DATA621 Homework 1 - 1970 Cigarette Sales Analysis

Erik Nylander

February 2, 2016

1 Data Exploration

In this analysis, we use the 1970 cigarette usage in the United States study. The data has been broken into a training data set containing data on 45 states and the District of Columbia. The remaining 5 states have been placed in an evaluation data set that we will use for prediction. We use the data to examine the effects of three predictor variables [Age, Income, Price(per pack)] on the per-capita Sales of packs of cigarettes in state.

1.1 Summary Statistics

Upon receiving the data, we began the analysis by looking at some basic summary statistics for the data. We first note that there was no missing data in the training or evaluation data. However, there is no Sales data in the evaluation set that we received so we will not be able to evaluate our final sales predictions versus that data set. In Table 1 we can see the summary statistics from our training data set. We note that there does appear to be evidence for a right skew in the sales data which could lead to issues in our regression modeling we will look at this further in the next section.

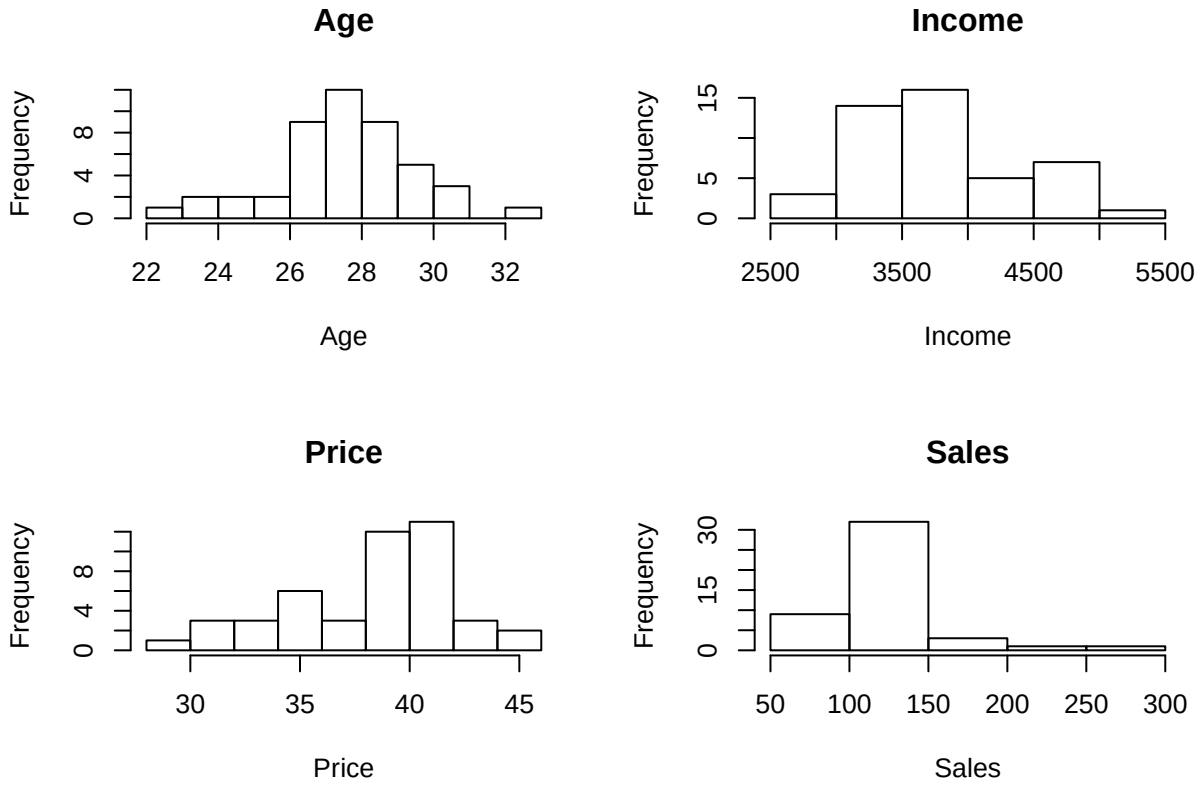
Table 1: Mean, Median, and Standard Deviations

	Mean	Median	Standard Deviation
Age	27.54	27.50	1.899
Income	3766	3744	597.626
Price	38.28	38.90	4.076
Sales	120	115.5	37.397

Finally, we explored the correlations between all of the different variables. The largest correlation value that we found was a correlation of 0.338 between Income and Sales. This does indicate that Income will likely be a significant variable in our final model.

