STAT 425 Applied Regression and Design Term Project

Chuanyue Shen
Lingzhu Gu
Lijun Zhang

University of Illinois at Urbana–Champaign

Content

I Introduction

II Data Preprocessing

III Variable Processing

IV Outcome Transformation

V Model Selection

VI Diagnostic

VII Visualization

I. Introduction

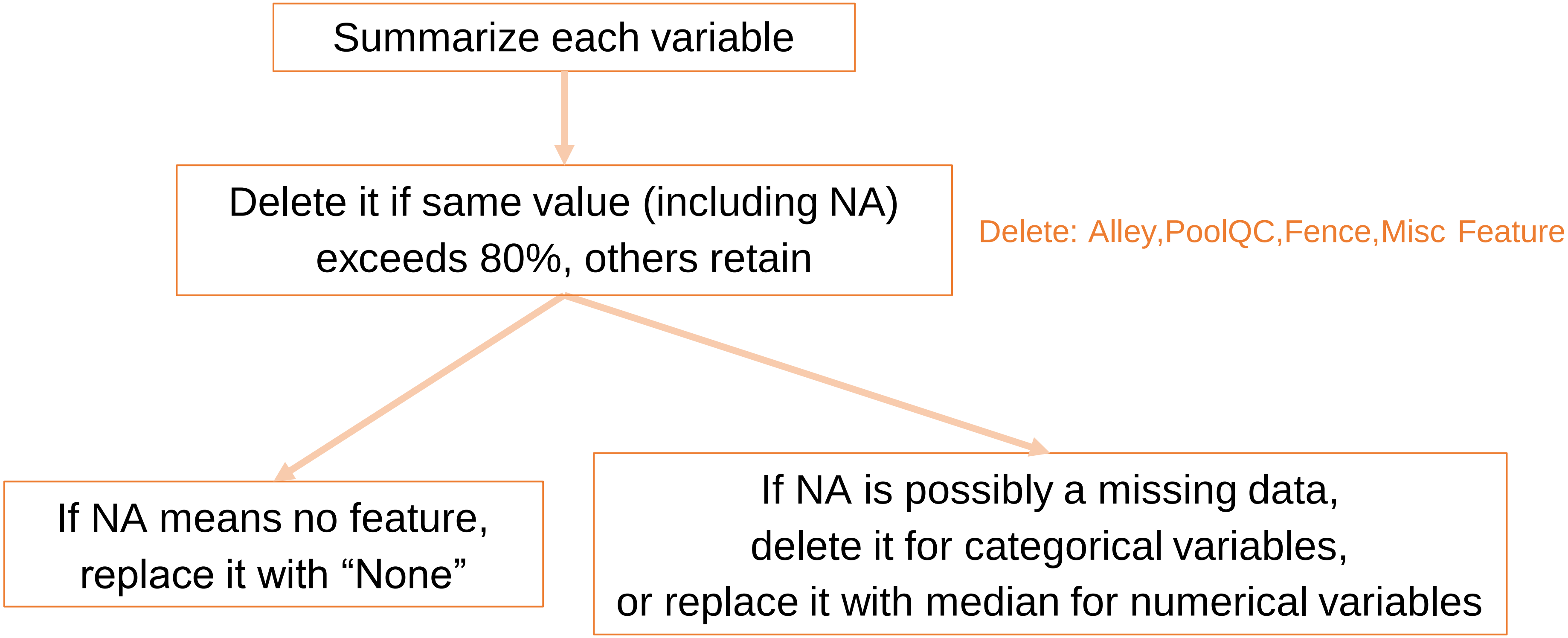
AmesHousing data set

- R package [AmesHousing]
- Housing price in Ames, Iowa from 2006 to 2010
- 2930 observations
- 82 variables including SalePrice and geo-info

