Timila Kulkarni

Group 31

STAT 311

Final Project Part 2

Problem 1: Hypothesis Test

I conducted a hypothesis test, a two sample z-test for the difference between means, to test the hypothesis that furnished units have lower prices than unfurnished units, on average. I used a z-test since the sample was *very* large, and I calculated the standard deviations in R.

The statistical hypotheses are:

$H_0: \mu_F - \mu_{UF} = 0$

$H_A: \mu_F - \mu_{UF} < 0$

In words, the null hypothesis is that the difference between mean prices for furnished and unfurnished units is 0, or there is no difference. The alternative hypothesis is that the difference in mean prices for furnished and unfurnished units is less than 0, or the mean prices for furnished units are lower than the mean prices for unfurnished units.

I performed the hypothesis test in R. I used a 95% confidence level, i.e. a significance level of 0.05. I first subsetted the data for furnished and unfurnished, and then found the standard deviations of prices of furnished and unfurnished units: 66461.95 and 4573591 respectively. Then, I used these values to carry out the z test in R, using `z.test`.

The test statistic for this was -0.97075, with a p-value of 0.1658.

Since the test statistic of -0.97075 is greater than the z-critical cutoff value of -1.96, we fail to reject the null hypothesis.

Since the p-value of 0.1658 is greater than the significance level of 0.05, we fail to reject the null hypothesis.

Hence, we do not have enough evidence to reject the null hypothesis. So, there is insufficient evidence to reject the claim that there is no difference between average prices of furnished and unfurnished units. Hence, we cannot conclude the alternative hypothesis that the difference in mean prices is greater than zero.

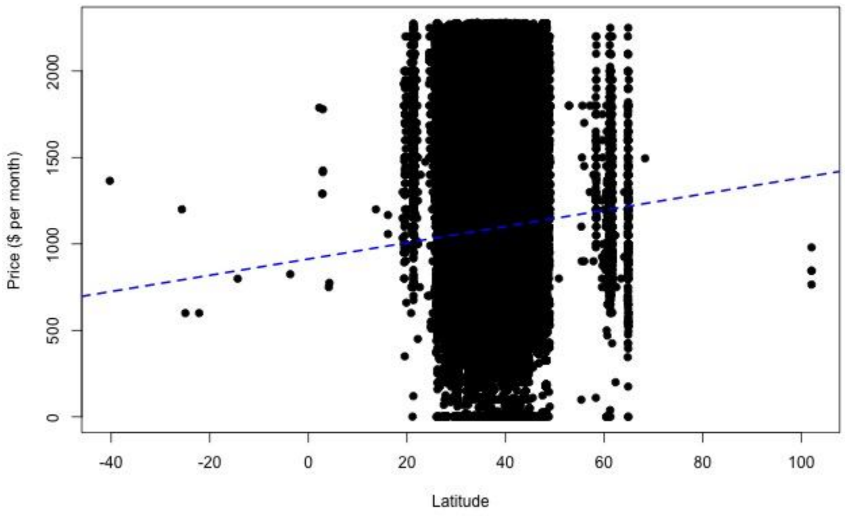
Problem 2: Regression Models:

***Timila*:**

a) Building the regression model

As proposed in Part 1, I am trying to build a regression for price on latitude. The data for this comes from the Housing data set. In Part 1, I described how I decided to clean up my data by removing outliers in order to carry out a meaningful regression. For this regression, I decided to remove the extreme outliers for price since there were a few that were very high and skewing the data a lot. However, I kept everything else.

After carrying out the regression using `lm.out` in R, I plotted the regression line with a scatterplot:



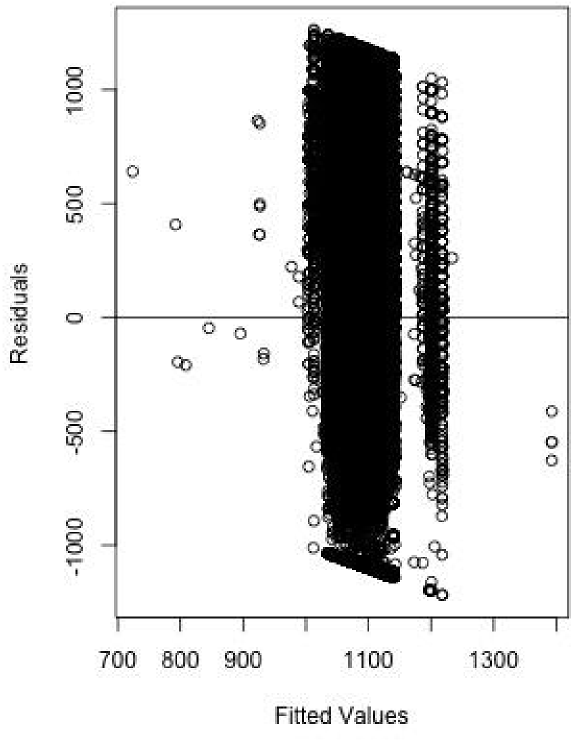
The corresponding regression equation was:

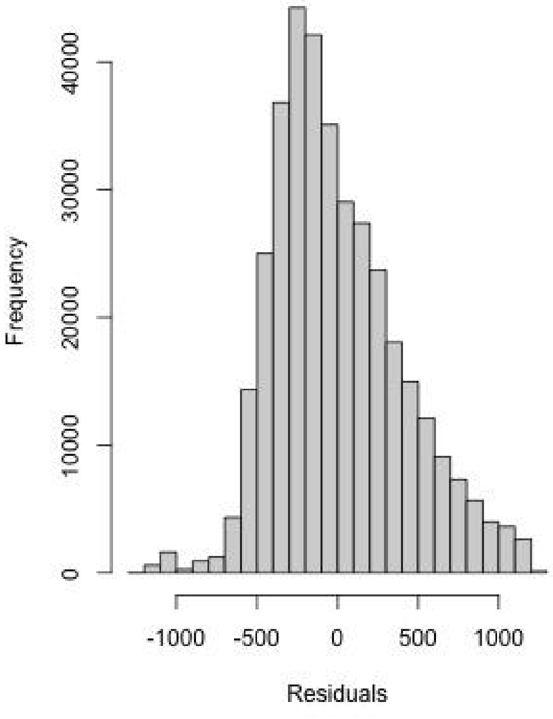
$$\hat{y} = 912.7873 + 4.6999x$$

The regression slope coefficient, 4.6999, is the change in price for every unit increase in latitude. So, for every one degree increase in latitude, there is a $4.6999 in price.

The R-squared value of 0.004241 means that about 0.4% of the variation in price can be explained by the latitude of the housing unit.

b) I ran diagnostics on the model, with the following residual plot and histogram of the residuals:





From the residual plot, we can see that the points do bounce around on either side of the line. However, although there is no definite pattern, they do seem to be clustered in the center.

However, since this is not a “systematic” pattern, we can say the ***linearity assumption*** has been met. However, since the residuals are not evenly spread out, and they spread out more as we move along the x-axis, the ***constant variance assumption*** is violated.

The histogram of the residuals looks right-skewed, hence, the ***normality assumption*** has been violated. Finally, from the residual plot, there seems to be a possible correlation between the residuals and fitted values, hence, the independence assumption may have been violated.

c) The 95% confidence interval for the regression slope is 4.4661 to 4.9337. In context, this means that we are 95% confident that on average, for each 1 degree increase in latitude, the price will change by between $4.4661 to $4.9337.

d) The 95% confidence interval for the mean price when the latitude is equal to the 75th percentile, 41.1963, is $1104.819 to $1107.996. In context, this means that we are 95% confident that the mean price of a house whose latitude is 41.1963 will fall between $1104.819 and $1107.996.

e) The 95% prediction interval for a new randomly selected value of price when latitude is equal to the 75th percentile, 41.1963, is $325.0927 to $1887.722. In context, this means that there is a 95% probability that a randomly selected housing unit whose latitude is 41.1963 will have its price fall between $325.0927 and $1887.722.

Problem 3: Group Synthesis

a) The biggest difficulty that our group had with our “Housing” dataset was the overall bulk and size of the dataset. The issues that arose from this dilemma was the multiple filtering processes that we had to perform in order to get rid of outliers such as the “0” values which would skew the graphs and visualizations. Some of the outliers were very high prices that skewed the entire dataset. Additionally, an issue with this was deciding whether or not to remove outliers for the regression analysis, and if so, how far to go with that. Since the outliers skewed the data so much, they were an important consideration in the regression. At the same time, they caused the scatterplot to be very skewed; therefore, for a meaningful regression, some outliers had to be removed. With the larger dataset, there was also a delay in processing speed as rstudio had to intake huge amounts of data in order to present viable outputs. Finally, there weren't too many quantitative variables that we could use for a regression analysis, since the number of beds/baths clearly didn't work; we were left with square footage and latitude, which weren't perfect either.

b) Our regression models were all for the regression of price on various other variables. Timila used the latitude variable, while Matthew and Mikayla used the square footage variable. The best regression model for our specific dataset was the regression for price on square footage, done by Mikayla and Matthew. These clearly fit the data better, with the R-squared values being much higher than Timila's for the price on latitude regression. Additionally, the residuals showed that the data was better suited for inference, since most assumptions were satisfied, with this, the regression residuals showcased the linear relationships between the housing unit prices and square footage. On the other hand, for the latitude and price regression, although there was a rough linear correlation, the R-squared value was very low, less than 1%. Additionally, the residuals showed that most of the assumptions for inference were not met. Even visually, the regression line for price on square footage was clearly a better fit than the regression line for price on latitude. This makes sense, because latitude usually doesn't affect price too much if you think about it – something like zip code might have been more appropriate, but you can't use a regression analysis on this. On the other hand, the square footage of a house matters, because larger houses usually cost more.