HOMEWORK 4

Andrew Guo

2/28/2022

PROBLEM 1

This question should be answered using the `Carseats` data set. Information can be found here: `https://rdrr.io/cran/ISLR/man/Carseats.html`

(a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`. Summarize the results.

```
Seats = read.csv('Carseats.csv')
attach(Seats)

lm.fit.Sales = lm(Sales~Price + Urban + US)
summary(lm.fit.Sales)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## 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
```

(b) Provide an interpretation of each coefficient β_j in the model with respect to a unit change in the predictor. Be careful – some of the variables in the model are qualitative!

The Intercept states that when the carseat is both not from an Urban store nor from a US store, a one unit increase in x0 leads to a 13.043469 increase in sales, holding all other variables constant.

The Price coefficent states that for a one unit increase in x1, the sales will decrease by -0.054459 units, holding all other predictors constant.

The UrbanYes coefficent states that for a one unit increase in x2, provided that the carseat was sold in an Urban store, the sales will decrease by -0.021916 units, holding all other predictors constant. This is also equal to the difference in the expected mean value of sales when the carseat is both sold in a store that is in the US and in an Urban area and the expected mean value of sales when the carseat is not sold in an Urban area, but is sold in the U.S.

The USYes coefficeint states that for a one unit increase in x3, provided that the carseat was sold in the U.S., the sales will increase by 1.200573 units, holding all other predictors constant. This is also equal to the difference in the expected mean value of sales when the carseat is both sold in a store that is in the US and in an Urban area and the expected mean value of sales when the carseat is sold in an Urban area, but is not sold in the U.S.

(c) Use the results of the fit to determine model utility. Is at least one of the predictors associated with the response? Justify your answer.

With a global utility F-statistic of 41.52 and its associated p-value being far less than 0.05, we conclude that the model is useful. There is convincing evidence that at least one of the predictors is related to the response.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$? Justify your answer.

If a predictor contains a p-value that is below a = 0.05, then we can rejecth the null hypothesis that betahat = 0, because this indicates the presence of a statistically significant predictor. All of our predictors, with the exception of UrbanYes, meet this description, and thus, it is those predictors that we can reject the null hypothesis $H_0 : \beta_j = 0$ for.

(e) On the basis of your response to the previous question, fit a reduced model that only uses the predictors for which there is evidence of association with the outcome. Summarise the results.

```
lm.fit.Sales2 = lm(Sales~Price + US)
summary(lm.fit.Sales2)
```

```
##
## Call:
```

```
## lm(formula = Sales ~ Price + US)
##
## 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
```

(f) How well do the models in (a) and (e) fit the data? Justify your answer using the appropriate metrics.

```
summary(lm.fit.Sales)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## 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
```

```
summary(lm.fit.Sales2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US)
##
## 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
```

Because the number of observations in our data is significantly larger than the number of predictors in the model, it is safe to use r-squared and adjusted r-squared as interpreters for how good of a fit the model is for showcasing the relationship between the predictor and the response.

In the model from (a), there is an adjusted R-squared of 0.2335, meaning that only 23.35% of the variation in our data on Sales can be explained by the model on Price, Urban status, and US status, taking into account both the sample size and the number of parameters in the model.

In model from (e), there is an adjusted R-squared of 0.2354, meaning that only 23.54% of the variation in our data on Sales can be explained by the model on Price and US status, taking into account both the sample size and the number of parameters in the model.

With such a low adjusted r-squared value, this means that in both cases, there is a weak relationship between the response and predictors. Thus, the models in (a) and (e) do not fit the data very well.

(g) Using the model from (e), obtain 95% confidence intervals for the significant coefficient(s).

```
confint(lm.fit.Sales2, 'Price')
```

