**Project 1B Dataset Analysis**

**Kayla Klassen, Mason Lorber and Cole Streich**

**DSCI 429: Predictive Analytics**

**Introduction**

TabletCo is an online producer and supplier of tablets, that has recently experienced a considerable lag in profits. Because of this, they are considering a strategic shift by changing the price that they charge. While TabletCo exclusively sells their product online, it pins the price it charges to the location of the consumer, in a form of regional pricing. To support their pricing decision, the company has collected monthly data on their tablets for eight distinctive regions in which the tablets have been sold for the past four years.

**Problem Presentation**

How will an adjustment in price alter TabletCo’s recent lag in profit?

**Specification and Design**

Within the monthly data that was collected by TabletCo, ten independent variables were provided. These independent variables are listed and explained in Table 1. From these ten independent variables, only average age and price were utilized in our predictive model. These two variables were chosen after a linear regression model ran each independent variable on the dependent variable, profit per capita (profit per capita is representative of the profit per capita for TabletCo tablets in the Region during the Month). The model showed that the average age variable was moderately correlated with profit per capita and that price was strongly correlated, leading us to use them as model predictors.

**Table 1.** The list of independent variables provided in the monthly data collected by TabletCo. Avg. Age and Price are the two variables we chose to use in our predictive model.

|  |  |
| --- | --- |
| ***Independent Variables*** | ***Description*** |
| **Month** | Month during which observation was recorded |
| **Region** | Geographic region where observation was recorded |
| **Population(in thousands)** | Number of people living in the Region during the Month |
| **Avg. Education** | Average education level of people living in the Region during the Month |
| **Avg. Income** | Average income level of people living in the Region during the Month |
| **Avg. Age** | Average age level of people living in the Region during the Month |
| **Avg, Household Size** | Average size of all households in the Region during the Month |
| **Unemployment** | Unemployment rate in the Region during the Month |
| **Rainfall** | Recorded rainfall in the Region during the Month |
| **Price** | Average price charged for TabletCo tablets in the Region during the Month |

**Descriptive Results**

**Table 2.** The Six Number Summary table consists of the Minimum, 1st Quarter, Median, Mean, 3rd Quarter and Maximum statistics for Average Age, Price and Profit per Capita.

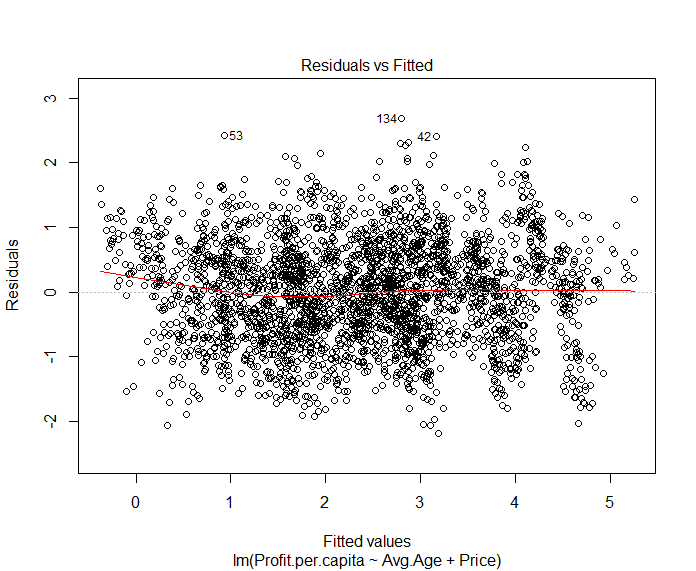
|  |  |  |  |
| --- | --- | --- | --- |
| ***Six Number Summary*** | | | |
|  | **Avg Age** | **Price** | **Profit per Capita** |
| **Min** | 34.80 | 585.40 | -1.727 |
| **1st. Quarter** | 38.80 | 707.30 | 1.320 |
| **Median** | 40.90 | 761.70 | 2.333 |
| **Mean** | 40.98 | 760.50 | 2.373 |
| **3rd Quarter** | 43.23 | 811.50 | 3.399 |
| **Max** | 49.70 | 94.10 | 6.691 |

**Table 3.** The Correlation table shows the correlation that exists between each variable included (Avg. Age, Price, Profit per Capita).

