STAT170 Final Project Part 2

Ryan Kitagawa, Lucas Chin, Riley Menter, Justin Tran

2025-09-25

Introduction

This dataset was created, cleaned, and modified by the Town of Cary GIS Group. Our goal was to explore the practical applications of GIS in real estate and investigate how various factors influence housing prices.

Our research question was: What factors are associated with higher and lower housing prices in Wake County? We hypothesized that total sales values for houses in Wake County are correlated with specific property characteristics. This research is motivated by the practical applications of GIS in real estate and the need for insights into housing prices.

The National Academic Press in GIS for Housing and Urban Development (2003) highlights that "agencies such as state, local, and federal governments need geographic information to carry out missions, such as resource conservation, infrastructure planning, and land-use analysis. HUD uses geographic information to increase homeownership, support community development, and improve access to affordable housing free from discrimination."

For this project, we analyzed a real estate dataset from three counties in North Carolina, provided by the Town of Cary GIS Group. The dataset was originally cleaned and modified to suit the town's needs. It contained 285,000 observations and 62 variables, but we narrowed our focus to a subset to improve processing speed and model interpretability. Our motivation to work on this dataset was to see the practical uses and applications of GIS in real estate. We also wanted to investigate how different variables affect housing prices.

Understanding predictors of housing prices is vital for buyers, sellers, urban planners, and policymakers. Using exploratory data analysis (EDA), we employed descriptive statistics and visualizations to examine the data's structure and distributions. Our response variable, TotalSaleValue, initially exhibited a right-skewed distribution. To meet the assumptions of linear regression, we applied a Box-Cox transformation and settled on a square-root transformation to approximate normality, as illustrated in the histogram of the transformed variable.

Key variables in our analysis included:

TotalSaleValue: The sale price of the property per $1000 (response variable). TotalBldgSqft: The total building area per sqft. DeedAcres: The acreage of the property per acre. BuiltBefore1990: Whether the building was constructed before 1990 (binary variable). LandValue: The assessed value of the land per $1000. BldgValue: The assessed value of the property per $1000. IsDetachedUnit (1 for True and 0 for False): If the primary structure on the property is a detached unit or not.

EDA

We added the above predictors to our initial model and checked to make sure that at least one of the predictors was significant. Our initial fit model had an adjusted R-squared value of 0.4521 and a P value of <2.2e-16. Our histogram was unimodal and is skewed to the right. From here, we briefly checked our model assumptions. Our residual vs fitted plot shows a slight pattern where values to the right had lower residuals, so the linearity condition and the constant variance conditions were not satisfied. Our normal Q-Q plot deviated from the line of best fit towards the edges, and our Shapiro-Wilkes test gave us a very small

number so we concluded that our current model violated the normality assumption. Our Durbin-Watson test gave a non-significant value and the VIF test values were under 10 for all variables so we concluded that our model had independently collected samples and no multicollinearity issues.

Regression Analysis

Since our model did not satisfy the normality conditions, we used the Box-Cox test and determined from the curve that a square root transformation for our response variable Total Sales Value was necessary. We double-checked this by using the box-cox power test and found that the value was around 0.5, which confirms that square root transformation is necessary. We checked on the model assumptions again. Our histogram was more normal and the Q-Q normal plot followed the line of best fit a little better, but still trailed off around the edges. After following through with the Shapiro-Wilkes test and getting a very small value, we concluded that our model still violated the normality assumption.

In order to weed out any unnecessary variables, we conducted stepwise regression twice for consistency. We started with only the intercept for the first step model and started with all of the predictors for our second model. Both approaches resulted in the same final model with the five variables Building value, land value, total building square feet, acres in deed, and whether the house was built before 1990 or not. We ran the Box-Cox model on the new step model to ensure that the response variable should still have a square root transformation applied. Our box-cox curve showed a value around 0.3, confirmed by the box-cox power test.

We checked our model assumptions once more from here and found that our normality assumption still wasn't satisfied. We also ran residuals for each predictor and saw that land value and acres in deed both had an exponential curve trend. Because of this we applied a log transformation on these two predictors and ran the residuals on them again to see that they had less of a pattern. From here, we checked on all of the assumptions again and saw that the normality assumption was still not satisfied, so we turned to higher-order terms and interaction terms to increase our R-squared values and meet our assumptions

After refining our residuals, the adjusted R-squared value increased to 0.4001. This suggests that a higher-order model with interaction terms could improve the fit.

The final model that we obtained is:

$TotalSaleValue(Y) = 6.488{+}0.106(BuildingValue){+}0.249(LandValue){-}1.160e{-}2(TotalBuildingSquareFeet){+}$
$5.125(RecordedAcres) + 0.362(BuiltBefore1990) - 9.795e-6(BuildingValue)2 - 1.360e-2(LandValue) *$
$(TotalBuildingSquareFeet) - 0.975(BuildingValue) * (RecordedAcres)$

The final regression model was selected because it captures the relationship between total sale value and the other predictor variables without being too complex. The higher-order term and interaction effects were considered to account for the nonlinearity of the variable Building Value. The different interactions were guesses of potential variables that sounded like they would interact with each other.

The different interactions that we considered in our fit models before we found our final model were the following:

- Model 1: Model with first-order variables
- Model 2: Testing higher order of bldgvalue and interactions between bldgvalue, landvalue, and totalbldgsqft
- Model 2.5: Removed bldgvalue and totalbldgsqft interaction
- Model 3: Testing interaction between deedacres and bldgvalue, landvalue, totalbldgsqft.
- Model 3.5: Removed interaction between deedacres and totalbldgsqft, deedacres and bldgvalue
- Model 4: Testing interaction between built_before_1990 and all other predictors

We conducted ANOVA tests between each model to see the significance of each interaction term in improving the overall performance of the model. Fit model 3 had the highest adjusted R^2 value of 0.4704. However, after dropping two interactions, we found that fit model 3.5 had a lower adjusted R^2 value of 0.4683 but is slightly less complex than model 3. Using an ANOVA test to test the significance of the two dropped variables, we get a p-value of 0.05773, indicating that the two interaction variables are not statistically significant contributors to the model's overall performance. This is how we concluded that fit model 3.5 was

the best model out of our other models. Every single interaction in fit model 4 had individual t-test p-values greater than 0.35 so fit model 4 did not provide meaningful improvements to the predictive performance of our model. Therefore, we decided not to pursue it further.

Conclusion

Our analysis highlighted several findings:

Surprising Insignificance: TotalBldgSqft and IsDetachedUnit were not significant predictors, countering the expectation that larger or detached homes would equate to higher prices. DeedAcres Trend: Contrary to prediction, properties with larger acreage tended to have lower sale values. Limited Interactions: Interactions between predictors, such as BuiltBefore1990 and LandValue, had minimal impact on model fit. Predictive Power: The adjusted R-squared value of 0.4001 reflects moderate predictive ability but highlights the need for more complex modeling.

Our findings indicate that the selected predictors don't fully capture the variability in housing prices. Buyers and sellers can use this model as a starting point to estimate reasonable prices and identify influential features. However, more robust modeling—including non-linear terms and additional predictors—is needed for greater accuracy. We do think a higher R^2 is possible, but it would require a more complex model and more time to explore combinations.

Project Limitations

Several factors limited our analysis:

Subset Size: Since there were 280,000 observations initially, reducing the dataset improved computation speed but may have excluded relevant patterns. The loss of potentially relevant patterns likely had a large effect on the R^2 of our model. Temporal Accuracy: House sale values are based on the most recent listings, meaning that some values could misrepresent current market conditions, e.g., a house that was last listed and sold 20 years ago. Model Fit: Violations of linear regression assumptions and a low adjusted R-squared value both indicate that the model could be improved Variable Exclusion: Variables like location and owner were excluded for simplicity, and might hold predictive value.

Appendix

R Markdown

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(GGally)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(dplyr)
library(leaps)
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

Roles:

```
##              Introduction    Analysis 1     Analysis 2      Final Project
## Script       Riley Menter    Lucas Chin     Ryan Kitagawa   Lucas Chin
## Script       Lucas Chin      Riley Menter   Justin Tran     Ryan Kitagawa
## LDA          Ryan Kitagawa   Justin Tran    Lucas Chin      Justin Tran
## Facilitator  Justin Tran     Ryan Kitagawa  Riley Menter    Riley Menter
```

##INTRODUCTION

```
data <- read.csv2("property.csv")
```

Selecting a subset of data

```
set.seed(2) # For consistency

data <- data %>% filter(phycity == "Cary" | phycity=="CARY" | phycity=="cary") # Only Cary county in NC

data <- drop_na(data) # Drop rows with missing values

data <- sample_n(data, size = 999, replace = FALSE) # Sample 999 observations

data <- data %>% filter(totalsalevalue > 0) # Ignore rows with negative sales values


#data <- sample_n(data, size = 999, replace = FALSE) # Randomly sample 999 UNIQUE rows

data <- data %>% dplyr::select(totalsalevalue, bldgvalue, landvalue, # looking at ONLY these columns
                        totalbldgsqft, deedacres, yearbuilt,
                        apastructuredesc, phycity)

data <- data %>% mutate(totalbldgsqft = as.integer(totalbldgsqft), deedacres = as.double(deedacres)) #

head(data)
```

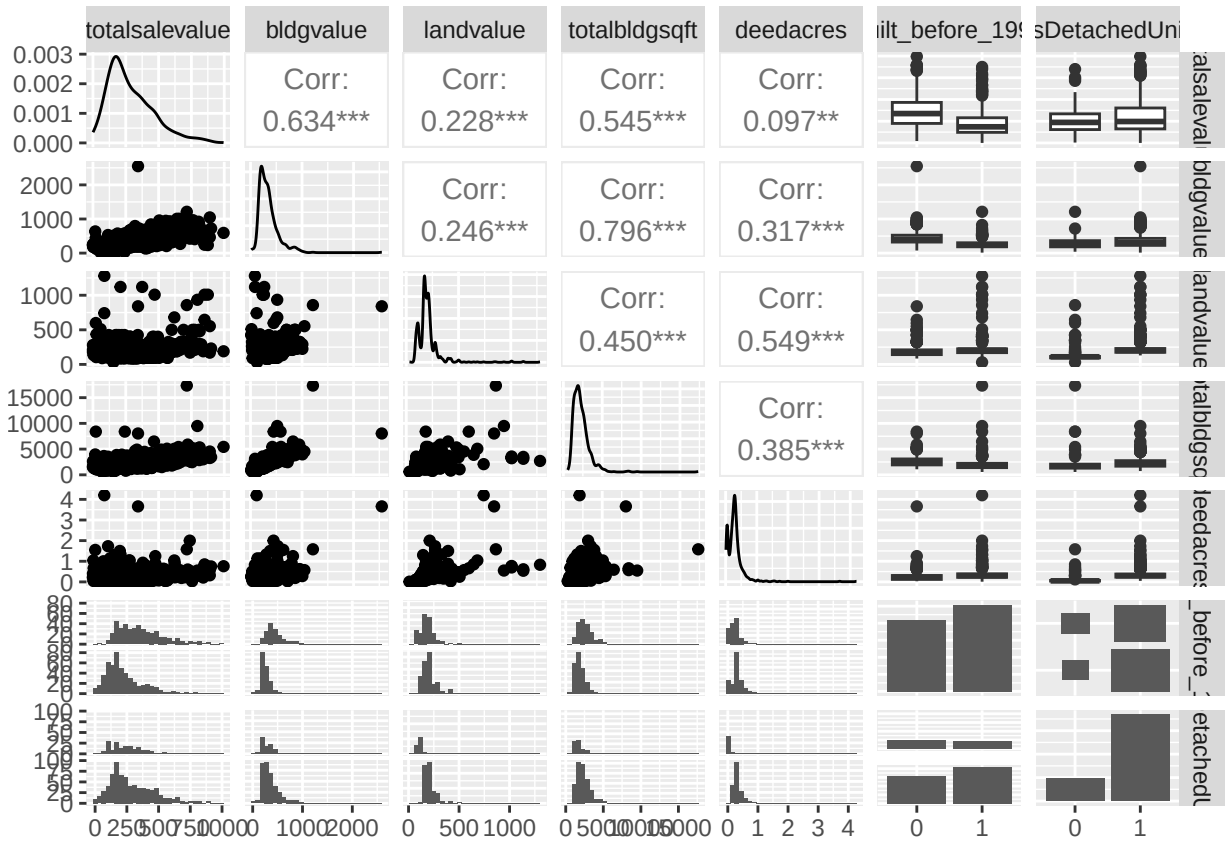
```
##   totalsalevalue bldgvalue landvalue totalbldgsqft deedacres yearbuilt
## 1         138000  329201.0  175000.0          2243      0.19      1991
## 2         810000  493975.0  230000.0          2863      0.34      1990
```

```
## 3        441500  654248.0  170000.0          3866      0.28      2014
## 4        310000  225089.0  140000.0          1496      0.11      1997
## 5        255000  302514.0  195500.0          1651      0.16      1993
## 6        370000  240044.0  135000.0          1484      0.04      1987
##   apastructuredesc phycity
## 1   Detached Units    Cary
## 2   Detached Units    Cary
## 3   Detached Units    Cary
## 4   Detached Units    Cary
## 5   Detached Units    Cary
## 6       Townhouses    Cary
```

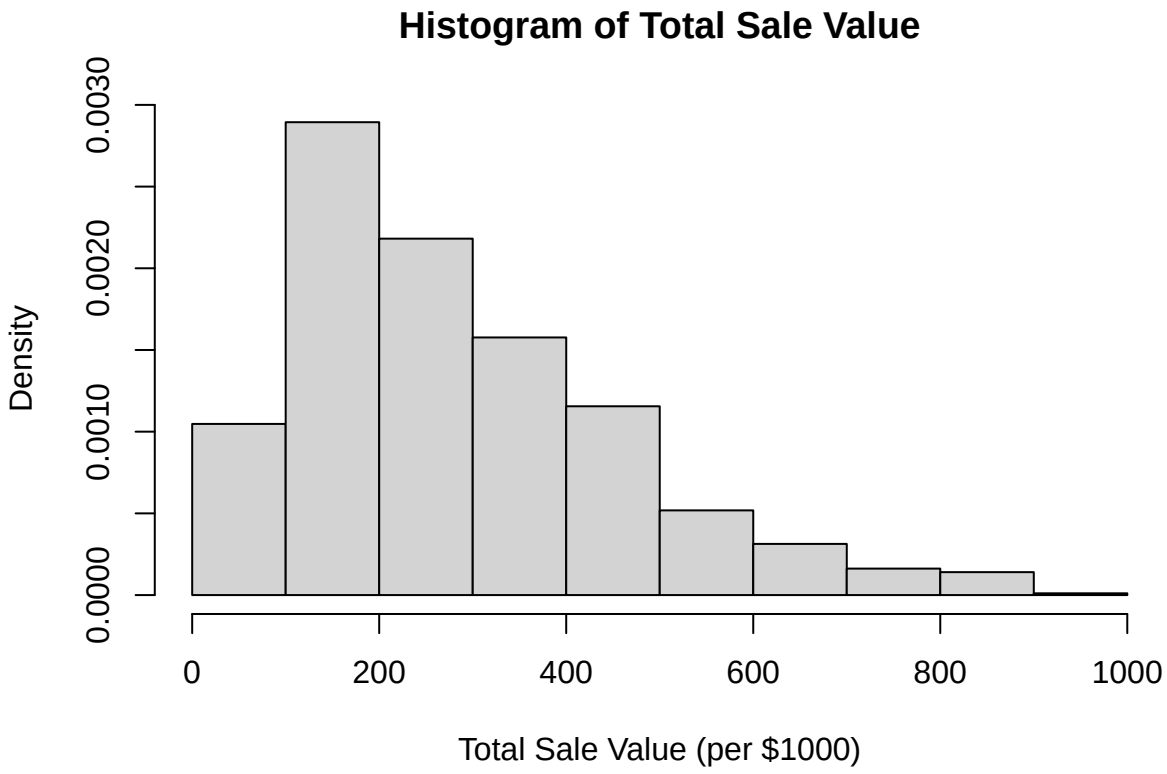
```
# Convert qualitative to quantitative data AND rename the variables we mutated
cleaned_data1 <- data %>% mutate(yearbuilt = if_else(yearbuilt <1990, "1", "0"),
                                apastructuredesc = if_else(apastructuredesc != "Detached Units", "0", "
                rename(built_before_1990 = yearbuilt,
                        IsDetachedUnit = apastructuredesc)

# Scale variables to thousands of dollars, convert data types, and exclude sale values below 1000
cleaned_data1 <- cleaned_data1 %>% mutate(totalsalevalue = as.double(totalsalevalue) / 1000,
                        bldgvalue = as.double(bldgvalue) / 1000,
                        landvalue = as.double(landvalue) / 1000,
                        totalbldgsqft = as.integer(totalbldgsqft),
                        deedacres = as.double(deedacres)) %>%
                  filter(totalsalevalue < 1000)

cleaned_data <- cleaned_data1 %>% dplyr::select(totalsalevalue, bldgvalue, landvalue,
                        totalbldgsqft, deedacres, built_before_1990,
                        IsDetachedUnit)

#cleaned_data
```

```
ggpairs(cleaned_data)
```

```
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```




