STAT 231 Spring 2021 Final Project

SOLUTIONS

Submissions that do not populate the provided template will receive a 10% penalty.

Answers **must be typed** with the exception of certain question parts (this is indicated in the solutions). **Outside of these exceptions, submitted answers which are not typed will not be marked and given a mark of zero.**

The report must be professionally formatted and written as if presented to an audience without detailed R knowledge. No R code should be included in the report. **Submissions that include R code will receive a 10% penalty.**

Values submitted are to be checked in R for all booklets – answers which do not match receive 0 marks for that part of the question.

Projects which are left as a single document and not uploaded to the appropriate places in Crowdmark will be assigned a 10% penalty.

Booklet numbers that require a 10% penalty overall (issues in red above) should be reported to Matthew and he can apply them.

Marks for Final Project:

Section	Marks
Analysis Q1	23
Analysis Q2	15
Analysis Q3	22
Analysis Q4	17
Conclusion	10
PDF, R file, Dataset LEARN dropboxes	3
Total	90

ANALYSIS QUESTION 1 (23 MARKS)

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

a) (3 marks) The table below contains the numerical summaries for the MRT_distance variate.

Minimum	23.38	IQR	1163.946
1st Quartile	287.30	Range	6464.64
Sample Median	490.52	Sample Mean = $\bar{y}$	1084.71
3rd Quartile	1451.24	Sample Standard Deviation = s	1294.16
Maximum	6488.02	Sample Skewness	1.96

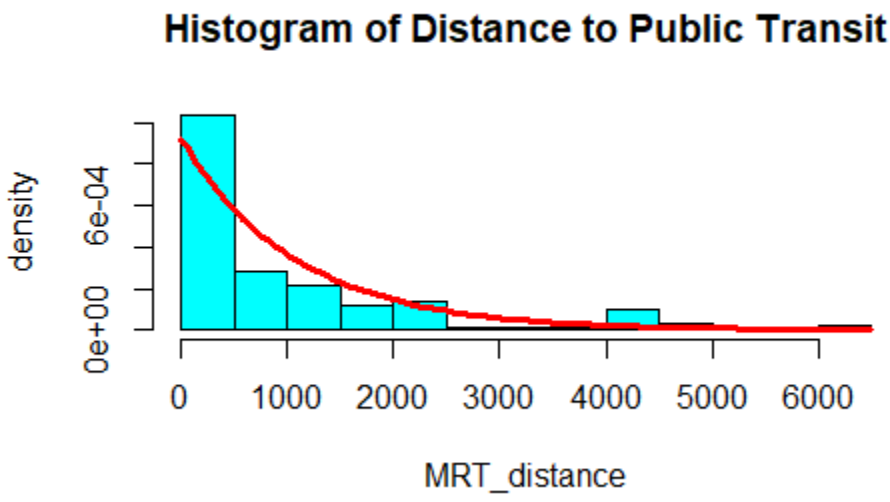
b) (2 marks)

The maximum likelihood estimate for theta is the sample mean, or 1084.71 meters.

This is the mean distance to the nearest metro station for houses sold in Sindian Dist,, New Taipei City, Taiwan from 2012-2014.

c) (2 marks)

The image below shows the histogram for the distance to the nearest metro station with an Exponential(θ = 1084.71) pdf overlaid.



d) (7 marks) 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)	58	50.5
2	[200, 400)	71	42.0
3	[400, 600)	42	34.9
4	[600, 800)	17	29.1
5	[800, 1000)	10	24.2
6	[1000, 1200)	12	20.1
7	≥ 1200	90	99.2
Total		300	300

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

We find the probability of an observation falling into each of the bins using the Exponential(1084.71) cdf.

For example, for [0, 200):

$$P(X \leq 200) = 1 - e^{-\frac{200}{1084.71}} = 0.1684$$

For [200, 400):

$$P(X \leq 400) - P(X \leq 200) = e^{-\frac{200}{1084.71}} - e^{-\frac{400}{1084.71}} = 0.1400$$

And similarly for the other bins.

