# Multiple Linear Regression

Taylor Shrode

3/29/2020

MSDS 660

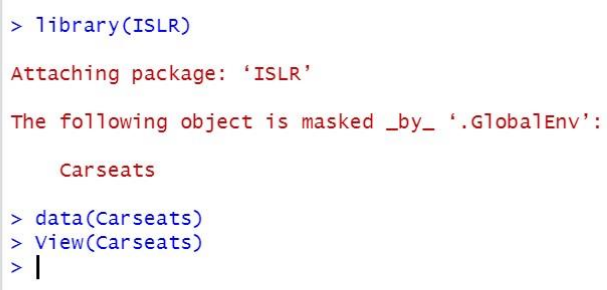
Multiple linear regression, MLR, is a statistical technique that uses more than one explanatory (independent, or predictor) variable to predict the outcome of a response (dependent, or response) variable (Kenton, 2019). The goal is to model the linear relationship between the independent and dependent variables (Kenton, 2019). The generic formula for multiple linear regression is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where $Y$is the dependent variable, $x_k$is the explanatory variables, $\beta_0$is the y-intercept, and $\beta_k$is the slope coefficients for each explanatory variable (Kenton, 2019). There are several assumptions that must be met in multiple linear regression. These assumptions include (Assumptions of multiple linear regression, n.d.):

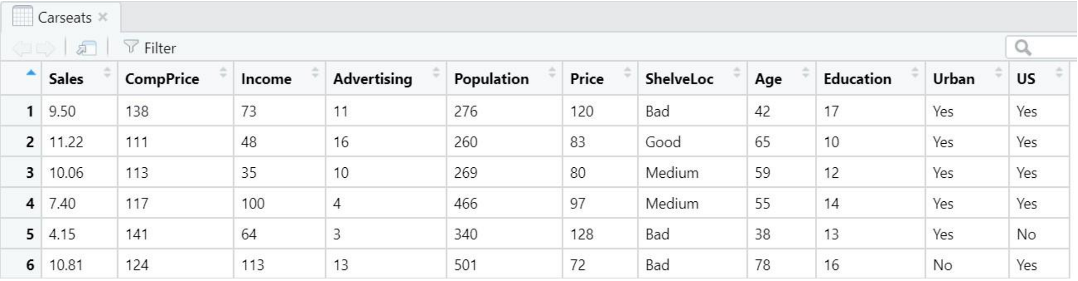
1. There must be a linear relationship between the independent variables and the dependent variable. This can be shown by using scatterplots.
2. The residuals must be normally distributed.
3. The independent variables are not highly correlated with each other. In other words, there is no multicollinearity.
4. The variance of error terms are similar across the independent variables. This can be shown by plotting the standardized residuals versus predicted values and verifying that the points are equally distributed.

For this assignment, we will be performing multiple linear regression using a dataset called *Carseats* located in the **ISLR** library, which is the same dataset we used last week. To begin, we will load the **ISLR** library and then load the *Carseats* dataset. To ensure the data was loaded properly, we use the **View()** command to view the data.



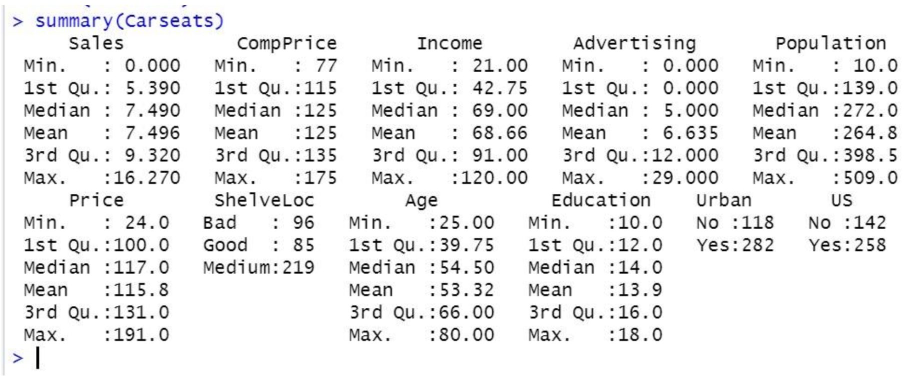
```
> library(ISLR)

Attaching package: ‘ISLR’

The following object is masked _by_ ‘.GlobalEnv’:

    Carseats

> data(Carseats)
> View(Carseats)
> 
```

As we can see below, we have successfully loaded the dataset and can continue.



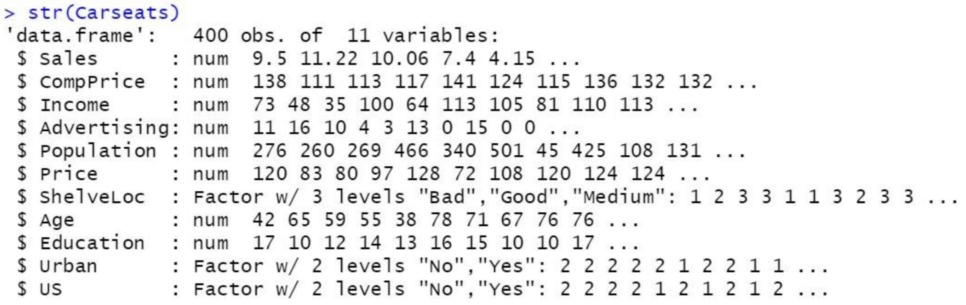
| | Sales | CompPrice | Income | Advertising | Population | Price | ShelveLoc | Age | Education | Urban | US |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9.50 | 138 | 73 | 11 | 276 | 120 | Bad | 42 | 17 | Yes | Yes |
| 2 | 11.22 | 111 | 48 | 16 | 260 | 83 | Good | 65 | 10 | Yes | Yes |
| 3 | 10.06 | 113 | 35 | 10 | 269 | 80 | Medium | 59 | 12 | Yes | Yes |
| 4 | 7.40 | 117 | 100 | 4 | 466 | 97 | Medium | 55 | 14 | Yes | Yes |
| 5 | 4.15 | 141 | 64 | 3 | 340 | 128 | Bad | 38 | 13 | Yes | No |
| 6 | 10.81 | 124 | 113 | 13 | 501 | 72 | Bad | 78 | 16 | No | Yes |

Before we begin fitting a linear model to our dataset, we will first look at the summary of our data.