SalePrice ~ **81 Features**

II. Data Preprocessing

Preprocessing process



Summarize each variable

Delete it if same value (including NA) exceeds 80%, others retain

Delete: Alley,PoolQC,Fence,Misc Feature

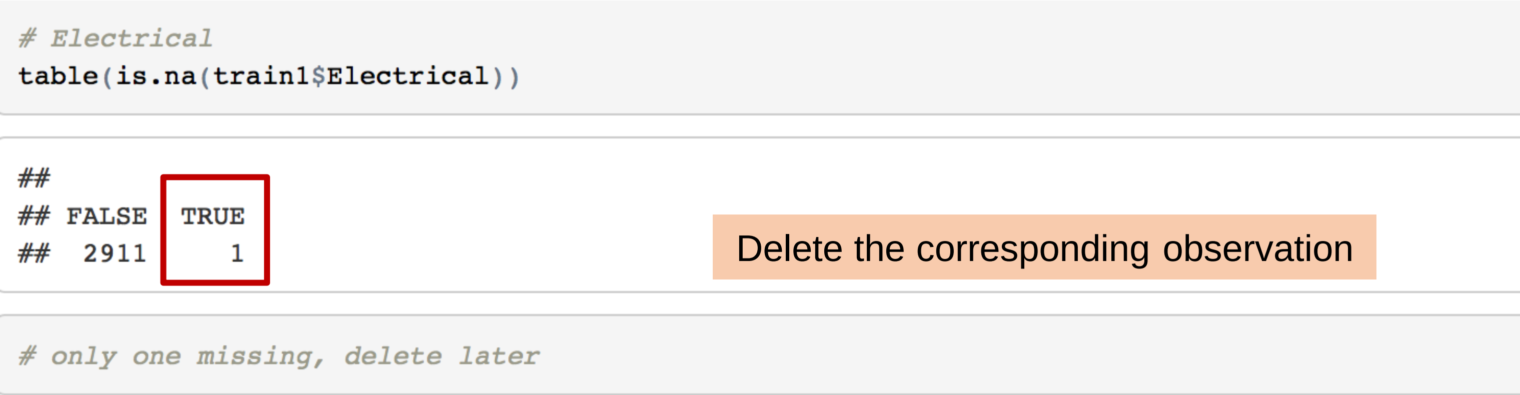
If NA means no feature, replace it with “None”

If NA is possibly a missing data,
delete it for categorical variables,
or replace it with median for numerical variables

II. Data Preprocessing

Example

- Categorical variable



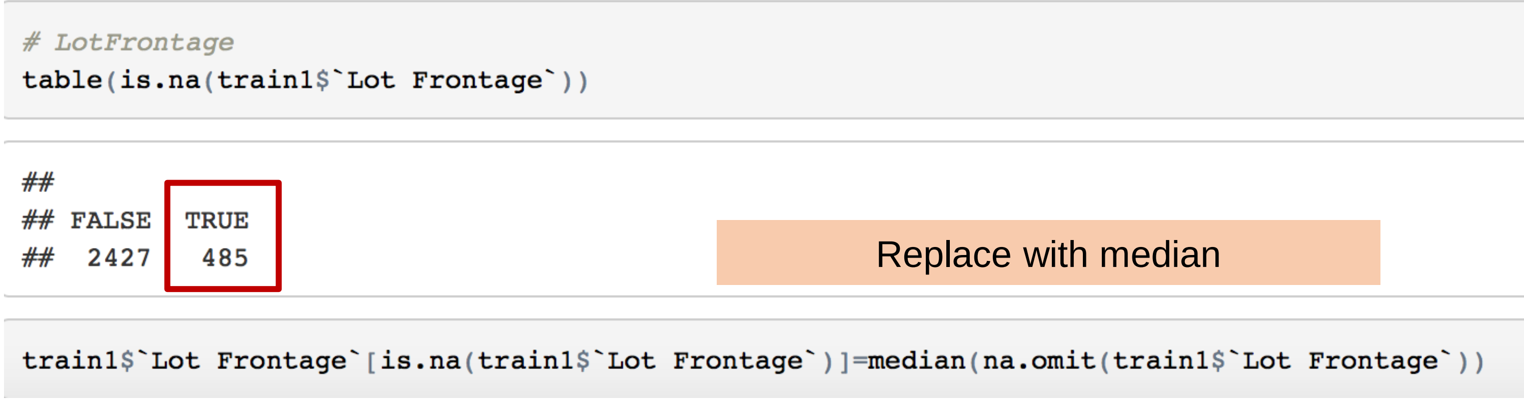
```
# Electrical
table(is.na(train1$Electrical))
```

```
##
## FALSE  TRUE
##  2911     1
```

