Name: Amy Lai

ANALYSIS PART 1

Is distance to nearest metro station well described by an exponential model?

a) The table below contains the numerical summaries for the MRT_distance variate.

Minimum	23.38284	IQR	1137.383
1st Quartile	289.32480	Range	6372.9
Sample Median	492.23130	Sample Mean = $\bar{y}$	1085.84
3rd Quartile	1426.70800	Sample Standard Deviation = s	1288.32
Maximum	6396.28300	Sample Skewness	1.878554

b)

Assume an exponential model Exponential(θ) is reasonable to model the variable MRT_distance. The maximum likelihood estimate for θ is given by $\hat{\theta} = \bar{y} = 1085.84$. Here, θ corresponds to average distance from each of the studied population to their nearest metro station measured in meters, where the studied population is the real estates in Sindian Dist, New Taipei City, Taiwan recorded from approximately late 2012 to early 2014.

c)

The following graph is a plot of the relative frequency histogram for metro distance with superimposed Exponential($\hat{\theta}$) probability density function

Name: Amy Lai

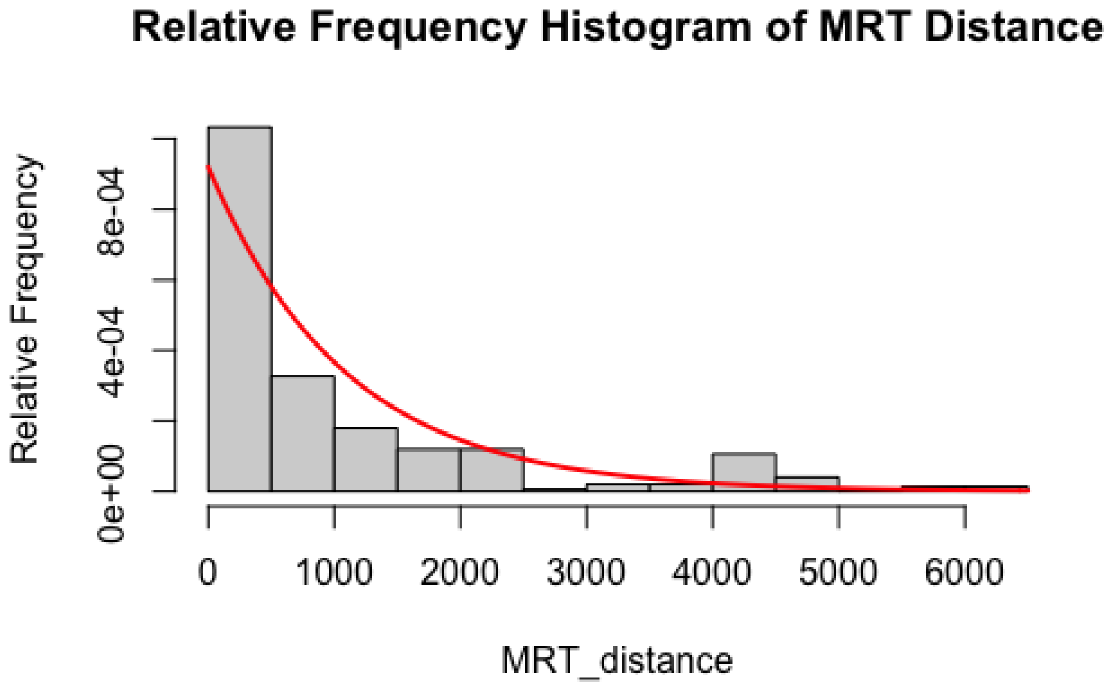


Figure 1. Relative Frequency Histogram for MRT Distance

d) The following table summarizes observed and expected frequencies assuming an Exponential($\hat{\theta}$) model.

j	Distance (meters) to Public Transit Group x_j	Observed Frequency f_j	Expected Frequency e_j
1	[0, 200)	50	50.465
2	[200, 400)	75	41.975
3	[400, 600)	50	34.913
4	[600, 800)	16	29.039
5	[800, 1000)	13	24.153
6	[1000, 1200)	8	20.090
7	≥ 1200	88	99.350
Total		300	300

The expected frequencies were calculated using the following formulas/procedure:

Let x_i denote the observed distance from the i-th sampled house to their nearest metro station. The expected frequency in the interval $[a_{j-1}, a_j]$ is calculated using

Name: Amy Lai

$$e_j = 300 \int_{a_{j-1}}^{a_j} \frac{1}{\hat{\theta}} e^{-x_i/\hat{\theta}} dx_i = 300(e^{-\frac{a_{j-1}}{1085.84}} - e^{-\frac{a_j}{1085.84}})$$

Where 300 is the total number of houses sampled and we have by $\hat{\theta} = 1085.84$ from part b.

When j=1, the interval is [0, 200), thus we have

$$e_1 = 300\left(e^{-\frac{0}{1085.84}} - e^{-\frac{200}{1085.84}}\right) = 50.465$$

When j=2, the interval is [200,400), thus we have

$$e_2 = 300\left(e^{-\frac{200}{1085.84}} - e^{-\frac{400}{1085.84}}\right) = 41.975$$

When j=3, the interval is [400, 600), thus we have

$$e_3 = 300\left(e^{-\frac{400}{1085.84}} - e^{-\frac{600}{1085.84}}\right) = 34.913$$

When j=4, the interval is [600, 800), thus we have

$$e_4 = 300\left(e^{-\frac{600}{1085.84}} - e^{-\frac{800}{1085.84}}\right) = 29.039$$

When j=5, the interval is [800,1000), thus we have

$$e_5 = 300\left(e^{-\frac{800}{1085.84}} - e^{-\frac{1000}{1085.84}}\right) = 24.153$$

When j=6, the interval is [1000, 1200), thus we have

$$e_6 = 300\left(e^{-\frac{1000}{1085.84}} - e^{-\frac{1200}{1085.84}}\right) = 20.090$$

When j=7, the interval is [1200, ∞), thus we have

$$e_7 = 300 \int_{1200}^{\infty} \frac{1}{1085.84} e^{-\frac{x_i}{1085.84}} dx_i = 99.350$$

e)

Step 1

The null hypothesis is H_0: An exponential model fits the MRT_distance data.

Mathematically, this is $H_0: p_j(\theta) = e^{-\frac{a_{j-1}}{\theta}} - e^{-\frac{a_j}{\theta}}$ for j=1,…,7

The alternative hypothesis is H_A: An exponential model does not fit the MRT_distance data.

