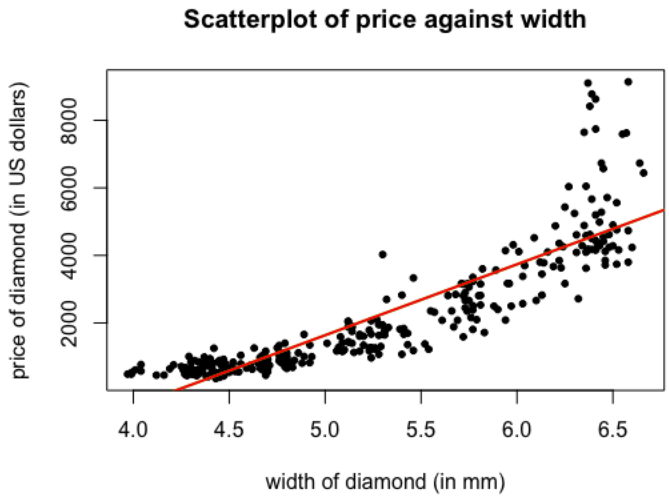
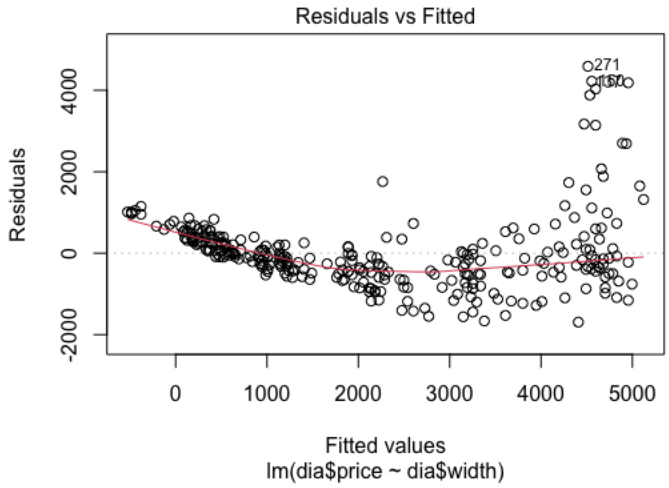
ST221 Assignment 1
Student Number: 2106983

Question 1

a)



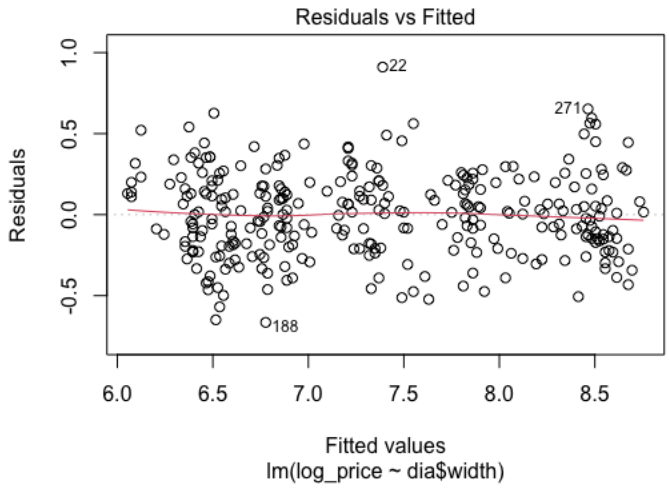
b)



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.

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.



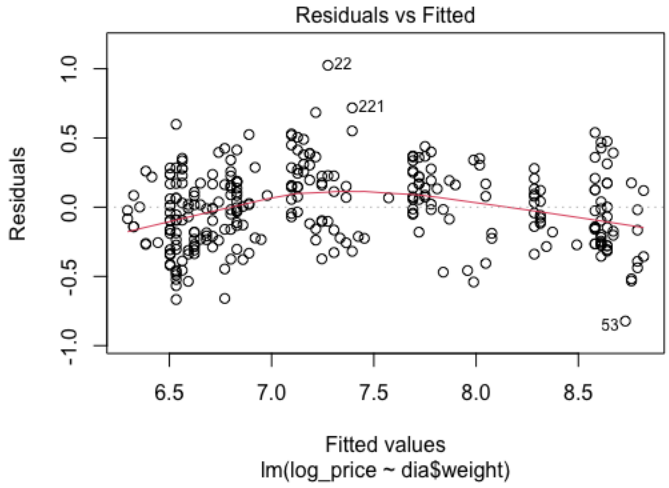
Above is the residual plot for the new model which is definitely an acceptable plot as the smoother is very close to a horizontal line at 0 and the spread of the data across the fitted values is fairly constant so neither model assumptions of linearity and homoscedasticity are violated.

d)

As the slope for the model is 1 for each 20% increase in width, the average price of a diamond is expected to multiply by e^0.2 so increase by about 22 percent.

Question 2

a)



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.

b)

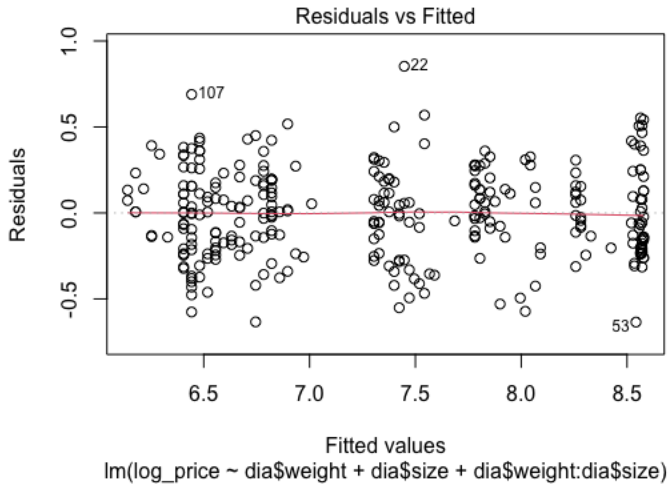
$$\log(price)_j = \mu + \alpha_M x_{jM} + \alpha_L x_{jL} + \beta weight_j + \gamma_M(x_{jM} * weight_j) + \gamma_L(x_{jL} * weight_j) + \epsilon_j$$

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

$$\log(price)_j =$$

$\mu + \beta weight_j + \epsilon_j$	if diamond j is of size small
$(\mu + \alpha_M) + (\beta + \gamma_M) weight_j + \epsilon_j$	if diamond j is of size medium
$(\mu + \alpha_L) + (\beta + \gamma_L) weight_j + \epsilon_j$	if diamond j is of size large

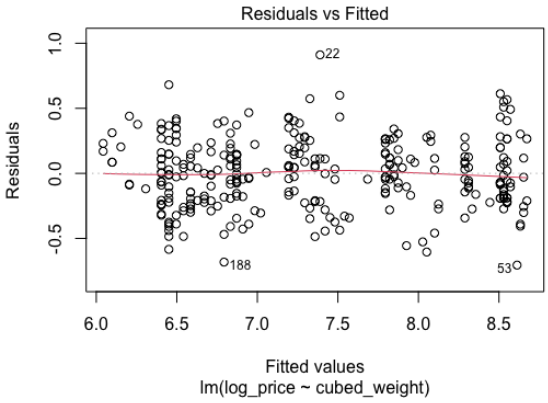
c)



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.

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.