4: Multiple Linear Regression Model

Cheng Peng

Class Note: STA321 Topics of Advanced Statistics

Contents

1 Introduction

The general purpose of multiple linear regression (MLR) is to identify a relationship in an explicit functional form between explanatory variables or predictors and the dependent response variable. This relationship will be used to achieve two primary tasks:

- **Association analysis** - understanding the association between predictors and the response variable. As a special association, the causal relationship can be assessed under certain conditions.
- **Prediction** - the relationship between predictors and the response can be used to predict the response with new out-of-sample values of predictors.

2 The structure of MLR

Let $\{x_1, x_2, \cdots, x_k\}$ be k explanatory variables and y be the response variables. The general form of the multiple linear regression model is defined as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

This is a very special form in that y is linear in both parameters and predictors. The actual linear regression only assumes that y is linear only in parameters but not predictors since the value of predictors will be observed in data.

2.1 Assumptions based on the above special form.

- The function form between y and $\{x_1, x_2, \cdots, x_k\}$ must be correctly specified.
- The residual ϵ must be normally distributed with $\mu = 0$ and a constant variance σ^2.
- An implicit assumption is that predictor variables are non-random.

2.2 Potential Violations

There are various potential violations of the model assumptions. The following is a shortlist of potential violations of the model assumptions.

- The potential incorrect functional relationship between the response and the predictor variables.
  - the correct form may have power terms.
  - the correct form may have a cross-product form.
  - the correct form may need important variables that are missing.
- The residual term does not follow the normal distribution $N(0, \sigma^2)$. That means that
  - ϵ is not normally distributed at all.
  - ϵ is normally distributed but the variance is not a constant.

2.3 Variable types

Since there will be multiple variables involved in the multiple regression model.

- All explanatory variables are continuous variables - classical linear regression models.
- All explanatory variables are categorical variables - analysis of variance (ANOVA) models.
- The model contains both continuous and categorical variables - analysis of covariance (ANCOVA) model.

2.4 Dummy and discrete numerical explanatory variables

- Categorical variables with n ($n > 2$) categories MUST be dichotomized into $n - 1$ dummy variables (binary indicator variables).
  - **Caution**: categorical variable with a numerical coding - need to use R function **factor()** to automatically define a sequence of dummy variables for the category variable.
- **Discrete (numerical) variable** - If we want to use discrete numerical as a categorical variable, we must dichotomize it and define a sequence of dummy variables since interpretations of the two types of variables are different. In a categorical variable case, the coefficient of a dummy variable is the relative

contribution to the response compared with the baseline category. In the discrete case, the regression coefficient is the relative contribution to the response variable compared with adjacent values and the relative contribution is constant across all adjacent values of the discrete predictor variable.

2.5 Interpretation of Regression Coefficients

A _multiple linear regression model with k predictor variables has k+1 unknown parameters: intercept parameters (β_0), slope parameters ($\beta_i, i = 1, 2, \cdots, k$), and the variance of the response variable (σ^2). The key parameters of interest are the slope parameters since they capture the information on whether the response variable and the corresponding explanatory variables are (linearly) associated.

- If y and x_i are not linearly associated, that is, $\beta_i = 0, i = 1, 2, \cdots, k$, then β_0 is the mean of y.
- If $\beta_i > 0$, then y and x_i are positively linearly correlated. Furthermore, β_i is the increment of the response when the explanatory variable increases by one unit.
- We can similarly interpret β_i when it is negative.

3 Model building

Modeling building is an iterative process for searching for the best model to fit the data. An implicit assumption is that the underlying data is statistically valid.

3.1 Data structure, sample size, and preparation for MLR

In the model-building phase, we assume data is valid and has sufficient information to address the research hypothesis.

- Data records are independent - collected based on a cross-sectional design.
- The sample size should be large enough such that each regression coefficient should have 14 distinct data points to warrant reliable and robust estimates of regression coefficients.
- Imbalanced categorical variables and extremely distributed continuous explanatory variables need to be treated to a warrant valid estimate of regression coefficients. This includes combining categories in meaningful and practically interpretable ways and discretizing extremely skewed continuous variables.
- New variable definition - sometimes we can extract information from several variables to define new variables to build a better model. This is an active area in machine learning fields and data science. There are many different methods and algorithms in literature and practice for creating new variables based on existing ones.
  - Empirical approach - based on experience and numerical pattern.
  - Model-based approach - this may require a highly technical understanding of algorithms and modeling ideas. This is not the main consideration in this course.

3.2 Candidate models and residual diagnostics

- Consider only the multiple linear regression models that have a linear relationship between response and predictor variables.
- Perform residual analysis
  - if a curve pattern appears in residual plots, identify a curve linear relationship between the response and the individual predictor variable
  - if non-constant variance appears in the residual plots, then perform an appropriate transformation to stabilize the constant variance - for example, Box-cox transformation.
  - if the QQ plot indicates non-normal residuals, try transformations to convert it to a normal variable.

