Determining a Fair Apartment Price

Stat 410 Final Project

Jack Dzialo, Devin Abraham, Tom Vincent

Purpose

Why choose our dataset?

- Aimed to address an actual problem that we face
- Currently looking for housing next year

Challenge: How to determine if price is fair?

Proposal

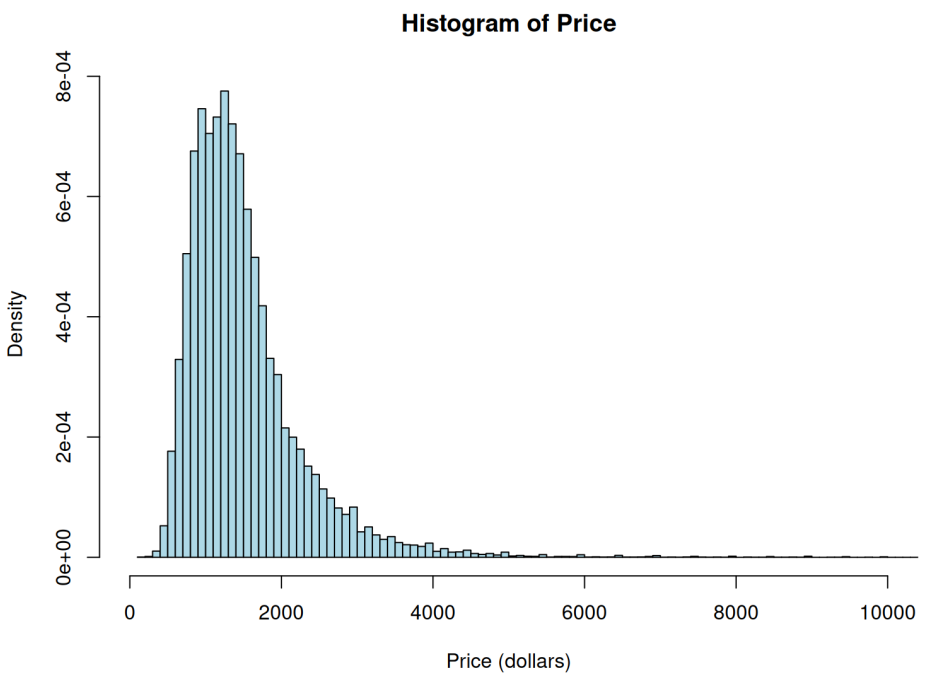
- Chose to use Forward Stepwise Search, using F tests
  - → Tested interaction terms once individual predictors were determined
- Read up upon different ways to determine model accuracy
- Use confidence intervals generated by the model to see if posted price for an apartment is fair

Definitions

1. **AIC** - Measures trade off between model fit and complexity with a penalty for more complexity
2. **AICc** - Corrected version of AIC with a larger penalty for regressors
3. **BIC** - Penalizes model complexity heavily and favors simple model with large sample sizes
4. **PRESS** - Predicted Residual Error Sum of Squares; predicts how accurate a model is at predicting data it hasn't seen before

Data Collection

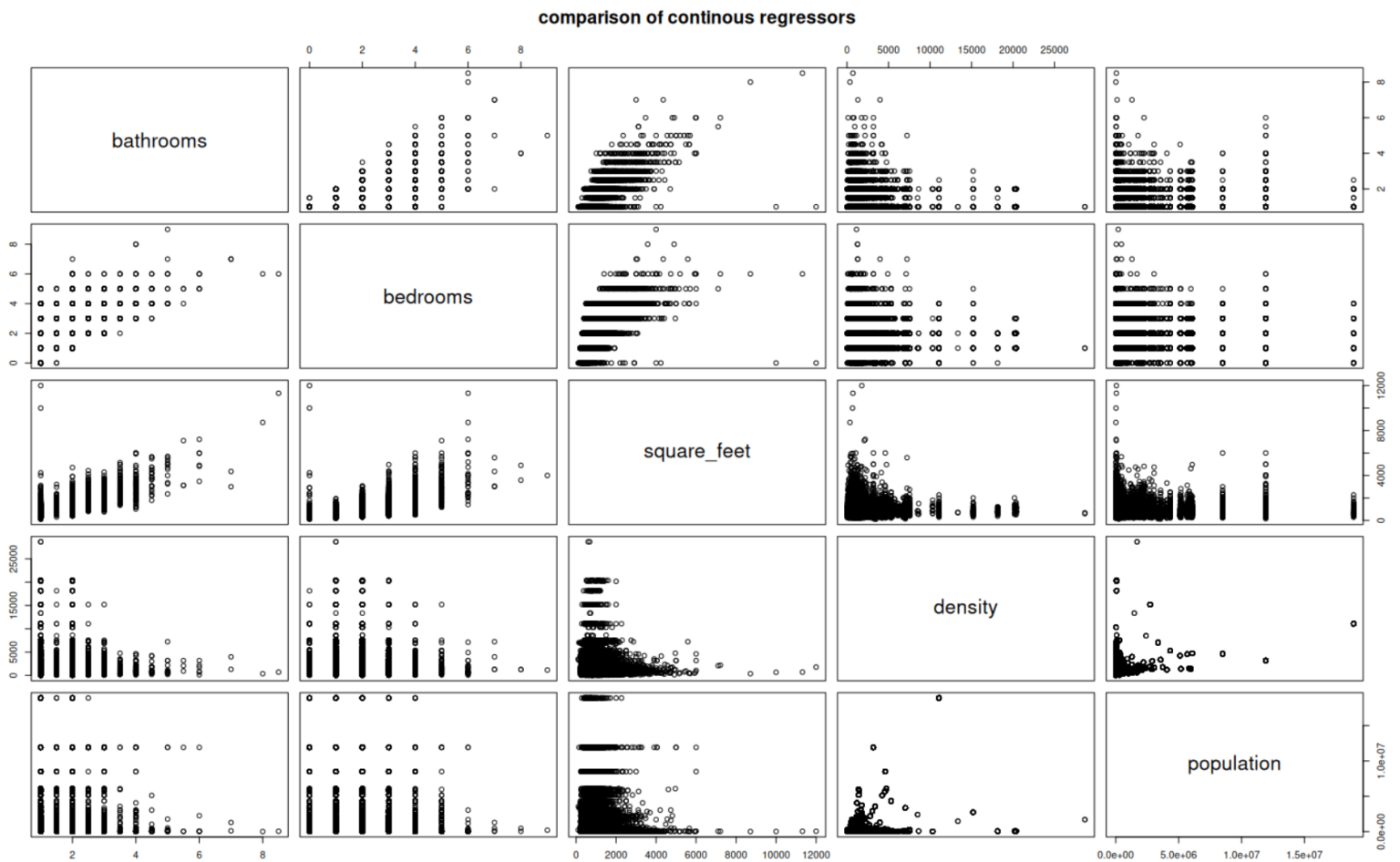
- Dataset of apartments for rent, sourced from UCI dataset repository
- 52,000 rows, each representing a different apartment
- 22 columns, each representing a descriptor for an apartment listing
  - → Square feet, apartment type, number of bathrooms, has photos



