Data Analysis Assignment Lesson 2

Maureen Renaud

5/29/2020

INTRODUCTION:

With nearly 4 million babies born in the US every year[1], the baby care industry has a massive and constantly changing customer base to which products are marketed. Through targetting these parents, eager to protect and care for their new baby, the global industry is expected to reach 16.78 billion by 2025 according to Grandview Research[2].

While many of the marketed products are unnecessary, there are many things considered essential. Car seats are one of these must-haves, as they are required to even take your baby home from the hospital. Car seat sales reached 4 billion in 2018 and are expected to reach nearly 6 billion dollars by 2024 according to Market Watch[3]. In order to capitalize on this market share, car seat manufacturing companies need to take a close look at the data in order to determine where to focus their marketing efforts.

This particular study looked at the model:

carseats.lm = lm(Sales~ Price + Urban + US, data=Carseats)

In this model, Price is the price the company charges for car seats at each site, Urban is a factor with levels No and Yes to indicate whether the store is in an Urban or Rural location, US is a factor with levels No and Yes to indicate whether the store is in the US or not, and Sales is unit sales (in thousands) at each location.

After running a linear regression, it was found that regression coefficients for Price and having a US based store were not equal to zero and thus there is a relationship between these variables and Sales. We could not conclude that the regression coefficient for having a store located in an Urban versus Rural area was not zero so we cannot infer a relationship between this variable and Sales.

DATA

The data for this study was obtained from the Carseats dataset in the ISLR library and consists of 400 observations of 11 variables.

Doing a check for missing values shows us that this is a complete data set:

```
sum(is.na(Carseats))
```

```
## [1] 0
```

Below is a summary of the variables considered in this study:

```
vars <- c("Sales","Price","Urban","US")
new.carseats <- Carseats[vars]
summary(new.carseats)
```

```
##      Sales            Price       Urban       US
##  Min.   : 0.000   Min.   : 24.0   No :118   No :142
##  1st Qu.: 5.390   1st Qu.:100.0   Yes:282   Yes:258
##  Median : 7.490   Median :117.0
##  Mean   : 7.496   Mean   :115.8
##  3rd Qu.: 9.320   3rd Qu.:131.0
##  Max.   :16.270   Max.   :191.0
```

ANALYSES

I chose to run a linear regression because according to the text *An Introduction To Statistical Learning*[4], a linear model is a simple yet useful tool for predicting a quantitative response. In this case, we are interested in the quantitative response of Sales. First, our linear model:

```
new.carseats.lm = lm(Sales~ Price + Urban + US, data=new.carseats)
summary(new.carseats.lm)
```

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

This summary provides several vauable pieces of information.

First, extracting the coefficients for each variable gives us the equation:

Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

This means that for every five cent decrease in price, one thousand additional car seats were sold. A store located in an urban area will see a 20 unit decrease in sales and a store located in the United States will see a 1020 unit increase in sales.

With p-values near zero, we can reject the null hypothesis and conclude that regression coefficient for Price and US location are not zero and are therefore related to Sales. Based on a p-value of 0.936, we fail to reject the null hypothesis and we cannot conclude that the regression coefficient for a Urban vs. Rural store is not equal to zero. Therefore, this variable is not linearly related to Sales.

We can safely remove the Urban predictor from the model and examine the reduced model with only Price and US location as predictor variables:

```
reduced.carseats.lm = lm(Sales~ Price + US, data=new.carseats)
summary(reduced.carseats.lm)
```

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

I continue to reject the null hypothesis and conclude that the regrssion coefficient for Price and US location are not zero. A scatterplot of this data, provided in *Figure 1* clearly shows that decreasing price and US based stores are associated with increased Sales.

In order to determine how well each of these two models fit the data, I will first consider the R-sqared value. Both models have an R-squared value of 0.2393. We know that R-squared increases when more variables are added. So, the fact that there is no difference in the R-squared values between the two models indicates that including Urban Location in the model adds no venefit.

I will next consider the Residual Standard Error (RSE), which is the average amount the resopnse will deviate from the regression line. The full model has an RSE of 2.472 and the reduced model has an RSE of 2.469. So, the reduced model is a slightly better fit despite having fewer variables.

Given that the R-squared values are the same for each model, the RSE is slightly better in the reduced model, and the reduced model is simpler with two variables instead of three, the reduced model is the logical choice as the best fit for the data.

Using the reduced model, the 95% confidence intervals for the coefficients are as follows:

```
confint(reduced.carseats.lm, level = 0.95)
```

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

Note that the confidence interval for both Price and US location do not contain zero. This is another clear indicator that the regresssion coefficients are not equal to zero and there is a relationship between these variables and Sales.

Finally, to check for outliers and high leverage points.

