|  |  |
| --- | --- |
| Real estate valuation  New Taipei City, Taiwan | daniel White  Friday, 12 PM  kenneth liu  Friday, 9 AM  Sophia Krafcik  Thursday, 2 PM |

**ABSTRACT**

This project focuses on building real estate valuation models, with six different predictors used in our data set to predict the house price of unit area in New Taipei City, Taiwan. The predictors include the following: transaction date, house age (year), distance to the nearest MRT station (meter), the number of convenience stores in the living circle on foot (integer), latitude (degree), and longitude (degree). One question of interest we answer in our regression analysis is which (if any) of the 6 predictors are useful in predicting the house price per unit in New Taipei City, Taiwan. Once we have concluded which predictors should be included to predict house price per unit, we predict what the house price per unit would be if we were valuing a house with the average values for transaction date, house age, distance to nearest MRT station, number of convenience stores, latitude and longitude based on the model. Under the same assumptions for the predictors, we then calculate a confidence interval for where the estimate of the true median of house price per unit lies.

**INTRODUCTION**

This project discusses the house price of unit area in Sindian Dist., New Taipei City, Taiwan and the factors, or predictors, that influence the price of the housing in this area. The following six factors that the project will focus on are the transaction date (in years), the house age (in years), the distance to the nearest MRT station (in meters), the number of convenience stores in the living area, latitude (in degrees), and longitude (in degrees). The purpose of this project is to compare the factors and see which factors create the most impact to the prices of the housing. In addition, we want to see if there are any factors that are not very influential. This is important to know because if a factor does not really influence the price that much, people will not have to strongly consider the insignificant factor when trying to find housing. Therefore, based on which factors are important, we can predict the most accurate, representative prices for the people living there. Furthermore, the project can give information about the factors to the people living there so that the people can determine which areas within the housing are the most suitable to them.

**QUESTIONS OF INTEREST**

1. From the 6 predictors given in the data set (transaction date, house age, distance to the nearest MRT station, number of convenience stores, latitude, and longitude), which are useful in predicting house price per unit in New Taipei City, Taiwan?
2. Given the average of transaction date, house age, distance to the nearest Mass Rapid Transit (MRT) station, number of convenience stores, latitude, and longitude, can we predict the house price per unit?
3. Based on our model, can we estimate the true “average” house price per unit in New Taipei City, Taiwan?

**DATA AND REGRESSION METHODS**

Our source of our data is in the UC Irvine Machine Learning Repository, and the data is about predicting the price of housing in Sindian Dist., New Taipei City, Taiwan by six factors, or predictors. The data is created by Prof. I-Cheng Yeh, who is in the Department of Civil Engineering in Tamkang University, Taiwan. There are seven variables involved in the data: the response variable, which is the house price per unit area (10000 New Taiwan Dollar/Ping where 1 Ping = 3.3 meter2) and the six predictor variables, which are the transaction date - a discrete variable, house age (year) - a continuous variable, distance to nearest Mass Rapid Transit (MRT) station (meter) - a continuous variable, number of convenience stores in the living circle on foot (integer) - a discrete variable, geographical latitude (degree [coordinate]) - a continuous variable, and geographical longitude (degree [coordinate]) - a continuous variable.

For the first question of interest, we answer it by first checking if the initial model with all the predictor variables and the response variable unchanged and not transformed satisfies the four assumptions of linear regression by comparing added-variable plots, residual vs fitted values plots, and Q-Q plots of each individual predictor to the response variable. The four assumptions of linear regression are linearity, independence, equal variance, and normality. In this case, we must assume independence of the data is not violated. We check marginal relationships with scatterplots of each variable, linearity with added-variable plots, constant variance with residual vs fitted values plots, and normality with Q-Q plots. If the initial model violates any of the assumptions, we will use statistical methods such as BoxCox function to transform the variables to fix the violations of the assumptions. Once we have a model that does not violate any of the assumptions, we will use a statistical test called the global F-test to check if at least one predictor is useful. Finally, we will use backward selection and use the regsubsets function to test if there are any simpler models that are better than the full model, and therefore we can conclude by our final model which predictors are useful.

For the second question of interest, after getting our final model and finding the average of each individual predictor, we can predict the house price per unit by using the predict function with the newly fitted transformed model and the means of each individual predictor to calculate the predicted average house price per unit.

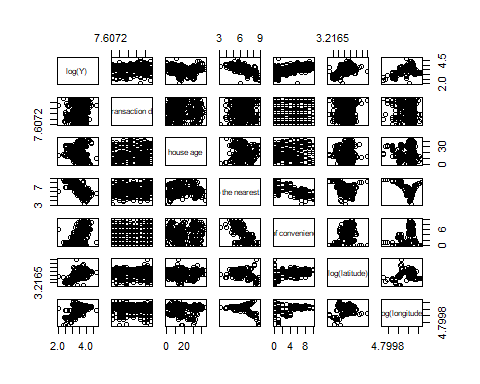
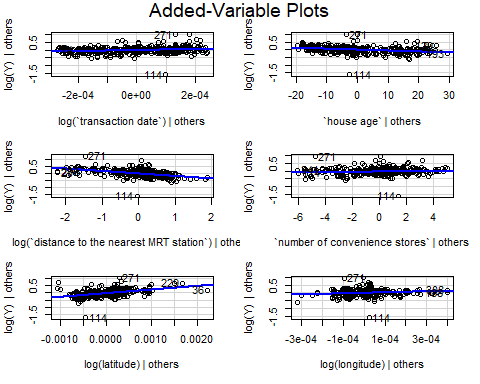
For the third question of interest, we will use the predict function to find a 95 percent confidence interval to give the range of the “average” house price per unit for a house in Taipei City, Taiwan with average transaction date, house age, distance to nearest MRT station, number of convenience stores, latitude, and longitude.

Based on the residual vs fitted values plots and Q-Q plots found in Appendix 2, when comparing each individual predictor variable to the response variable in the initial model and the scatterplots with pairs function and added-variable plots, there will be a need to do variable transformations. For example, after creating scatterplots with pairs function, there are no marginal relationships between any pair of variables. In addition, the added-variable plots show that the untransformed predictor variables and response variable have weak, if any, linear relationships in between each individual predictor and response variable. Therefore, the linearity assumption is violated.  Based on the residual vs fitted values plots when comparing each individual predictor variable to the response variable in the initial model, constant variance assumption is violated among the predictor variables: ‘house age’, ‘distance to nearest MRT station’, ‘latitude’, and ‘longitude’. Based on the Q-Q plots when comparing each individual predictor variable to the response variable, the normality assumption is violated among all the predictor variables due to outliers and having most of the plots being skewed, especially to the right. Therefore, we need to do transformations on the variables due to violations of linearity, constant variance, and normality. From the BoxCox function plot, we use logarithmic transformations, due to lambda being close to 0, first on the response variable. After producing the same plots to test assumptions, we notice the same violations. We then proceed to transform all predictor variables except ‘house age’ and ‘number of convenience stores’ due to both of these predictor variables having 0’s in their data, which falls outside of the domain of the logarithmic scale. After running the tests for a third time, we find that the only predictor still violating the assumptions is longitude, which heavily violates constant variance and normality. We decide to remove it from our model while keeping all other predictors which do not violate assumptions.

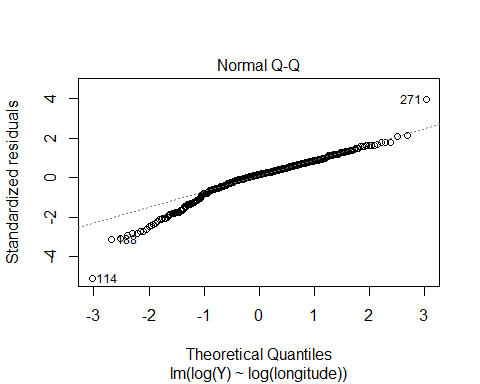
From the exploratory analysis and Appendix 2, after transforming the response variable and all the predictor variables except ‘house age’ and ‘number of convenience stores’ using log, it is shown that the violations of linearity, equal variance, and normality assumptions have all been fixed. For example, based on the pairs plots in Appendix 1, we can see positive relationships between log of longitude and log of house price per unit, log of latitude and log of house price per unit, and number of convenience stores and log of house price per unit. There seems to be a negative relationship between log of distance to the nearest MRT station and log of house price per unit. In addition, based on the added variable plots in Appendix 1, we can see an improvement in linearity for each predictor against the log of house price per unit while controlling all other predictors. We can see a strong negative linear relationship between log of distance to the nearest MRT station and the log of house price per unit. There is also a strong positive relationship between log of latitude and log of house price per unit. The other predictors have less obvious slopes, but there is only a linear pattern between the predictors and response. Therefore, from the added variable plots and pairs plots, the violation of the linearity assumption has been fixed. Based on the residual vs fitted values plots of the transformed model between the response variable toward each individual predictor variable in Appendix 1, we can see that variance among all the relationship between response variable and predictor variables have all become more constant except for longitude, which we have to remove from the model because of the violation of constant variance and normality. Based on the Q-Q plots of the transformed model of the individual predictor variables to the response variable in Appendix 1, normality has greatly improved among all predictor variables. Although there are some outliers in the Q-Q plots, we can test for the outliers by using the influenceIndexPlot function to get Cook’s Distance plot. Based on the Cook’s Distance plot, we can remove points 114, 149, and 271 since these points have high Cook’s Distance values, which implies these points are outliers. We also remove high-leverage points 36 and 229. After removing these points, we refit the new transformed model with the data set without the influential points. Therefore, linearity, equal variance, and normality assumptions are no longer violated.