Mathematically, this is $H_A: p_j(\theta) \neq e^{-\frac{a_{j-1}}{\theta}} - e^{-\frac{a_j}{\theta}}$ for j=1,…,7

Step 2

$$\Lambda = 2\sum_{j=1}^{7} X_j \log\,(\frac{X_j}{E_j})$$

Let λ be the observed value of Λ, then

$$\lambda = 2\sum_{j=1}^{7} x_j \log\left(\frac{x_j}{e_j}\right)$$

$$= 2\left[50\log\left(\frac{50}{50.47}\right) + 75\log\left(\frac{75}{41.98}\right) + 50\log\left(\frac{50}{34.92}\right) + 16\log\left(\frac{16}{29.04}\right) + 13\log\left(\frac{13}{24.16}\right) + 8\log\left(\frac{8}{20.09}\right) + 88\log\left(\frac{88}{99.35}\right)\right]$$

$$= 50.733$$

Step 3

Since n=300 is large, if H_0 is true then the distribution of Λ a Chi-squared distribution with degree of freedom k-1-p. In this scenario, k=7 since there are 7 intervals and p=1 because there is only one parameter θ being estimated under the hypothesized Exponential(θ) model. Therefore, the degree of freedom equals 7-1-1=5. The approximate p-value is

$$p - value = P(\Lambda\ \geq 50.766)$$

$$\approx P(W\ \geq 50.766) \quad \text{where } W\sim_{X^2}(5)$$

$$= 1 - P(W\ \leq 50.766)$$

$$= 9.808029 \times 10^{-10}$$

Step 4

Since $9.808029 \times 10^{-10} < 0.001$, we conclude that there is very strong evidence against the hypothesis that an Exponential model fits the MRT_distance data based on the observed data.

f)

Part a) provides numerical summaries to examine the suitability of exponential model for the distance to the nearest metro station variate. Firstly, for exponential model, mean should be greater than median. The sample mean of my data is 1085.84 and the sample median is 492.23130. Since sample mean is greater than sample median, the distance to the nearest metro station can be modelled by an exponential model. Secondly, the skewness of an exponential model should be positive. The sample skewness of my data is 1.878554, thus the distribution of the sample is close to an exponential model. Thirdly, the mean and the standard deviation of an exponential model should be close in value. However, the sample mean of my data is 1085.84 and the sample standard deviation is 1288.32. The two differs by around 200. Thus, this suggests a difference between the distribution of the distance to the nearest metro station variate and an exponential model.

From part c), we can observe some differences between the relative frequency histogram and the superimposed exponential curve. In particular, the second and third bar from the left is significantly lower than the expected exponential curve. When the distance is between 2500 and 4500, we see an increase in relative frequency, which contradicts the decreasing behaviour of the exponential curve. The bar at 4000 to 4500 is also significantly taller than the exponential curve. Therefore, the curve demonstrates several differences between the distribution of the distance to the nearest metro station variate and an exponential model. These differences are also displayed numerically in part d) where the expected frequency and observed frequency differed by around 30 when j=2 and around 15 when j=3 and j=6.

Lastly, the p-value from part e) is very small. It provides very strong evidence against the null hypothesis that an exponential model is a good fit for the MRT_distance variate.

Altogether, we have more evidence against the exponential model being suitable than evidence in support. Therefore, we conclude that distance to the nearest metro station is not well described by an Exponential model.

Name: Amy Lai

ANALYSIS PART 2

Is price explained by age of the house?

a)

The graph below is a scatterplot of the price per unit area versus house age.

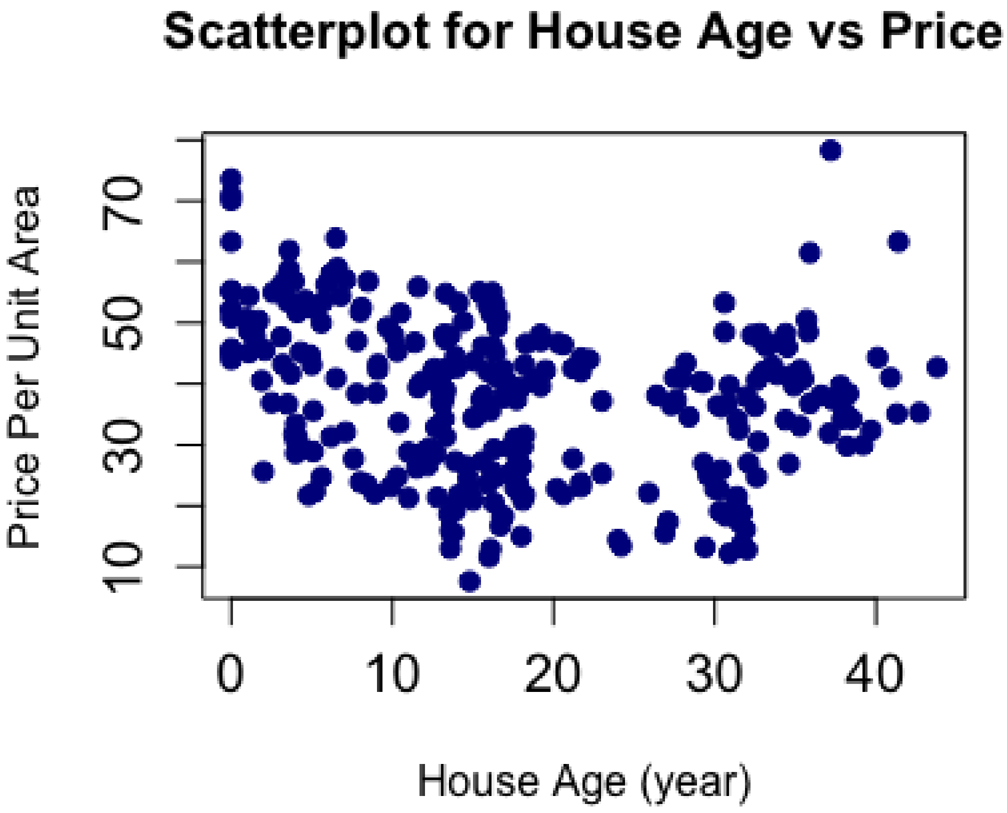


Figure 2. Scatterplot for House Age vs Price

b)

The sample correlation for price and age of house is -0.246543.

c)

Below is a two-way table for the observed frequencies of house age and price level.

Observed Frequencies	Price Category (per unit area)			
Age Group (in yrs)	Low (< 20)	Medium ([20,40))	High (≥ 40)	Total

< 15	7	51	79	137
[15,30)	11	47	36	94
≥ 30	8	34	27	69
Total	26	132	142	300

d)

Below is a two-way table for the expected frequencies of house age and price level under the null hypothesis of independence.

Expected Frequencies	Price Category (per unit area)			
Age Group (in yrs)	Low (< 20)	Medium ([20,40))	High (≥ 40)	Total
< 15	11.873	60.280	64.847	137
[15,30)	8.146	41.360	44.493	94
≥ 30	5.980	30.360	32.660	69
Total	26	132	142	300

The expected number of observations in the age less than 15 years and low price category is calculated as follows:

Let $\alpha_i = P(house\ in\ ith\ age\ group)$ and $\beta_j = P(house\ in\ jth\ price\ group)$
Where I = 1 (less than 15),…, 3 (greater than or equal to 30)
And j = 1 (low),…, 3 (high)

The maximum likelihood estimates of α_i is given by the total houses in that age group divided by the total number of houses. The maximum likelihood estimates of β_j is given by the total number of houses in that price category divided by the total number of houses.