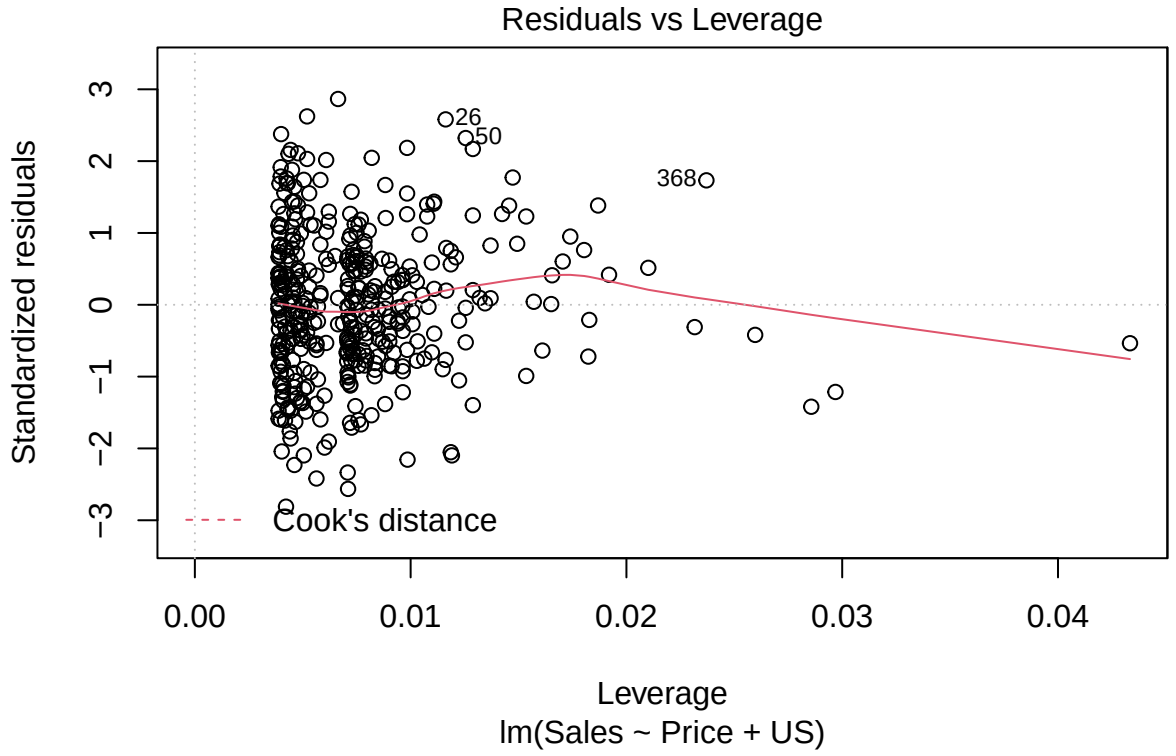
```
##              2.5 %      97.5 %
## Price -0.06475984 -0.04419543
```

```
confint(lm.fit.Sales2, 'USYes')
```

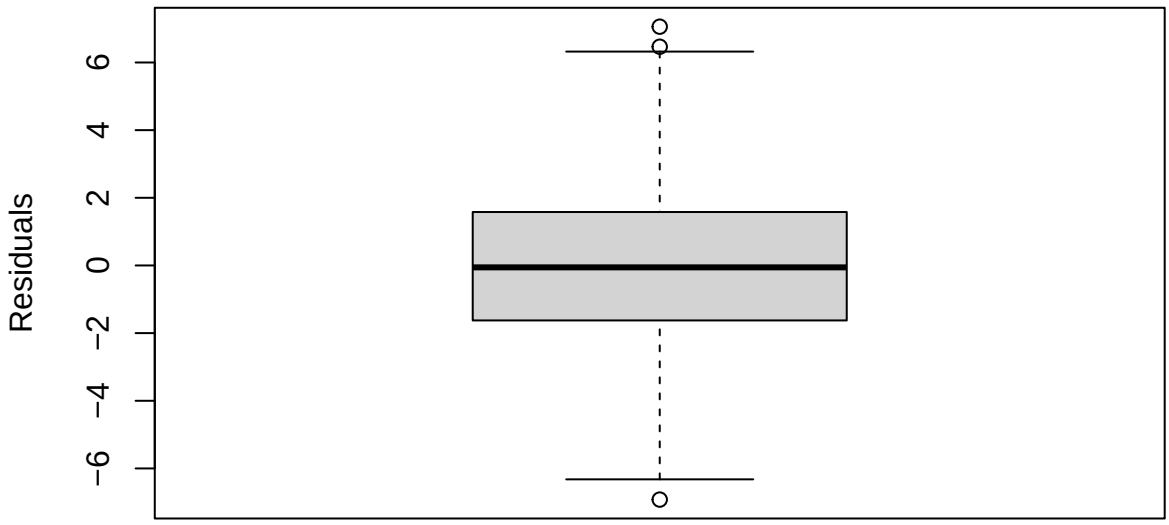
```
##              2.5 %   97.5 %
## USYes 0.6915196 1.707766
```

(h) Is there evidence of outliers or high leverage observations in the model from (e)? Justify your answer using the appropriate plot(s)

```
plot(lm.fit.Sales2, which = c(5))
```




```
boxplot(lm.fit.Sales$residuals, ylab="Residuals")
```



The residuals vs leverage plot also does not contain many concerning points; no observation falls outside of +-3 standard deviations, and even those points that are near +-3 standard deviations are extremely low in leverage. Thus, there are no high leverage observations.

