Simple Linear Regression

Taylor Shrode

3/22/2020

MSDS 660

The purpose of linear regression is to "model a continuous variable, *y*, as a mathematical function of one or more *x* variable(s), so that we can use this regression model to predict the *y* when only the *x* is known (Prabhakaran, 2016)." This mathematical equation can be generalized as

$$Y = \beta_1 + \beta_2 X$$

where β_1 is the intercept and β_2 is the slope (Prabhakaran, 2016).

Simple linear regression is "is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables (What is simple linear regression, n.d.)." One variable, denoted *x*, is considered the predictor or the independent variable and the other variable, denoted *y*, is considered the response or the dependent variable (What is simple linear regression, n.d.).

For this assignment, we will be performing simple linear regression using a dataset called *Carseats* located in the **ISLR** library. The *Carseats* dataset contains sales of child car seats at 400 different stores, on the following 11 variables (R: sales of child car seats, n.d.):

1. Sales: Unit Sales (in thousands) at each location
2. CompPrice: Price charged by competitor at each location
3. Income: Community income level (in thousands of dollars)
4. Advertising: Local advertising budget for company at each location (in thousands of dollars)
5. Population: Population size in region (in thousands)
6. Price: Price company charges for car seats at each site
7. ShelveLoc: A factor with levels *Bad*, *Good* and *Medium* indicating the quality of the shelving location for the car seats at each site
8. Age: Average age of the local population
9. Education: Education level at each location
10. Urban: A factor with levels *No* and *Yes* to indicate whether the store is in an urban or rural location
11. US: A factor with levels *No* and *Yes* to indicate whether the store is in the US or not

To begin, we need to install the **ISLR** package, load the library, and then load the *Carseats* dataset. To ensure the data was loaded properly, we use the **View()** command to view the data.

```
> install.packages("ISLR")
WARNING: Rtools is required to build R packages but is not currently installed. Ple
ase download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/07hoc/OneDrive/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/ISLR_1.2.zip'
Content type 'application/zip' length 2924248 bytes (2.8 MB)
downloaded 2.8 MB

package 'ISLR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\07hoc\AppData\Local\Temp\RtmpYf8vbg\downloaded_packages
> library(ISLR)
> data("Carseats")
> View(Carseats)
```

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

Carseats

Filter

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

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 ...
```

We can see that our dataset is a data.frame. Now, we can begin performing data exploration by using graphical analysis techniques. For this assignment, we aim to build a simple linear regression model to predict price (Price) by establishing a statistically significant relationship with the amount of sales (Sales) (Prabhakaran, 2016). We can view the linear relationship between these variables by constructing a scatter plot. The scatter plot below was created using the command **plot(sales, advertising): plot(Carseats$Sales, Carseats$Price, main = 'Sales vs Advertising', xlab = 'Sales', ylab = 'Price').**

When performing linear regression, simple or multiple, we aim to establish a linear relationship between the predictor variable and the response variable (Prabhakaran, 2016). As we can see from the plot above, we can visualize the linear relationship between the predictor (Price) and the response (Sales). This plot shows a linearly decreasing relationship between the two variables. In other words, we can see that there is a decrease in sales of car seats as the price of the car seat increases. Using this conclusion from the visual above, we should be able to predict the Price of a car seat based on how many Sales there are.

Another visualization technique is producing a matrix of scatter plots. The matrix plot below looks at the scatterplots of Sales, Age, Income, and Price and groups the points by ShelveLoc. We can group the points by US, ShelveLoc, or Urban. For the purpose of this example, we group the points by ShelveLoc. The purpose of the matrix of scatter plots is to visualize multiple scatter plots on a single view to relationships of specified variables (Scatter plot matrices, n.d.). The base function **pairs()** is used, and since we are grouping points on the scatter plot, we need to specify the colors of each group of points. This can be done by using the command **my_cols <- c('green', 'darkorchid', 'firebrick1').** Next, we create the matrix of scatter plots with specified variables, without the lower panel, using the command **pairs(~Sales+Age+Income+Price, data = Carseats, lower.panel = NULL, col = my_cols[Carseats$ShelveLoc], pch = 20, cex = .5)**

As we can see above, of the few variables we chose to analyze, the only set of variables that have a clear linear relationship is Sales and Price. Next, we need to perform a correlation test between two variables. To do this, we compute a correlation matrix which is used to investigate the dependence between multiple variables at the same time (Correlation matrix, n.d.). The results is a table containing the correlation coefficients between each variable. The correlation coefficient value is always between +1 and -1. A -1 value indicates a perfect negative linear relationship while a +1 indicates a perfect positive linear relationship (Rumsey, n.d.). A 0 indicates that no linear relationship exists. To compute a correlation matrix, all data must be numeric. Thus, the variables Urban, ShelveLoc, and US need to be converted. This can be done using the commands below.

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

To view the correlation matrix of the whole *Carseats* dataset, we use the command **cor(Carseats).**