1.2 Summary Plots

For the predictor variables we found by looking at the plots of the histograms that the data for the predictor variables, Age, Income, and Price seem to be normally distributed. However, we can see that there are some issues with the Sales data itself we seem to have some serious outliers that are causing a sever right skew to the data. We will attempt to clear these up below.

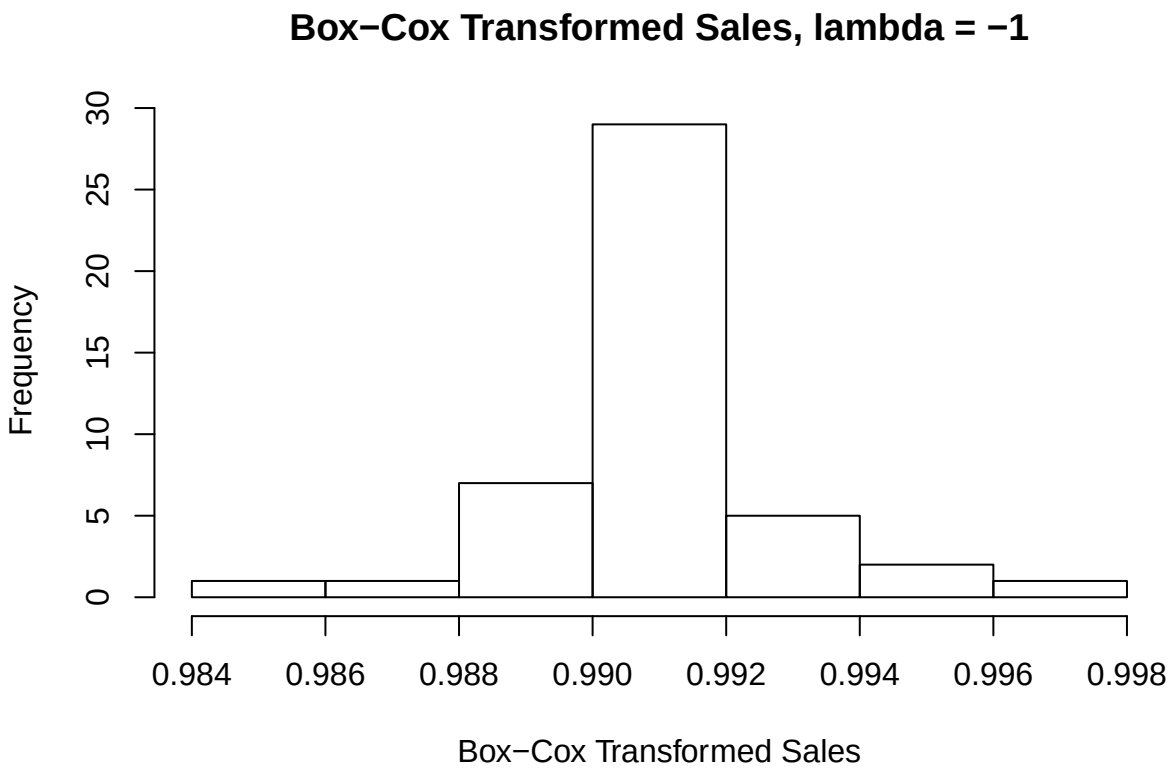


2 Data Preperation

After examining the data and determining that New Hampshire, Nevada, and the District of Columbia are causing the right skew of the data, we have decided to apply a transformation to the data to attempt help with the right skew to the data. We tried both log and square root transformations but found that they did little to normalize the data set and reduce the effect of right skew. Finally, we decided to perform a Box-Cox transformation. Using the *forecast()* package in R we found a λ value of -0.95. We round this value to -1 and apply the following transformation:

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda}$$

While this transformation did not eliminate the outliers it did improve the balance of the data and give us a mean = 0.9913 and median of 0.9912 thus we have helped to remove the skew and will move forward with the Box-Cox transformed data. We can see that under the Box-Cox transformation our data assumes a more normal distribution. There are still outliers to the data but we have balanced them on each side of the mean through the transformation.



