Project Report

Hao Gao, Trisha Trupng, Karen Hovhannisyan, Chenhao Li, ZiYao Cui

December 1, 2018

General Background

We decided to do our project on housing prices. Buying a house is a big event for anybody and a home is usually ones biggest asset in life. Therefore, trying to find a house that fits one's needs and wants but also is within budget, is a problem that begs for a solution. It is also important to be able to calculate the sale price of a home based on certain features, so that buyers can know the cost of their desired attributes. Because a sale price is a reflection of all aspects of the house it is hard to discern at first glance how significant the various factors are. Creating a regression model to achieve this insight would surely be useful for home buyers and the real estate world so they can forecast the market. It is also important for students our age to know this because of how soon we will be entering the real world and looking for houses of our own.

Data Outline

Our dataset contains about 2274 observations of housing price with at first 28 explanatory variables describing the aspect of the house. Those variables include:
Lotfront: Linear feet of street connected to property
Lotarea: Lot size in square feet
YearBuilt: The year the house was built
YearRemodADD: The year the house was remoded
MasVnrArea: Masonry veneer area in square feet
BsmtFinSF1: Type 1 finished square feet
BsmtFinSF2: Type 2 finished square feet
.
.
.
Age: The house age
Agerem: Age of the house since re modeled
Sf: Total square footage
Fsf: First floor square footage
Ssf: Second floor square footage
Bath: Number of Bathroom
Bedroom: Number of Bedroom
Garage: Size of garage in square feet
The original data set came from Ames, Iowa where the Assessor's office of the city compiled date from all house sales in the city between 2006 and 2010, with 2930 observations and 80 different variables. We removed most of the discrete variables from the data set and considered the ones that seemed to be the most important. We also delete all samples those contains missing values, such as thosw with "NA" or "None" in the predictor entries.

Linear Regression Model

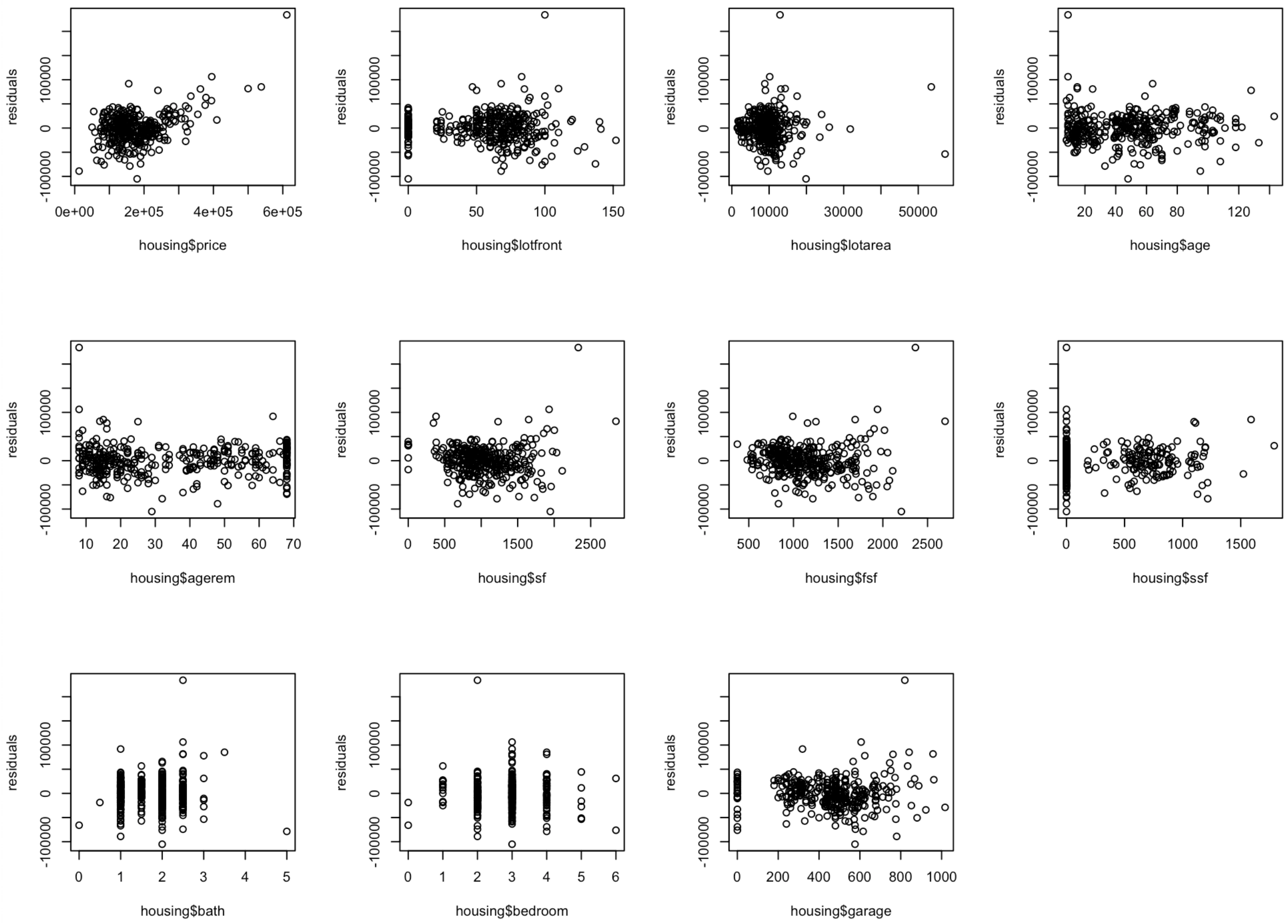
At first we applied the first order linear model to fit our data set:
$\hat{Y}_i = \mathbf{XB}$