```
> cor(Carseats)
                Sales   CompPrice       Income  Advertising   Population
Sales       1.00000000  0.06407873  0.151950979  0.269506781  0.050470984
CompPrice   0.06407873  1.00000000 -0.080653423 -0.024198788 -0.094706516
Income      0.15195098 -0.08065342  1.000000000  0.058994706 -0.007876994
Advertising 0.26950678 -0.02419879  0.058994706  1.000000000  0.265652145
Population  0.05047098 -0.09470652 -0.007876994  0.265652145  1.000000000
Price      -0.44495073  0.58484777 -0.056698202  0.044536874 -0.012143620
ShelveLoc   0.15746968  0.02334991 -0.067677735  0.008543504 -0.044771998
Age        -0.23181544 -0.10023882 -0.004670094 -0.004557497 -0.042663355
Education  -0.05195524  0.02519705 -0.056855422 -0.033594307 -0.106378231
Urban      -0.01541944  0.06659440  0.037967176  0.042034534 -0.052024635
US          0.17709327  0.01686887  0.089601328  0.684460204  0.060563613
                Price    ShelveLoc           Age    Education       Urban
Sales      -0.44495073  0.157469677 -0.231815440 -0.051955242 -0.01541944
CompPrice   0.58484777  0.023349910 -0.100238817  0.025197050  0.06659440
Income     -0.05669820 -0.067677735 -0.004670094 -0.056855422  0.03796718
Advertising 0.04453687  0.008543504 -0.004557497 -0.033594307  0.04203453
Population -0.01214362 -0.044771998 -0.042663355 -0.106378231 -0.05202463
Price       1.00000000  0.014632653 -0.102176839  0.011746599  0.04701613
ShelveLoc   0.01463265  1.000000000  0.056488485  0.001491726 -0.06397765
Age        -0.10217684  0.056488485  1.000000000  0.006488032  0.02847860
Education   0.01174660  0.001491726  0.006488032  1.000000000 -0.03309377
Urban       0.04701613 -0.063977646  0.028478596 -0.033093770  1.00000000
US          0.05786126 -0.039759746  0.008652076 -0.078250434  0.04708496
                  US
Sales       0.177093268
CompPrice   0.016868865
Income      0.089601328
Advertising 0.684460204
Population  0.060563613
Price       0.057861264
ShelveLoc  -0.039759746
Age         0.008652076
Education  -0.078250434
Urban       0.047084963
US          1.000000000
> |
```

To simplify this matrix, we save the output in a variable called **res** and then we can use the **round()** function to round to a specified number of decimal places. Below, we round to 3 decimal places.

