hw1

Ken Ye

2023-09-17

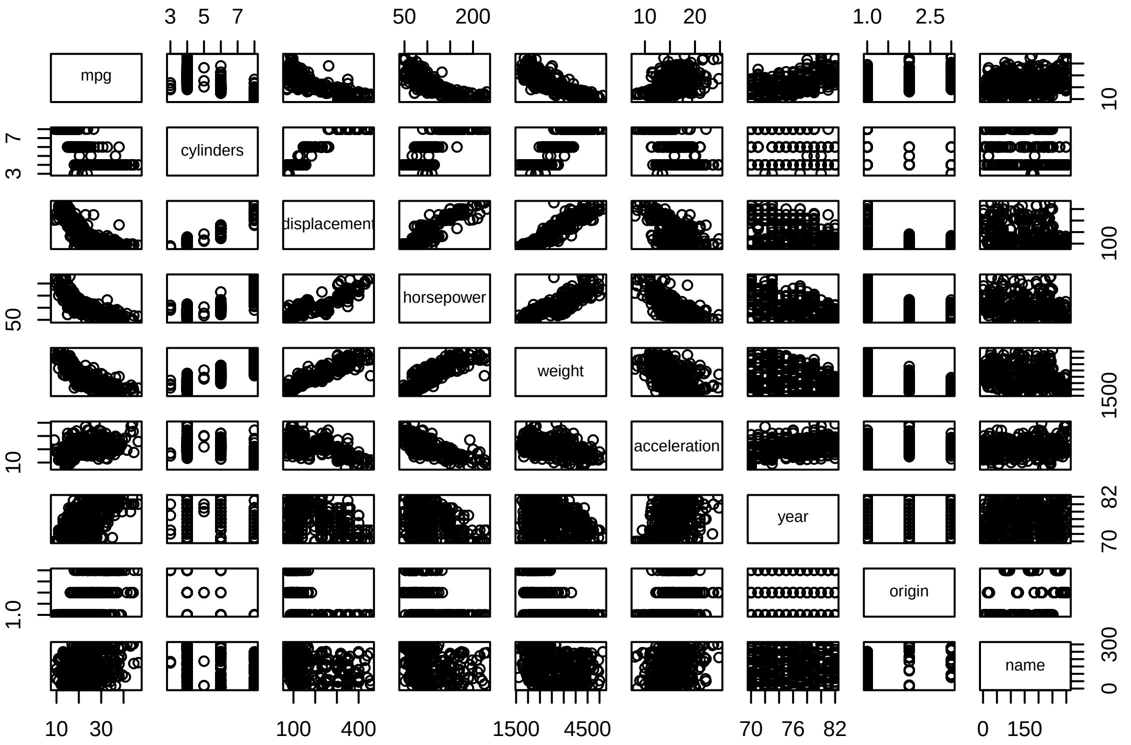
Question 5

This question involves the use of multiple linear regression on the Auto data set.

```r
library(ISLR)
data(Auto)
attach(Auto)
```

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

```r
pairs(Auto)
```



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

```r
cor(Auto[,-9])
```

```
##                     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
```

(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results.

```r
mlr <- lm(mpg ~ . -name, data = Auto)
summary(mlr)
```

```
##
## 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
```

Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

There does seem to be a relationship between the predictors and the response, indicated by the large F-statistic and the near-zero p-value. In addition, the adjusted R-squared is 0.8182, which is quite high and indicates that 81.82% of the model variability is explained by the predictors.

ii. Which predictors appear to have a statistically significant relationship to the response?

