Mike Brodecki
Econ 0150
12/6/25

Capstone Paper

For my research question, I decided to test if there was any difference in ratings for hotel rooms due to price or room classification (economy or luxury). Using a CSV data set with 108 observations from 14 different hotel chains, I was able to support my question with three simple variables. My dependent variable or output was the rating that each hotel room received, with an independent variable of price and a categorical variable describing the classification of each room, whether economy or luxury.

The summary statistics for my model include price, rating, class, and hotel chain, providing me with information surrounding my question. The data came from one of my other classes, where the teacher previously found the CSV and used it for an example in class that sparked my interest in the topic. There may be some credibility and reliability issues because the professor found the data set, yet I believe that my other professor is a credible source to me. Each observation was assigned its own unique identification factor correlating to each point, allowing you to look at the specific price and rating if interested while looking through the dataset. The only data cleaning that took place within my model was creating new variables for luxury and economy, to be able to place the categorical variables within my GLM.

Looking at my GLM, I ran a regression with rating, price, and class (class being defined as Z or categorical). My output is a rating that each hotel room receives, which is represented by an intercept, the price of the room, and its classification variable as luxury or economy. Our intercept can be calculated in two different ways, depending on which classification you use as your baseline category. For this capstone report, I'm going to choose economy as my baseline (Z

= 0). Intercept received a value of 60.99, my slope or beta attached to price (in $) was 0.038, which increases my rating by that amount for every additional dollar; finally, my categorical beta attached to Z is 17.48, which only increases or adds value to luxury (Z = 1).

Provided at the end is a summary chart from my regression model, giving us the coefficients or betas on each variable, a t-stat, and a p-value for each variable. For interpreting what all of these numbers mean, we can start by looking at the F-stat at the bottom of our output, which tells us that our entire regression model with price and class is significant. Looking more in depth, all of the numbers provided for each coefficient or beta, we see that they all have a very small p-value, smaller than 0.05, while their t-stats are all significantly large, with both measures proving that each beta is significant on its own towards the output or rating of each hotel room.

Attached at the end is a residual plot, showing that normality and homoscedasticity are satisfied. There is no variation from a clear line, like a cone or X/U shape pattern, within my residuals, helping prove normality, and the residuals are evenly represented from 0-500 on my fitted values, satisfying the homoscedasticity requirements.

Concluding my capstone paper, we can see within the data how there is a significant difference in rating between two different classifications of hotel rooms, allowing us to fail to reject our Null Hypothesis (I said the opposite during my presentation, but this is my true conclusion). I found that when testing for the categorical variable or class, there is a significant difference between economy and luxury, proving that the categorical variable is required or significant towards my final regression.

References to data can be found within the folder submitted with my CSV and code, but information was provided by a different course at the University of Pittsburgh. As for the data, there were 108 observations used towards the model from 14 different hotel chains, giving

numerous variations between chains and no questions about the reliability of conclusions drawn from the data.

```
Call:
lm(formula = rating ~ price + Luxury, data = hotels)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2191 -1.7405 -0.3445  1.6879  7.6935

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.986337   0.709975  85.899   <2e-16 ***
price        0.037826   0.003867   9.782   <2e-16 ***
Luxury      17.480761   0.998614  17.505   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.904 on 105 degrees of freedom
Multiple R-squared:  0.941,     Adjusted R-squared:  0.9399
F-statistic: 837.1 on 2 and 105 DF,  p-value: < 2.2e-16
```