House Prices Prediction using Linear Regression Model

IOE 591 Final Project

Jinxiang Ma, Shucen Zhao, Tianze Qu, Zitong Li

Group 11, December 2023

Overview

Dataset Introduction

In the realm of real estate, determining the factors that contribute to the pricing of residential properties is a complex undertaking. This dataset, encompassing 18 distinct predictors, showing the multifaceted nature of house price determination.

- Spatial and Locational Features
- Quality of Living Metrics
- Educational and Socio-Economic Indicators
- Infrastructure and Services
- Environmental Considerations
- Recreational and Green Spaces

This dataset aims to empower potential customers, urban planners and policymakers with a deeper understanding of the factors that collectively influence house prices. Our objective is selecting significant predictors to build robust linear regression model which plays an significant role for making informed decision.

Dataset Description

The following table describe the predictor in the house price dataset.

crime_rate	Crime rate in that neighborhood	dist4	Distance from employment hub 4 (miles)
resid_area	Proportion of residential area in the town	teachers	Number of teachers per thousand population
air_qual	Quality of air in that neighborhood	poor_prop	Proportion of poor population in the town
room_num	Average number of rooms in houses	n_hos_beds	Number of hospital beds per 1000 population in the town
age	How old is the house construction in years	n_hot_rooms	Number of hotel rooms per 1000 population in the town
dist1	Distance from employment hub 1 (miles)	rainfall	The yearly average rainfall (centimeters)
dist2	Distance from employment hub 2 (miles)	parks	Proportion of land assigned as parks in the town
dist3	Distance from employment hub 3 (miles)		

Table: Numerical Variables

Airport	Is there an airport in the city? (Yes/No)
Waterbody	What type of natural fresh water source is there in the city (lake/ river/ both/ none)
bus_ter	Is there a bus terminal in the city? (Yes/No)

Table: Categorical Variables

There are 498 observations in the house price dataset. The response variable is house price(per \$10k).

EDA

1. Set categorical variables:
   - two level - airport, bus terminal (Yes, No)
   - four level - waterbody (None, River, Lake, River and Lake)
2. Remove useless data: All "YES" in predictor bus terminal
3. Remove missing value (NA) from dataframe
4. Remove truncated data: R^2 increases

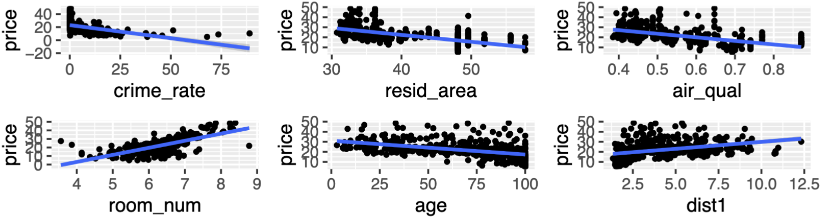


FIGURE: Plots of raw data

EDA

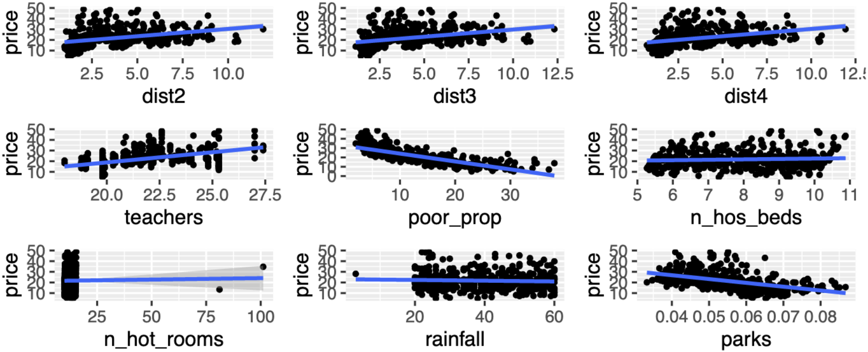
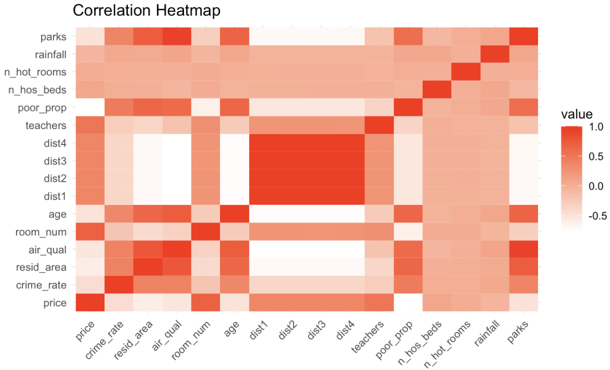