However, the boxplot does indicate that there are three outliers. However, this should not be too much of a concern considering how few in number they are and that the overall shape of the boxplot is fairly symmetrical and the median is located roughly in the center of the plot.

PROBLEM 2

In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor `x` and a response `y` as follows.

```
> set.seed(1)
> x <- rnorm(100)
> y <- 2 * x + rnorm(100)
```

```
set.seed(1)
x <- rnorm(100)
y <- 2*x + rnorm(100)
```

(a) Perform a simple linear regression of `y` onto `x`, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the command `lm(y~x+0)`.)

```
lm.fit.Y = lm(y~x + 0)
summary(lm.fit.Y)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

Coefficient estimate $\hat{\beta}$: 1.9939

Standard error: 0.1065

t-statistic: 18.73

p-value associated with the null hypothesis $H_0 : \beta = 0$: <2e-16

With a p-value less than a = 0.05, we reject the null. There is convincing evidence that y is a statistically significant predictor for x.

(b) Now perform a simple linear regression of `x` onto `y` without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

```
lm.fit.X = lm(x~y + 0)
summary(lm.fit.X)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

Coefficient estimate $\hat{\beta}$: 0.39111

Standard error: 0.02089

t-statistic: 18.73

p-value associated with the null hypothesis $H_0 : \beta = 0$: <2e-16

With a p-value less than a = 0.05, we reject the null. There is convincing evidence that x is a statistically significant predictor for y. Additionally, we also notice that the t-statistic for an SLR of x onto y is the same as an SLR of y onto x.

(c) Compare the results obtained in (a) and (b). Which aspects are similar? Which are dissimilar?

The standard error and coefficeint estimates are different, but the t-statistics and p-values are identical. Both predictors in each of the models are also statistically significant.

(d) Why is the slope for part (a) accurately estimated, in contrast to part (b)? Explain.

In part (b), we're essentially flipping the x and the y variable positions, changing the slopes, and creating a completely different function. The response in part (a) took the values that were in y while the predictor in (b) was the values that were in x and a slope was estimated based on the individual x values as predictors for each of its respective individual y values.

Part (b) is completely different, the response variable took on the values that are in y, while the predictor took on the values that are in x, which would create a completely different slope because the values in 2(rnorm(100)) + rnorm(100) are being used to regress rnorm(100). Two completely different slopes are generated as a result.

PROBLEM 3

In *Chance* (Fall 2000), statistician Scott Berry built a multiple regression model for predicting total number of runs scored by a Major League Baseball team during a season. Using data on all teams over a 10-year period (a sample of $n = 234$), the following results were obtained:

Variable	$\hat{\beta}$	$\hat{SE}(\hat{\beta})$	Variable	$\hat{\beta}$	$\hat{SE}(\hat{\beta})$
Intercept	3.70	15.00	Home Runs (x_5)	1.51	0.05
Walks (x_1)	0.34	0.02	Stolen Bases (x_6)	0.26	0.05
Singles (x_2)	0.49	0.03	Caught Stealing (x_7)	-0.14	0.14
Doubles (x_3)	0.72	0.05	Strikeouts (x_8)	-0.10	0.01
Triples (x_4)	1.14	0.19	Outs (x_9)	-0.10	0.01

(a) Conduct a test of $H_0 : \beta_j = 0$ against $H_a : \beta_j \neq 0$ for each coefficient at $\alpha = .05$. Summarize the results in the following table. Note that for a multiple linear regression model, the $df = n - (p + 1))$ for the t-statistic.

df = 234-(10) = 224

Variable	t-statistic	p-value	Variable	t-statistic	p-value
Intercept	0.246	0.8059075	Home Runs	30.02	1.836923e-80
Walks	17	3.500235e-42	Stolen Bases	5.2	4.488363e-07
Singles	16.33	5.207177e-40	Caught Stealing	-1	0.3183895
Doubles	14.4	1.021056e-33	Strikeouts	-10	1.057151e-19
Triples	5.863	1.615733e-08	Outs	-10	1.057151e-19

2*pt(t, df, lower.tail = FALSE/TRUE)

(b) Form a 95% confidence interval for β_5. Interpret the results.

