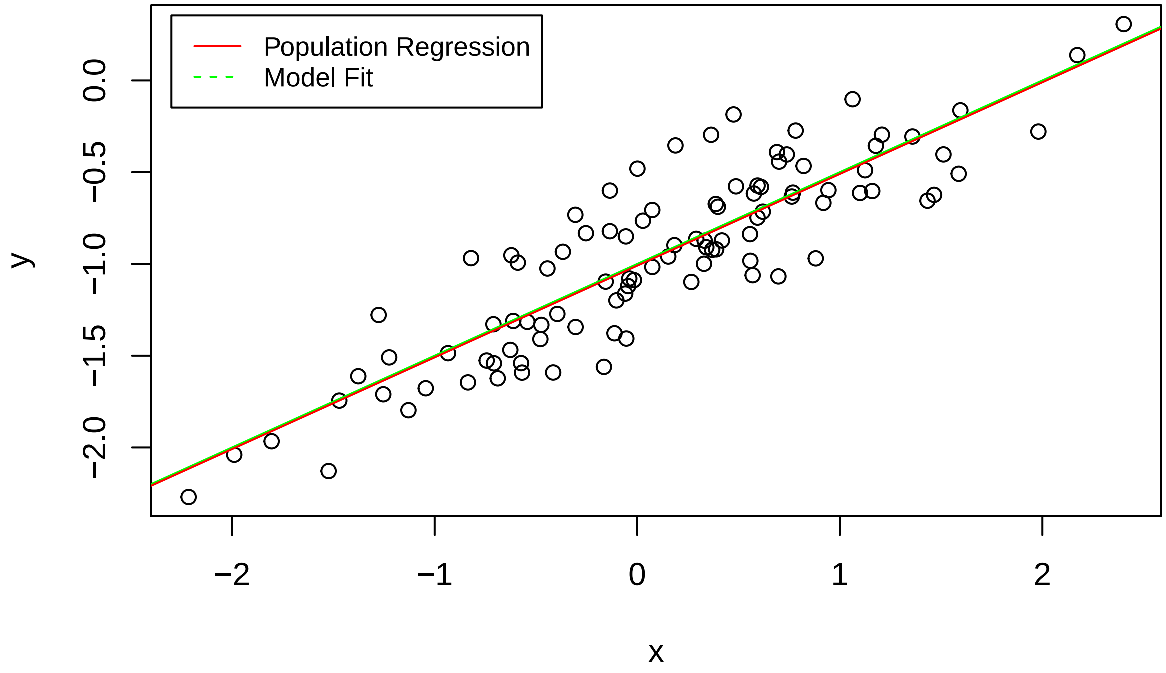
```
fit_lm <- lm(y ~ x)
summary(fit_lm)
```

```
##
## 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
```

The linear regression fits a model that is very close to the true values of the coefficents that was constructed in 13c). Additionally, on running F-test for overall significance, we would reject the null, and conclude that the predictor, x does explain y fairly well.

13f)

```
plot(x,y)
abline(-1, 0.5, col="green")  # ground truth model
abline(fit_lm, col="red")     # regression fitted on data
legend(x= "topleft", inset = 0.02, legend=c("Population Regression", "Model Fit"),
       col=c("red", "green"), lty=1:2, cex=0.8)
```