3 Building Models

Using the training data that was given to us with a Box-Cox transformation on the Sales data, we will now start building linear regression models. For this analysis with three predictor values we will test all seven of the possible linear regression models.

3.1 All Variables Included

For our first model we will use all three of the predictor variables. We find all three of them to be significant (p-value $< .05$). The residual standard error (RSE) of the model was 0.0015 (42 degrees of freedom) with an R^2 value of 0.4054. Table 2 s hows the model coefficient estimates.

Table 2: Linear Regression with all Predictors

Intercept	Age	Income	Price
0.9827	0.0004085	0.00000111	-0.0001797

3.2 Two Variables Included

For our next groups of models we will just include two of the predictor variables.

Age and Income We find in this model that that both of these predictor values are significant (p-value < .05). The RSE of the model is 0.0016 (43 degrees of freedom) with R^2 value of 0.2625. Table 3 shows the model coefficient estimates.

Table 3: Linear Regression with Age and Income

Intercept	Age	Income
0.9785	0.0003259	0.00000099

Age and Price We find in this model that that both of these predictor values are significant (p-value < .05). The RSE of the model is 0.0016 (43 degrees of freedom) with R^2 value of 0.2887. Table 4 Shows the model coefficient estimates.

Table 4: Linear Regression with Age and Price

Intercept	Age	Price
0.9841	0.0004879	-0.0001646

Price and Income We find in this model that that both of these predictor values are significant (p-value < .05). The RSE of the model is 0.0017 (43 degrees of freedom) with R^2 value of 0.2517. Table 5 shows the model coefficient estimates.

Table 5: Linear Regression with Price and Income

Intercept	Price	Income
0.991341	-0.0001406	0.000001395

3.3 Single Variable Models

Finally we will look at the three possible single variable linear models.

Age We find in this model that that the predictor value is significant (p-value < .05). The RSE of the model is 0.0017 (44 degrees of freedom) with R^2 value of 0.1677. Table 6 shows the model coefficient estimates.

Table 6: Linear Regression with Age

Intercept	Age
0.980091	0.000404

Price We find in this model that that the predictor value is not significant (p-value = 0.107). The RSE of the model is 0.0018 (44 degrees of freedom) with R^2 value of 0.05788. This model will not be included in our final calculations. Table 7 shows the model coefficient estimates.

Table 7: Linear Regression with Price

Intercept	Price
0.9955	-0.0001106

Income We find in this model that that the predictor value is significant (p-value < 0.05). The RSE of the model is 0.0017 (44 degrees of freedom) with R^2 value of 0.1602. This model will not be included in our final calculations. Table 8 shows the model coefficient estimates.

Table 8: Linear Regression with Income

Intercept	Price
0.9955	-0.0001106

3.4 Model Discussion

Now that we have constructed all of the possible models we wee that there appears to be no real improvement from using the models other than the full model. We will explore this further in section 4. One thing that we notice is that the Age and Income predictor variables each explain approximately 15% of the variation in the model. The price on it's own, however, only explains about 5% of the variability in the model. However when we combine all three of the variables together in the full model it serves to explains 40% of the variability in the data.

We next calculate the variance inflation factors (VIF) to check for collinarity between the predictor variables and we found that all of the variables had VIF values below 10 so there is no issue with colliniarity.

Finally we will look at each of the of the coefficients of the model. The thing that we need to keep in mind is that we have conducted a Box-Cox transformation with $\lambda = -1$. This has transformed all of our sales values to be close to 1 so our coefficients are explaining the small changes in these values. When we go to predict our test values we will have to transform these results to get actual sales values. The one thing that we can discuss, is the sign of the coefficients and relative sizes of the coefficients in our full model.

