Mini Project 2: Predicting Housing Prices in Ames, Iowa

Chris Corona, Farshina Nazrul, Gak Roppongi, Nolan Walker

2022-11-03

Abstract

In this mini-project, our task is to select predictors, fit, and evaluate several different types of models for regression. The data set is from the Ames Assessor's Office house in computing assessed values for individual residential properties sold in Ames, IA, from 2006 to 2010. This dataset has 2930 observations with 82 predictors (2930 * 82). 82 predictors include 23 nomial (categorical) variables, 23 ordinal variables, 14 discrete variables, and 20 continuous variables with two additional observation identifiers. We tried to predict the price of the residential properties based the predictors. After fitting various models, we compared the OOO to determine which model has the best performance.

1 Summary

This data set presented a challenging problem with so much missing data to clean and then so many variables to sort through. To deal with the 74 variables, we approached the task in several ways. First, we tried to reduce the variables through subset selection - choosing the most important variables one at a time in a stepwise fashion. This only achieved a minor reduction and left us with a somewhat confusing interpretation. The neighborhood, square footage of different parts of the house, condition, quality, roofing material, are all important when predicting price - not much of a reduction. Next we tried two different shrinkage methods, lasso and ridge regression. Ridge regression shrinks the coefficients towards zero, but not exactly zero so this actually does not perform any reduction on the variables. Lasso on the other hand shrinks many of the coefficients exactly to zero. The best lasso model left us with just 9 non-zero coefficients which were related to area of the house, year built/remodeled, and condition. This is a much more interpretable result. And finally, we tried Principal Component Analysis which allowed us to reduce the model down to just five variables, a huge reduction. However, PCA is known for even more confusing interpretations because the principal components are derivations from all the other variables. That being said, the first two PCs could be interpreted as relating to overall quality, area of the house, and year built/remodeled. After these analyses, we can say that the most important variables in predicting the price of a house in Ames, IA between the years of 1872-2010 are area, year built/remodeled, and overall quality.

The final step in our analysis was using just one predictor, year built, to predict price using three different methods. Simple linear regression gives us simple results. The regression line does not capture the non-linearity of the data, and therefore is not a good model to predict price (particularly for very old or very new houses). Next, we tried polynomial regression which the 2nd order polynomial gave the best results. This makes sense because the curvature of the data most closely matches a parabola. And finally, we used smoothing spline to fit a model. The best smooth spline appears to greatly overfit the data as evidenced by how wiggly and flexible the fit is. In the end, the best model is the one that does the best job at fitting to the underlying truth of the data. The challenge is to not underfit or overfit to the data. This is more of an art than a science.

2 Data Preprocessing

The data set is from the Ames Assessor's Office house in computing assessed values for individual residential properties sold in Ames, IA, from 2006 to 2010. There are 2930 observations. The prices were between $12,789 and $755,000 with the mean value of $180,796.0600682594, the median value of $160,000, and the standard deviation of 79886.692356665. "Pool quality", "Miscellaneous Feature", "Alley", "Fence", "Fireplace quality", and "Lot Frontage" features were removed because most of the values in these columns were missed. We filled the missing value with the most common response for "Masonry veneer type", "Garage Condition", "Garage Quality", "Garage Finish", "Garage Year Built", "Garage Type", "Basement Condition", "Basement Quality", "Basement Exposure", "Basement Finish Type. 1", and "Basement Finish Type. 2". As a preprocessing step, we split out data into training and test sets. We performed 10 times cross validation. First 10% of the dataset was used as the test set and the other 90% was used as the training set for the first iteration, the second 10% was used as the test set and the other 90% was used as the training set for the second iteration, and we repeated this for 10 times to complete the 10 times cross validation. Since this is a regression dataset, the datasets were split according to the index.

3 Exploratory Analysis

In our Explanatory Data Analysis (EDA), we rigorously reviewed the data for any issues and reduced the number of predictors from 80 to 74. We first checked for any missing data. 6 features were removed because most of the values were missed in these columns. Missing values in 11 features were replaced with the most common response in each of the feature. We followed the instruction from the data description and removed the observations with the area more than 4000 squared feet. Next, we analyzed the variable correlation. The highest correlation between a variable and the response variable was between area and price at ~0.8. Then, we performed the forward and backward selection. However,