FIGURE: Plots of raw data

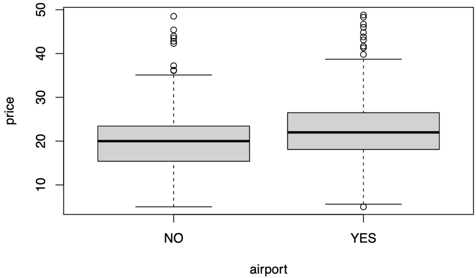
Correlation

1. dist1, dist2, dist3, and dist4 have pretty high covariance (close to 1)
   - Consider dropping dist2, dist3 and dist4.
2. parks and air quality highly correlated (close to 1)
   - Consider dropping parks



Categorical Predictors Analysis: Airport

1. Boxplot: response variable price in terms of two groups categorical predictor - Is there an airport in the city? (Yes/No)



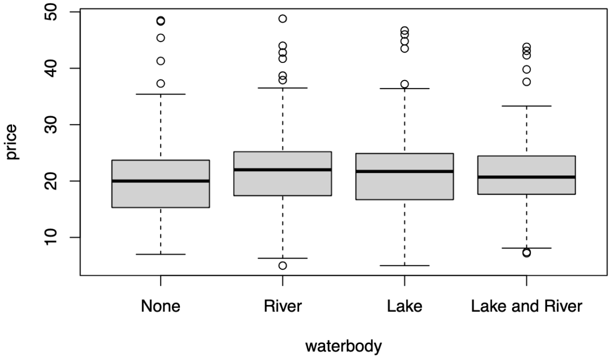
2. t-test for the difference between groups

- $t = -2.9851$, $df = 480$, p-value $= 0.00298$
- 95 percent confidence interval: (-3.5147529, -0.7243667)

Conclusion: true difference in means between group (airport) NO and YES is not equal to 0

Categorical Predictors Analysis: Waterbody

1. Boxplot: response variable price in terms of four groups categorical predictor: type of waterbody in the city (lake/ river/ both/ none)



2. Set interaction term between numerical and categorical predictors
   - - Reference level: AirportNO and waterbodyNone
   - - Based on significant level 0.05, room number:airport is significant

Variable Selection - Testing-Based

1. Backward Elimination (significant level 0.05)

```
##                                    Elimination Summary
## -------------------------------------------------------------------------------------------
##        Variable                                  Adj.
## Step          Removed              R-Square    R-Square     C(p)         AIC          RMSE
## -------------------------------------------------------------------------------------------
##    1    waterbody.Lake.and.River     0.7681      0.7607    15.0486    2681.2868    3.8347
##    2    waterbody.River              0.7681      0.7611    13.1450    2679.3867    3.8310
##    3    rainfall                     0.7678      0.7613    11.8069    2678.0722    3.8296
##    4    n_hot_rooms                  0.7674      0.7615    10.4771    2676.7652    3.8283
##    5    waterbody.Lake               0.7666      0.7611    10.1581    2676.4990    3.8311
##    6    n_hos_beds                   0.7655      0.7605    10.3180    2676.7177    3.8358
## -------------------------------------------------------------------------------------------
```

Figure: Summary of Backward Elimination

2. Forward Selection (significant level 0.05)

```
##
##                                Selection Summary
## -----------------------------------------------------------------------------------------
##         Variable                                Adj.
## Step         Entered             R-Square    R-Square      C(p)          AIC         RMSE
## -----------------------------------------------------------------------------------------
##    1     poor_prop                 0.5821      0.5812    360.1834    2937.2306     5.0725
##    2     room_num                  0.6621      0.6607    201.7272    2836.8095     4.5660
##    3     teachers                  0.7158      0.7140     96.1099    2755.4641     4.1922
##    4     air_qual                  0.7278      0.7256     73.8717    2736.5235     4.1064
##    5     dist1                     0.7474      0.7447     36.6909    2702.6228     3.9604
##    6     crime_rate                0.7547      0.7516     23.9579    2690.4000     3.9065
##    7     resid_area                0.7587      0.7551     18.0085    2684.5478     3.8789
##    8     age                       0.7615      0.7575     14.3859    2680.9158     3.8604
##    9     room_num.airportYES       0.7634      0.7589     12.6273    2679.1138     3.8493
##   10     airport.YES               0.7655      0.7605     10.3180    2676.7177     3.8358
##   11     n_hos_beds                0.7666      0.7611     10.1581    2676.4990     3.8311
##   12     waterbody.Lake            0.7674      0.7615     10.4771    2676.7652     3.8283
## -----------------------------------------------------------------------------------------
```