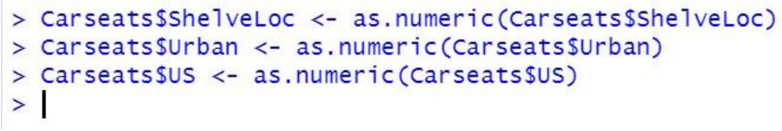
```
> summary(Carseats)
     Sales          CompPrice       Income        Advertising       Population   
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0  
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0  
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0  
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8  
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5  
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0  
     Price        ShelveLoc        Age          Education    Urban       US     
 Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0   No :118   No :142  
 1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0   Yes:282   Yes:258  
 Median :117.0   Medium:219   Median :54.50   Median :14.0                      
 Mean   :115.8                Mean   :53.32   Mean   :13.9                      
 3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0                      
 Max.   :191.0                Max.   :80.00   Max.   :18.0                      
> |
```

Next, we can display the structure of our dataset by using the **str()** command.



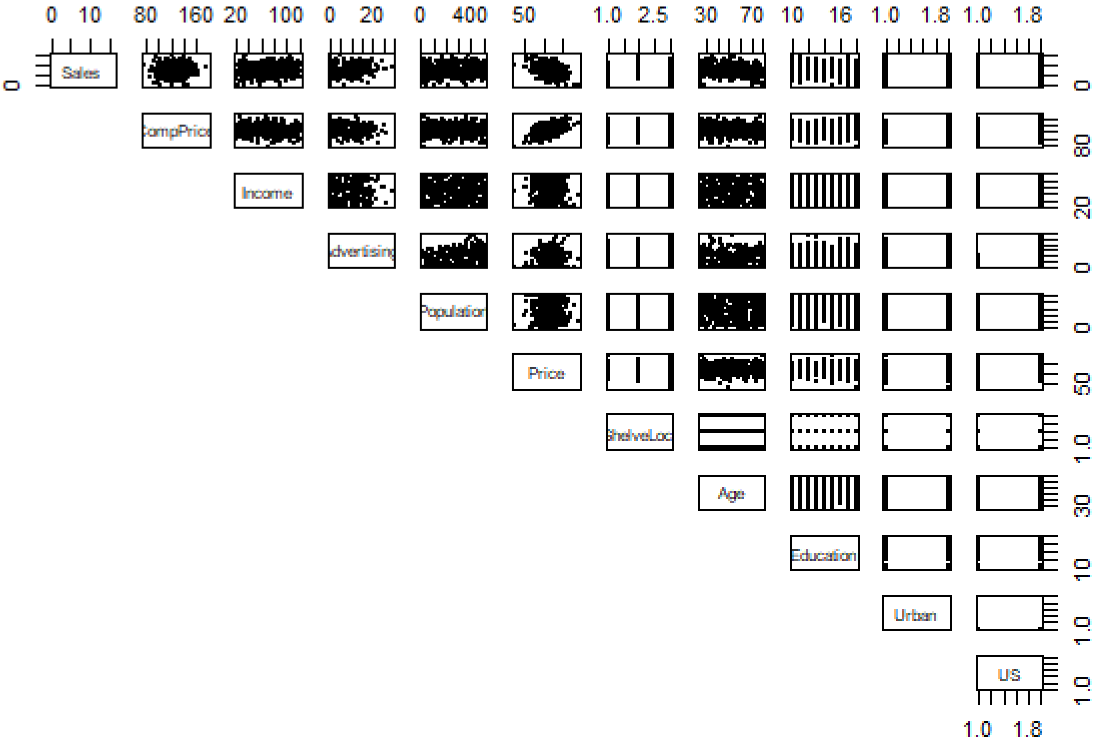
```
> str(Carseats)
'data.frame':   400 obs. of  11 variables:
 $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
 $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
 $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
 $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
 $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

As stated above, one of the assumptions for MLR is that there must be a linear relationship between the independent (predictor) variables and the dependent (response) variable, which can be detected with the use of scatter plots. First, all of the variables need to be in numeric format. Thus, **ShelveLoc, US,** and **Urban** need to be converted. This can be done using the **as.numeric** function.

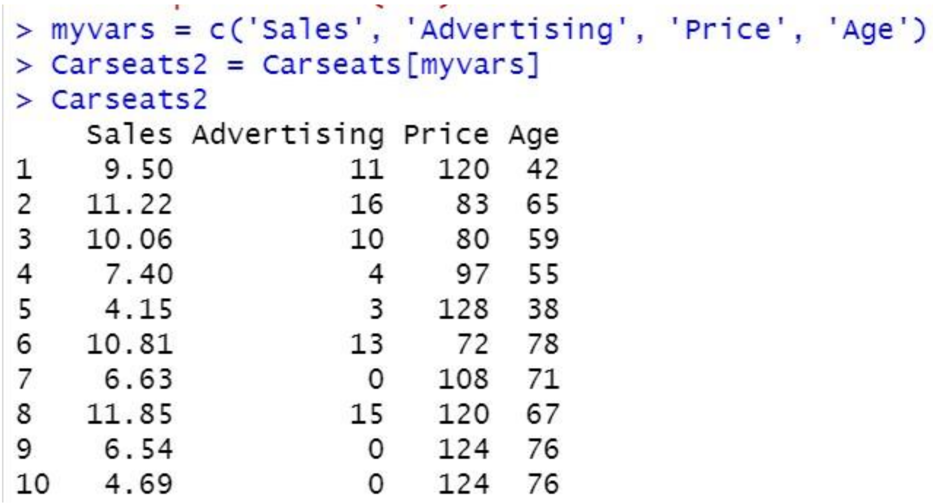