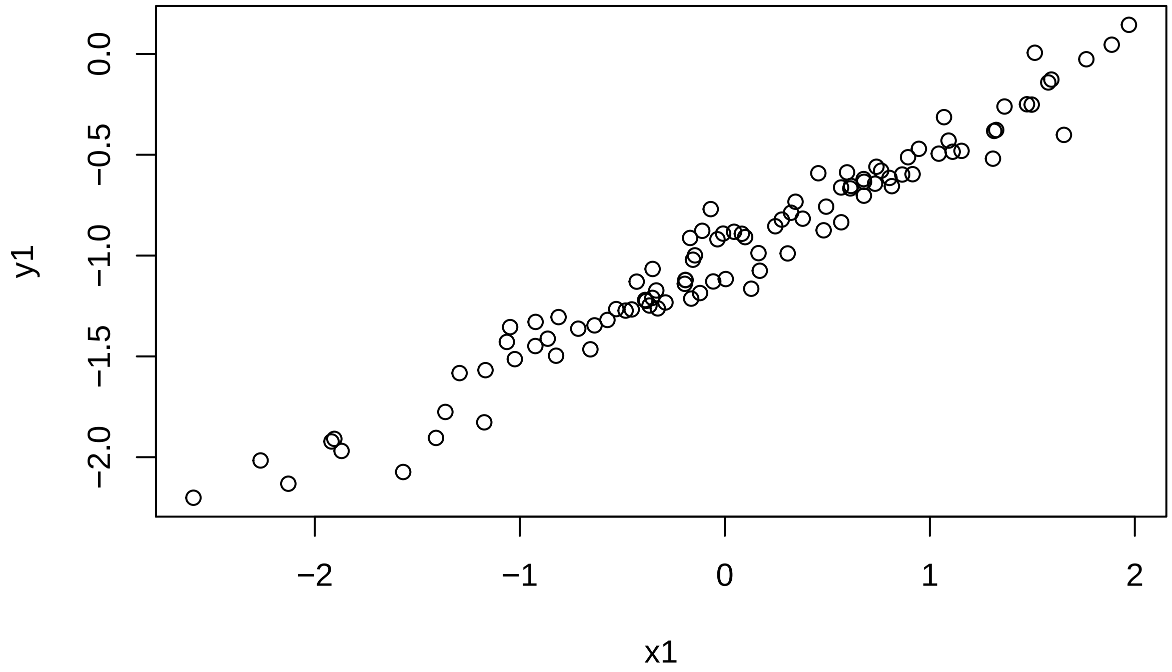
```r
lm_quad <- lm(y ~ x + I(x^2))
summary(lm_quad)
```

```
##
## 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
```

Although the R^2 and RSE have slightly increased over the training data, the large F-statistic implies that the addition of the quadratic term has resulted in the significance of the new model to be significantly related to y.

13h)

```r
smaller_eps = rnorm(100, 0, 0.10)
x1 = rnorm(100)
y1 = -1 + 0.5*x1 + smaller_eps
plot(x1, y1)
```

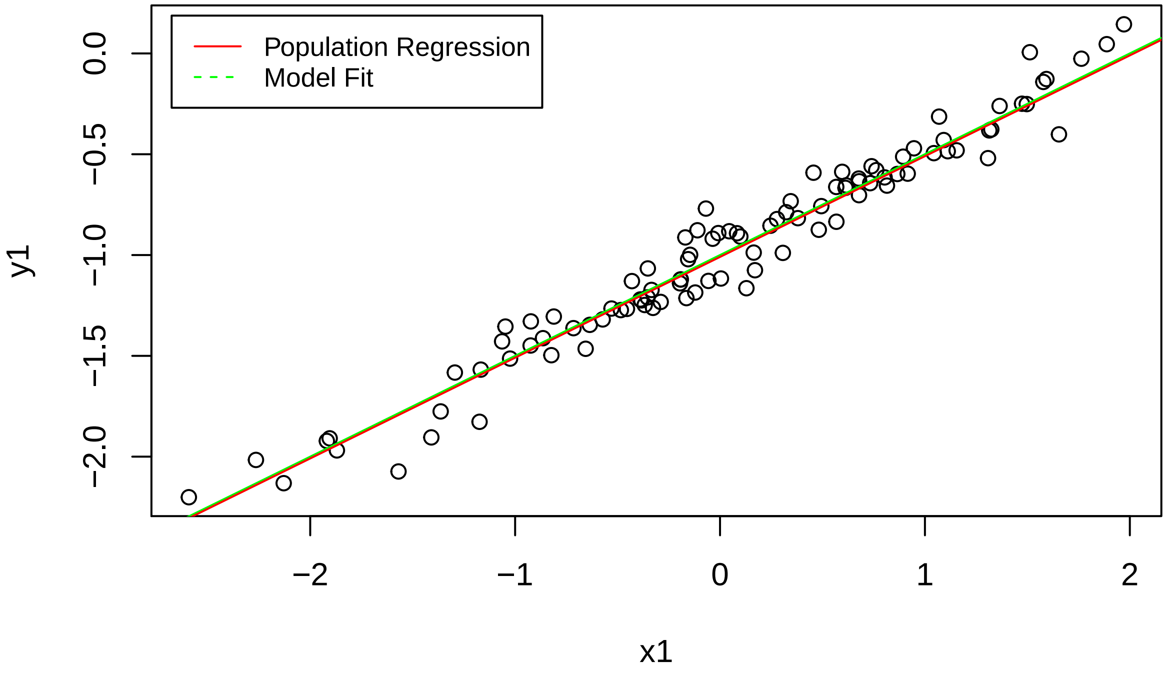


```
small_eps_fit = lm(y1~x1)
summary(small_eps_fit)
```

```
##
## Call:
## lm(formula = y1 ~ x1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.273134 -0.056528  0.002025  0.063101  0.263341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.99763    0.01035  -96.43   <2e-16 ***
## x1           0.51155    0.01047   48.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1033 on 98 degrees of freedom
## Multiple R-squared:  0.9606, Adjusted R-squared:  0.9602
## F-statistic:  2387 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x1,y1)
abline(-1, 0.5, col="green")  # ground truth model
abline(fit_lm, col="red")     # regression fitted on data
```

