**HOUSING PRICE: USING ADVANCED REGRESSION TECHNIQUES**

**PRESENTED BY JIANWEN WU, NITASHA TULI AND THOMAS KWOK**

**M.A STATISTICS, HUNTER COLLEGE**

**Table of contents**

**1. Introduction**

**2. Data Analysis**

1. **Data Analysis**
2. **Data Integration**
3. **Variable Description**

**3. Model I – Regression Analysis**

1. **Variable Selection** 
   1. **Correlation Matrix**
   2. **VIF**
2. **Model output**
3. **Model Diagnostic**
4. **Model Prediction**
5. **Model Forecasting**

**4. Discussion**

**INTRODUCTION:**

Housing has always been seen as one of the major stepping stones in adulthood; people graduate college, get a job, start a family, and save enough money to buy a house. The question though has always been how much does it cost to buy a house and what factors into the price of a house. In this project, we looked into dataset from 1460 houses bought in Ames Iowa to create a regression model that best predicts the price of a house in Ames, Iowa. The variable of interest that we study is Sales Price and in our dataset we have eighty predictor variables that are utilized to predict the cost of a house. Of the eighty predictors, we have 23 nominal, 23 ordinal, 14 discrete, and 20 continuous and these range from everything from Total Plot Area to Fire Place and Pool. Our interest in this project is to select the most important variables to study and create a model that best predicts the Sales Price of a house. In evaluating our model, we will look into the bias of our model to the actual sales price, the maximal deviation, mean iabsolute deviation, and mean square error to conclude how accurately we created a model to predict the Sales Price. We will use 1060 observations in our training data set and compare our model to the 400 observations in the test data set to see how our modeled Sales Price compares to the actual Sales Price of the 400 observations.

From our dataset, we will look into normalizing the Sales Price by using the log function, and we will convert some of our predictors into dummy variables. From the list of variables, we will focus on several that we believe are highly correlated with SalesPrice, and then variable selection techniques also to see the maximum number of variables for predicting a highly accurate model without overfitting the data. The purpose is to create a model that can best predict the Sales Price from our validation data set, but also that could be used on other data sets and still get a good prediction.

**DATA PREPROCESSING**

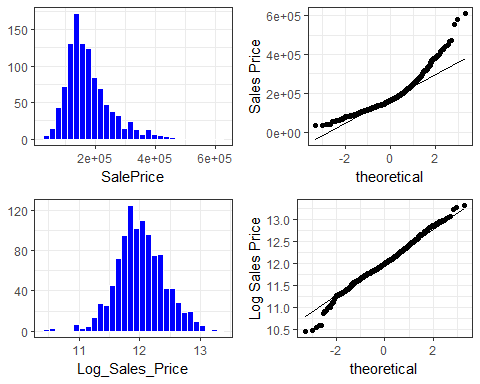
Before we start to apply our regression model to the data set, it is essential for us to have a clean and tide training data, the better training data we can get, the better performance of our regression model could behave. In data preprocessing we are mostly detecting and dealing with missing values, handling outliers and replacing the values with the median or most frequent value Treatment given to each variable is described below.

* + Missing values in variables like PoolQC, MiscFeature, Alley, Fence, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, MasVnrType, MSSubClass are replaced by None based on data description.
  + Missing values in LotFrontage is replace by median of the data series.
  + Missing values in GarageYrBlt, GarageArea, GarageCars, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath, MasVnrArea are replaced with zero.
  + Electrical which denotes type of electrical sytem is replace by “SBrkr” because it was most frequent data value in that series.
  + Rest all categorical variables are converted into factors.