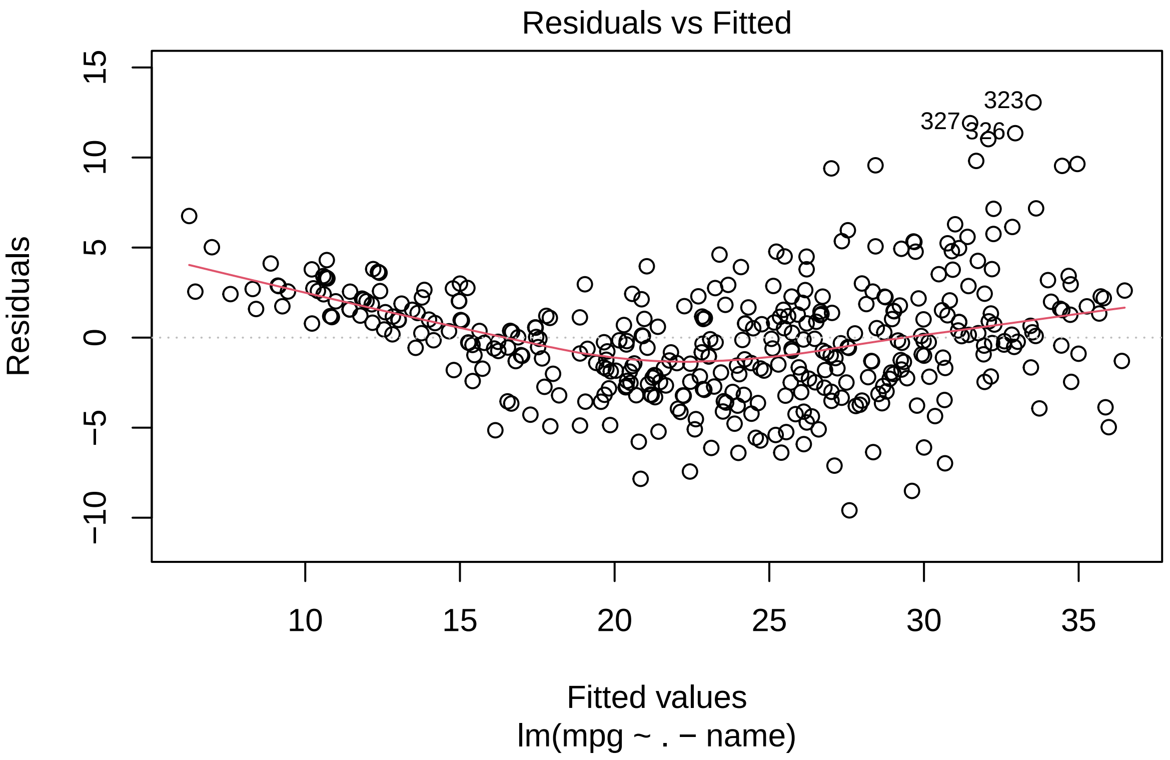
These predictors appear to have a statistically significant relationship to the response as their p-values are all < 0.05: displacement, weight, year, and origin.

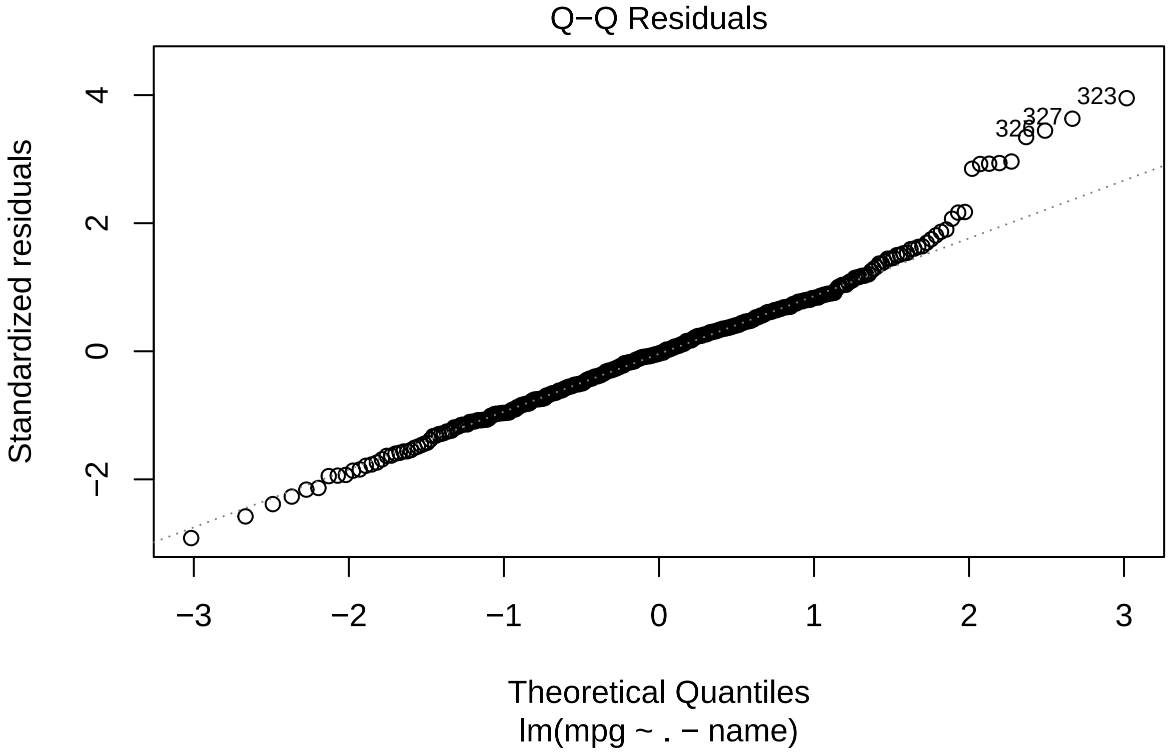
iii. What does the coefficient for the year variable suggest?