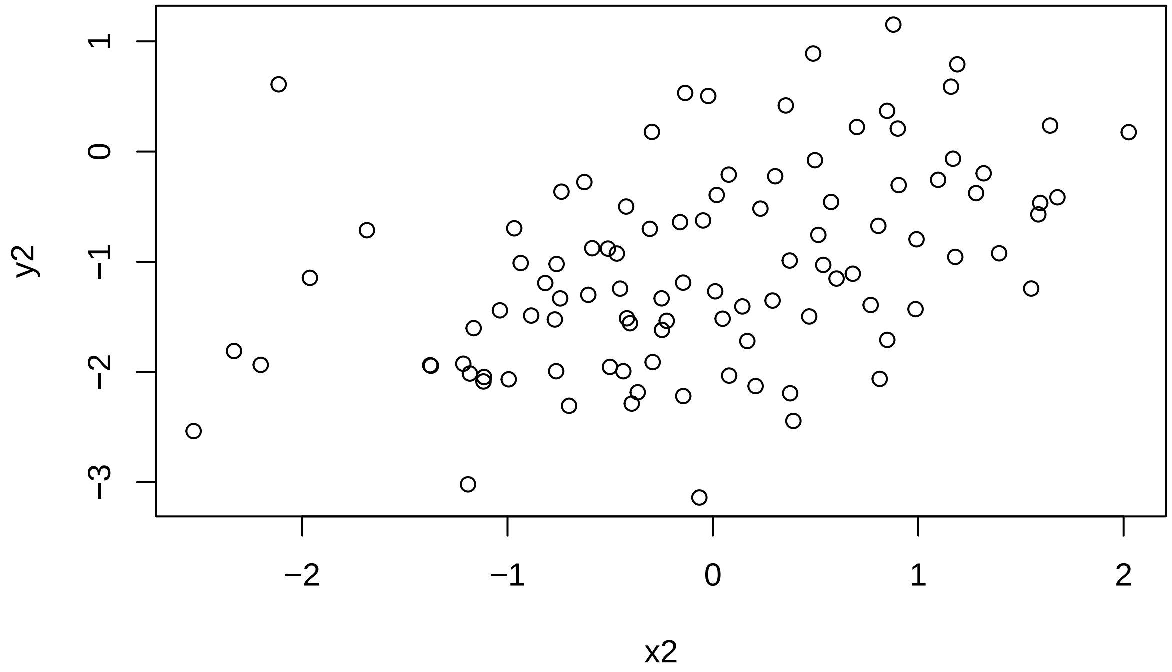
```
legend(x= "topleft", inset = 0.02, legend=c("Population Regression", "Model Fit"),
       col=c("red", "green"), lty=1:2, cex=0.8)
```



The error observed in R^2 and RSE decrease significantly. Additionally, the points are more tightly clustered around the regression lines.

13i)

```
larger_eps = rnorm(100, 0, 0.7)
x2 = rnorm(100)
y2 = -1 + 0.5*x2 + larger_eps
plot(x2, y2)
```