```
hist(cleaned_data$totalsalevalue, probability = TRUE, main = "Histogram of Total Sale Value", xlab = "To
```



Response variable: totalsalevalue has a right skewed distribution

---

##Model Selection

```
# Linear model for all predictors
fit <- lm(totalsalevalue ~ . , data = cleaned_data)
summary(fit)
```

```
##
## Call:
## lm(formula = totalsalevalue ~ ., data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -896.12  -84.33  -20.80   64.30  603.37
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.310e+01  1.579e+01   4.631 4.17e-06 ***
## bldgvalue            5.030e-01  4.502e-02  11.173  < 2e-16 ***
## landvalue            2.583e-01  5.135e-02   5.029 5.93e-07 ***
## totalbldgsqft        1.434e-02  7.327e-03   1.957   0.0506 .
## deedacres           -1.253e+02  2.091e+01  -5.992 2.98e-09 ***
## built_before_19901 -2.758e+01  1.126e+01  -2.450   0.0145 *
## IsDetachedUnit1      1.078e+01  1.185e+01   0.909   0.3634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131.6 on 919 degrees of freedom
## Multiple R-squared:  0.442,  Adjusted R-squared:  0.4383
## F-statistic: 121.3 on 6 and 919 DF,  p-value: < 2.2e-16
```

Explanation of model selection procedure

The H0 for the F-test is that all of the parameters are zero, and since the P value is very low, we reject the null hypothesis and conclude that at least one of the parameters is nonzero.

The t-test for Totalbldgsqft, yearbuilt, Indivibillclass, IsDetachedUnit1, and phycityCARY are detached are nonsignificant when significance = 0.05.

Generally, you would expect that the more area a building has, the more it would cost. The data here says otherwise for acres but not bldgsqft.

Multicollinarity: Variance Inflation Factor

```
vif(fit)
```

```
##         bldgvalue         landvalue     totalbldgsqft         deedacres
##          3.841729          1.752010          3.432226          1.745862
## built_before_1990    IsDetachedUnit
##          1.678220          1.220020
```

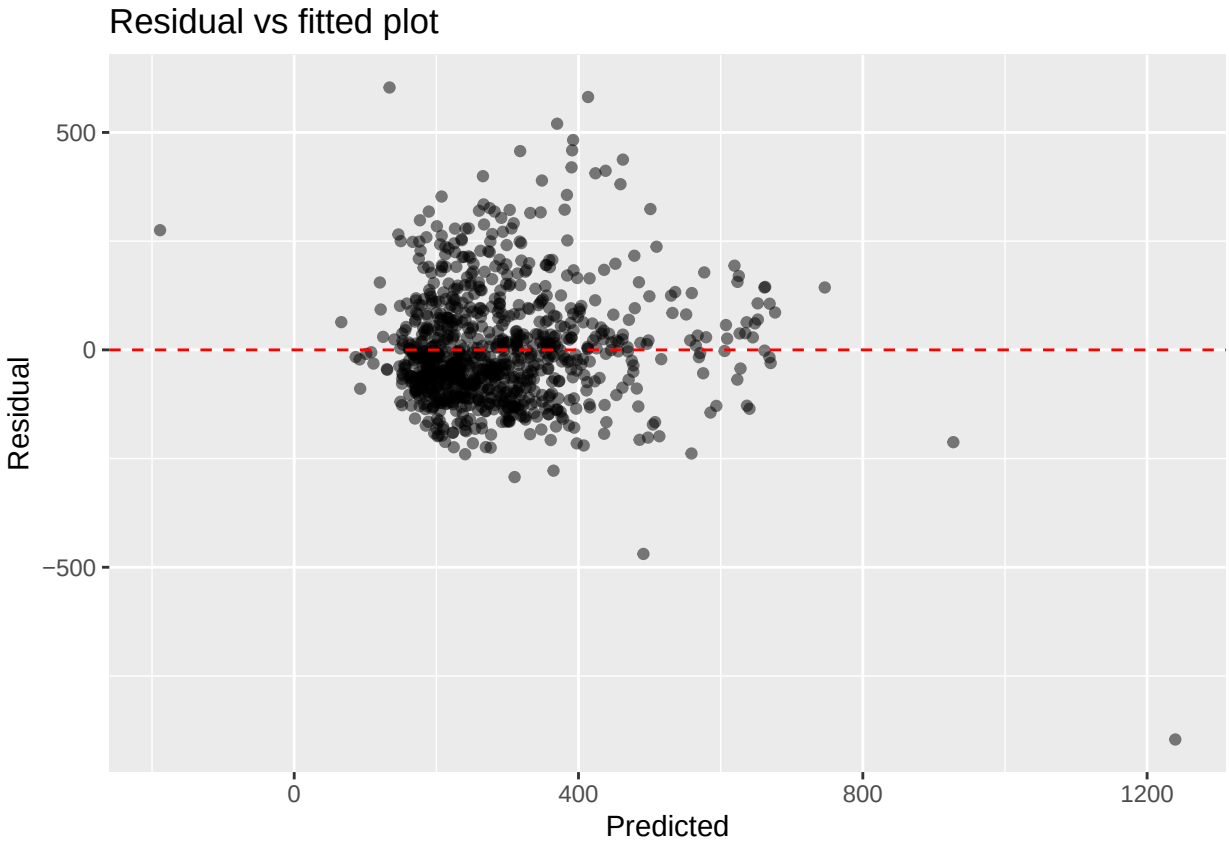
Since there are no parameters greater than 10, we can conclude that there is no major multicollinearity.

##Residual Diagnostic on Fit

Linearity assumption and constant variance assumption: Residual vs fitted

```r
# Residuals vs Fitted Plot of the fitted model
ggplot(data = fit, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title= "Residual vs fitted plot", x = "Predicted", y = "Residual")
```

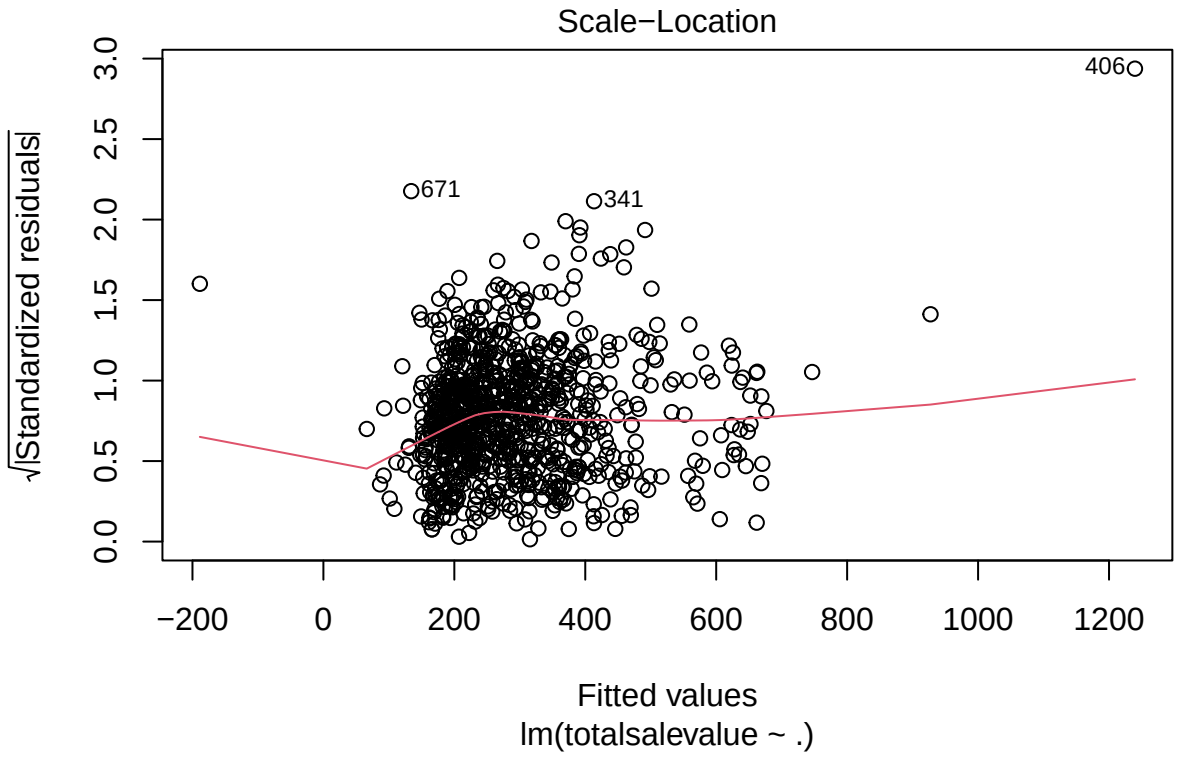
```
## Warning: `fortify(<lm>)` was deprecated in ggplot2 3.6.0.
## i Please use `broom::augment(<lm>)` instead.
## i The deprecated feature was likely used in the ggplot2 package.
##   Please report the issue at <https://github.com/tidyverse/ggplot2/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



The plot of residuals vs. predicted shows a slight pattern. The linearity condition is not satisfied. The vertical spread of the residuals is nearly constant across the plot but shows a slight negative slope after predicted 400. The constant variance condition is not satisfied.

Residual vs Fitted

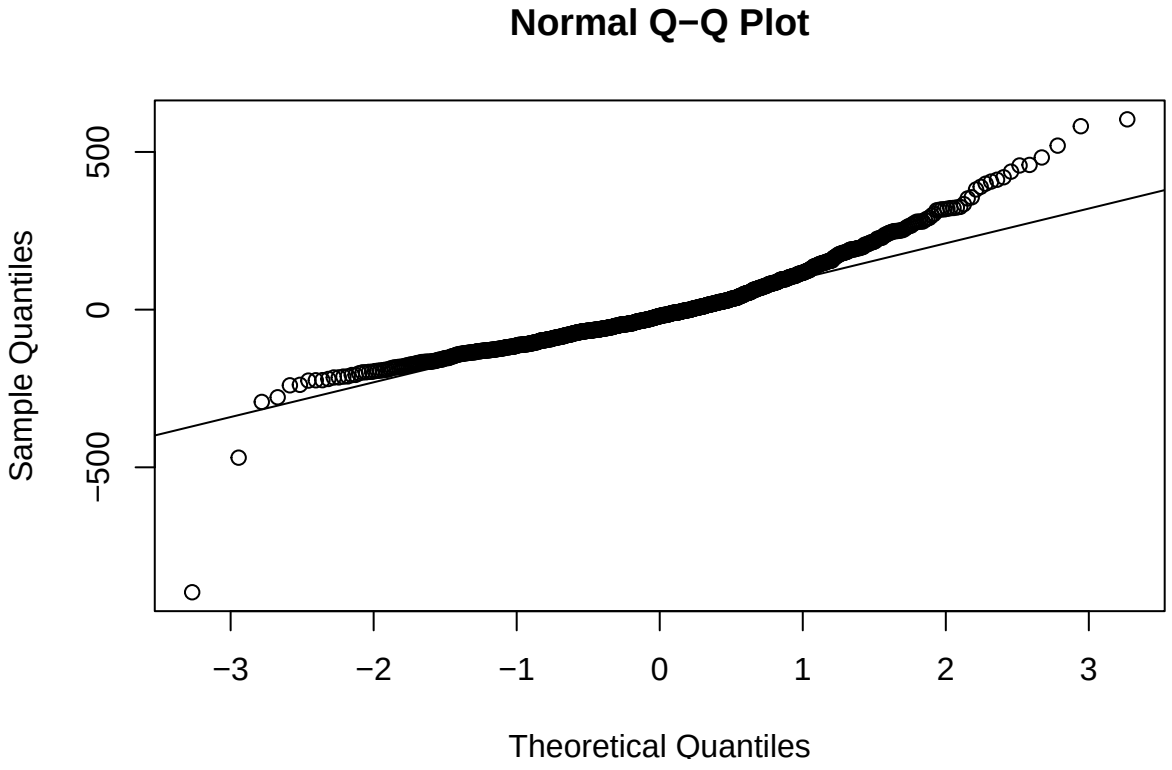
```r
# (which = 1): Looks for non-linearity (residuals vs fitted)
# (which = 2): Looks at normality of residual (QQ Plot)
# (which = 3): Looks at homoscedasticity (constant variance of residuals).
plot(fit, which = 3)
```



INTERPRETATION: The red curve has a slight upward trend on the right side of the curve. The variance isn't entirely constant especially as we approach larger fitted values. This slight pattern indicates heteroscedasticity (violating the assumption of constant variance). This means that transformations are necessary.

Normality Test: Q-Q Plot

```
qqnorm(resid(fit))
qqline(resid(fit)) # Reference line to compare to
```



The residuals follow the reference line closely for the middle data points. However, the data points on the left and right tails deviate largely from the reference line. This indicates a violation in the normality assumption.

Normality Test: Shapiro Test

```
shapiro.test(resid(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit)
## W = 0.93298, p-value < 2.2e-16
```

The Shapiro-Wilk test has a p-value of 2.2e-16. This, along with the histogram and QQ plot suggest that residuals appear to be non-normal.