```
> Carseats$ShelveLoc <- as.numeric(Carseats$ShelveLoc)
> Carseats$Urban <- as.numeric(Carseats$Urban)
> Carseats$US <- as.numeric(Carseats$US)
> |
```

Now, after creating a scatter plot matrix of our *Carseats* dataset, we can identify linear relationships. The command to create the plot below is **plot(Carseats, lower.panel = NULL, pch = 19, cex = 0.5).**

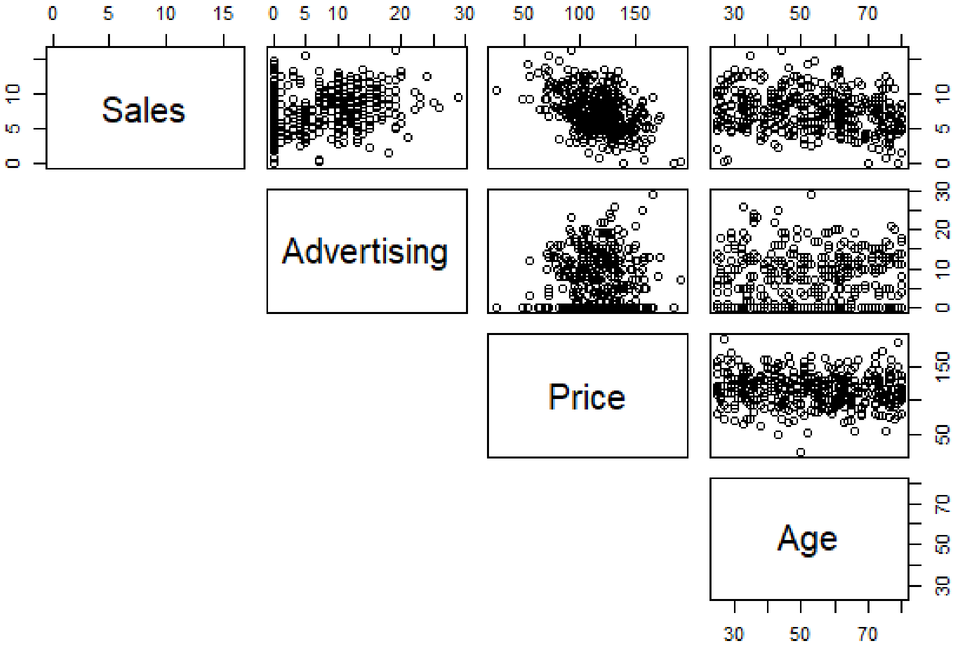


From this, many of the linear relationships appear to be paired with the **Sales** variable. Using this variable as our dependent (response) variable, we can then choose our independent (predictor) variables. To fit the first multiple linear model, we will use the variables **Advertising, Price,** and **Age** as the independent variables. Using the same **plot**() function, we can view the scatterplot matrix that includes only the variables we have chosen, similar to above. First, we need to create a new dataframe with only the data we need.



```
> myvars = c('Sales', 'Advertising', 'Price', 'Age')
> Carseats2 = Carseats[myvars]
> Carseats2
   Sales Advertising Price Age
1   9.50          11   120  42
2  11.22          16    83  65
3  10.06          10    80  59
4   7.40           4    97  55
5   4.15           3   128  38
6  10.81          13    72  78
7   6.63           0   108  71
8  11.85          15   120  67
9   6.54           0   124  76
10  4.69           0   124  76
```

Now, we can create the plot.



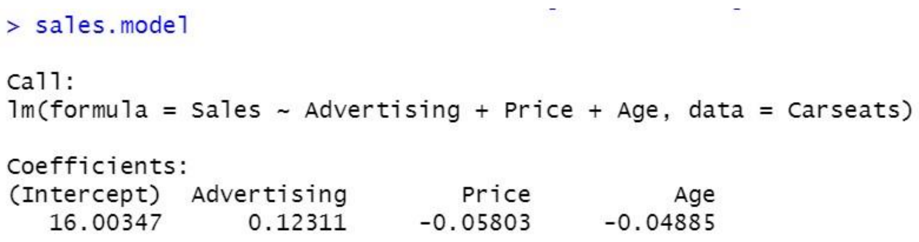
Focusing on the relationships with **Sales**, we see that the **Sales-Advertising** relationship appears to be slightly positive, the **Sales-Price** relationship is negative, and the **Sales-Age** relationship is slightly negative. Thus, we have confirmed that the explanatory variables all have a linear relationship with our dependent variable.

Now that we have chosen the variables to be used in our linear model, we need to identify the null and alternative hypotheses. The null hypothesis states that there is no linear relationship between the variables, which means that the coefficients are all 0 (WorldClass FTE). The alternative hypothesis states that at least one of the coefficients is not 0. This means that at least one of the independent variables affects the dependent variable (WorldClass FTE). In other words,

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

$$H_A: At\ least\ one\ \beta\ is\ not\ zero.$$

We can now create our linear model using the three predictors chosen above. To create the model, we can use the **lm()** function, which takes in a formula argument and a data argument. Generically, the function looks like **lm(YVAR~XVAR1 + XVAR2 + … + XVARk, data)**, where **YVAR** is the dependent variable, the **XVAR's** are the independent variables, and **data** is the variable that contains the data set (Quick, 2009). Thus, our function looks like **lm(Sales ~ Advertising + Price + Age, Carseats)**. After saving this model in a variable, we have called **sales.model**, we can display the output of the function above.

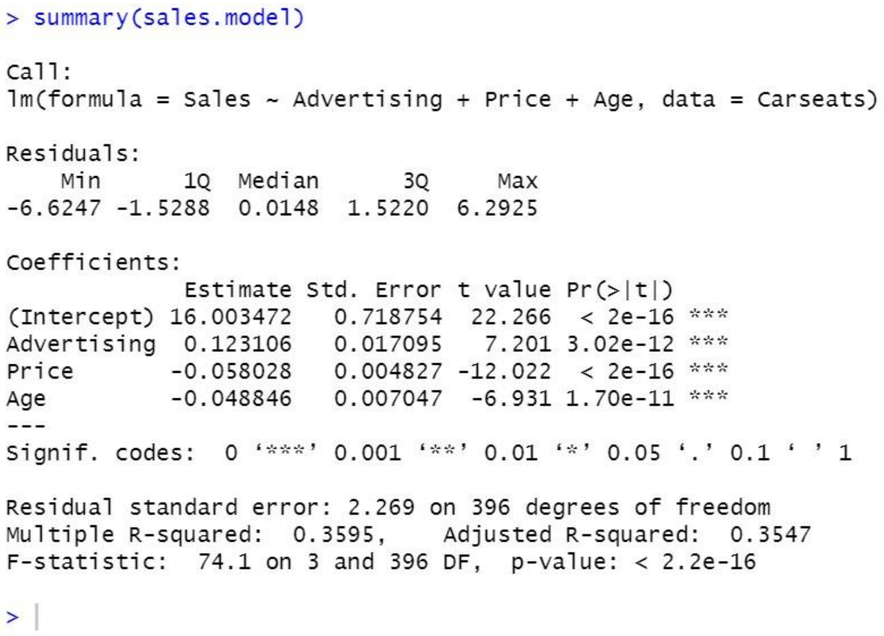