where $\mathbf{B} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_{28} \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} X_{i1} \\ X_{i2} \\ ... \\ X_{i3} \\ X_{i28} \end{pmatrix}$. In the model selection and interpretation step, we will drop certain predictors that might have zero coefficients and seek the influence of iteractions.

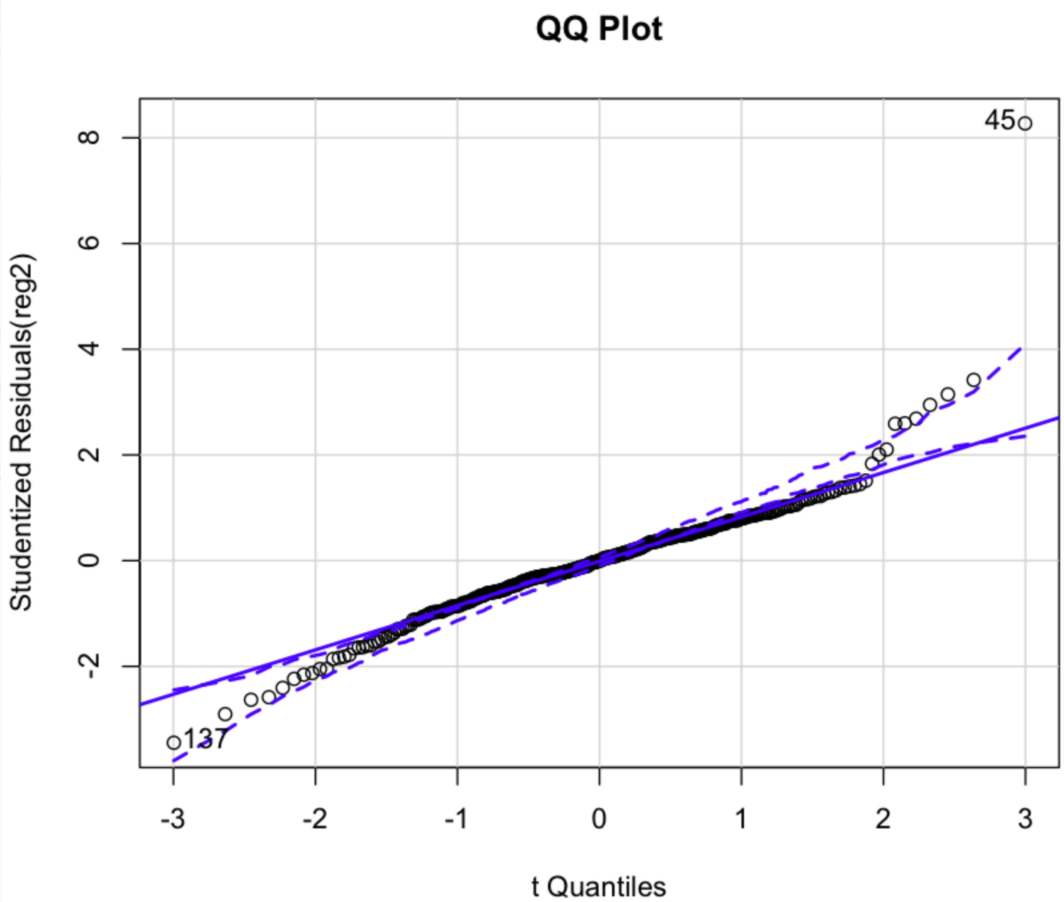
```
Call:
lm(formula = Price ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-537996  -21053   -2933   17926  379320

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.077e+05  1.285e+06  -0.084 0.933253
LotFrontage    1.042e+00  4.524e+01   0.023 0.981619
LotArea        6.222e-01  1.510e-01   4.122 3.90e-05 ***
YearBuilt      1.985e+02  5.755e+01   3.450 0.000572 ***
YearRemodAdd   5.565e+02  5.659e+01   9.835  < 2e-16 ***
MasVnrArea     5.644e+01  5.385e+00  10.481  < 2e-16 ***
BsmtFinSF1     1.463e+01  2.921e+00   5.007 5.94e-07 ***
BsmFinSF2      4.528e-01  5.462e+00   0.083 0.933937
TotalBsmtSF    3.765e+01  2.868e+00  13.131  < 2e-16 ***
BsmtFullBath   6.462e+03  2.343e+03   2.759 0.005852 **
BsmtHalfBath  -3.978e+03  3.711e+03  -1.072 0.283895
FullBath       1.880e+04  2.322e+03   8.098 9.09e-16 ***
HalfBath       1.200e+04  2.064e+03   5.812 7.07e-09 ***
Bedroom       -8.842e+03  1.523e+03  -5.806 7.29e-09 ***
Kitchen       -5.176e+04  4.925e+03 -10.510  < 2e-16 ***
TotRmsAbvGrd   1.296e+04  9.618e+02  13.475  < 2e-16 ***
Fireplaces     1.245e+04  1.543e+03   8.071 1.12e-15 ***
GarageYrBlt   -6.393e+01  6.854e+01  -0.933 0.351099
GarageCars     1.209e+04  2.630e+03   4.596 4.56e-06 ***
GarageArea     4.534e+01  9.045e+00   5.012 5.80e-07 ***
WoodDeckSf     3.274e+01  7.457e+00   4.391 1.18e-05 ***
OpenPorchSF    1.360e+01  1.406e+01   0.967 0.333656
EnclosedPorch  3.316e+01  1.436e+01   2.310 0.021003 *
`3SsnPorch`    1.783e+01  3.313e+01   0.538 0.590561
ScreenPorch    7.807e+01  1.488e+01   5.245 1.71e-07 ***
PoolArea      -4.338e+01  2.315e+01  -1.874 0.061037 .
Miscval       -1.500e+01  1.677e+00  -8.945  < 2e-16 ***
Mosold         2.162e+00  3.117e+02   0.007 0.994465
YrSold        -6.153e+02  6.395e+02  -0.962 0.336084
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39480 on 2245 degrees of freedom
  (656 observations deleted due to missingness)
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7756
F-statistic: 281.6 on 28 and 2245 DF,  p-value: < 2.2e-16
```