**DATA EXPLORATION**

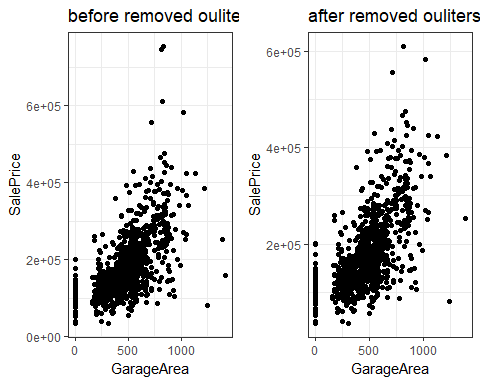
After data cleaning and transformation we began with some exploratory analysis. A histogram plot shows the distribution of the target variable ‘SalePrice’ as being was right-skewed. So we decided to take log of sale price in a way to obtain normal distribution for sale price.

Transformation of Sale Price



We also noticed that most of the variables are positively skewed so as the sale price like 1stfloor surface area, second floor surface area, BedroomabvGr (bedrooms above basement level), LotFrontage, OpenporchSF etc.

While data exploration we noticed that some features had outliers like sale price was very huge for medium garage area. We have removed those outliers from the data.

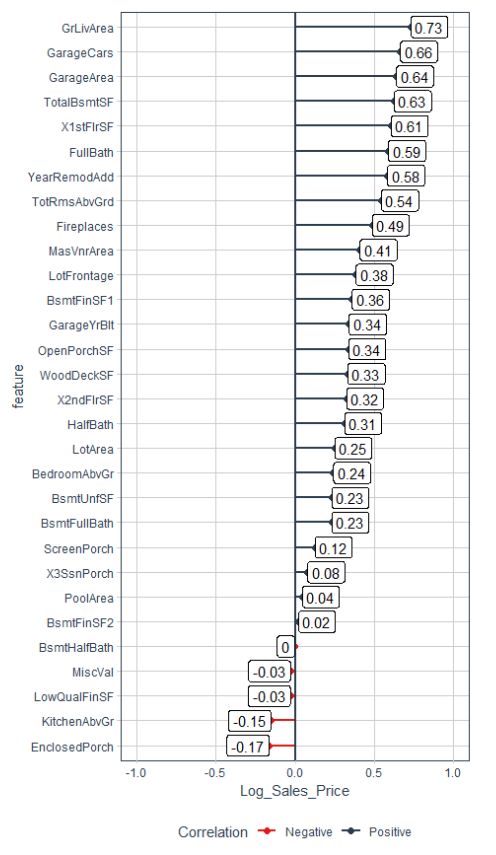


**VARIABLE SELECTION**

When selecting our variables, we decided first to see which predictors would someone who was buying a house choose. The ones that stood out right away were Living Area, Bathroom, number of bedrooms, and total rooms. When most people speak about houses, one of the first things they mention is the number of bedrooms and bathrooms so we assumed those would be important factors in buying a house. For Living Area and total Basement Square Footage, we knew we did not need to use the 1st and 2nd floor breakdowns because they were added for total living area and total basement square footage. Below graph shows comparison between log of sale price and above the ground living area square feet. We can see some strong pattern between them.

Next we included garage area and garage cars because of location, Iowa is a state that most people drive in so there will be a need for cars. Also, Iowa snows regularly, so an enclosed space, such as a garage, would be very beneficial for whomever is buying the house. This factor also made us include heating quality because Iowa is known to get very cold in the winters.

**CORRELATION PLOT**



The above chart shows correlation plot between numerical variables and log of sale price. We choose top 10 variables which had correlation above 50 percent to reduce our total number of variables to enter into the model.

**NEW VARIABLE CREATION AND TRANSFORMATION**

The dataset contains rich information about all the factors which a home buyer will consider while buying a house. We created new intuitive variables to draw more meaningful insights.

* Total number of bathrooms : we decided to create our own variable for bathrooms, by adding up basement half and full bathroom, and above ground half and full bathroom. The reason why we decided to create a new variable called total bathroom, was because we felt they each were important but we also just didn’t need four separate variables for bathroom.
* Two categories for year of built: Year of built had many different years. We decided to split it into two variables, old and new. If a house was built before 1950, it was considered old and if it was built after 1950, we considered it new. The reason we chose 1950 was because when we compared year built with sales price, we saw a huge increase in sale price after 1950.