β_5 = Home Runs = 1.51

```
qt(0.025, 224, lower.tail = FALSE)
```

```
## [1] 1.970611
```

1.51 +- 1.970611(0.05)

1.51 +- 0.09853055

[1.41146945, 1.60853055]

I am 95% confident that the true value of β_5 lies between 1.41146945 and 1.60853055.

Problem 4

Because the coefficient of determination R^2 always increases when a new independent variable is added to the model, it is tempting to include many variables in a model to force R^2 to be near 1. However, doing so reduces the degrees of freedom available for estimating σ^2 , which adversely affects our ability to make reliable inferences. As an example, suppose you want to use the responses to a survey consisting of 16 demographic, social, and economic questions to model a college student's intelligence quotient (IQ). You fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{16} x_{16} + \epsilon$$

where $y =$ IQ and $x_1, x_2, \ldots, x_{16}$ are the 16 independent variables. Data for only 20 students ($n = 20$) are used to fit the model, and you obtain $R^2 = .95$.

(a) Use the omnibus F-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n - (p + 1))} = \frac{R^2/p}{(1 - R^2)(n - (p + 1))}$$

to see whether this impressive looking R^2 is large enough for you to infer that at least one term in the model is important for predicting IQ. Use $\alpha = .05$.

$$F = \frac{0.95/16}{(1 - 0.95)(20 - (16 + 1))}$$

$$F = \frac{0.95/16}{(0.05)(3)}$$

$$F = \frac{0.06}{0.15} = 0.4$$

```
pf(0.4, 16, 3, lower.tail = FALSE)
```

```
## [1] 0.9034632
```

$$p(F > 0.4) = 0.9034632$$

With a p-value greater than 0.05, we fail to reject the null. There is not enough convincing evidence that at least one of the parameters is a statistically significant predictor for a college student's IQ.

(b) Calculate the adjusted R^2, R_a^2, using

$$R_a^2 = 1 - \frac{(n-1)}{(n-(p+1))}(1-R^2)$$

and interpret its value.

$$R_a^2 = 1 - \frac{(20-1)}{(20-(16+1))}(1-0.95)$$

$$R_a^2 = 1 - \frac{19}{3}(0.05)$$

$$R_a^2 = 1 - \frac{0.95}{3}$$

$$R_a^2 = 1 - 0.316666666 = 0.683$$

With an adjusted R-squared of approximate 0.683, meaning that only 68.3% of the variation in our data on Sales can be explained by the model, taking into account both the sample size and the number of parameters in the model. While this value still indicates a moderately strong relationship between the response and the predictors, this is far lower than the R-squared value of 0.95 without taking into consideration sample size and number of parameters.

Problem 5

This question should be answered using the `FuelConsumptionCo2` data set. Information can be found here: `https://github.com/kvinlazy/Dataset/blob/master/FuelConsumptionCo2.csv`