Then we multiply by the total observations n = 300. I.e.

$$e_1 = 300 * 0.1684 = 50.52$$
$$e_2 = 300 * 0.1400 = 42$$

Etc.

e) (4 marks)

Step 1: Our null hypothesis is H_0: An exponential model fits this data
And the alternate hypothesis is H_A: An exponential model is not a good fit for this data.

Step 2: The test statistic is $\lambda = 2\sum_{j=1}^{7} f_j \, ln\left(\frac{f_j}{e_j}\right)$ = 40.20

Step 3: The p-value is calculated as
$p = P(W \geq 40.20)$ for $W \sim \chi^2(5)$
Where the 5 degrees of freedom arise from the 7 – 1 – 1 for the 7 bins, one restriction on the multinomial framework and the estimation of theta for the exponential model.

Step 4: We get p-value = 1.3×10^{-7} and therefore conclude there is very strong evidence against the hypothesis that the data is well described by an exponential model.

f) (5 marks)

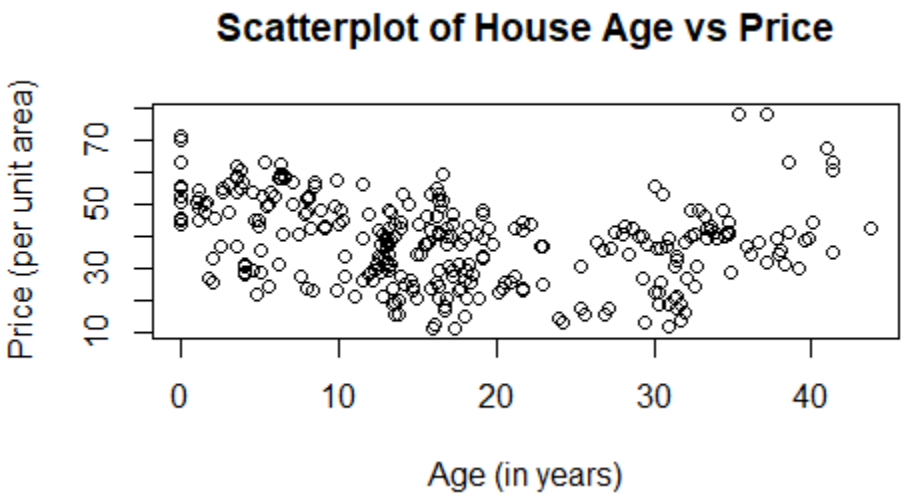
In the numerical summaries above we see some indications that the Exponential model might be appropriate: the median is less than the mean, and the skewness is positive or right skewed. However, when examining the histogram, we see that the agreement with the Exponential pdf is not great. There is more density near 0 and the far right of the tail, and less in the middle of the range, than would be expected for an Exponential distribution. Further, the goodness of fit test provides very strong evidence against the null hypothesis of an Exponential model being appropriate for the data. Therefore, we conclude that the Distance to Nearest Metro Station variate is **not** well described by the Exponential distribution.

ANALYSIS QUESTION 2 (15 MARKS)

Is price explained by age of the house?

a) (1 marks)

The following is a scatterplot of house price versus age.



b) (1 mark)

The correlation between house age and price is -0.23.

c) (2 marks)

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	4	53	82	139
[15,30)	13	50	33	96
≥ 30	8	31	26	65
Total	25	134	141	300

d) (3 marks)

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	11.6	62.1	65.3	139
[15,30)	8.0	42.9	45.1	96
≥ 30	5.4	29.0	30.6	65
Total	25	134	141	300

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

To find the expected frequencies under the null hypothesis of independence, we use the result that if $P(A_iB_j) = P(A_i)P(B_j)$, then $e_{ij} = \frac{r_i c_j}{n}$. Or, we multiply the row total by the column total for the corresponding rows and columns of the category and divide by total observations.

Thus, for <15 year age and <20 price/unit area, we have (139*25)/300 = 11.583.

e) (4 marks)

Step 1: Our null hypothesis is H_0: The house age and price per unit area are independent
And the alternate hypothesis is H_A: House age and price per unit area are not independent

Step 2: The test statistic is $\lambda = 2\sum_{i=1}^{3}\sum_{j=1}^{3} f_{ij}\, ln\left(\frac{f_{ij}}{e_{ij}}\right)$ = 21.25
Where f_{ij} are the observed frequencies and e_{ij} are the expected frequencies.