Figure: Summary of Forward Selection

Variable Selection - Testing-Based

3. Stepwise Selection (significant level 0.05)

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	poor_prop	addition	0.582	0.581	360.1830	2937.2306	5.0725
2	room_num	addition	0.662	0.661	201.7270	2836.8095	4.5660
3	teachers	addition	0.716	0.714	96.1100	2755.4641	4.1922
4	air_qual	addition	0.728	0.726	73.8720	2736.5235	4.1064
5	dist1	addition	0.747	0.745	36.6910	2702.6228	3.9604
6	crime_rate	addition	0.755	0.752	23.9580	2690.4000	3.9065
7	resid_area	addition	0.759	0.755	18.0080	2684.5478	3.8789
8	age	addition	0.761	0.757	14.3860	2680.9158	3.8604
9	room_num.airportYES	addition	0.763	0.759	12.6270	2679.1138	3.8493
10	room_num.airportYES	removal	0.761	0.757	14.3860	2680.9158	3.8604
11	airport.YES	addition	0.763	0.758	13.4560	2679.9551	3.8526
12	airport.YES	removal	0.761	0.757	14.3860	2680.9158	3.8604

Figure: Summary of Stepwise Selection

Variable Selection - Criterion-Based

1. Akaike information criterion (AIC)
   - AIC $= n\ln(RSS/n)+2(p+1)$
   - Selected Model:

```
Step:  AIC=1306.64
price ~ crime_rate + resid_area + air_qual + room_num + age +
    dist1 + teachers + poor_prop + airport.YES + n_hos_beds +
    room_num.airportYES

                     Df Sum of Sq    RSS    AIC
<none>                            6898.3 1306.6
- n_hos_beds          1     31.83 6930.1 1306.9
- airport.YES         1     63.43 6961.7 1309.0
- room_num.airportYES 1     75.83 6974.1 1309.9
- age                 1    103.27 7001.6 1311.8
- resid_area          1    103.78 7002.1 1311.8
- crime_rate          1    244.69 7143.0 1321.4
- air_qual            1    278.87 7177.2 1323.7
- dist1               1    746.77 7645.0 1354.2
- room_num            1    766.93 7665.2 1355.5
- poor_prop           1   1020.27 7918.6 1371.1
- teachers            1   1221.89 8120.2 1383.2

Call:
lm(formula = price ~ crime_rate + resid_area + air_qual + room_num +
    age + dist1 + teachers + poor_prop + airport.YES + n_hos_beds +
    room_num.airportYES, data = hp_data)

Coefficients:
        (Intercept)           crime_rate           resid_area             air_qual             room_num                  age                dist1
            3.81829             -0.09627             -0.12288            -12.81830              3.31533             -0.02858             -1.05308
           teachers            poor_prop          airport.YES           n_hos_beds  room_num.airportYES
            0.86919             -0.35912             -7.09212              0.17581              1.23765
```

2. Bayes information criterion (BIC)
   - $BIC = n\ln(RSS/n) + (p+1)\ln(n)$
   - Selected Model:

```
Step:  AIC=1348.11
price ~ crime_rate + resid_area + air_qual + room_num + dist1 +
    teachers + poor_prop

             Df Sum of Sq    RSS    AIC
<none>                    7131.8 1348.1
- resid_area  1    117.13 7249.0 1349.8
- crime_rate  1    215.12 7347.0 1356.3
- air_qual    1    424.99 7556.8 1369.8
- dist1       1    704.86 7836.7 1387.4
- teachers    1   1312.70 8444.5 1423.4
- poor_prop   1   1421.29 8553.1 1429.5
- room_num    1   1836.76 8968.6 1452.4

Call:
lm(formula = price ~ crime_rate + resid_area + air_qual + room_num +
    dist1 + teachers + poor_prop, data = hp_data)

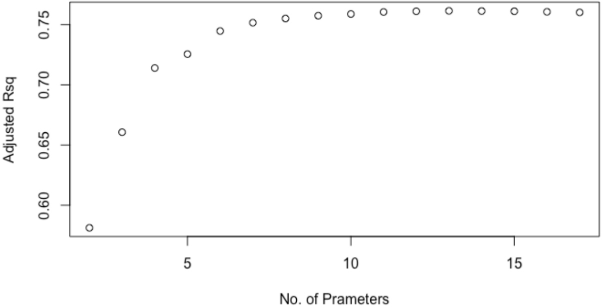
Coefficients:
(Intercept)   crime_rate   resid_area     air_qual     room_num        dist1     teachers    poor_prop
    1.18069     -0.08956     -0.12973    -15.21682      3.87622     -0.94369      0.89594     -0.39570
```

Conclusion: Since BIC penalized larger model more heavily than AIC, it tends to select fewer predictors.