(a) Read in the data and fit a multiple linear regression model with all interactions using `CO2EMISSIONS` as the response and the following predictors: `ENGINESIZE`, `CYLINDERS`, `FUELCONSUMPTION_CITY`, `FUELCONSUMPTION_HWY`.

```
Fuel = read.csv('FuelConsumptionCo2.csv')
attach(Fuel)
```

```
lm.fit.Fuel = lm(CO2EMISSIONS~ENGINESIZE*CYLINDERS*FUELCONSUMPTION_CITY*FUELCONSUMPTION_
summary(lm.fit.Fuel)
```

```
##
## Call:
## lm(formula = CO2EMISSIONS ~ ENGINESIZE * CYLINDERS * FUELCONSUMPTION_CITY *
##     FUELCONSUMPTION_HWY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.711  -4.568  -1.801   5.180  72.825
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                      64.42394
## ENGINESIZE                                                      -76.90080
## CYLINDERS                                                       -59.67919
## FUELCONSUMPTION_CITY                                             24.62148
## FUELCONSUMPTION_HWY                                              15.34749
## ENGINESIZE:CYLINDERS                                             15.74195
## ENGINESIZE:FUELCONSUMPTION_CITY                                   0.73619
## CYLINDERS:FUELCONSUMPTION_CITY                                    3.31023
## ENGINESIZE:FUELCONSUMPTION_HWY                                   11.68828
## CYLINDERS:FUELCONSUMPTION_HWY                                     4.73772
## FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                         -3.10700
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY                        -0.83842
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_HWY                         -1.83263
## ENGINESIZE:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY              -0.13984
## CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY               -0.10190
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY     0.07589
##                                                                Std. Error
## (Intercept)                                                      87.31467
## ENGINESIZE                                                       30.26527
## CYLINDERS                                                        19.89850
## FUELCONSUMPTION_CITY                                              7.82102
```

```
## FUELCONSUMPTION_HWY                                                 12.96205
## ENGINESIZE:CYLINDERS                                                 4.64351
## ENGINESIZE:FUELCONSUMPTION_CITY                                      2.27864
## CYLINDERS:FUELCONSUMPTION_CITY                                       1.62683
## ENGINESIZE:FUELCONSUMPTION_HWY                                       4.08181
## CYLINDERS:FUELCONSUMPTION_HWY                                        2.91465
## FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                             0.66061
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY                            0.32800
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_HWY                             0.57133
## ENGINESIZE:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                  0.18985
## CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                   0.11848
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY        0.02561
##                                                                   t value Pr(>|t|)
## (Intercept)                                                         0.738 0.460779
## ENGINESIZE                                                         -2.541 0.011200
## CYLINDERS                                                          -2.999 0.002771
## FUELCONSUMPTION_CITY                                                3.148 0.001690
## FUELCONSUMPTION_HWY                                                 1.184 0.236668
## ENGINESIZE:CYLINDERS                                                3.390 0.000725
## ENGINESIZE:FUELCONSUMPTION_CITY                                     0.323 0.746696
## CYLINDERS:FUELCONSUMPTION_CITY                                      2.035 0.042125
## ENGINESIZE:FUELCONSUMPTION_HWY                                      2.864 0.004273
## CYLINDERS:FUELCONSUMPTION_HWY                                       1.625 0.104360
## FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                           -4.703  2.9e-06
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY                          -2.556 0.010724
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_HWY                           -3.208 0.001379
## ENGINESIZE:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                -0.737 0.461532
## CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                 -0.860 0.389943
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY       2.964 0.003109
##
## (Intercept)
## ENGINESIZE                                                        *
## CYLINDERS                                                         **
## FUELCONSUMPTION_CITY                                              **
## FUELCONSUMPTION_HWY
## ENGINESIZE:CYLINDERS                                              ***
## ENGINESIZE:FUELCONSUMPTION_CITY
## CYLINDERS:FUELCONSUMPTION_CITY                                    *
## ENGINESIZE:FUELCONSUMPTION_HWY                                    **
## CYLINDERS:FUELCONSUMPTION_HWY
## FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY                          ***
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY                         *
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_HWY                          **
## ENGINESIZE:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY
## CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY
```

```
## ENGINESIZE:CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.18 on 1051 degrees of freedom
## Multiple R-squared:  0.9189, Adjusted R-squared:  0.9177
## F-statistic: 793.7 on 15 and 1051 DF,  p-value: < 2.2e-16
```

(b) Based on the results of the summary, are there any terms you would eliminate from the model? Explain.

The global omnibus test shows a p-value of less than a =0.05, thus indicating that at least one of the predictors is statistically significant. However, this does not mean that all of the predictors are statistically significant. I would begin by removing all insignificant interaction terms (those with a corresponding p-value > 0.05). These terms are: ENGINESIZE:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY, CYLINDERS:FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY, FUELCONSUMPTION_CITY:FUELCONSUMPTION_HWY, and ENGINESIZE:FUELCONSUMPTION_CITY. There is one main term that is statistically significant (FUELCONSUMPTION_HWY), but because this variable contains interaction terms that are statistically significant (for example: ENGINESIZE:FUELCONSUMPTION_HWY and ENGINESIZE:CYLINDERS:FUELCONSUMPTION_HWY), I will keep this term in the model.