Assumtions

We assume that there exist a linear relashionship along the response variable "Price" and all other predictor variables. We also hope the residuals has constance variance. To check for such two assumptions, we plot residuals agains response variable, as well as residuals against some other variables in our model. Here are 11 of residual plots:



By oberseve the residual plot against response variable and residual plot agianst some other predictors, we notice that most plot doesn't show up any systematically pattern. however for some plot such as residuals against lotfront and residuals against price, there seems to have some pattern. The former implies some degree of non linearity, the latter implies some degree of non-constancy of variance. Such observations give us a reason to consider more terms in our model, meanwhile we also might want remove certain terms. The model selection selection recorded our method to remove predictors and include interactions.
We also want our residuals in the current model follow the normal distrubution, we use QQNorm plot for checking such assumption:



The QQ plot seems reasonably good. Hence we concluded that the residuals followed normal distribution.

Data transformations

We didn't preprocessing data in the original first order linear model. However in latter model we added some interaction terms, we centered each predictor variables to avoid internel computational errors.

Model Selection

We want to know if certain regression coefficients should be zero, which means weather certain predictor should be droped out from our model. We decided to use F-test with $\alpha = 0.01$on $H_0 : \beta_i = 0, H_1 \beta_i neq 0$ for each β_i. The reason we prefer F-test rather than T test is that the possible existance of muticuvlinear might affence the accuracy of T test. To achieve F test for each single coefficient, we use type 2 SS:

```
Analysis of Variance Table

Response: Price
               Df     Sum Sq    Mean Sq   F value    Pr(>F)    
LotFrontage     1 1.9494e+12 1.9494e+12 1250.8491 < 2.2e-16 ***
LotArea         1 3.9418e+11 3.9418e+11  252.9255 < 2.2e-16 ***
YearBuilt       1 4.3515e+12 4.3515e+12 2792.1518 < 2.2e-16 ***
YearRemodAdd    1 8.3138e+11 8.3138e+11  533.4541 < 2.2e-16 ***
MasVnrArea      1 1.4801e+12 1.4801e+12  949.7322 < 2.2e-16 ***
BsmtFinSF1      1 4.4285e+11 4.4285e+11  284.1516 < 2.2e-16 ***
BsmFinSF2       1 1.8179e+09 1.8179e+09    1.1664  0.280253    
TotalBsmtSF     1 6.8575e+11 6.8575e+11  440.0082 < 2.2e-16 ***
BsmtFullBath    1 2.7746e+09 2.7746e+09    1.7803  0.182244    
BsmtHalfBath    1 1.2120e+10 1.2120e+10    7.7769  0.005337 ** 
FullBath        1 5.2607e+11 5.2607e+11  337.5545 < 2.2e-16 ***
HalfBath        1 4.0187e+11 4.0187e+11  257.8622 < 2.2e-16 ***
Bedroom         1 1.0333e+07 1.0333e+07    0.0066  0.935109    
Kitchen         1 8.9105e+10 8.9105e+10   57.1743 5.777e-14 ***
TotRmsAbvGrd    1 4.6244e+11 4.6244e+11  296.7227 < 2.2e-16 ***
Fireplaces      1 1.4826e+11 1.4826e+11   95.1301 < 2.2e-16 ***
GarageYrBlt     1 3.2563e+10 3.2563e+10   20.8939 5.118e-06 ***
GarageCars      1 2.3936e+11 2.3936e+11  153.5820 < 2.2e-16 ***