The expected frequency is given by $e_{ij} = n\widehat{\alpha_{\iota}}\widehat{\beta_{\iota}}$
Thus, the expected number of observations in the age less than 15 years and low price category is

$$\frac{number\ of\ house\ with\ low\ price \times number\ of\ house\ with\ age\ less\ than\ 15}{total\ number\ of\ houses} = \frac{26\times137}{300} = 11.873$$

e)

Step 1
The null hypothesis is H_0: The age of house and price per unit area of house are independent

The alternative hypothesis is H_A: The age of house and price per unit area of house are dependent

Step 2

The observed value of the likelihood ratio statistic is

$$\lambda = 2\sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij}\log\left(\frac{y_{ij}}{e_{ij}}\right)$$

$$= 2[7\log\left(\frac{7}{11.873}\right) + 51\log\left(\frac{51}{60.280}\right) + 79\log\left(\frac{79}{64.847}\right) + 11\log\left(\frac{11}{8.146}\right) + 47\log\left(\frac{47}{41.360}\right) + 36\log\left(\frac{36}{44.493}\right) + 8\log\left(\frac{8}{5.980}\right) + 34\log\left(\frac{34}{30.360}\right) + 27\log\left(\frac{27}{32.660}\right)]$$

$$= 12.1963$$

Step 3

Notice that there are 3 intervals for age and 3 intervals for price. The degree of freedom for the Chi-squared approximation are (3-1)(3-1)=4 and the p-value is given by

$$p - value = P(\Lambda \geq \lambda;\ H_0)$$

$$\approx P(W \geq 12.1963) \quad \text{where } W \sim_{X^2}(4)$$

$$= 0.01594951$$

Step 4

Since 0.01 < p-value < 0.05, we conclude that there is evidence against the hypothesis that the age of house and price per unit area of house are independent based on the observed data.

f)

From the figure in part a, we observe that as the age of house increase, the price per unit area first decrease then increase. The points of the scatterplot exhibit a parabolic shape, though the pattern is not obvious as the points are somewhat scattered. Thus, there may exist a weak positive quadratic relationship between the age and price per unit area of the house. The sample correlation between the two variates is -0.246543. We know that if the sample correlation is close to -1, then there exists a strong negative relationship between the two variates; if the sample correlation is close to 0, then there is no relationship between the two variates. -0.246543 is close two 0, hence the sample correlation demonstrates that there may exist a negative correlation between the two variates, but the relationship is very weak. In fact, the increasing trend on the right end of the scatterplot suggests that a negative linear relationship is unlikely for these two variates. Thus, the two likely has a positive quadratic relationship rather than negative linear relationship. Lastly, from the hypothesis test, we have that $0.01 < p\text{-value} < 0.05$ and we conclude that there is evidence against the hypothesis that the age of house and price per unit area of house are independent based on the observed data.

Altogether, we conclude that the two variates may have a weak positive quadratic relationship.

Name: Amy Lai

ANALYSIS PART 3

Is price explained by distance to nearest metro station?

a)

The graph below is a scatter plot for distance to nearest metro station vs price for the houses sampled.

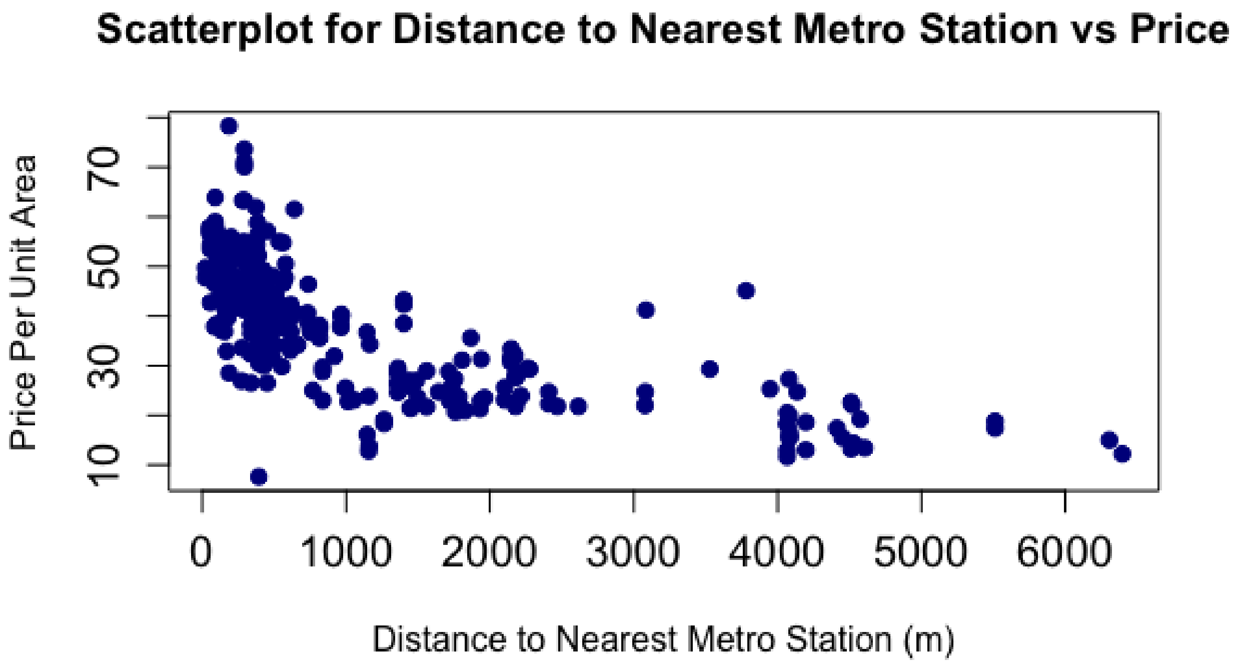


Figure 3. Scatterplot for Distance to Nearest Metro Station vs Price

The correlation coefficient is -0.7037109.
Based on the graph, we see that as distance to nearest metro station increases, the price per unit area decreases constantly with some inconstant variability. Thus, there is possibly a negative linear relationship between distance to nearest metro station and price per unit area. Since the correlation coefficient is negative and close to -1, this also suggests that there exists a fairly strong negative linear relationship between the two variates. Overall the scatterplot and correlation coefficient suggest a possible negative linear association between the two variates.

b)

The following is a summary of key values from the fitted model:

maximum likelihood estimate of the intercept α	45.38016

Name: Amy Lai

maximum likelihood estimate of the slope β	-0.007065337
unbiased estimate of σ	9.20546
estimate of the standard deviation of $\tilde{\beta}$	0.0004132
estimate of the standard deviation of $\tilde{\alpha}$	0.6955547

c)

Step 1

The null hypothesis is $H_0: \beta = 0$. In other words, the null hypothesis is that there is no relationship between the distance to the nearest metro station and the price of the house.

The alternative hypothesis is $H_A: \beta \neq 0$. In other words, the alternative hypothesis is that there is a relationship between the distance to the nearest metro station and the price of the house.