```
> 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
Education       1.000 -0.033 -0.078
Urban          -0.033  1.000  0.047
US             -0.078  0.047  1.000
>
```

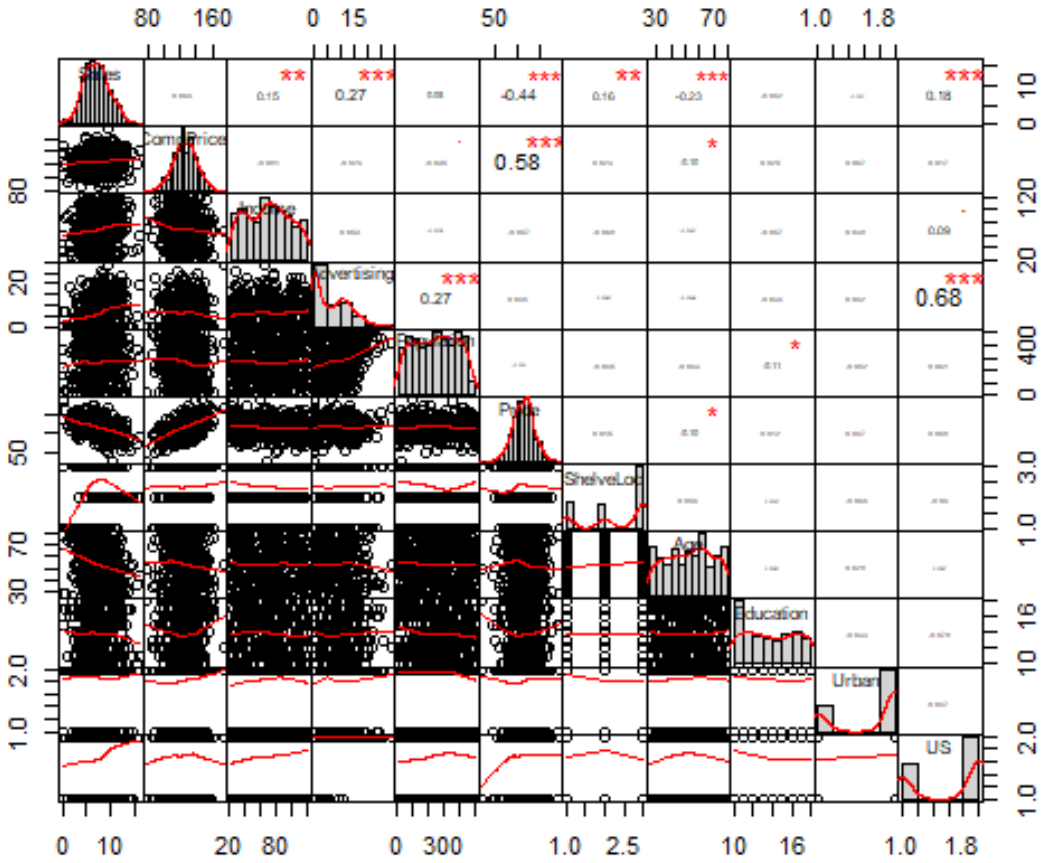
To further simplify the correlation matrix, we can use the **dplyr** library to use the **select()** function which allows us to select which variables we want in our correlation matrix. Below, we include Sales, Age, Income, and Price and round the coefficients to 3 decimal places.