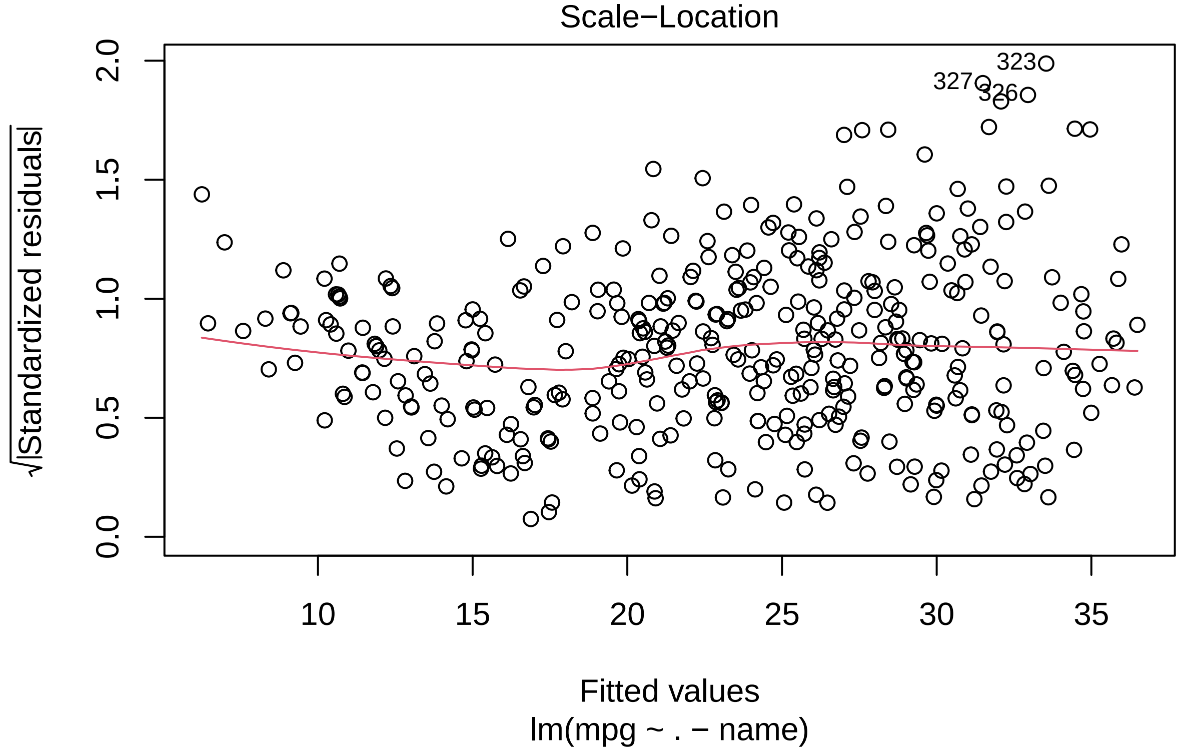
For each unit increase in year, the mpg is expected to increase by 0.75 units, holding all other variables constant.