Step 2

Since $H_0: \beta = 0$, we use the test statistic $D = \frac{|\tilde{\beta}-0|}{S_e/\sqrt{S_{xx}}}$, with the observed value $d = \frac{|\hat{\beta}-0|}{s_e/\sqrt{S_{xx}}}$. From the summary function in R, we have $\frac{s_e}{\sqrt{S_{xx}}} = 0.0004132$. Thus, $d = \frac{|-0.007065337-0|}{0.0004132} = 17.09898$. We can also get this value from the R summary command by taking the absolute value of the entry in second row, third column. For the sake of accuracy, we will use d=17.09898.

Step 3

The p-value for testing $H_0: \beta = 0$ is

$$\begin{aligned} \text{p-value} &= 2P(T \geq d) \qquad\qquad \text{T} \sim \text{t(300-2)} = \text{T} \sim \text{t(298)} \\ &= 2P(T \geq 17.09898) \\ &= 2(1 - P(T \leq 17.09898)) \\ &= -2 \times 10^{-16} \\ &\approx 0 \end{aligned}$$

This result can also be obtained from the summary function in R.

Step 4

Since the p-value is less than 0.001, there is very strong evidence against the hypothesis $H_0: \beta = 0$ or the hypothesis of no relationship between the distance to the nearest metro station and the price per unit area.

d) The 90% confidence interval for the slope β is [-0.007747151, -0.006383524]. This means that we are 90% confident that the interval [-0.007747151, -0.006383524] contains the true value of the slope β. In other words, suppose the experiment which was used to estimate β was conducted a large number of times and each time a 90% confidence interval for β was constructed, then approximately 90% of these constructed intervals would contain the true value of β.

e)

A 90% confidence interval for $\mu(1000)$ is [37.43594, 39.1937]. This means that we are 90% confident that the interval [37.43594, 39.1937] contains the true value of the mean house price that has the nearest metro station in 1000m. In other words, suppose the experiment which was used to estimate the mean house price that has the nearest metro station in 1000m was conducted a large number of times and each time a 90% confidence interval was constructed, then approximately 90% of these constructed intervals would contain the true value of the mean house price that has the nearest metro station in 1000m.

f)

The 90% prediction interval for $\mu(1000)$ is [23.10056, 53.52908]. This means that we are 90% confident that the interval [23.10056, 53.52908] contains the true but unknown mean house price that has the nearest metro station in 1000m.

g) Below find the diagnostic plots for this model:

Name: Amy Lai

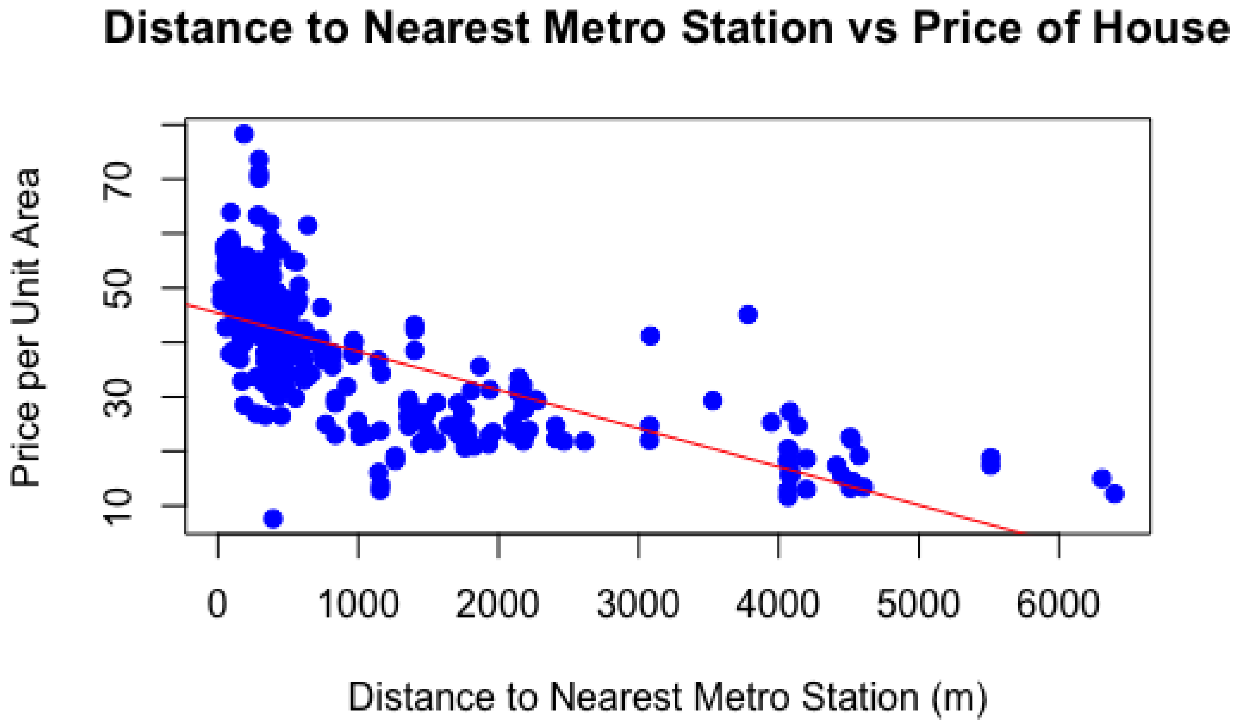
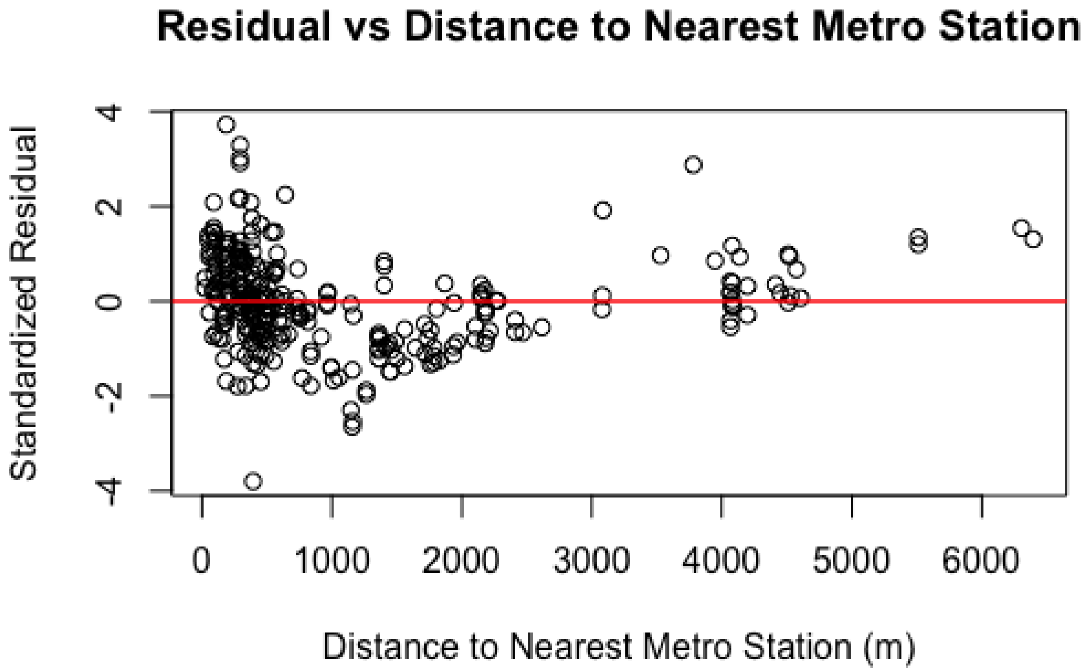