GarageArea      1 4.4256e+10 4.4256e+10   28.3971 1.087e-07 ***
WoodDeckSf      1 1.5172e+10 1.5172e+10    9.7350  0.001831 ** 
OpenPorchSF     1 1.2247e+08 1.2247e+08    0.0786  0.779257    
EnclosedPorch   1 3.3133e+09 3.3133e+09    2.1260  0.144960    
`3SsnPorch`     1 1.3416e+08 1.3416e+08    0.0861  0.769246    
ScreenPorch     1 4.1709e+10 4.1709e+10   26.7626 2.504e-07 ***
PoolArea        1 4.6503e+09 4.6503e+09    2.9839  0.084236 .  
Miscval         1 1.2534e+11 1.2534e+11   80.4244 < 2.2e-16 ***
Mosold          1 4.6694e+07 4.6694e+07    0.0300  0.862594    
YrSold          1 1.4427e+09 1.4427e+09    0.9257  0.336084    
Residuals    2245 3.4988e+12 1.5585e+09                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the result of the anova table, we concluded that we can actually drop out night predictors while only keeping 19 predictors.

Now we have a more compact model. As the residual plot implies, we are also interested in considiering the potential interaction among those 19 predictors. We first oberseved the correlation matrix of the predictors:

	(Intercept)	LotArea	YearBuilt	YearRemodAd	MasVnrArea	BsmtFinSF1	TotalBsmtSF	BsmtFullBath	FullBath
(Intercept)	1	-0.110338	-0.4594488	-0.6822193	-0.002874	-0.0384003	0.1937306	0.1177937	0.4373111
LotArea	-0.110338	1	0.0840981	0.0459729	0.0208766	-0.0269775	-0.1050763	-0.0384701	-0.0181522
YearBuilt	-0.4594488	0.0840981	1	-0.3344973	-0.0680197	-0.0310604	-0.2093123	-0.0668079	-0.2939065
YearRemodAd	-0.6822193	0.0459729	-0.3344973	1	0.065325	0.0650234	-0.0423661	-0.0722482	-0.2231572
MasVnrArea	-0.002874	0.0208766	-0.0680197	0.065325	1	-0.1072764	-0.1534344	0.0480433	-0.0284983
BsmtFinSF1	-0.0384003	-0.0269775	-0.0310604	0.0650234	-0.1072764	1	-0.3146508	-0.5543696	0.0093159
TotalBsmtSF	0.1937306	-0.1050763	-0.2093123	-0.0423661	-0.1534344	-0.3146508	1	0.0076304	0.0011809
BsmtFullBath	0.1177937	-0.0384701	-0.0668079	-0.0722482	0.0480433	-0.5543696	0.0076304	1	0.1235539
FullBath	0.4373111	-0.0181522	-0.2939065	-0.2231572	-0.0284983	0.0093159	0.0011809	0.1235539	1
HalfBath	0.2624465	0.0080725	-0.2725759	-0.0559431	-0.1029426	-0.0187676	0.3112716	0.0302962	0.1940034
Bedroom	-0.1846033	-0.0495017	0.0292999	0.1630499	0.058152	0.0396469	-0.0001525	0.0364068	-0.1590992
Kitchen	-0.2371853	0.0532831	0.0590134	0.1732725	0.0520891	0.0401097	0.0232389	-0.1204634	-0.1629864
TotRmsAbvGr	-0.0418232	-0.0464111	0.21818	-0.1385127	-0.1157128	0.0291625	-0.1695584	0.0201559	-0.237413
Fireplaces	-0.1037784	-0.1234533	0.0504057	0.069291	-0.0520212	-0.1053868	-0.1018776	-0.0100692	-0.0730762
GarageCars	0.1412306	0.015331	-0.1540634	-0.026923	-0.0013225	0.0699352	0.0465142	-0.0350167	-0.132894
GarageArea	-0.008855	-0.0682055	0.0100703	-0.0001593	-0.0627424	-0.0765941	-0.1388141	0.0320474	0.0600199
WoodDeckSf	0.0575951	-0.0727192	-0.0053229	-0.058842	-0.0037132	-0.039181	-0.0177821	-0.0764635	-0.0313568
EnclosedPorc	-0.262108	0.0048382	0.3022076	0.0262261	0.0356663	0.0414258	-0.0616546	-0.0345249	-0.0446756
ScreenPorch	-0.108415	0.0001838	0.0959364	0.0335907	-0.0126316	-0.0311402	-0.0342101	-0.0147007	-0.0082371