Delete the corresponding observation

```
# only one missing, delete later
```

- Numerical variable



```
# LotFrontage
table(is.na(train1$`Lot Frontage`))
```

```
##
## FALSE  TRUE
##  2427   485
```

Replace with median

```
train1$`Lot Frontage`[is.na(train1$`Lot Frontage`)]=median(na.omit(train1$`Lot Frontage`))
```

II. Data Preprocessing

Clean data

- 2880 observations
- 77 features including geo-info
- 42 categorical variables
- 35 numerical variables

III. Variable Processing

Variable Category

1 Numerical Variable: 35

2 Categorical Variable: 42

Collinearity

1 Criteria: 0.9

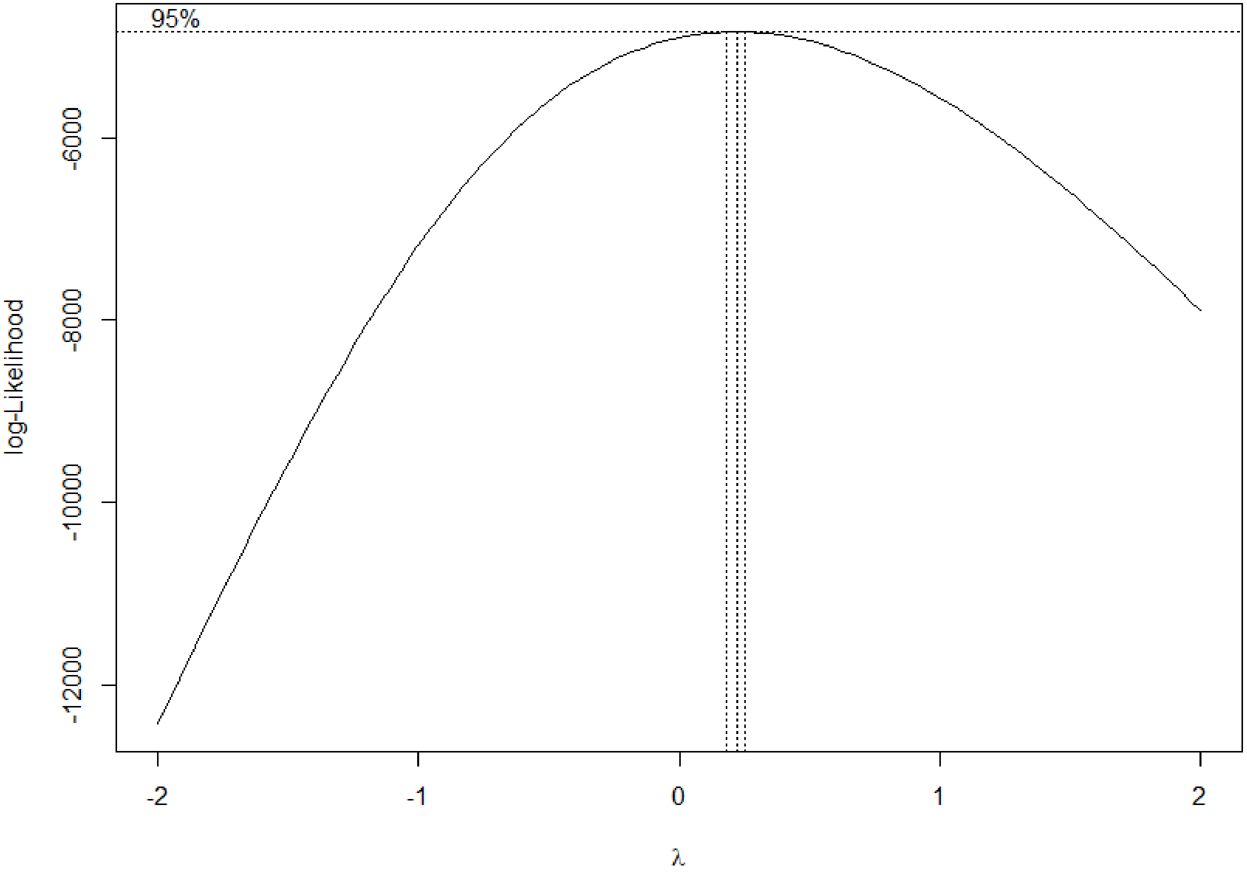
2 Only for 35 numerical variables

3 No variable is deleted

IV. Outcome Transformation

BoxCox Transformation

- λ for maximum likelihood is 0.22, thus use $\lambda = 0$
- No change for outcome Y