Step 3: The p-value is calculated as
$p = P(W \geq 21.25)$ for $W \sim \chi^2(4)$
Where the 4 degrees of freedom arise from 9 – 1 – 2 – 2 or (3-1)(3-1).

Step 4: We get p-value = 2.83×10^{-4} and therefore conclude there is very strong evidence against the hypothesis that house age and price are independent.

f) (4 marks)

In the scatterplot, we see a reasonable amount of random scatter and the correlation indicates a negative but somewhat weak linear relationship between age of house and

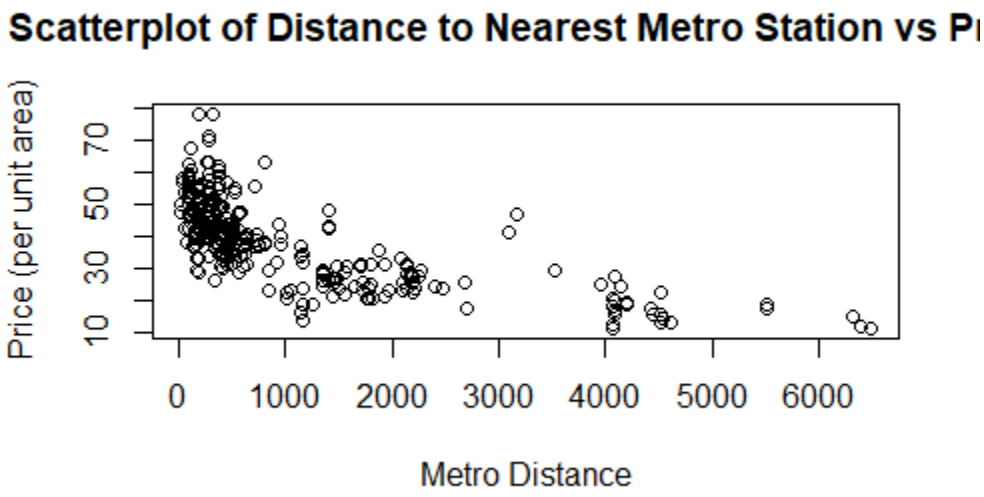
price. However, the scatterplot shows a quadratic relationship might be more appropriate. Further, there was strong evidence against the null hypothesis of independence between the two variates. Therefore, I conclude there is a relationship between age of house and price per unit area, and that the relationship is quadratic.

ANALYSIS QUESTION 3 (22 MARKS)

Is price explained by distance to nearest metro station?

a) (2 marks)

Examining the following scatterplot of the two variates:



Combined with the sample correlation value of -0.71, we conclude that there is a strong negative **linear** relationship between the price of a house and the distance to public transit.

b) (3 marks)

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

maximum likelihood estimate of the intercept α	45.96
maximum likelihood estimate of the slope β	-0.007
unbiased estimate of σ	9.29
estimate of the standard deviation of $\tilde{\beta}$	0.0004
estimate of the standard deviation of $\tilde{\alpha}$	0.7006

c) (4 marks)

Step 1: Our null hypothesis is H_0: $\beta = 0$ or no relationship between distance to transit and price of a house in Taiwan

And the alternate hypothesis is H_A: $\beta \neq 0$

Step 2: The test statistic is $d = \frac{|\widehat{\beta}-0|}{s_e/\sqrt{S_{xx}}}$ = 17.32 (-17.32 in R where the absolute value is dropped)

Step 3: The p-value is calculated as
$p = 2 * P(T \geq 17.32)$ for $T \sim t(298)$, i.e. t distribution on 298 degrees of freedom.

Step 4: We get p-value ≈ 0 therefore conclude there is very strong evidence against the hypothesis of no relationship between distance to public transit and house price.

d) (1 mark)

A 90% confidence interval for β is (-0.0079, -0.0065).