Statistical test on Independence Assumption

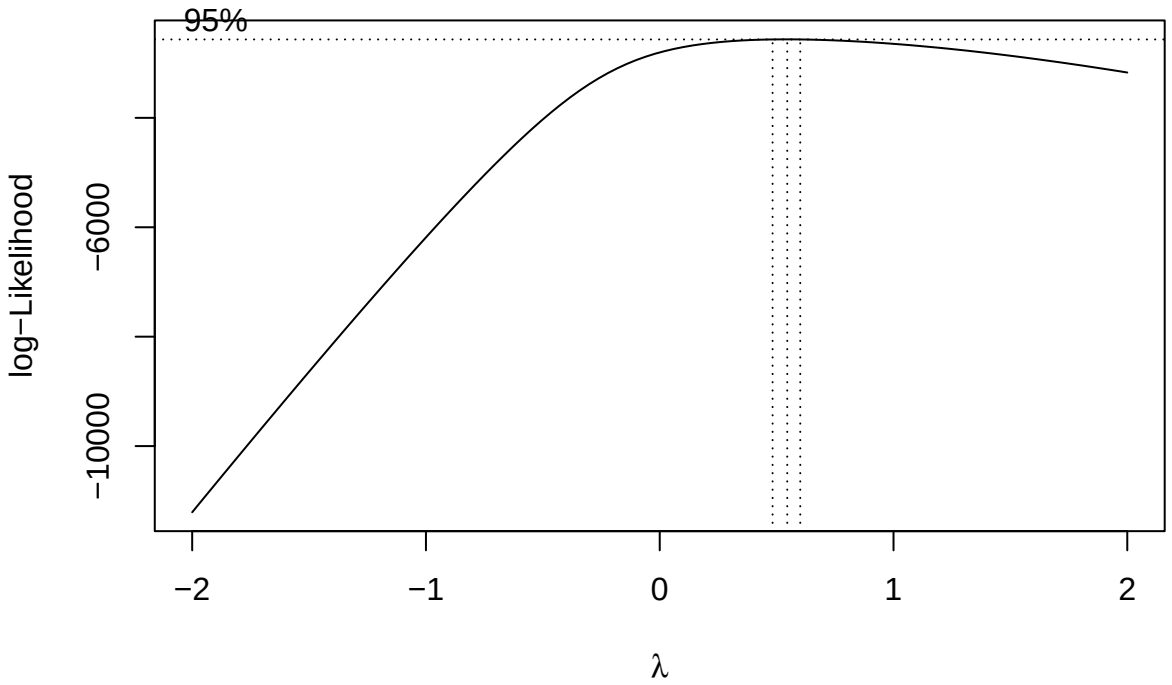
```
dwt(fit)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       -0.076292      2.150722   0.022
##  Alternative hypothesis: rho != 0
```

Since the p-value for the Durbin-Watson test is .758 (greater than .05), we do not reject the null hypothesis, and cannot conclude that the residuals in this regression model are autocorrelated.

##Transformation

```
bc <- boxcox(fit)
```




```
bc.power <- bc$x[which.max(bc$y)] # Get the power value
bc.power
```

```
## [1] 0.5454545
```

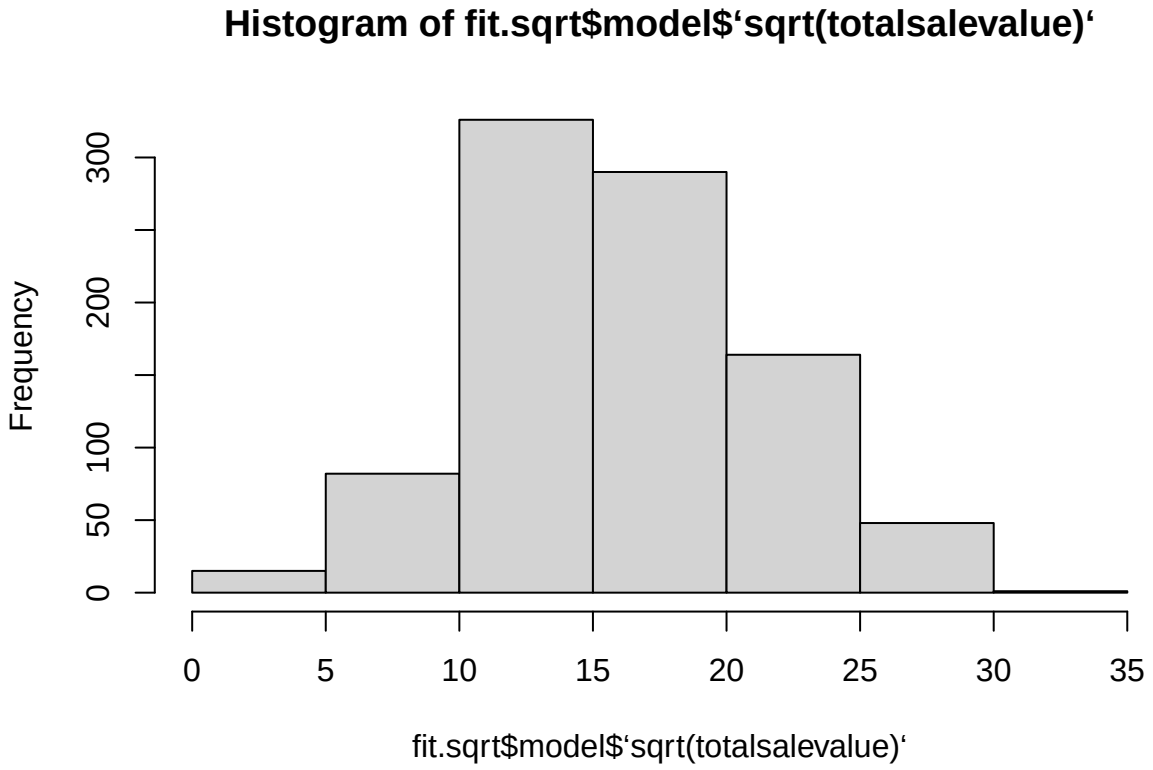
The optimal Box-Cox transformation suggests a power of 0.3434343, indicating a preference for a square root transformation.

```
# New fit model with a square root transformation applied to the dependent variable
fit.sqrt <- lm(sqrt(totalsalevalue) ~., data=cleaned_data)
summary(fit.sqrt)
```

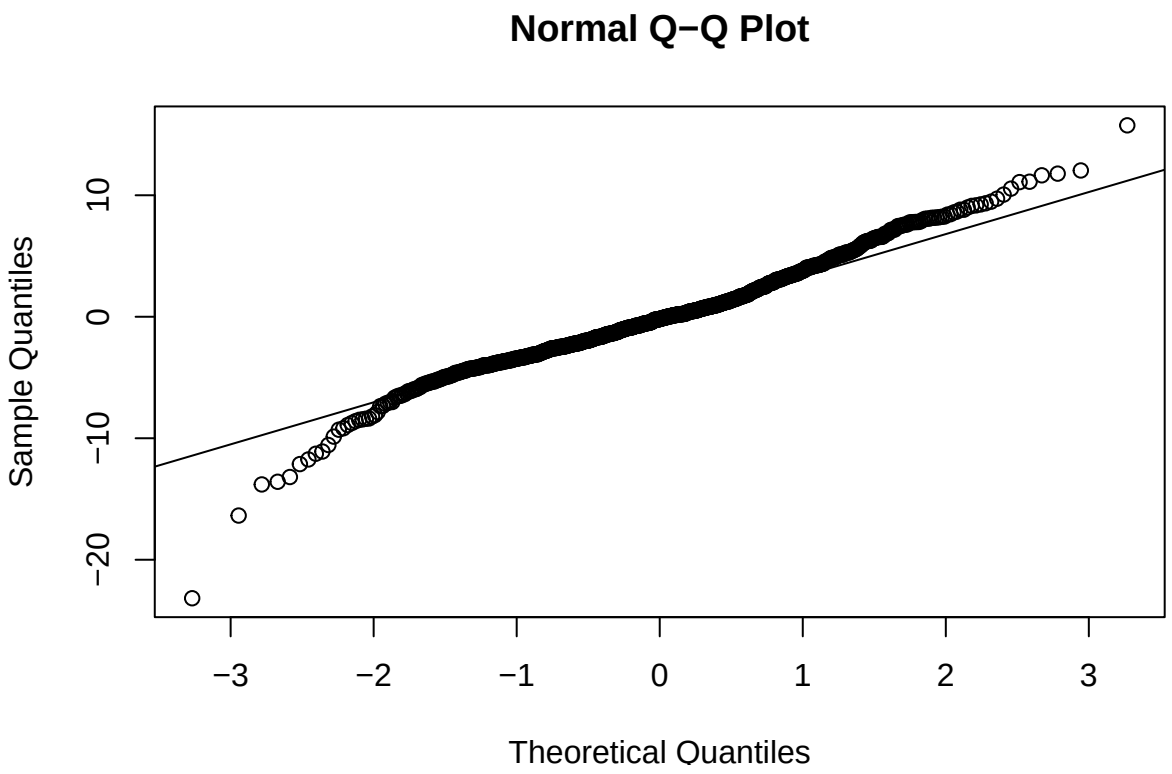
```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ ., data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1683  -2.4520  -0.1789   2.2173  15.7555
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.6736329  0.4779615  22.332  < 2e-16 ***
## bldgvalue           0.0142611  0.0013631  10.462  < 2e-16 ***
## landvalue           0.0056782  0.0015548   3.652 0.000275 ***
## totalbldgsqft       0.0003553  0.0002218   1.601 0.109621
## deedacres          -3.6483904  0.6329671  -5.764 1.12e-08 ***
## built_before_19901 -1.1643695  0.3408519  -3.416 0.000663 ***
## IsDetachedUnit1     0.4106252  0.3588982   1.144 0.252869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.985 on 919 degrees of freedom
## Multiple R-squared:  0.4126, Adjusted R-squared:  0.4088
```

```
## F-statistic: 107.6 on 6 and 919 DF,  p-value: < 2.2e-16
```

```
# Distribution of the square root values of totalsalevalue
hist(fit.sqrt$model$`sqrt(totalsalevalue)`)
```

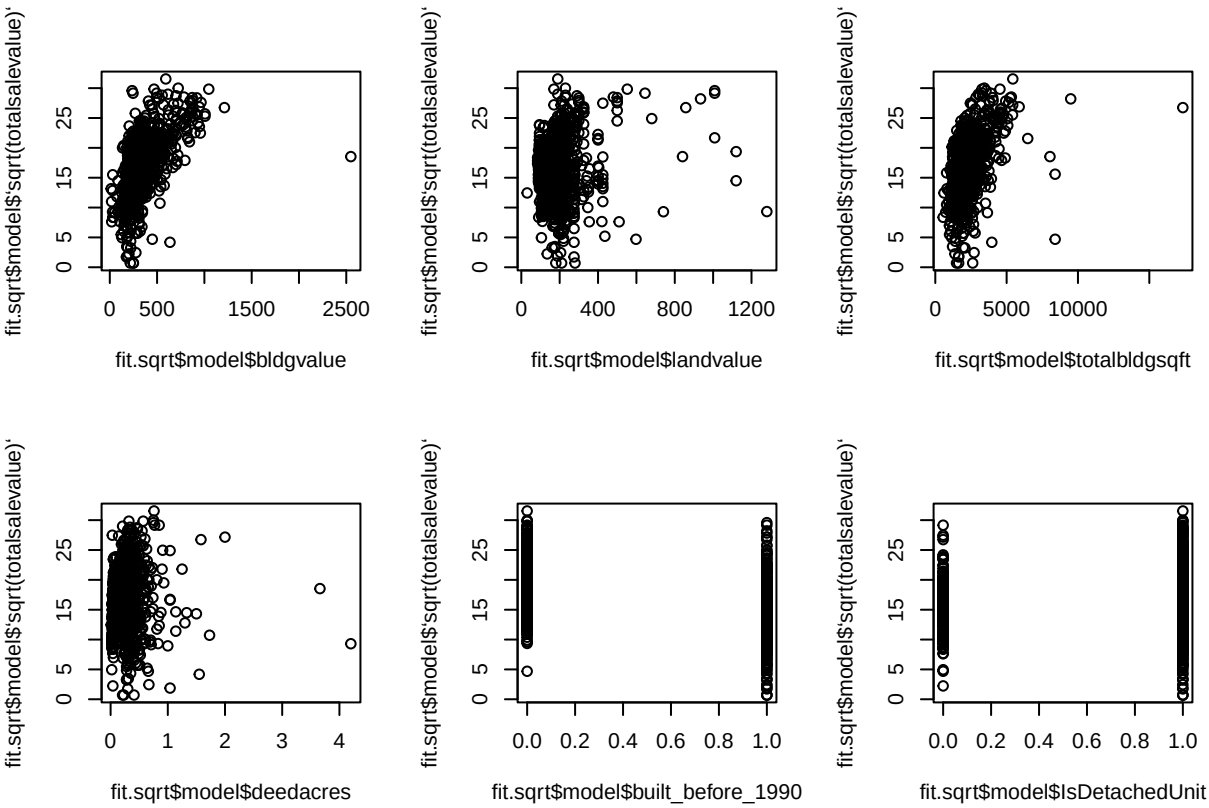



```
qqnorm(resid(fit.sqrt))
qqline(resid(fit.sqrt))
```



The residuals follow the reference line closely for the middle data points. However, the data points on the left and right tails deviate from the reference line. This indicates a violation in the normality assumption.