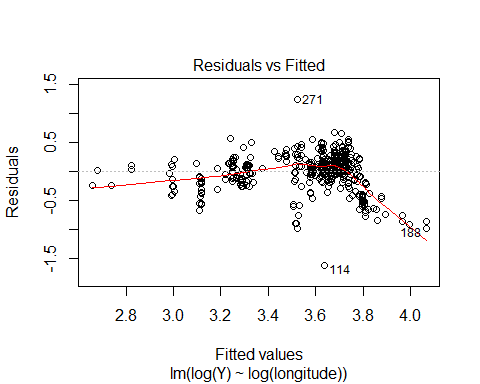
**REGRESSION ANALYSIS, RESULTS, AND INTERPRETATION**

*Question 1*:

Refer to Appendix 2 to see all previous diagnostic checks prior to arriving at the final transformation of the data. We decide linearity is not violated by our final transformations for reasons mentioned based on the plots of the marginal relationships, as well as the added-variable plots, which shows relationships between a predictor and the response while controlling all other predictors:

From our diagnostics, we know that after the transformation of the response and variables, we exclude longitude from the model based on the following residual vs fit and Q-Q plot:





Here, we can see that constant variance and normality are both blatantly violated by longitude.

With the new model that does not violate the four assumptions of linear regression, we will use the global F-test to determine whether there is at least one predictor that is useful in predicting the house price per unit. The null hypothesis is that all the predictors equal 0 while the alternative hypothesis is at least one predictor is not equal to 0. Based on the summary table of the new model without the outliers and transformed variables, the p-value is < 2.2e-16, which is less than any reasonable significant value. Therefore, we reject the null hypothesis. We then conclude that there is at least one predictor that is useful. Next, we use backward selection to determine if there are smaller models that are better than the full model.  After using backward selection, we determine that the full model is the best since the lowest and only AIC is -1389.22. Next, we use regsubsets to verify if the full model is the best model or not. Based on the regsubsets data from the regsubsets function in Appendix 1, we conclude that Model 5 of summary.reg$which, or the full model, is the best since it has the smallest mean square error of 17.93521, the largest adjusted R-squared value of 0.7146166, and a very small BIC value of -488.0111. Since Model 5’s BIC value is very close to the smallest BIC value in Model 4, which is -489.7180, we therefore still choose Model 5, or the full model, as the best model. Next, we perform a partial F-test between the reduced model, Model 4, with all predictors except ‘number of convenience stores’ and the full model to determine if ‘number of convenience stores’ predictor is useful or not. The null hypothesis is that ‘number of convenience stores’ equals 0 while the alternative hypothesis is that ‘number of convenience stores’ is not equal to 0. Since the F-statistic is large (9.8948) and the p-value of the ANOVA table between the reduced model without ‘number of convenience stores’ predictor and the full model is 0.00178 which is less than significance level 0.05, we can therefore reject the null hypothesis. Therefore, we can conclude that ‘number of convenience stores’ predictor is useful, and therefore we should pick the full model. Therefore, the answer to the first question of interest is that based on the final model, the first five predictors, which are ‘transaction date’, ‘house age’, ‘distance to the nearest MRT station’, ‘number of convenience stores’, and latitude are the useful predictors for predicting the price of housing per unit. Based on summary(newmod.trans2), our final model is as follows:

Yi(hat) = -3333 + 322.6x1 - 0.006118x2 - 0.1884x3 + 0.01369x4 + 274.5x5

where Yi(hat) is the estimate of log of the price of housing per unit area, x1 is log of transaction date, x2 is house age, x3 is log of distance to the nearest MRT station, x4 is number of convenience stores, and x5 is log of latitude.

*Question 2*:

Now that we have the final model, we can now answer the second question of interest. To find the average predicted value, we need to first find the mean of each predictor variable and fit the response variable and the predictor variables of the final model. Next, we find the mean of each predictor variable and put the means into a data frame. Next, we use the predict function and plug in the fit of the final model and the mean data frame and set interval = ‘prediction’ and level to 0.95. After computing the predict function, our predicted average house price per unit is e^(3.455273)\*10000, or 316669.3 New Taiwan Dollar per Ping. In addition, for our prediction interval, we are 95 percent confident that a house in New Taipei City, Taiwan with average transaction date, house age, distance to nearest MRT station, number of convenience stores, and latitude falls between e^(3.097448)\*10000, or 221,413.7 New Taiwan Dollar per ping and e^(3.813098)\*10000, or 452,905.3 New Taiwan Dollar per Ping. Note: the response of the data set is originally in 10,0000 New Taiwan Dollar per Ping.

*Question 3*:

To calculate the estimate of the true “average” house price per unit in New Taipei City, Taiwan, we find the 95 percent confidence interval for the mean of each predictor through the predict function by using the data frame of all the means of the predictor variables and by using the fit of the final model. After computing the predict function, we can say that we are 95% confident that the estimated median house price per unit of a house with average transaction date, house age, distance to nearest MRT station, number of convenience stores, and latitude is between e^(3.432954)\*10000, or 309,679.9 New Taiwan Dollar per Ping and e^(3.477592)\*10000, or 323,816.5 New Taiwan Dollar per Ping.

**CONCLUSION**

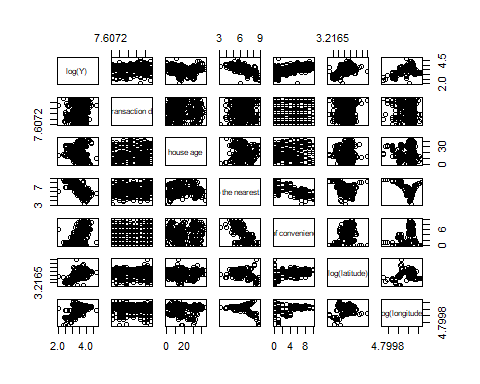
From the analysis of the data set of the house price per unit area and the house price’s factors, we can conclude that the house’s price per unit area is most affected by five factors, which are the transaction date (in years),  the house age (in years), the distance to the nearest MRT station (in meters), the number of convenience stores, and latitude (in degrees). Though, we can conclude that longitude does not have a significant effect on the price of housing in Taipei City, Taiwan. In addition, we saw that the closer a house is to a MRT station or the more convenience stores there are around the house, the higher the price of the house becomes. Furthermore, the older the house becomes, the lower of the price of the house is. Based on the amount of evidence supporting these results, we can trust these results such as the five predictors having the most impact on the housing price. However, these results can only apply to the housing in Taipei City, Taiwan since housing in other cities or areas may have different factors and variables affecting the price of housing in other areas. Finally, we are 95 percent sure that the average price of the housing in Taipei City is between 309,679.9 New Taiwan Dollar per Ping and 323,816.5 New Taiwan Dollar per Ping. We can also predict that a house with average transaction date, house age, distance to nearest MRT station, number of convenience stores, and latitude may be 316669.3 New Taiwan Dollar per Ping as time goes by.

Appendix 1

From Appendix 2, we see violations of assumptions that led us to transform our response and some of our predictors. We cannot take the log of house age or log of number of convenience stores as some values are 0, which falls outside of the domain of the logarithmic scale.