We first analyze the residuals plots, shown in *Figure 2*. A visual scan of the plots show no obvious abnormalities. In particular, looking at the Residuals Vs. Leverage plot, we look for points in the lower or upper right hand corners or points outside the Cook's Distance dashed line. The points are so tightly clustered, that the Cook's Distance dashed line doesn't even appear in the plot. Another check is to determine if the Cook's Distance is high for any point. Using a reference value of one, it appears that there are no points with a Cook's Distance greater than one. Therefore, I feel safe concluding that there are no outliers or high leverage points in this data set.

```
which(cooks.distance(reduced.carseats.lm) > 1)
```

```
## named integer(0)
```

PLOTS AND TABLES

```
library(ggplot2)
scatter.plot <- ggplot(new.carseats, aes(x = Price, y = Sales))

scatter.plot + geom_point(aes(shape = US, color = US)) +
  geom_smooth(aes(color = US, fill = US), method = lm) +
  labs(title="Carseat Sales By Price Based on Location",
       x ="Price", y = "Unit Sales (In Thousands)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Figure 1

```
par(mfrow=c(2,2))
plot(reduced.carseats.lm)
```

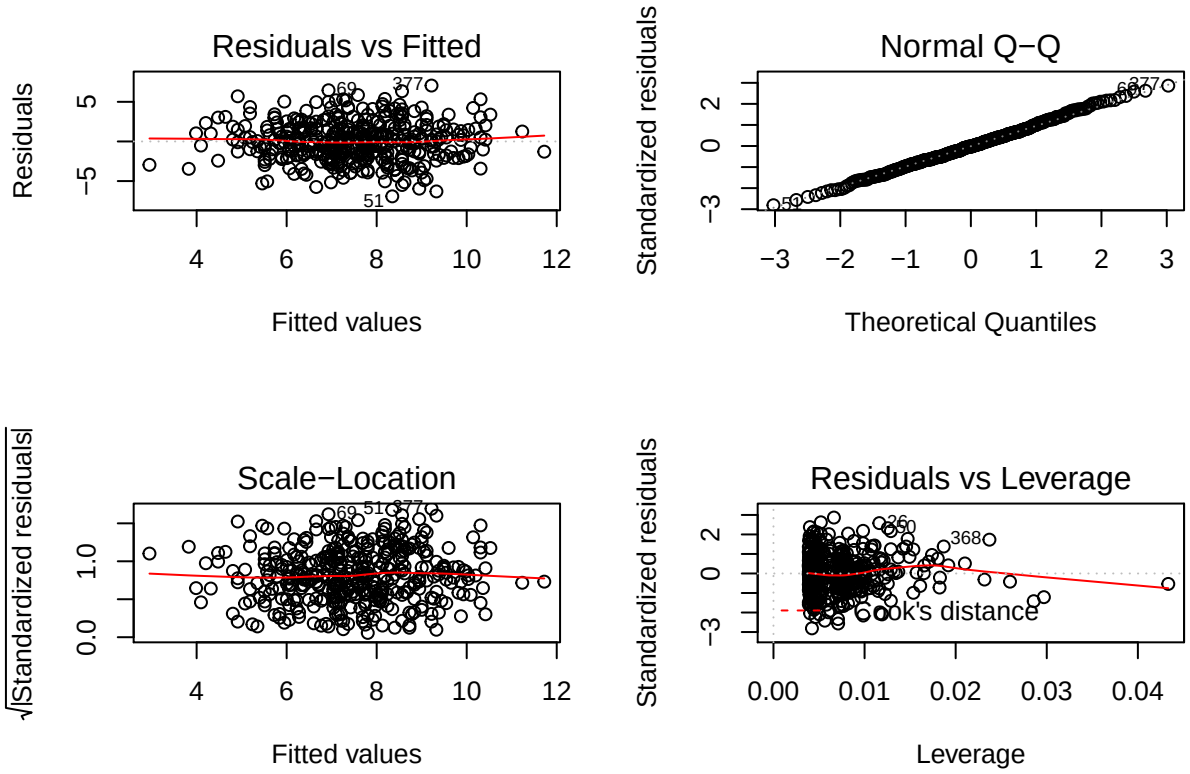

Figure 2

CONCLUSIONS

Based on analyzing the relationship between the predictor variables Price, Urban vs. Rural location, and US vs global location and the response variable Sales, we see that regression coefficients for Price and US vs. global location are not zero and therefore there is a relationship between these variables and sales. We cannot conclude that the coefficient variable for Urban vs. Rural is not zero and therefore we can conclude that location is not related to Sales.

Specifically, having a lower Price and a US based location are related with having a greater Sales volume.

With this knowledge,companies looking to maximize their sales and share of the baby car seat market should focus on minimizing their prices and on their US based store locations.

SOURCES

1) https://www.cdc.gov/nchs/fastats/births.htm

2) https://www.grandviewresearch.com/press-release/global-baby-products-market

3) https://www.marketwatch.com/press-release/baby-car-seat-market-size-share-2020-by-product-types-and-application-top-manufacturer-regional-analysis-and-forecasts-to-2024-says-market-reports-world-2020-03-11

4) James, G et al. (2017). *An Introduction To Statistical Learning.* New York, NY.