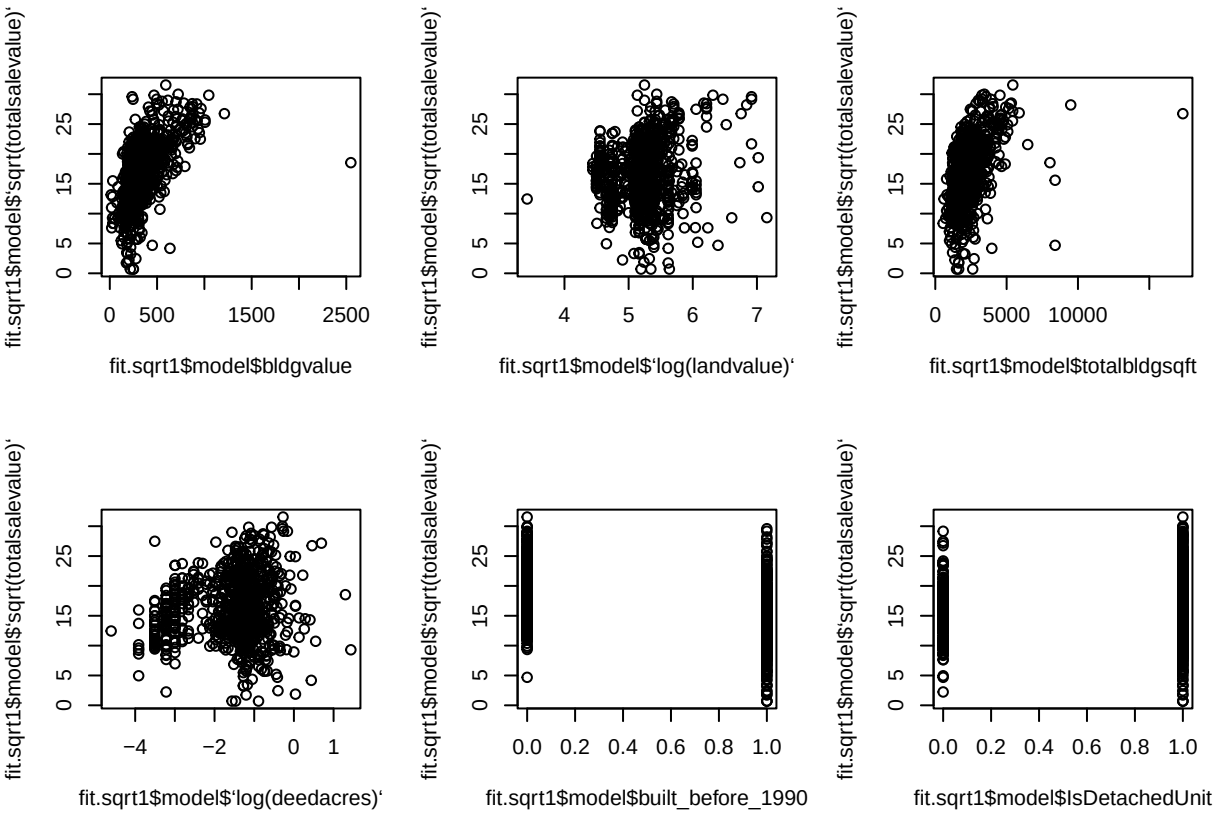
```
# Plots comparing EACH predictor variable with the square root (transformed) dependent variable (totals

par(mfrow=c(2,3))

plot(fit.sqrt$model$bldgvalue, fit.sqrt$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt$model$landvalue, fit.sqrt$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt$model$totalbldgsqft, fit.sqrt$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt$model$deedacres, fit.sqrt$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt$model$built_before_1990, fit.sqrt$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt$model$IsDetachedUnit, fit.sqrt$model$`sqrt(totalsalevalue)`)
```



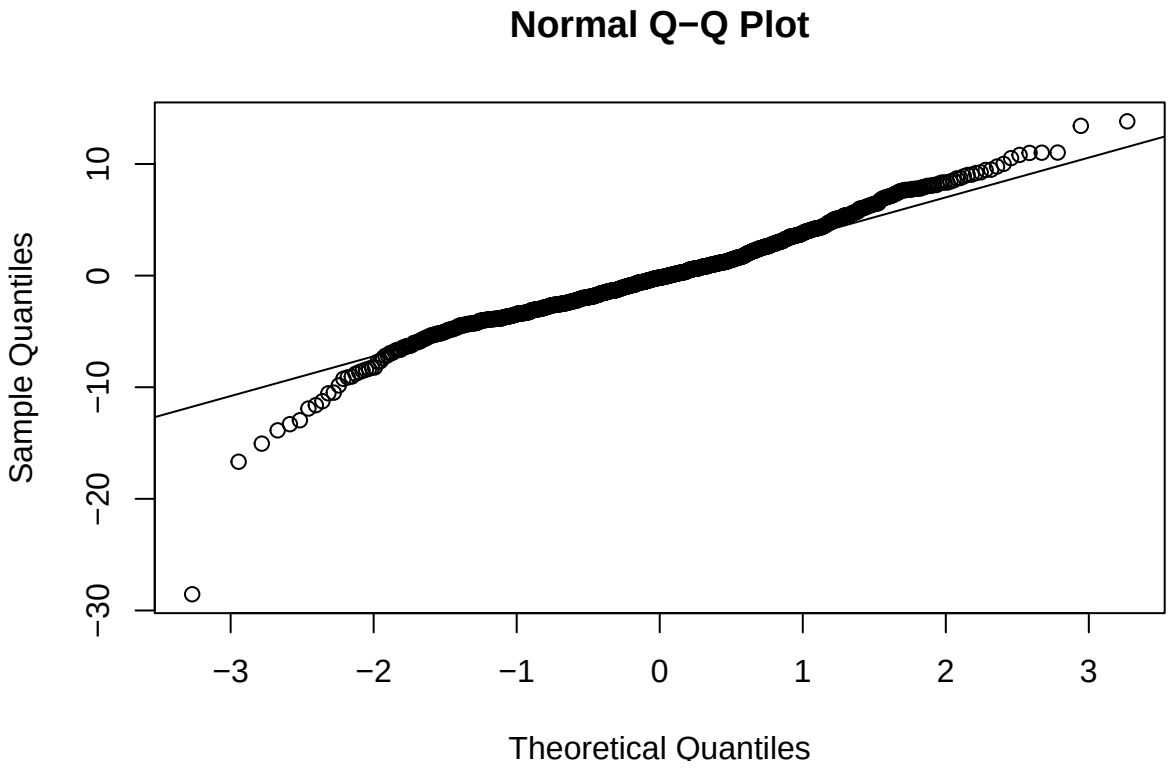

```
# UPDATED fit model with log transformations applied to landvalue and deedacres.
fit.sqrt1 <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft + log(deedacres) + bu
summary(fit.sqrt)
```

```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ ., data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1683  -2.4520  -0.1789   2.2173  15.7555
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         10.6736329  0.4779615  22.332  < 2e-16 ***
## bldgvalue            0.0142611  0.0013631  10.462  < 2e-16 ***
## landvalue            0.0056782  0.0015548   3.652 0.000275 ***
## totalbldgsqft        0.0003553  0.0002218   1.601 0.109621
## deedacres           -3.6483904  0.6329671  -5.764 1.12e-08 ***
## built_before_19901 -1.1643695  0.3408519  -3.416 0.000663 ***
## IsDetachedUnit1      0.4106252  0.3588982   1.144 0.252869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.985 on 919 degrees of freedom
## Multiple R-squared:  0.4126, Adjusted R-squared:  0.4088
## F-statistic: 107.6 on 6 and 919 DF,  p-value: < 2.2e-16
```

```r
# Plots comparing EACH predictor variable with the square root (transformed) dependent variable (totals
# This time, for the fit.sqrt1 model

par(mfrow=c(2,3))

plot(fit.sqrt1$model$bldgvalue, fit.sqrt1$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt1$model$`log(landvalue)`, fit.sqrt1$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt1$model$totalbldgsqft, fit.sqrt1$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt1$model$`log(deedacres)`, fit.sqrt1$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt1$model$built_before_1990, fit.sqrt1$model$`sqrt(totalsalevalue)`)

plot(fit.sqrt1$model$IsDetachedUnit, fit.sqrt1$model$`sqrt(totalsalevalue)`)
```




```r
qqnorm(resid(fit.sqrt1))
qqline(resid(fit.sqrt1))
```



The residuals follow the reference line closely for the middle data points. However, the data points on the left and right tails deviate from the reference line. This indicates a violation in the normality assumption.

```
shapiro.test(resid(fit.sqrt1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit.sqrt1)
## W = 0.9679, p-value = 2.043e-13
```

The Shapiro-Wilk test has a p-value of 8.99e-10. This, along with the QQ plot above suggest that residuals appear to be non-normal.

##Building a Step wise Regression Model:

Use a step wise regression to add predictors to the model one by one until no additional benefit is seen.

```
# specify a null model w/ no predictors
null_model <- lm(totalsalevalue ~ 1, data = cleaned_data)

# specify the full model using all of the potential predictors
full_model <- lm(totalsalevalue ~ ., data = cleaned_data)

# Use a step wise algorithm to build a parsimonious model
step_model1 <- step(null_model, scope = list(lower = null_model, upper = full_model), direction = "both
```

```
## Start:  AIC=9572.86
## totalsalevalue ~ 1
##
##                    Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## + bldgvalue         1  11451118 17081932 9099.8 619.4166 < 2.2e-16 ***
```

```
## + totalbldgsqft      1   8475497 20057553 9248.5 390.4444 < 2.2e-16 ***
## + built_before_1990  1   4393698 24139352 9420.0 168.1809 < 2.2e-16 ***
## + landvalue          1   1477219 27055832 9525.6  50.4494  2.44e-12 ***
## + deedacres          1    269038 28264013 9566.1   8.7953  0.003098 **
## + IsDetachedUnit     1    256555 28276495 9566.5   8.3835  0.003876 **
## <none>                           28533050 9572.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=9099.78
## totalsalevalue ~ bldgvalue
##
##                     Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## + deedacres          1    342378 16739555 9083.0  18.8783 1.548e-05 ***
## + built_before_1990  1    165921 16916012 9092.7   9.0532  0.002694 **
## + landvalue          1    155591 16926342 9093.3   8.4844  0.003668 **
## + totalbldgsqft      1    128896 16953037 9094.8   7.0177  0.008209 **
## <none>                           17081932 9099.8
## + IsDetachedUnit     1       581 17081352 9101.7   0.0314  0.859417
## - bldgvalue          1  11451118 28533050 9572.9 619.4166 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=9083.03
## totalsalevalue ~ bldgvalue + deedacres
##
##                     Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## + landvalue          1    654003 16085552 9048.1  37.4865 1.361e-09 ***
## + totalbldgsqft      1    257801 16481754 9070.7  14.4215 0.0001557 ***
## + IsDetachedUnit     1     38763 16700792 9082.9   2.1400 0.1438457
## <none>                           16739555 9083.0
## + built_before_1990  1     34520 16705035 9083.1   1.9053 0.1678253
## - deedacres          1    342378 17081932 9099.8  18.8783 1.548e-05 ***
## - bldgvalue          1  11524458 28264013 9566.1 635.4455 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=9048.13
## totalsalevalue ~ bldgvalue + deedacres + landvalue
##
##                     Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## + built_before_1990  1     88823 15996729 9045.0   5.1139   0.02397 *
## + totalbldgsqft      1     48313 16037239 9047.3   2.7746   0.09611 .
## <none>                           16085552 9048.1
## + IsDetachedUnit     1      5991 16079561 9049.8   0.3432   0.55816
## - landvalue          1    654003 16739555 9083.0  37.4865 1.361e-09 ***
## - deedacres          1    840790 16926342 9093.3  48.1928 7.281e-12 ***
## - bldgvalue          1  10938540 27024092 9526.5 626.9809 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=9045
## totalsalevalue ~ bldgvalue + deedacres + landvalue + built_before_1990
##
```

```
##                     Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## + totalbldgsqft      1     59976 15936753 9043.5   3.4623   0.06310 .
## <none>                           15996729 9045.0
## + IsDetachedUnit     1      7936 15988794 9046.5   0.4566   0.49938
## - built_before_1990  1     88823 16085552 9048.1   5.1139   0.02397 *
## - deedacres          1    616361 16613090 9078.0  35.4865 3.652e-09 ***
## - landvalue          1    708306 16705035 9083.1  40.7802 2.698e-10 ***
## - bldgvalue          1   5858475 21855204 9332.0 337.2974 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=9043.52
## totalsalevalue ~ bldgvalue + deedacres + landvalue + built_before_1990 +
##     totalbldgsqft
##
##                     Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## <none>                           15936753 9043.5
## + IsDetachedUnit     1     14327 15922426 9044.7   0.8269   0.36340
## - totalbldgsqft      1     59976 15996729 9045.0   3.4623   0.06310 .
## - built_before_1990  1    100486 16037239 9047.3   5.8009   0.01621 *
## - landvalue          1    481688 16418441 9069.1  27.8070 1.671e-07 ***
## - deedacres          1    611147 16547900 9076.4  35.2804 4.045e-09 ***
## - bldgvalue          1   2236535 18173288 9163.1 129.1111 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The 6 predictors kept from the piecewise regression using the null model (step_model1) are yearbuilt, landvalue, IsDetachedUnit, phycity, bldgvalue, and totalbldgsqft.

```
step_model2 <- step(full_model, scope = list(lower = null_model, upper = full_model), direction = "both
```

```
## Start:  AIC=9044.69
## totalsalevalue ~ bldgvalue + landvalue + totalbldgsqft + deedacres +
##     built_before_1990 + IsDetachedUnit
##
##                     Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## - IsDetachedUnit     1     14327 15936753 9043.5   0.8269   0.36340
## <none>                           15922426 9044.7
## - totalbldgsqft      1     66368 15988794 9046.5   3.8306   0.05063 .
## - built_before_1990  1    103997 16026423 9048.7   6.0024   0.01447 *
## - landvalue          1    438160 16360586 9067.8  25.2895 5.932e-07 ***
## - deedacres          1    622031 16544457 9078.2  35.9020 2.976e-09 ***
## - bldgvalue          1   2162732 18085158 9160.6 124.8271 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=9043.52
## totalsalevalue ~ bldgvalue + landvalue + totalbldgsqft + deedacres +
##     built_before_1990
##
##                     Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## <none>                           15936753 9043.5
## + IsDetachedUnit     1     14327 15922426 9044.7   0.8269   0.36340
## - totalbldgsqft      1     59976 15996729 9045.0   3.4623   0.06310 .
## - built_before_1990  1    100486 16037239 9047.3   5.8009   0.01621 *
```

```
## - landvalue          1    481688 16418441 9069.1  27.8070 1.671e-07 ***
## - deedacres          1    611147 16547900 9076.4  35.2804 4.045e-09 ***
## - bldgvalue          1   2236535 18173288 9163.1 129.1111 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(step_model1)
```

```
##
## Call:
## lm(formula = totalsalevalue ~ bldgvalue + deedacres + landvalue +
##     built_before_1990 + totalbldgsqft, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -919.31  -85.05  -19.13   62.94  598.16
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.861e+01  1.458e+01   5.392 8.86e-08 ***
## bldgvalue           5.079e-01  4.470e-02  11.363  < 2e-16 ***
## deedacres          -1.210e+02  2.037e+01  -5.940 4.04e-09 ***
## landvalue           2.665e-01  5.054e-02   5.273 1.67e-07 ***
## built_before_19901 -2.708e+01  1.124e+01  -2.408   0.0162 *
## totalbldgsqft       1.353e-02  7.272e-03   1.861   0.0631 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131.6 on 920 degrees of freedom
## Multiple R-squared:  0.4415, Adjusted R-squared:  0.4384
## F-statistic: 145.4 on 5 and 920 DF,  p-value: < 2.2e-16
```

```
summary(step_model2)
```

```
##
## Call:
## lm(formula = totalsalevalue ~ bldgvalue + landvalue + totalbldgsqft +
##     deedacres + built_before_1990, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -919.31  -85.05  -19.13   62.94  598.16
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.861e+01  1.458e+01   5.392 8.86e-08 ***
## bldgvalue           5.079e-01  4.470e-02  11.363  < 2e-16 ***
## landvalue           2.665e-01  5.054e-02   5.273 1.67e-07 ***
## totalbldgsqft       1.353e-02  7.272e-03   1.861   0.0631 .
## deedacres          -1.210e+02  2.037e+01  -5.940 4.04e-09 ***
## built_before_19901 -2.708e+01  1.124e+01  -2.408   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131.6 on 920 degrees of freedom
```