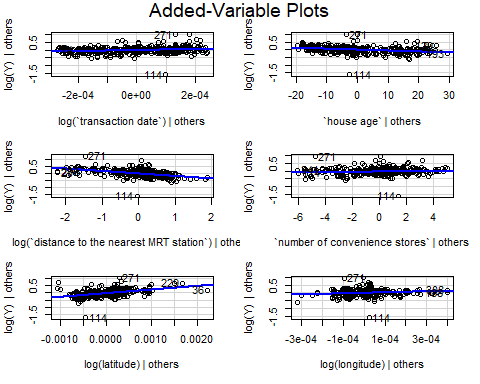
mod.trans2 <- lm(log(`Y`) ~ log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) + `number of convenience stores` + log(latitude) + log(longitude))

# We use scatterplots to look for marginal relationships  
pairs(~log(`Y`) + log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) + `number of convenience stores` + log(latitude) + log(longitude))



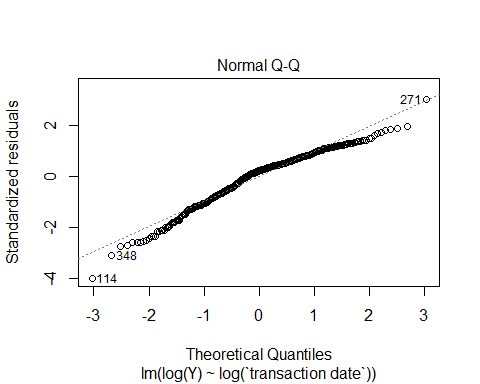
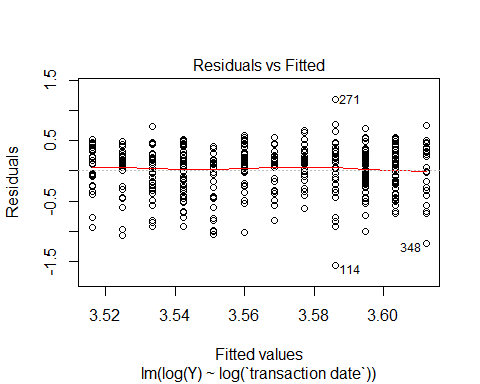
We can see positive relationships between log of longitude and log of house price per unit, log of latitude and log of house price per unit, and number of convenience stores and log of house price per unit. There seems to be a negative relationship between log of distance to the nearest MRT station and log of house price per unit.

# We use added-variable plots to check for linearity of each predictor while controlling all other predictors  
avPlots(mod.trans2)



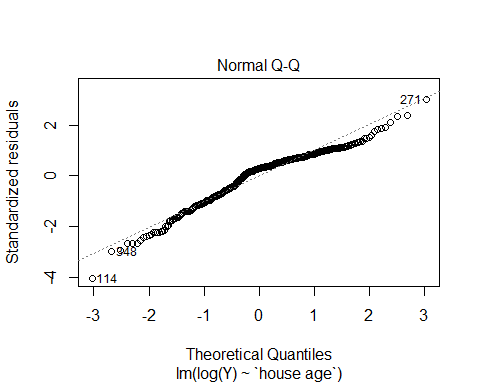
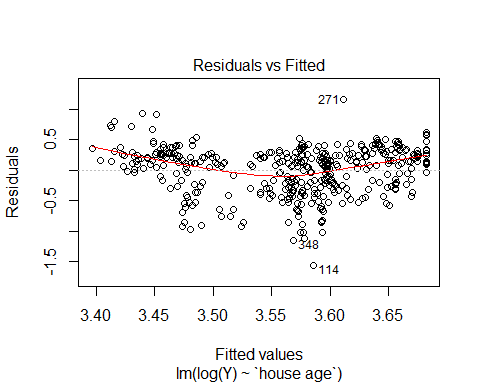
From the Added-Variable Plots, we can see an improvement in linearity for each predictor against the log of house price per unit while controlling all other predictors. We can see a strong negative linear relationship between log of distance to the nearest MRT station and the log of house price per unit. There is also a strong positive relationship between log of latitude and log of house price per unit. The other predictors have less obvious slopes, but there is only a linear pattern between the predictors and response.

# We use residual vs fit plot to determine constant/equal variance for each predictor  
# We use Q-Q plot to determine normality for each predictor  
  
# Predictor: log of transaction date  
fit1 <- lm(log(`Y`)~ log(`transaction date`))  
plot(fit1, which = c(1,2))



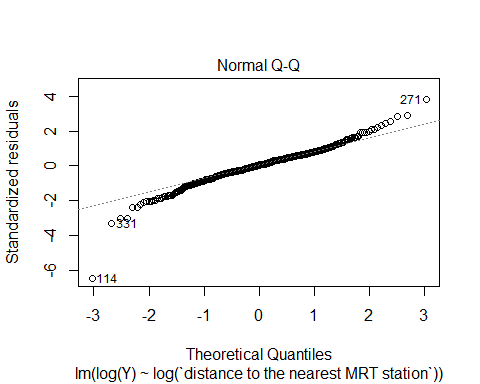
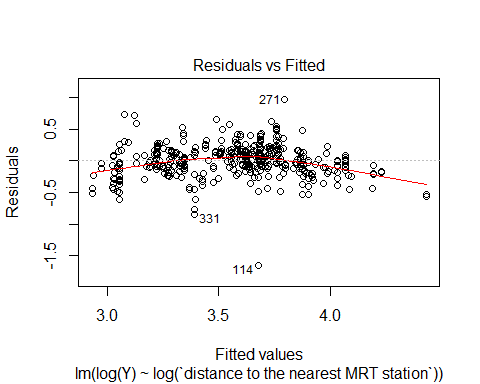
From the Residuals vs Fitted Plot for log of transaction date vs log of house price per unit (other predictors controlled), we can say that the variance remains mainly constant except for a couple outlying points now and the assumption is not violated. Based on the Q-Q Plot, we see that normality has improved greatly. Any outliers can be tested for and removed if necessary.

# Predictor: house age  
fit2 <- lm(log(`Y`)~`house age`)  
plot(fit2, which = c(1,2))



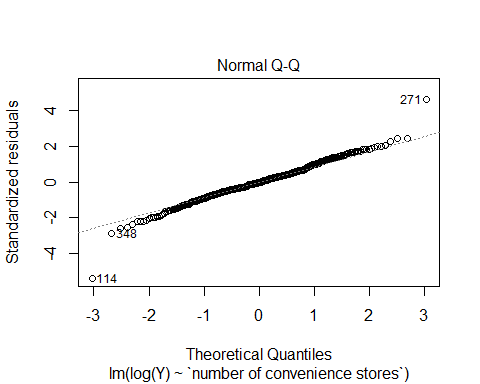
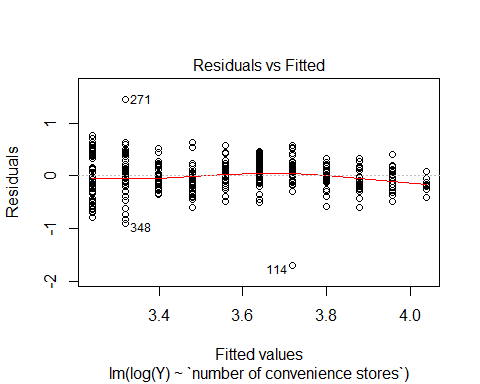
From the Residuals vs Fitted Plot for house age vs log of house price per unit (other predictors controlled), we can see that the residuals form a slight quadratic pattern, but the unit size of the residuals remain small. Variance has greatly improved from initial model. Based on the Q-Q Plot, we see that normality has also improved. Any outliers can be tested for and removed if necessary.

# Predictor: log of distance to the nearest MRT station  
fit3 <- lm(log(`Y`)~log(`distance to the nearest MRT station`))  
plot(fit3, which = c(1,2))



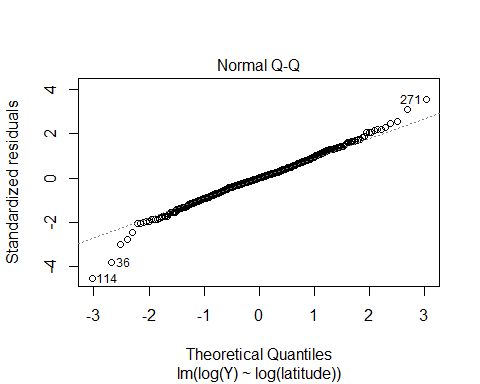
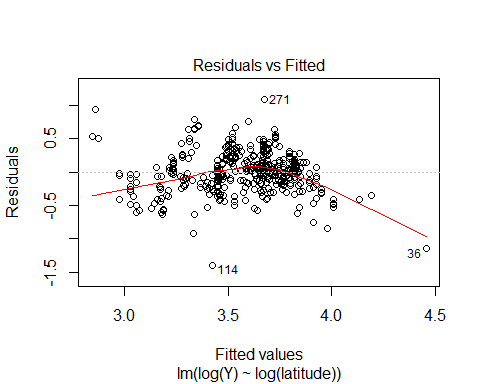
From the Residuals vs Fitted Plot for log of distance to the nearest MRT station vs log of house price per unit (other predictors controlled), we can see that the variance has improved from the transformation and the unit size of the residuals is small. Based on the Q-Q Plot, we see that normality has also improved. Any outliers can be tested for and removed if necessary.

