Statistical Modelling Project

Shweta Saini 251387415
Maximilian Ho 251267689

Introduction

The goal of this project was to train a multiple linear regression model to predict house prices in an American city (Ames, Iowa) using information on the houses such as floor area and number of rooms. Real estate professionals could save time by using this model instead of manually estimating a house's price.

Our model has the form:

$$\text{House Price} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

where x_j is the j-th predictor and β_j is the j-th predictor's coefficient. x_j can be a function of the original predictors.

The dataset used to train the model contained 2930 observations (houses) and 80 predictors (variables describing each house). The predictors included both numeric and categorical variables. The professor instructed us to only use the first 1000 observations.

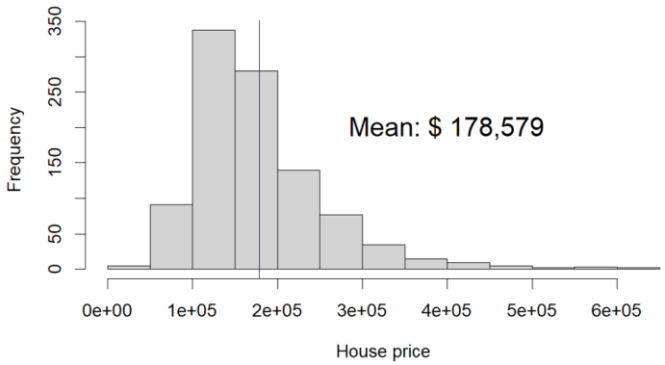
Summary statistics and data visualization

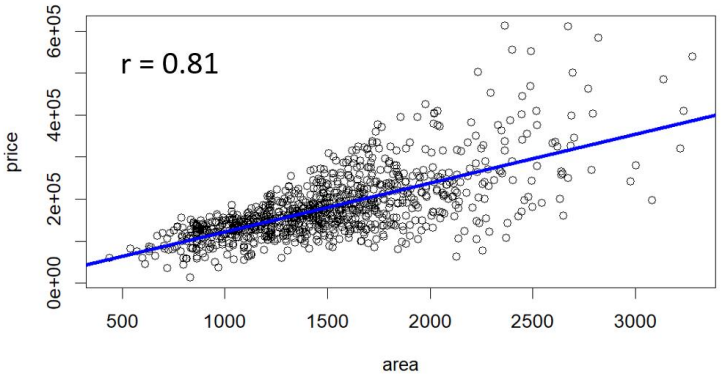
Variable selection using Lasso

To make our analysis manageable, we used **lasso regression** to select the best 13 predictors. Lasso regression is a regularization technique that penalizes the magnitudes of the regression coefficients to reduce overfitting. It can set coefficients to zero. The amount of regularization is controlled by the tuning parameter λ, which we manually adjusted to get the following 13 predictors:

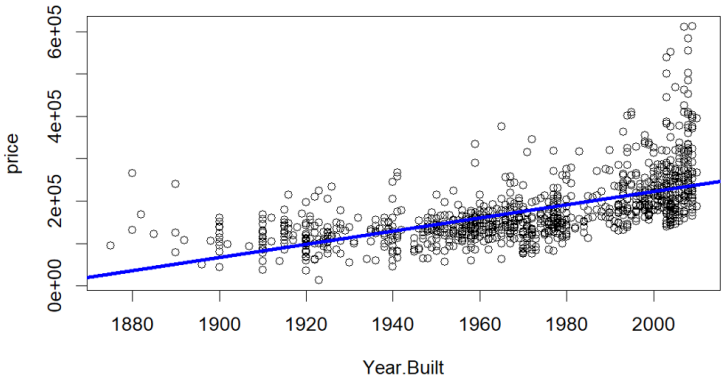
area	floor area above ground	**Exter.QualTA**	whether the construction quality of the exterior of the house is average (yes/no)
Lot.Area	area of the land that comes with the house	**BsmtFin.SF.1**	area of finished parts of basement
Neighborhood NridgHt	whether the house is in the Northridge Heights neighborhood (yes/no)	**TotalBsmt.SF**	total area of basement
Overall.Qual	construction quality of the house (1 - 10)	**X1st.Flr.SF**	area of first floor
Year.Built	year the house was built	**Garage.Cars**	how many cars can fit in the garage
Year.Remod.Add	year the house was remodeled (same as Year.Built if house was never remodeled)	**Garage.Area**	area of the garage
Mas.Vnr.TypeStone	whether the house has a stone masonry veneer (yes/no)		