```
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> res2 <- cor(select(Carseats,Sales,Age,Income,Price))
> round(res2, 3)
        Sales    Age Income  Price
Sales   1.000 -0.232  0.152 -0.445
Age    -0.232  1.000 -0.005 -0.102
Income  0.152 -0.005  1.000 -0.057
Price  -0.445 -0.102 -0.057  1.000
>
```

As we can see in our correlation matrix above, the strongest linear relationship is located in the Sales-Price pair. With a correlation coefficient of -0.445, this indicates a moderate negative relationship, which we also saw in the scatter plots above (Rumsey, n.d.).

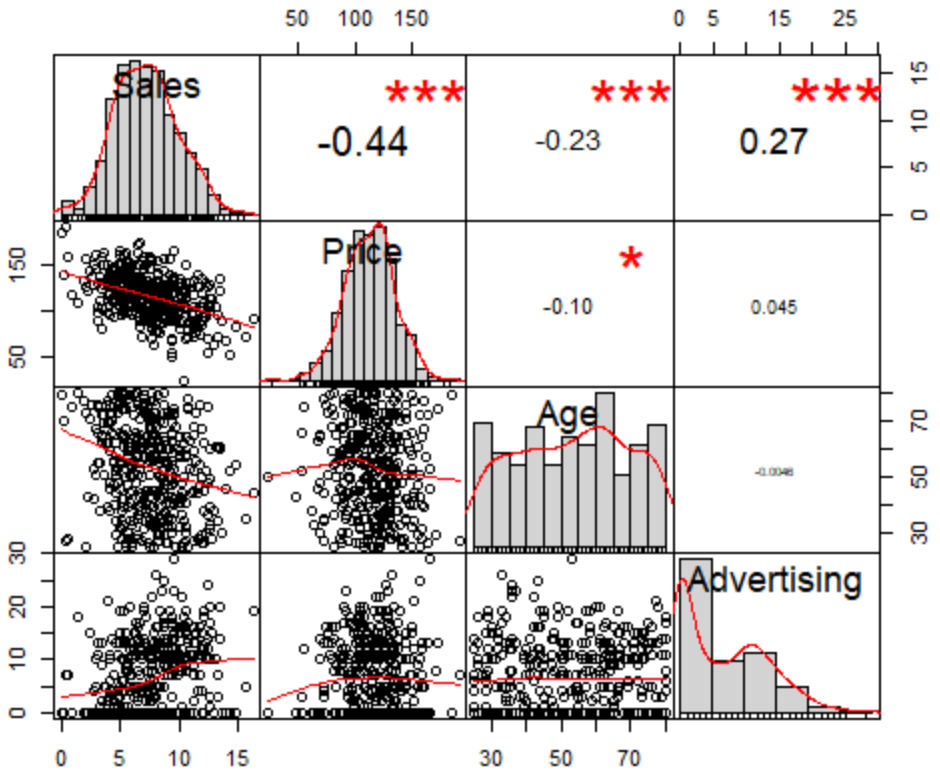
We can also use the function **chart.correlation,** located in the package **PerformanceAnalytics,** to display a chart of the correlation matrix (Correlation matrix, n.d.). This chart includes the distribution of each variable (on the diagonal), the scatterplots with fitted lines (bottom of diagonal), and the value of the correlation coefficients with significance levels as stars (top of

diagonal) (Correlation matrix, n.d.). Below is the correlation chart for the entire dataset, which is created using the command **chart.Correlation(Carseats, histogram=TRUE,pch=19)**.



To simplify this chart, we can specify the variables we want included. Below is the correlation chart for Sales, Age, Income, and Price, which is created with the following commands:

1. data <- Carseats[,c(1,6,8,4)] (to specify variables)
2. chart.Correlation(data, histogram=TRUE,pch=19)



Now that we have seen the linear relationship between our predictor variable (Price) and the response variable (Sales) using scatter plots and computing correlation, we can begin building our linear model.

First, we need to construct a null and alternative hypothesis to test. 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,

$$H_0: \beta_1 = 0$$

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

The **lm()** function is used to build linear models. It takes in two arguments: Formula and Data. The function looks like **lm(YVAR ~ XVAR), dataset),** where **YVAR** is the predictor and **XVAR** is response (Quick, 2009). Applying this function to our variables, Sales and Price, we can determine our regression coefficients (intercept and slope) which is used in the mathematical equation mentioned above.

```
> lm(Sales~Price, Carseats)

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

Coefficients:
(Intercept)        Price
   13.64192     -0.05307
```

From this, we can construct our mathematical equation:

$$Y = \beta_1 + \beta_2 X$$

$$Sales = 13.64192 - 0.05307 * Price.$$

This equation tells us that the predicted number of sales will decrease by -0.05307 as price increases. Before we use this model, we need to ensure that our model is statistically significant. To do this, we begin by printing the summary of our model, as seen below.