4 Model development and performance evaluation

Subset Selection

Best Subset and step-wise selection are two model selection methods that reduce the number of terms in a model, by comparing AICs of models with different number of terms. Best subset compares all models with 1 to N number of terms, where N is the total number of variables in the data set, against each other and returns the best model for each value of n. This approach is computationally impractical for data sets where there are a large number of features, as it scales exponentially with the number of features. Step-wise selection is a model selection approach that instead takes into account the performance of the model at each step after adding or removing a term from the model, for forward or backward selection respectively. It returns the model with terms that only increased the performance at the addition or removal at each step of a term, that stopped when performance started to decrease relative to the last best performing model.

```
##                            Estimate      Pr(>|t|)
## Roof.Matl_ClyTile      -6.541570e+05 3.734999e-115
## area                    4.899541e+01  4.637292e-66
## Neighborhood_StoneBr    4.712845e+04  2.603557e-35
## Overall.Qual            7.496563e+03  3.343027e-33
## Neighborhood_NridgHt    3.111856e+04  4.878000e-33
## Neighborhood_NoRidge    3.858718e+04  4.019759e-32
## BsmtFin.SF.1            3.070232e+01  8.188474e-32
## Misc.Val               -8.727790e+00  5.785672e-27
## Overall.Cond            5.600154e+03  3.474018e-26
```

```
## Bldg.Type_1Fam          2.426120e+04  1.311599e-25
## Neighborhood_Somerst    2.413950e+04  1.272028e-22
## Kitchen.Qual_Ex         2.256745e+04  3.589225e-21
## Bsmt.Exposure_Gd        1.813739e+04  8.666704e-21
## Bsmt.Qual_Ex            1.964933e+04  1.412907e-19
## Condition.2_PosN       -1.111750e+05  5.654908e-19
## Lot.Area                6.096945e-01  7.031330e-16
## BsmtFin.SF.2            3.290237e+01  1.797417e-15
## Neighborhood_GrnHill    1.263057e+05  1.820325e-14
## Exter.Qual_Ex           2.387125e+04  1.273928e-13
## Roof.Matl_CompShg      -6.064824e+04  2.610131e-12
## Condition.1_Norm        9.791548e+03  2.700137e-12
## Year.Built              2.407314e+02  1.643738e-10
## Neighborhood_Crawfor    1.747126e+04  2.492725e-10
## Exterior.1st_BrkFace    1.686944e+04  4.098810e-10
## Bsmt.Unf.SF             1.340007e+01  1.598979e-08
## House.Style_1Story      1.043999e+04  1.821048e-08
## Roof.Matl_WdShake      -6.412097e+04  2.295463e-08
## Screen.Porch            4.447029e+01  2.428694e-08
## Pool.Area               6.824306e+01  1.207118e-07
## Bsmt.Exposure_Av        7.407956e+03  2.526550e-07
## Mas.Vnr.Area            1.920143e+01  2.571985e-07
## Sale.Type_New           8.790949e+03  3.230372e-06
## `Roof.Matl_Tar&Grv`    -4.861543e+04  5.654868e-06
## Condition.2_PosA        5.465662e+04  1.394226e-05
## Sale.Condition_Abnorml -7.826559e+03  1.427642e-05
## Fireplaces              3.693406e+03  1.453961e-05
## Neighborhood_Edwards   -8.206289e+03  5.269414e-05
## Land.Contour_Bnk       -9.393813e+03  8.221926e-05
## X2nd.Flr.SF             1.381680e+01  8.519175e-05
## Condition.1_PosN        1.620394e+04  1.129208e-04
## Garage.Qual_Ex          6.981785e+04  3.256009e-04
## Lot.Config_CulDSac      6.572861e+03  5.786654e-04
## Neighborhood_NPkVill    1.907924e+04  6.843556e-04
## Land.Slope_Mod          2.477835e+04  7.913510e-04
## Bedroom.AbvGr          -2.621373e+03  8.395246e-04
## Neighborhood_BrDale     1.822136e+04  9.115680e-04
## Land.Contour_Low       -1.216109e+04  1.257092e-03
## Bldg.Type_2fmCon        1.322837e+04  2.110695e-03
## Exterior.1st_PreCast    7.144042e+04  2.691044e-03
## Exterior.1st_MetalSd    4.046507e+03  2.916924e-03
## Garage.Cond_Ex         -5.537691e+04  3.104841e-03
## BsmtFin.Type.2_LwQ     -8.159345e+03  3.190949e-03
## Full.Bath               4.027983e+03  3.309251e-03
## Mas.Vnr.Type_CBlock    -6.737885e+04  3.561786e-03
## Garage.Area             1.418483e+01  3.606275e-03
## MS.Zoning_RL            5.102231e+03  3.628292e-03
## Heating.QC_Ex           3.270043e+03  3.780935e-03
## House.Style_2.5Fin     -2.587233e+04  3.910683e-03
## BsmtFin.Type.2_Rec     -7.601937e+03  4.028942e-03
## Mas.Vnr.Type_BrkFace   -3.702238e+03  4.061867e-03
## Roof.Style_Flat        -2.540958e+04  5.501468e-03
## Kitchen.AbvGr          -1.028292e+04  5.824926e-03
## Sale.Type_Con           2.803748e+04  6.586380e-03
```

```
## Functional_Maj1        -1.501322e+04  6.830701e-03
## Functional_Min2        -7.743513e+03  7.984123e-03
## Garage.Yr.Blt           7.107056e+01  9.443831e-03
## Functional_Mod         -1.053419e+04  9.825312e-03
## Neighborhood_MeadowV    1.220903e+04  1.008264e-02
## Neighborhood_NWAmes    -6.114098e+03  1.196597e-02
## Garage.Qual_Gd          1.228178e+04  1.260600e-02
## Year.Remod.Add          8.497792e+01  1.353905e-02
## Neighborhood_BrkSide    6.650979e+03  1.362639e-02
## Bldg.Type_Twnhs        -7.787479e+03  1.379081e-02
## Heating_OthW           -4.047283e+04  1.391204e-02
## Street_Grvl            -1.762185e+04  1.425590e-02
## Garage.Finish_RFn      -2.574093e+03  1.563387e-02
## Condition.2_RRAe        5.849103e+04  1.572026e-02
## Land.Slope_Gtl          1.763194e+04  1.690963e-02
## Exter.Qual_Fa           1.101013e+04  1.840714e-02
## BsmtFin.Type.1_LwQ     -4.805857e+03  2.193252e-02
## Bldg.Type_Duplex        1.082977e+04  2.302073e-02
## Roof.Matl_Roll         -5.497535e+04  2.577453e-02
## Lot.Config_FR2         -5.813948e+03  2.579927e-02
## Functional_Maj2        -1.739224e+04  2.668491e-02
## Functional_Sal         -3.780071e+04  2.670581e-02
## Neighborhood_Blmngtn    1.105346e+04  2.719552e-02
## Neighborhood_Greens     1.927611e+04  3.018307e-02
## Garage.Cars             3.047177e+03  3.198328e-02
## BsmtFin.Type.2_BLQ     -6.519595e+03  3.220404e-02
## Neighborhood_Mitchel   -5.111395e+03  3.482812e-02
## Functional_Sev         -3.550133e+04  3.514866e-02
## Land.Contour_HLS        5.346337e+03  3.613726e-02
## BsmtFin.Type.1_Rec     -3.367571e+03  3.666833e-02
## Exterior.2nd_VinylSd    2.688500e+03  3.779772e-02
## Lot.Shape_IR1          -2.092068e+03  3.974367e-02
## BsmtFin.Type.1_GLQ      2.718602e+03  4.295348e-02
## Sale.Condition_AdjLand  1.406146e+04  4.749378e-02
```