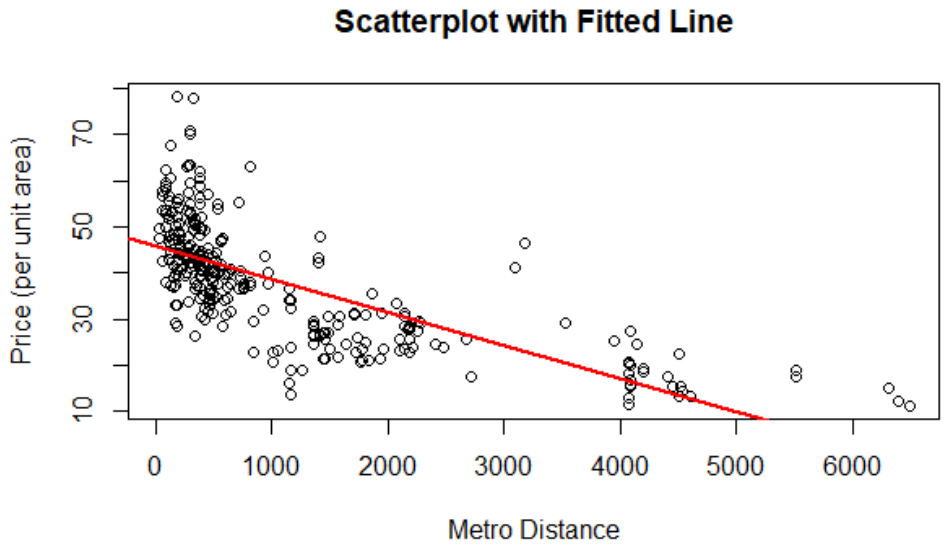
e) (2 marks)

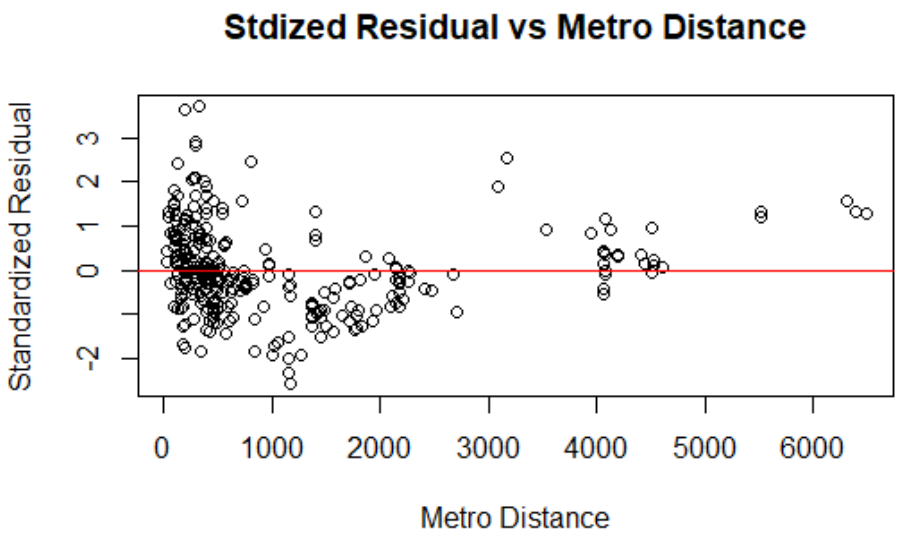
The estimate for the average price per unit area when the distance to the nearest metro station is 1000m is 38.77. A 90% confidence interval for this value is (37.88, 39.65).