Figure 4. Scatterplot for Distance to Nearest Metro Station vs Price with line of best fit



Name: Amy Lai

Figure 5. Plot of Standardized Residuals versus the Distance to Nearest Metro Station

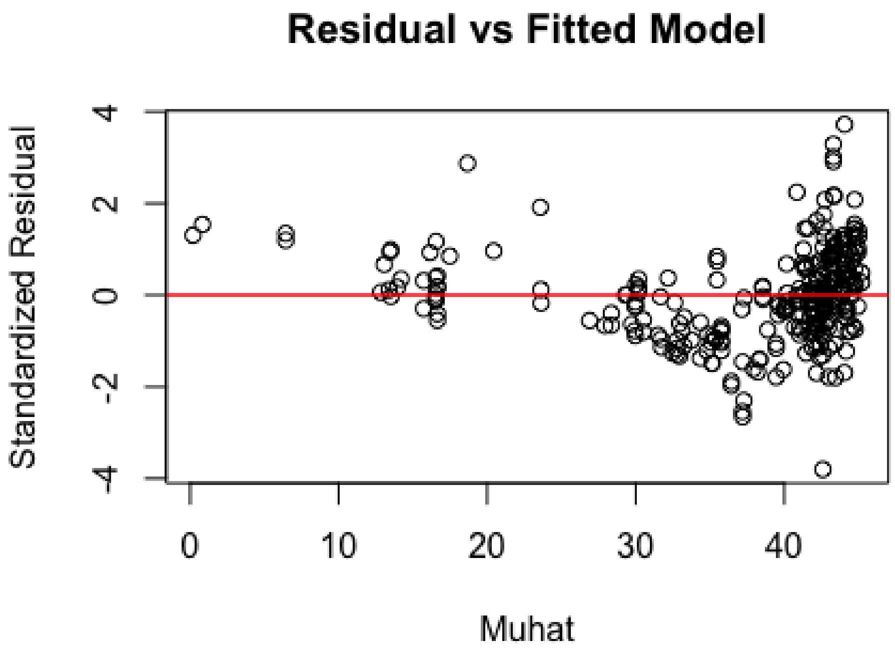


Figure 6. Plot of the Standardized Residuals versus the Fitted Values

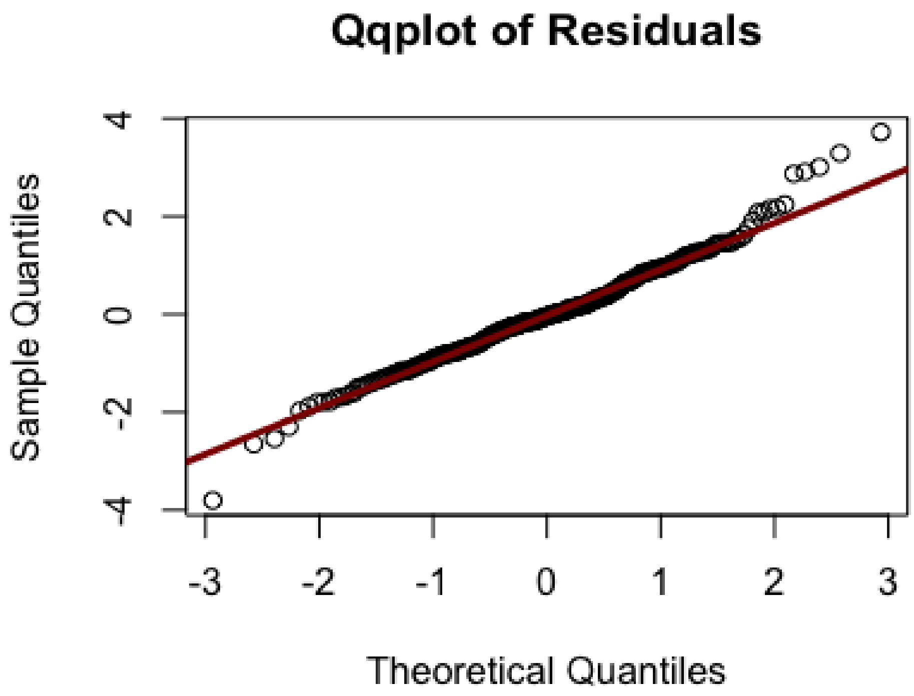


Figure 7. Qqplot for the Standardized Residuals

h)

The scatterplot with fitted line checks if the price per unit can be modeled by a random variable whose mean is a linear function of the distance to nearest metro station and whose standard deviation is constant over the range of distance to nearest metro station. In other words, it checks whether a linear association is ideal for modelling. We hope to see the data points aligned to form straight line relationship (either positive or negative) with the variability about the fitted line being reasonably constant over the range of price. In my data, it can be observed that most points lie on the line of best fit with some obvious variability on the left of the graph and small variability on the right of the graph.

The plot of the standardized residual versus the explanatory variate checks whether the distance to nearest metro station can be modeled by a Gaussian random variable whose mean is a linear function of the distance to nearest metro station and whose standard deviation is constant over the range of the distance to nearest metro station values. In other words, it is used to check for mean of 0, constant variance, and whether a linear relationship is appropriate. If the model is satisfactory, we would expect the points to lie roughly within a horizontal band of constant width between -3 and 3. Approximately half the points should lie on either side of the line where standardized residuals = 0. In my plot, we observe that half of the points are above the line where standardized residuals = 0 and half of the points are below it. However, there are some points with standardized residual higher than 3 or lower than -3. Moreover, there is the issue that the data is not evenly variable across the range of price. A linear association appears appropriate since the data does not display another obvious trend such as quadratic.

Similarly, the plot of the standardized residual versus muhat checks whether the distance to nearest metro station can be modeled by a Gaussian random variable whose mean is a linear function of the distance to nearest metro station and whose standard deviation is constant over the range of the distance to nearest metro station values. In other words, it is used to check for mean of 0, constant variance, and whether a linear relationship is appropriate. If the model is satisfactory, we would expect the points to lie roughly within a horizontal band of constant width between -3 and 3. Approximately half the points should lie on either side of the line where standardized residuals = 0. In my plot, we observe that half of the points are above the line where standardized residuals = 0 and half of the points are below it. However, there are some points with standardized residual higher than 3 or lower than -3. Moreover, there is the issue that the data is not evenly variable across the range of price. A linear association appears appropriate since the data does not display another obvious trend such as quadratic.