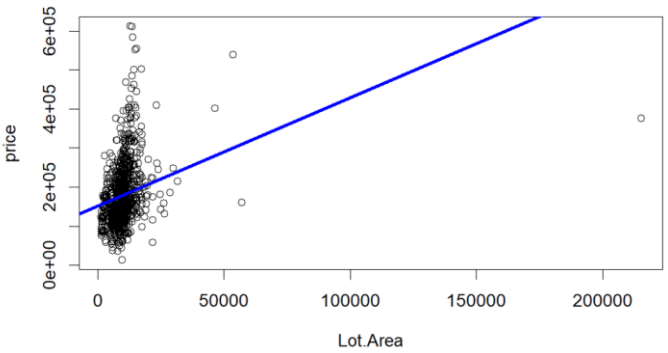
Relationships between predictors and house price





The target (house price) is positively skewed due to a few very expensive houses (left image). Of all the predictors, “area” was the most strongly correlated with house price (r = 0.81), indicating they have a strong linear relationship (right image).



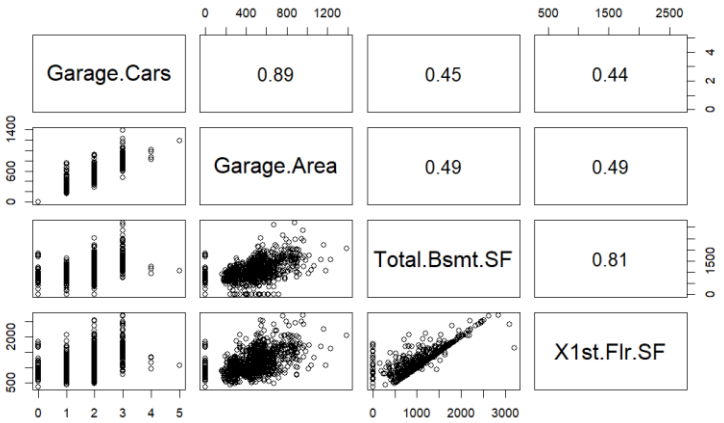


Some of the predictors had a non-linear relationship with house price, such as “Year.Built” (left image). There was also a house with a very large “Lot.Area” value (right image). We will address these issues below.

Methods

Removing multicollinearity

Mulicollinear predictors are correlated with each other. This increases the variance of their estimated coefficients. Hence, we want to remove multicollinear predictors to improve the interpretability of our model. A multicollinear predictors has a high **variance inflation factor** (VIF). The following predictors had large VIFs: “Garage.Cars” (5.4), “Garage.Cars” (5.0), “Total.Bsmt.SF” (3.7), and “X1st.Flr.SF” (3.6).

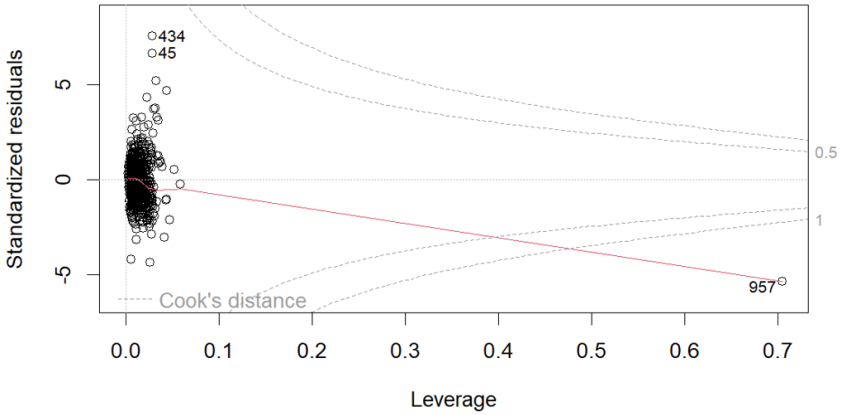


From the pair plot, we see that “Garage.Cars” and “Garage.Area” are highly correlated. This is because a larger garage can fit more cars. Similarly,”Total.Bsmt.SF” and”X1st.Flr.SF” are highly correlated. This is because a

house with a large 1st floor tends to have a large basement. We will remove ”Garage.Area” and ”X1st.Flr.SF” since they are less correlated with house price. After removing them, all the remaining predictors were statistically significant and had small VIFs.

