MATH3821 Assignment 2 - Presentation

Kirat Kounsal, Samuel McLeod, Martin Tran

(z5163354), (z5061746), (z5330510)

24th July 2022

Section 1

Goal of Statistical Analysis

Housing in New York City

- Continual housing crisis as there is excess demand and short supply
- Rent control and stabilisation measures are applied to 45% of apartments
- The effects of rent policies are debated
  - Put upward pressure on non-regulated apartments
  - Discourage construction of affordable housing

Do rent control measures have any effect overall?

- What factors affect house and rent prices
- Do they affect them in the same way?
- Aim to create a model predicting rent prices, so we can determine if housing prices have stayed in line with them over time

Section 2

Data Collection and Exploration

The New York City Housing and Vacancy Survey

- Survey conducted every 3 years on various New York properties
- Captures 35 different variables, reflecting various physical, social, economic, and demographic factors
- 102218 total observations from 1991 to 2017

Data Cleaning

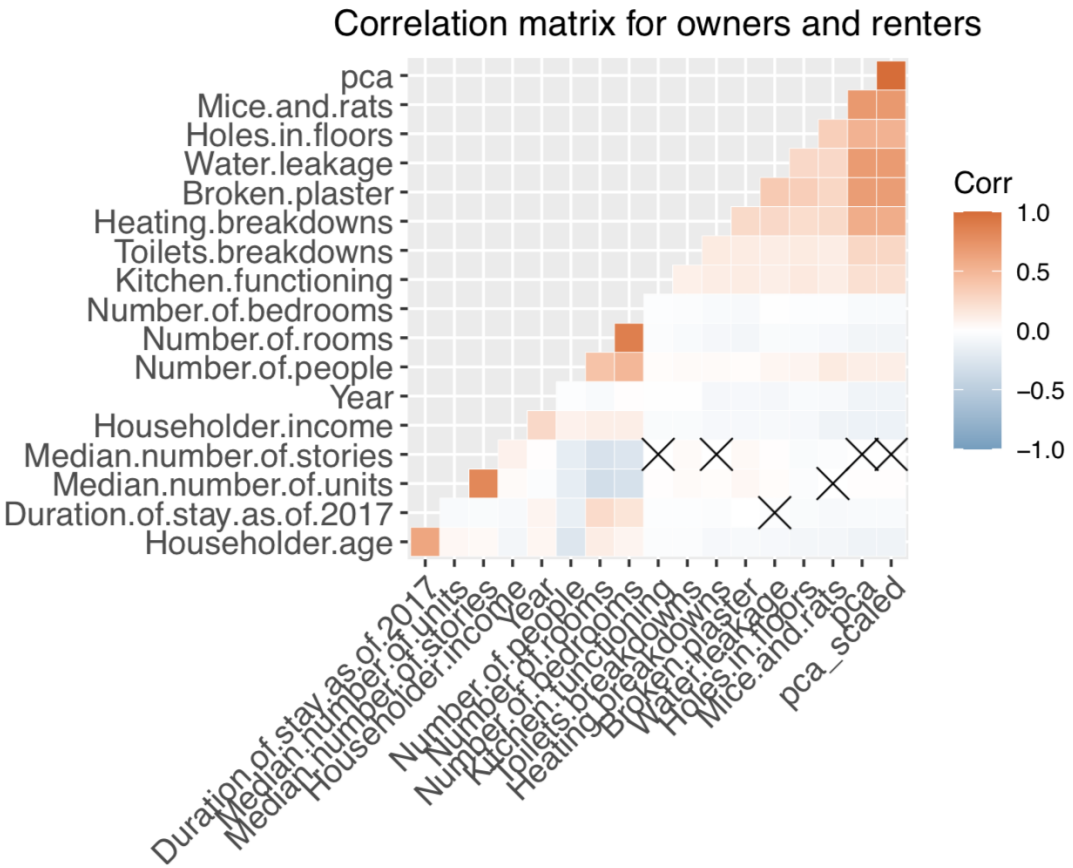
- Renaming variables appropriately
- Replacing categorical variables with dummy variables
- Omitting NA entries
- Changing categorical range variables to medians
- Splitting into two: renters and owners