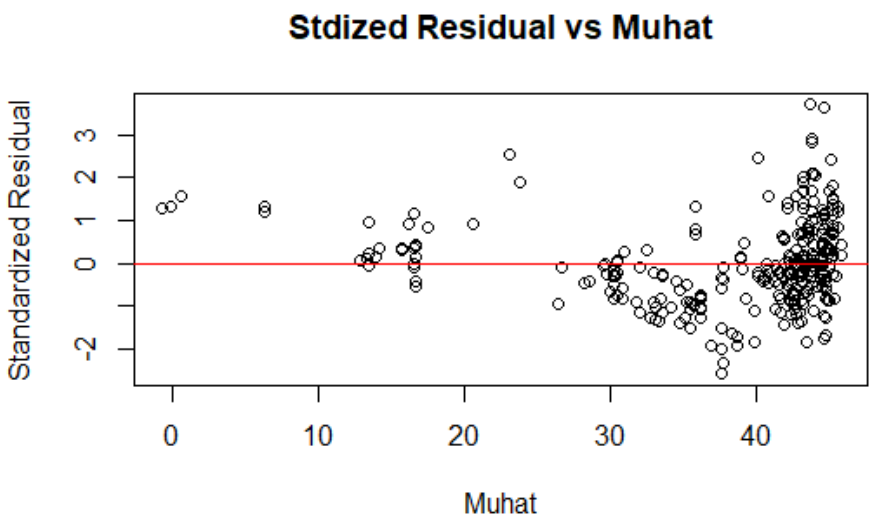
f) (2 marks)

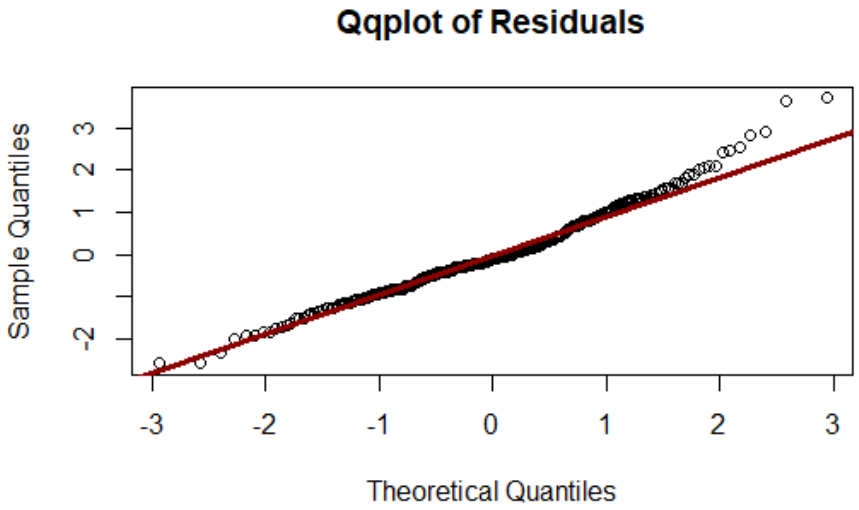
The estimate for price per unit area of an individual house that is 1000m from the nearest metro station is also 38.77. The corresponding 90% prediction interval is (23.41, 54.13).

g) (2 marks) Below find the diagnostic plots for this model:




Stdized Residual vs Metro Distance
Standardized Residual
-2
0
1
2
3
0
1000
2000
3000
4000
5000
6000
Metro Distance


Stdized Residual vs Muhat
Standardized Residual
-2
0
1
2
3
0
10
20
30
40
Muhat


Qqplot of Residuals
Sample Quantiles
-2
0
1
2
3
-3
-2
-1
0
1
2
3
Theoretical Quantiles

h) (4 marks)

While the scatterplot indicates a linear relationship between the two variates, the assumption of constant variance is clearly violated. In the residuals vs x or fitted values, we see points that are not randomly scattered: there is significant heteroskedasticity and grouping of points. There are potentially a few outliers to investigate based on some values do not lie evenly between (-3,3) in the residual plots and the one tail of the qqplot, but the mean of 0 and normal assumption are relatively reasonable. However, due to the issue with the variance, we conclude that the linear regression model is not a good fit to the data.

i) (2 marks)