  - if there are serial patterns in the residual plot, we need to remove the serial pattern with an appropriate method.
  - if some clusters appear in the residual plot, create a group variable to capture the clustering information.

3.3 Significant test, goodness-of-fit, and Variable selection

Significant tests and goodness-of-fit measures are used to identify the final model. Please keep in mind that a good statistical model must have the following properties;

- Interpretability
- parsimony
- Accuracy
- Scalability

3.3.1 Significant Tests

Significant tests are used for selecting statistically significant variables to include in the model. However, in practical applications, some practically important variables should always be included in the model regardless of their statistical significance. The t-test is used for selecting (or dropping) individual statistically significant variables.

3.3.2 Variable (model) selection criteria

There are many different methods for model selection.

- R^2 - coefficient of determination. It explains the variation explained by the underlying regression model. R^2 is used to compare two candidate models. Adjusted R^2 is used when there are many predictor variables in the model.
- Likelihood ratio χ^2 test - comparing two candidate models with a hierarchical relationship.
- Information criteria - likelihood-based measures: AIC and SBC.
- Mallow's Cp - a residual-based measure that is used for comparing two models that do not necessarily have a hierarchical structure.
- F tests - testing the overall significance of a group of regression coefficients.

3.3.3 Variable selection methods

- Step-wise Procedures
- Criterion-based procedures

This short note summarized the above two methods for Variable Selection(click the link to view the text).

4 Case Study -Factors That Impact the House Sale Prices

We present a case study to implement various model-building techniques.

4.1 Data Description

The data in this note was found from Kaggle. I renamed the original variables and modified the sales dates to define the sales year indicator. The modified data set was uploaded to the course web page at https://raw.githubusercontent.com/pengdsci/sta321/main/ww03/w03-Realestate.csv.

- ObsID

- TransactionYear(X1): transaction date
- HouseAge(X2): house age

- Distance2MRT(X3): distance to the nearest MRT station
- NumConvenStores(X4): number of convenience stores
- Latitude(X5): latitude

- Longitude(X6): longitude

- PriceUnitArea(Y): house price of unit area

4.2 Practical Question

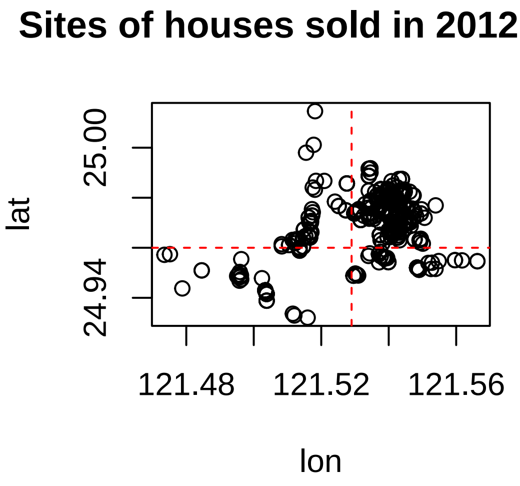
The primary question is to identify the association between the house sale price and relevant predictor variables available in the data set.

4.3 Exploratory Data Analysis

We first explore the pairwise association between the variables in the data set. Since longitude and latitude are included in the data set, we first make a map to see if we can define a variable according to the sales based on the geographic regions.

To start, we load the data to R.

```
realestate0 <- read.csv("https://raw.githubusercontent.com/pengdsci/sta321/main/ww03/w03-Realestate.csv
realestate <- realestate0[, -1]
# longitude and latitude will be used to make a map in the upcoming analysis.
lon <- realestate$Longitude
lat <- realestate$Latitude
plot(lon, lat, main = "Sites of houses sold in 2012-2013")
abline(v=121.529, h=24.96, col="red", lty=2)
```



We use longitude and latitude to define a group variable, **geo.group**, in the following.

geo.group = **TRUE** if longitude > 121.529 AND latitude > 24.96; **geo.group** = **FALSE** otherwise.

From the map representation of the locations of these houses given below (generated by Tableau Public), we can see that **geo. group** is an indicator