Variables in the Dataset

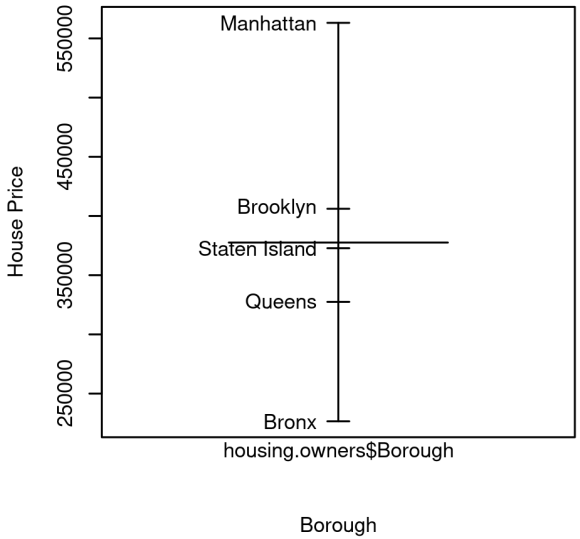
- Social factors e.g. sex, age, race
- Fixed features e.g. apartment size, density
- Physical features e.g. building condition, wear and tear

Correlation Matrix and PCA

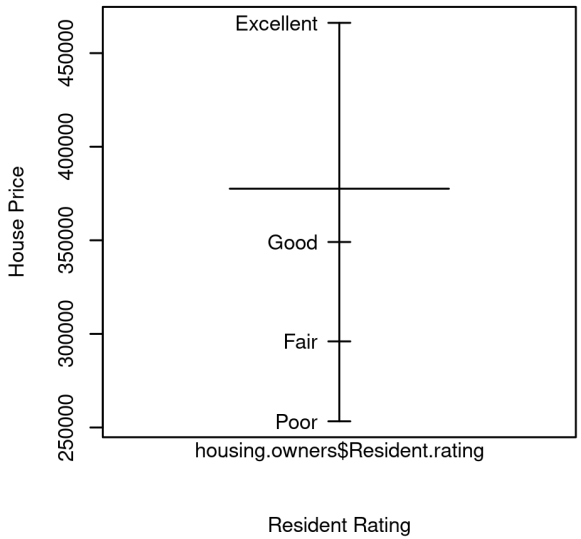
- Many fixed and physical features are correlated
- PCA is a good proxy for all the physical features