Statistic	Value
Mean	1524
Max	52500
Min	100
Median	1349

Data Collection

The plot below shows each of the continuous predictors plotted against each other, in order to determine if multicollinearity exists.



Data Analysis - Model Building

Forward Stepwise Search

First Iteration

For the first iteration of our stepwise search, we start off with a linear model that only consists of an intercept, β_0.

Model:

$$Y = \beta_0$$

After fitting the basic model with every predictor, we find that the Square Feet variable results in the lowest P value of the model.

Statistic	Values
Predictor	square_feet
P-Value	0
Adjusted R Squared	0.174501333784464
AIC	703090.036256909
AICc	703090.039689993
BIC	703107.771144924
PRESS	34891683588.1365

Forward Stepwise Search

Second Iteration

For the second iteration of our stepwise search, we use the linear model from the previous iteration, being the model with an intercept β_0 and a single predictor Square Feet, denoted as X_1.

Model:

$$Y = \beta_0 + X_1\beta_1$$

After fitting the model with every predictor, we find that the Density variable results in the lowest P value of the model.

Statistic	Values
Predictor	density
P-Value	0
Adjusted R Squared	0.318251588121893
AIC	693057.659341886
AICc	693057.666284873
BIC	693084.261673907
PRESS	28826984608.1955

Forward Stepwise Search

Second Iteration

Current:

Statistic	Values
Predictor	density
P-Value	0
Adjusted R Squared	0.318251588121893
AIC	693057.659341886
AICc	693057.666284873
BIC	693084.261673907
PRESS	28826984608.1955

Previous:

Statistic	Values
Predictor	square_feet
P-Value	0
Adjusted R Squared	0.174501333784464
AIC	703090.036256909
AICc	703090.039689993
BIC	703107.771144924
PRESS	34891683588.1365

The model improved when including density, while **AIC**, **AICc**, **BIC**, **PRESS** got smaller.

Forward Stepwise Search

Third Iteration

For the third iteration, we now include both predictors: Square Feet (denoted by X_1), and density (denoted by X_2).

Model:

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2$$

After fitting the model with every predictor, we find that the Population variable results in the lowest P value of the model.

Statistic	Values
Predictor	population
P-Value	0
Adjusted R Squared	0.338395188475126
AIC	691485.834329277
AICc	691485.846003542
BIC	691521.304105305
PRESS	27977452423.3496

Forward Stepwise Search

Third Iteration

Current:

Statistic	Values
Predictor	population
P-Value	0
Adjusted R Squared	0.338395188475126
AIC	691485.834329277
AICc	691485.846003542
BIC	691521.304105305
PRESS	27977452423.3496

Previous:

Statistic	Values
Predictor	density
P-Value	0
Adjusted R Squared	0.318251588121893
AIC	693057.659341886
AICc	693057.666284873
BIC	693084.261673907
PRESS	28826984608.1955

Negligible improvement

Forward Stepwise Search

Completed

Hence, we have found our linear model, with the predictor variables being Square Feet (X_1), and density (X_2).