The qqplot checks whether the distance to the nearest metro station can be modeled by a Gaussian random variable whose mean is a linear function of the price per unit area and whose standard deviation is constant over the range of values of the price per unit area. If the model is satisfactory, then the points in the qqplot should lie roughly along a straight line with more variability in the tails. In my plot, the points lie in a straight line. Overall, we can say that the Gaussian assumption is ok.

Based on what all the plots are displaying, the price per unit area appears to have a negative linear association with the distance the nearest metro station. However, there is still some problem with a few points in the residual plot being outside the horizontal band of 3 and -3, and variability is not constant, which can be observed in the scatterplot and the residual plots.

i)

There is a negative linear relationship between the variates. Where we find that as the distance to nearest metro station increases, the price per unit area decreases by $0.007065337. I am not surprised by the result because I would expect that the further the nearest metro station, the more inconvenient it is for the residents. Hence, it is reasonable for the price to decrease as the distance to nearest metro station increases.

ANALYSIS PART 4

Is price explained by number of convenience stores within walking distance?

a)

The graph below is the scatter plot for the price per unit area versus the number of nearby stores.

Name: Amy Lai

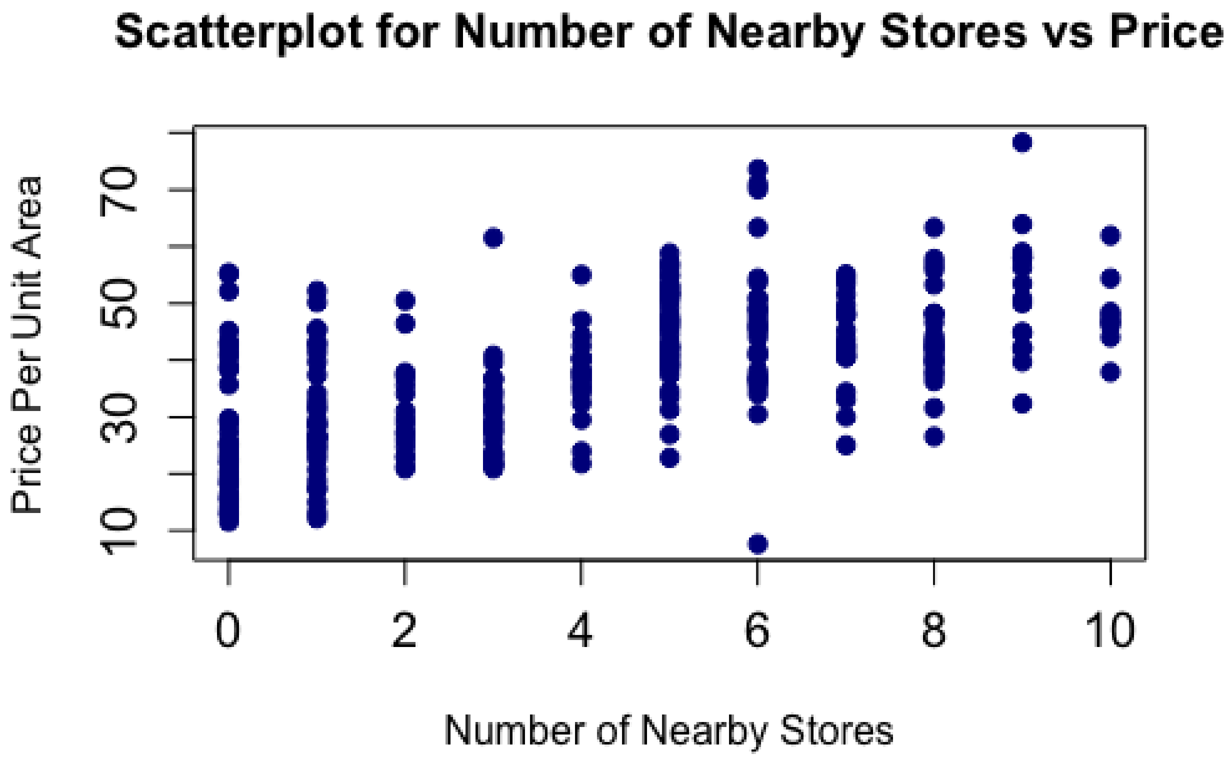


Figure 8. Scatterplot for Number of Nearby Stores vs Price

The correlation coefficient is 0.6175849.
Based on the graph, we see that as the number of nearby store increases, the general trend of price per unit area also increases at a constant rate. Hence, there is a possible positive linear relationship between the number of nearby stores and price per unit area. This can also be evidenced by the correlation coefficient. Since 0.6175849 is positive and close to 1, we say that there is a positive linear relationship between the two variates. Notice that there are some variabilities in the graph and the correlation coefficient is not perfectly 1, thus the relationship is not perfectly positively linear.

b)

The following is a summary of key values from the fitted model:

maximum likelihood estimate of the intercept α	26.73732
maximum likelihood estimate of the slope β	2.65642
unbiased estimate of σ	10.19039
estimate of the standard deviation of $\tilde{\beta}$	0.196

estimate of the standard deviation of $\tilde{\alpha}$	1.001

c)

Step 1
The null hypothesis is $H_0: \beta = 0$. In other words, the null hypothesis is that there is no relationship between the number of nearby stores and the price of the house.

The alternative hypothesis is $H_A: \beta \neq 0$. In other words, the alternative hypothesis is that there is a relationship between the number of nearby stores and the price of the house.

Step 2
Since $H_0: \beta = 0$, we use the test statistic $D = \frac{|\tilde{\beta}-0|}{S_e/\sqrt{S_{xx}}}$, with the observed value $d = \frac{|\hat{\beta}-0|}{s_e/\sqrt{S_{xx}}}$. From the summary function in R, we have $\frac{s_e}{\sqrt{S_{xx}}} = 0.196$. Thus, $d = \frac{|2.65642-0|}{0.196} = 10.19039$. We can also get this value from the R summary command by taking the absolute value of the entry in second row, third column. For the sake of accuracy, we will use d=13.55102.

Step 3
The p-value for testing $H_0: \beta = 0$ is

$$\begin{aligned} \text{p-value} &= 2P(T \geq d) \qquad\qquad T \sim t(300\text{-}2) = T \sim t(298) \\ &= 2P(T \geq 13.55102) \\ &= 2(1 - P(T \leq 13.55102)) \\ &= -2 \times 10^{-16} \\ &\approx 0 \end{aligned}$$

This result can also be obtained from the summary function in R.