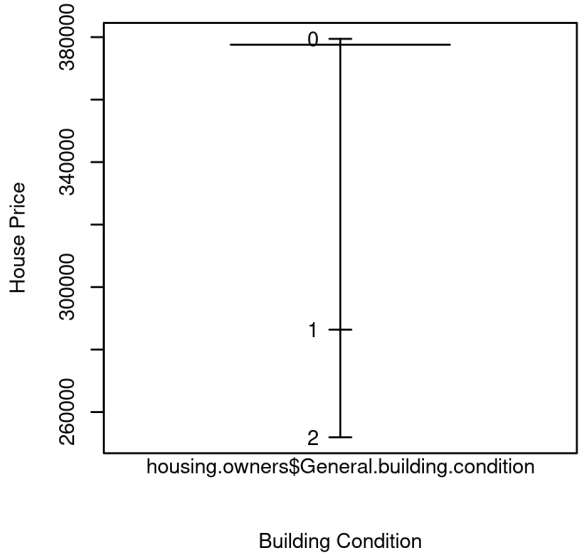
Factors to consider



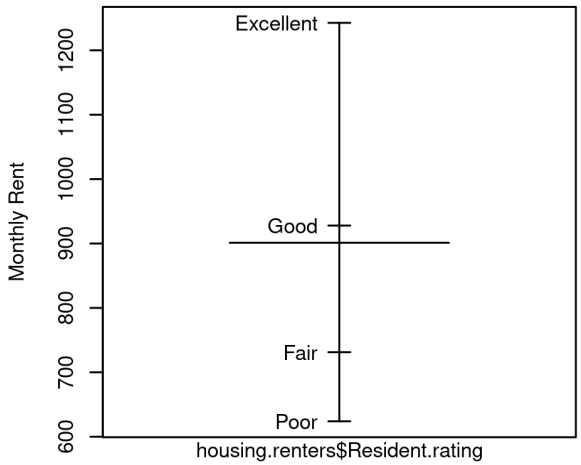
Borough



Resident Rating



Building Condition



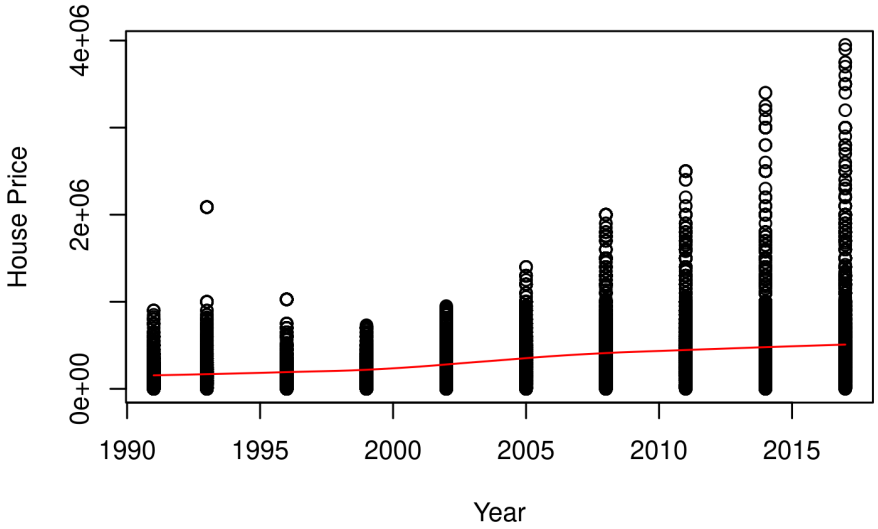
Section 3

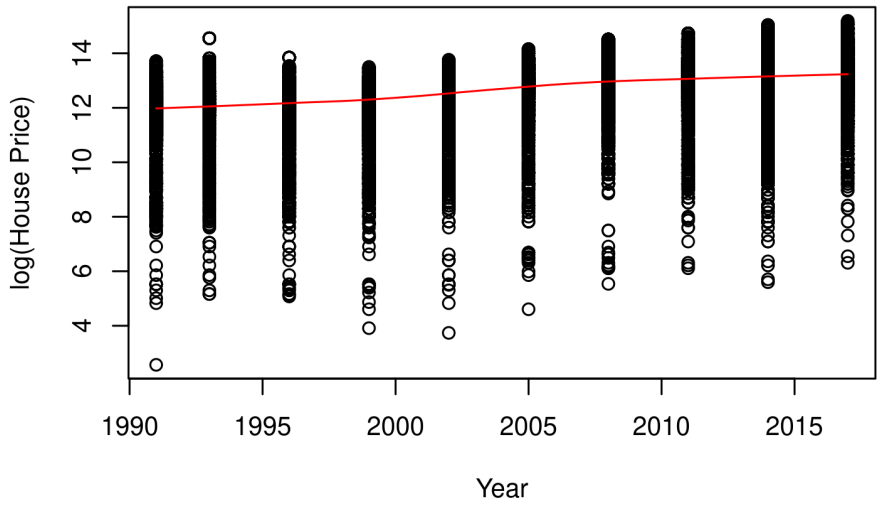
Model Choice

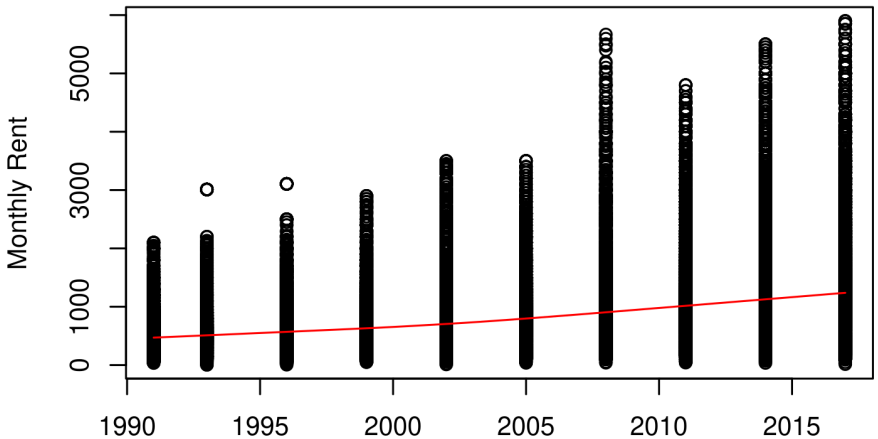
Initial thoughts on Model Selection

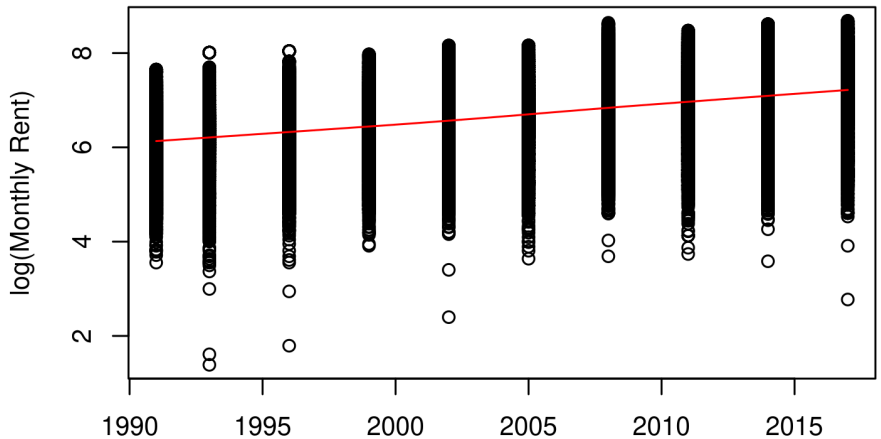
- Linear model is appropriate for continuous data
- Residual plots of base model alarming
- Transformation is necessary

Transformations of the response variable









Additional benfits of log-transformation

- Improvement in distribution of Variance
- Improvement in Residual plots
- Larger Adjusted R-squared values
- Eases interpretation
- Avoids over-fitting

Conclusion of Model Selection

- log(HousePrice) and log(MonthlyRent) as responses against linear predictors

Household Value Preliminary Model

```
own.full.lm <-
  lm(log(Household.value) ~ Householder.age + Householder.hispanic.
       Householder.race + Householder.income + Householder.female +
       Duration.of.stay.as.of.2017 + Year + Borough +
       Median.number.of.units + Median.number.of.stories +
       Number.of.rooms + Number.of.bedrooms + Plumbing.facilities +
       Kitchen.facilities + Resident.rating + Number.of.people +
       Severity.walls + Severity.windows + Severity.stairways +
       Severity.floors + Mice.and.rats + Broken.plaster +