V. Model Selection

AIC

Step Function

“Forward”

- model 2.1
- **49** variables,
- **22** numerical & **27** categorical

“Backward”

- model 2.2
- **51** variables,
- **23** numerical & **28** categorical

“Both”

- model 2.3
- The same as model 2.2

BIC

Step Function

“Forward”

- model 3.1
- **27** variables,
- **15** numerical & **12** categorical

“Backward”

- model 3.2
- **26** variables,
- **14** numerical & **12** categorical

“Both”

- model 3.3
- the same as model 3.2

ANOVA

ANOVA

ANOVA

- model 2.1 & model 2.2
- **Select model 2.1.**

ANOVA

- model 2.1 & model 3.1
- **Select model 2.1**

ANOVA

- model 2.1 & model 3.2
- **Select model 2.1**

VI. Diagnostic

High Leverage

1. Criteria: leverage > 0.054
2. 1156 high leverage points
3. Keep them

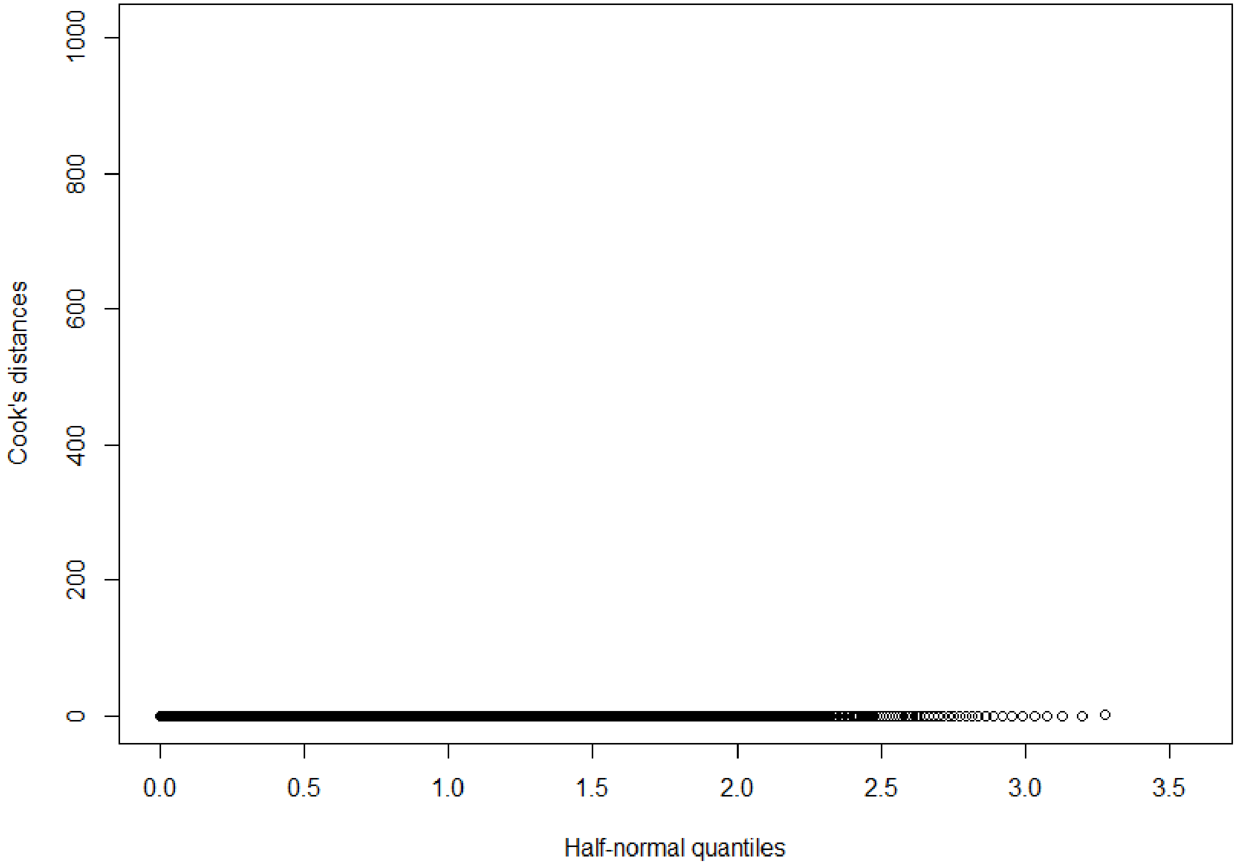
Outlier

1. Criteria: studentized residual < 4.304
2. 15 outliers
3. Delete them

VI. Diagnostic

High Influential Point

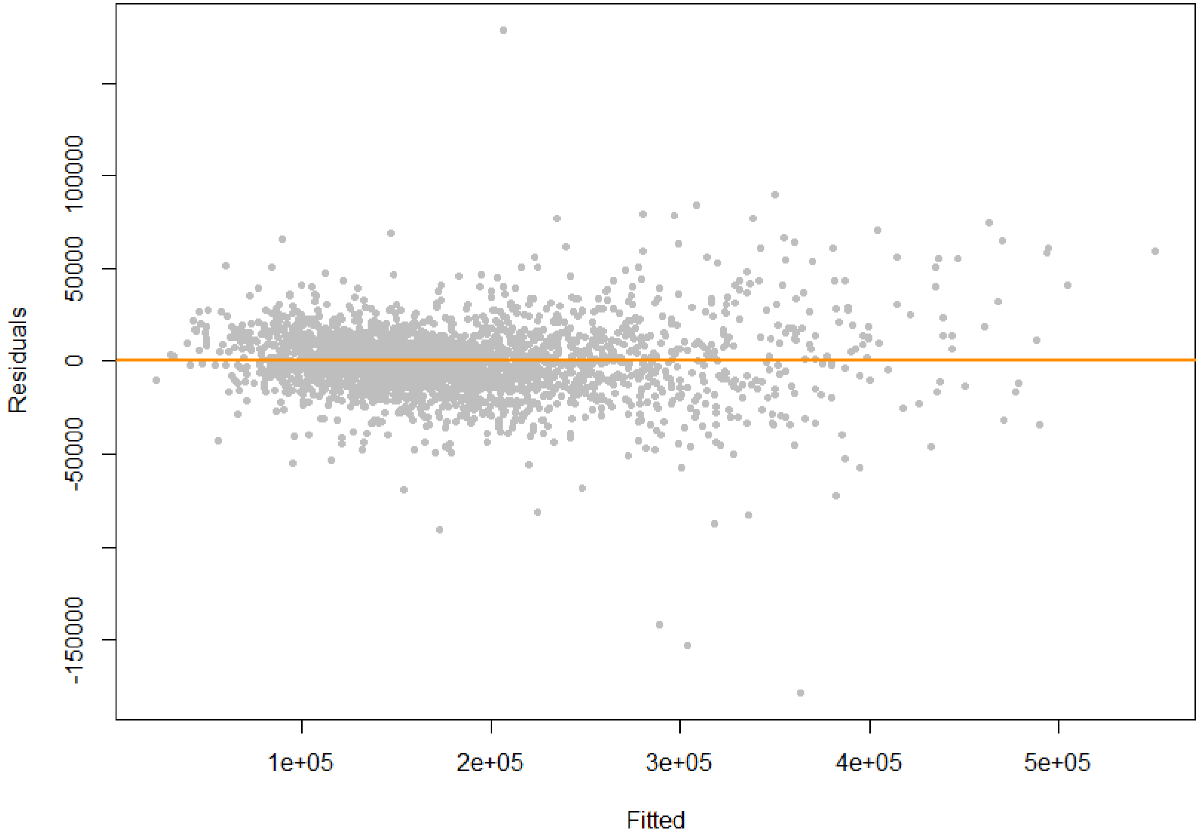
- Criteria: cook distance > 1
- 12 high influential points, delete them → model 4



Fitted vs. Residuals Plot

- Check for Linearity & Constant Variance

Residuals roughly centered at 0 → Good linearity
No uniform spread of the residuals along fitted value → Reject constant variance assumption.