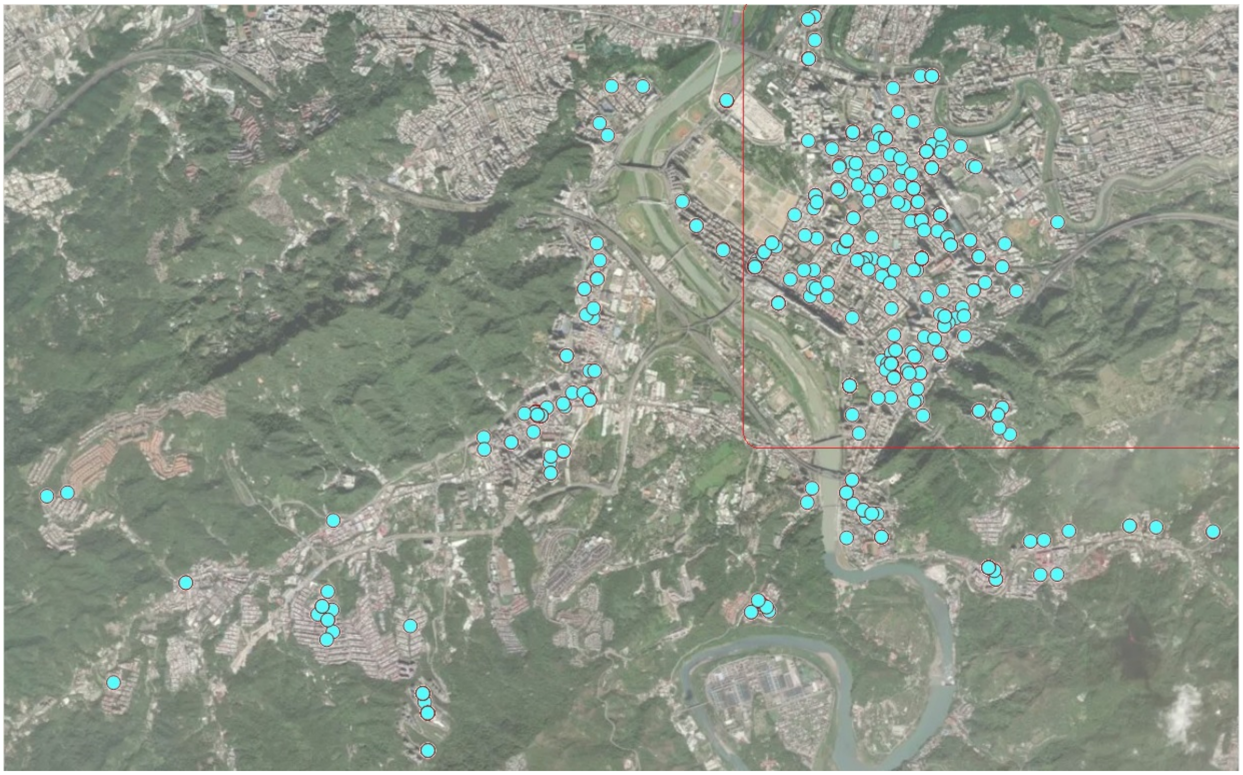


Figure 1: Locations of houses for sale

We also turn the variable **TransactionYear** into an indicator variable. At the same time, we scale the distance from the house to the nearest MRT by defining **Dist2MRT = Distance2MRT/1000**.

```
geo.group = (lon > 121.529) & (lat > 24.96)      # define the geo.group variable
                                                 # top-right region = TRUE, other region = FALSE
realestate$geo.group = as.character(geo.group)   # convert the logical values to character values.
realestate$sale.year = as.character(realestate$TransactionYear) # convert transaction year to dummy.
realestate$Dist2MRT.kilo = (realestate$Distance2MRT)/1000    # re-scale distance: foot -> kilo feet
final.data = realestate[, -c(1,3,5,6)]           # keep only variables to be used in the candidate
                                                 # models
kable(head(final.data))
```

HouseAge	NumConvenStores	PriceUnitArea	geo.group	sale.year	Dist2MRT.kilo
32.0	10	37.9	TRUE	2012	0.0848788
19.5	9	42.2	TRUE	2012	0.3065947
13.3	5	47.3	TRUE	2013	0.5619845
13.3	5	54.8	TRUE	2013	0.5619845
5.0	5	43.1	TRUE	2012	0.3905684
7.1	3	32.1	FALSE	2012	2.1750300

4.4 Fitting MLR to Data

We start the search process for the final model.

4.4.1 Full model and diagnostics

We start with a linear model that includes all predictor variables.

```
full.model = lm(PriceUnitArea ~ ., data = final.data)
kable(summary(full.model)$coef, caption ="Statistics of Regression Coefficients")
```

Table 2: Statistics of Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.9559432	1.7134680	20.984310	0.0000000
HouseAge	-0.3000749	0.0390329	-7.687735	0.0000000
NumConvenStores	1.0707846	0.1897962	5.641761	0.0000000
geo.groupTRUE	7.5447005	1.3420266	5.621871	0.0000000
sale.year2013	3.0332784	0.9447021	3.210831	0.0014283
Dist2MRT.kilo	-3.6589504	0.5317107	-6.881469	0.0000000

Next, we conduct residual diagnostic analysis to check the validity of the model before making an inference about the model.

```
par(mfrow=c(2,2))
plot(full.model)
```

We can see from the residual plots that there are some minor violations:

- the variance of the residuals is not constant.
- the QQ plot indicates the distribution of residuals is slightly off the normal distribution.
- The residual plot seems to have a weak curve pattern.