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2$$

32% of the variation in price explained by model

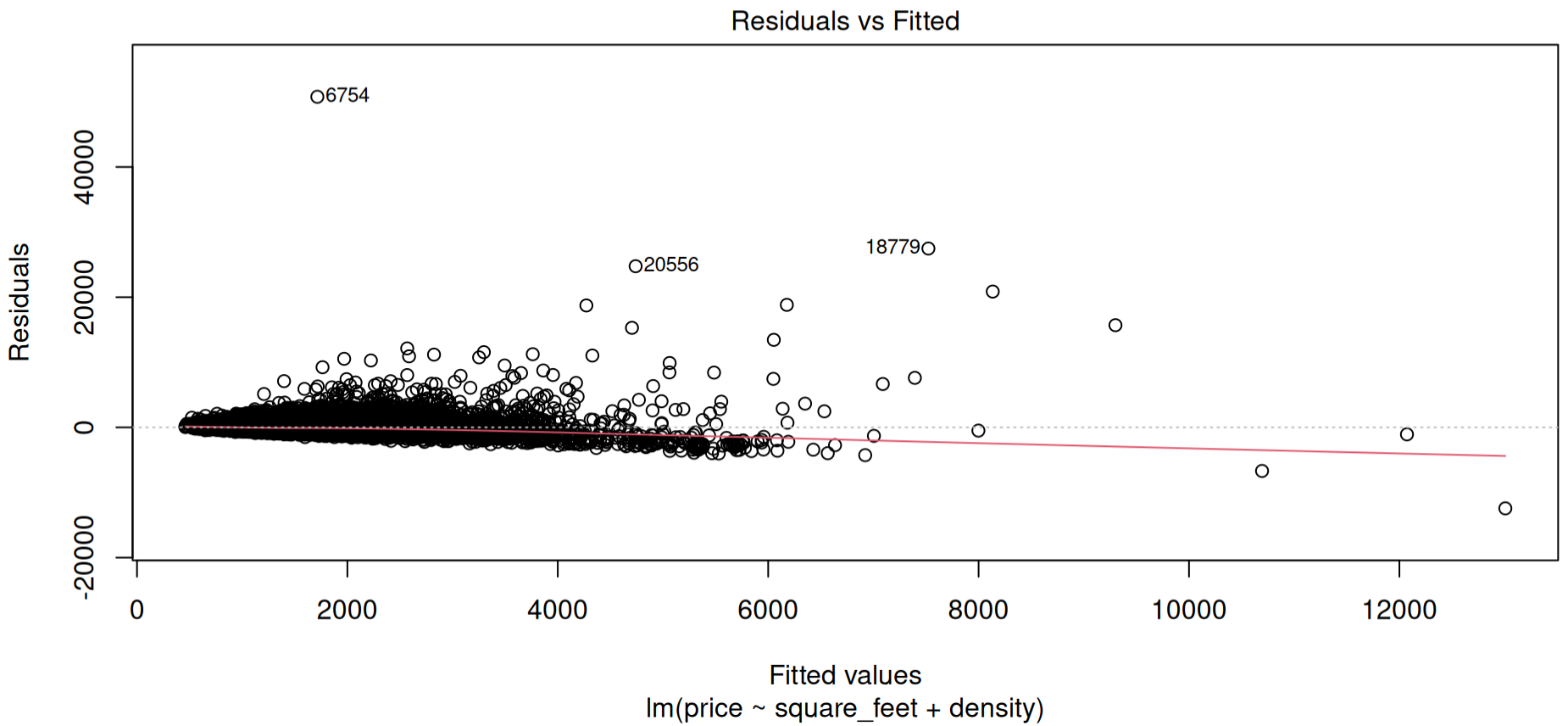
SUMMARY OF MODEL

Statistic	Values	SE	P_Value	R_Squared
Intercept	195.2076	9.4232	0	0.3183
Square Feet	1.0359	0.0084	0	
Density	0.2131	0.0020	0	