VI. Diagnostic

Breusch-Pagan Test

1 Check for Constant Variance

Fitted vs. Residuals plot gives an idea about homoscedasticity, but a more formal test is preferred

2 Null & Alternative Hypothesis

H_0: Homoscedasticity. The errors have constant variance about the true model
H_a: Heteroscedasticity. The errors have non-constant variance about the true model

3 BP Test for model 4

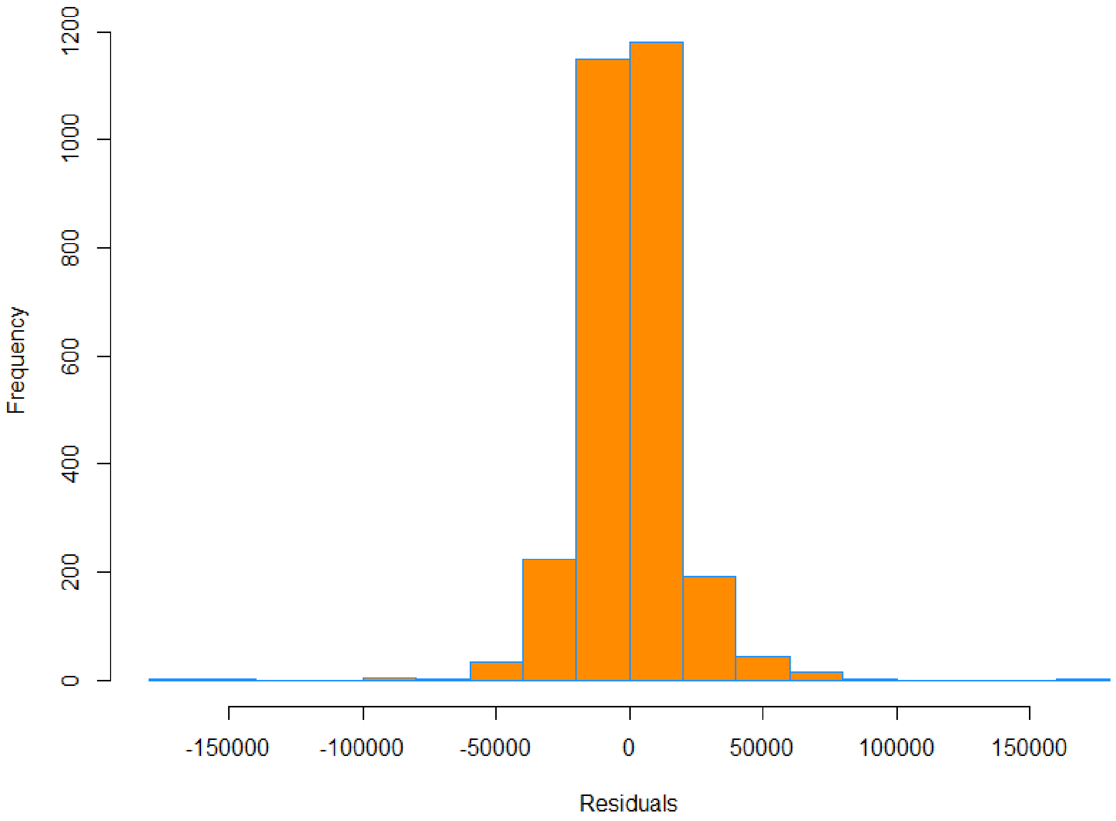
P-value is less than 0.05, reject null hypothesis, indicating constant variance assumption is not satisfied

VI. Diagnostic

Histograms

- Check for Normality

Rough bell shape, but has a very sharp peak → not clear whether the model satisfies normality assumption