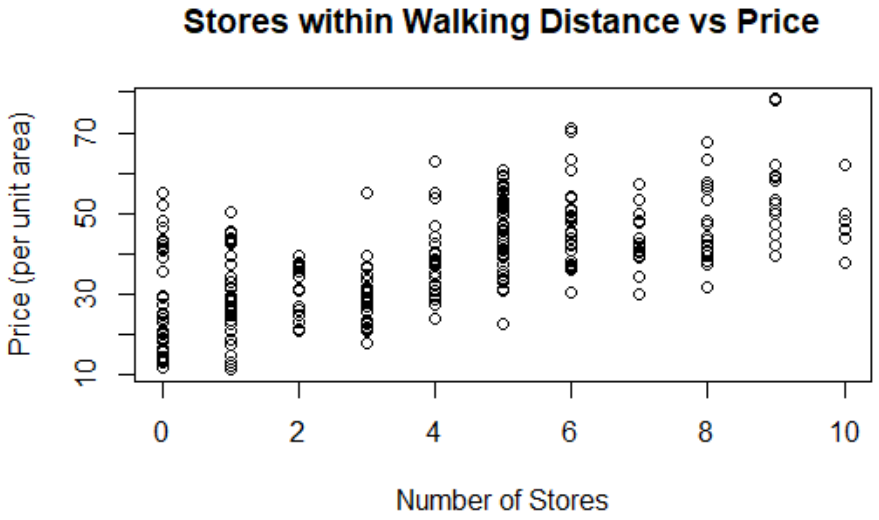
There is a negative linear relationship between the price per unit area of a house and the distance to the nearest metro station. Our estimate of beta indicates that for every 1-meter increase in the distance to public transit, the **average** price per unit area decreases by $0.007. This is not surprising as one would expect houses that are closer to public transit to be more desirable and thus cost more to purchase.

ANALYSIS QUESTION 4 (17 MARKS)

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

a) (2 marks)

We have the following scatterplot for the price per unit area of a house versus the number of convenience stores nearby.



And we find that the sample correlation is 0.64.

Thus, we conclude that there is strong positive **linear** relationship between the two variates.

b) (3 marks)

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

maximum likelihood estimate of the intercept α	26.32
maximum likelihood estimate of the slope β	2.90
unbiased estimate of σ	10.14
estimate of the standard deviation of $\tilde{\beta}$	0.20
estimate of the standard deviation of $\tilde{\alpha}$	1.01

c) (4 marks)

Step 1: Our null hypothesis is H_0: $\beta = 0$ or no relationship between number of nearby stores and price of a house in Taiwan
And the alternate hypothesis is H_A: $\beta \neq 0$

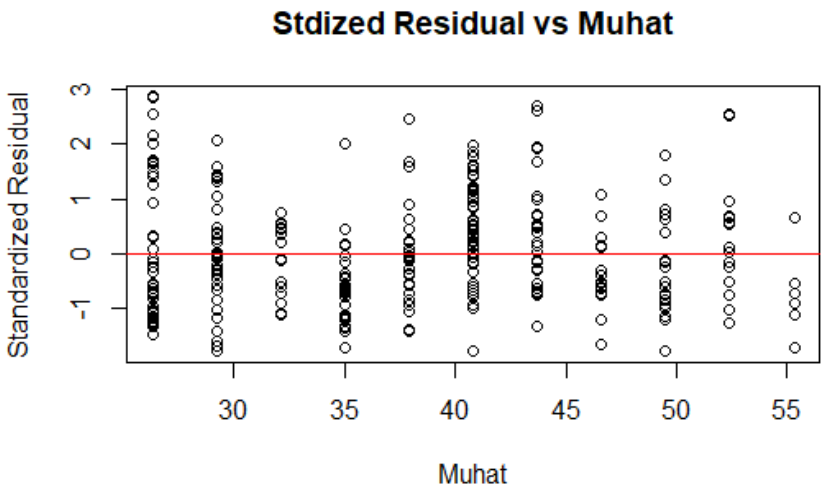
Step 2: The test statistic is $d = \frac{|\hat{\beta}-0|}{s_e/\sqrt{S_{xx}}}$ = 14.30