**MODEL**

Our final list of variables is as follows

* GrLivArea(1st + 2nd SF) : Aboe the ground living surface area.
* GarageArea : Size of garage in square feet.
* TotalBsmtSF(BSMT 1st + 2nd) : Total square feet of basement area.
* Add 4 bathroom variables up: Total Bathroom
* GarageCars: Size of garage in car capacity.
* TotRmsAbvGrd : Total rooms above the ground does not include bathroom
* BedroomAbvGr : Number of bedrooms
* YearBuilt : Divided this variable in two categories old and new based.
* ExterQual : Exterior material quality
* Neighborhood : Physical location
* BldgType : Type of dwelling
* OverallQuality : Overall material and finish quality
* HeatingQC : Heating quality and condition
* SaleType = COD, Contract, New, WD/

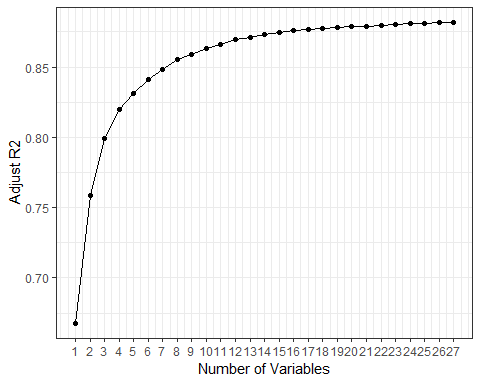
We decide to spilt categorical variables into dummy variables. In total we had fourteen variables chosen, but after making dummy variables, this number increased to 43.

**MODEL DEVELOPMENT.**

We developed a regression model with all the raw variables, dummy variable and newly created variables. The preliminary model had numerous statistically significant variables. However, variables were dropped iteratively to optimize the number of variables present in the final model. We measured the test error and fine-tuned the model to

reduce complexity while retaining predictive power.

In the below chart we plotted all the variables corresponding to adjusted R square and we noticed that as we increased number of variables adjusted R square also increases and becomes stagnant after adding 27 variables. Hence we decided to move forward with 27 variables.