Table 1: List of coefficients and p-values from best model using backward selection and one-hot encoding

A Simple linear model was fit to the data using all of the variables and backwards selection was performed, performance was evaluated using cross validation. The model obtained using backwards selection had a lower RMSE and used 53 of the 72 variables in the data(RMSE 34524 vs 35120). However the model still had 172 terms, since some of the variables were categorical with multiple levels each getting their own term in the model. To try to reduce the number of terms, the categorical variables were one hot encoded, and backwards selection was performed on the full model with all of the one hot encoded variables, this resulted in a model with fewer terms, 123, and the lowest RMSE of any model tested(RMSE=26755 vs 35120). There were over 40 terms in the reduced model with p-values below .001, most with very large coefficients, so the best model is one that with a lot of terms. Some of the most significant terms were, in order, clay tile roofing(categorical), area, stone brook neighborhood(categorical), Overall quality, Finished Basement Square Foot, Value of miscellaneous features, single family building type(categorical), Excellent Kitchen Quality(categorical), and so on. While many of the terms describe similar features, the types of terms that increase value are the ones for total area and quality of each floor/room/area, the location within and which neighborhood the house is in, having various features like a pool, a fireplace, a finished basement/garage, the Year built/remodeled/additions added and the type/size of area outside the house. Interestingly, almost all of the coefficients for terms that describe roofing material are negative.

Lasso and Ridge Regression

Both Lasso and Ridge are shrinkage or regularization methods that estimates the coefficients of 'not-so-useful' predictors towards zero (ridge) or exactly zero (lasso) based on the shrinkage penalty factor, λ. In this project, we used the 10-fold Cross validation for model performance evaluation. However, we only used the numeric predictors. A plot of the cross-validated prediction RMSE for each value of λ is shown in the Figures for both ridge and lasso. The best λ for lasso was selected as 1151.4 and their corresponding RMSE was 36490.15, whereas the best λ for ridge was selected as 3727.59 and their corresponding RMSE was 35716.38.