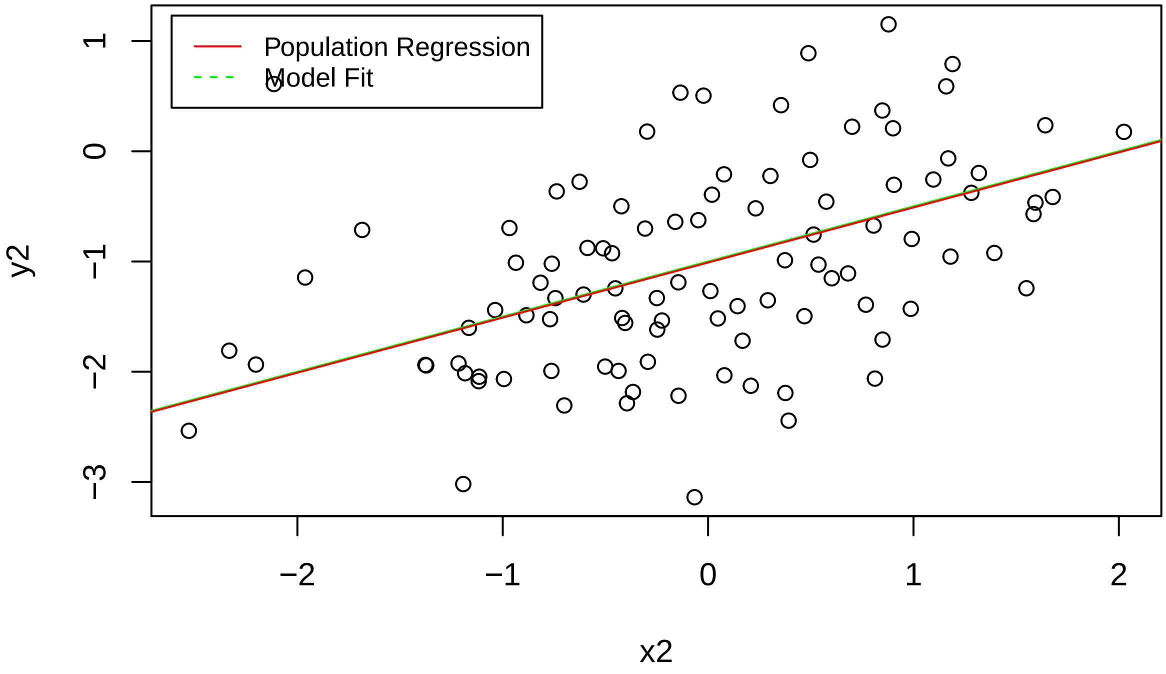


```
large_eps_fit = lm(y2~x2)
summary(large_eps_fit)
```

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07988 -0.48826 -0.08821  0.45852  2.53545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.03082    0.08204 -12.565  < 2e-16 ***
## x2           0.42312    0.08525   4.963 2.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8195 on 98 degrees of freedom
## Multiple R-squared:  0.2009, Adjusted R-squared:  0.1927
## F-statistic: 24.64 on 1 and 98 DF,  p-value: 2.922e-06
```

```
plot(x2,y2)
abline(-1, 0.5, col="green")  # ground truth model
abline(fit_lm, col="red")    # regression fitted on data
```

```
legend(x= "topleft", inset = 0.02, legend=c("Population Regression", "Model Fit"),
       col=c("red", "green"), lty=1:2, cex=0.8)
```



The error observed in Rˆ2 and RSE increase significantly. Additionally, the points are more scattered, and distant from the regression lines.

13j)

```
confint(fit_lm)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0575402 -0.9613061
## x            0.4462897  0.5531801
```

```
confint(small_eps_fit)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0181590 -0.9770987
## x1           0.4907754  0.5323322
```

```
confint(large_eps_fit)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1936184 -0.8680151
## x2           0.2539492  0.5922848
```

All intervals seem to be centered on approximately 0.5, with the model with the smaller epsilon's interval being narrower than the orifinal model's interval and the model with the larger epsilon's interval being wider than the original model's interval. Generally, we can say that confidence intervals will be tighter for populations with smaller variances.

ISLR Applied Exercise 14)

14a)

```
set.seed(1)
x1 = runif(100)
x2 = 0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