Miscval	-0.0045067	-0.0469572	0.0121816	-0.0035468	-0.0048348	-0.0928078	-0.0503803	0.0862509	0.0301321

	HalfBath	Bedroom	Kitchen	TotRmsAbvGr	Fireplaces	GarageCars	GarageArea	WoodDeckSf	EnclosedPorc	ScreenPorch	Miscval
(Intercept)	0.2624465	-0.1846033	-0.2371853	-0.0418232	-0.1037784	0.1412306	-0.008855	0.0575951	-0.262108	-0.108415	-0.0045067
LotArea	0.0080725	-0.0495017	0.0532831	-0.0464111	-0.1234533	0.015331	-0.0682055	-0.0727192	0.0048382	0.0001838	-0.0469572
YearBuilt	-0.2725759	0.0292999	0.0590134	0.21818	0.0504057	-0.1540634	0.0100703	-0.0053229	0.3022076	0.0959364	0.0121816
YearRemodAd	-0.0559431	0.1630499	0.1732725	-0.1385127	0.069291	-0.026923	-0.0001593	-0.058842	0.0262261	0.0335907	-0.0035468
MasVnrArea	-0.1029426	0.058152	0.0520891	-0.1157128	-0.0520212	-0.0013225	-0.0627424	-0.0037132	0.0356663	-0.0126316	-0.0048348
BsmtFinSF1	-0.0187676	0.0396469	0.0401097	0.0291625	-0.1053868	0.0699352	-0.0765941	-0.039181	0.0414258	-0.0311402	-0.0928078
TotalBsmtSF	0.3112716	-0.0001525	0.0232389	-0.1695584	-0.1018776	0.0465142	-0.1388141	-0.0177821	-0.0616546	-0.0342101	-0.0503803
BsmtFullBath	0.0302962	0.0364068	-0.1204634	0.0201559	-0.0100692	-0.0350167	0.0320474	-0.0764635	-0.0345249	-0.0147007	0.0862509
FullBath	0.1940034	-0.1590992	-0.1629864	-0.237413	-0.0730762	-0.132894	0.0600199	-0.0313568	-0.0446756	-0.0082371	0.0301321
HalfBath	1	-0.0941143	0.0761891	-0.2597586	-0.1237043	-0.0571134	0.0320089	-0.03161	-0.0270029	-0.0434123	-0.0383414
Bedroom	-0.0941143	1	0.0394658	-0.5603706	0.1270641	0.0560022	0.0091067	0.0092712	0.0074364	-0.0059624	0.0546583
Kitchen	0.0761891	0.0394658	1	-0.2647015	0.1493839	-0.0187197	0.0402306	0.0718605	0.0592574	0.052851	-0.0271615
TotRmsAbvGr	-0.2597586	-0.5603706	-0.2647015	1	-0.1701371	-0.0607385	-0.0335591	-0.0366696	-0.0145978	0.0108689	-0.0612463
Fireplaces	-0.1237043	0.1270641	0.1493839	-0.1701371	1	-0.1083528	0.0748645	-0.1014419	-0.0492078	-0.1374767	0.0485432
GarageCars	-0.0571134	0.0560022	-0.0187197	-0.0607385	-0.1083528	1	-0.8209634	-0.0082377	-0.0045037	0.0112865	0.0443183
GarageArea	0.0320089	0.0091067	0.0402306	-0.0335591	0.0748645	-0.8209634	1	-0.0228176	-0.0312304	-0.0383445	-0.0121917
WoodDeckSf	-0.03161	0.0092712	0.0718605	-0.0366696	-0.1014419	-0.0082377	-0.0228176	1	0.0756043	0.105349	-0.046006
EnclosedPorc	-0.0270029	0.0074364	0.0592574	-0.0145978	-0.0492078	-0.0045037	-0.0312304	0.0756043	1	0.1122774	-0.0093717
ScreenPorch	-0.0434123	-0.0059624	0.052851	0.0108689	-0.1374767	0.0112865	-0.0383445	0.105349	0.1122774	1	-0.0025982
Miscval	-0.0383414	0.0546583	-0.0271615	-0.0612463	0.0485432	0.0443183	-0.0121917	-0.046006	-0.0093717	-0.0025982	1

As the matric indicated there indeed exists some high correlated terms. However its infeasible to hiearchly adding all of the interaction terms in our model and test them one by one, since there are a total of 190 interaction terms. We decided to first included all interaction in our model, and drop those terms that failed to pass the F test with $\alpha = 0.01$. We first center all the predictor variables by their means to avoid some rcomputational rounding error when internelly solving inverse matrix :