Lasso had 9 non-zero coefficients. Most of them are related to the total area, year built or remodeled, quality of the house, room areas etc. The coefficients that were 'dropped' were mostly related to basement, porch, garage and month or year sold. Ridge, by definition, has all non-zero coefficients, but some of them were very close to zero.

In between the two methods, ridge had overall lowest prediction RMSE because it is essentially using more predictors or features, so it has a little bit better estimation than lasso. On the other hand, lasso is computationally and storage-wise much more efficient as it requires fewer number of predictors (reduces the dimension).

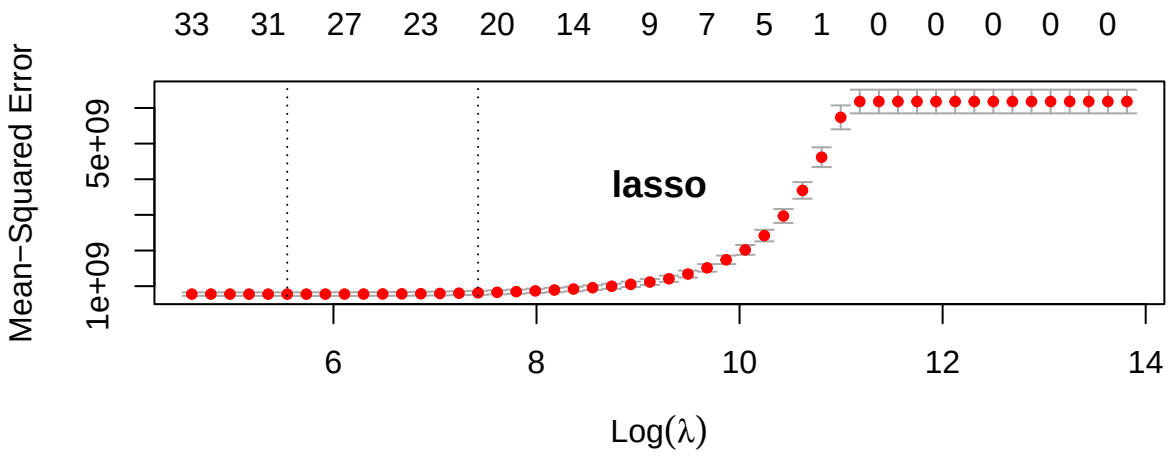

Figure 1: Plot of RMSE across various lambda using lasso shrinkage method

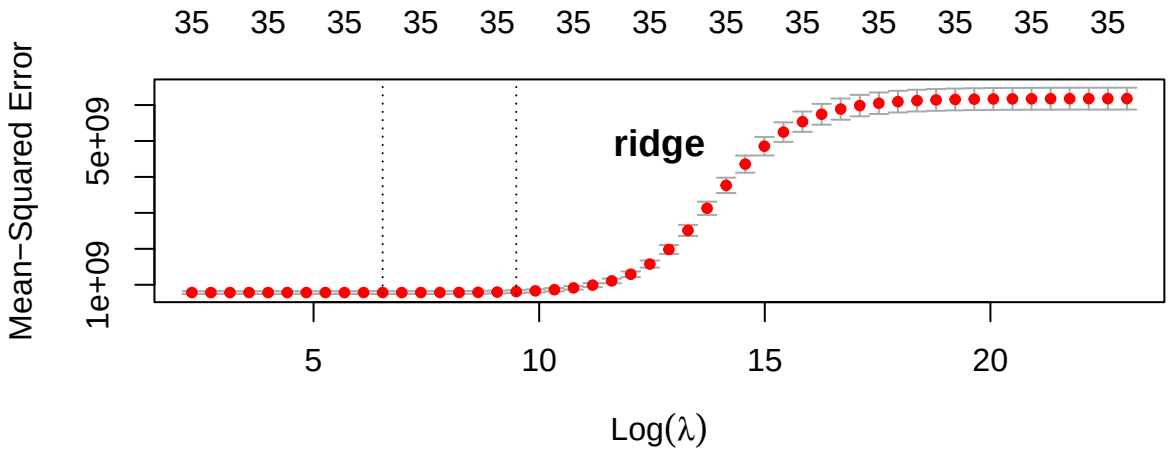

Figure 2: Plot of RMSE across various lambda using ridge regression shrinkage method