```
## Multiple R-squared:  0.4415, Adjusted R-squared:  0.4384
## F-statistic: 145.4 on 5 and 920 DF,  p-value: < 2.2e-16
```

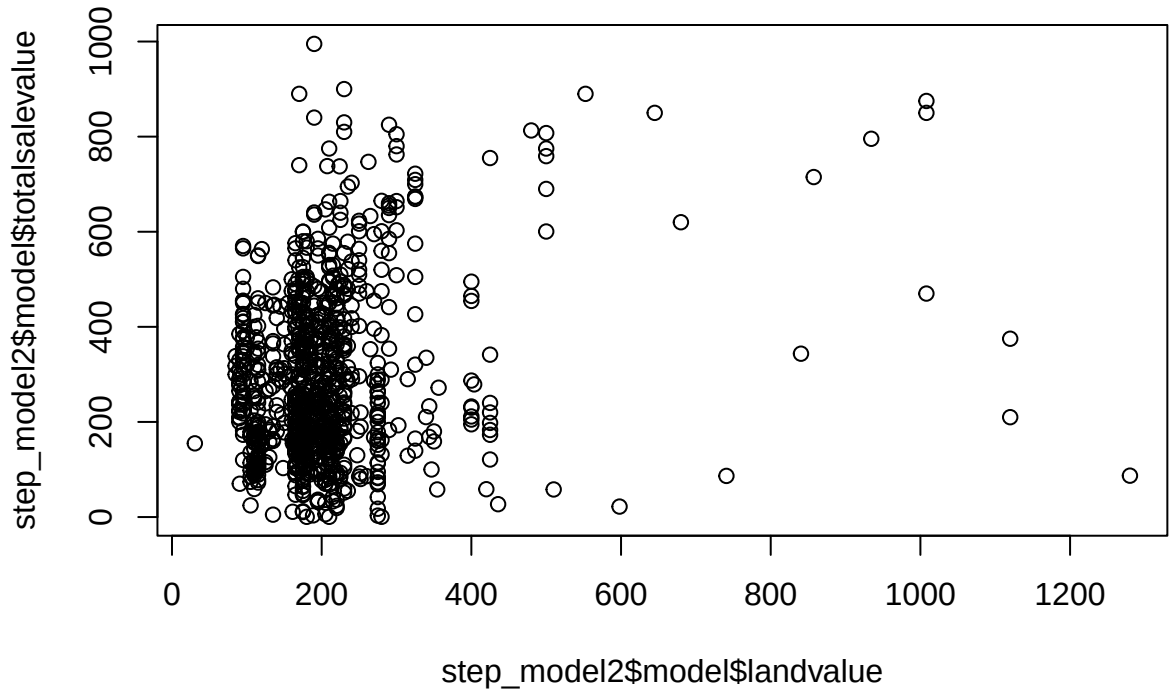
Starting with no predictors (step_model1) and using the step function yields the same results as starting with all predictors (step_model2), so we will proceed with the step_model2 since its variables are organized in a cleaner order.

Quantitative: - bldgvalue (Building Value) - landvalue (Land Value) - deedacres (land size in Acres in deed) Qualitative: - ResLessThan101 (If the land is residential and < 10 acres) - IndivBillClass1. (If billing class for taxes is individual)

```
step_model2
```

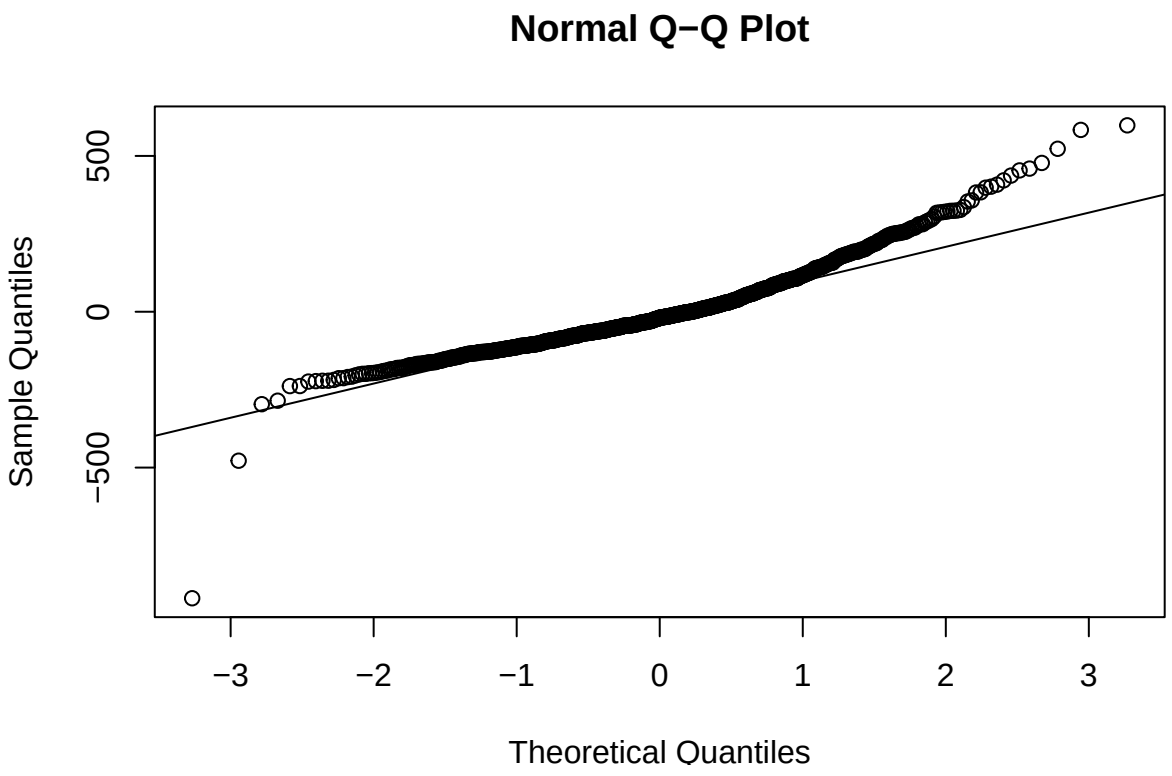
```
##
## Call:
## lm(formula = totalsalevalue ~ bldgvalue + landvalue + totalbldgsqft +
##     deedacres + built_before_1990, data = cleaned_data)
##
## Coefficients:
##        (Intercept)           bldgvalue           landvalue       totalbldgsqft
##           78.60865             0.50789             0.26651             0.01353
##          deedacres  built_before_19901
##         -121.00428           -27.07995
```

```
# To observe relationship between landvalue and totalsalevalue.
plot(step_model2$model$landvalue, step_model2$model$totalsalevalue)
```



There doesn't seem to be a clear trend.

```
qqnorm(resid(step_model2))
qqline(resid(step_model2))
```



The residuals follow the reference line closely for the middle data points. However, the data points on the left and right tails deviate from the reference line. This indicates a violation in the normality assumption.

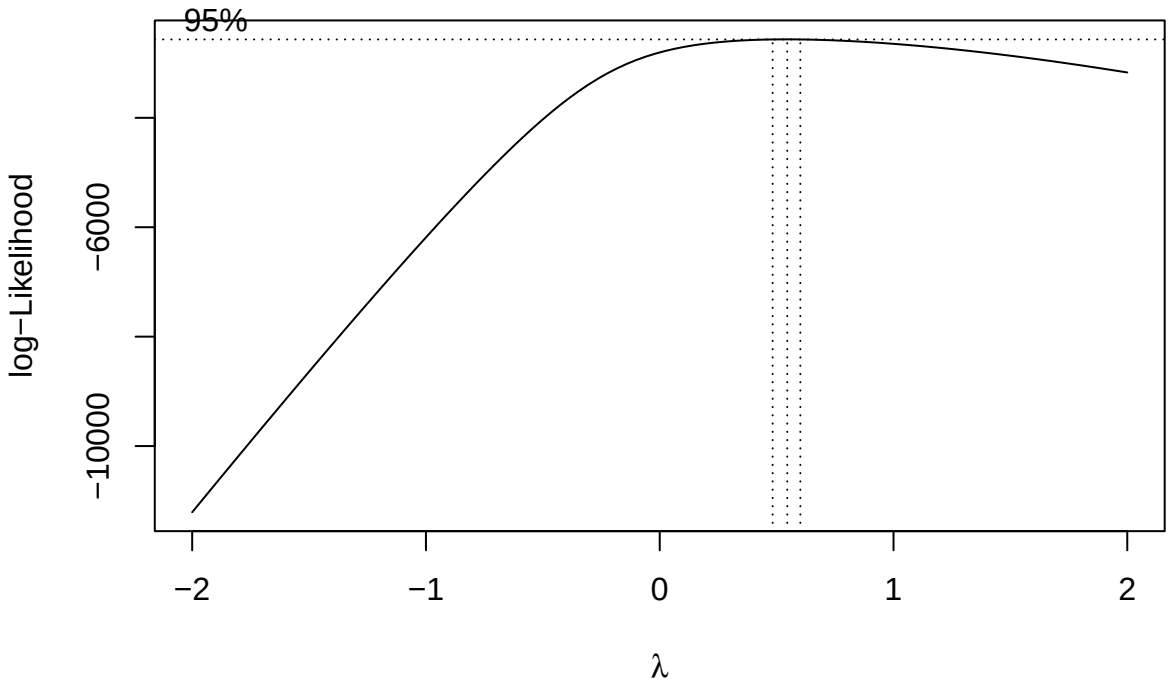
Multicollinarity: Variance Inflation Factor

```
vif(step_model2)
```

```
##         bldgvalue         landvalue     totalbldgsqft         deedacres
##          3.787292          1.697233          3.381650          1.658048
## built_before_1990
##          1.674181
```

Since there are no parameters greater than 10, we can conclude that there is no major multicollinearity.

```
bc <- boxcox(step_model2)
```




```
bc.power <- bc$x[which.max(bc$y)]
bc.power
```

```
## [1] 0.5454545
```

The optimal Box-Cox transformation suggests a power of 0.3434343, indicating a preference for a square root transformation.

```
# Output the fit & step_model2 models. Then make a square root transformation model of step_model2 & ou
fit
```

```
##
## Call:
## lm(formula = totalsalevalue ~ ., data = cleaned_data)
##
## Coefficients:
##        (Intercept)           bldgvalue           landvalue       totalbldgsqft
##           73.10440             0.50302             0.25826             0.01434
##          deedacres  built_before_19901     IsDetachedUnit1
##         -125.26802           -27.58219            10.77963
```

```
step_model2
```

```
##
## Call:
## lm(formula = totalsalevalue ~ bldgvalue + landvalue + totalbldgsqft +
##     deedacres + built_before_1990, data = cleaned_data)
##
## Coefficients:
##        (Intercept)           bldgvalue           landvalue       totalbldgsqft
##           78.60865             0.50789             0.26651             0.01353
##          deedacres  built_before_19901
##         -121.00428           -27.07995
```

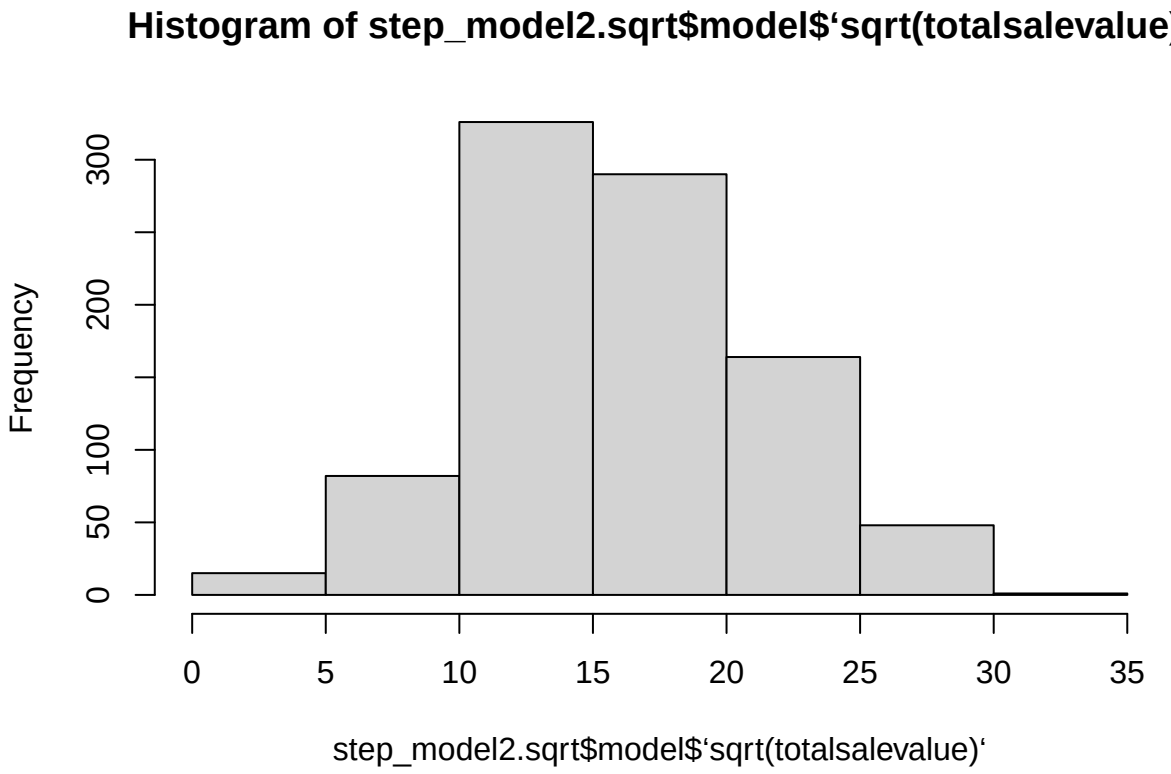
```r
step_model2.sqrt <- lm(sqrt(totalsalevalue) ~., data = step_model2$model)
summary(step_model2.sqrt)
```

```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ ., data = step_model2$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.0516  -2.4827  -0.1565   2.2831  15.5569
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         10.8833045  0.4415074  24.650  < 2e-16 ***
## bldgvalue            0.0144468  0.0013536  10.672  < 2e-16 ***
## landvalue            0.0059928  0.0015306   3.915 9.69e-05 ***
## totalbldgsqft        0.0003245  0.0002202   1.473 0.141036
## deedacres           -3.4859729  0.6169466  -5.650 2.13e-08 ***
## built_before_19901 -1.1452379  0.3404987  -3.363 0.000802 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.986 on 920 degrees of freedom
## Multiple R-squared:  0.4118, Adjusted R-squared:  0.4086
## F-statistic: 128.8 on 5 and 920 DF,  p-value: < 2.2e-16
```

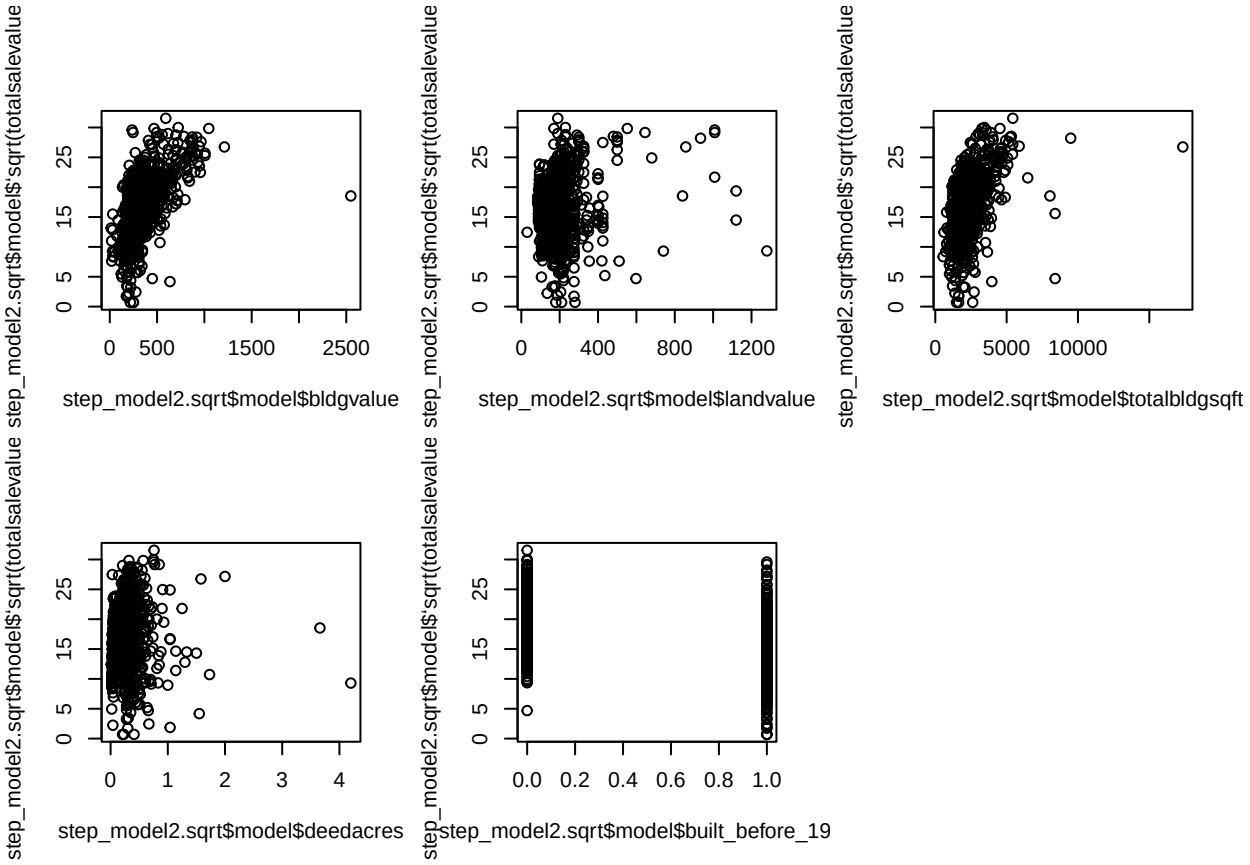
```r
summary(step_model2)
```