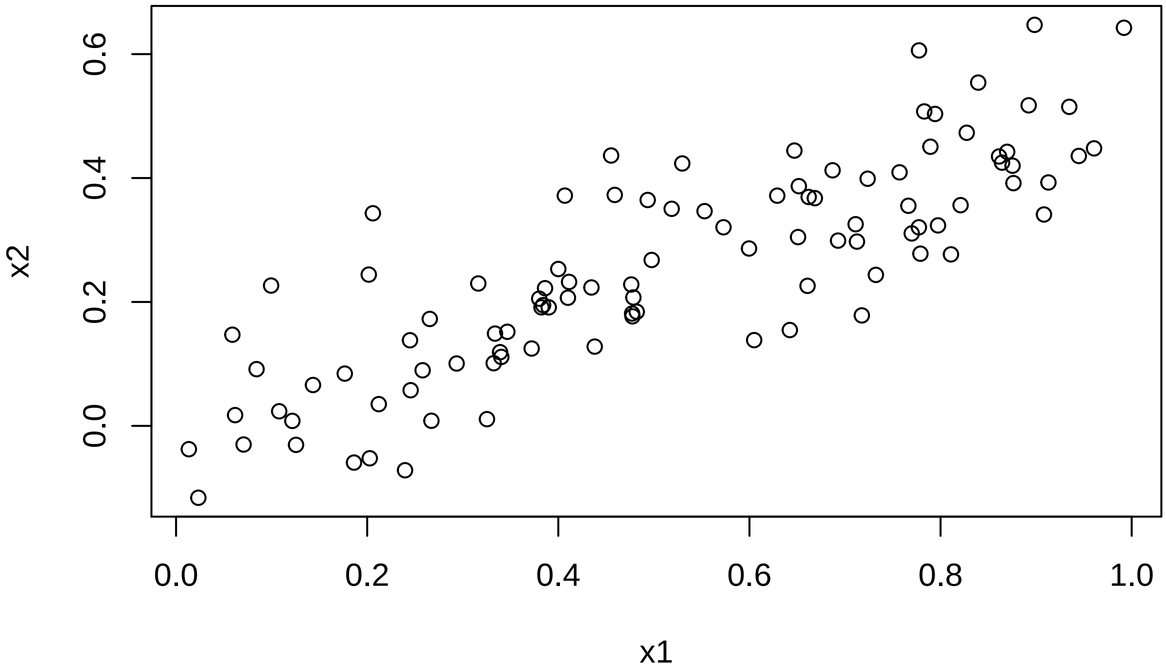
Population regression is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where $\beta_0 = 2$, $\beta_1 = 2$ and $\beta_2 = 0.3$

14b)

```
cor(x1,x2)
```

```
## [1] 0.8351212
```

```
plot(x1,x2)
```



x1 and x2 have a strong, positive linear relationship.

14c)

```
lm_ex14 <- lm(y ~ x1 + x2)
summary(lm_ex14)
```

```
##
## 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
```

$\beta_0 = 2.1305$, $\beta_1 = 1.4396$, $beta_2 = 1.0097$

The coefficient for x1 is statistically significant, but the coefficient for x2 is not statistically significant given that x1 is already in the model. These betas try to estimate the population betas: $\hat{\beta}_0$ is close (rounds to 2), $\hat{\beta}_1$ is 1.44 instead of 2 with a high standard error and $\hat{\beta}_2$ is farthest off.

14d)

```
fit_x1 <- lm(y ~ x1)
summary(fit_x1)
```

```
##
## 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
```

We can reject the null hypothesis, and conclude that x1 has a significant relationship with y.

14e)

```
fit_x2 <- lm(y ~ x2)
summary(fit_x2)
```

```
##
## 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
```

We can reject the null hypothesis, and conclude that x2 has a significant relationship with y.

14f)

No. Without the presence of other predictors, both β_1 and β_2 are statistically significant. In the presence of other predictors, β_2 is no longer statistically significant. This may imply that x1 may explain something that x2 already does, adding x2 may have introduced collinearity, or perhaps x2 is a confounding variable.

14g)

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y,6)

fit_lm_new <- lm(y ~ x1 + x2)
fit_x1_new <- lm(y ~ x1)
fit_x2_new <- lm(y ~ x2)
summary(fit_lm_new)
```