Variable Selection - Criterion-Based

3. Adjusted R^2

   - Plot of No. parameters vs. Adjusted R^2

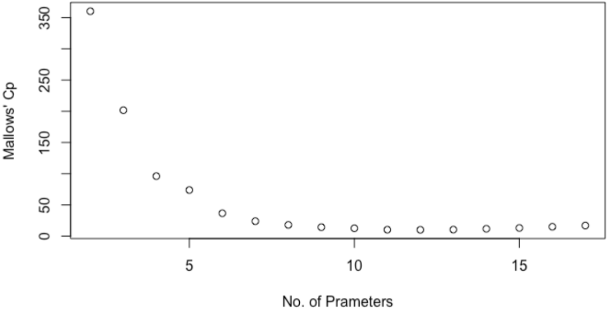


   - Selected Model (12 predictors) with largest adjusted R^2

```
       crime_rate               resid_area                 air_qual                 room_num                      age                    dist1
              "*"                      "*"                      "*"                      "*"                      "*"                      "*"
         teachers                poor_prop              airport.YES               n_hos_beds              n_hot_rooms          waterbody.River
              "*"                      "*"                      "*"                      "*"                      " "                      " "
  waterbody.Lake waterbody.Lake.and.River                 rainfall        room_num.airportYES
              "*"                      " "                      " "                      "*"
```

4. Mallows' C_p

    - Plot of No. parameters vs. Mallows' C_p



   - Selected Model (11 predictors) with smallest Mallows' C_p

```
      crime_rate            resid_area              air_qual              room_num                   age                 dist1
             "*"                   "*"                   "*"                   "*"                   "*"                   "*"
        teachers             poor_prop           airport.YES           n_hos_beds           n_hot_rooms       waterbody.River
             "*"                   "*"                   "*"                   "*"                   " "                   " "
  waterbody.Lake waterbody.Lake.and.River              rainfall   room_num.airportYES
             " "                   " "                   " "                   "*"
```

Variable Selection Summary

1. Summary of result: Several selection methods give very similar fit

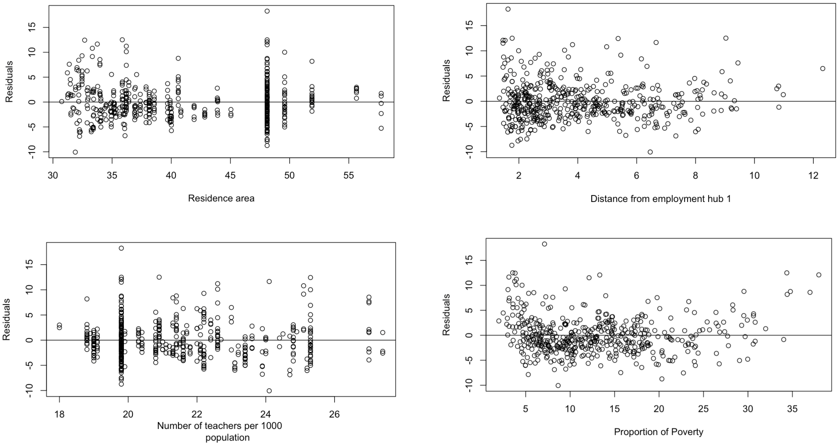
Methods	Crime_rate	resid_area	air_qual	room_num	age	dist1	teachers	poor_prop
Backward	√	√	√	√	√	√	√	√
Forward	√	√	√	√	√	√	√	√
Stepwise	√	√	√	√	√	√	√	√
AIC	√	√	√	√	√	√	√	√
BIC	√	√	√	√		√	√	√
Adjust R2	√	√	√	√	√	√	√	√
Mallows' Cp	√	√	√	√	√	√	√	√

Methods	airportYes	n_hos_beds	n_hot_ room	waterbody river	waterbody lake	waterbody lake&river	rainfall	room_num: airportYes
Backward	√							√
Forward	√	√			√			√
Stepwise								
AIC	√	√						√
BIC								
Adjust R2	√	√			√			√
Mallows' Cp	√	√						√

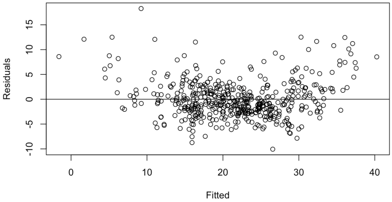
2. Similar fit model leads to similar fit, the data are not ambiguous.
3. Generally, criterion-based methods are preferred
4. Based on preference of criterion-based methods and similar conclusions from different models, we choose AIC model for further analysis.