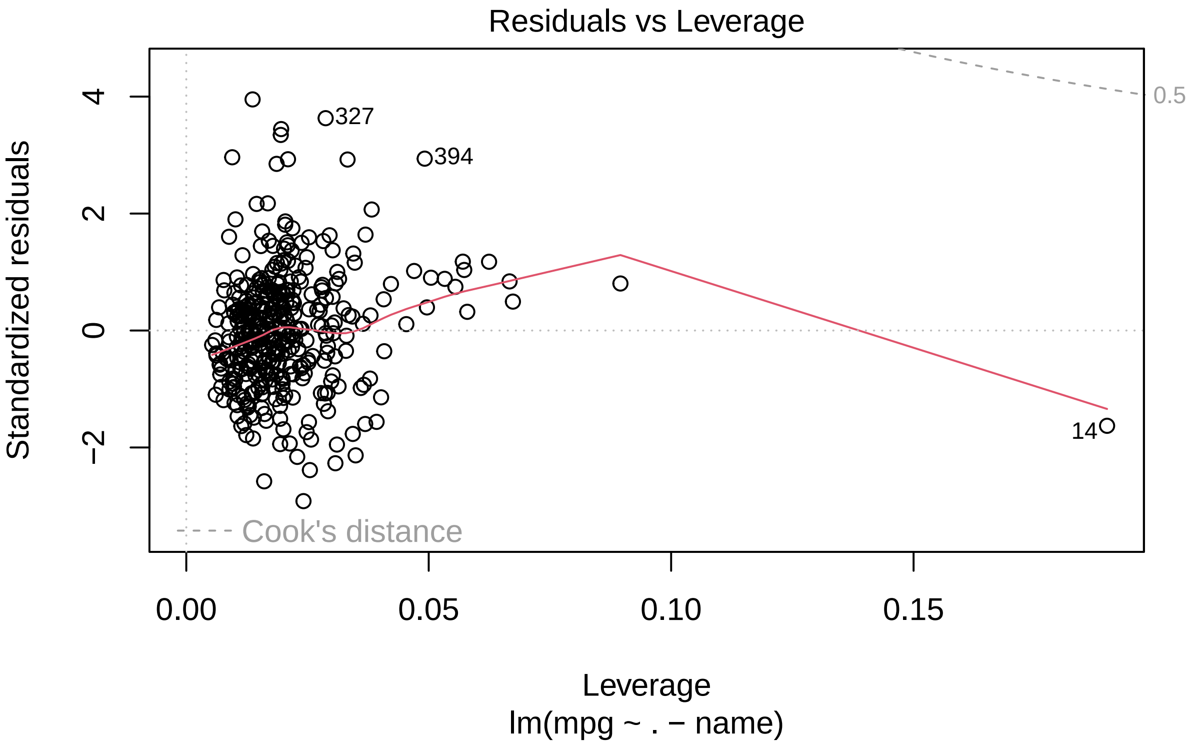
(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```r
plot(mlr)
```




Q−Q Residuals
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
323
327
326
Theoretical Quantiles
lm(mpg ~ . − name)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
10
15
20
25
30
35
323
327
326
Fitted values
lm(mpg ~ . − name)



Looking at the "Residuals vs Fitted" as well as the "Scale-Location" graphs, the residuals seem to be slightly cone-shaped (larger residuals with larger fitted values), indicating possible heteroscedasticity. R highlights observations 323, 326, and 327 as potential outliers as they have unusually high residual, which requires further investigation.

The "Q-Q Residuals" plot validates the normality assumption as the residual generally follows a normal distribution (thought right skewed at the tail) except for very large observations such as 323, 326, and 327, which are potential outliers.

In the "Residuals vs Leverage" plot, R identifies observations 14, 327, and 394 as influential points. Among them, observation 14 has the highest leverage (around 0.2).

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
mlr2 <- lm(mpg ~ (. - name)^2, data = Auto)
summary(mlr2)
```

```
##
## Call:
## lm(formula = mpg ~ (. - name)^2, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.548e+01  5.314e+01   0.668  0.50475
## cylinders                   6.989e+00  8.248e+00   0.847  0.39738
## displacement               -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                  5.034e-01  3.470e-01   1.451  0.14769
## weight                     4.133e-03  1.759e-02   0.235  0.81442
## acceleration              -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                       6.974e-01  6.097e-01   1.144  0.25340
## origin                    -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders:displacement    -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower       1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight           3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration     2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year            -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin           4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower   -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight        2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year          5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin        2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight         -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year           -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin          2.233e-03  2.930e-02   0.076  0.93931
## weight:acceleration        2.346e-04  2.289e-04   1.025  0.30596
## weight:year               -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin             -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year          5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin        4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

In this model including all main effects and all two-way interactions, these interaction terms are statistically significant (p-value < 0.05): displacement:year, acceleration:year, and acceleration:origin.

(f) Try a few different transformations of the variables, such as log(X), sqrt(X), X^2. Comment on your findings.