```
## 
## 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
```

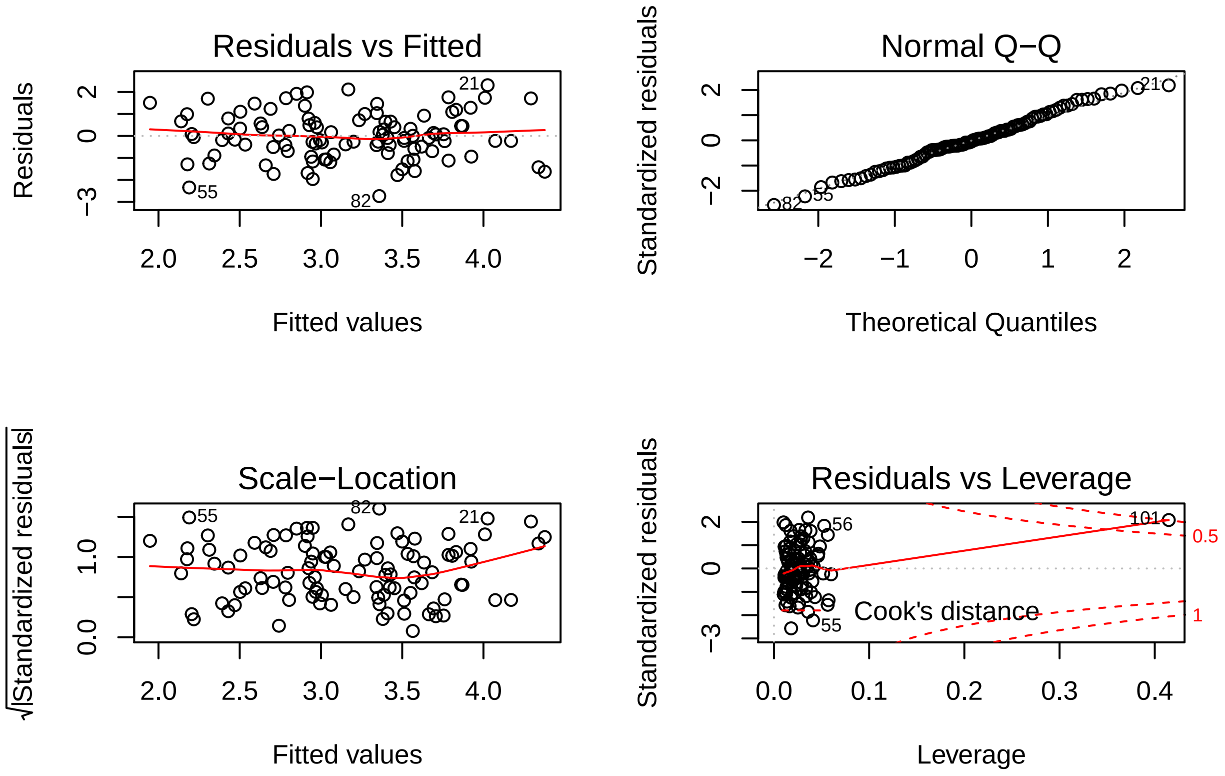
```r
summary(fit_x1_new)
```

```
##
## 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: 4.295e-05
```

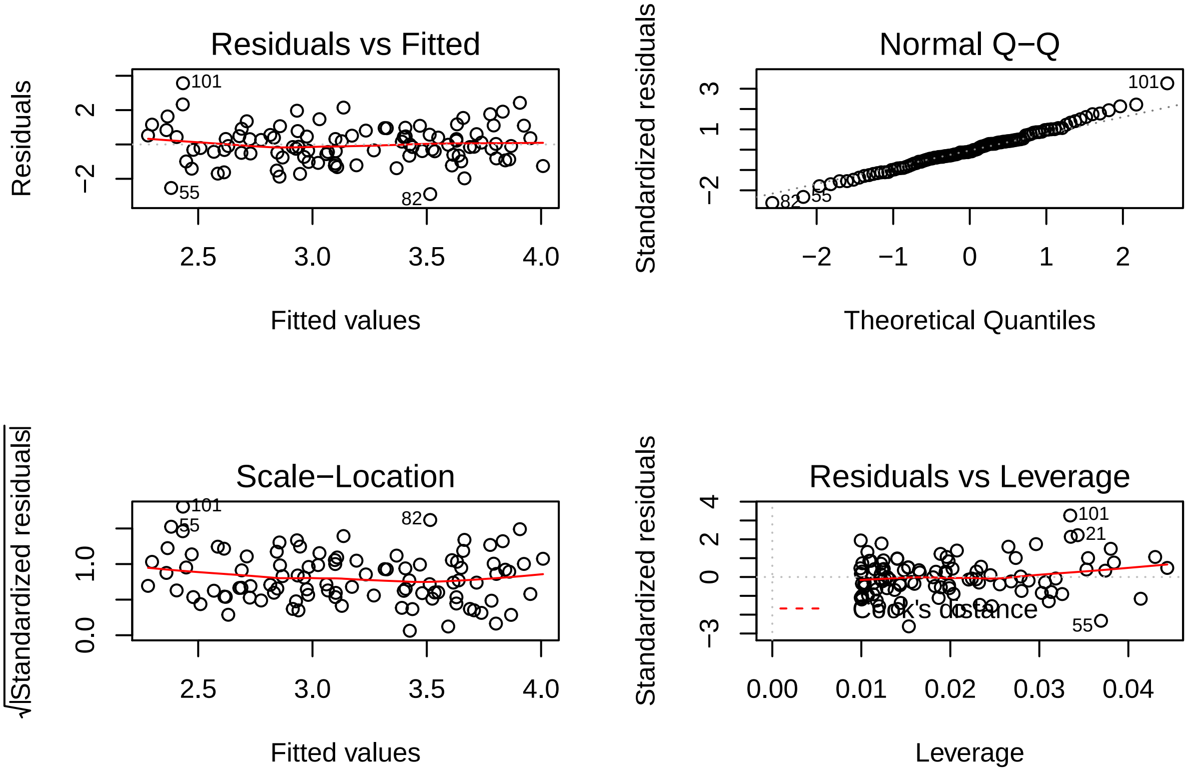
```r
summary(fit_x2_new)
```