```
##
## Call:
## lm(formula = totalsalevalue ~ bldgvalue + landvalue + totalbldgsqft +
##     deedacres + built_before_1990, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -919.31  -85.05  -19.13   62.94  598.16
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.861e+01  1.458e+01   5.392 8.86e-08 ***
## bldgvalue            5.079e-01  4.470e-02  11.363  < 2e-16 ***
## landvalue            2.665e-01  5.054e-02   5.273 1.67e-07 ***
## totalbldgsqft        1.353e-02  7.272e-03   1.861   0.0631 .
## deedacres           -1.210e+02  2.037e+01  -5.940 4.04e-09 ***
## built_before_19901 -2.708e+01  1.124e+01  -2.408   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131.6 on 920 degrees of freedom
## Multiple R-squared:  0.4415, Adjusted R-squared:  0.4384
## F-statistic: 145.4 on 5 and 920 DF,  p-value: < 2.2e-16
```

```r
# Look at distribution of square root transformed values of totalsalevalue.
hist(step_model2.sqrt$model$`sqrt(totalsalevalue)`)
```



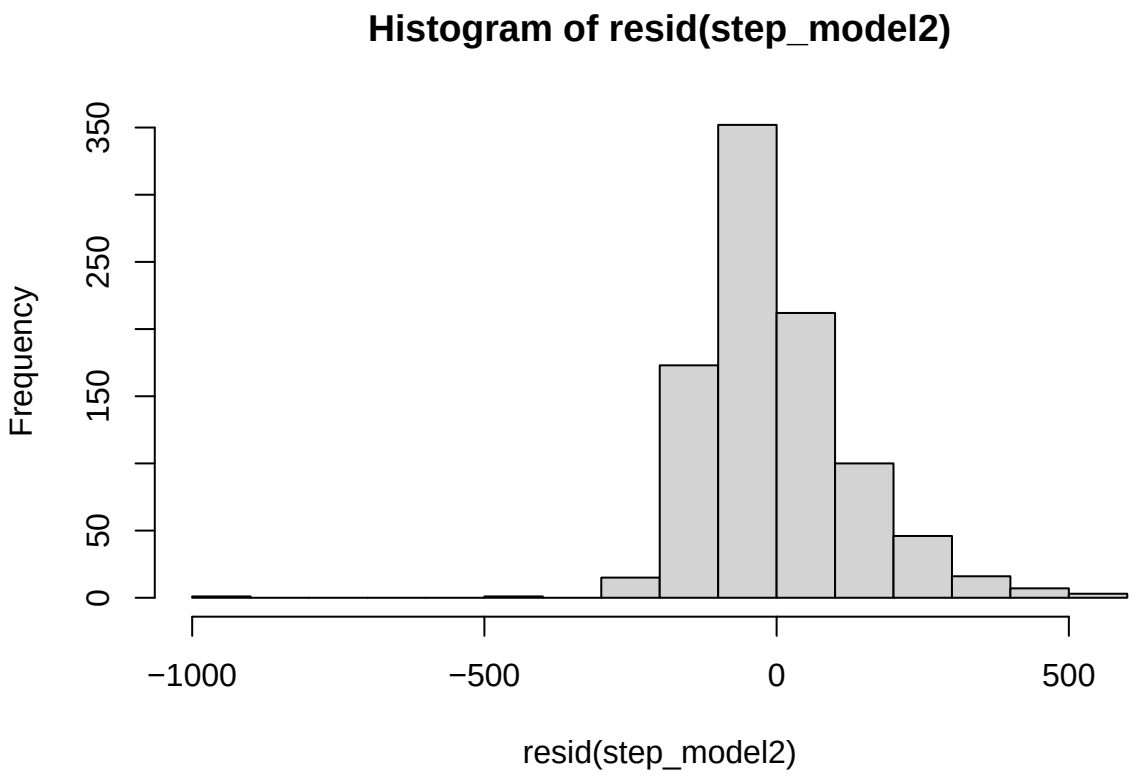

```
par(mfrow=c(2,3))
plot(step_model2.sqrt$model$bldgvalue, step_model2.sqrt$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt$model$landvalue, step_model2.sqrt$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt$model$totalbldgsqft, step_model2.sqrt$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt$model$deedacres, step_model2.sqrt$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt$model$built_before_1990, step_model2.sqrt$model$`sqrt(totalsalevalue)`)
```



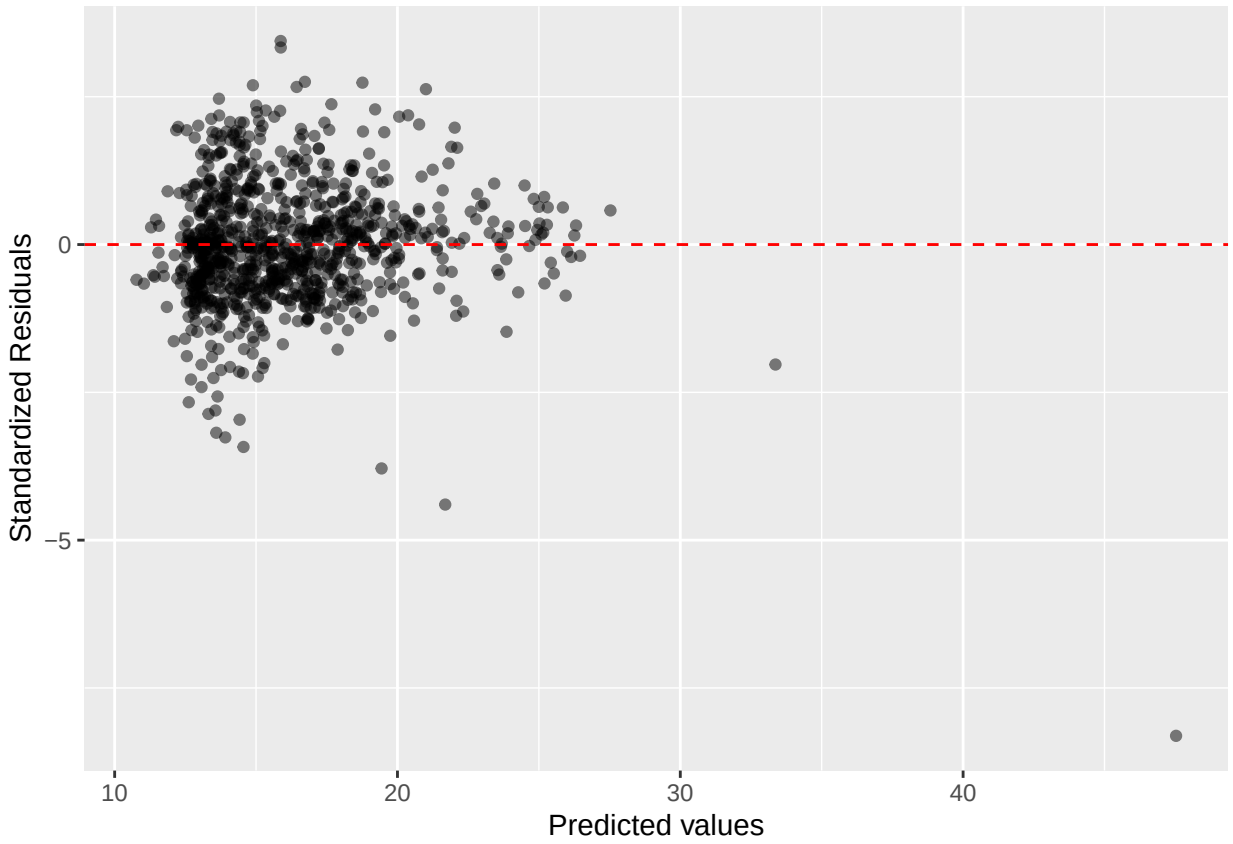

```
step_model2.sqrt1 <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft+ log(deedacres
summary(step_model2.sqrt1)
```

```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) +
##     totalbldgsqft + log(deedacres) + built_before_1990, data = step_model2$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.9970  -2.5161  -0.1859   2.2593  13.7090
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.8183946  3.0629635   1.900   0.0578 .
## bldgvalue            0.0125195  0.0013384   9.354  < 2e-16 ***
## log(landvalue)       1.0327485  0.5484328   1.883   0.0600 .
## totalbldgsqft        0.0004152  0.0002176   1.908   0.0566 .
## log(deedacres)      -0.3732355  0.2380919  -1.568   0.1173
## built_before_19901 -1.6303420  0.3434232  -4.747 2.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.052 on 920 degrees of freedom
## Multiple R-squared:  0.392,  Adjusted R-squared:  0.3887
## F-statistic: 118.6 on 5 and 920 DF,  p-value: < 2.2e-16
```

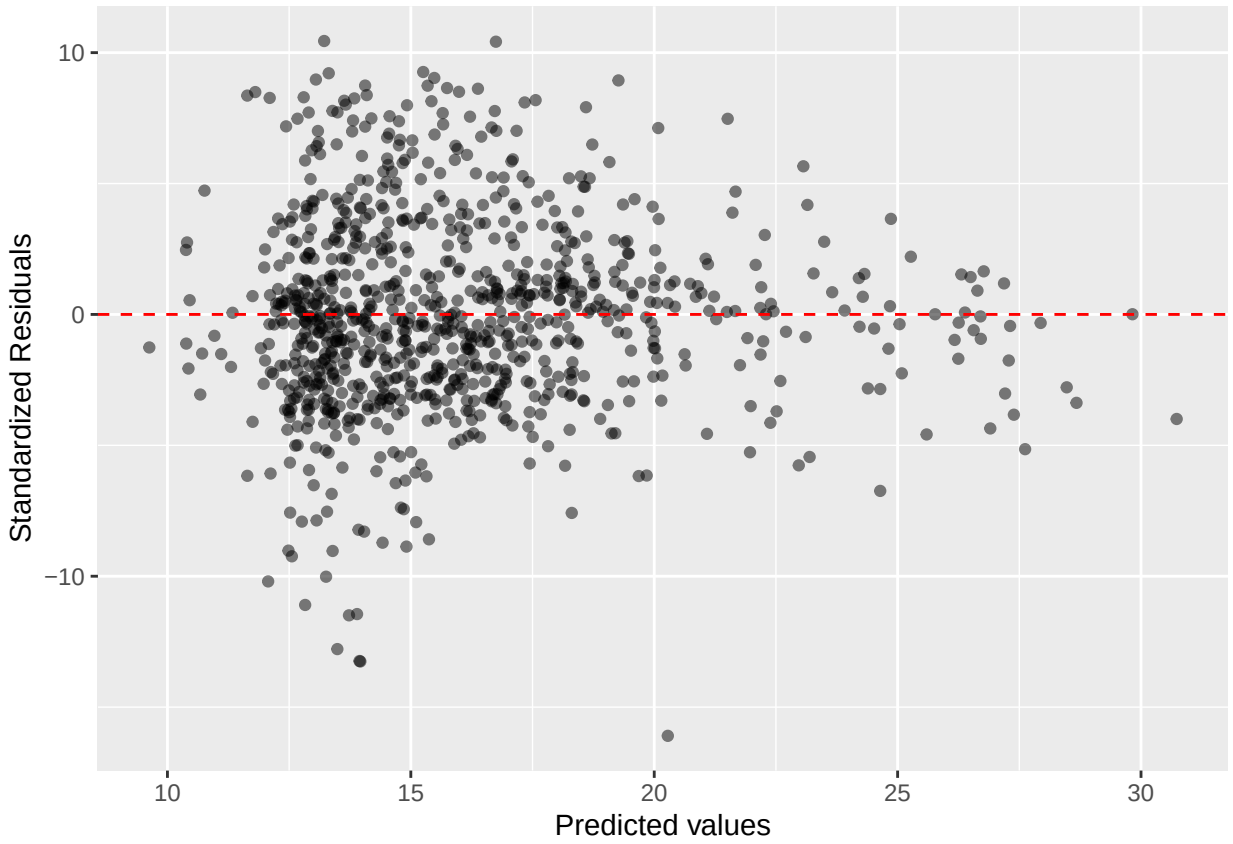
```r
hist(resid(step_model2))
```