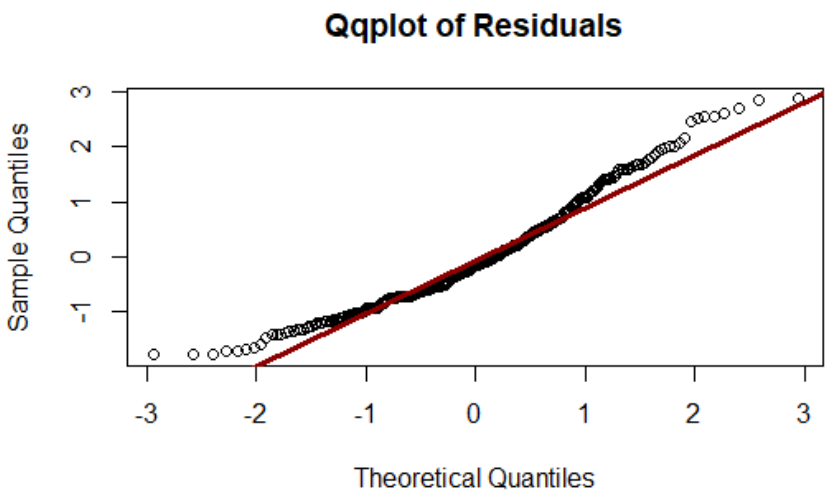
Step 3: The p-value is calculated as
$p = 2 * P(T \geq 14.30)$ for $T \sim t(298)$, i.e. t distribution on 298 degrees of freedom.

Step 4: We get p-value ≈ 0 therefore conclude there is very strong evidence against the hypothesis of no relationship between number of nearby stores and house price.

d) (2 marks) Below find the diagnostic plots for this model:


Stdized Residual vs Stores
Standardized Residual
-1
0
1
2
3
0
2
4
6
8
10
Stores in Walking Distance


Stdized Residual vs Muhat
Standardized Residual
-1
0
1
2
3
30
35
40
45
50
55
Muhat


Qqplot of Residuals
Sample Quantiles
-1
0
1
2
3
-3
-2
-1
0
1
2
3
Theoretical Quantiles

e) (4 marks)

In checking the scatterplot and residual vs stores or residual vs fitted plots, we see that the assumptions of a linear relationship and constant variance appear satisfied (note that the grouping occurs since stores are a discrete variate, so this grouping is not a concern). The assumption of normality is a little more concerning given the residuals do not appear to be evenly distributed around 0/ within (-3,3), and the qqplot has a slight "u" shape potentially. The deviations are not too extreme in both cases, so I conclude that the model is okay.

Note: some students may suggest a transformation is necessary to correct for the issues with the normal assumption, and that would also be accepted as correct.

f) (2 marks)

There is a positive linear relationship between the price per unit area of a house and the number of convenience stores within walking distance. Our estimate of beta indicates that for every additional store nearby, the **average** price per unit area increases by $2.9. This is not surprising as one would expect houses that are closer to stores to be more desirable and thus cost more to purchase.

CONCLUSION (10 marks):

In this study we examined data on house prices and potential explanatory variates for the price of a house in Taiwan around the years 2012 – 2014. In the analysis above we first established that the distance to the nearest metro station was not well modeled by an Exponential model. However, we then focused on our main objective of investigating the relationship between price and potential explanatory variates of age of the house, the distance to the nearest metro station, or the number of convenience stores nearby. All 3 of these variates showed a strong relationship with the price per unit area. For the age of the house, we saw that a quadratic relationship may be appropriate and concluded that there was very strong evidence against the hypothesis of independence between house age and price. For the distance to the nearest metro station, we again saw a strong relationship. It was clear there was a negative linear relationship between the variates, but the use of a simple linear regression model may be inappropriate based on the model diagnostics. The relationship between house price and nearby stores was strong and a positive linear relationship. For this pairing, the simple linear regression model was deemed appropriate.

That said, we can only conclude there are associations between the explanatory variates examined and the price of a house, we cannot conclude direct causal relationships. I.e. it is inappropriate to claim that being further from public transit causes the house price to be lower. There are many reasons this is inappropriate. The biggest issue is that the study was observational and not experimental, so the explanatory variates were not controlled. This creates issues since the association could be due to both the explanatory and response variates responding to a change in another unobserved or lurking variate. Or, there might be other confounding variates that make it impossible to separate out the effects one of the explanatory variates. To establish causation with an observational study, we would need to replicate the results in many different observation studies which attempt to control for the confounding variates.