```
> sales.model

Call:
lm(formula = Sales ~ Advertising + Price + Age, data = Carseats)

Coefficients:
(Intercept)  Advertising        Price          Age  
   16.00347      0.12311     -0.05803     -0.04885  
```

From this output, we can determine our coefficients for our multiple linear regression formula. Our formula is

$$Sales \;=\; 16.00347 + 0.12311^{*}Advertising - 0.05803^{*}Price - 0.04885^{*}Age.$$

This equation tells us that the predicted number of sales of car seats will increase as the amount of advertising increases, the price of car seats decreases, and when the age of the buyer decreases. Now, we can determine the validity of the model by analyzing the summary of the model.



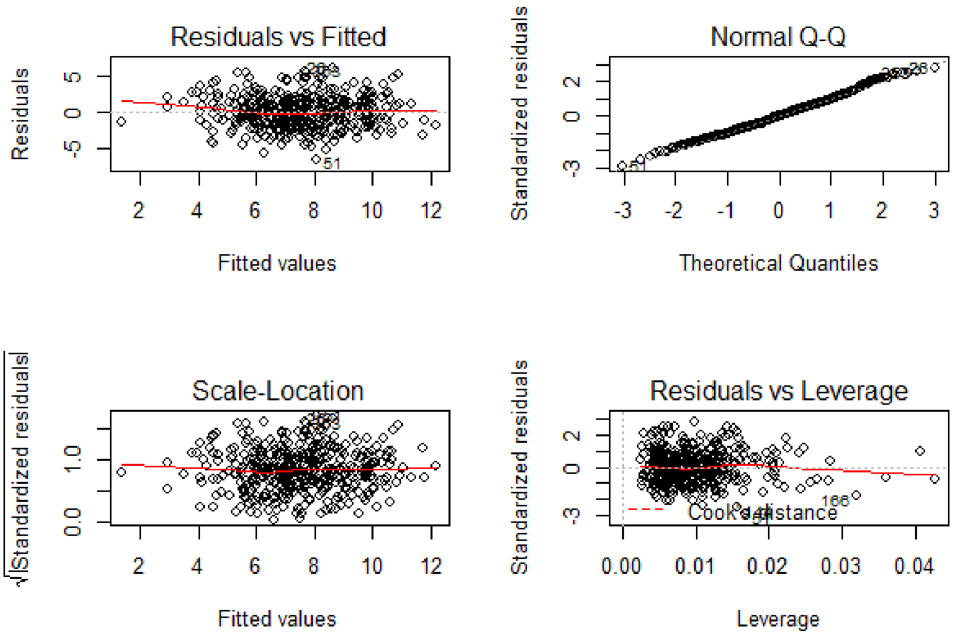
```
> summary(sales.model)

Call:
lm(formula = Sales ~ Advertising + Price + Age, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6247 -1.5288  0.0148  1.5220  6.2925 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.003472   0.718754  22.266  < 2e-16 ***
Advertising  0.123106   0.017095   7.201 3.02e-12 ***
Price       -0.058028   0.004827 -12.022  < 2e-16 ***
Age         -0.048846   0.007047  -6.931 1.70e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.269 on 396 degrees of freedom
Multiple R-squared:  0.3595,	Adjusted R-squared:  0.3547 
F-statistic:  74.1 on 3 and 396 DF,  p-value: < 2.2e-16