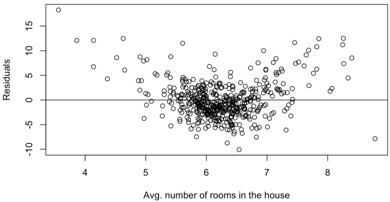
DIAGNOSTICS - NON-LINEARITY AND CONSTANT VARIANCE

Heteroskedasticity:



No distinct patterns occurred in the residual plots.

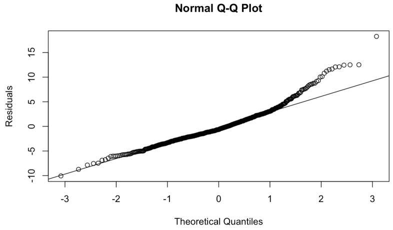




Conjecture: The number of rooms in the house has a significant impact on the price.

Normality Check

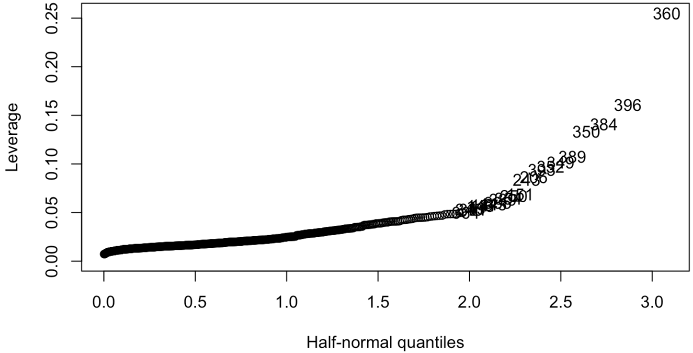
1. QQ Plot and Histogram of Residuals



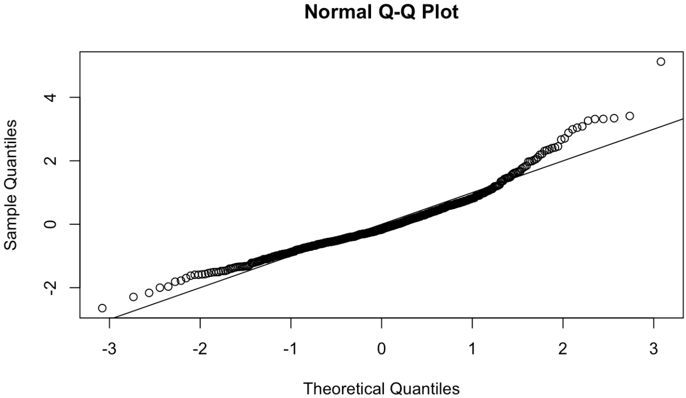
2. Shapiro-Wilk Test
   - W = 0.95024, p-value = 1.169e-11;
   - Normality assumption failed.

Find Large leverage points

Hat matrix: $H = X(X^TX)^{-1}X^T$.
Leverage: $h_i = H_{ii}$.
Rule of thumb: Leverages greater than $2(p+1)/n$ are considered high.



Use Studentized residuals to find outliers

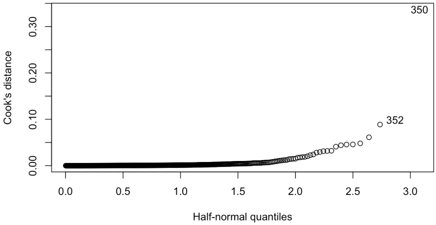


Find outlier(s) with Bonferroni correction:

- Point 350.