Analysis of Residuals - Model Improvement

Residuals

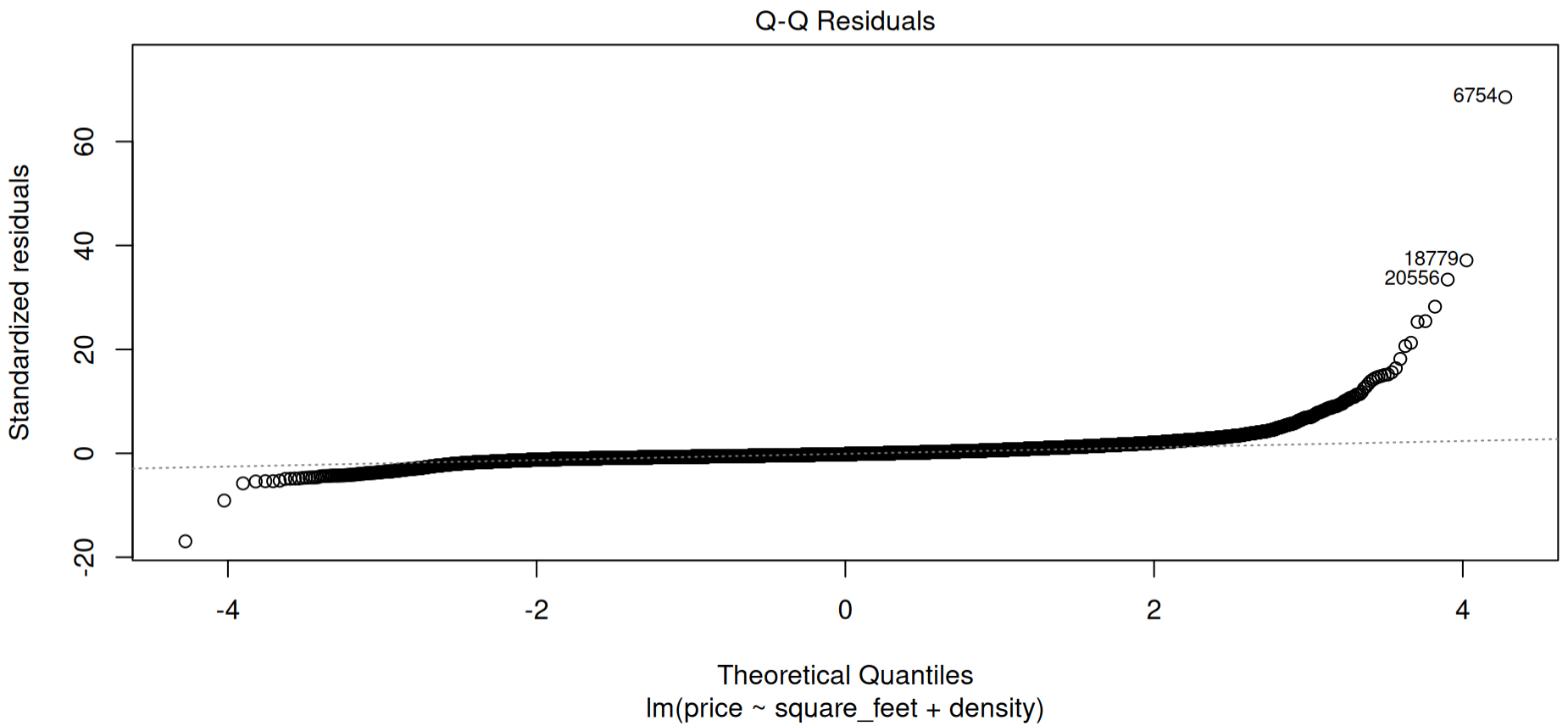
Residuals vs. Fitted Values



Slight heteroskedasticity

Residuals

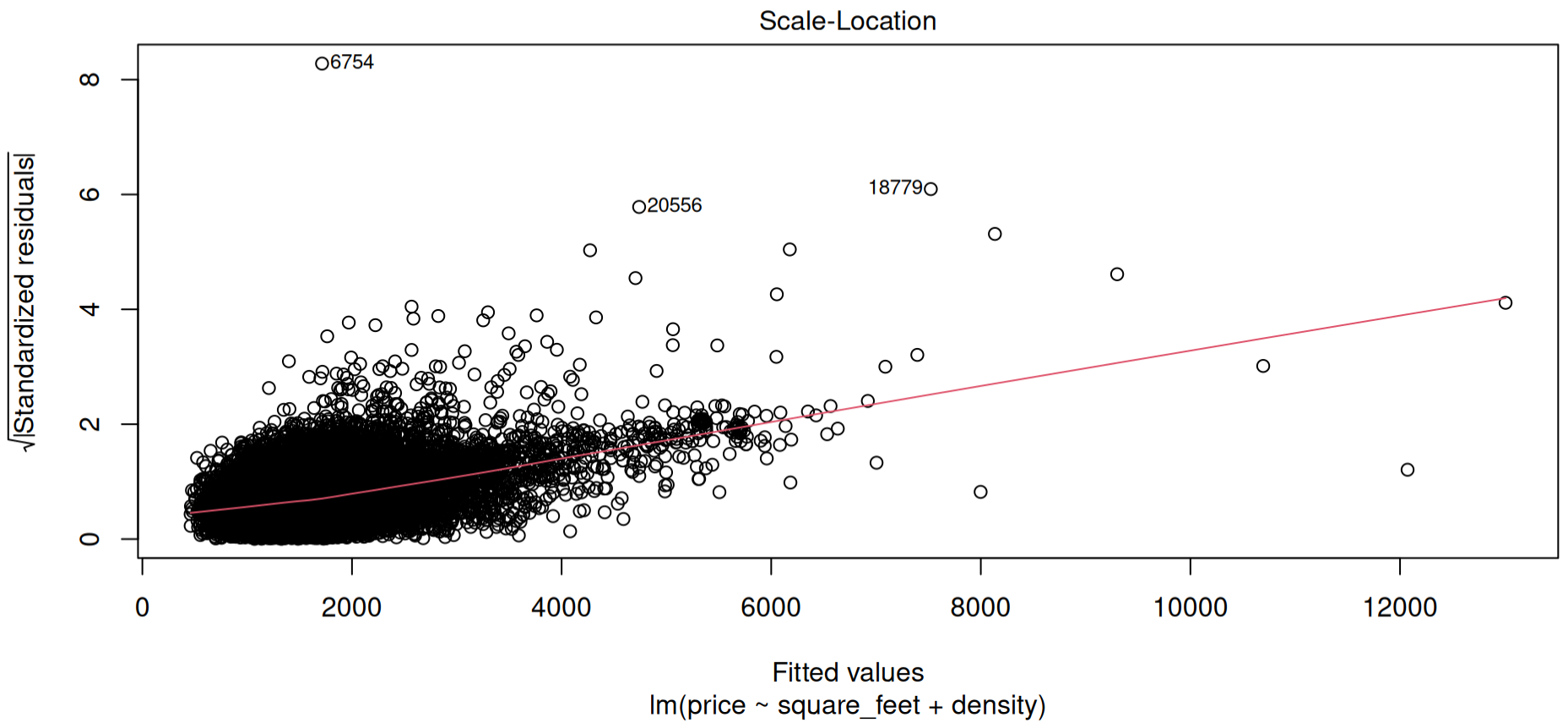
Quantile-Quantile



Relatively normal, room for improvement

Residuals

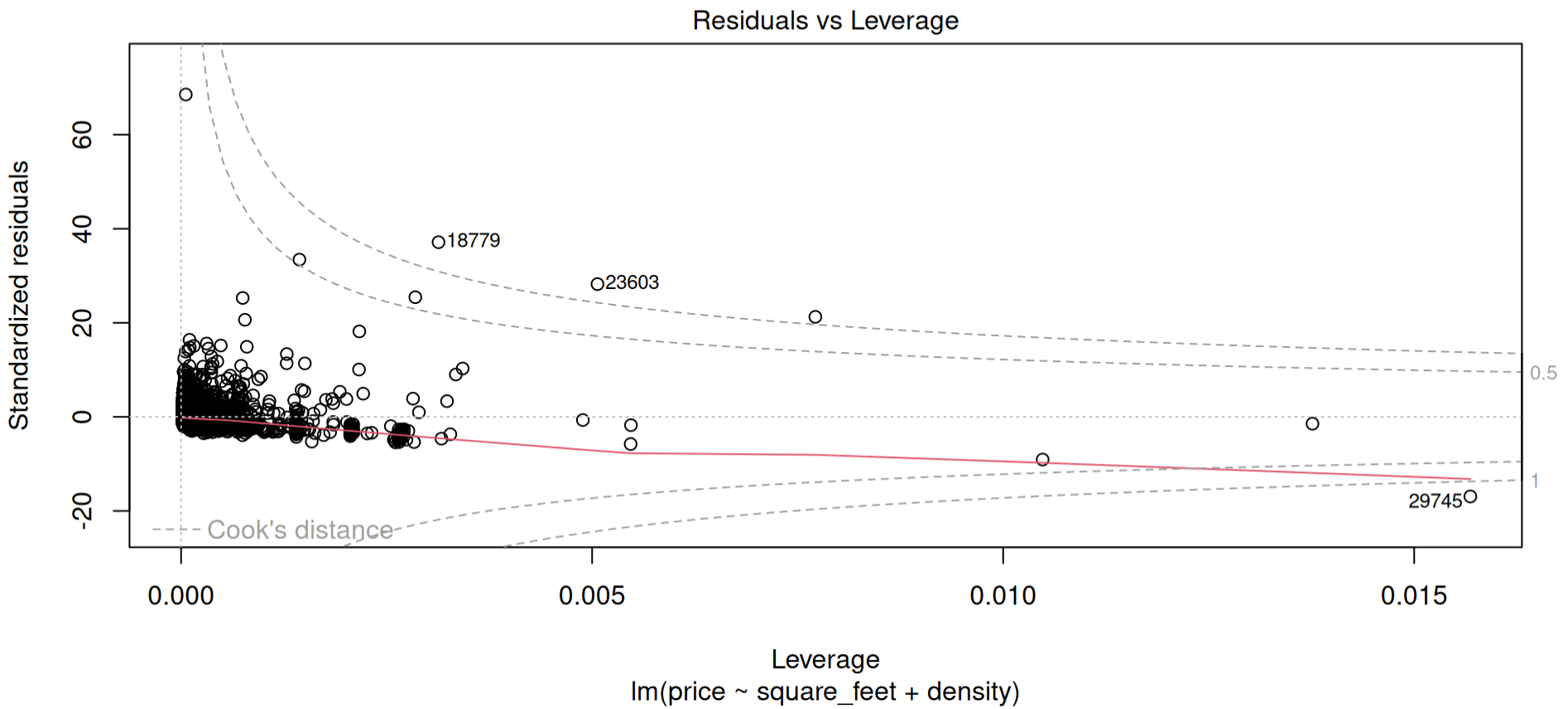
Scale-Location



Another heteroskedasticity indicator

Residuals

Residuals vs. Leverage

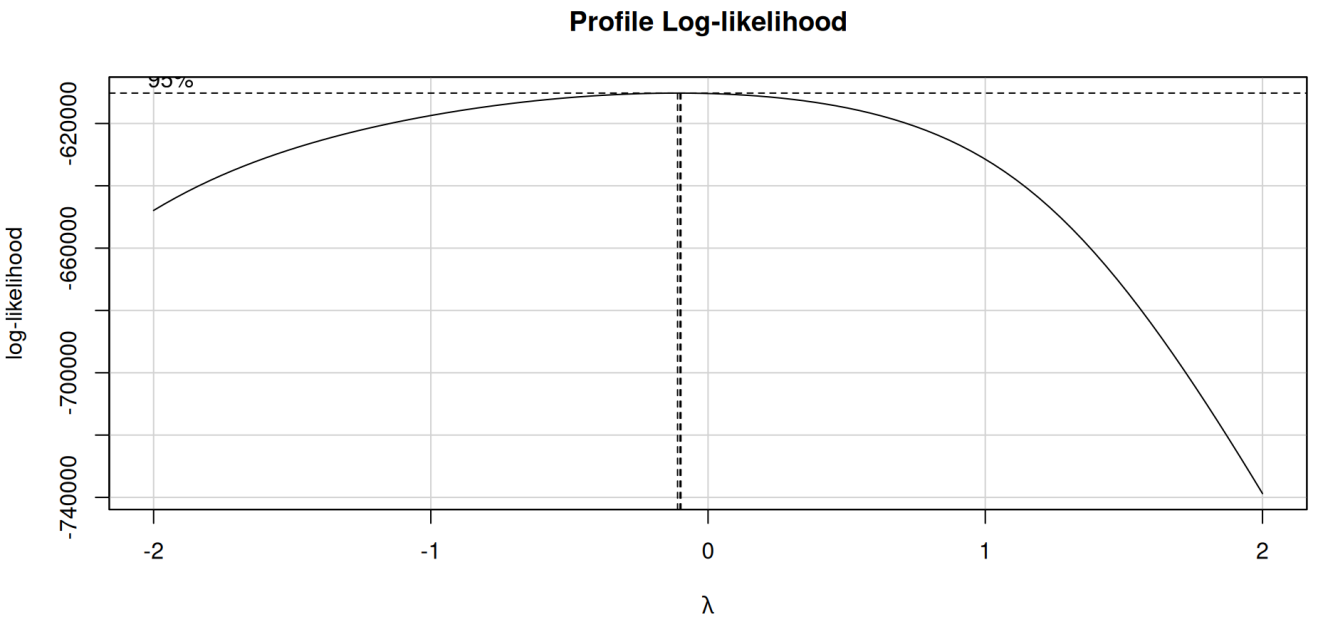


Multiple high leverage points

Adjusted Model