> |
```

First, we need to examine the F-statistic and the associated p-value (Multiple linear regression, 2018). The further the F-statistic is from 1, the better (Rego, 2015). Above, we see that our F-statistic is 74.1, which is far away from 1, indicating that our model is statistically significant. The associated p-value is < 2.2e-16 which is highly significant as it is less than 0.05, which means at least one of the independent variables affects the dependent variable.

To determine which predictors are significant, we look at the coefficients in our summary output. The t-statistic (t value) evaluates if there is significant association between the predictor and the outcome variable or not (Multiple linear regression, 2018). The t-statistic determines how many standard deviations the coefficient is from 0 (Rego, 2015). As we can see above, all three predictor variables are far from 0. Also, all of the p-values (Pr(>|t|)) for our coefficients are less than 0.05 which indicates that we can reject the null hypothesis and conclude that there is a relationship between our predictors and response variables.

To check the accuracy of the model, we can examine the R-squared values (Multiple linear regression, 2018). The R-squared value represents the “proportion of variance, in the outcome variable, that may be predicted by knowing the value of the predictor variables. An R-squared value close to 1 indicates that the model explains a large portion of the variance in the outcome variable (Multiple linear regression, 2018).” Our R-squared value in the model above is a 0.3595. In other words, 35.95% of the variance in the measure of sales can be predicted by advertising, price, and income.

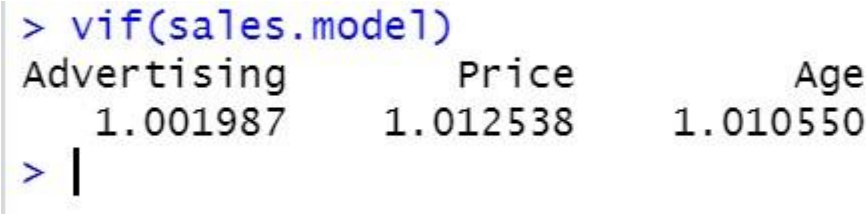
Next, we plot the model diagnostics to validate the assumptions of the model.



To verify that our linear relationship assumption we made above is correct, we can use the Residuals vs Fitted plot (Linear regression assumptions and diagnostics, 2018). The horizontal line and no distinct pattern that is shown above indicates that there is a linear relationship. Now, we need to check the assumption that the residuals are normally distributed. The Normal Q-Q plot is used to determine this (Linear regression assumptions and diagnostics, 2018). As we can see above, the residuals points follow the straight dashed line, which indicates that the residuals are normally distributed. Next, we can check the assumption that there is homogeneity of variance of the residuals, or homoscedasticity. This is done by using the Scale-Location plot (Linear regression assumptions and diagnostics, 2018). In the plot above, we see that we have a horizontal line which indicates homoscedasticity. The Residuals vs Leverage is used to identify extreme values that might alter the regression results (Linear regression assumptions and diagnostics, 2018). As shown above, we see that the plot identified the influential observations as

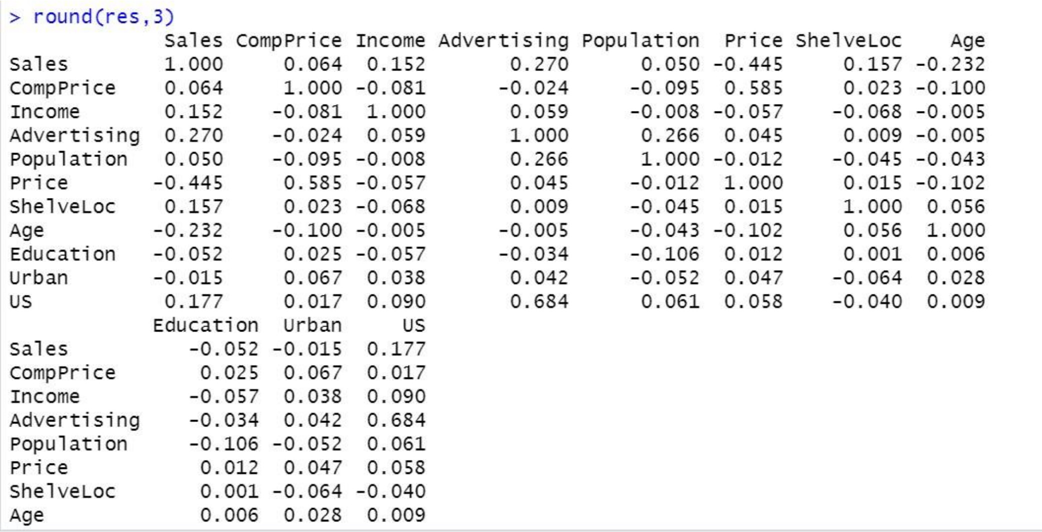
#166. In the future, the model would be reconstructed without this data point and then the model would be analyzed.

Finally, we need to ensure there is no multicollinearity. Multicollinearity occurs when there are high correlations between two or more predictor variables (Stephanie, 2015). To detect multicollinearity, we can calculate the variance inflation factor (VIF) of the model or we can create a correlation matrix (WorldClass FTE). First, we look at the VIF of our predictors in **sales.model.** A value of 1 means that the predictor is not correlated with the other variables, and as the VIF value gets higher, the correlation between the variables is greater (Bock, 2018). We can use the **vif()** function in the **cars** package to calculate the variance inflation variables (Multicollinearity essentials and vif, 2018).



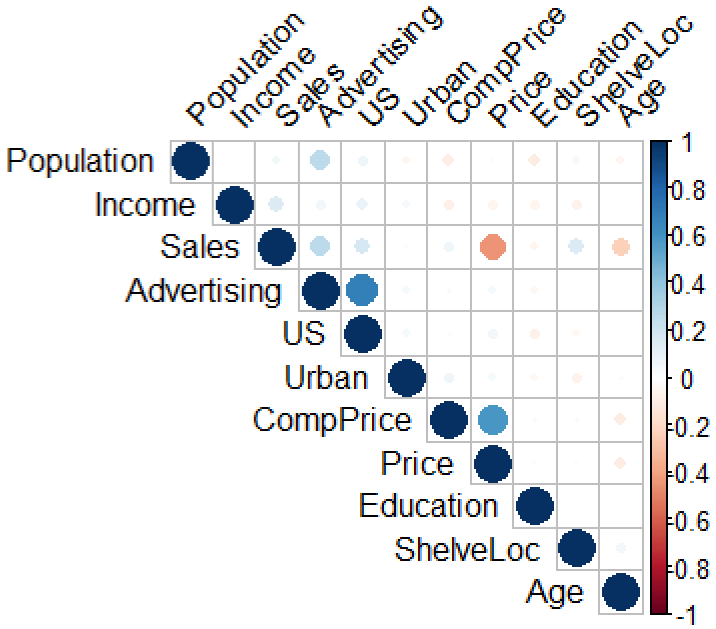
```
> vif(sales.model)
Advertising       Price         Age
   1.001987    1.012538    1.010550
> |
```

All of the variables have a value of ~1, meaning there is no multicollinearity. To verify this conclusion, we can also use a correlation matrix. We do this by using the **cor()** function.



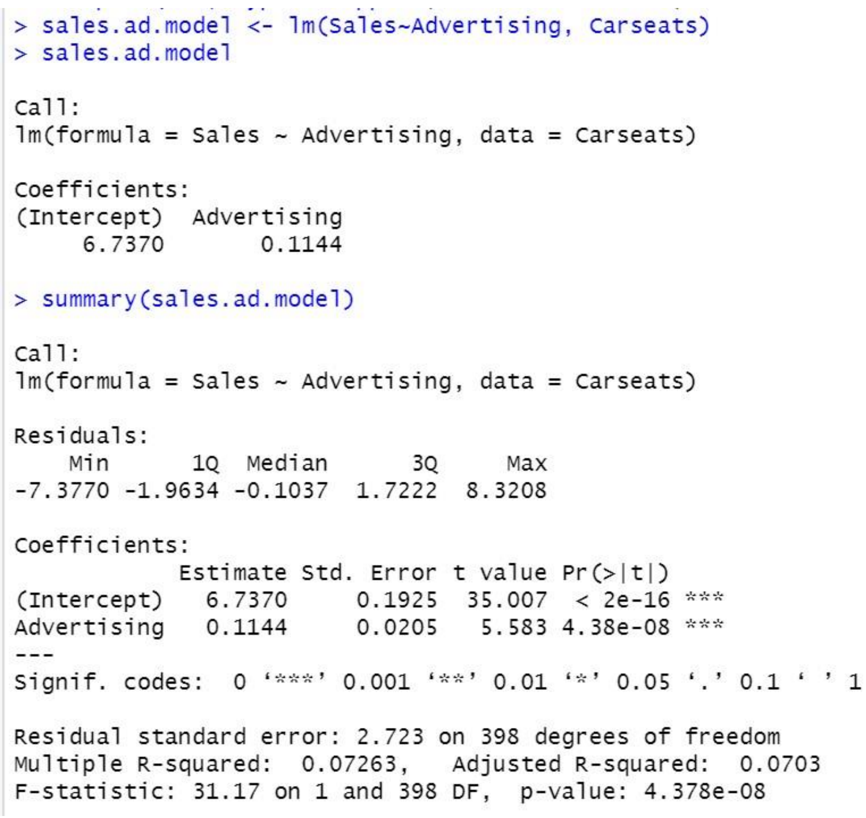
```
> round(res,3)
             Sales CompPrice Income Advertising Population  Price ShelveLoc    Age
Sales        1.000     0.064  0.152       0.270      0.050 -0.445     0.157 -0.232
CompPrice    0.064     1.000 -0.081      -0.024     -0.095  0.585     0.023 -0.100
Income       0.152    -0.081  1.000       0.059     -0.008 -0.057    -0.068 -0.005
Advertising  0.270    -0.024  0.059       1.000      0.266  0.045     0.009 -0.005
Population   0.050    -0.095 -0.008       0.266      1.000 -0.012    -0.045 -0.043
Price       -0.445     0.585 -0.057       0.045     -0.012  1.000     0.015 -0.102
ShelveLoc    0.157     0.023 -0.068       0.009     -0.045  0.015     1.000  0.056
Age         -0.232    -0.100 -0.005      -0.005     -0.043 -0.102     0.056  1.000
Education   -0.052     0.025 -0.057      -0.034     -0.106  0.012     0.001  0.006
Urban       -0.015     0.067  0.038       0.042     -0.052  0.047    -0.064  0.028
US           0.177     0.017  0.090       0.684      0.061  0.058    -0.040  0.009
            Education  Urban     US
Sales          -0.052 -0.015  0.177
CompPrice       0.025  0.067  0.017
Income         -0.057  0.038  0.090
Advertising    -0.034  0.042  0.684
Population     -0.106 -0.052  0.061
Price           0.012  0.047  0.058
ShelveLoc       0.001 -0.064 -0.040
Age             0.006  0.028  0.009
```

To make this matrix easier to interpret, we can create a correlogram using the **corrplot()** in the **corrplot** library. We can create this plot using the command **corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45),** where **res** is the correlation matrix above, **type = “upper”** displays only the upper triangle of the matrix, **order = “hclust”** reorders the matrix according to the correlation coefficient, **tl.col** defines the text label, and **tl.srt** rotates the text (Correlation matrix, n.d.).



A value of 1 indicates high correlation, which is not present in any of the variables above. Thus, multicollinearity does not exist in our model.

Now, we can test the significance of each explanatory variable. Recall, last week we found statistical significance between **Sales ~ Price**. Thus, the **Sales ~ Advertising** model and the **Sales ~ Age** model will be discussed below. First, we look at the **Sales ~ Advertising** model.