We first perform Box-Cox transformation to correct the non-constant variance and correct the non-normality of the QQ plot.

4.4.2 Models Based on Box-Cox transformation

We first perform Box-Cox transformation and then choose appropriate transformations for both response and predictor variables to build candidate regression models.

4.4.2.1 Box-Cox Transformations Since non-constant variance, we perform the Box-Cox procedure to search for a transformation of the response variable. We perform several tried Box-Cox transformations with different transformed

```
library(MASS)
par(pty = "s", mfrow = c(2, 2), oma=c(.1,.1,.1,.1), mar=c(4, 0, 2, 0))
##
boxcox(PriceUnitArea ~ HouseAge + NumConvenStores + sale.year +  log(Dist2MRT.kilo)
       + geo.group, data = final.data, lambda = seq(0, 1, length = 10),
       xlab=expression(paste(lambda, ": log dist2MRT")))
##
boxcox(PriceUnitArea ~ HouseAge + NumConvenStores + sale.year +  Dist2MRT.kilo  +
       geo.group, data = final.data, lambda = seq(-0.5, 1, length = 10),
       xlab=expression(paste(lambda, ": dist2MRT")))
##
boxcox(PriceUnitArea ~ log(1+HouseAge) + NumConvenStores + sale.year +  Dist2MRT.kilo  +
      geo.group, data = final.data, lambda = seq(-0.5, 1, length = 10),
       xlab=expression(paste(lambda, ": log-age")))
```