Find Influential Points

1. Compute Cook's distance



2. Compare coefficients of models

- Original Model

(Intercept)	crime_rate	resid_area	air_qual	room_num	age	dist1	teachers
3.818	-0.096	-0.123	-12.818	3.315	-0.029	-1.053	0.869
poor_prop	airport.YES	n_hos_beds	room_num.airportYES				
-0.359	-7.092	0.176	1.238				

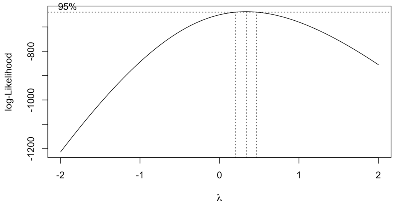
- Model without Influential Points

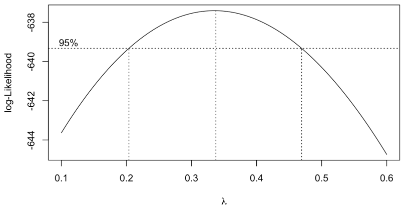
(Intercept)	crime_rate	resid_area	air_qual	room_num	age	dist1	teachers
-1.450	-0.106	-0.118	-13.527	4.079	-0.035	-1.027	0.877
poor_prop	airport.YES	n_hos_beds	room_num.airportYES				
-0.275	-8.403	0.148	1.433				

Summary of Diagnostics

1. Non-linearity assumption is almost obeyed.
2. There is some heteroscedasticity problem shown in the residual plots. In the latter section, we will adopt weighted least squares and robust regression to try to solve it.
3. Some unusual points are found. Coefficients don't change too much after removal of them.
4. The dataset doesn't follow the normal distribution. We are going to use the Box-Cox method to handle it.

Box-Cox method

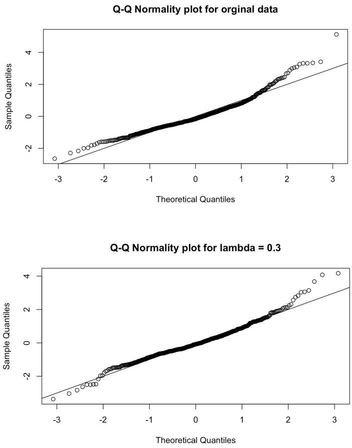




Box-Cox transformation is needed.

- The optimal λ: 0.3434343.

Box-Cox method



Weighted least squares

Since the unequal variance occurs in diagnostics, we applied iteratively reweighted least squares (IRWLS) to solve this problem (based on AIC model selected before).

```
Intercept: 3.818291
     crime_rate       resid_area         air_qual         room_num
    -0.09627480      -0.12288455     -12.81830336       3.31532896
            age            dist1         teachers        poor_prop
    -0.02857932      -1.05308371       0.86918735      -0.35912101
    airport.YES       n_hos_beds room_num.airportYES
    -7.09211736       0.17580648       1.23764740
Number of iterations: 1
```

Robust regression - Least Squares

Choose the loss function as $L(z) = z^2$,

```
Call:
lm(formula = price ~ crime_rate + resid_area + air_qual + room_num +
    age + dist1 + teachers + poor_prop + airport.YES + n_hos_beds +
    room_num.airportYES, data = hp_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.062  -2.345  -0.644   1.915  18.274

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.81829    4.53359   0.842  0.40009
crime_rate           -0.09627    0.02358  -4.083 5.22e-05 ***
resid_area           -0.12288    0.04621  -2.659  0.00810 **
air_qual            -12.81830    2.94071  -4.359 1.61e-05 ***
room_num              3.31533    0.45864   7.229 1.99e-12 ***
age                  -0.02858    0.01077  -2.653  0.00826 **
dist1                -1.05308    0.14764  -7.133 3.74e-12 ***
teachers              0.86919    0.09526   9.124  < 2e-16 ***
poor_prop            -0.35912    0.04307  -8.338 8.45e-16 ***
airport.YES          -7.09212    3.41146  -2.079  0.03817 *
n_hos_beds            0.17581    0.11939   1.473  0.14154
room_num.airportYES   1.23765    0.54452   2.273  0.02348 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Robust regression - Least Absolute Deviations

Choose the loss functions as $L(z) = |z|$,

```
Call: rq(formula = price ~ crime_rate + resid_area + air_qual + room_num +
    age + dist1 + teachers + poor_prop + airport.YES + n_hos_beds +
    room_num.airportYES, data = hp_data)

tau: [1] 0.5

Coefficients:
                    coefficients lower bd  upper bd
(Intercept)          -1.18281    -17.24150 14.61471
crime_rate           -0.12283     -0.13847 -0.07464
resid_area           -0.05077     -0.14035  0.02719
air_qual            -12.36123    -17.76264 -7.04860
room_num              3.70837      1.83361  5.91333
age                  -0.02698     -0.04739 -0.01024
dist1                -0.78087     -1.13053 -0.59252
teachers              0.77480      0.66089  0.95947
poor_prop            -0.32565     -0.41542 -0.23272
airport.YES          -7.61772    -19.64258  7.71649
n_hos_beds            0.07557     -0.10851  0.29735
room_num.airportYES   1.33030     -1.13276  3.31881
```

Robust regression - Huber's Method

Combining the above two methods together (apply LS when z is close to zero and apply LAD when z is far away from zero),