```
data_centered = data.frame(matrix(ncol = 0, nrow = 2274))
for (i in 1:19){
  data_centered[,i] = data2[,i] - mean(data2[,i], na.rm = TRUE)
}

names(data_centered) = c("LotArea", "YearBuilt", "YearRemodAdd", "MasVnrArea", "BsmtFinSF1", "TotalBsmtSF", "BsmtFullBath", "FullBath",
                         "HalfBath", "Bedroom", "Kitchen", "TotRmsAbvGrd", "Fireplaces", "GarageCars", "GarageArea", "WoodDeckSf", "
                         EnclosedPorch", "ScreenPorch", "Miscval")

data_centered[, "Price"] = data2[, "Price"]
```

Then we fit the orignal predictors with all their pairwise interaction terms in the first order model:

```
reg_c = lm(Price ~ .^2, data = data_centered)
```

Finally we run the deletion procedure:

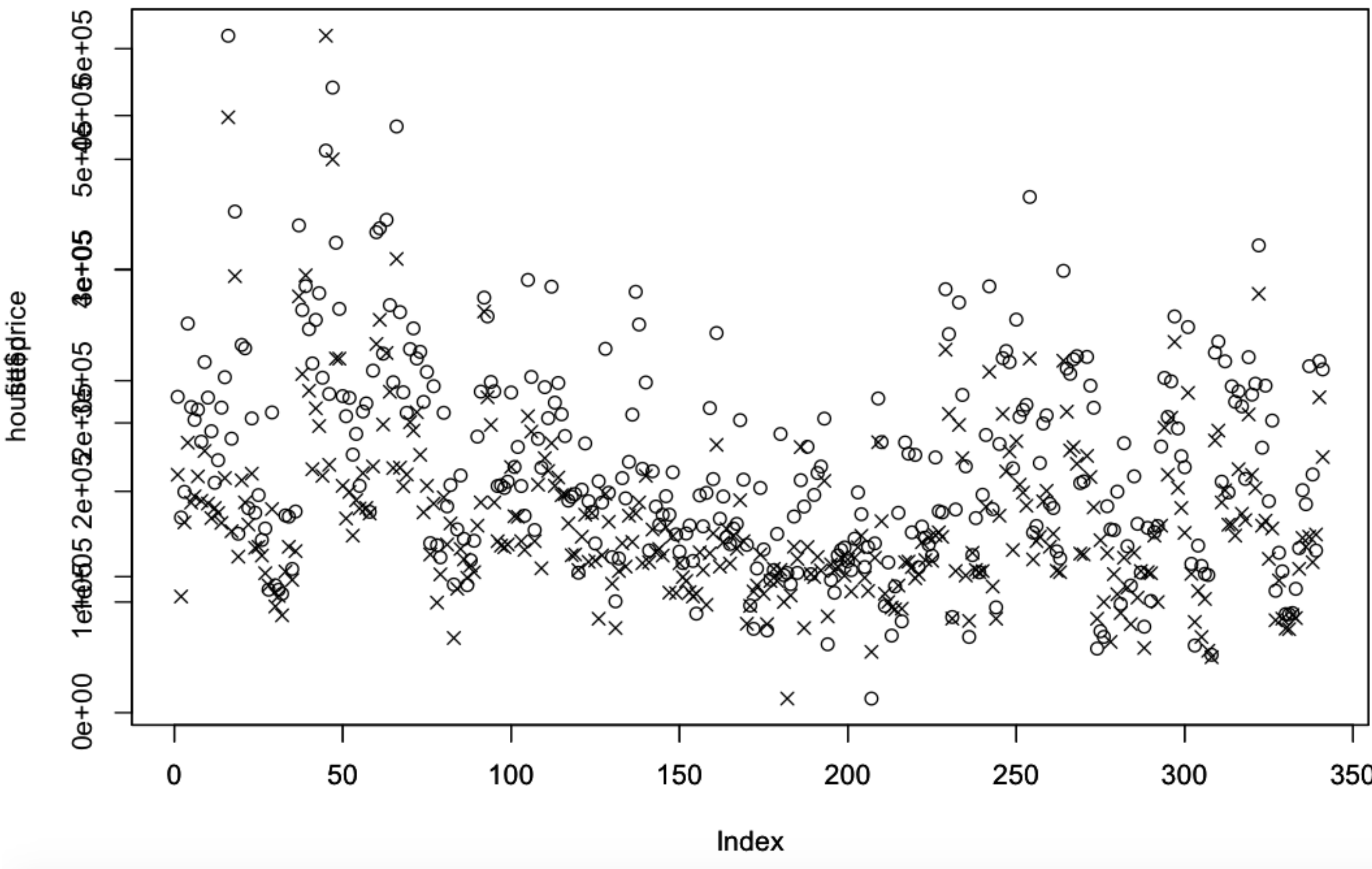
```r
anova_c = Anova(reg_c, type = 2)
data_with_interaction = data_centered
for(i in 1:18){
  helper = 0
  for(k in 1:i){
    helper = helper + 19-(k-1)
  }
  for(j in (i+1):19){
    if(anova_c[helper + j-i, 4] <= 0.01){
      data_with_interaction[, rownames(anova_c)[helper + j-i]] = data_centered[,i]*data_centered[,j]
    }
  }
}
```

The new model summary is as follows:

```
Call:
lm(formula = Price ~ ., data = data_with_interaction)

Residuals:
    Min      1Q  Median      3Q     Max 
-279120  -16203   -1265   14898  255605 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               1.685e+05  9.636e+02 174.846  < 2e-16 ***
LotArea                   9.831e-01  1.869e-01   5.259 1.58e-07 ***
YearBuilt                 3.461e+02  4.361e+01   7.935 3.29e-15 ***
YearRemodAdd              5.716e+02  5.086e+01  11.239  < 2e-16 ***
MasVnrArea                9.990e+00  5.257e+00   1.900 0.057525 .  
BsmtFinSF1                2.027e+01  2.329e+00   8.702  < 2e-16 ***
TotalBsmtSF               4.083e+01  2.465e+00  16.567  < 2e-16 ***
BsmtFullBath              6.039e+03  1.769e+03   3.413 0.000653 ***
FullBath                  1.982e+04  1.872e+03  10.588  < 2e-16 ***
HalfBath                  2.081e+04  1.715e+03  12.137  < 2e-16 ***
Bedroom                  -4.068e+03  1.281e+03  -3.174 0.001522 ** 
Kitchen                  -4.752e+04  4.427e+03 -10.734  < 2e-16 ***
TotRmsAbvGrd              9.359e+03  7.870e+02  11.892  < 2e-16 ***
Fireplaces                1.197e+04  1.280e+03   9.355  < 2e-16 ***
GarageCars                6.172e+03  2.150e+03   2.871 0.004132 ** 
GarageArea                4.161e+01  7.063e+00   5.892 4.40e-09 ***
WoodDeckSf                1.388e+01  5.995e+00   2.315 0.020702 *  
EnclosedPorch             2.491e+01  1.147e+01   2.172 0.029974 *  
ScreenPorch               6.125e+01  1.202e+01   5.097 3.74e-07 ***
Miscval                  -1.166e+01  1.455e+00  -8.015 1.76e-15 ***
`LotArea:YearBuilt`       3.323e-02  6.508e-03   5.106 3.56e-07 ***
`LotArea:MasVnrArea`      1.885e-03  6.922e-04   2.723 0.006525 ** 
`LotArea:BsmtFinSF1`     -4.167e-03  3.125e-04 -13.335  < 2e-16 ***
`LotArea:BsmtFullBath`    1.731e+00  2.684e-01   6.449 1.38e-10 ***
`LotArea:FullBath`       -1.518e+00  3.019e-01  -5.028 5.36e-07 ***
`LotArea:Bedroom`         6.983e-01  1.891e-01   3.693 0.000227 ***
`LotArea:Fireplaces`      6.018e-01  2.313e-01   2.602 0.009334 ** 
`LotArea:GarageCars`      1.938e+00  3.820e-01   5.073 4.24e-07 ***
`LotArea:GarageArea`     -4.636e-03  1.289e-03  -3.596 0.000330 ***
`YearBuilt:MasVnrArea`    1.483e+00  2.021e-01   7.336 3.06e-13 ***
`YearBuilt:TotalBsmtSF`   6.293e-01  7.530e-02   8.357  < 2e-16 ***
`YearBuilt:Fireplaces`   -3.311e+02  5.074e+01  -6.526 8.35e-11 ***
`YearRemodAdd:FullBath`   3.917e+02  8.184e+01   4.786 1.81e-06 ***
`MasVnrArea:WoodDeckSf`   8.087e-02  3.110e-02   2.601 0.009369 ** 
`BsmtFinSF1:Fireplaces`   1.080e+01  3.036e+00   3.555 0.000386 ***
`TotalBsmtSF:Kitchen`    -4.070e+01  6.463e+00  -6.297 3.65e-10 ***
```