Principal Components Analysis

For this data set, we carried out a Principal Components Analysis (PCA). PCA can only be performed on quantitative variables, of which there are 35 in the cleaned data set. We split the data set into a train and test set, then ran the `princomp` function on the 35 quantitative predictors and the response. From this function, we are able to get the PC loadings and scores from all 35 principle components. We built 35 regression models each using cumulative PC scores as predictors (ie model 1 uses PC1 only, model 2 uses PC1 and PC2, model 3 uses PC1, PC2, and PC3, etc). With each of these models, we predicted the price for the test set (after calculating the PC scores for the test set using the loadings). The RMSE was computed for each of these models. We repeated these steps on 10 folds for k-fold cross-validation. A plot of the cross-validated RMSE for each of the 35 models is shown in Figure 3. I would recommend 5 PCs because the slope of the RMSE drops off significantly after that. 5 PCs is the best balance of dimensionality reduction and improvement in RMSE. The minimum RMSE achieved was with 32 PCs, however we have only reduced the dimensionality by three predictors - hardly any improvement. The whole effort of PCA is to reduce dimensionality. If we choose not to reduce the dimensionality by very much, there is little point. Looking at the PCA loadings, we can determine (somewhat) which variables are the most important. We will only focus on the first 2 PCs since these two together explain the most variation (~14%). The first principal component puts the most emphasis on the overall quality, total area of the house, garage size, number of full bathrooms, year built/remodeled, basement and 1st floor square footage. These are all typical factors that home buyers value. The second principal component puts the most emphasis on the second floor square footage, number of bedrooms/rooms above ground, total area, number of half bathrooms, and kitchen size. Interestingly, there are also variables that weigh negatively on this second PC: variables dealing with basement size. So we can think of the 2nd PC as favoring houses without basements. Figure 4 shows a biplot of the different variables and how they relate to the first two principal components. The first PC has many variables associated with it that are not at all associated with the 2nd PC and are therefore too hard to read. But we can clearly see the variables associated with the 2nd PC (and not with the 1st PC) such as X2nd.Flr.SF, Bedroom.AbvGr, and TotRms.AbvGrd.

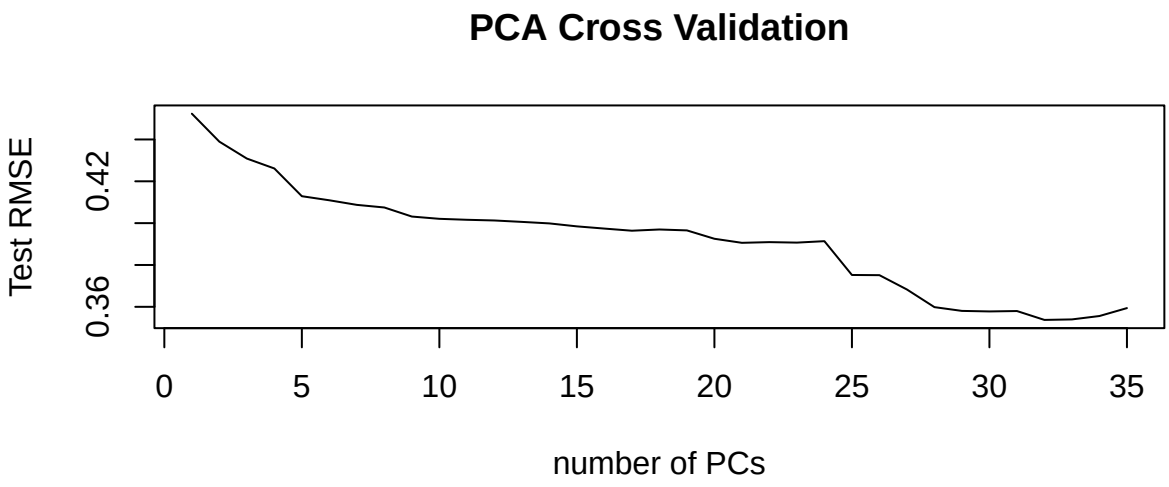

Figure 3: Plot of 10-fold cross-validation of PCA regression.