VI. Diagnostic

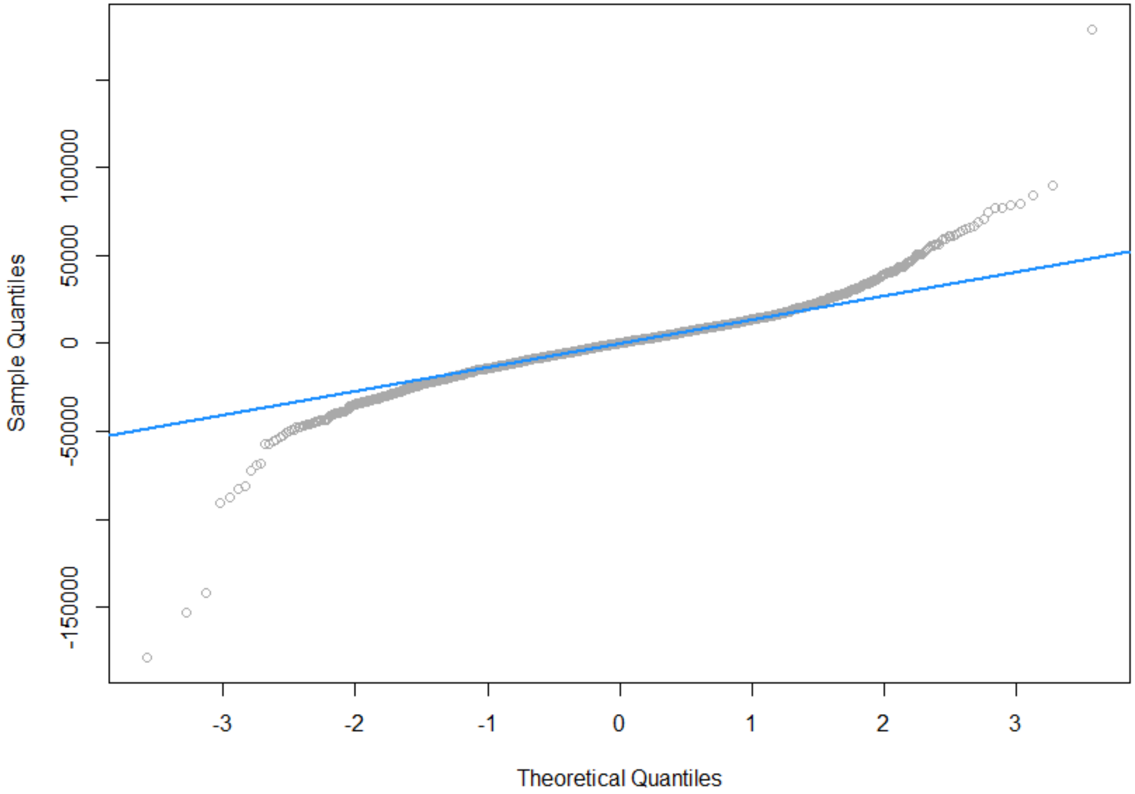
QQ Plot & Shapiro-Wilk Test

- Check for Normality

Points of the plot do not perfectly follow a straight line, suggesting that the errors may not follow a normal distribution

- Shapiro-Wilk Test

H_0: Normality
H_a: Non-normality
P-value is less than 0.05
→ only a small probability that the data sampled from a normal distribution



VII. Visualization

Three Findings

1 Prediction for Sale Price in Next Few Years

- Sale Price vs. Year Built
- Based on the trend line and forecasting, we have 95% confidence to say that the average sale price will stay stable around $260,000 in next five years

2 Location Choice for Economic Sale Price

- Sale Price vs. Neighborhood
- Neighborhood NridgHt neighborhood has the highest sale price, while the BrDale has the lowest sale price
- Sale price in NirdgHt is most sensitive to first floor area

3 Influence of House Available and House Condition

- Sale Price vs. Basement Exposure/Kitchen Quality/Sale Condition
- The highest price is the one with excellent kitchen, good basement and abnormal sale condition, it will be cheaper if you could trade off between basement and kitchen
- The is also constrains for number of house available in different conditions