ST221 Assignment 1

Student Number: 2106983

# Question 1

# a)

Chart, scatter chart

Description automatically generated

# b)

Chart, scatter chart

Description automatically generated

This is not an acceptable plot as there is a large amount of residual data points considerably further away from the smoother for larger fitted values than smaller ones. This means that the variance of the residuals is clearly not constant so the model assumption of homoscedasticity is violated. tend to be positive for small fitted values and negative for intermediate so convex curve rather than horizontal line, issue with linearity assumption.

# c)

I would suggest an improved model would be to log transform the price (response variable) and fit a simple linear regression of log price on width. I would suggest to do this because the scatterplot in part a) resembles an exponential curve with values of price greatly increasing with larger width.

Chart, scatter chart

Description automatically generated

Above is the residual plot for the new model which is definitely an acceptable plot as the smoother is very close to a horizontal line at 0 and the spread of the data across the fitted values is fairly constant so neither model assumptions of linearity and homoscedasticity are violated.

# d)

As the slope for the model is 1.0043 for each 1mm increase in width, the average price of a diamond is expected to multiply by e^1=2.73 so increase by about 173 percent.

# Question 2

# a)

Chart, scatter chart

Description automatically generated

Whilst the residuals are fairly evenly spread around the smoother which means the model assumption of homoscedasticity is not violated, the smoother itself slightly deviates from a horizontal line at 0 so one could argue that the linearity assumption is violated. negative for small and large fitted but positive for intermediate

# b)

Where and are the values of the indicators variables of being of size medium and large respectively for the jth observation. Or equivalently:

if diamond j is of size small

if diamond j is of size medium

if diamond j is of size large

# c)

Chart, scatter chart

Description automatically generated

# d)

For a 20% increase in weight the average increase of price in a large diamond is expected to be e^-0.9 times the average increase of price in a small diamond so around 0.41 times. The expected change of price for a small diamond with a 20% increase in weight is e^0.756 times so about 2.13 times so about 113 percent increase. Therefore expected change in price for a large diamond is about 0.87 times so about 13 percent decrease. Didn’t need to give a quantitative interpretation just needed to say represents the difference in slope between the regression line for large diamonds and the regression line for small diamonds. Difference in effect of weight on log-price between large diamonds and small diamonds. I had considered writing this but left it with the numbers.

Chart, scatter chart

Description automatically generated# e)

# f)

One difference is that the model in (b) introduces categories for the weight to rectify issues in the model where it over/underestimates values by considering the whole range. This is advantageous because we can see more directly how a “type” of weight more directly affects the price and also within each type how the weight affects the price. The model in (e) is limited by considering the weight as all one type so we cannot see the intricacies of how weight affects price as much, however one benefit of this is that it is a simpler model and therefore the more parsimonious model which makes it preferrable.

When looking at the residual plots for the models they are both acceptable as the smoothers closely resembe horizontal lines at 0 and the residuals are evenly spread across the fitted values. Therefore both models satisfy the linearity and homoscedasticity assumptions and there is no benefit to picking either model in this sense. Choice of size categories arbitrary

An advantage of the model in (b) is that it might be more easily understood by non- statisticians as a linear model that allows for different pricing structures for diamonds in the different size groups. However, the choice of size groups is somewhat subjective and the fitted mean function for log-price is a piece-wise linear function with discontinuities at the size boundaries. Furthermore, we note that the range of weights for large diamonds is rather limited and that the relationship between the log-price and the weight of large diamonds is negative, which is counter-intuitive.