Final model output

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 10.73 | 0.149 | 71.99 | 0 |
| **OverallQual2** | 0.1014 | 0.1645 | 0.6164 | 0.5377 |
| **OverallQual3** | 0.3856 | 0.1503 | 2.565 | 0.01046 |
| **OverallQual4** | 0.4933 | 0.1485 | 3.321 | 0.0009277 |
| **OverallQual5** | 0.59 | 0.1489 | 3.963 | 7.913e-05 |
| **OverallQual6** | 0.6613 | 0.1491 | 4.435 | 1.022e-05 |
| **OverallQual7** | 0.7453 | 0.1497 | 4.979 | 7.485e-07 |
| **OverallQual8** | 0.8337 | 0.1509 | 5.527 | 4.145e-08 |
| **OverallQual9** | 0.9797 | 0.1532 | 6.395 | 2.437e-10 |
| **OverallQual10** | 0.9975 | 0.1597 | 6.248 | 6.109e-10 |
| **GrLivArea** | 0.000247 | 1.652e-05 | 14.95 | 7.292e-46 |
| **YearBuilt\_Old** | -0.1066 | 0.01748 | -6.095 | 1.55e-09 |
| **TotalBsmtSF** | 0.0001171 | 1.343e-05 | 8.719 | 1.125e-17 |
| **GarageArea** | 0.0002069 | 2.765e-05 | 7.484 | 1.559e-13 |
| **Total\_Bathroom** | 0.0602 | 0.006669 | 9.027 | 8.624e-19 |
| **Neighborhood\_Crawfor** | 0.1851 | 0.02549 | 7.263 | 7.511e-13 |
| **HeatingQC\_TA** | -0.0742 | 0.01164 | -6.373 | 2.807e-10 |
| **BldgType\_Twnhs** | -0.1449 | 0.03064 | -4.73 | 2.565e-06 |
| **BldgType\_Duplex** | -0.1187 | 0.02278 | -5.212 | 2.258e-07 |
| **Neighborhood\_IDOTRR** | -0.1622 | 0.03145 | -5.157 | 3.015e-07 |
| **Neighborhood\_OldTown** | -0.08257 | 0.02128 | -3.881 | 0.0001108 |
| **Neighborhood\_ClearCr** | 0.1245 | 0.03154 | 3.947 | 8.44e-05 |
| **HeatingQC\_Fa** | -0.1058 | 0.02493 | -4.246 | 2.376e-05 |
| **Neighborhood\_BrDale** | -0.1592 | 0.04734 | -3.362 | 0.0008012 |
| **Neighborhood\_MeadowV** | -0.1075 | 0.04406 | -2.439 | 0.0149 |
| **Neighborhood\_NridgHt** | 0.1014 | 0.02392 | 4.239 | 2.446e-05 |
| **Neighborhood\_Veenker** | 0.2062 | 0.07009 | 2.942 | 0.003333 |
| **HeatingQC\_Gd** | -0.03576 | 0.01303 | -2.745 | 0.006166 |
| **Neighborhood\_Edwards** | -0.03334 | 0.01841 | -1.811 | 0.07043 |
| **ExterQual\_Fa** | -0.1066 | 0.04307 | -2.475 | 0.01349 |
| **BldgType\_TwnhsE** | -0.06654 | 0.01942 | -3.427 | 0.0006356 |
| **BedroomAbvGr** | -0.01939 | 0.007505 | -2.583 | 0.009923 |
| **Neighborhood\_StoneBr** | 0.1272 | 0.04506 | 2.822 | 0.004859 |
| **Neighborhood\_Timber** | 0.05862 | 0.02692 | 2.177 | 0.02969 |
| **Neighborhood\_Somerst** | 0.03764 | 0.02095 | 1.796 | 0.07272 |
| **Neighborhood\_BrkSide** | 0.04589 | 0.02637 | 1.74 | 0.08217 |

Fitting linear model: Log\_Sales\_Price ~ .

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 1056 | 0.1389 | 0.8869 | 0.8812 |

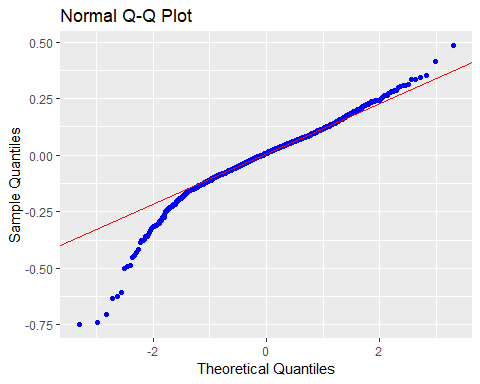
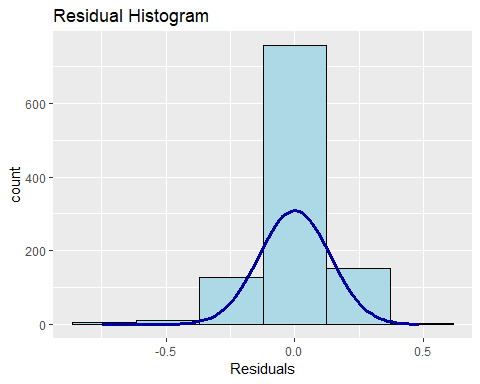
**MODEL EQUATION :**