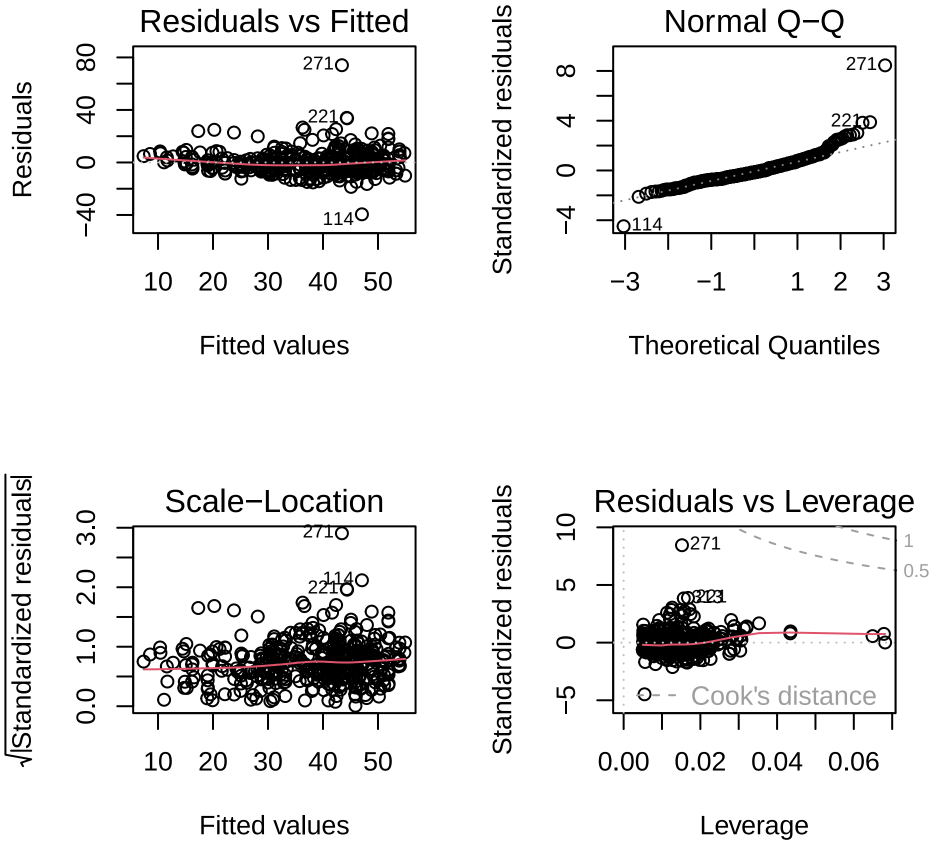
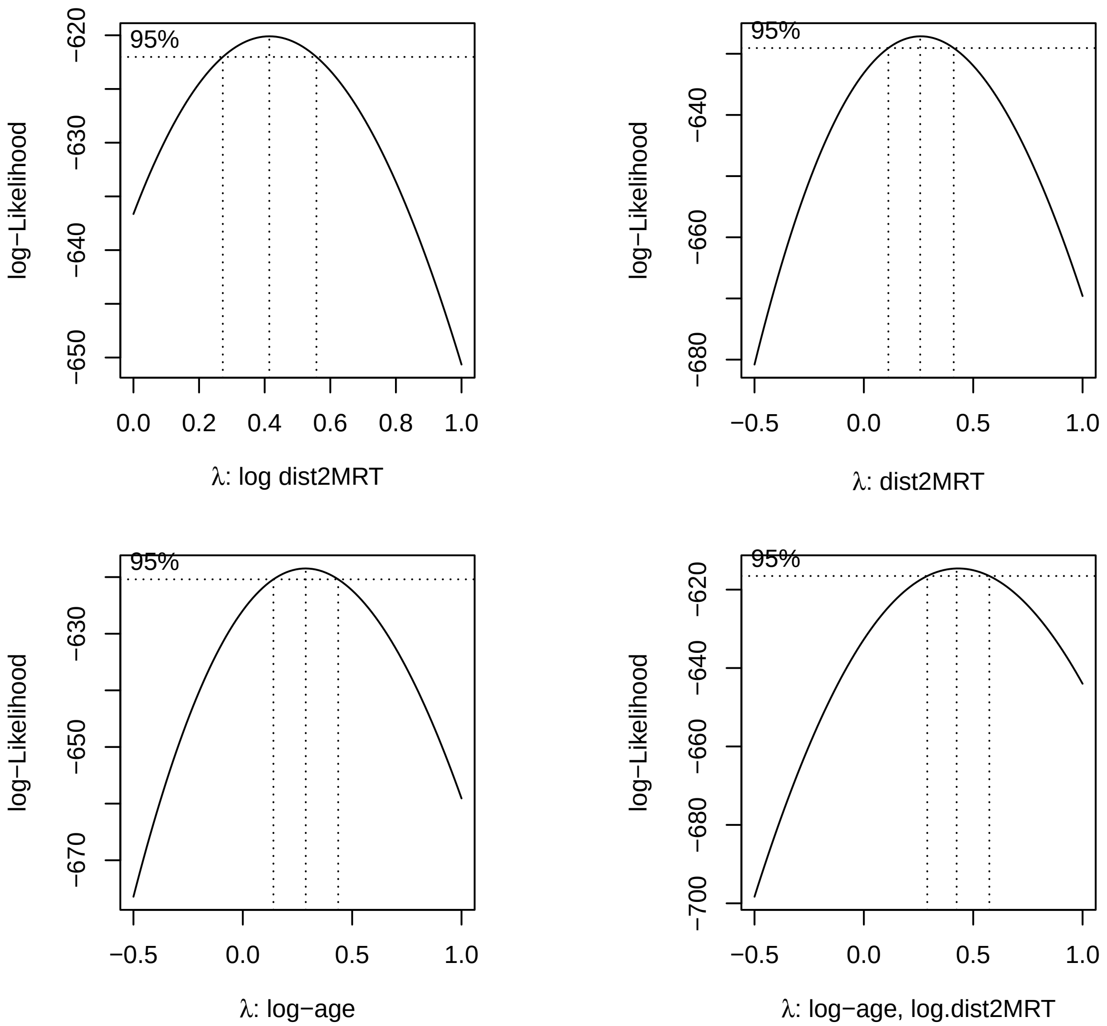


Figure 2: Residual plots of the full model

```
##
boxcox(PriceUnitArea ~ log(1+HouseAge) + NumConvenStores + sale.year +  log(Dist2MRT.kilo)  +
      geo.group, data = final.data, lambda = seq(-0.5, 1, length = 10),
      xlab=expression(paste(lambda, ": log-age, log.dist2MRT")))
```



The above Box-cox transformation plots indicate the optimal λ under different transformed predictor variables. log-Transformed distance from MRT impacts the coefficient of the power transformation: λ.

As a special power transformation, if $\lambda = 0$, the transformation degenerates to log transformation.

4.4.2.2 Square-root Transformation We perform Box-Cox transformation with log-transformed distance to the nearest MRT in the following.