       General.building.condition + Toilets.breakdowns +
       Heating.breakdowns + Kitchen.functioning +
       Holes.in.floors + Water.leakage,
     data = housing.owners)
```

Monthly Rent Preliminary Model

```
rent.full.lm <-
  lm(log(Monthly.rent) ~ Householder.age + Householder.hispanic.orig
       Householder.race + Householder.income + Householder.female +
       Duration.of.stay.as.of.2017 + Year + Borough +
       Median.number.of.units + Median.number.of.stories +
       Number.of.rooms + Number.of.bedrooms + Plumbing.facilities +
       Kitchen.facilities + Resident.rating + Number.of.people +
       Severity.walls + Severity.windows + Severity.stairways +
       Severity.floors + Mice.and.rats + Broken.plaster +
       General.building.condition + Toilets.breakdowns +
       Heating.breakdowns + Kitchen.functioning +
       Holes.in.floors + Water.leakage,
     data = housing.renters)
```

Section 4

Model Fitting

Selecting subset of useful predictors

- Aim: find subset of predictors that attain balance of fit and simplicity
- Possible due to results of ANOVA tests and factor box plots
- AIC used to find best subset of predictors

Stepwise Forward Selection (Homeowner model)

```
forward.AIC.own <- stepAIC(own.intercept.lm,
                           scope = list(lower = own.intercept.
                                        upper = own.full.lm),
                           direction = 'forward')
```