Outliers and Data Transformation

Address non-normality through Box-Cox transformation



- Optimal λ close to 0

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(y), & \text{if } \lambda = 0 \end{cases}$$

Adjusted Model

Summary

Our adjusted model is: $\ln(Y) = \beta_0 + X_1\beta_1 + X_2\beta_2$

SUMMARY OF MODEL

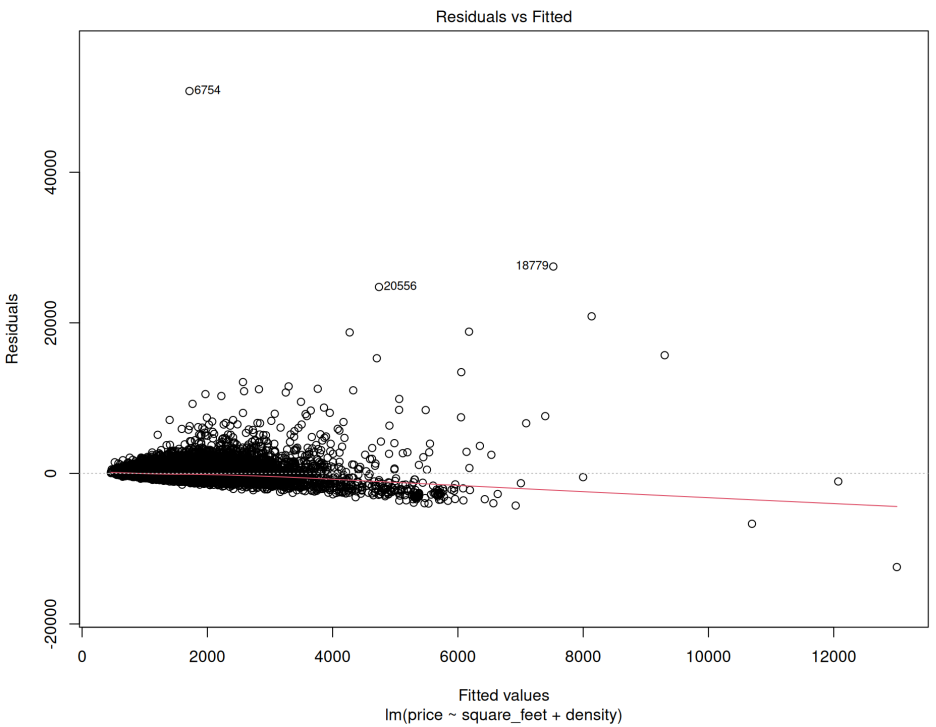
Statistic	Values	SE	P_Value	R_Squared
Intercept	6.5582	0.0046	0	0.3339
Square Feet	0.0005	0.0000	0	
Density	0.0001	0.0000	0	