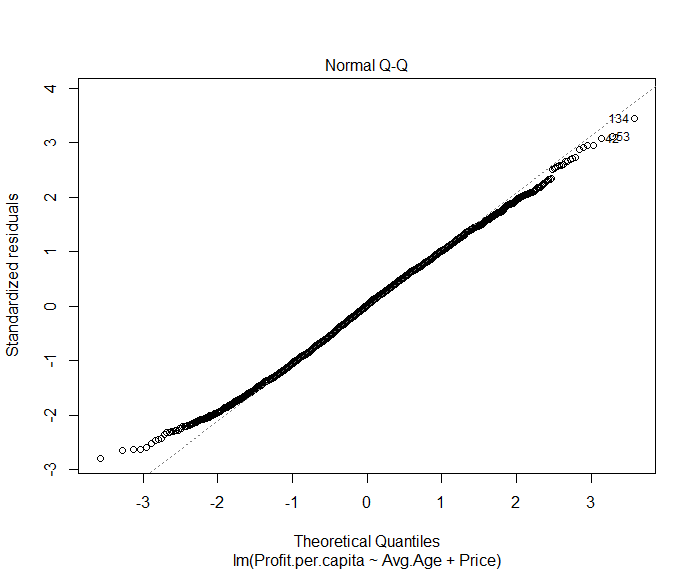
|  |  |  |  |
| --- | --- | --- | --- |
| ***Correlation*** | | | |
|  | **Avg. Age** | **Price** | **Profit per Capita** |
| **Avg. Age** | 1 | - | - |
| **Price** | -.53 | 1 | - |
| **Profit Per Capita** | -.43 | .83 | 1 |

**Table 4.** The Coefficient table includes data regarding the coefficient, the standard error, the T-Stat Value, the probability that the value falls within the 95% confidence interval, the total standard error with the included degrees of freedom, the R2 value, the Adjusted R2 value, the F-Statistic with the included degrees of freedom and the existing P-Value.

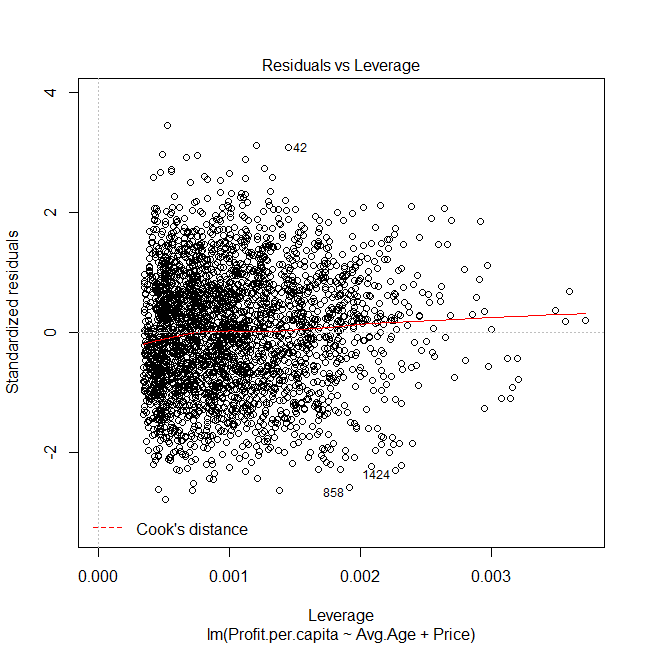
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | | | | |
|  | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| **Intercept** | -.1055e+01 | 3.633e-01 | -29.051 | <2e-16\*\*\* |
| **Avg. Age** | 1.403e-02 | 5.77e-03 | 2.428 | 0.0152\* |
| **Price** | 1.642e-02 | 2.318e-04 | 70.061 | <2e-16\*\*\* |
| **Standard Error** | 0.7803 and 2877 degrees of freedom | | | |
| **Multiple R-squared** | 0.6973 | | **Adjusted R-squared** | 0.6971 |
| **F-Statistic** | 3314 on 2 and 2877 DF | | **p-value** | <2.2e-16 |

**

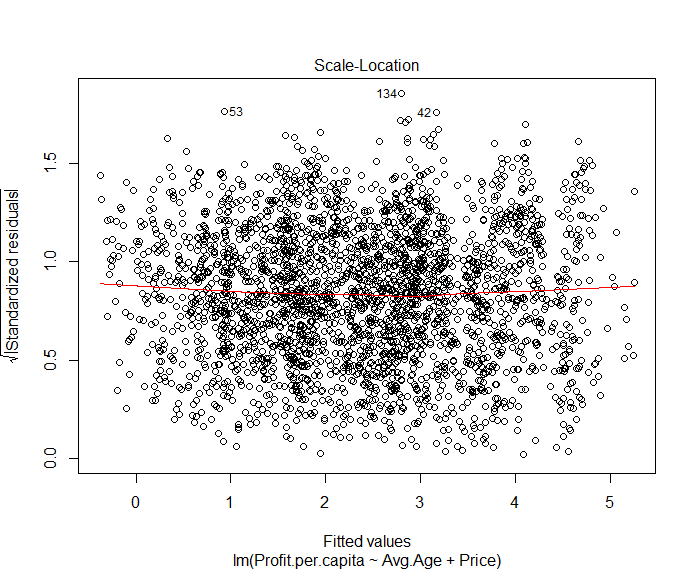
**Plot 1.** As the fitted values increase from 0 to 5, the variance of the plot stays relatively constant, replicating the data to be somewhat homoscedastic. Also, the red trendline of this plot stays relatively flat, indicating a strong line of fit.