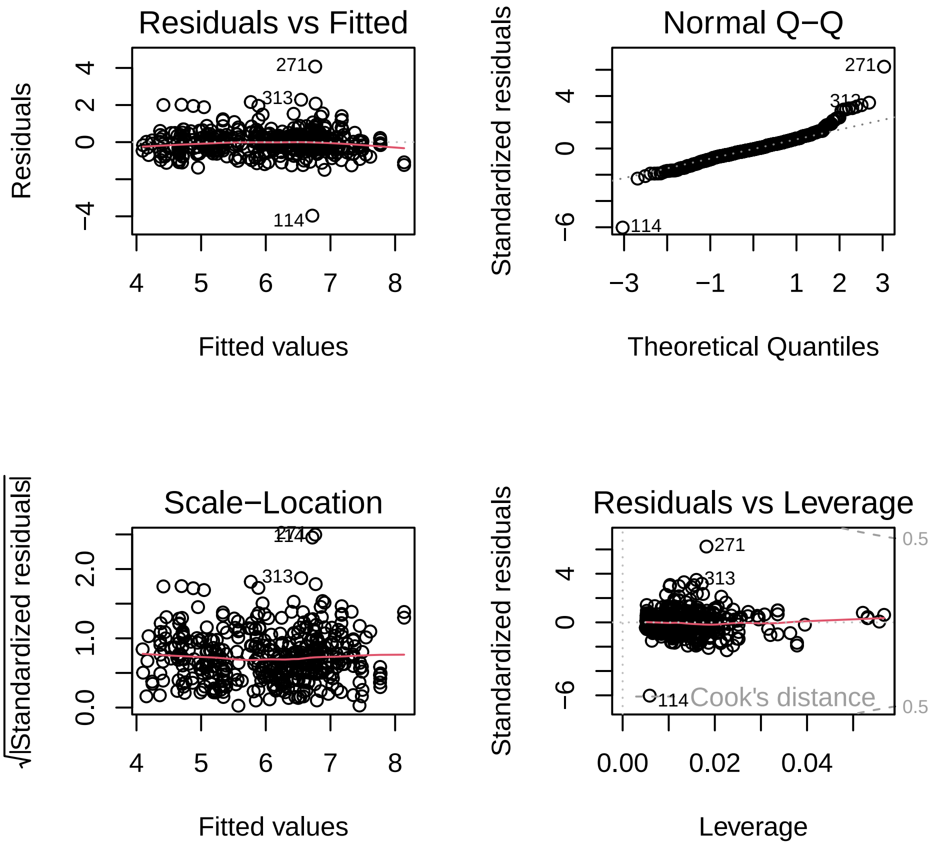
```
sqrt.price.log.dist = lm((PriceUnitArea)^0.5 ~ HouseAge + NumConvenStores + sale.year +  log(Dist2MRT.k
kable(summary(sqrt.price.log.dist)$coef, caption = "log-transformed model")
```

Table 3: log-transformed model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.3978808	0.0923831	58.429342	0.0000000
HouseAge	-0.0209198	0.0029561	-7.076954	0.0000000
NumConvenStores	0.0513220	0.0153946	3.333767	0.0009351
sale.year2013	0.2951406	0.0707905	4.169214	0.0000374
log(Dist2MRT.kilo)	-0.4905637	0.0481611	-10.185897	0.0000000
geo.groupTRUE	0.5709120	0.0973966	5.861723	0.0000000

Residual plots are given below.

```
par(mfrow = c(2,2))
plot(sqrt.price.log.dist)
```



There are two improvements in the above residual diagnostic plots: (1) the weak curve pattern has been removed from the residual plot; (2) the non-constant variance has also been corrected. However, the violation of the normality assumption is still an issue.

4.4.2.3 Log-Transformation We take the log transformation of the sale price according to the Box-Cox transformation and then build a linear regression based on the log price.

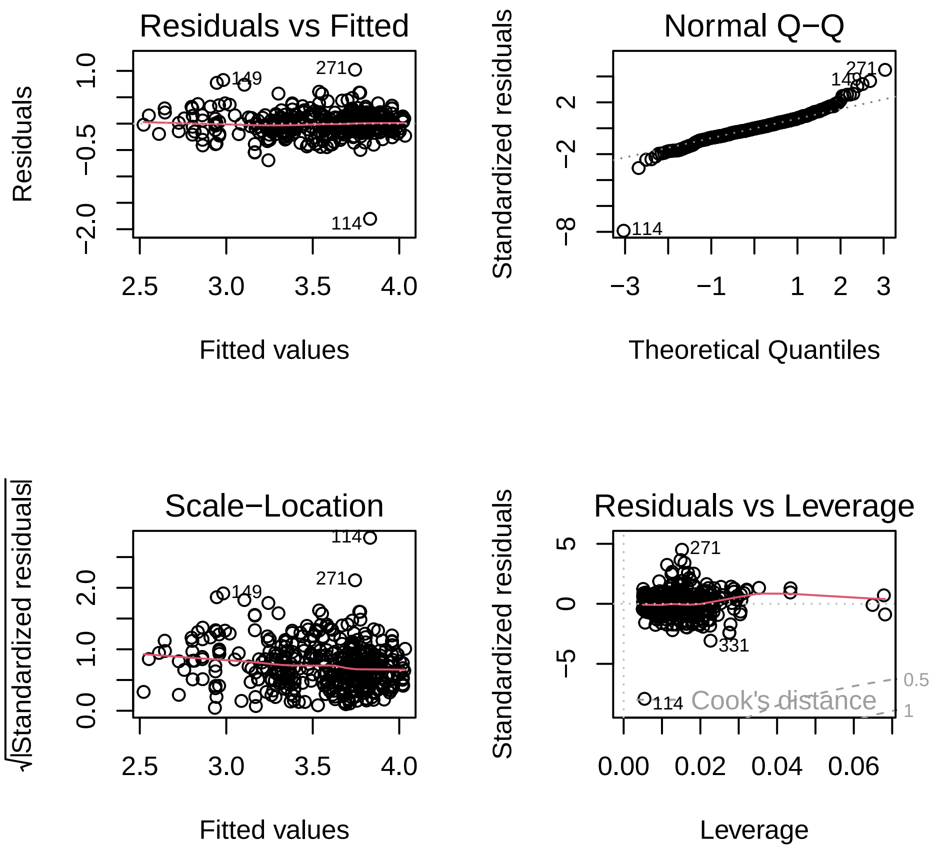
```
log.price = lm(log(PriceUnitArea) ~ HouseAge + NumConvenStores + sale.year +  Dist2MRT.kilo  + geo.grou
kable(summary(log.price)$coef, caption = "log-transformed model")
```