```
Call: rlm(formula = price ~ crime_rate + resid_area + air_qual + room_num +
    age + dist1 + teachers + poor_prop + airport.YES + n_hos_beds +
    room_num.airportYES, data = hp_data)
Residuals:
    Min      1Q  Median      3Q     Max
-9.7481 -2.0303 -0.2095  2.2308 20.2465

Coefficients:
                    Value    Std. Error t value
(Intercept)           1.6011   3.9512     0.4052
crime_rate           -0.1149   0.0206    -5.5892
resid_area           -0.0947   0.0403    -2.3501
air_qual            -11.8568   2.5629    -4.6262
room_num              3.5144   0.3997     8.7921
age                  -0.0313   0.0094    -3.3332
dist1                -0.9156   0.1287    -7.1160
teachers              0.8083   0.0830     9.7362
poor_prop            -0.3299   0.0375    -8.7876
airport.YES          -9.3530   2.9732    -3.1457
n_hos_beds            0.1220   0.1041     1.1729
room_num.airportYES   1.5964   0.4746     3.3639

Residual standard error: 3.08 on 470 degrees of freedom
```

Robust regression - Least Trimmed Squares

Least Trimmed Squares method will minimizes the sum of squares of q of n smallest residues.

```
(Intercept)          crime_rate          resid_area            air_qual
-1.21721013         -0.40551785         -0.16498579          2.86747041
   room_num                 age               dist1            teachers
 2.73108622         -0.06435727         -0.77025761          0.88879160
  poor_prop         airport.YES          n_hos_beds room_num.airportYES
-0.12900842          1.63044620          0.04930835         -0.25616092
```

Summary of Problem solving

Comparison between all the introduced methods.

Methods	Intercept	Crime_rate	resid_area	air_qual	room_num	age
Box-Cox (lambda = 0.3)	3.8366	-0.0191	-0.0089	-1.6099	0.2577	-0.0021
Box-Cox (lambda = 0.34)	3.9995	-0.021	-0.0104	-1.8118	0.2992	-0.0025
Box-Cox (lambda = 0.4)	4.2431	-0.0242	-0.0131	-2.1633	0.374	-0.0031
Weighted least squares	3.8183	-0.0963	-0.1229	-12.8183	3.3153	-0.0286
Least Squares	3.8183	-0.0963	-0.1229	-128183	3.3153	-0.0286
Least Absolute Deviations	-1.1828	-0.1228	-0.0508	-12.3612	3.7083	-0.027
Huber's Method	1.6011	-0.1149	-0.0947	-11.8568	3.5144	-0.0313
Least Trimmed Squares	-1.217	-0.4055	-0.165	2.8675	2.7311	-0.0644

Methods	dist1	teachers	poor_prop	airportYes	n_hos_beds	room_num: airportYes
Box-Cox (lambda = 0.3)	-0.1072	0.0917	-0.0552	-0.6073	0.0108	0.1102
Box-Cox (lambda = 0.34)	-0.1222	0.1042	-0.0614	-0.6983	0.0128	0.1264
Box-Cox (lambda = 0.4)	-0.1485	0.1263	-0.0721	-0.8611	0.0165	0.1553
Weighted least squares	-1.0531	0.8692	-0.3591	-7.0921	0.1758	1.2376
Least Squares	-1.0531	0.8692	-0.3591	-7.0921	0.1758	1.2376
Least Absolute Deviations	-0.7809	0.7748	-0.3257	-7.6177	0.0756	1.3303
Huber's Method	-0.9156	0.8083	-0.3299	-0.953	0.122	1.5964
Least Trimmed Squares	-0.7703	0.8888	-0.129	1.6304	0.0493	-0.2562

Summary of Problem solving

1. The Box-Cox transformation is needed. After comparison, the results of lambda choosing different values near the optimum are not sensitive. For convenience, we choose $\lambda = 0.34$ in the following steps.
2. The Box-Cox transformation is a function of logarithm, so the inverse transformation is exponential, both of them have good properties, which is very convenient.

Prediction

Based on our previous analysis, we selected the box-cox transformed AIC model($\lambda = 0.34$) with following predictors:

- crime_rate
- resid_area
- air_qual
- room_num
- age
- dist1
- teachers
- poor_prop
- airport.YES
- n_hos_beds
- room_num.airportYES

Prediction

There is 482 observations in the house price dataset. We selected 80% of observations as training data, and 20% of observations as test data.