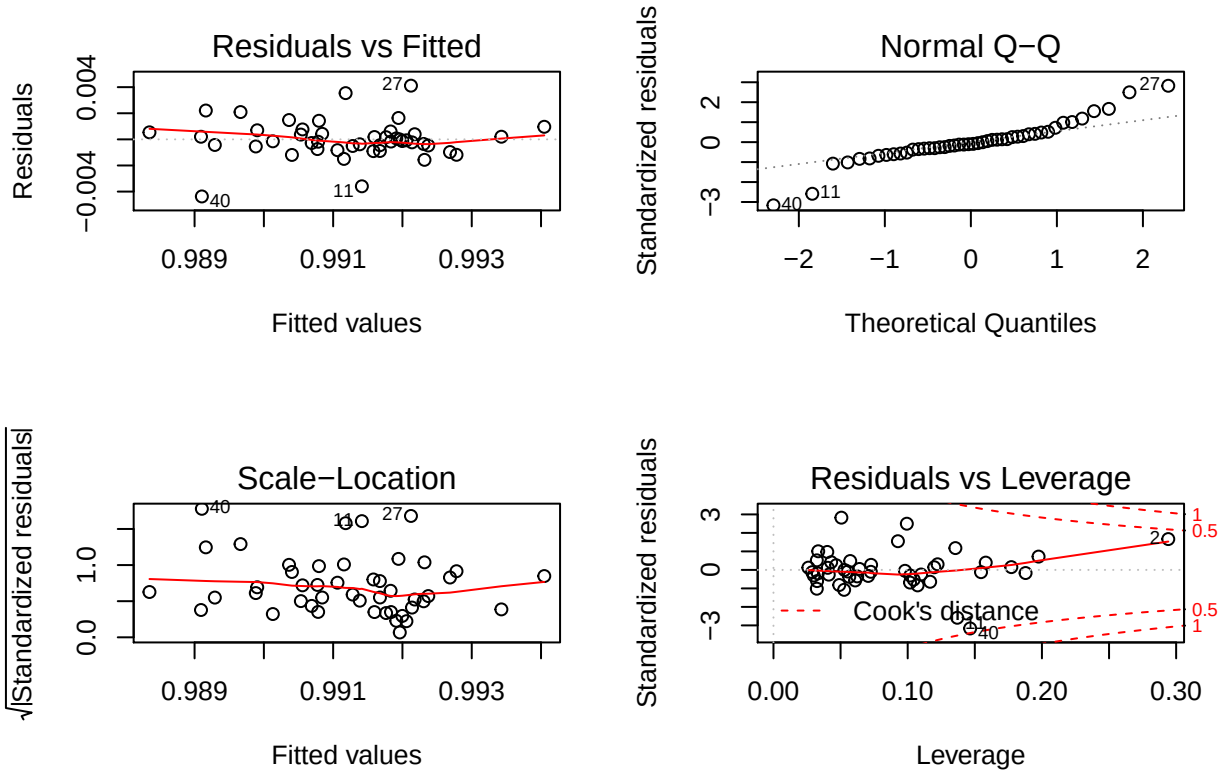
For our full model we note that as both Age and Income are increased there is an increase in Sales. We also see that the coefficient of Age is 0.0004085 and the coefficient of Income is 0.00000111. From these values we can see that when all of the other variables are held constant the Age has a larger impact on sales than the Income does. Looking at the coefficient of Price we get a value of -0.0001797, this implies that as we increase the price of a pack of cigarettes we should expect the sales to decrease. The magnitude of the effect is similar to age and much larger then income. These values do seem to make sens to the problem as we would expect that there is more cigarette use in the older population. I am somewhat surprised by the increase in Sales with income but it may be a case of more available income results in more sales. Finally I would expect that the increase in price per pack would lead to a decrease in sales.

4 Selecting a Model

Finally we need to select a final model and make our predictions. To decide on the best model we will be looking at the mean square error, the residual standard error and adjusted R^2 values. These are calculated by using our model to predict the values in the training data set and comparing them to the actual values. Our best model ended up being:

$$Box - Cox(Sales) = 0.0004085 * Age + 0.00000111 * Income - 0.0001797 * Price$$

Using this full model we get the following values; The Adjusted $R^2 = 0.363$, MSE = 0.000002042, RSE = 0.001495. Using this model we also get an R^2 value of 0.4054 and an F-statistic of 9.547 with a p-value of ~0. Our last step before making predictions from the data is to look at the residual and diagnostic plots. WE notice in the figure below that there is no discernible pattern to the residuals and the variance of the residuals by observation appear to be consistent. While we do see some outliers in the residuals they appear to be evenly distributed. We would not be comfortable making predictions near the edges of our data but we may be able to make decent predictions near the center of our data.