Log of Sales Price = 10.73 + 0.1014 \* OverallQual2 + 0.3856 \* OverallQual3 + 0.4933 \* OverallQual4 + 0.59 \* OverallQual5 + 0.6613 \* OverallQual6 + 0.7453 \* OverallQual7 + 0.8337 \* OverallQual8 + 0.9797 \* OverallQual9 + 0.9975 \* OverallQual10 + 0.000247 \* GrlivArea - 0.1066 \* YearBuilt\_old + 0.0001171 \* TotalBsmtSF + 0.0002069 \* GarageArea + 0.0602 \* Total\_Bathroom + 0.1851 \* Neighborhood\_Crawfor – 0.0742 \* HeatingQC\_TA – 0.1449 \* BldgType\_Twnhs – 0.1187 \* BldgType\_Duplex – 0.1622 \* Neighborhood\_IDOTRR – 0.08257 \* Neighborhood\_OldTown + 0.1245 \* Neighborhood\_ClearCr - 0.1058 \* HeatingQC\_Fa – 0.1592 \* Neighborhood\_BrDale – 0.1075 \* Neighborhood\_MeadowV + 0.1014 \* Neighborhood\_NridgHt + 0.2062 \* Neighborhood\_Veenker – 0.03576 \* HeatingQC\_Gd – 0.03334 \* Neighborhood\_Edwards – 0.1066 \* ExterQual\_Fa – 0.06654 \* BldgType\_TwnhsE – 0.01939 \* BedroomAbvGr + .1272 \* Neighborhood\_StoneBr + 0.05862 \* Neighborhood\_Timber + 0.03764 \* Neighborhood\_Somerst + 0.04589 \* Neighborhood\_BrkSide

We can see from the modeling equation that neighbourhood like StoneBr, Timber, Brkside, Somrest shows a positive relationship with sale price whereas neighbor like meadow and Brdale shows negative relationship with sale price,

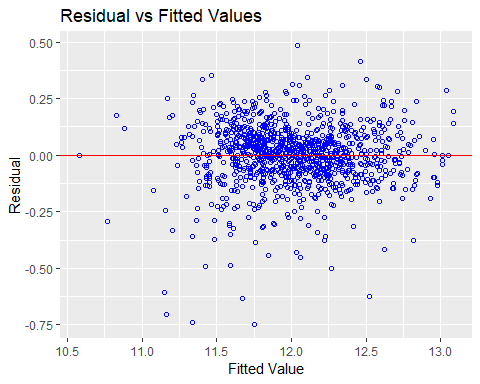
**MODEL DIAGNOSTICS AND FORECASTING**

**Residual analysis**

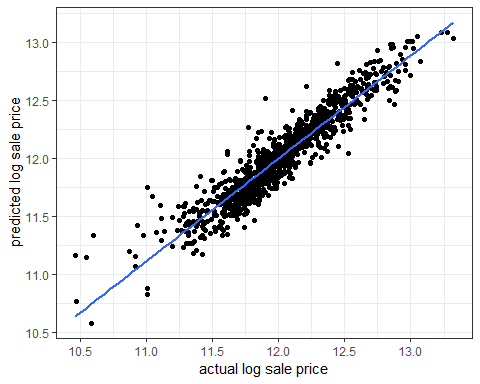
Looking at the QQ plot and Histogram of the residuals we can say that residuals of the model are normally distributed with mean 0.

**Residuals vs fitted values.**



Looking at the above plot we can see that residuals doesn’t have any relationship with fitted values and its randomly distributed. Residuals spread roughly around zero indicating linear relationship and homogeneity of error variance.

**Correlation between actual log of Sale Price vs predicated log of Sale Price**



The correlation between actual and predicated sale price came out to be 94%.

|  |  |
| --- | --- |
| measure | value |
| test\_bias | 751.8 |
| test\_Max\_Dev | 94781 |
| test\_Mean\_Abs\_Dev | 15563 |
| test\_Mean\_Sq\_Err | 448613544 |

**CONCLUSION**

For sales price, we decided to normalize the data by taking the log and then converted it back to regular price by using the exponential afte