```
> sales.ad.model <- lm(Sales~Advertising, Carseats)
> sales.ad.model

Call:
lm(formula = Sales ~ Advertising, data = Carseats)

Coefficients:
(Intercept)  Advertising
     6.7370       0.1144

> summary(sales.ad.model)

Call:
lm(formula = Sales ~ Advertising, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3770 -1.9634 -0.1037  1.7222  8.3208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7370     0.1925  35.007  < 2e-16 ***
Advertising   0.1144     0.0205   5.583 4.38e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.723 on 398 degrees of freedom
Multiple R-squared:  0.07263,   Adjusted R-squared:  0.0703
F-statistic: 31.17 on 1 and 398 DF,  p-value: 4.378e-08
```

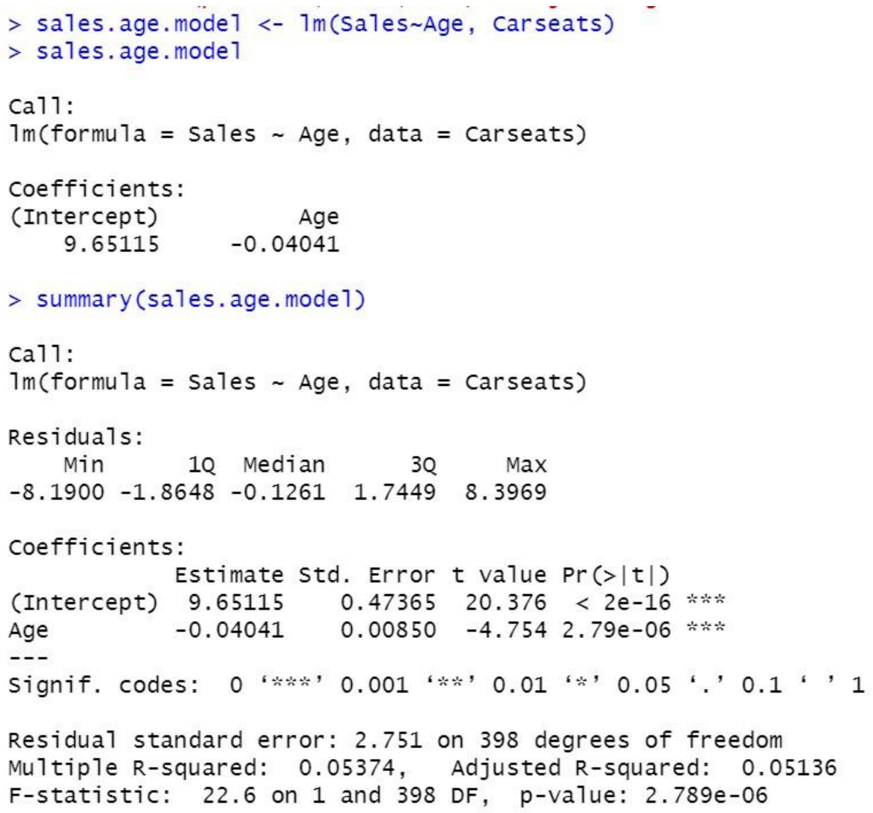
First, we define the null and alternative hypothesis.  If there is a significant linear relationship between the independent and dependent variable, the slope will not equal zero (Significance test for linear regression, n.d.). Thus,

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0.$$

From the summary above, we can see that our p-value is nearly zero, which indicates the model is statistically significant. The F-statistic is also high with an associated p-value that is near zero, which indicates there is a relationship between sales and advertising. Thus, our model indicates that it is statistically significant and we can reject the null hypothesis.

Next, we look at the **Sales ~ Age** model.