```r
# log(weight)
mlr3 <- lm(mpg ~ . - name - weight + log(weight), data = Auto)
summary(mlr3)
```

```
##
## Call:
## lm(formula = mpg ~ . - name - weight + log(weight), data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.382 -1.973 -0.016  1.681 12.803
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  133.138912  11.531157  11.546 < 2e-16 ***
## cylinders     -0.432749   0.299977  -1.443 0.149946
## displacement   0.020977   0.006825   3.074 0.002265 ** 
## horsepower    -0.010072   0.012546  -0.803 0.422593    
## acceleration   0.135179   0.089798   1.505 0.133051    
## year           0.788784   0.047596  16.573  < 2e-16 ***
## origin         1.011407   0.262262   3.856 0.000135 ***
## log(weight)  -21.858785   1.651400 -13.237  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.091 on 384 degrees of freedom
## Multiple R-squared:  0.8459, Adjusted R-squared:  0.8431 
## F-statistic: 301.2 on 7 and 384 DF,  p-value: < 2.2e-16
```

This model including all variables except name but with log(weight) has a slightly higher adjusted R-squared than the original mlr model (0.8431 > 0.8182), meaning more variability is explained by the predictors.

```r
# sqrt(horsepower)
mlr4 <- lm(mpg ~ . - name - horsepower + sqrt(horsepower), data = Auto)
summary(mlr4)
```

```
##
## Call:
## lm(formula = mpg ~ . - name - horsepower + sqrt(horsepower), 
##     data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.5240 -1.9910 -0.1687  1.8181 12.9211 
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -6.0373910  5.5460041  -1.089 0.277012    
## cylinders        -0.5222540  0.3166839  -1.649 0.099938 .  
## displacement      0.0220542  0.0071987   3.064 0.002341 ** 
## weight           -0.0054593  0.0006842  -7.979 1.72e-14 ***
## acceleration     -0.1021239  0.1038565  -0.983 0.326070    
## year              0.7240379  0.0501791  14.429  < 2e-16 ***
## origin            1.5173206  0.2703470   5.612 3.83e-08 ***
## sqrt(horsepower) -1.1434906  0.3113771  -3.672 0.000274 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.277 on 384 degrees of freedom
## Multiple R-squared:  0.8269, Adjusted R-squared:  0.8237 
## F-statistic:   262 on 7 and 384 DF,  p-value: < 2.2e-16
```

This model including all variables except name but with sqrt(horse) has a slightly higher adjusted R-squared than the original mlr model (0.8237 > 0.8182), meaning more variability is explained by the predictors.

Question 6

This problem focuses on the collinearity problem.

(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?

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

The linear model is

$$Y = 2 + 2X_1 + 0.3X_2 + \epsilon$$

where

$$\epsilon_i \sim i.i.d.N(0, 100)$$

.

The coefficients are

$$\beta_0 = 2$$

,

$$\beta_1 = 2$$

, and

$$\beta_2 = 0.3$$

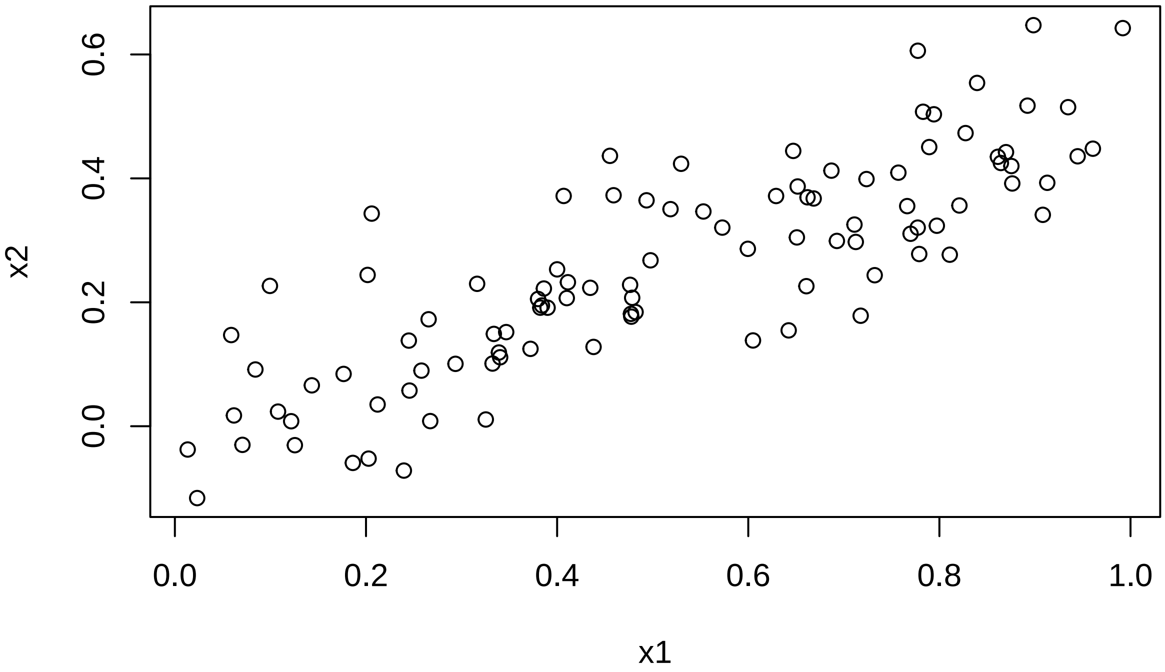
.

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

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

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

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



The correlation between x1 and x2 is 0.835, which is quite high. In addition, from the scatter plot we can learn that there is a strong positive linear relationship b/w x1 and x2.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are ˆ B0, ˆ B1, and ˆ B2? How do these relate to the true B0, B1, and B2? Can you reject the null hypothesis H0 : B1 = 0? How about the null hypothesis H0 : B2 = 0?

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