	Df	Sum of Sq	RSS	AIC
<none>			9910.6	-6732.8
+ Water.leakage	1	1.25415	9909.4	-6732.8
+ Broken.plaster	1	1.05449	9909.6	-6732.5
+ Toilets.breakdowns	1	0.58104	9910.1	-6731.7
+ Householder.age	1	0.41451	9910.2	-6731.5
+ pca	1	0.17394	9910.5	-6731.1
+ pca_scaled	1	0.17394	9910.5	-6731.1
+ Plumbing.facilities	1	0.13057	9910.5	-6731.0
+ Holes.in.floors	1	0.00410	9910.6	-6730.9
+ Severity.stairways	2	0.68985	9909.9	-6729.9
+ Severity.walls	2	0.48914	9910.1	-6729.6
+ Severity.windows	3	0.51639	9910.1	-6727.6

Stepwise Forward Selection (Renters Model)

	Df	Sum of Sq	RSS	AIC
+ Year	1	8163.8	26044	-8101.0
+ Resident.rating	3	2295.4	31912	-1343.2
+ Borough	4	1002.3	33205	-21.0
+ Mice.and.rats	1	790.5	33417	184.4
+ General.building.condition	2	539.2	33668	435.3
+ Broken.plaster	1	454.5	33753	516.9
+ Heating.breakdowns	1	344.1	33863	625.4
+ Severity.windows	3	343.2	33864	630.3
+ Severity.stairways	2	285.5	33922	684.9
+ Holes.in.floors	1	274.7	33933	693.5
+ Water.leakage	1	259.3	33948	708.5
+ Severity.floors	2	175.8	34032	792.2
+ Toilets.breakdowns	1	125.0	34082	839.8
+ Kitchen.functioning	1	118.0	34089	846.6
+ Severity.walls	2	53.0	34154	911.9
+ Number.of.rooms	1	25.4	34182	936.7
+ Plumbing.facilities	1	22.4	34185	939.7
+ Kitchen.facilities	1	9.8	34197	951.9
<none>			34207	959.4

Selection - Adjusted R^2

- Measures goodness-of-fit but penalizes models with more variables
- Used to find the subset models with the highest Adjusted R^2

Adjusted R^2: Homeowners Model

```
> best.adj.r2.owners
```

```
Number of Parameters:  19 Adj R2:  0.4152284
```

```
Parameters include Year, Number.of.rooms, BoroughB,
General.building.condition, Severity.stairways,
Kitchen.functioning, Heating.breakdowns, Mice.and.rats,
Holes.in.floors, Broken.plaster, Resident.rating,
Median.number.of.units, Median.number.of.stories
```

- Adjusted R^2 using `regsubsets`: 0.4152284
- Adjusted R^2 using `stepAIC`: 0.4152284

Adjusted R^2: Monthly Rent Model

```
> best.adj.r2.owners
```

```
Number of Parameters:  28 Adj R2:  0.3165994
```

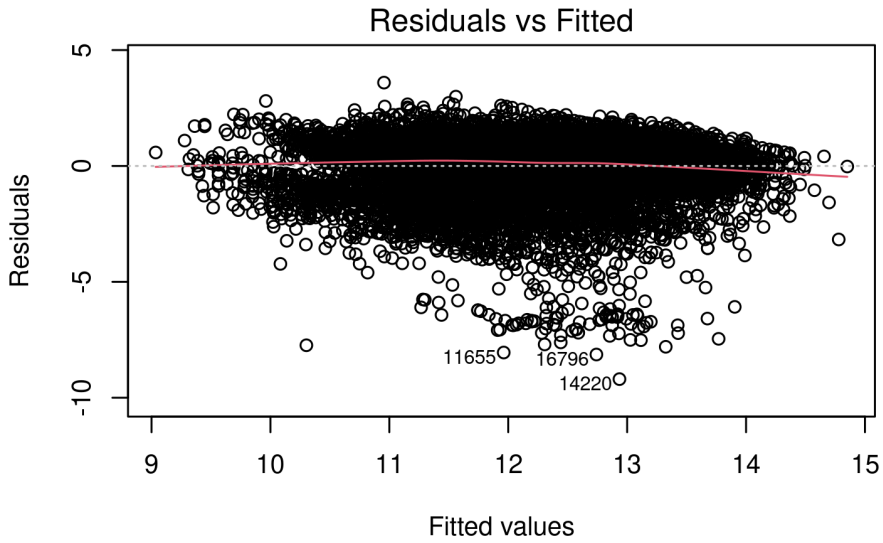
```
Parameters include Year, Number.of.rooms, BoroughB,
General.building.condition, Severity.walls,
Severity.windows, Severity.stairways, Severity.floors,
Toilets.breakdowns, Kitchen.functioning, Mice.and.rats,
Broken.plaster, Water.leakage, Resident.rating,
Plumbing.facilities, Kitchen.facilities,
Median.number.of.units, Median.number.of.stories
```