# Predictor: number of convenience store  
fit4 <- lm(log(`Y`)~`number of convenience stores`)  
plot(fit4, which = c(1,2))



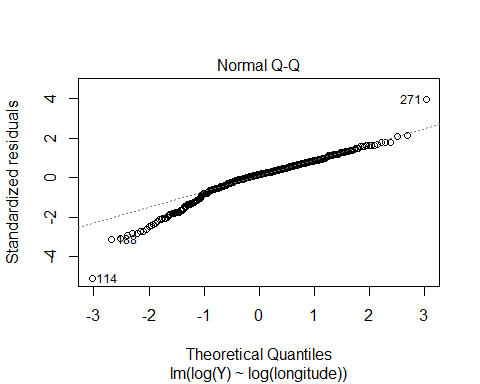
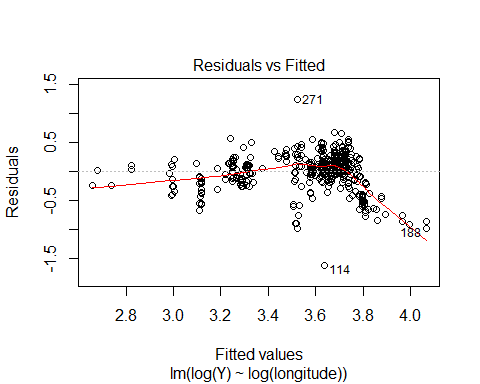
From the Residuals vs Fitted Plot for number of convenience stores vs log of house price per unit (other predictors controlled), we can say that, With the exception of a few outlying points, the variance remains mainly constant and the assumption is not violated. Based on the Q-Q Plot, we see that normality is not violated assuming any outliers can be tested for and removed if necessary.

# Predictor: log of latitude  
fit5 <- lm(log(`Y`)~log(latitude))  
plot(fit5, which = c(1,2))



From the Residuals vs Fitted Plot for log of latitude vs log of house price per unit (other predictors controlled), we can see that variance is improved through the transformation. Based on the Q-Q Plot, we see that normality has greatly improved. Any outliers can be tested for and removed if necessary.

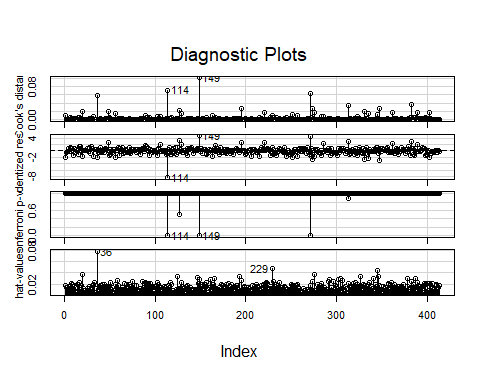
# Predictor: log of longitude  
fit6 <- lm(log(`Y`)~log(longitude))  
plot(fit6, which = c(1,2))



From the Residuals vs Fitted Plot for log of longitude vs log house price per unit (other predictors controlled), we can see that variance is clearly not constant based on the clustering and dipping pattern of the residuals. Therefore, the assumption is still violated after the transformation. Based on the Q-Q Plot, we see that normality is violated due to the outliers and tail ends of the plot. Although both variance and normality have been improved by this transformation, those two assumptions are still clearly violated. Therefore, we will exclude longitude from our model.

Normality is violated mainly by outliers. So we will test for outliers, high leverage points, and influential points to know if we can remove those points from the data set. We could then proceed in finding our model.

mod.trans2 <- lm(log(`Y`) ~ log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) + `number of convenience stores` + log(latitude))  
influenceIndexPlot(mod.trans2)



We see that points 114, 149, and 271 have high Cook’s Distance values. We can also note that points 36 and 229 are high-leverage points.

We now remove our outliers.

# Delete influential points  
data.noinfluence <- data1[c(1:35,37:113,115:148,150:228,230:270,272:414),]  
  
# New fit without influential points  
newmod.trans2 <- lm(log(`Y`)~log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) + `number of convenience stores` + log(latitude), data = data.noinfluence)  
  
# Compare the initial fit and the new fit  
summary(mod.trans2)

##   
## Call:  
## lm(formula = log(Y) ~ log(`transaction date`) + `house age` +   
## log(`distance to the nearest MRT station`) + `number of convenience stores` +   
## log(latitude))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.61013 -0.10834 0.01186 0.10986 0.95996   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -3.413e+03 5.673e+02 -6.015  
## log(`transaction date`) 3.370e+02 7.461e+01 4.517  
## `house age` -6.068e-03 9.180e-04 -6.610  
## log(`distance to the nearest MRT station`) -1.944e-01 1.333e-02 -14.581  
## `number of convenience stores` 1.027e-02 4.966e-03 2.068  
## log(latitude) 2.652e+02 2.395e+01 11.076  
## Pr(>|t|)   
## (Intercept) 4.00e-09 \*\*\*  
## log(`transaction date`) 8.21e-06 \*\*\*  
## `house age` 1.21e-10 \*\*\*  
## log(`distance to the nearest MRT station`) < 2e-16 \*\*\*  
## `number of convenience stores` 0.0392 \*   
## log(latitude) < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2097 on 408 degrees of freedom  
## Multiple R-squared: 0.7181, Adjusted R-squared: 0.7146   
## F-statistic: 207.8 on 5 and 408 DF, p-value: < 2.2e-16

summary(newmod.trans2)

##   
## Call:  
## lm(formula = log(Y) ~ log(`transaction date`) + `house age` +   
## log(`distance to the nearest MRT station`) + `number of convenience stores` +   
## log(latitude), data = data.noinfluence)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.59990 -0.10939 0.01375 0.11298 0.66818   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -3.333e+03 4.952e+02 -6.731  
## log(`transaction date`) 3.226e+02 6.497e+01 4.966  
## `house age` -6.118e-03 7.970e-04 -7.677  
## log(`distance to the nearest MRT station`) -1.884e-01 1.172e-02 -16.083  
## `number of convenience stores` 1.369e-02 4.351e-03 3.146  
## log(latitude) 2.745e+02 2.211e+01 12.417  
## Pr(>|t|)   
## (Intercept) 5.83e-11 \*\*\*  
## log(`transaction date`) 1.01e-06 \*\*\*  
## `house age` 1.25e-13 \*\*\*  
## log(`distance to the nearest MRT station`) < 2e-16 \*\*\*  
## `number of convenience stores` 0.00178 \*\*   
## log(latitude) < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1817 on 403 degrees of freedom  
## Multiple R-squared: 0.7769, Adjusted R-squared: 0.7741   
## F-statistic: 280.6 on 5 and 403 DF, p-value: < 2.2e-16

From the summary of the original fit, we note a residual standard error of 0.2097 and an adjusted R-squared value of 0.7146. In the summary of the new fit with the influential points removed, we note that the residual standard error has decreased to 0.1817 and the adjusted R-squared value has increased to 0.7741. Therefore, we know that the fit without the influential points is a better for the model.