```
##
## 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
```

The fitted linear model is

$$\hat{Y} = 2.13 + 1.44X_1 + 1.01X_2$$

where

$$\hat{\beta}_0 = 2.13$$

,

$$\hat{\beta}_1 = 1.44$$

, and

$$\hat{\beta}_2 = 1.01$$

.

Comparing these to the true values,

$$\beta_0$$

is off by 2 - 2.13 = -0.13,

$$\beta_1$$

is off by 2 - 1.44 = 0.56, and

$$\beta_2$$

is off by 0.3 - 1.01 = -0.71. This indicates a moderate bias b/w the estimated and the true parameters.

We reject the null hypothesis that

$$\beta_1 = 0$$

because p-value = $0.0487 < 0.05$ threshold.

We fail to reject the null hypothesis that

$$\beta_2 = 0$$

because p-value = $0.3754 > 0.05$ threshold.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 : B1 = 0?

```
lsr2 <- lm(y ~ x1)
summary(lsr2)
```

```
##
## 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
```

The fitted linear model is

$$\hat{Y} = 2.11 + 1.98X_1$$

where

$$\hat{\beta}_0 = 2.11$$

, and

$$\hat{\beta}_1 = 1.98$$

.

Both

$$\hat{\beta}_0$$

and

$$\hat{\beta}_1$$

are closer to the true values than the previous lsr model.

We reject the null hypothesis that

$$\beta_1 = 0$$

because p-value = 2.66e-06 $<$ 0.05 threshold.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis H0 : B1 = 0?

```
lsr3 <- lm(y ~ x2)
summary(lsr3)
```

```
##
## 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
```

The fitted linear model is

$$\hat{Y} = 2.39 + 2.90X_2$$

where

$$\hat{\beta}_0 = 2.39$$

, and

$$\hat{\beta_1} = 2.90$$

.

Both

$$\hat{\beta_0}$$

and

$$\hat{\beta_1}$$

are farther from the true values than the previous lsr model.

We reject the null hypothesis that

$$\beta_1 = 0$$

because p-value = 1.37e-05 < 0.05 threshold.

(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

We know that x1 and x2 has a high correlation = 0.835. As a result, if we use both of them as predictors in the regression model, only one is statistically significant due to the high collinearity, which is the case in (c). However, when we split them into separate models, they each have strong relationship (as indicated by the near-zero p-values) with the response variable Y since the problem of collinearity is avoided. Therefore, the results from (d) and (e) both make sense as well.