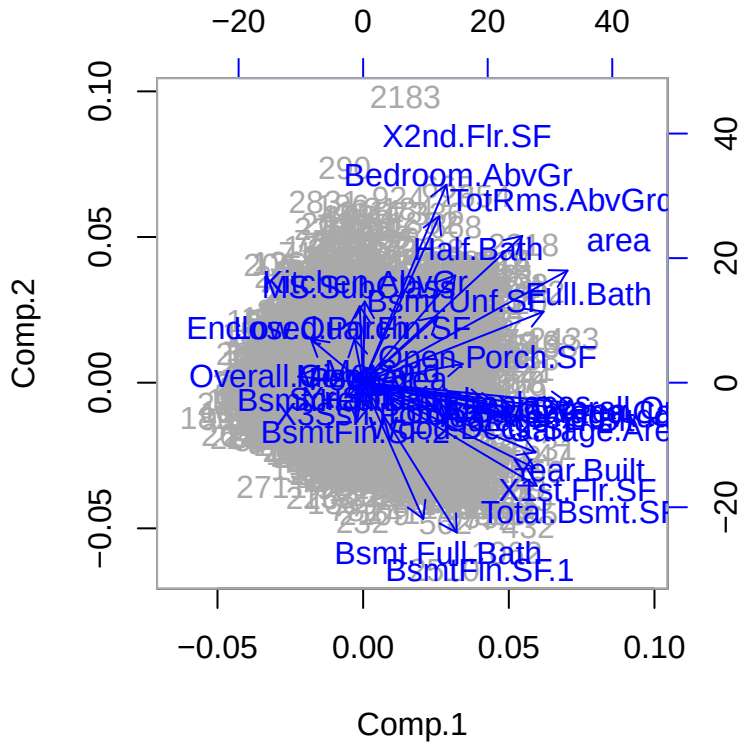

Figure 4: Biplot of first two principal components

Simple Linear Model

A linear model with only year built was fit against price. Ten fold cross validation was performed and as expected, the RMSE was a lot higher than for any of the other linear models fit with more predictors.

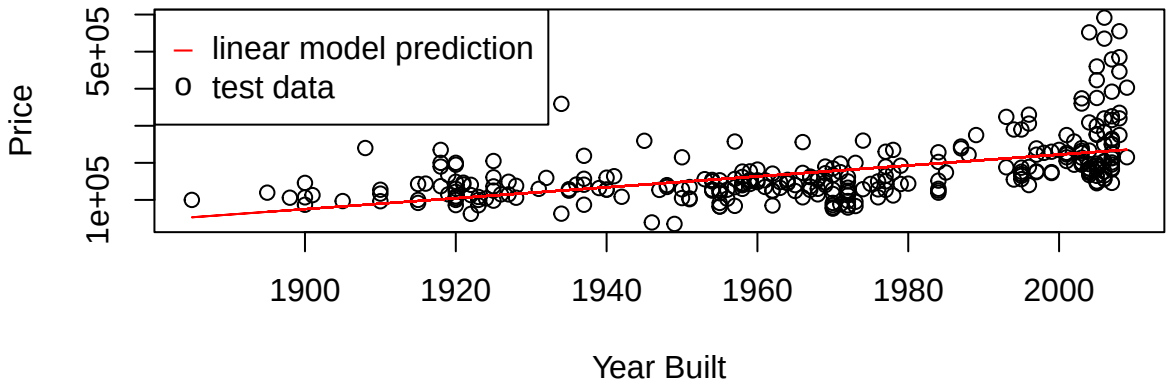

Figure 5: Plot of simple linear regression `price ~ year built`

Polynomial

We used the year the house was built as the only predictor to fit various polynomials to the data set. First we split the data set into a train and test set. On the training set, we built 5 models each with `Year.Built` and adding the next order polynomial of `Year.Built` to subsequent models up to 5th order (ie model 1 `price ~ Year.Built`, model 2 `price ~ Year.Built + Year.Built^2`, and so on). We used these models to predict on the test set and computed the RMSE. We repeated these steps on 10 folds and averaged the RMSE for k-fold cross-validation. I would recommend the 2nd order polynomial because the slope of the RMSE drops off significantly after that. There is minimal improvement in RMSE with higher order polynomials. Also with higher orders, there is the risk of overfitting to the training data. A plot of the cross-validated RMSE for each of the 5 models is shown in Figure 6. The 1st-5th order polynomial fits are shown in Figure 7. The 1st order polynomial (linear) does not capture the curvature of the data. The 2nd order polynomial is able to capture this curvature well. Beyond the 2nd order there is not much difference, the lines are crowded together. Notice some of the polynomial fits appear to be missing - they are not, the the other polynomials predict the exact same values and therefore plots perfectly on top, covering it.