Model 1: Original Predictors

After fitting the model, we checked for **influential observations**. An observation is influential if its deletion significantly changes the fitted model. An observation is influential if it is both an **outlier** and has high **leverage**. An outlier has an abnormal target value given its predictor values. Meanwhile, a high-leverage observation has an abnormal set of predictor values.



Observation 957 had a very large influence on the fitted model. This was the house with the very large ”Lot.Area” value we identified in our exploratory data analysis. It had a large negative residual (-$82,684), indicating the model greatly underestimated its price. To improve the fit on the other observations, we removed this observation from our training set and re-fit our model.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.036e+06  1.186e+05  -8.738  < 2e-16 ***
area               4.974e+01  2.481e+00  20.047  < 2e-16 ***
Lot.Area           1.953e+00  2.136e-01   9.146  < 2e-16 ***
NeighborhoodNridgHt 3.203e+04  4.390e+03   7.296 6.10e-13 ***
Overall.Qual       1.381e+04  1.106e+03  12.485  < 2e-16 ***
Year.Built         2.278e+02  4.309e+01   5.287 1.53e-07 ***
Year.Remod.Add     2.724e+02  5.867e+01   4.643 3.90e-06 ***
Mas.Vnr.TypeStone  1.776e+04  3.382e+03   5.251 1.85e-07 ***
Exter.QualTA      -1.035e+04  2.448e+03  -4.227 2.59e-05 ***
BsmtFin.SF.1       2.808e+01  2.352e+00  11.940  < 2e-16 ***
Total.Bsmt.SF      2.360e+01  2.821e+00   8.365  < 2e-16 ***
Garage.Cars        7.569e+03  1.643e+03   4.607 4.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value of each coefficient corresponds to the hypothesis test:

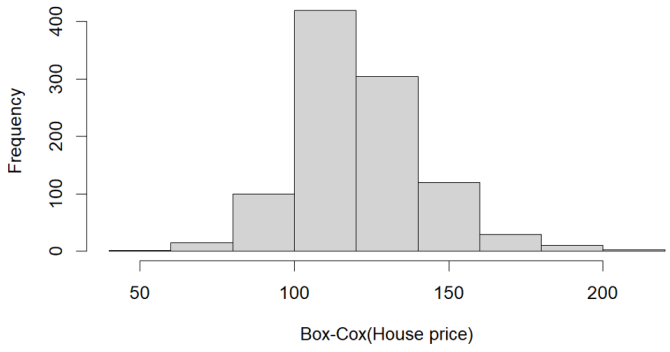
$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0$$

If a predictor’s p-value is smaller than 0.05, it indicates that the predictor is linearly related to the target, given that all the other predictors are used in the model.

Model 2: Box-Cox transformation

After conducting model diagnostics, we found that our model violated several of the **LINE** assumptions. These are the assumptions linear regression makes about the data. When they are violated, we cannot trust our linear regression coefficients. The assumptions are **linearity** (*y* has a linear relationship with each predictor), **independence** (the observations are independent), **normality** (the residuals follow a normal distribution), and **equal variance** (the variance of the residuals is equal for all $\hat{y}$).