Table 4: log-transformed model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5724282	0.0443235	80.599030	0.0000000
HouseAge	-0.0075712	0.0010097	-7.498507	0.0000000
NumConvenStores	0.0274872	0.0049096	5.598667	0.0000000
sale.year2013	0.0805519	0.0244373	3.296272	0.0010655
Dist2MRT.kilo	-0.1445122	0.0137541	-10.506820	0.0000000
geo.groupTRUE	0.1825871	0.0347151	5.259583	0.0000002

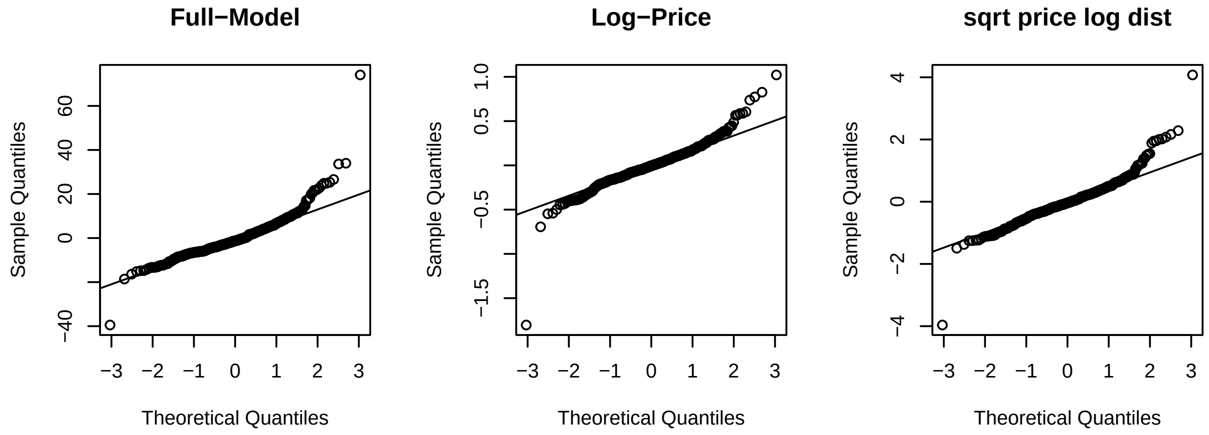
Residual plots are given below.

```
par(mfrow = c(2,2))
plot(log.price)
```



The above residual diagnostic plots are similar to that of the previous model. The Q-Q plots of all three models are similar to each other, this means that the assumption of normal residuals is not satisfied for all three models.

```
#define plotting area
par(pty = "s", mfrow = c(1, 3))
#Q-Q plot for original model
qqnorm(full.model$residuals, main = "Full-Model")
qqline(full.model$residuals)
#Q-Q plot for Box-Cox transformed model
qqnorm(log.price$residuals, main = "Log-Price")
qqline(log.price$residuals)
#display both Q-Q plots
qqnorm(sqrt.price.log.dist$residuals, main = "sqrt price log dist")
qqline(sqrt.price.log.dist$residuals)
```