We will perform a Global F-Test to ensure we have at least one useful predictor.

summary(newmod.trans2)

##   
## Call:  
## lm(formula = log(Y) ~ log(`transaction date`) + `house age` +   
## log(`distance to the nearest MRT station`) + `number of convenience stores` +   
## log(latitude), data = data.noinfluence)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.59990 -0.10939 0.01375 0.11298 0.66818   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -3.333e+03 4.952e+02 -6.731  
## log(`transaction date`) 3.226e+02 6.497e+01 4.966  
## `house age` -6.118e-03 7.970e-04 -7.677  
## log(`distance to the nearest MRT station`) -1.884e-01 1.172e-02 -16.083  
## `number of convenience stores` 1.369e-02 4.351e-03 3.146  
## log(latitude) 2.745e+02 2.211e+01 12.417  
## Pr(>|t|)   
## (Intercept) 5.83e-11 \*\*\*  
## log(`transaction date`) 1.01e-06 \*\*\*  
## `house age` 1.25e-13 \*\*\*  
## log(`distance to the nearest MRT station`) < 2e-16 \*\*\*  
## `number of convenience stores` 0.00178 \*\*   
## log(latitude) < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1817 on 403 degrees of freedom  
## Multiple R-squared: 0.7769, Adjusted R-squared: 0.7741   
## F-statistic: 280.6 on 5 and 403 DF, p-value: < 2.2e-16

: = = = = = 0, : at least one 0 for j = 1, 2, 3, 4, 5. After running a Global F-Test, we see that our p-value is 2.2e-16 which is less than a significance level of 0.05. Therefore, we reject our null hypothesis and can assume that at least one of our predictors is useful in predicting the log of house unit per price.

We will now use Backward Selection as it is the ideal way to find the best model.

# Create the most simple model

basemod <- lm(log(`Y`)~1)  
  
model.backward <- step(newmod.trans2, scope = list(lower = basemod, upper = newmod.trans2), direction = 'backward')

## Start: AIC=-1389.22  
## log(Y) ~ log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) +   
## `number of convenience stores` + log(latitude)  
##   
## Df Sum of Sq RSS AIC  
## <none> 13.300 -1389.2  
## - `number of convenience stores` 1 0.3265 13.626 -1381.3  
## - log(`transaction date`) 1 0.8138 14.114 -1366.9  
## - `house age` 1 1.9448 15.245 -1335.4  
## - log(latitude) 1 5.0880 18.388 -1258.7  
## - log(`distance to the nearest MRT station`) 1 8.5368 21.837 -1188.4

Backward Selection suggests that the best model is the full model. The lowest and only AIC is -1389.22. Backward Selection stops when AIC can no longer decrease.

We will use best subset regression to verify whether or not the full model is the best option for the data.

library(leaps)  
mod.reg <- regsubsets(cbind(log(`transaction date`), `house age`, log(`distance to the nearest MRT station`), `number of convenience stores`, log(latitude)), log(`Y`), data = data.noinfluence)  
  
summary.reg <- summary(mod.reg)  
  
summary.reg$which

## (Intercept) house age number of convenience stores   
## 1 TRUE FALSE FALSE TRUE FALSE FALSE  
## 2 TRUE FALSE FALSE TRUE FALSE TRUE  
## 3 TRUE FALSE TRUE TRUE FALSE TRUE  
## 4 TRUE TRUE TRUE TRUE FALSE TRUE  
## 5 TRUE TRUE TRUE TRUE TRUE TRUE

summary.reg$rsq

## [1] 0.5789156 0.6713779 0.6997028 0.7151151 0.7180716

summary.reg$rss

## [1] 26.78778 20.90568 19.10376 18.12329 17.93521

summary.reg$adjr2

## [1] 0.5778936 0.6697788 0.6975055 0.7123289 0.7146166

summary.reg$cp

## [1] 199.383153 67.573926 28.582894 8.278594 6.000000

summary.reg$bic

## [1] -346.0260 -442.6410 -473.9313 -489.7180 -488.0111

Biggest increase in R-squared: Model 2

Smallest mean square: Model 5

Largest adjusted R-squared: Model 5

Smallest Cp (besides full model): Model 4

Smallest BIC: Model 4, but this BIC value is extremely close to the BIC value for Model 5

Generally, we would want to opt for the most simple model. In this case that would be model 4, which has the smallest Cp value and smallest BIC value. We perform a Partial F-Test to test if the simpler model is better than the full model (model 5), which has the smallest mean square value, largest adjusted R-squared value, and smallest AIC value.

red.lm <- lm(log(`Y`)~log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) + log(latitude), data = data.noinfluence)  
anova(red.lm, newmod.trans2)

## Analysis of Variance Table  
##   
## Model 1: log(Y) ~ log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) +   
## log(latitude)  
## Model 2: log(Y) ~ log(`transaction date`) + `house age` + log(`distance to the nearest MRT station`) +   
## `number of convenience stores` + log(latitude)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 404 13.626   
## 2 403 13.300 1 0.32655 9.8948 0.00178 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

: = 0, : 0. From the ANOVA table, we see that the F-statistic is large (9.8948) and the P-value is 0.00178 is small. Therefore, we can reject our null hypothesis and assume that our fourth predictor (number of convenience stores) is a useful predictor. We can now conclude that the full model is our optimal model.

Now that we have our model, we can use it to calculate a Prediction Interval for the mean of each predictor.

data(data.noinfluence)

attach(data.noinfluence)

x1 <- log(`transaction date`)  
x2 <- `house age`  
x3 <- log(`distance to the nearest MRT station`)  
x4 <- `number of convenience stores`  
x5 <- log(latitude)  
Y1 <- log(`Y`)  
mean1 <- mean(`transaction date`)  
mean2 <- mean(`house age`)  
mean3 <- mean(`distance to the nearest MRT station`)  
mean4 <- mean(`number of convenience stores`)  
mean5 <- mean(latitude)  
  
predict(lm(Y1~x1 + x2 + x3 + x4 + x5), data.frame(x1 = log(mean1), x2 = mean2, x3 = log(mean3), x4 = mean4, x5 = log(mean5)), interval = 'prediction', level = 0.95)

## fit lwr upr  
## 1 3.455273 3.097448 3.813098

# Interpretation Calculations  
  
# Prediction  
exp(3.455273)

## [1] 31.66693

# Lower bound  
exp(3.097448)

## [1] 22.14137

# Upper bound  
exp(3.813098)

## [1] 45.29053

We predict that a house in New Taipei City, Taiwan with average transaction date, house age, distance to nearest MRT station, number of convenience stores, and latitude is 316,669.3 New Taiwan Dollar per Ping. We are 95% confident that a house in New Taipei City, Taiwan with average transaction date, house age, distance to nearest MRT station, number of convenience stores, and latitude falls between 221,413.7 New Taiwan Dollar per ping and 452,905.3 New Taiwan Dollar per Ping. Note: the response is originally in 10,0000 New Taiwan Dollar per Ping.

We will also calculate a Confidence Interval for mean of each predictor to find an interval for true “average”.

data(data.noinfluence)

attach(data.noinfluence)

predict(lm(Y1~x1 + x2 + x3 + x4 + x5), data.frame(x1 = log(mean1), x2 = mean2, x3 = log(mean3), x4 = mean4, x5 = log(mean5)), interval = 'confidence', level = 0.95)

## fit lwr upr  
## 1 3.455273 3.432954 3.477592

# Interpretation Calculations  
  
# Lower Bound  
exp(3.432954)

## [1] 30.96799

# Upper Bound  
exp(3.477592)

## [1] 32.38165

We are 95% confident that the estimated median house price per unit of a house with average transaction date, house age, distance to nearest MRT station, number of convenience stores, and latitude is between 309,679.9 New Taiwan Dollar per Ping and 323,816.5 New Taiwan Dollar per Ping.

Appendix 2

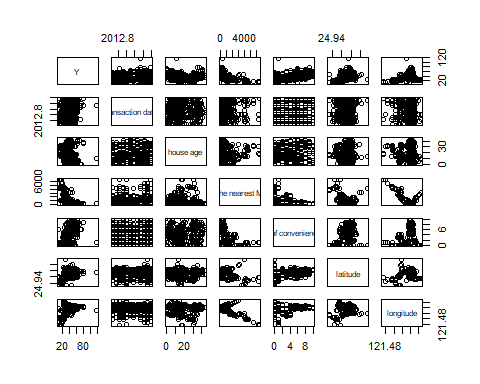
Import Data set

library(readxl)  
Real\_estate\_valuation\_data\_set <- read\_excel("Real estate valuation data set.xlsx")  
data1 <- Real\_estate\_valuation\_data\_set  
data(data1)

attach(Real\_estate\_valuation\_data\_set)

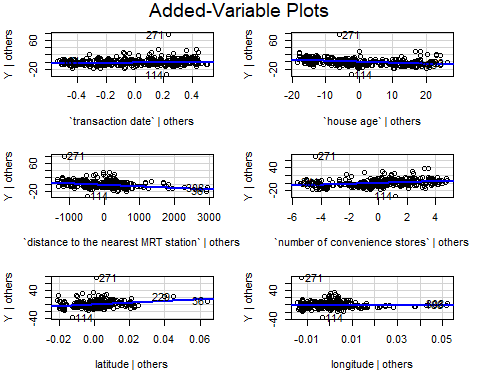
Make Plots with All 7 Variables to check LINE Assumptions