```
##
## 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
```

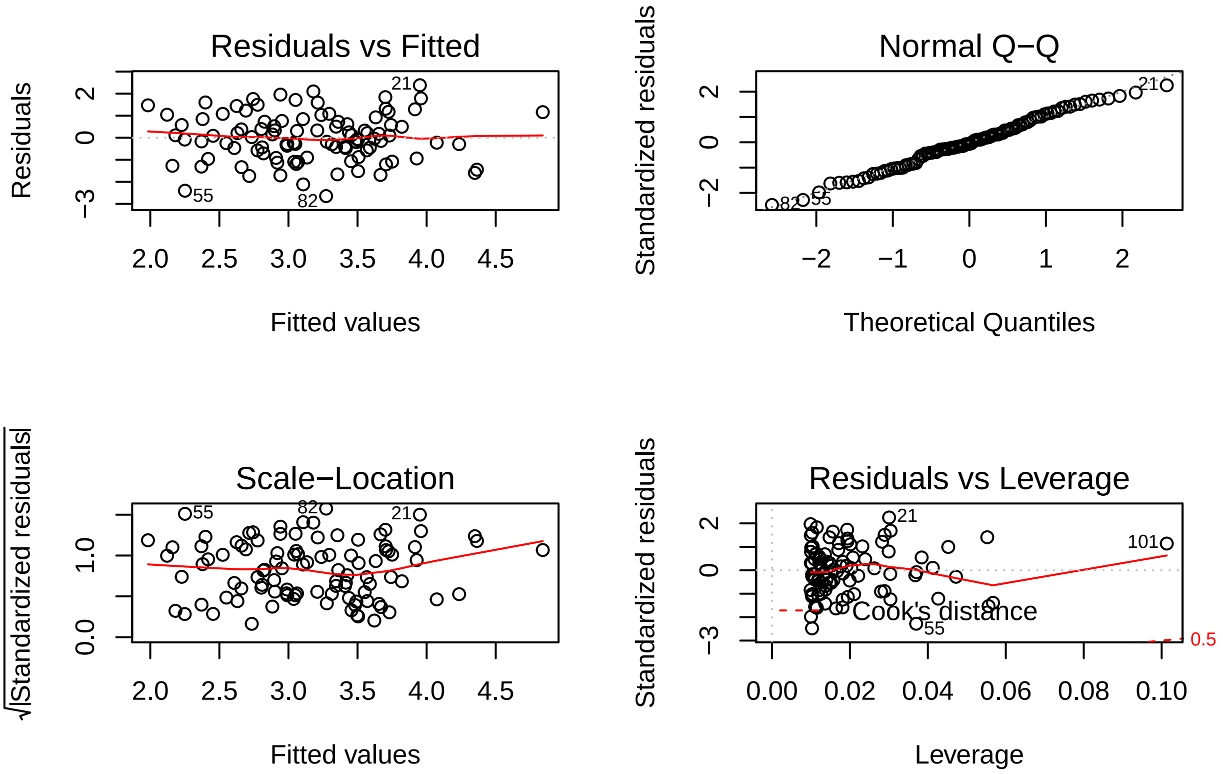
```r
par(mfrow=c(2,2))
plot(fit_lm_new)
```