```
> lm_model <- lm(Sales~Price, Carseats)
> abline(lm(Sales~Price, Carseats), col = 'red')
> par(mfrow=c(2,2))
> plot(lm_model)
> summary(lm_model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-6.5224 -1.8442 -0.1459  1.6503  7.5108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.641915   0.632812  21.558   <2e-16 ***
Price       -0.053073   0.005354  -9.912   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.532 on 398 degrees of freedom
Multiple R-squared:  0.198,     Adjusted R-squared:  0.196
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16

>
```

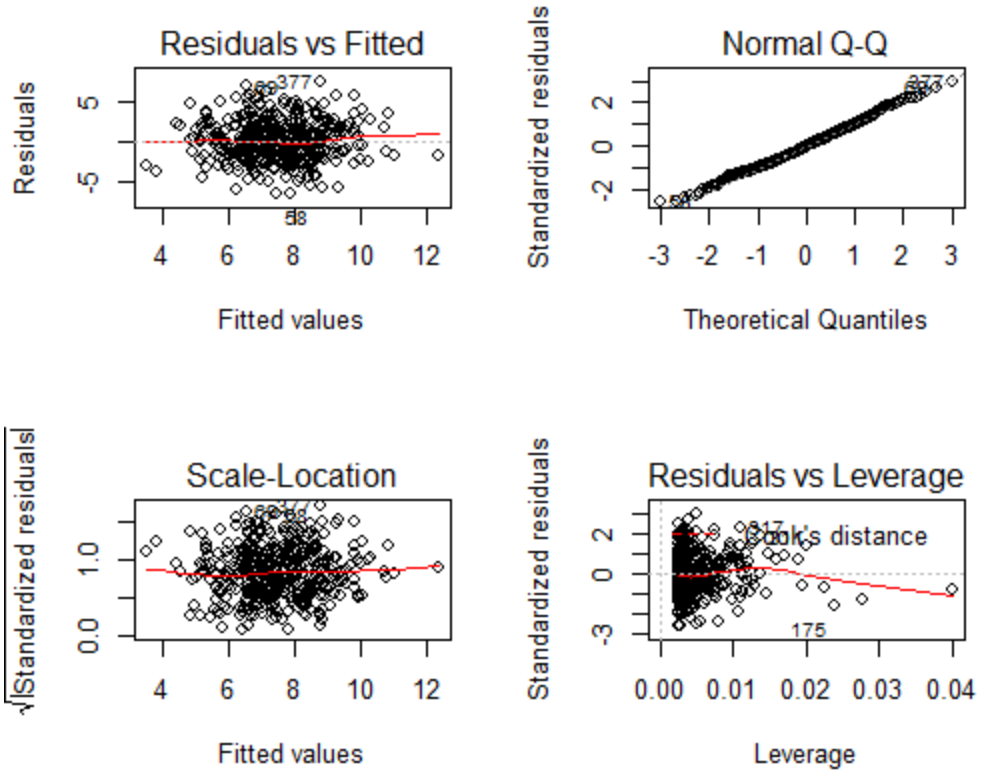
We can see that the summary shows the model formula, residual quartiles, coefficient estimate with standard error, and a significance test, multiple and adjusted R-square, and F-test for model fit (Rago, 2015). A description of these components are located below (Rago, 2015).

1. The model formula is our **lm_model** formula defined above.
2. The residual quartiles are defined as the difference between the actual observed response values and response values that the model predicted.
3. In simple linear regression, the coefficients are the two unknown constants that represent the slope and intercept in the linear model.
    a. Coefficient-Estimate

        i. For the intercept, is the expected value of *Sales* required for a store when considering the average *Price* of all the car seats in the dataset.
        ii. For the slope, is the effect *Price* has on *Sales*.
    b. Coefficient-Standard Error measures the average amount that the coefficient estimates vary from the average value of our *Sales* variable. The lower the number, the better.
    c. Coefficient-*t* value is a measure of how many standard deviations our coefficient estimate is far away from 0. The farther away from 0 indicates we could reject the null hypothesis.
    d. Coefficient-Pr(>|*t*|), or p-value, indicates the probability of observing any value equal or larger than *t*.
        i. Note the 'signif. Codes' associated with each estimate. Three asterisks represent a highly significant p-value.
4. Residual Standard of error is a measure of the quality of a linear regression fit. This is the average amount our response variable will deviate from the true regression line.
5. R-Squared provides a measure of how well the model is fitting the actual data. This is a measure of the linear relationship between the predictor and the response.
6. The F-statistic determines whether there is a relationship between the predictor and the response. The further the F-statistic is from 1, the better.

From the summary of our model above, we can see that our p-value is nearly zero, which is much less than 0.05, and also has 3 stars which indicates high significance. This means that it is unlikely that no relationship exists between Sales and Price. We can also see that the F-statistic is very high, also indicating that a relationship between Sales and Price exists. However, the R-squared and Adj R-squared values are small, which indicates our model explains virtually none of the variability.