```
`TotalBsmtSF:Fireplaces`  1.087e+01  3.592e+00   3.027 0.002500 **
`TotalBsmtSF:WoodDeckSf` -9.241e-03  1.327e-02  -0.696 0.486345
`FullBath:Bedroom`        9.994e+03  1.505e+03   6.639 3.96e-11 ***
`FullBath:Fireplaces`     1.484e+04  2.586e+03   5.740 1.08e-08 ***
`FullBath:GarageArea`     6.860e+01  8.083e+00   8.488  < 2e-16 ***
`HalfBath:Kitchen`       -2.683e+04  5.782e+03  -4.640 3.69e-06 ***
`HalfBath:Fireplaces`     1.259e+04  2.273e+03   5.541 3.36e-08 ***
`Bedroom:Fireplaces`     -9.338e+03  1.641e+03  -5.689 1.44e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31460 on 2230 degrees of freedom
Multiple R-squared:  0.8602,    Adjusted R-squared:  0.8575
F-statistic: 319.2 on 43 and 2230 DF,  p-value: < 2.2e-16
```

By the result of F test, we droped most of the interaction terms while only keeping 24 of them. And we randomly plot around 350 observations and their corresponding three hundred fitted value, denoted by symbol "x" and "o", to visualize our model's performance:



Interpretations

The y-intercept for our model was estimated at $16850 indicating that if all variables were set to zero, we expect this to be the starting price of a house. Our positive beta values show that these

corresponding variables have a positive, linear relationship to housing price. If we increase these aspects, our house price is expected to increase as well. Variables with negative beta values have the opposite effect on the house price. These variables include age of the home and the years since the house has been remodeled. Housing prices will increase if they have been built or remodelled in recent years. Our r-squared value is calculated at 0.8602. This means that 82.29% of the variability in housing prices are explained by the x-variables in our model. Although a high r-squared value does not indicates that this is a good model to estimate the price of a house, it still indicates that our model make sense in some way.
By including the interaction terms, the mean square error were reduced about ten times. Although its still a considerably large number. One potential reason for might due to the distribution of the response variable is very sparse, and the range is numerically large. In selecting our model, we chose different variables that we believed would influence the pricing of homes. We put our data into R and the ANOVA table and F-test revealed that some of our original variables did not significantly contribute to our model. From here, we decided to remove this variable which led us to the model that we have presented. We chose a multiple linear regression model with interaction terms because we believed that there would be several highly influential factors in pricing a home rather than prices relying on a single variable.

Possible Problems

The biggest problem of our dataset is that it will be most applicable to the town of Ames, Iowa. Our data was retrieved from this specific area and, therefore, is best used to estimate prices of the houses here. This would not be a good model to use for other cities that have large economic differences from Iowa. We believe that factors such as crime rate, income, etc. have a significant effect on housing prices as well, so we cannot rely on this model to give accurate estimates for houses in areas that are not similar to Ames. For example, a house price in a city, such as Boston, will have greater beta values as these houses tend to sell at higher prices than Ames, on average. Another problem that we noticed is that the beta values for number of bedrooms and kitchens is negative. We expected that the more bedrooms and bathrooms a house contains, the higher the price of the house would be. However, this was not the case in our data analysis. This is one aspect of our model that we are questioning and it may be tied with variable dependence even though we have included interaction terms. If we included more interactions and higher order terms in our model, we might achieve a better model. However a linear model with too high order suffer a lot from overfitting: It often perform well in obervation set, but perfromed poorly for new data.
We also delete tens of categorical variables from original dataset, which might also leads problem. Those categorical variables might be the key for having a good model.
The other major problems is that a fitness of a overall linear model remain questionaire. In the persepective of predicting, there might be the model that perform much better than linear model, such as decision tree or neural network... Or even adding l_1 or l_2 regularizations to linear model might make it perfrom much better.

Conclusion

Overall our linear regression model had some problems and it was not perfect but it should work as a game model to estimate housing prices. It cal be a useful tool in helping individuals find a house that is fitting for their budget. We can alter different aspects of our dream home and find the perfect combination that work best with our needsand wallets. Overall, linear model overperfrom many other model such as decision trees, in the persepective of interpretation. By our interpretation among interaction terms, we obtain a lot of useful information that might reveal the potential relashionship of response variables and predictors.