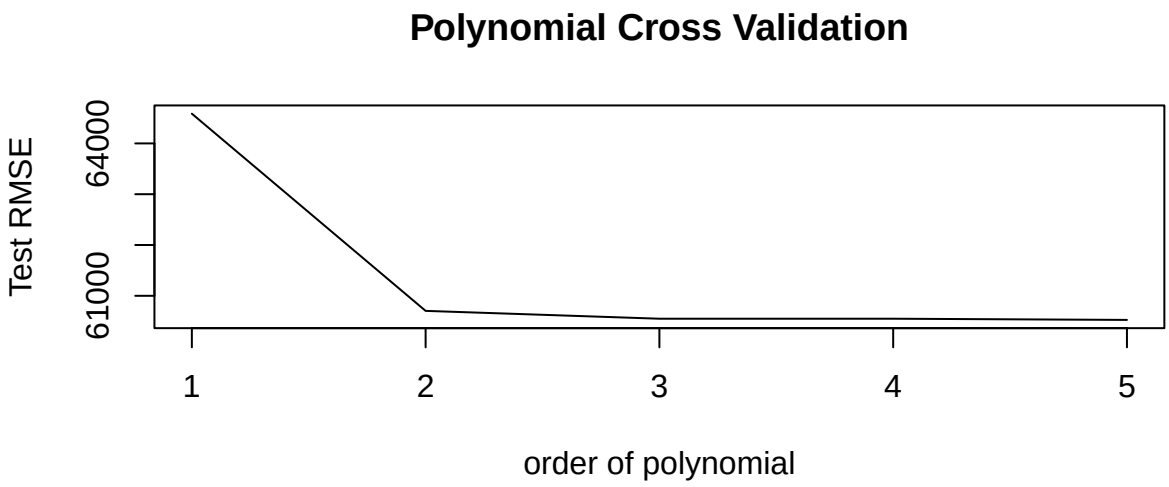

Figure 6: Plot of 10-fold cross-validation of polynomial regression

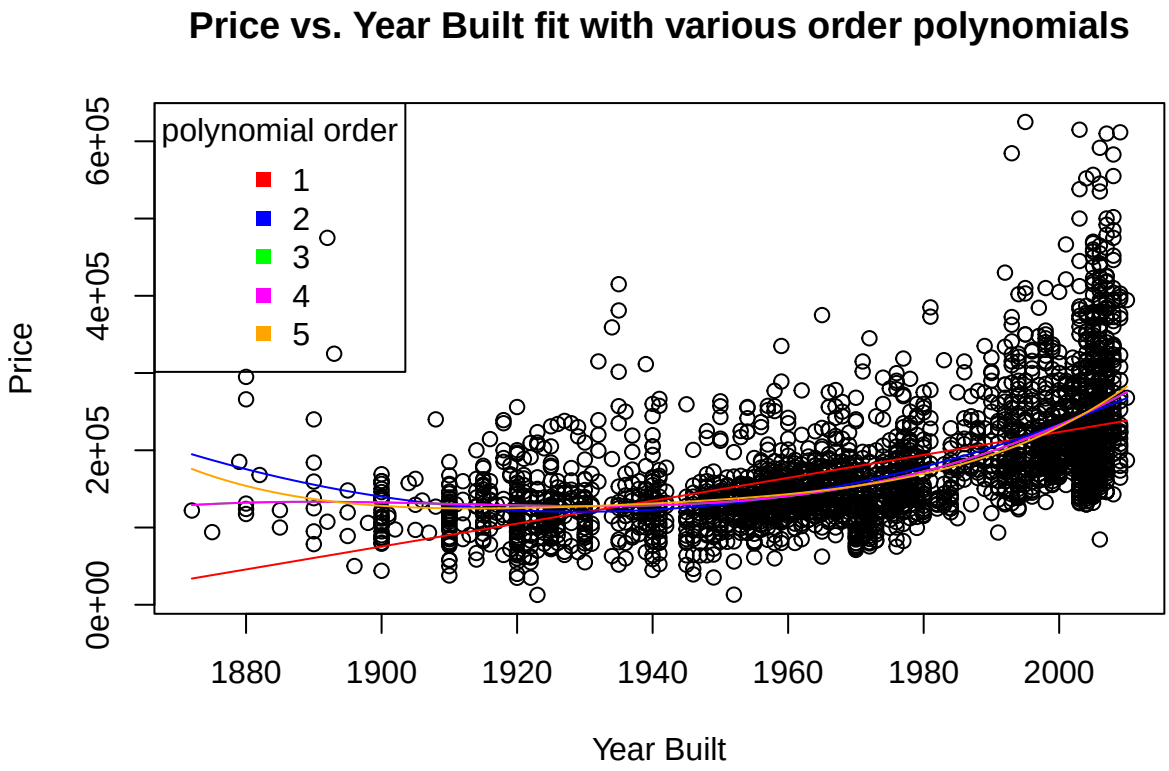

Figure 7: Plot of price vs year built and fit of 1st-5th order polynomials

Smoothing Spline

Splines are a type of interpolation or function estimation techniques. Based on a set of noisy observations, it fits the data into piece-wise polynomial curves called basis functions, in such a way that minimizes the total

error between the observed data and the estimated data points. Similar to lasso and ridge regularization, the smoothing spline approach also has a tuning parameter λ that helps penalize the variability in the basis function. λ effectively controls the degree of freedom and hence the number of knots. A larger value of λ forces the function to be smoother, but at the cost of underfitting, whereas $\lambda \to 0$ causes overfitting. Therefore, we tried 50 different values of λ within a logarithmic range from 10^{-10} to 10^{4}.