```r
plot(fit_x1_new)
```



```
plot(fit_x2_new)
```



In the first model, the new observation shifts `x1` to be statistically insignificant and shifts `x2` to be statistiscal significant from the change in p-values between the two linear regressions.

The new point is an outlier for x2 and has high leverage for both x1 and x2.

Looking at the regression with `x1` and `x2`, the residual vs. leverage plot shows that observation 101 is standing out.

Looking at the regression with `x1` only, the new point has high leverage, but doesn't cause issues, since the new point is not an outlier for `x1` or `y`.

Looking at the regression with `x2` only, the new point has high leverage, but doesn't cause issues, since it falls close to the regression line.

Problem 4)

4a)

```
set.seed(2)
df.train <- data.frame(matrix(rnorm(25*25,0,1),ncol = 25))
names(df.train)[25] <- "y"
```

4b)

```
df.test <- data.frame(matrix(rnorm(25*25,0,1),ncol = 25))
names(df.test)[25] <- "y"
```

4c)

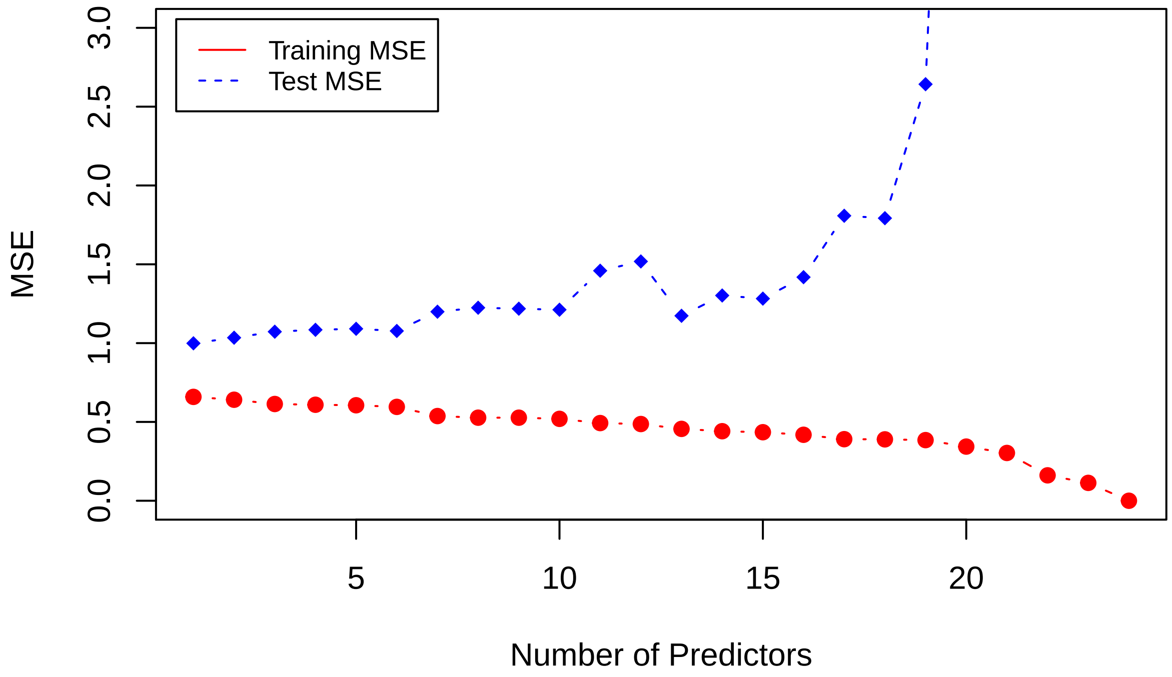
```
mse.train <- rep(0,24)
mse.test <- rep(0,24)
for(i in 1:24){

  mod <- lm(y~.,data = df.train[,c(1:i,25)])
  mse.test[i] <- mean((predict(mod,newdata = df.test[,c(1:i,25)])-df.test[,"y"])^2)
  mse.train[i] <- mean((predict(mod)-df.train[,"y"])^2)

}
```

4d)

```
plot(c(1:24), mse.train, type="b", pch=19, col="red", xlab="Number of Predictors",
     ylab="MSE", ylim = range(0, 3))
# Add a line
lines(c(1:24), mse.test, pch=18, col="blue", type="b", lty=2)
# Add a legend
legend(x= "topleft", inset = 0.02, legend=c("Training MSE", "Test MSE"),
       col=c("red", "blue"), lty=1:2, cex=0.8)
```



4e)

What happens to the training error as more predictors are added to the model? What about the test error?

The training error appears to decrease, while the test error appears to increase.