```r
step_model2.sqrt1$.stdresid <- rstandard(step_model2.sqrt1)

ggplot(data =step_model2.sqrt1, aes(x = .fitted, y = .stdresid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Standardized Residuals")
```

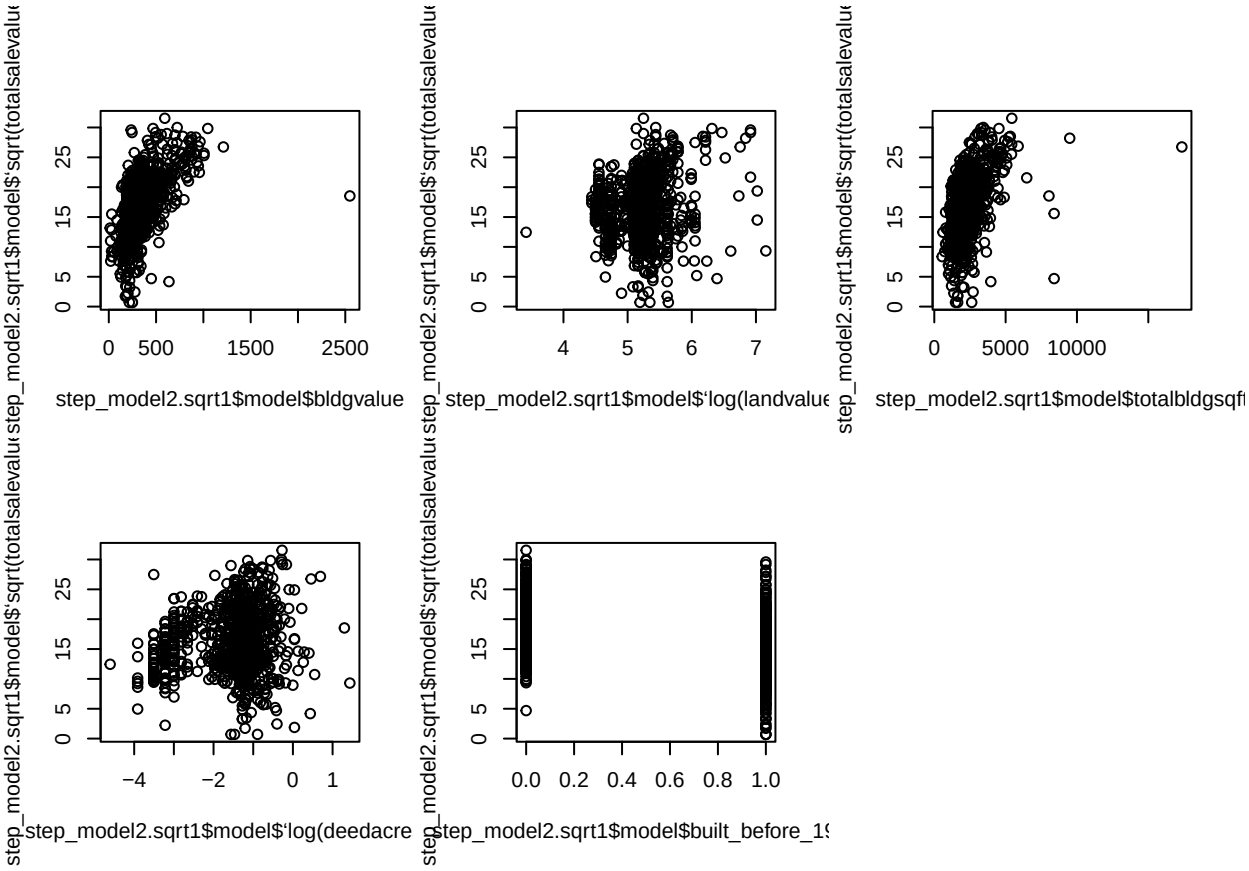



```
step_model2.test <- step_model2
step_model2.test$model$.stdresid <- rstandard(step_model2.test)

step_model2.test$model <- step_model2.test$model %>% filter(.stdresid <= 3 & .stdresid >= -3)
# step_model2.test$model # To check data

step_model2.test <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft+ log(deedacres)

step_model2.test$.stdresid <- rstandard(step_model2.test)

ggplot(data =step_model2.test, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Standardized Residuals")
```



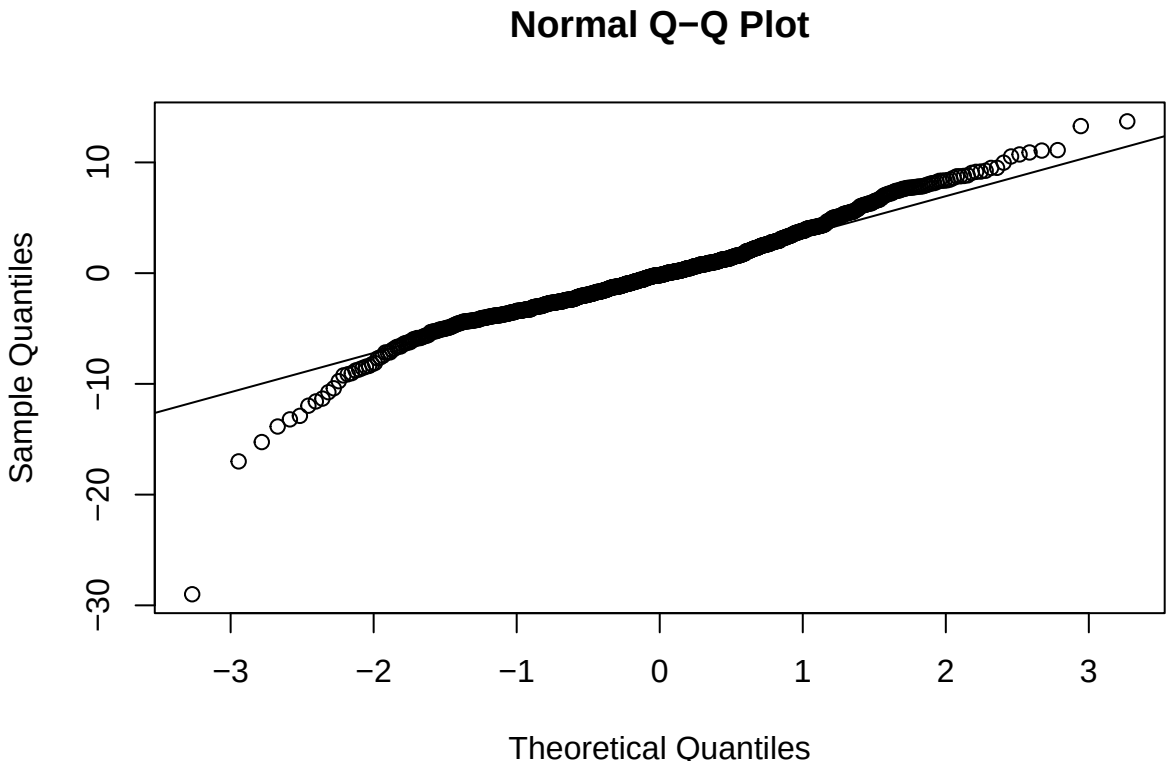
We can quickly check that our R-squared is 0.4051 and adjusted r-squared is 0.4001. This model does not seem too bad. However, 40.01% is not too high so we should look into higher-order and interaction terms for our next models.

```
par(mfrow=c(2,3))
plot(step_model2.sqrt1$model$bldgvalue, step_model2.sqrt1$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt1$model$`log(landvalue)`, step_model2.sqrt1$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt1$model$totalbldgsqft, step_model2.sqrt1$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt1$model$`log(deedacres)`, step_model2.sqrt1$model$`sqrt(totalsalevalue)`)

plot(step_model2.sqrt1$model$built_before_1990, step_model2.sqrt1$model$`sqrt(totalsalevalue)`)
```



Checking normality again

```
qqnorm(resid(step_model2.sqrt1))
qqline(resid(step_model2.sqrt1))
```




```
shapiro.test(resid(step_model2.sqrt1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(step_model2.sqrt1)
## W = 0.96646, p-value = 9.114e-14
```

```
vif(step_model2.sqrt1)
```

```
##         bldgvalue    log(landvalue)     totalbldgsqft    log(deedacres)
##          3.581930          2.594430          3.192301          2.440295
## built_before_1990
##          1.647525
```

Fit models

Model 1: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + B_5 X_5$

```
summary(step_model2.sqrt1)
```

```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) +
##     totalbldgsqft + log(deedacres) + built_before_1990, data = step_model2$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.9970  -2.5161  -0.1859   2.2593  13.7090
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.8183946  3.0629635   1.900   0.0578 .
## bldgvalue           0.0125195  0.0013384   9.354  < 2e-16 ***
## log(landvalue)      1.0327485  0.5484328   1.883   0.0600 .
## totalbldgsqft       0.0004152  0.0002176   1.908   0.0566 .
## log(deedacres)     -0.3732355  0.2380919  -1.568   0.1173
## built_before_19901 -1.6303420  0.3434232  -4.747 2.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.052 on 920 degrees of freedom
## Multiple R-squared:  0.392,  Adjusted R-squared:  0.3887
## F-statistic: 118.6 on 5 and 920 DF,  p-value: < 2.2e-16
```

The assumption checks for model 1 is shown above at the end of analysis 1.

Looking at the stepwise model for bldgvalue, we should check it as a potential higher-order. Bldgvalue, landvalue, and totalbldgsqft seem like they can also interact with each other.

Model 2: Testing higher-order of X1 and X3 and interactions between x1, x2, and x3 $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1^2 + \beta_7 X_1 X_2 + \beta_8 X_2 X_3 + \beta_9 X_1 X_3$

```
fitmodel2 <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft + log(deedacres) +
                  built_before_1990 + I(bldgvalue^2) +
                  bldgvalue:log(landvalue) +
                  log(landvalue):totalbldgsqft +
                  bldgvalue:totalbldgsqft, data=step_model2$model)

summary(fitmodel2)
```

```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) +
##     totalbldgsqft + log(deedacres) + built_before_1990 + I(bldgvalue^2) +
##     bldgvalue:log(landvalue) + log(landvalue):totalbldgsqft +
##     bldgvalue:totalbldgsqft, data = step_model2$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6100  -2.3345  -0.2091   2.2121  12.1225
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -4.308e+00  5.432e+00  -0.793  0.42787
## bldgvalue                     6.184e-02  1.464e-02   4.225 2.62e-05 ***
## log(landvalue)                2.159e+00  1.011e+00   2.135  0.03300 *
## totalbldgsqft                -3.675e-03  1.963e-03  -1.872  0.06149 .
## log(deedacres)               -7.157e-01  2.490e-01  -2.874  0.00414 **
## built_before_19901           -3.221e-01  3.571e-01  -0.902  0.36729
## I(bldgvalue^2)               -5.565e-06  2.230e-06  -2.496  0.01273 *
## bldgvalue:log(landvalue)     -6.146e-03  2.706e-03  -2.271  0.02337 *
## log(landvalue):totalbldgsqft  6.645e-04  3.528e-04   1.884  0.05993 .
## bldgvalue:totalbldgsqft      -6.953e-07  6.146e-07  -1.131  0.25820
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.853 on 916 degrees of freedom
## Multiple R-squared:  0.4528, Adjusted R-squared:  0.4474
## F-statistic: 84.22 on 9 and 916 DF,  p-value: < 2.2e-16
```

Adjusted R-squared is higher than model1

```
vif(fitmodel2)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                    bldgvalue               log(landvalue)
##                   473.834145                    9.760437
##                totalbldgsqft               log(deedacres)
##                   287.532718                    2.953002
##            built_before_1990               I(bldgvalue^2)
##                     1.970549                   22.310847
##     bldgvalue:log(landvalue) log(landvalue):totalbldgsqft
##                   576.008678                  363.253467
##      bldgvalue:totalbldgsqft
##                    38.251771
```

```
anova(step_model2.sqrt1, fitmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990
## Model 2: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990 + I(bldgvalue^2) + bldgvalue:log(landvalue) +
##     log(landvalue):totalbldgsqft + bldgvalue:totalbldgsqft
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    920 15109
## 2    916 13597  4    1511.8 25.462 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H0 : \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ $Ha : \beta_i \neq 0, i = 6, 7, 8, 9$ Test statistic: F-test = 29.709 p-value: <2.2e-16 Conclusion: Since the p value is less than 0.05, we can reject the null hypothesis and conclude that at least one of the tested variables is not zero. This suggests that at least one of the interactions or higher-order variables are significant contributors to the model.

Looking at the p-values from the individual t-tests, we can see that the interaction of bldgvalue(X1) and totalbldgsqft (X3) has a p-value 0.632800 so we will exclude it in the next model. The variable deedacres seems like it would interact with totalbldgsqft, bldgvalue, and landvalue so we will test it in model 3.

Model 2 adjusted: Removed X1 and X3 interaction $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1^2 + \beta_7 X_1 X_2 + \beta_8 X_2 X_3$

```
fitmodel2adj <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft + log(deedacres) +
                   built_before_1990 + I(bldgvalue^2) +
                   bldgvalue:log(landvalue) +
                   log(landvalue):totalbldgsqft, data=step_model2$model)

summary(fitmodel2adj)
```

```
##
## Call:
```

```
## lm(formula = sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) +
##     totalbldgsqft + log(deedacres) + built_before_1990 + I(bldgvalue^2) +
##     bldgvalue:log(landvalue) + log(landvalue):totalbldgsqft,
##     data = step_model2$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5565  -2.3554  -0.2206   2.1395  12.6748
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -7.049e+00  4.862e+00  -1.450  0.14747
## bldgvalue                      6.053e-02  1.459e-02   4.148 3.66e-05 ***
## log(landvalue)                 2.757e+00  8.628e-01   3.195  0.00145 **
## totalbldgsqft                 -2.791e-03  1.801e-03  -1.550  0.12157
## log(deedacres)                -7.572e-01  2.463e-01  -3.074  0.00217 **
## built_before_19901            -3.429e-01  3.567e-01  -0.961  0.33660
## I(bldgvalue^2)                -7.226e-06  1.679e-06  -4.304 1.86e-05 ***
## bldgvalue:log(landvalue)      -5.966e-03  2.702e-03  -2.208  0.02748 *
## log(landvalue):totalbldgsqft  4.388e-04  2.910e-04   1.508  0.13191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.853 on 917 degrees of freedom
## Multiple R-squared:  0.452,  Adjusted R-squared:  0.4473
## F-statistic: 94.56 on 8 and 917 DF,  p-value: < 2.2e-16
```

Model 3: Testing interaction between X4 and X1, X2, X3. $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1^2 + \beta_7 X_1 X_2 + \beta_8 X_2 X_3 + \beta_9 X_4 X_2 + \beta_{10} X_4 X_3 + \beta_{11} X_4 X_1$

```
fitmodel3 <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft + log(deedacres) +
                  built_before_1990 + I(bldgvalue^2) +
                  bldgvalue:log(landvalue) +
                  log(landvalue):totalbldgsqft +
                  log(deedacres):log(landvalue) + log(deedacres):totalbldgsqft +
                  log(deedacres):bldgvalue, data=step_model2$model)

summary(fitmodel3)
```

```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) +
##     totalbldgsqft + log(deedacres) + built_before_1990 + I(bldgvalue^2) +
##     bldgvalue:log(landvalue) + log(landvalue):totalbldgsqft +
##     log(deedacres):log(landvalue) + log(deedacres):totalbldgsqft +
##     log(deedacres):bldgvalue, data = step_model2$model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.578  -2.242  -0.249   2.066  13.650
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.039e+00  6.784e+00   0.301  0.76380
## bldgvalue                      4.602e-02  1.831e-02   2.514  0.01212 *
```

```
## log(landvalue)                 1.419e+00  1.186e+00   1.197  0.23153
## totalbldgsqft                 -5.226e-03  2.734e-03  -1.912  0.05620 .
## log(deedacres)                -1.034e+00  1.762e+00  -0.587  0.55735
## built_before_19901            -2.767e-01  3.580e-01  -0.773  0.43986
## I(bldgvalue^2)                -6.267e-06  2.031e-06  -3.086  0.00209 **
## bldgvalue:log(landvalue)      -3.807e-03  3.084e-03  -1.235  0.21727
## log(landvalue):totalbldgsqft   8.010e-04  4.286e-04   1.869  0.06194 .
## log(landvalue):log(deedacres)  3.306e-01  3.607e-01   0.917  0.35963
## totalbldgsqft:log(deedacres)  -5.249e-04  3.617e-04  -1.451  0.14711
## bldgvalue:log(deedacres)      -9.537e-04  1.985e-03  -0.481  0.63096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.843 on 914 degrees of freedom
## Multiple R-squared:  0.4567, Adjusted R-squared:  0.4502
## F-statistic: 69.85 on 11 and 914 DF,  p-value: < 2.2e-16
```

```
anova(fitmodel2adj, fitmodel3)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990 + I(bldgvalue^2) + bldgvalue:log(landvalue) +
##     log(landvalue):totalbldgsqft
## Model 2: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990 + I(bldgvalue^2) + bldgvalue:log(landvalue) +
##     log(landvalue):totalbldgsqft + log(deedacres):log(landvalue) +
##     log(deedacres):totalbldgsqft + log(deedacres):bldgvalue
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    917 13616
## 2    914 13500  3    116.04 2.6188 0.04972 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H0 : \beta_9 = \beta_10 = \beta_11 = 0$ $Ha : \beta_i \neq 0, i = 9, 10, 11$ Test statistic: F-test = 0.0002073 p-value: 6.5901

Conclusion: Since the p value is less than 0.05, we can reject the null hypothesis and conclude that at least one of the tested terms had a significant impact on the model. Looking at the p-values of the individual t-tests, we can see that the interaction between deedacres(X4) and totalbldgedqft(X3) is 0.737561 so we will exclude the interaction in the next model.

Since we haven't seen the interactions between predictor built_before_1990 and other variables, we will check them.

Model 3adj: Removed interaction between X4 and X3, X4 and X1 $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1^2 + \beta_7 X_1 X_2 + \beta_8 X_2 X_3 + \beta_9 X_4 X_2$