Improved R^2

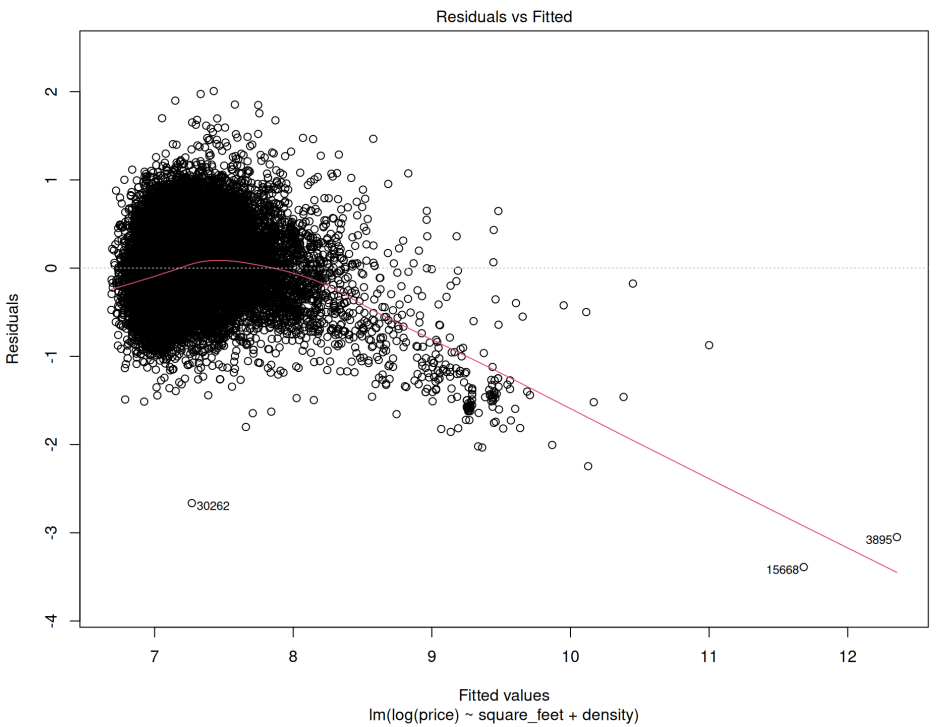
Adjusted Model Residuals

Residuals vs. Fitted Values

Previous



Adjusted

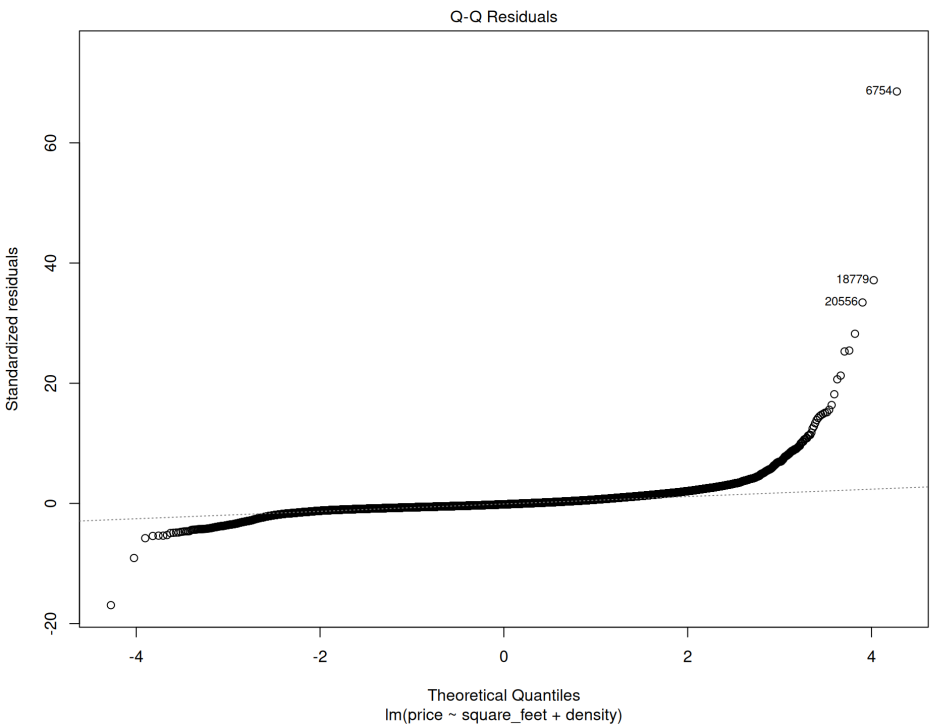


Normallity improved slightly

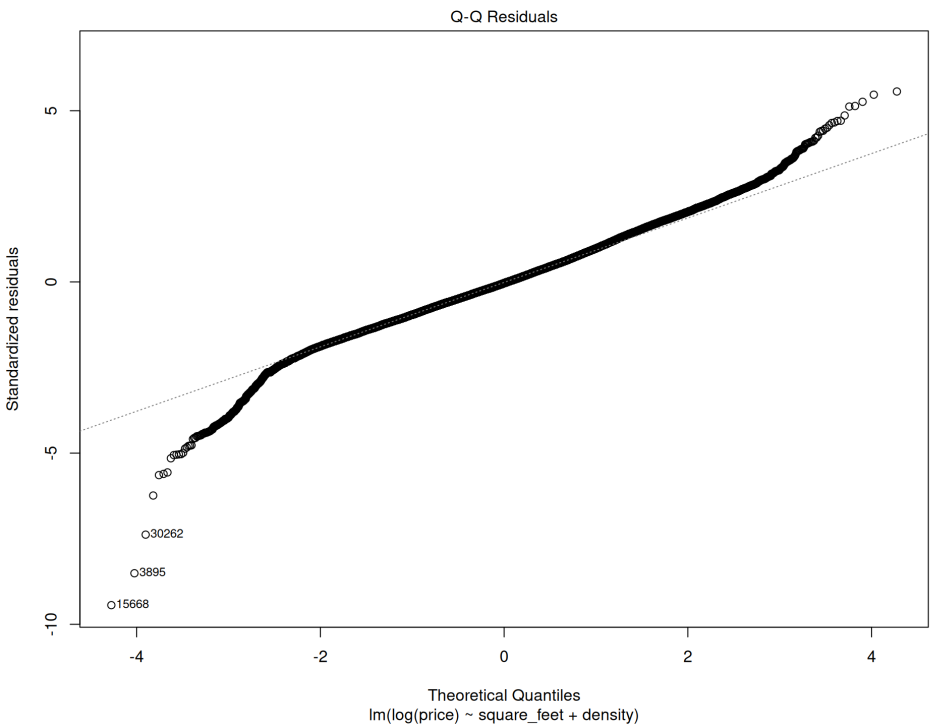
Adjusted Model Residuals

Quantile-Quantile

Previous



Adjusted

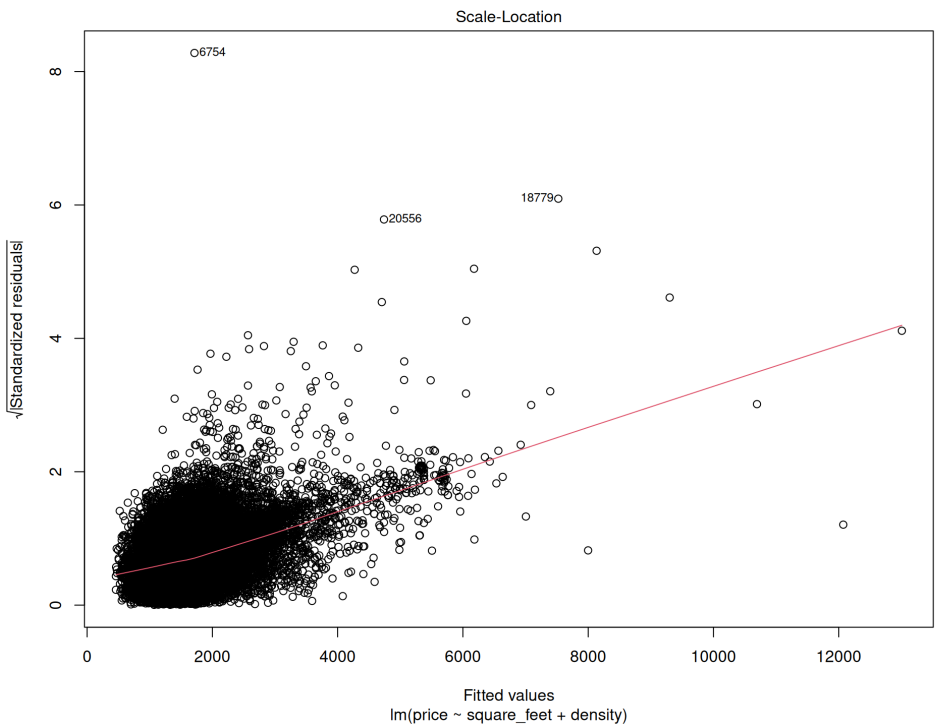


Significant improvement

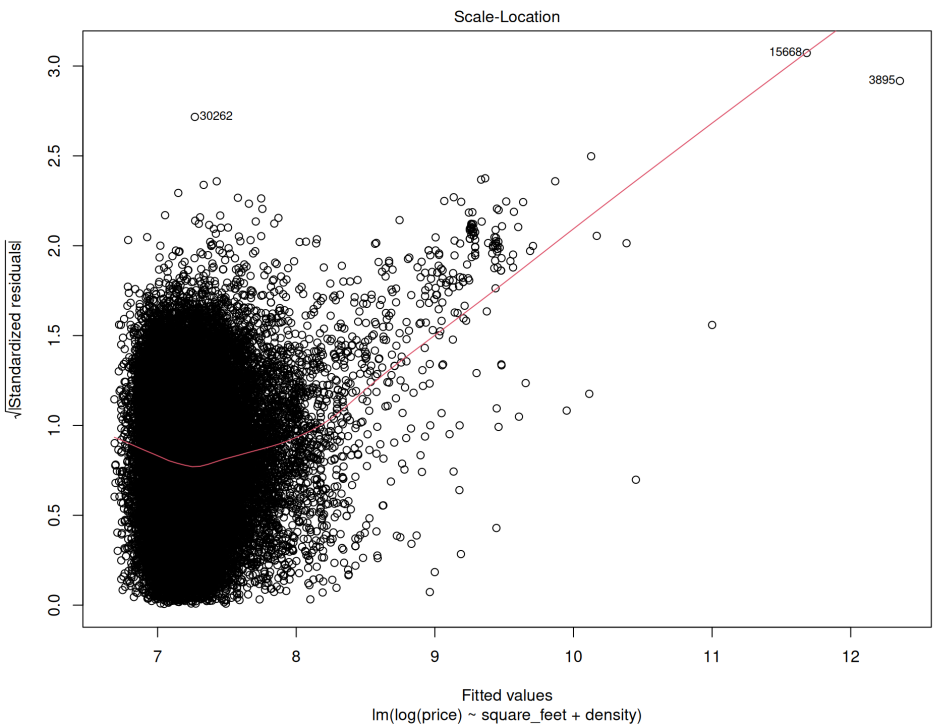
Adjusted Model Residuals

Scale-Location

Previous



Adjusted

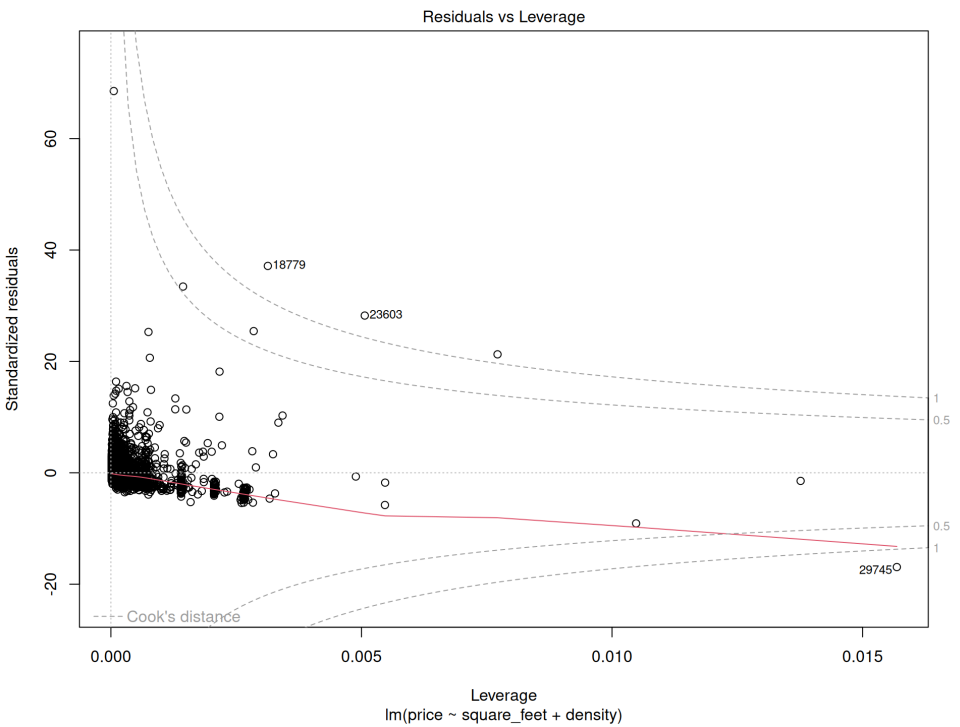


Slight homoscedasticity