**

**Plot 2.** The Normal Q-Q plot shows that the majority of the data is normally distributed while a small portion of the data at the head and tail of the plot, stray off the typical trend of the data.

**

**Plot 3.**  The data doesn’t stray outside of the Cook’s Distance line, reflecting that there is a lack of outliers and that the data falling within the confidence intervals.

**

**Plot 4.** The red trend line stays flat through as the fitted values increase but starts to trend up near the end of the fitted values. Similarly, to Plot 2, this plot shows the data is somewhat homoscedastic because of the similar variation throughout the data.

**Book Questions**

**1.) Predict what will happen if TabletCo raises or lowers its price.**

According to our regression model equation *Profit per Capita = -10.55 + 0.014AvgAge + 0.016Price + U*. Price and profit per capita are positively correlated, meaning they move in the same directions. The price coefficient of 0.016 implies for every dollar increase in the price that TabletCo charges, profit per capita increases by roughly a cent. This also means that for every dollar decrease in the price, profit per capita decreases by roughly a cent.

**2.) Detail the data-generating process you assumed to arrive at your results. Explain how you chose features of the determining function, e.g., its functional form and the variables in it.**

We began the data-generating process by running a multiple linear regression in RStudio for all the given independent variables on the dependent variable, profit per capita. From the output of the multiple linear regression, we saw that the model explained 77% of the variation in profit per capita. We also noticed that only two of the ten variables were moderately or strongly correlated with profit per capita: average age and price. Using this information, we constructed our determining function with two independent variables:

*Profit per Capita = -10.55 + 0.014AvgAge + 0.016Price + U*

This linear determining function, with two independent variables explains 69.71% of the variation in profit per capita.

**3.) Explain the estimation method you used to arrive at your estimates for the data-generating process.**

We used Ordinary Least Squares as the estimation method to produce the determining function. Using this method allowed us to produce a best fit line for the TabletCo data, where the seven assumptions of the Ordinary Least Squares method were met:

1. The regression model is linear in the coefficients and the error term
2. The error term has a conditional mean of zero
3. All independent variables are uncorrelated with the error term
4. Observations of the error term are uncorrelated with each other
5. The error term has a constant variance
6. No independent variable is a perfect linear function of another explanatory variable
7. The error term is normally distributed

We also assumed the population regression equation is equal to the determining function of the data-generating process.

**4.) After applying your estimation method, explain:**

1. **What do the point estimates mean?**

Point estimates use sample data to calculate a single value which is to serve as a “best guess” of an unknown population parameter.

1. **What do the p-values mean?**

The p-values for price and average age indicate the significance of each variable. In our regression model, we can say that each independent variable has an effect on the dependent variable and the effect is statistically significant.

**c. What do the upper and lower bounds for the 95% confidence intervals mean?**

The upper and lower bounds for the 95% confidence intervals mean we are 95% sure that the true population mean lies between the lower and upper bound.

**5.) Detail the line of reasoning necessary to make your prediction in Question 1.**

1. The data-generating process for an outcome, Y, can be expressed as Yi = α + β1Xi + … + βKXKi + Ui
2. {Y, X1i, … XKi} Ni = 1 is a random sample. We are choosing a random sample to eliminate systematic bias
3. E[U] = E[U(X1)]...E[U(XK)] = 0 meaning the population regression equation is equal to the determining function of the data-generating process.

**6.) Identify at least one argument why your estimates may not be suitable for making the active prediction you made in Question 1. Be sure to highlight where your line of reasoning breaks down if this opposing argument is correct.**

Our estimates may not be suitable for making active predictions because no factors in the error term, U, are fixed, which could generate endogeneity. We do not have demand data, so we do not have any information about the change in demand for the product or the variation in demand. We also used correlation as a means of determining which independent variables should be utilized in our model instead of using p-values to determine statistical significance. We can explain less of the variation in profit per capita using our simplified regression model than the model using all ten independent variables.

**Causality**

We cannot assume causality because our passive model lead to a passive prediction. Because we used lag data, our model cannot be an active one. We also cannot exogenously alter the variables to produce an active model. We did find that the data presented correlation between the independent variables and profit per capita. The independent variables in their entirety attempt to predict profit per capita, accounting for 77% of the variation in the data. This model was unnecessarily large and we wanted to reduce the amount of treatments and find a high correlation within the data. After conducting trials to reduce the number of independent variables, we produced our final regression model, which predicts about 70% of the variation in the dependent variable.

**Conclusion**

Our linear regression model predicts 69.71% of the variation of the Profit per Capita dependent variable. After running a regression of all the independent variables on Profit Per Capita, we narrowed down the variables used for the final linear model. We chose average age and price to utilize in our final prediction model based on their high correlation to Profit per Capita.