Step 4
Since the p-value is less than 0.001, there is very strong evidence against the hypothesis $H_0: \beta = 0$ or the hypothesis of no relationship between the number nearby stores and the price per unit area.

d) Below find the diagnostic plots for this model:

Name: Amy Lai

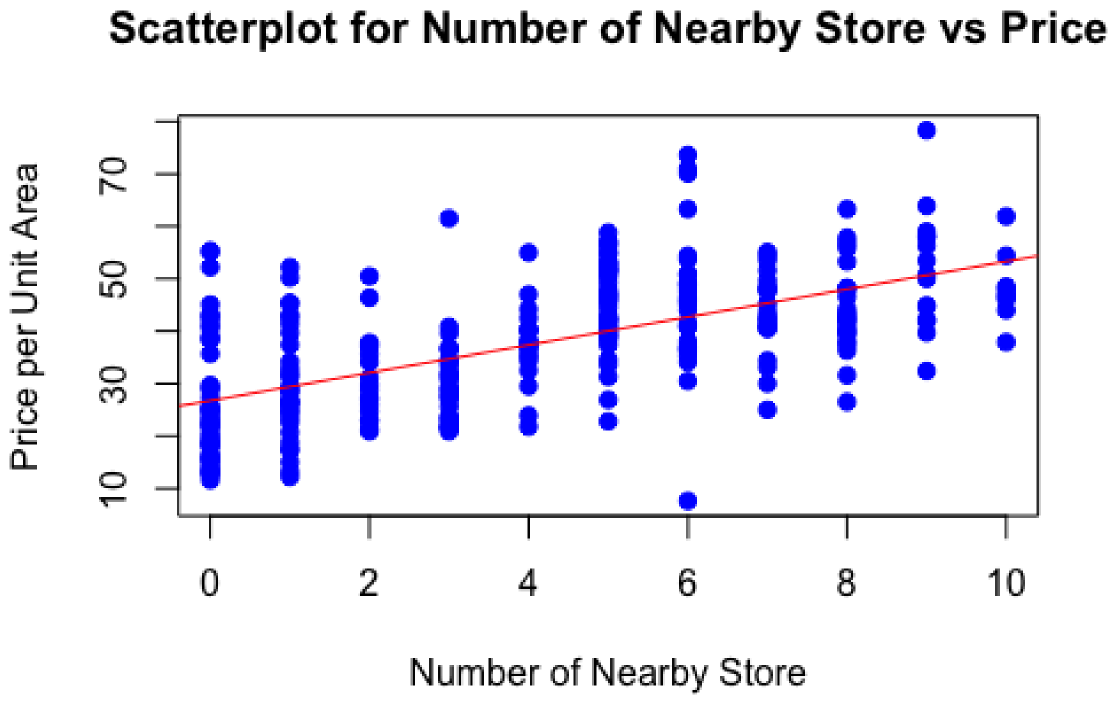


Figure 9. Scatterplot for Number of Nearby Store vs Price

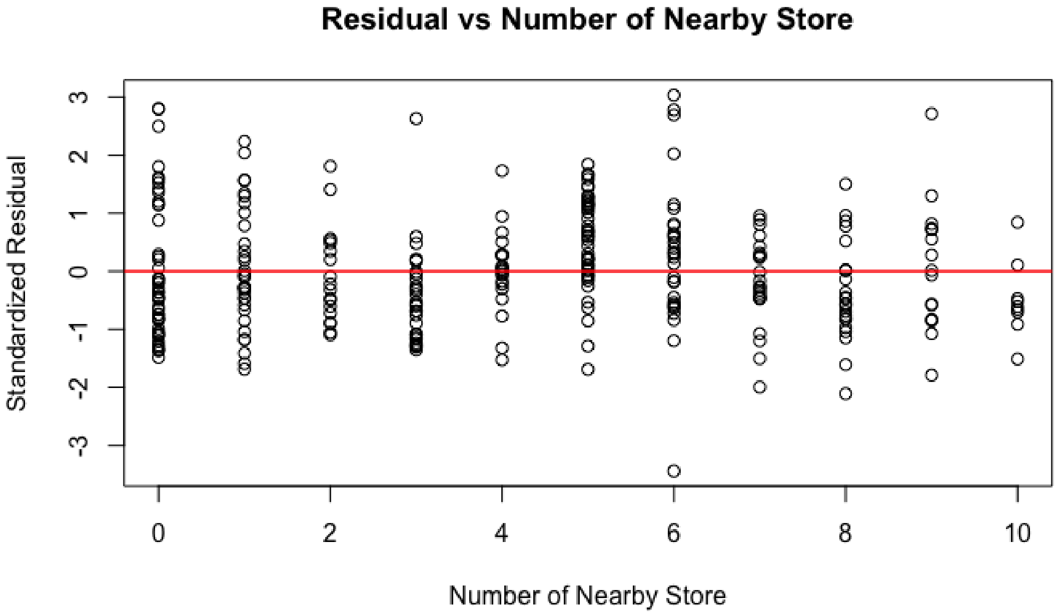


Figure 10. Plot of the Standardized Residuals versus the Number of Nearby Station