```
> sales.age.model <- lm(Sales~Age, Carseats)
> sales.age.model

Call:
lm(formula = Sales ~ Age, data = Carseats)

Coefficients:
(Intercept)          Age  
    9.65115     -0.04041  

> summary(sales.age.model)

Call:
lm(formula = Sales ~ Age, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1900 -1.8648 -0.1261  1.7449  8.3969 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.65115    0.47365  20.376  < 2e-16 ***
Age         -0.04041    0.00850  -4.754 2.79e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.751 on 398 degrees of freedom
Multiple R-squared:  0.05374,	Adjusted R-squared:  0.05136 
F-statistic:  22.6 on 1 and 398 DF,  p-value: 2.789e-06
```

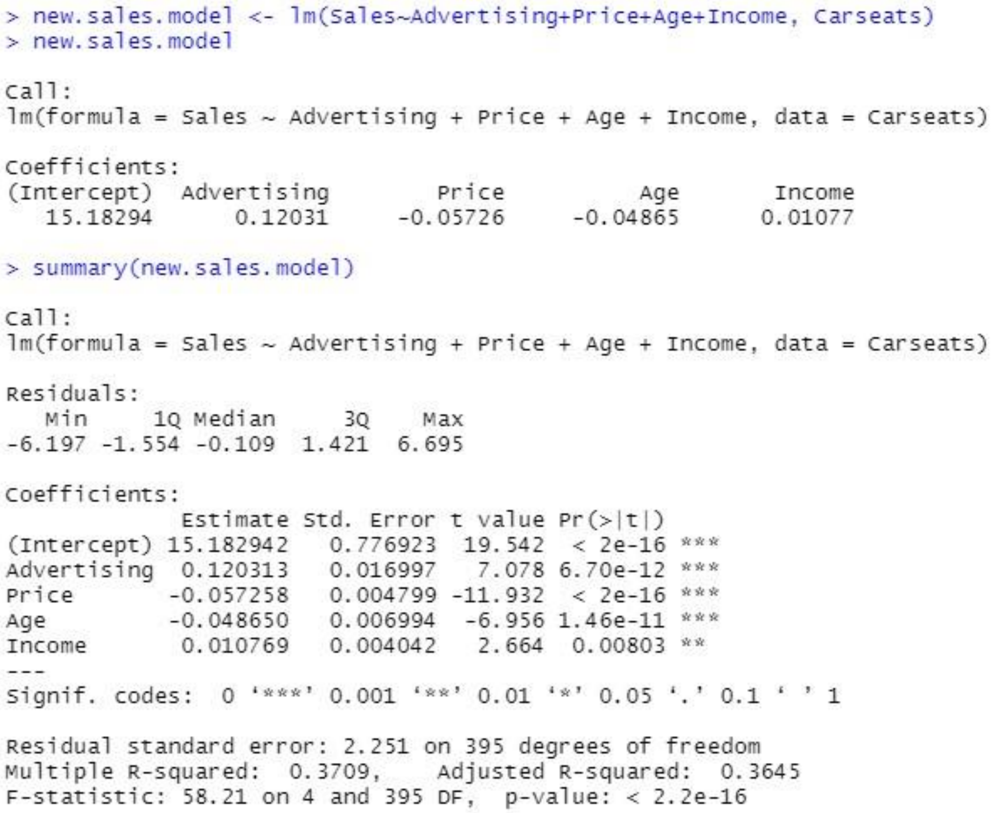
First, we define the null and alternative hypothesis. Thus,

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0.$$

From the summary above, we can see that our p-value is nearly zero, which indicates the model is statistically significant. The F-statistic is also high with an associated p-value that is near zero, which indicates there is a relationship between sales and age. Thus, our model indicates that it is statistically significant  and we can reject the null hypothesis. However, both the **Sales ~ Advertising** model and the **Sales ~ Age** model have a small R-squared value which indicates that both models explain virtually none of the variability.

To expand on the **sales.model**, we can add or remove predictors. We will add the variable **Income** to the model and analyze the new model. Similar to above, we create our new model, called **new.sales.model**, with the predictor variable **Income**.



```
> new.sales.model <- lm(Sales~Advertising+Price+Age+Income, Carseats)
> new.sales.model

Call:
lm(formula = Sales ~ Advertising + Price + Age + Income, data = Carseats)

Coefficients:
(Intercept)  Advertising        Price          Age       Income
   15.18294      0.12031     -0.05726     -0.04865      0.01077

> summary(new.sales.model)

Call:
lm(formula = Sales ~ Advertising + Price + Age + Income, data = Carseats)