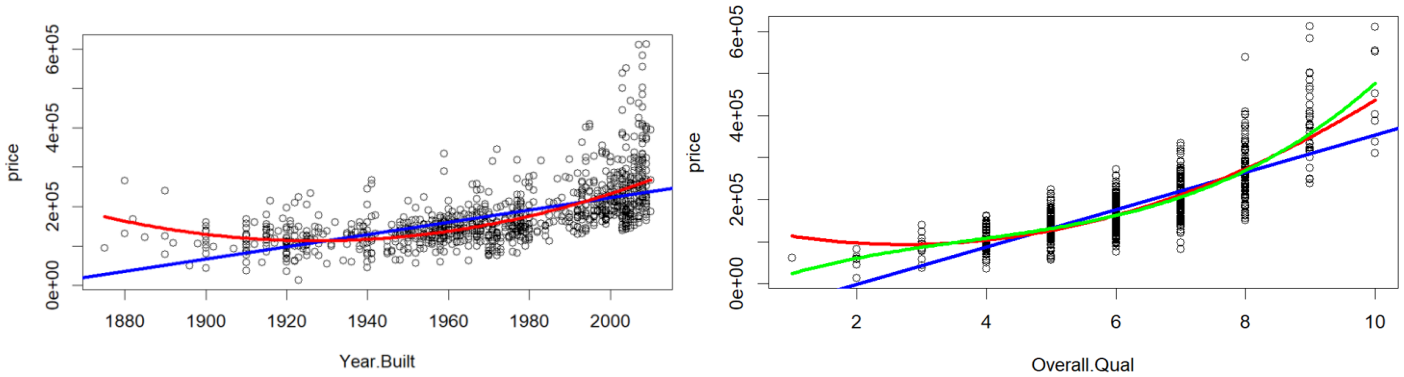
To try to fix the violations, we transformed the target using the **Box-Cox** transformation. It transforms the target such that its distribution resembles a normal distribution (left image).



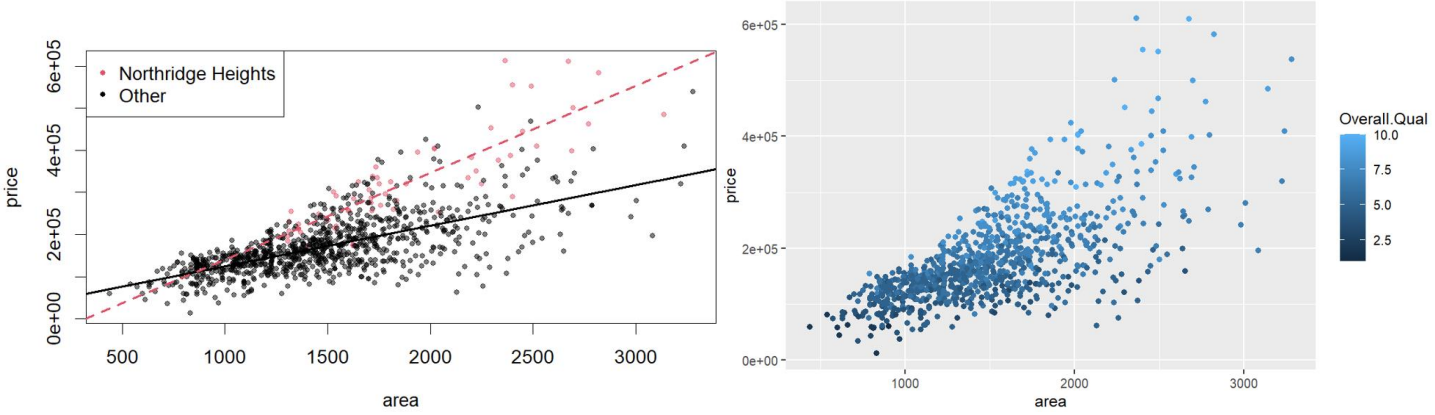

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -8.089e+02  7.064e+01 -11.450  < 2e-16 ***
area                 3.184e-02  1.478e-03  21.549  < 2e-16 ***
Lot.Area             1.214e-03  1.272e-04   9.543  < 2e-16 ***
NeighborhoodNridgHt  8.442e+00  2.614e+00   3.229 0.001282 **
Overall.Qual         1.032e+01  6.588e-01  15.664  < 2e-16 ***
Year.Built           2.314e-01  2.566e-02   9.016  < 2e-16 ***
Year.Remod.Add       2.546e-01  3.495e-02   7.285 6.59e-13 ***
Mas.Vnr.TypeStone    7.708e+00  2.014e+00   3.827 0.000138 ***
Exter.QualTA        -3.569e+00  1.458e+00  -2.447 0.014569 *
BsmtFin.SF.1         1.561e-02  1.401e-03  11.148  < 2e-16 ***
Total.Bsmt.SF        1.500e-02  1.680e-03   8.928  < 2e-16 ***
Garage.Cars          5.716e+00  9.784e-01   5.842 6.98e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model 3: Interaction terms

After the Box-Cox transformation, our model still violated the linear regression assumptions. Hence, we added some interaction terms based on our exploratory data analysis and domain knowledge.