4.4.2.4 Goodness-of-fit Measures Next, we extract several other goodness-of-fit from each of the three candidate models and summarize them in the following table.

```
select=function(m){ # m is an object: model
 e = m$resid                           # residuals
 n0 = length(e)                        # sample size
 SSE=(m$df)*(summary(m)$sigma)^2       # sum of squared error
 R.sq=summary(m)$r.squared             # Coefficient of determination: R square!
 R.adj=summary(m)$adj.r                # Adjusted R square
 MSE=(summary(m)$sigma)^2              # square error
 Cp=(SSE/MSE)-(n0-2*(n0-m$df))         # Mellow's p
 AIC=n0*log(SSE)-n0*log(n0)+2*(n0-m$df)          # Akaike information criterion
 SBC=n0*log(SSE)-n0*log(n0)+(log(n0))*(n0-m$df)  # Schwarz Bayesian Information criterion
 X=model.matrix(m)                     # design matrix of the model
 H=X%*%solve(t(X)%*%X)%*%t(X)          # hat matrix
 d=e/(1-diag(H))
 PRESS=t(d)%*%d   # predicted residual error sum of squares (PRESS)- a cross-validation measure
 tbl = as.data.frame(cbind(SSE=SSE, R.sq=R.sq, R.adj = R.adj, Cp = Cp, AIC = AIC, SBC = SBC, PRD = PRESS
 names(tbl)=c("SSE", "R.sq", "R.adj", "Cp", "AIC", "SBC", "PRESS")
 tbl
 }
```

```
output.sum = rbind(select(full.model), select(sqrt.price.log.dist), select(log.price))
row.names(output.sum) = c("full.model", "sqrt.price.log.dist", "log.price")
kable(output.sum, caption = "Goodness-of-fit Measures of Candidate Models")
```

Table 5: Goodness-of-fit Measures of Candidate Models

	SSE	R.sq	R.adj	Cp	AIC	SBC	PRESS
full.model	31831.19255	0.5836958	0.5785940	6	1809.7271	1833.8823	32792.58816
sqrt.price.log.dist	177.56943	0.6587132	0.6545308	6	-338.4528	-314.2976	182.95839
log.price	21.29945	0.6651882	0.6610851	6	-1216.4146	-1192.2594	21.94809

We can see from the above table that the goodness-of-fit measures of the third model are unanimously better than the other two models. Considering the interpretability, goodness-of-fit, and simplicity, we choose the last model as the final model.

4.4.3 Final Model

The inferential statistics of the final working model are summarized in the following table.

```
kable(summary(log.price)$coef, caption = "Inferential Statistics of Final Model")
```

Table 6: Inferential Statistics of Final Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5724282	0.0443235	80.599030	0.0000000
HouseAge	-0.0075712	0.0010097	-7.498507	0.0000000
NumConvenStores	0.0274872	0.0049096	5.598667	0.0000000
sale.year2013	0.0805519	0.0244373	3.296272	0.0010655
Dist2MRT.kilo	-0.1445122	0.0137541	-10.506820	0.0000000
geo.groupTRUE	0.1825871	0.0347151	5.259583	0.0000002

Since the sample size (414) is large, the argument for validating p-values is the Central Limit Theorem (CLT). all p-values are close to 0 meaning that all coefficients are significantly different from 0.

In this specific case study, there is no need to perform variable selection to determine the final model.

4.5 Summary of the model

We can explicitly write the final model in the following

$$\log(price) = 3.5723 - 0.0076 \times HouseAge + 0.0275 \times NumConvenStores+$$

$$0.0805 \times Sale.year2013 - 0.1445 \times Dist2MRT.kilo + 0.1826 \times geo.groupTRUE$$

Note that the estimated regression coefficients are based on log price. we now consider **the set of** all houses with ages x_0 and $x_0 + 1$ that are in the same conditions except for the sale prices. The exact practical interpretation is given below. Let p_{x_0} and p_{x_0+1} be the mean prices of houses with ages x_0 and $x_0 + 1$, respectively. Since the two types of houses are in the same conditions except for the age and the prices. Then

$$\log(p_{x_0+1}) - \log(p_{x_0}) = -0.0076 \rightarrow \log(p_{x_0+1}/p_{x_0}) = -0.0076 \rightarrow p_{x_0+1} = 0.9924 p_{x_0}$$

We re-express the above equation can be re-written as

$$p_{x_0+1} - p_{x_0} = -0.0076 p_{x_0} \rightarrow \frac{p_{x_0+1} - p_{x_0}}{p_{x_0}} = -0.076 = -0.76\%$$

That is, as the house age increases by one year, the house price **decreases** by 0.76%. We can similarly interpret other regression coefficients.

The distance to the nearest MRT is also negatively associated with the sale price. The rest of the factors are positively associated with house prices.

4.6 Discussions

We use various regression techniques such as Box-Cox transformation for response variables and other transformations of the explanatory variables to search for the final model in the case study. Since there are five variables in the data set and all are significant, we did not perform any variable selection procedure.

All candidate models have the same number of variables. We use commonly-used global goodness-of-fit measures as model selection criteria.

The interpretation of the regression coefficients is not straightforward since the response variable was transformed into a log scale. We used some algebra to derive the practical interpretation of the regression coefficients associated with the variables at their original scales.

The violation of the normal assumption of the residuals remains uncorrected. The inference on the regression coefficients is based on the central limit theorem. We will introduce bootstrap methods to construct bootstrap confidence intervals of the regression coefficients of the **final model**.