Residuals:
   Min     1Q Median     3Q    Max
-6.197 -1.554 -0.109  1.421  6.695

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.182942   0.776923  19.542  < 2e-16 ***
Advertising  0.120313   0.016997   7.078 6.70e-12 ***
Price       -0.057258   0.004799 -11.932  < 2e-16 ***
Age         -0.048650   0.006994  -6.956 1.46e-11 ***
Income       0.010769   0.004042   2.664  0.00803 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.251 on 395 degrees of freedom
Multiple R-squared:  0.3709,    Adjusted R-squared:  0.3645
F-statistic: 58.21 on 4 and 395 DF,  p-value: < 2.2e-16
```

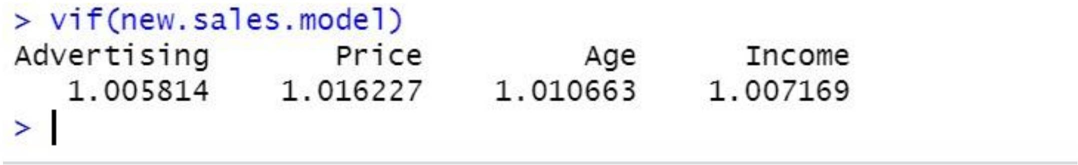
From this output, we can determine our coefficients for our multiple linear regression formula. Our formula is

$$\begin{aligned} Sales \; = \; & 15.18294 + 0.12031^{*}Advertising - 0.05726^{*}Price - 0.04865^{*}Age \\ & + 0.01077^{*}Income. \end{aligned}$$

This equation tells us that the predicted number of sales of car seats will increase as the amount of advertising increases, the price of car seats decreases, the age of the buyer decreases, and when the income of the buyer increases. Following the same steps as above to check for statistical significance, we come to the following conclusions:

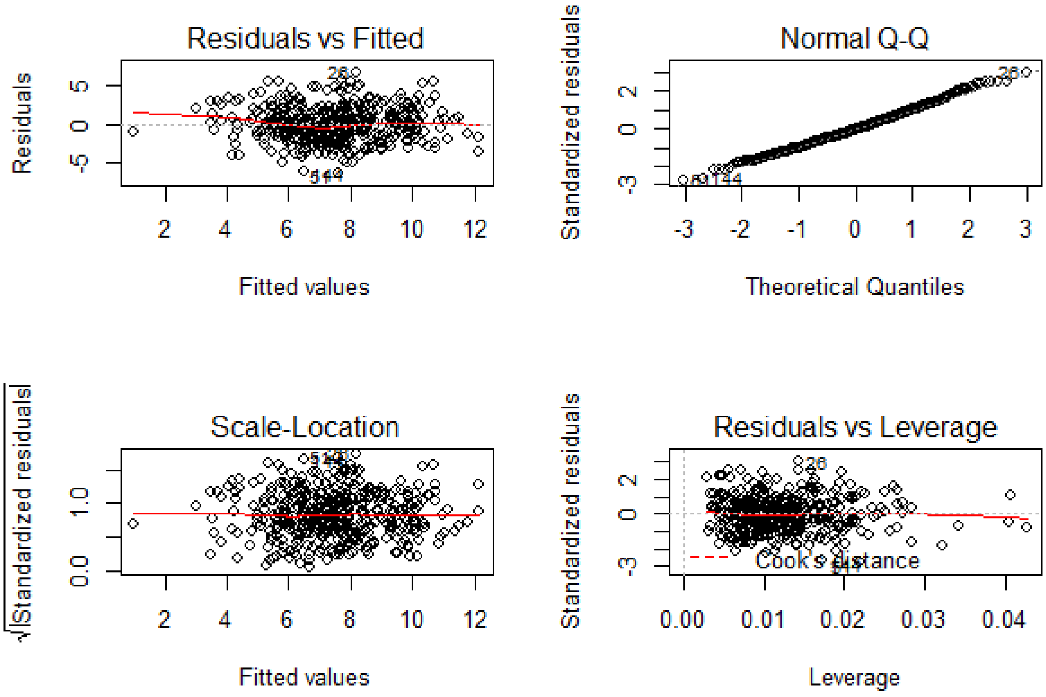
1. The F-statistic is less than the F-statistic in our **sales.model** (74.1), but it is still far from 1, thus we have a relationship between our predictors and the response variable.
2. The p-value associated with the F-statistic is < 2.2e-16, which is highly significant since it is less than 0.05, which indicates that at least one of the predictor variables is significantly related to the response variable.
3. All the coefficients t-statistics are from 0, indicating there is significant association between the predictors and the outcome variable.
4. Our R-squared value is a 0.3709, which indicates that 37.09% of the variance can be explained in this model, which is slightly better than our previous model.

Checking for multicollinearity, we use the **vif()** function.



```
> vif(new.sales.model)
Advertising       Price         Age      Income 
   1.005814    1.016227    1.010663    1.007169 
> |
```

All of the values are ~1, so we can conclude that there is no multicollinearity. Now, we can check our other multiple linear assumptions, following similar steps in the previous multiple linear regression model.



Similar to above, we see that we have a linear relationship (Residuals vs Fitted plot), the residuals are normally distributed (Normal Q-Q plot), and we have homoscedasticity (Scale-Location plot). Therefore, our **new.sales.model** is statistically significant.

As we have seen above, we have created two multiple linear regression models that were statistically significant. We used several methods to test for significance. To further test for statistical significance, we tested the significance of each explanatory variable. After testing the models for statistical significance, we conducted different tests to detect multicollinearity. After ensuring that multicollinearity doesn't exist, we checked that the other three multiple linear regression assumptions exist: linearity, normally distributed residuals, and homoscedasticity.

## Resources

Assumptions of multiple linear regression. (n.d.). *Statistics Solutions*. Retrieved March 28, 2020, from https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

Bock, T. (2018, April 6). What are variance inflation factors (Vifs)? | displayr. Com. *Displayr*. https://www.displayr.com/variance-inflation-factors-vifs/

*Correlation matrix*. (n.d.). Retrieved March 28, 2020, from http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

Kenton, W. (2019, April 14). *How multiple linear regression works*. Investopedia. https://www.investopedia.com/terms/m/mlr.asp

*Linear regression assumptions and diagnostics* . (2018, November 3). http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

*Multicollinearity essentials and vif* . (2018, November 3). http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/

*Multiple linear regression*. (2018, October 3). http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

Quick, J. M. (2009, December 8). R tutorial series: Multiple linear regression. *R-Bloggers*. https://www.r-bloggers.com/r-tutorial-series-multiple-linear-regression/

Rego, F. (2015, October 23). *Quick guide: Interpreting simple linear model output in r*. https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R

*Significance test for linear regression*. (n.d.). Retrieved March 22, 2020, from http://www.r-tutor.com/elementary-statistics/simple-linear-regression/significance-test-linear-regression

Stephanie. (2015, September 22). *Multicollinearity: Definition, causes, examples*. Statistics How To. https://www.statisticshowto.datasciencecentral.com/multicollinearity/