# We begin with a full model and find violations in order to apply appropriate transformation  
mod.full <- lm(`Y`~`transaction date` + `house age` + `distance to the nearest MRT station` + `number of convenience stores` + latitude + longitude)  
  
# We use scatterplots to initially look for marginal relationships  
pairs(~`Y` + `transaction date` + `house age` + `distance to the nearest MRT station` + `number of convenience stores` + latitude + longitude)

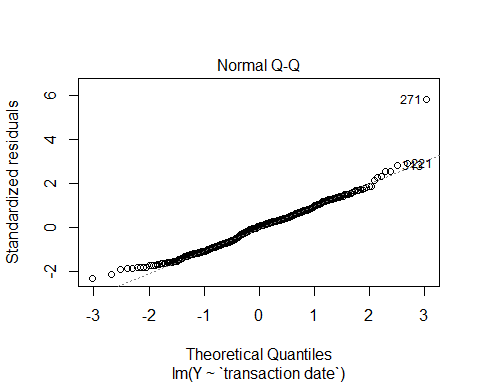
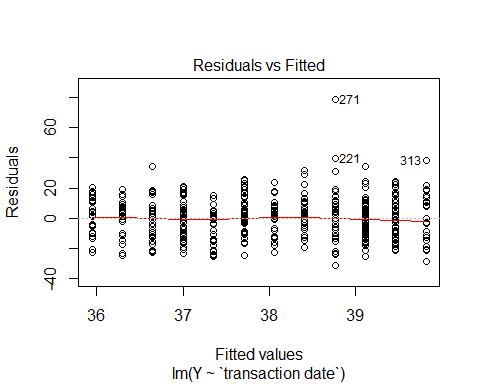
*COMMENT ON MARGINAL RELATIONSHIPS* From the scatterplots, it is hard to see any obvious marginal relationships between the variables of the data set.

library(car)

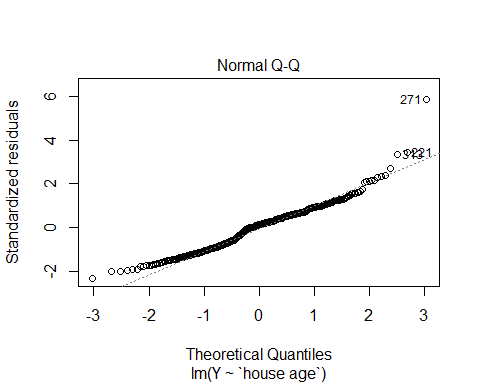
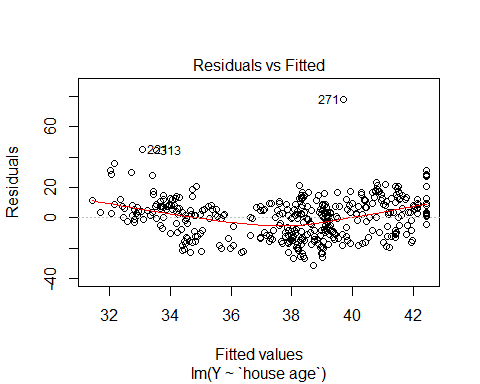
# We use added-variable plots to check for linearity of each predictor while controlling all other predictors  
avPlots(mod.full)

*COMMENT ON LINEARITY* From the Added-Variable Plots, we can check for linearity for each predictor while controlling all other predictors. We see that there is not much of a linear relationship between the response and house age and distance to the nearest MRT station, as the slope on those plots are near zero. There does seem to be a positive, slightly linear relationship relationship between transaction date and the response, number of convenience stores and the response, and latitude and the response. There is a negative, slightly linear relationship between longitude and the response. Overall, we can state that there is a weak, if any, linear relationship between each predictor and the response (house price per unit). Therefore, the linearity assumption as violated and we will eventually have to perform a transformation.

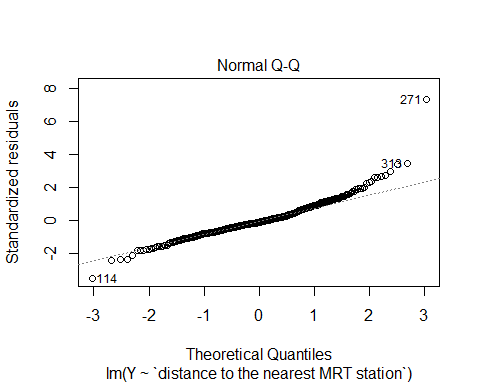
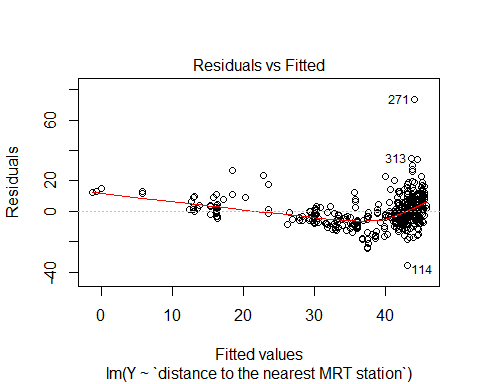
# We use residual vs fit plot to determine constant/equal variance for each predictor  
# We use Q-Q plot to determine normality for each predictor  
  
# Predictor: transaction date  
fit1 <- lm(`Y`~`transaction date`)  
plot(fit1, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for transaction date vs house price per unit (other predictors controlled), we can say that, With the exception of a few outlying points, the variance remains mainly constant and the assumption is not violated. Based on the Q-Q Plot, we see that normality is violated due to the outliers and the tail ends of the partially right-skewed plot.

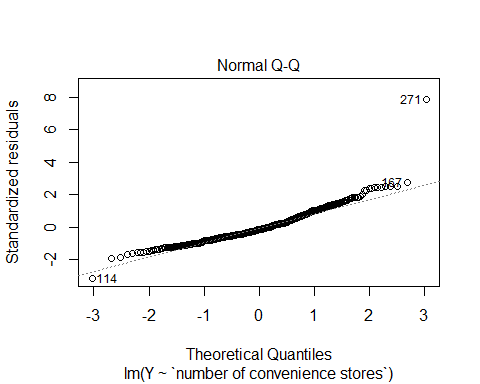
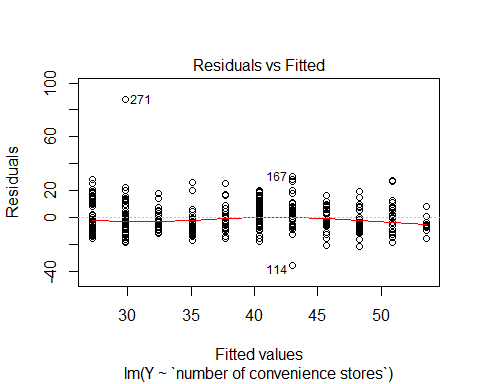
# Predictor: house age  
fit2 <- lm(`Y`~`house age`)  
plot(fit2, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for house age vs house price per unit (other predictors controlled), we can see that the residuals form a quadratic pattern, implying non-constant variance. Based on the Q-Q Plot, we see that normality is violated due to the outliers and the tail ends of the partially right-skewed plot.

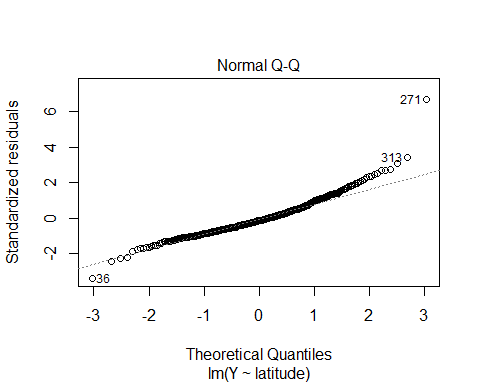
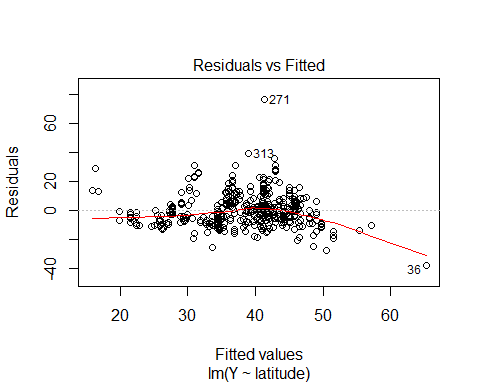
# Predictor: distance to the nearest MRT station  
fit3 <- lm(`Y`~`distance to the nearest MRT station`)  
plot(fit3, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for distance to the nearest MRT station vs house price per unit (other predictors controlled), we can see that the constant variance assumption is heavily violated based on the clustering of residuals towards the right side of the plot. Based on the Q-Q Plot, we see that normality is violated due to the outliers and the tail ends of the heavy-tailed plot.