Name: Amy Lai

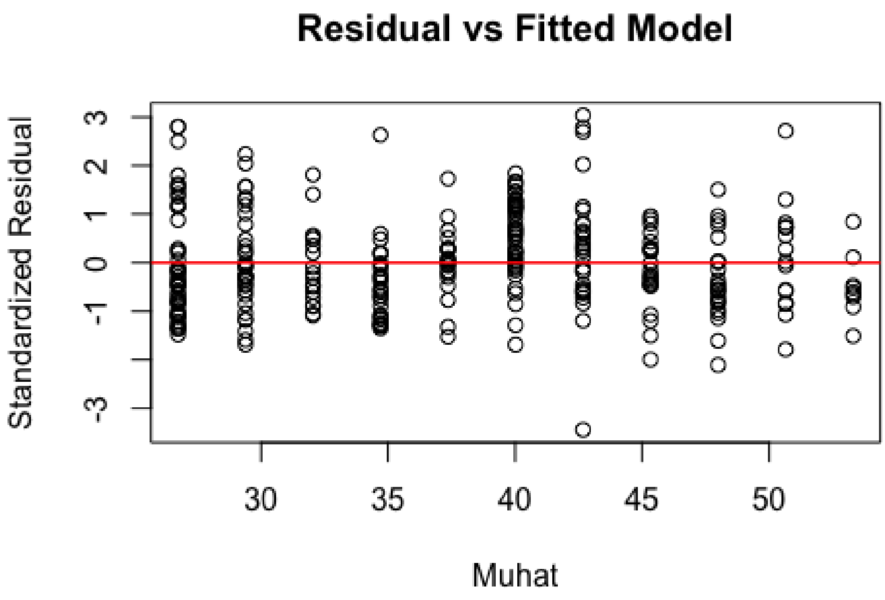


Figure 11. Plot of the Standardized Residuals versus the Fitted Values

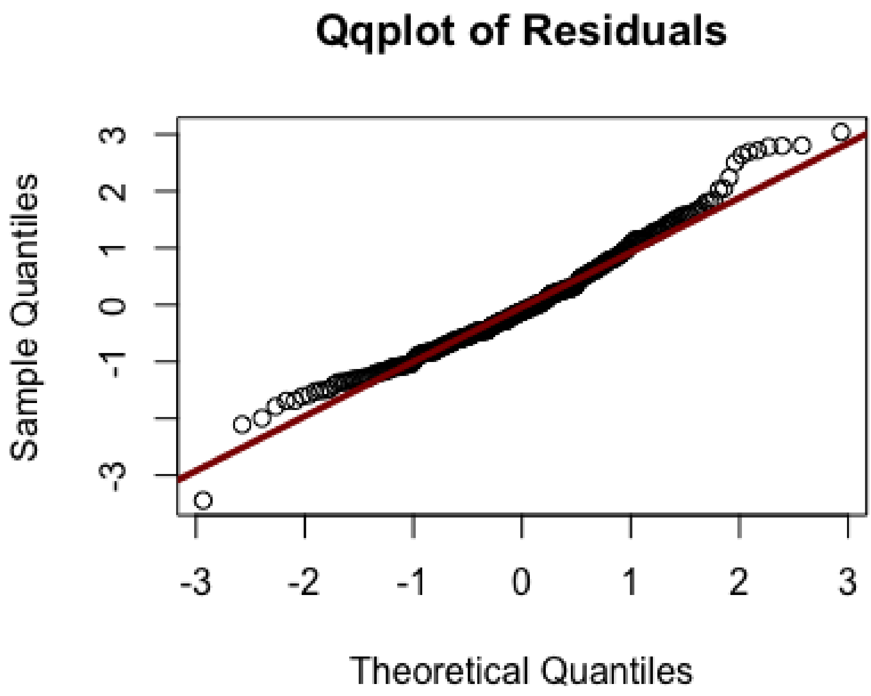


Figure 12. Qqplot with line of the standardized residuals

Name: Amy Lai

e)

The scatterplot with fitted line checks if price per unit area can be modeled by a random variable whose mean is a linear function of the number of nearby stores and whose standard deviation is constant over the range of number of nearby stores. In other words, it checks whether a linear association is ideal for modelling. We hope to see the mean price per unit at each number of nearby stores aligned to form straight line relationship (either positive or negative) with the variability about the fitted line being reasonably constant over the range of price. In my data, it can be observed that the line of best fit crosses the middle price value at each number of nearby stores. However, the variability is not constant as the number of nearby store changes. For example, the variability of price when the number of nearby stores is 6 is greater than the variability when the number of nearby stores Is 10. Overall, the scatterplot exhibits a positive linear association with some violation against constant variability.

The plot of the standardized residual versus the explanatory variate checks whether the price per unit area can be modeled by a Gaussian random variable whose mean is a linear function of the number of nearby stores and whose standard deviation is constant over the range of values of the number of nearby stores. In other words, it is used to check for mean of 0, constant variance, and whether a linear relationship is appropriate. If the model is satisfactory, we would expect the points to lie roughly within a horizontal band of constant width between -3 and 3. Approximately half the points should lie on either side of the line where standardized residuals = 0. In my plot, we observe that half of the points are above the line where standardized residuals = 0 and half of the points are below it. However, there are some points with standardized residual higher than 3 or lower than -3. Moreover, there is the issue that the data is not evenly variable across the range of price. A linear association appears appropriate since the data does not display another trend such as quadratic

Similarly, the plot of the standardized residual versus the fitted values checks whether the price per unit area can be modeled by a Gaussian random variable whose mean is a linear function of the number of nearby stores and whose standard deviation is constant over the range of values of the number of nearby stores. In other words, it is used to check for mean of 0, constant variance, and whether a linear relationship is appropriate. If the model is satisfactory, we would expect the points to lie roughly within a horizontal band of constant width between -3 and 3. Approximately half the points should lie on either side of the line where standardized residuals = 0. In my plot, we observe that half of the points are above the line where standardized residuals = 0 and half of the points are below it. However, there are some points with standardized residual higher than 3 or lower than -3. Moreover, there is the issue that the data is not evenly variable across the range of price. A linear association appears appropriate since the data does not display another trend such as quadratic.

The qqplot checks whether the price per unit area can be modeled by a Gaussian random variable whose mean is a linear function of the number of nearby stores and whose standard deviation is constant over the range of values of the number of nearby stores. If the model is satisfactory, then the points in the qqplot should lie roughly along a straight line with some variability in the tails. In my plot, the points lie in a straight line with some slight variability on both ends. Overall, we can say that the Gaussian assumption is ok.

Based on what all the plots are displaying, the price per unit area appears to have a positive linear association with the number of nearby stores. However, there is still some problem with a few points in the residual plot being outside the horizontal band of 3 and -3. Additionally, variability is not constant, which can be observed in the scatterplot and the residual plots.

f)

There is a positive linear relationship between the variates. Where we find that for every one additional nearby store, the price per unit area increases by $26.73732. I am not surprised by the result because I would expect that more stores will attract more interest for the house. Hence, it is reasonable for the price to increase as the number of nearby stores increases.

Name: Amy Lai

CONCLUSION

Based on the analysis, we conclude that the price of a house is associated with the age of the house, the distance to the nearest public transit, and the number of convenience stores within walking distance. In particular, the relationship between house age and price seems to be positively quadratic based on the scatterplot. This means that as house age increase, the price will first decrease then increase. The sample correlation for price and house age is also close to 0, thus a linear regression model is not appropriate for this relationship. On the other hand, the relationship between distance to nearest metro station and price is a negative linear relationship, indicating that as distance increase, the price will decrease constantly. This negative linear relationship is relatively strong because as the correlation coefficient is close to -1. In general, a linear regression model is appropriate for this relationship, however, it should be noted that the assumption of constant variability for a linear regression model seems to be violated. Likewise, the relationship between the number of nearby store and price is also linear, but it is positively linear. This means that as the number of nearby store increase, price will also increase at a constant rate. Based on the analysis, we conclude that a linear regression model is appropriate for this relationship. Though, there is, again, a slight violation against the constant variability assumption.

We cannot conclude that there are causal relationships between the explanatory variates and the response of house price from this observational study. It is generally hard or impossible to derive causative conclusions from observational study. In more detail, this study only examines 300 samples from one district of Taiwan. Thus, there is a chance that the observed associations are due to defect in the study or a peculiarity in the district. To establish a causal relationship, a strong association would need to be observed in numerous studies at different places of Taiwan. Moreover, the study did not control potential real “causes”, or other confounding variables, of the relationship. For example, the effect of distance to school significantly affects house prices. We do not know if an increase in price is due to the distance to school rather than the explanatory variates examined in this study. Similarly, we see a linear association between price and both the distance to nearest metro station and the number of stores nearby. Thus, we cannot be certain that a price increase is caused by closer distance to metro station because it can be caused by more stores nearby. In addition, we do not have scientific support or other pathways established to explain the causal relationship between the explanatory variates and the house price. For example, some houses in the suburban areas (away from metro stations) may have higher price because it is quieter. Thus, there is no real scientific explanation behind why the explanatory variates will necessarily cause the price to go up or down. Furthermore, we do not have consistent response between the house age and price. Hence, there is no causative relationship between these two variates. All in all, we cannot that there are causal relationships between the explanatory variates and the response of house price from this observational study.