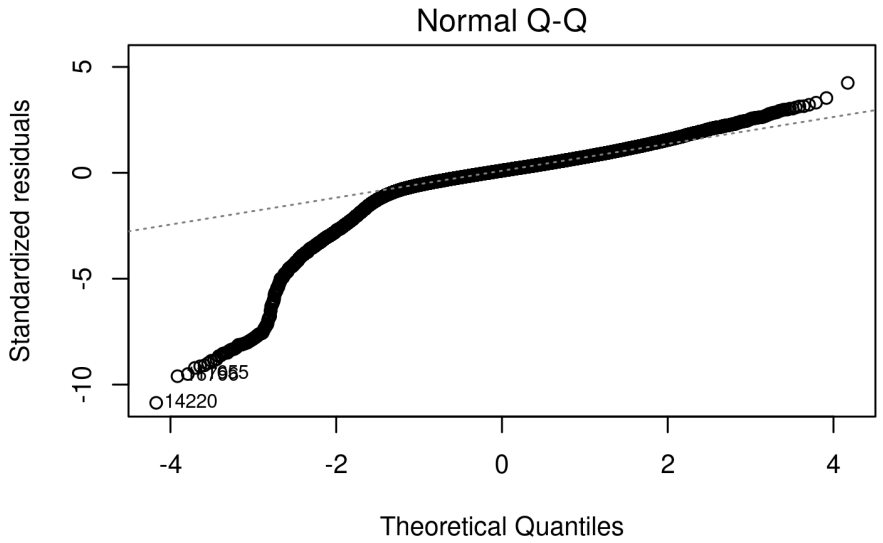
- Adjusted R^2 using `regsubsets`: 0.3165994
- Adjusted R^2 using `stepAIC`: 0.3165919