In this project, instead of leave-one-out cross-validation (LOOCV), we used the 10-fold Cross validation to be consistent with the other approaches. A plot of the cross-validated prediction RMSE for each value of λ is shown in Figure . The lowest prediction RMSE was obtained corresponding to $\lambda = 5.18 \times 10^{-5}$. Figure shows the best fit smoothing spline of year built vs price. It seems that after 1940, the spline follows the trend of the scatter plot. But before that, the data are significantly sparse so there are some variations in the curve. Intuitively, the total number of houses were much less (or we have access to very small amount of data for the years before 1940), and there were some exceptionally high-priced houses during that time.

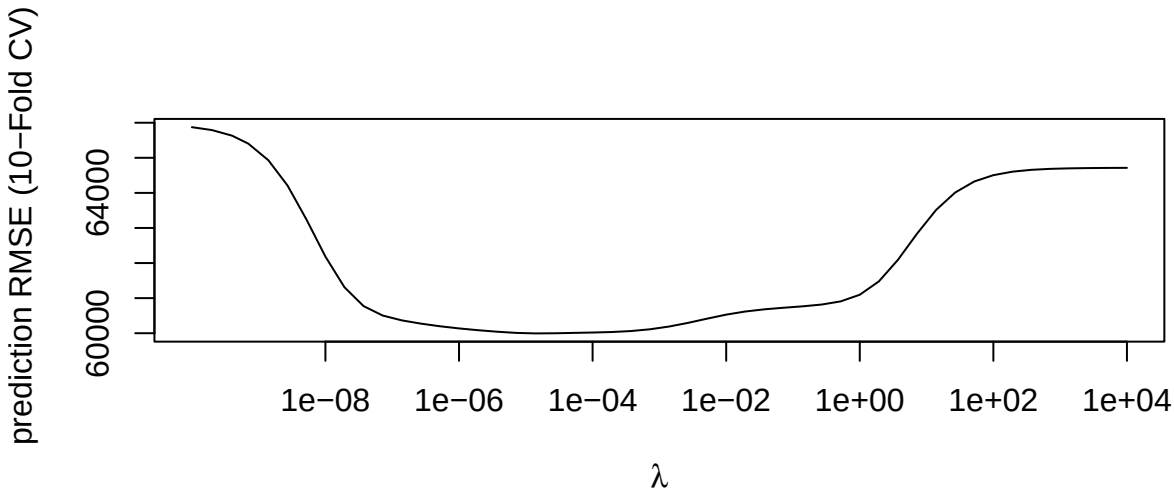

Figure 8: Plot of 10-fold cross-validation of smoothing spline

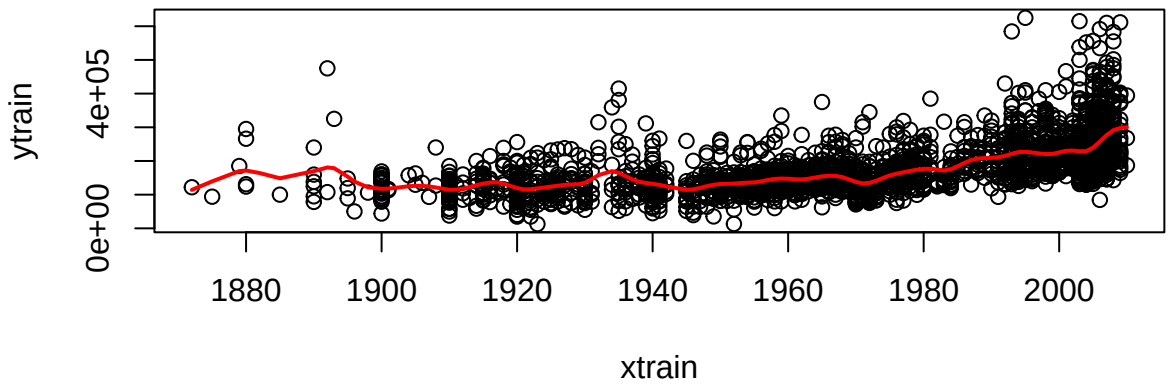

Figure 9: Plot of best smoothing spline model

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.