Finally we will make predictions for the evaluation data set. Applying our model to evaluation data results in the transformed sales figures for the states in the test data. We then invert our transformation to get our values. We can see the results in the Table 9 below.

Table 9: Predicted Sales in Packs of Cigarettes

DE	MO	MI	SC	NC
113.8	129.2	111.2	99.5	122.2

5 Concluding Remarks

In this analysis, we employed linear regression to fit a number of different models for predicting the sales of packs of cigarettes from the 1970 data set. The model with the best statistics was the linear model that included all three predictor variables. While we do have concerns about this model's predictions we were able to get values that seem appropriate. I am also unsure if this transformation is the best one that could be performed but it was an interesting skill to use.

6 Appendix

In this section, we provide the R code used in the analysis.

```
#Load the relevent libraries
library(foecast)
library(car)
```

```r
#Load the Data
cig <- read.csv("DATA621/cigarette-training-data.csv")
cig_test <- read.csv("DATA621/cigarette-evaluation-data.csv")

#Construct the Summary Plots
par(mfrow = c(2,2))
hist(cig$Age, main="Age", xlab="Age")
hist(cig$Income, main="Income", xlab="Income")
hist(cig$Price, main="Price", xlab="Price")
hist(cig$Sales, main="Sales", xlab="Sales")

#Transforming the Data using Box-Cox
cig$bcSales <- BoxCox(cig$Sales, -1)
hist(cig$bcSales, main="Box-Cox Transformed Sales, lambda = -1", xlab="Box-Cox Transformed Sales")  #Pl

# Fit the following models to the TRAINING set. For each model, extract the model
# coefficient estimates, predict the responses for the TRAINING set, and calculate
# the "mean prediction error" (and its standard error) in the TRAINING set.

#Building the full model
bcfit <- lm(bcSales ~ Age + Income + Price, data=cig) #Fitting the model
summary(bcfit)
plot(bcfit)
vif(bcfit)  #Checking for Collinearity
mean((cig$bcSales - predict(bcfit, cig))^2)
sd((cig$bcSales - predict(bcfit, cig))^2)/sqrt(46)

#Building the two variable models
# Age and Income
bcfit2 <- lm(bcSales ~ Age + Income, data=cig)
summary(bcfit2)
vif(bcfit2)
mean((cig$bcSales - predict(bcfit2, cig))^2)
sd((cig$bcSales - predict(bcfit2, cig))^2)/sqrt(46)

# Age and Price
bcfit3 <- lm(bcSales ~ Age + Price, data=cig)
summary(bcfit3)
vif(bcfit3)
mean((cig$bcSales - predict(bcfit3, cig))^2)
sd((cig$bcSales - predict(bcfit3, cig))^2)/sqrt(46)

# Price and Income
bcfit4 <- lm(bcSales ~ Price + Income, data=cig)
summary(bcfit4)
vif(bcfit4)
mean((cig$bcSales - predict(bcfit4, cig))^2)
sd((cig$bcSales - predict(bcfit4, cig))^2)/sqrt(46)

#Building the one variable models
# Age
bcfit5 <- lm(bcSales ~ Age, data=cig)
summary(bcfit5)
```

```r
mean((cig$bcSales - predict(bcfit5, cig))^2)
sd((cig$bcSales - predict(bcfit5, cig))^2)/sqrt(46)


# Price
bcfit6 <- lm(bcSales ~ Price, data=cig)
summary(bcfit6)
mean((cig$bcSales - predict(bcfit6, cig))^2)
sd((cig$bcSales - predict(bcfit6, cig))^2)/sqrt(46)


# Income
bcfit7 <- lm(bcSales ~ Income, data=cig)
summary(bcfit7)
mean((cig$bcSales - predict(bcfit7, cig))^2)
sd((cig$bcSales - predict(bcfit7, cig))^2)/sqrt(46)

# Predicting on the TEST data set and transforming them back to actual sales values.
test <- predict(bcfit, cig_test) # Predicting the data
predictions <- (test*(-1)+1)^(1/-1) # Converting back from Box-Cox
```