```
fitmodel3adj <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft + log(deedacres) +
                   built_before_1990 + I(bldgvalue^2) +
                   bldgvalue:log(landvalue) +
                   log(landvalue):totalbldgsqft +
                   log(deedacres):log(landvalue), data=step_model2$model)


summary(fitmodel3adj)
```

```
##
```

```
## Call:
## lm(formula = sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) +
##     totalbldgsqft + log(deedacres) + built_before_1990 + I(bldgvalue^2) +
##     bldgvalue:log(landvalue) + log(landvalue):totalbldgsqft +
##     log(deedacres):log(landvalue), data = step_model2$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6501  -2.3671  -0.2248   2.1441  12.5478
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -7.889e+00  5.693e+00  -1.386  0.16615
## bldgvalue                      5.984e-02  1.480e-02   4.043 5.71e-05 ***
## log(landvalue)                 2.918e+00  1.034e+00   2.824  0.00485 **
## totalbldgsqft                 -2.630e-03  1.888e-03  -1.393  0.16398
## log(deedacres)                -1.252e+00  1.760e+00  -0.712  0.47690
## built_before_19901            -3.510e-01  3.580e-01  -0.980  0.32712
## I(bldgvalue^2)                -7.360e-06  1.745e-06  -4.218 2.70e-05 ***
## bldgvalue:log(landvalue)      -5.821e-03  2.751e-03  -2.116  0.03458 *
## log(landvalue):totalbldgsqft   4.094e-04  3.089e-04   1.326  0.18532
## log(landvalue):log(deedacres)  9.985e-02  3.514e-01   0.284  0.77633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.855 on 916 degrees of freedom
## Multiple R-squared:  0.4521, Adjusted R-squared:  0.4467
## F-statistic: 83.98 on 9 and 916 DF,  p-value: < 2.2e-16
```

```
anova(fitmodel3adj, fitmodel3)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990 + I(bldgvalue^2) + bldgvalue:log(landvalue) +
##     log(landvalue):totalbldgsqft + log(deedacres):log(landvalue)
## Model 2: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990 + I(bldgvalue^2) + bldgvalue:log(landvalue) +
##     log(landvalue):totalbldgsqft + log(deedacres):log(landvalue) +
##     log(deedacres):totalbldgsqft + log(deedacres):bldgvalue
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    916 13615
## 2    914 13500  2    114.84 3.8876 0.02083 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 4: Testing interaction between X5 and all other predictors $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1^2 + \beta_7 X_1 X_2 + \beta_8 X_2 X_3 + \beta_9 X_4 X_2 + \beta_1 0 X_5 X_1 + \beta_1 1 X_5 X_2 + \beta_1 2 X_5 X_3 + \beta_1 3 X_5 X_4$

```
fitmodel4 <- lm(sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft + log(deedacres) +
                  built_before_1990 + I(bldgvalue^2) +
                  bldgvalue:log(landvalue) +
                  log(landvalue):totalbldgsqft +
                  log(deedacres):log(landvalue) +
                  built_before_1990:bldgvalue +
                  built_before_1990:log(landvalue) +
```

```
                  built_before_1990:totalbldgsqft +
                  built_before_1990:log(deedacres), data=step_model2$model)

summary(fitmodel4)
```

```
##
## Call:
## lm(formula = sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) +
##     totalbldgsqft + log(deedacres) + built_before_1990 + I(bldgvalue^2) +
##     bldgvalue:log(landvalue) + log(landvalue):totalbldgsqft +
##     log(deedacres):log(landvalue) + built_before_1990:bldgvalue +
##     built_before_1990:log(landvalue) + built_before_1990:totalbldgsqft +
##     built_before_1990:log(deedacres), data = step_model2$model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2096  -2.3415  -0.2739   2.1220  12.1584
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.513e+01  1.068e+01   1.416  0.15700
## bldgvalue                         2.892e-02  1.925e-02   1.502  0.13343
## log(landvalue)                   -1.538e+00  2.043e+00  -0.753  0.45168
## totalbldgsqft                    -1.354e-03  1.947e-03  -0.695  0.48693
## log(deedacres)                    8.600e-01  1.956e+00   0.440  0.66026
## built_before_19901               -1.925e+01  8.313e+00  -2.316  0.02079 *
## I(bldgvalue^2)                   -1.282e-05  2.388e-06  -5.367 1.01e-07 ***
## bldgvalue:log(landvalue)          1.678e-03  3.773e-03   0.445  0.65658
## log(landvalue):totalbldgsqft      6.344e-05  3.284e-04   0.193  0.84689
## log(landvalue):log(deedacres)    -1.910e-01  3.778e-01  -0.506  0.61320
## bldgvalue:built_before_19901     -1.182e-02  3.852e-03  -3.068  0.00222 **
## log(landvalue):built_before_19901  3.639e+00  1.548e+00   2.350  0.01898 *
## totalbldgsqft:built_before_19901   1.120e-03  5.488e-04   2.041  0.04157 *
## log(deedacres):built_before_19901 -8.233e-01  5.041e-01  -1.633  0.10281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.834 on 912 degrees of freedom
## Multiple R-squared:  0.4605, Adjusted R-squared:  0.4528
## F-statistic: 59.89 on 13 and 912 DF,  p-value: < 2.2e-16
```

```
anova(fitmodel3adj, fitmodel4)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990 + I(bldgvalue^2) + bldgvalue:log(landvalue) +
##     log(landvalue):totalbldgsqft + log(deedacres):log(landvalue)
## Model 2: sqrt(totalsalevalue) ~ bldgvalue + log(landvalue) + totalbldgsqft +
##     log(deedacres) + built_before_1990 + I(bldgvalue^2) + bldgvalue:log(landvalue) +
##     log(landvalue):totalbldgsqft + log(deedacres):log(landvalue) +
##     built_before_1990:bldgvalue + built_before_1990:log(landvalue) +
##     built_before_1990:totalbldgsqft + built_before_1990:log(deedacres)
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1    916 13615
## 2    912 13405  4     210.1 3.5736 0.006686 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H0 : \beta_10 = \beta_11 = \beta_12 = \beta_13 = 0$ $Ha : \beta_i \neq 0, i = 10, 11, 12, 13$ Test statistic: 0.7817 p-value: 0.5372

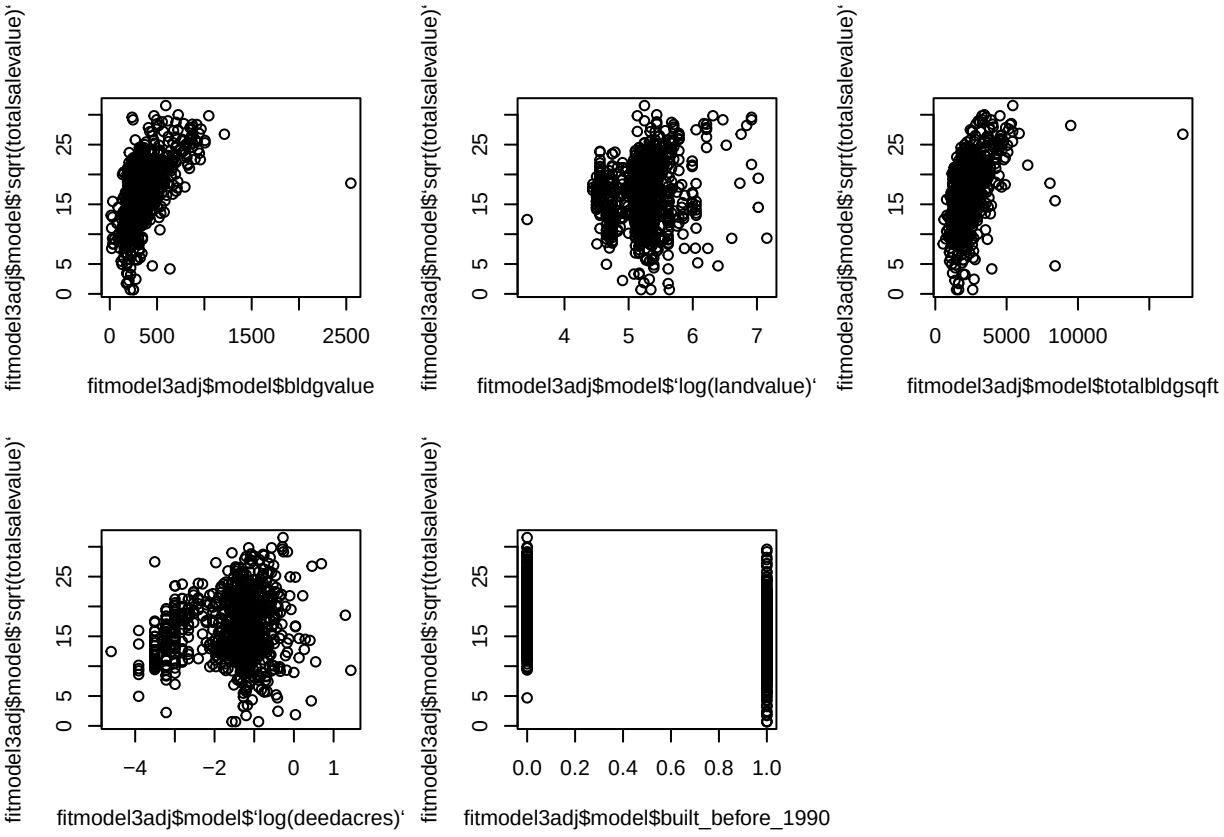
Conclusion: We fail to reject the null hypothesis and conclude that none of these tested terms have a significant impact on the model. Looking at the p-values of the individual t-tests, we can see that every single one had a p-value > 0.05. So we will exclude all these interactions in the final model.

Residual Assumption tests for final model

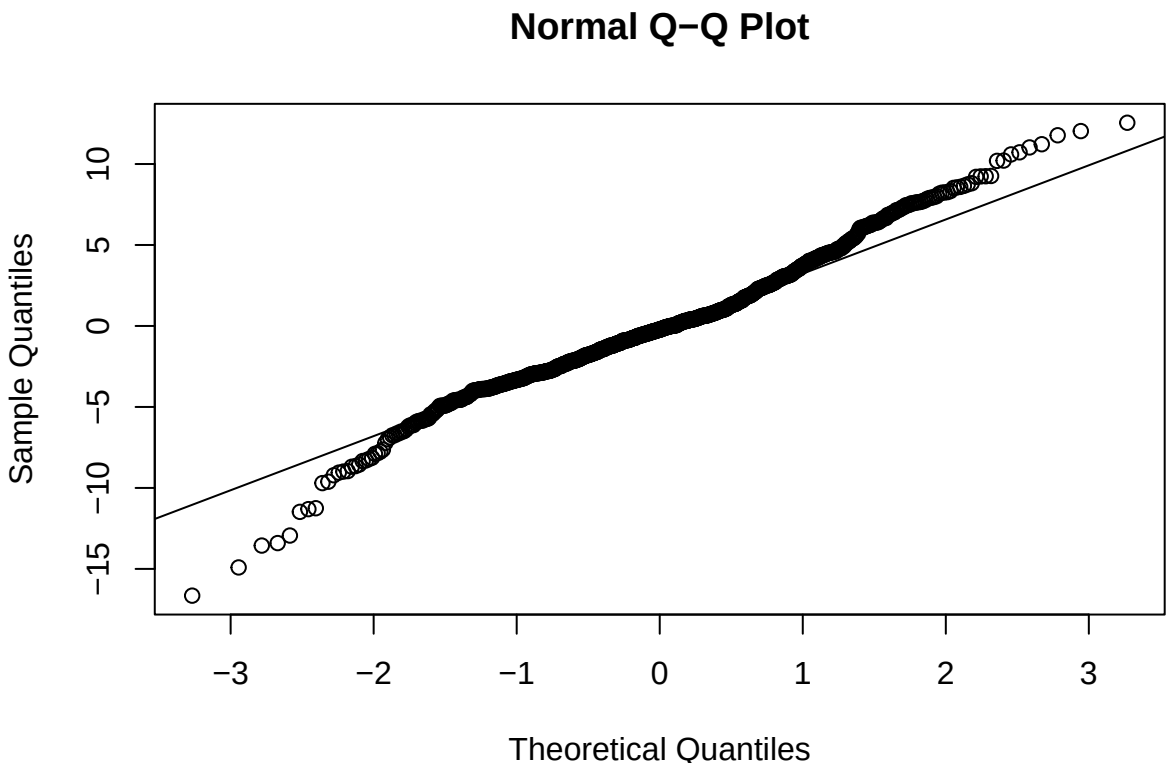
Final Model: $y = 6.468 + 0.106X_1 + 0.249X_2 - 1.160e-02X_3 + 5.125X_4 + 0.362X_5 - 9.795e-06X_1^2 - 1.360e-02X_1X_2 + 1.943e-03X_2X_3 - 0.975X_4X_2 - 2.572e-03X_4X_1$

```
# fitmodel3adj$model # To check
```

```
par(mfrow=c(2,3))
plot(fitmodel3adj$model$bldgvalue, fitmodel3adj$model$`sqrt(totalsalevalue)`)

plot(fitmodel3adj$model$`log(landvalue)`, fitmodel3adj$model$`sqrt(totalsalevalue)`)

plot(fitmodel3adj$model$totalbldgsqft, fitmodel3adj$model$`sqrt(totalsalevalue)`)

plot(fitmodel3adj$model$`log(deedacres)`, fitmodel3adj$model$`sqrt(totalsalevalue)`)

plot(fitmodel3adj$model$built_before_1990, fitmodel3adj$model$`sqrt(totalsalevalue)`)
```



```r
qqnorm(resid(fitmodel3adj))
qqline(resid(fitmodel3adj))
```



Some large deviations at the left tail, right tail isn't concerning.

```r
shapiro.test(resid(fitmodel3adj))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fitmodel3adj)
## W = 0.98129, p-value = 1.605e-09
```

Mild deviations from normality, P-value lower than 0.05 means data doesn't follow normal distribution.