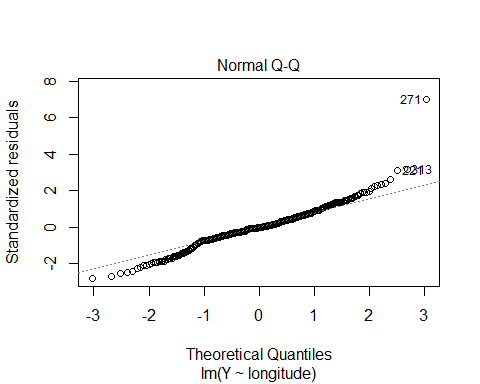
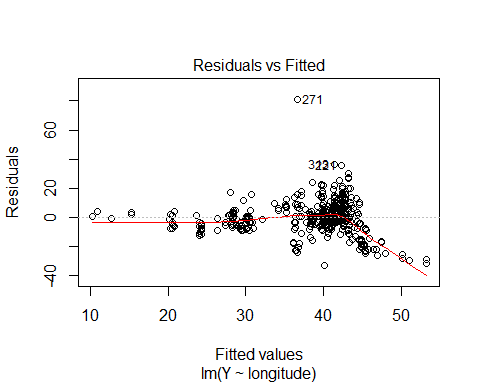
# Predictor: number of covenience stores  
fit4 <- lm(`Y`~`number of convenience stores`)  
plot(fit4, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for number of convenience stores vs house price per unit (other predictors controlled), we can say that, With the exception of a few outlying points, the variance remains mainly constant and the assumption is not violated. Based on the Q-Q Plot, we see that normality is violated due to the outliers and the tail ends of the right-skewed plot.

# Predictor: latitude  
fit5 <- lm(`Y`~latitude)  
plot(fit5, which = c(1,2))

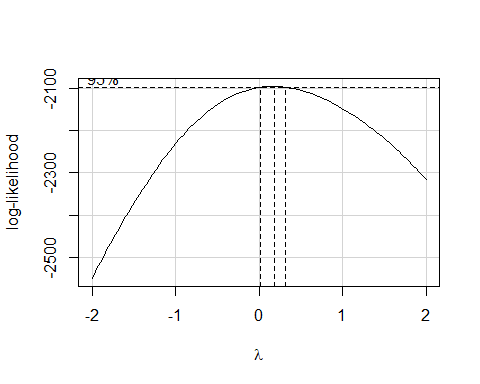
 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for latitude vs house price per unit (other predictors controlled), we can see that variance is not constant based on the quadratic pattern and the assumption is violated. Based on the Q-Q Plot, we see that normality is violated due to the outliers and the tail ends of the slightly heavy-tailed plot.

# Predictor: longitude  
fit6 <- lm(`Y`~longitude)  
plot(fit6, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for longitude vs house price per unit (other predictors controlled), we can see that variance is clearly not constant based on the clustering and dipping pattern of the residuals. Therefore, the assumption is violated. say that, With the exception of a few outlying points, the variance remains mainly constant and the assumption is not violated. Based on the Q-Q Plot, we see that normality is violated due to the outliers and the tail ends of the heavy-tailed plot.

Due to violations of linearity, constant variance, and normality, we must transform the data

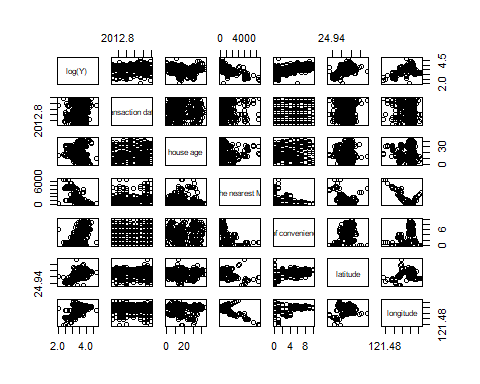
# First, we try Box-Cox Method  
boxCox(mod.full)

 We can see from the Box-Cox transformation that the ideal lambda is close to 0, therefore the appropriate scale to transform the data would be the logarithmic scale.

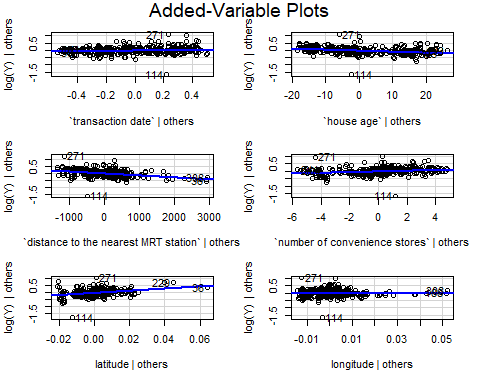
# Transform the response first  
mod.trans <- lm(log(`Y`)~`transaction date` + `house age` + `distance to the nearest MRT station` + `number of convenience stores` + latitude + longitude)

We check LINE assumptions again with the transformed response in order to determine whether or not the predictors also need to be transformed

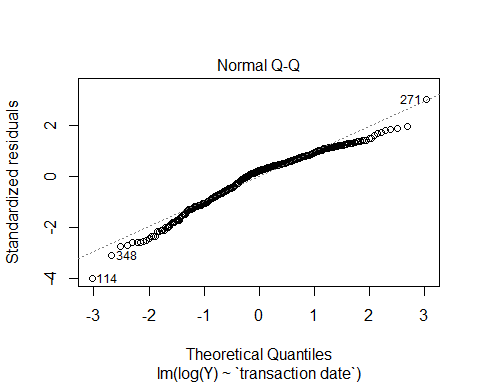
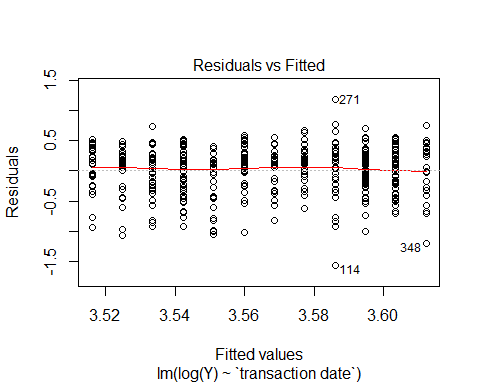
# We use scatterplots to look for marginal relationships  
pairs(~log(`Y`) + `transaction date` + `house age` + `distance to the nearest MRT station` + `number of convenience stores` + latitude + longitude)

 *COMMENT ON MARGINAL RELATIONSHIPS* From these scatterplots with the transformed response, we can see a negative relationship between distance to the nearest MRT station and the log of house price per unit. There also seems to be a positive relationship between latitude and log of the response, as well as longitude and log of the response.

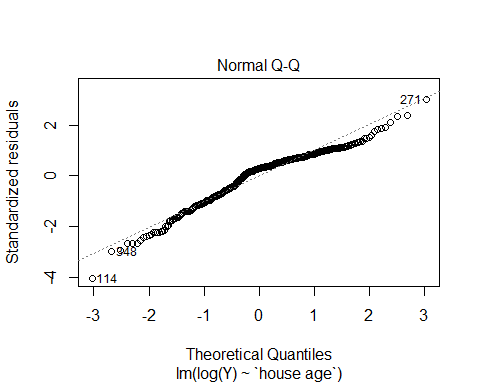
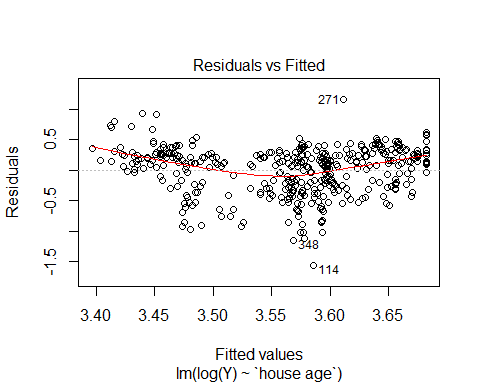
# We use added-variable plots to check for linearity of each predictor while controlling all other predictors  
avPlots(mod.trans)

 *COMMENT ON LINEARITY* From the Added-Variable Plots, we can check for linearity for each predictor against the log of house price per unit while controlling all other predictors. We see that now after the transformation there is not much of a linear relationship between the log of the response and the predictors: number of convenience stores, longitude, transaction date, and house age, as the slope on those plots are near zero. There does seem to be a positive, moderately linear relationship relationship between latitude and the log of the response and a negative, moderately linear relationship between distance to the nearest MRT station and the log of the response. Overall, we can state that after the initial transformation, there are less predictors with linear relationships with the log of the response, but the predictors that do show linearity have stronger linearity, making those predictors more useful.