Now, we need to plot the model. By plotting our model, we will produce 4 charts: residuals versus fitted, normal-QQ, Scale-Location, and Residuals versus Leverage. The values in the residuals versus fitted plot should be uncorrelated (Chouldehova, n.d.). In the Normal Q-Q plot, the points should form a line that is roughly straight. The Scale-Location is another version of the residuals vs fitted plot, so there should also be no trend in this plot (Chouldehova, n.d.). Finally, the Residuals vs Leverage measures how much an observation is influenced by the model fit.



We can see we have a relatively straight line in our Q-Q plot and both of the residuals plots show no trends, which leads us to believe that the model is statistically significant. We can also use the plots to verify model assumptions (e.g. linearity, normality, variance, etc.) (Using plots to check model assumptions, n.d.). The Residuals vs Fitted plot is used to check the linear relationship (Linear regression assumption, 2018). As we can see above, we have a horizontal line which is an indication of a linear relationship. The Q-Q plot is used to determine if the residuals are normally distributed (Linear regression assumption, 2018). Above, we see that the residual points follow the dashed, straight line, which suggests that the residuals are normally distributed. The Scale-Location plot is used to check the homogeneity of variance of the residuals (Linear regression assumption, 2018). We see a horizontal line, which suggests homogeneity of variance.

Therefore, with the information above, we can reject the null hypothesis and can conclude that the model is statistically significant.

Resources

Chouldechova, A. (n.d.). *Linear regression in R*. Retrieved March 22, 2020, from https://www.andrew.cmu.edu/user/achoulde/94842/lectures/lecture09/lecture09-94842.html

Prabhakaran, S. (2016). *Linear regression with r*. http://r-statistics.co/Linear-Regression.html

Correlation matrix. (n.d.). Retrieved March 22, 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

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

Quick, J. M. (2009, November 26). R tutorial series: Simple linear regression. *R-Bloggers*. https://www.r-bloggers.com/r-tutorial-series-simple-linear-regression/

R: sales of child car seats. (n.d.). Retrieved March 22, 2020, from https://vincentarelbundock.github.io/Rdatasets/doc/ISLR/Carseats.html

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

Rumsey, D. J. (n.d.). *How to interpret a correlation coefficient r*. Retrieved March 22, 2020, from https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/

Scatter plot matrices . (n.d.). Retrieved March 22, 2020, from http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

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

Using plots to check model assumptions. (n.d.). Retrieved March 22, 2020, from https://web.ma.utexas.edu/users/mks/statmistakes/modelcheckingplots.html

What is simple linear regression? . (n.d.). Retrieved March 22, 2020, from https://online.stat.psu.edu/stat462/node/91/