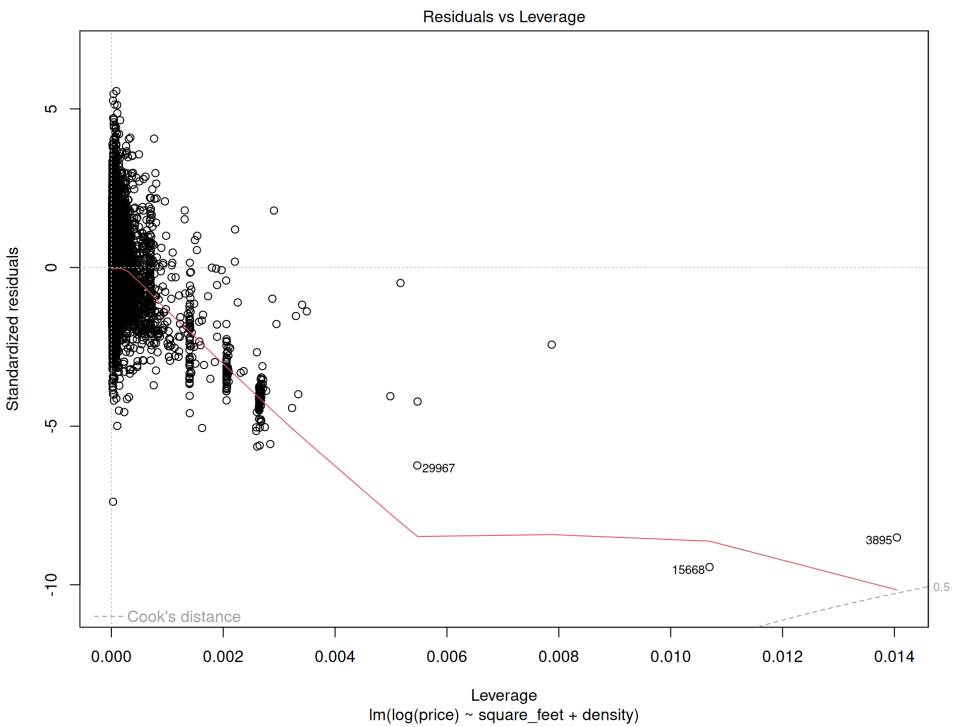
Adjusted Model Residuals

Residuals vs. Leverage

Previous



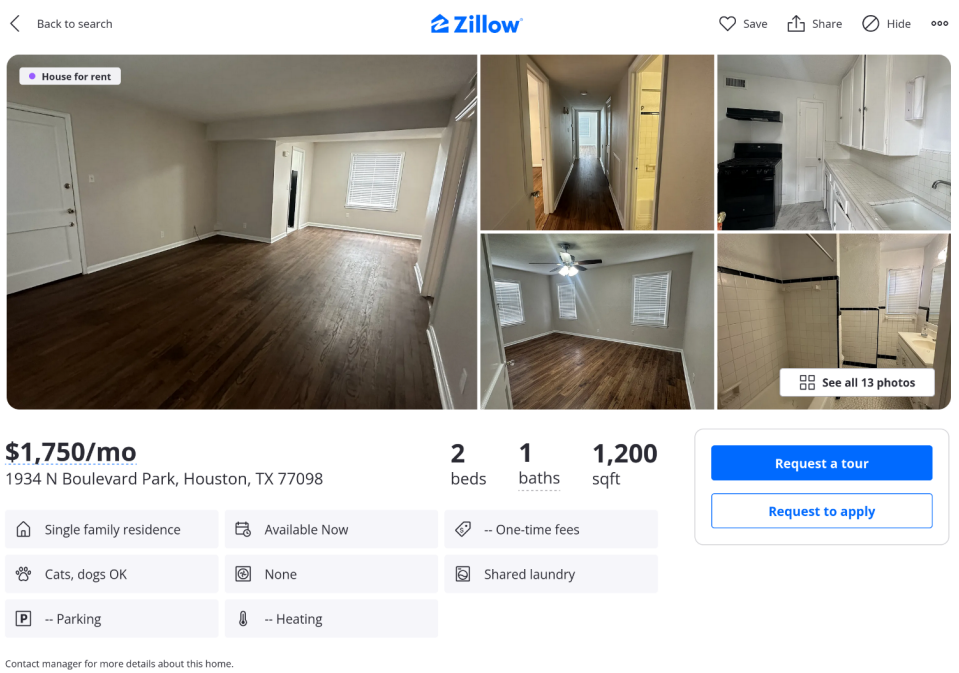
Adjusted



Less high leverage points/outliers

Confidence Interval

Real Apartment Listing



The confidence interval generated by the model:

fit	lwr	upr
1771.733	1763.942	1779.56

- Density metric was taken from the dataset
  - → Scaled to account for higher density

Good prediction!

Summary

- Explored the data and found 8 usable regressors 5 of which were continuous and 3 of which were categorical
- We used Forward Step wise Search to build our model and found 2 regressors for our final model, square_feet and density
- Adjusted our model to take the log of price which improved R^2

Limitations

- Since Forward Step Wise Search is a greedy algorithm so it makes the best choice at each step and does not guarentee the most optimal solution
- In the future we would like a data set with more regressors, our group hypothesizes that a regressor that represents the apartments luxuriousness would have a significant impact

Thank you for listening!