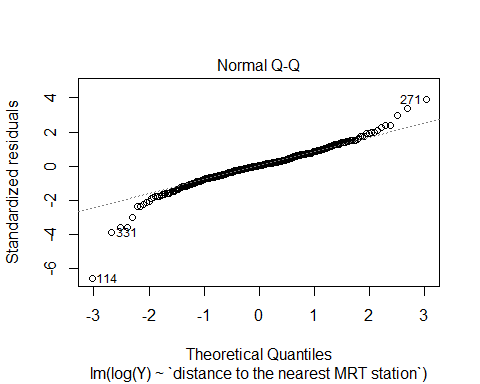
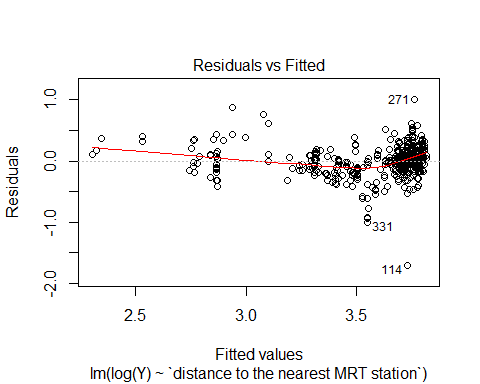
# We use residual vs fit plot to determine constant/equal variance for each predictor  
# We use Q-Q plot to determine normality for each predictor  
  
# Predictor: transaction date  
fit1 <- lm(log(`Y`)~`transaction date`)  
plot(fit1, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for transaction date vs log of house price per unit (other predictors controlled), we can say that the variance remains mainly constant except for a couple outlying points now and the assumption is not violated. Based on the Q-Q Plot, we see that normality is still violated due to the outliers and the tail ends of the slightly left-skewed plot.

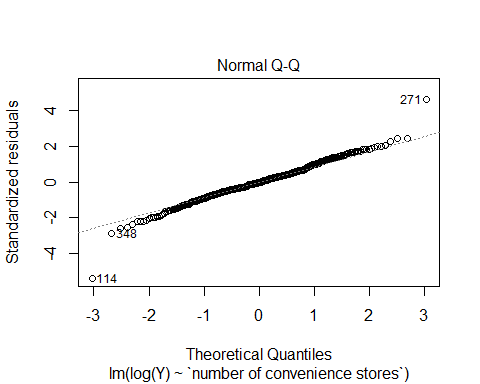
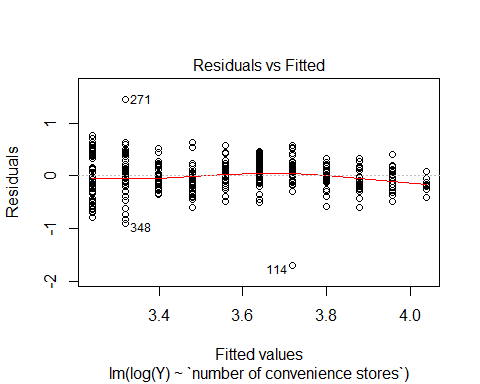
# Predictor: house age  
fit2 <- lm(log(`Y`)~`house age`)  
plot(fit2, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for house age vs log of house price per unit (other predictors controlled), we can see that the residuals still form a quadratic pattern, implying non-constant variance. Based on the Q-Q Plot, we see that normality is violated due to an outlier and the tail ends of the partially left-skewed plot.

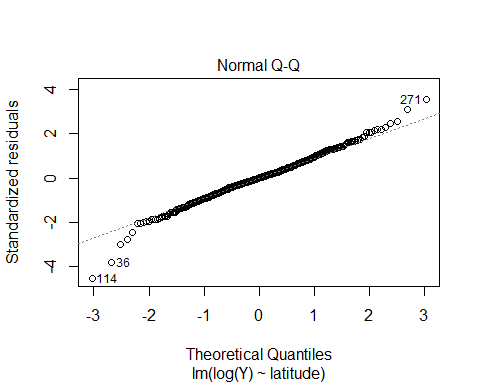
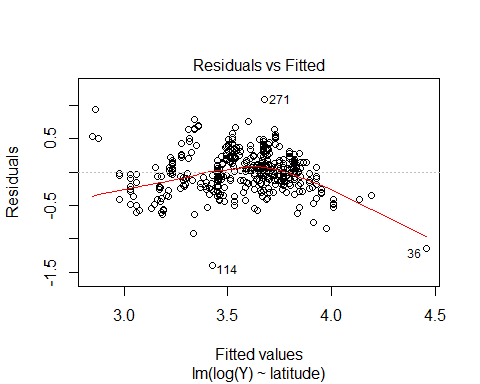
# Predictor: distance to the nearest MRT station  
fit3 <- lm(log(`Y`)~`distance to the nearest MRT station`)  
plot(fit3, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for distance to the nearest MRT station vs log of house price per unit (other predictors controlled), we can see that the constant variance assumption is still violated based on the clustering of residuals towards the right side of the plot. Based on the Q-Q Plot, we see that normality is still violated due to the outliers and the tail ends of the heavy-tailed plot.

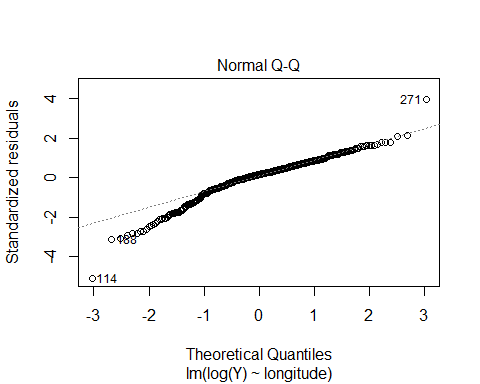
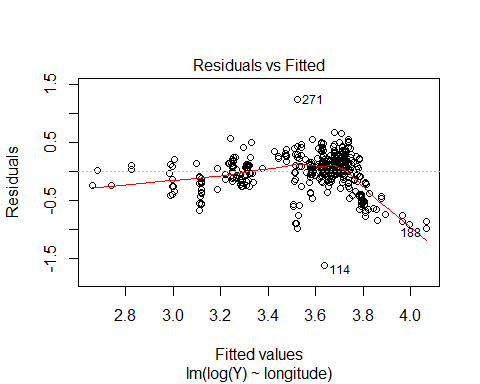
# Predictor: number of convenience stores  
fit4 <- lm(log(`Y`)~`number of convenience stores`)  
plot(fit4, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for number of convenience stores vs log of house price per unit (other predictors controlled), we can say that, With the exception of a couple outlying points, the variance still remains mainly constant and the assumption is not violated. Based on the Q-Q Plot, we see that normality is violated due to outliers and the tail ends of the right-skewed plot.

# Predictor: latitude  
fit5 <- lm(log(`Y`)~latitude)  
plot(fit5, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for latitude vs log of house price per unit (other predictors controlled), we can see that variance is still not constant based on the quadratic pattern and the assumption is violated. Based on the Q-Q Plot, we see that normality is still violated due to the outliers and the tail ends of the slightly heavy-tailed plot.

# Predictor: longitude  
fit6 <- lm(log(`Y`)~longitude)  
plot(fit6, which = c(1,2))

 *COMMENT ON VARIANCE AND NORMALITY* From the Residuals vs Fitted Plot for longitude vs log of house price per unit (other predictors controlled), we can see that variance is still not constant based on the clustering and dipping pattern of the residuals. Therefore, the assumption is violated. Based on the Q-Q Plot, we see that normality is still violated due to the outliers and the tail ends of the heavy-tailed plot.