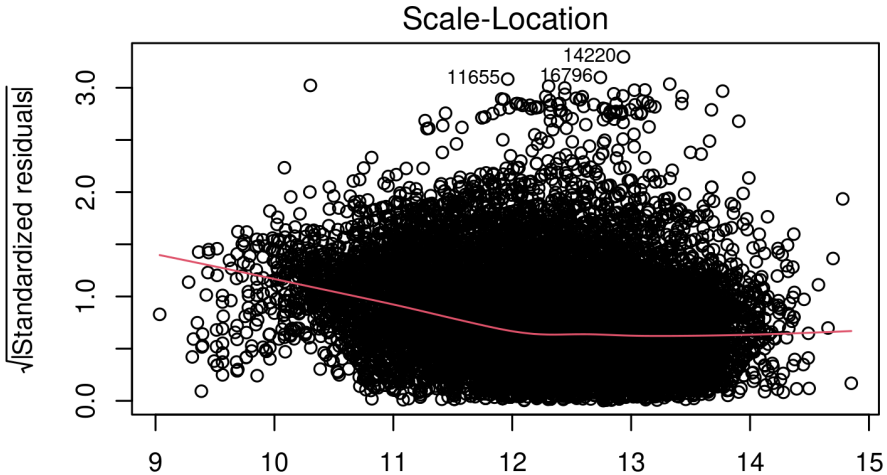
Section 5

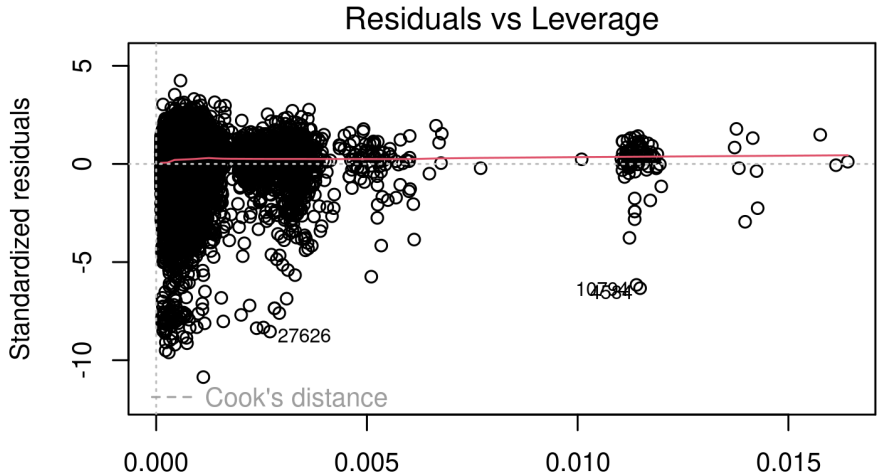
Diagnostics

Diagnostics for the Homeowners Model

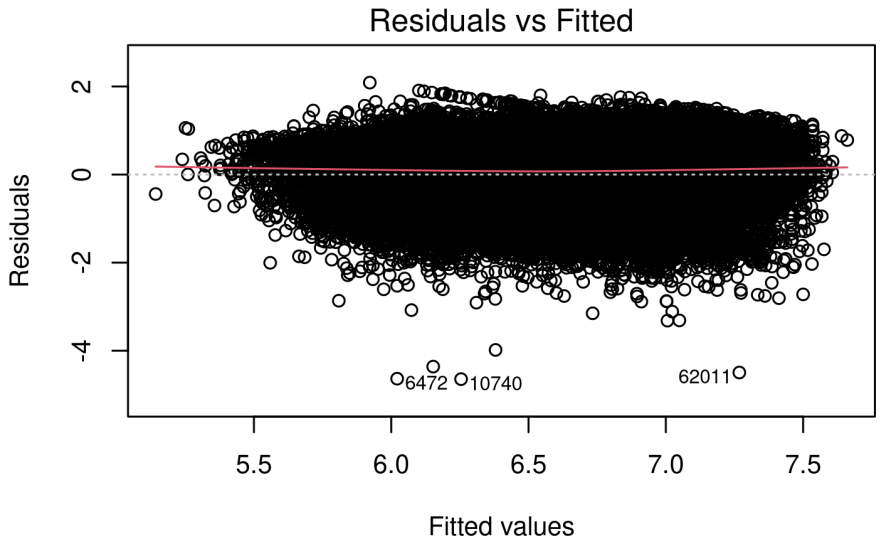


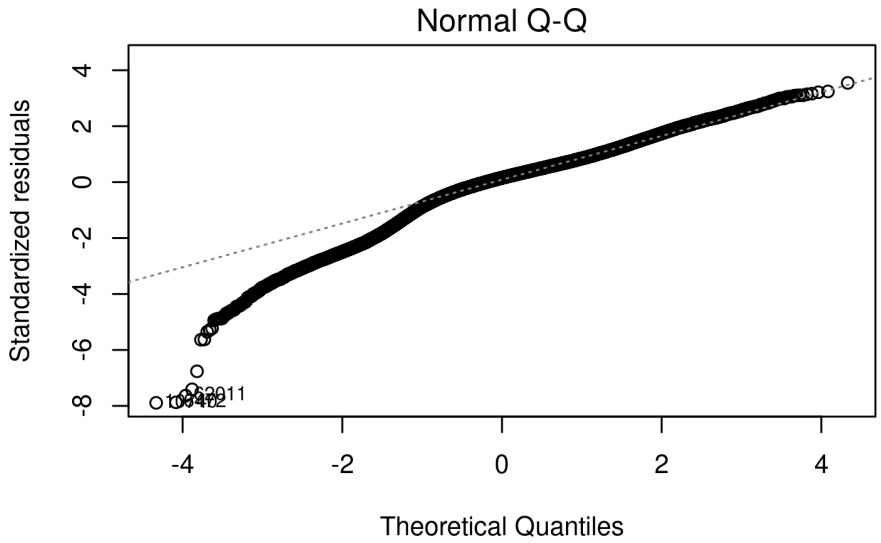


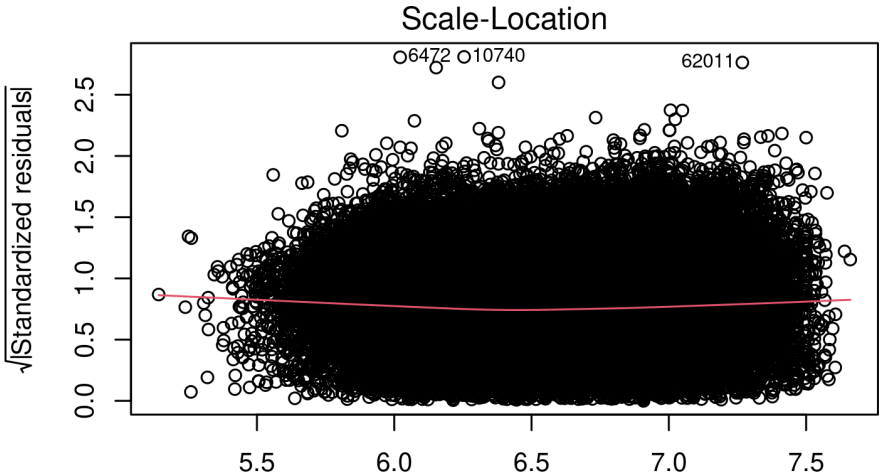


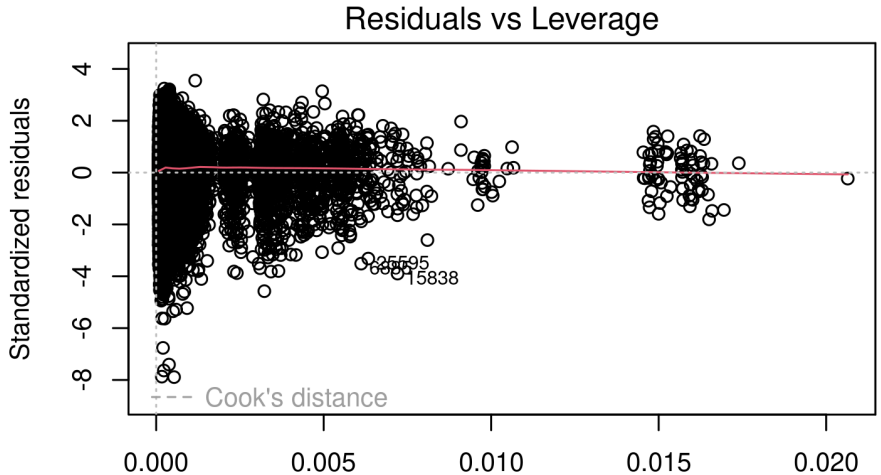


Diagnostics for the Monthly Rent Model









Section 6

Model Assessment

Model Assessment Techniques

- Predictive ability of the model can be assessed using Cross-Validation methods
- Run model on a subset of the data, and compare its predictions to observed data
- Can be computationally expensive for large datasets
- K-fold cross validation can be used to reduce computational expense

Model Limitations

- Predictive accuracy is an issue for varied datasets
- New York City housing data is very noisy, due to the city's large variation in housing types and qualities
- Wide confidence bands for point estimates

Section 7

Conclusion

Conclusion

- Both rent and housing prices rely on both fixed physical factors and economic factors
  - e.g. number of rooms, building condition, location
- Rental prices have more micro-level determinants than housing prices
  - e.g. damage severity, general wear and tear

Conclusion

- Used our predictive model to predict rental prices for home owners through the years
- Our predictive model showed on average, rent as a proportion of house prices has drastically decreased from 1991 to 2017
- House prices 463 $\times$ monthly rent in 1991, compared to 768 $\times$ in 2017
  - Standard errors of 23.8 and 39.9 respectively
  - 95% confidence intervals do not overlap
- Clear evidence that New York's rent control measures are indeed effective in dampening rent prices compared to real estate