```
##Training set RMSE
rmse((mod.lm$fit*0.34 + 1)^(1/0.34), tr$price)
```
```

[1] 3.423345
```

```
```{r}
#Test set RMSE
rmse((predict(mod.lm,newdata = te)*0.34 + 1)^(1/0.34), te$price)
```
```

```
[1] 3.589819
```

Figure: RMSE for Box-cox transformed AIC model

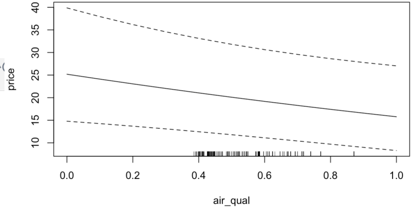
We fit the Box-cox transformed AIC model using our training data. Based on the R output, the training RMSE is 3.423345, and the test RMSE is 3.589819.

Prediction Interval

Question: What would be the 99% prediction interval for the house price for crime_rate = 0.5, resid_area = 30, air_qual = 0.5, room_num = 5, dist1 = 5, teacher = 20, poor_prop = 10, airport.YES = 1, n_hos_beds = 5, room_num.airportYES = 5?

```
(predict(mod.lm, x0, interval="prediction", level=0.99)*0.34 + 1)^(
```

```
       fit      lwr      upr
1 17.66325 10.75878 26.99399
```



Based on the result of prediction interval, the predicted house price for a community with given predictor values is approximately 17.66325. We are 99% confident that the house price with these predictor values will fall between approximately 10.75878 and 26.99399.

We would like to apply shrinkage method like ridge regression to prevent over-fitting during prediction. We need to find the best lambda that minimize the generalized cross-validation.

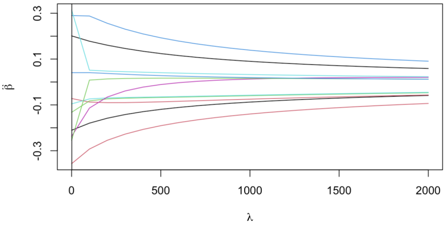
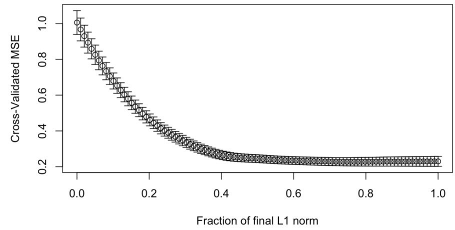


Figure: Select the best λ by LOOCV

From the above graph, we observed that the best lambda that yield smallest error is $\lambda = 0$, which implies that the penalty term has no effect and coefficient are the same as least squares regression.

Another shrinkage method we want to use is lasso regression. In order to control the strength of penalty we apply on absolute values of the coefficients, we need to find the fraction of the L1 penalty by cross-validation.




```{r}
cv.price$index[which.min(cv.price$cv)]
```

```
[1] 0.7474747
```

Based on the above graph, the best fraction of L1 norm that minimize cross validation error is 0.747474.

Shrinkage method - Lasso Regression

```{r}
pred.lars.price1 <- predict(lmod.price, trainX, s=0.7474747,
mode="fraction")
#Training RMSE - Lasso
print(rmse((pred.lars.price1$fit*0.34 + 1)^(1/0.34), tr$price))
```

```
[1] 3.393679
```

```{r}
testX = as.matrix(te[,-1])
pred.lars.price2 <- predict(lmod.price, testX, s=0.7474747,
mode="fraction")
#Test RMSE - Lasso
rmse((pred.lars.price2$fit*0.34 + 1)^(1/0.34), te$price)
```

```
[1] 3.711059
```

Figure: RMSE for Lasso

The test RMSE for Lasso Regression stands at 3.711059, which is higher than the test RMSE of the Box-Cox transformed AIC model. Consequently, Lasso Regression has no improvement on our model's performance.

Conclusion

Model	Test RMSE
Box-Cox Transformed Least Square model ($\lambda = 0.34$)	3.59
Ridge Regression(No effect, $\lambda = 0$)	3.59
Lasso Regression	3.71

Table: Model Comparison

In conclusion, we selected the Box-cox transformed least square model(with $\lambda \approx 0.34$) using AIC criterion and concluded that the following predictor have significant influence on house price, which includes: crime_rate, resid_area, air_qual, room_num, age, dist1, teachers, poor_prop, airport.YES, n_hos_beds, and the interaction term room_num*airport.YES.

That's all for our presentation.

Thank You!