A quadratic polynomial was sufficient to describe the relationship between “Year.Built” and house price (left image). Meanwhile, a cubic polynomial was sufficient to describe the relationship between “Overall.Qual” and house price (right image). We used the lowest-degree polynomial with a sufficient fit to prevent overfitting.



“area” and “NeighborhoodNridgHt” (categorical predictor) had an interactive effect on house price (left image). For the same unit increase in floor area, the price of houses in the Northridge Heights neighborhood increased more than houses in other neighborhoods. This is because Northridge Heights is a luxurious neighborhood.

Furthermore, “area” and “Overall.Qual” (numeric predictor) also had an interactive effect on house price (right image). For the same unit increase in floor area, the price of houses increased more when the construction quality of the house was high. This is because higher construction quality houses are more expensive to build.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -5.344e+03  2.719e+03  -1.965 0.049661 *
area                       5.715e-03  6.239e-03   0.916 0.359885
Lot.Area                   1.130e-03  1.238e-04   9.124  < 2e-16 ***
NeighborhoodNridgHt       -2.156e+01  1.028e+01  -2.097 0.036225 *
Overall.Qual               3.537e+01  8.030e+00   4.404 1.18e-05 ***
Year.Built                 4.783e+00  2.762e+00   1.732 0.083607 .
Year.Remod.Add             2.865e-01  3.610e-02   7.937 5.63e-15 ***
Mas.Vnr.TypeStone          7.904e+00  1.989e+00   3.973 7.61e-05 ***
Exter.QualTA              -5.320e+00  1.584e+00  -3.359 0.000812 ***
BsmtFin.SF.1               1.365e-02  1.384e-03   9.859  < 2e-16 ***
Total.Bsmt.SF              1.421e-02  1.634e-03   8.698  < 2e-16 ***
Garage.Cars                5.838e+00  9.543e-01   6.118 1.37e-09 ***
I(Year.Built^2)           -1.162e-03  7.074e-04  -1.643 0.100801
I(Overall.Qual^2)         -5.448e+00  1.427e+00  -3.818 0.000143 ***
I(Overall.Qual^3)          2.997e-01  7.890e-02   3.798 0.000155 ***
area:NeighborhoodNridgHt   1.354e-02  5.535e-03   2.446 0.014603 *
area:Overall.Qual          4.180e-03  1.008e-03   4.146 3.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results

Influential observations

We used the following definitions for unusual observations: An observation is influential if it has a Cook's distance greater than 4/(no. of observations). An observation is an outlier if it has a standardized residual greater than 2. An observation has high leverage if its leverage is greater than 2 times the average leverage of all observations.

	Model 1: Original	**Model 2: Box-Cox**	**Model 3: Higher-order terms**
No. of influential	75	62	60
No. of outliers	45	46	50
No. of high leverage	83	83	90

The Box-Cox transformation decreased the no. of influential observations. This is because the transformation compressed the large price differences between the very expensive houses. Adding the higher-order terms increased the no. of high leverage observations. This is because adding more predictors increased the dimensionality of the predictor space.

Model evaluation

The **leave one out cross-validation** (LOOCV) **root mean squared error** (RMSE) measures the test set prediction error of a model (the error for observations it was not trained on). Since the training set prediction error can be decreased to zero by overfitting the training set, we use the test set prediction error instead.

AIC, **BIC**, and **adjusted R-squared** can also be used to estimate a model's prediction performance when cross-validation is not possible. These metrics balance a model's goodness of fit and model complexity (no. of predictors). BIC prefers smaller models than AIC. The model with the smallest AIC or BIC is preferred. **R-squared** is the proportion of variation in the target that is explained by the predictors. Since, R-squared always increases as the no. of predictors increases, we use adjusted R-squared instead. The model with the largest adjusted R-squared is preferred.

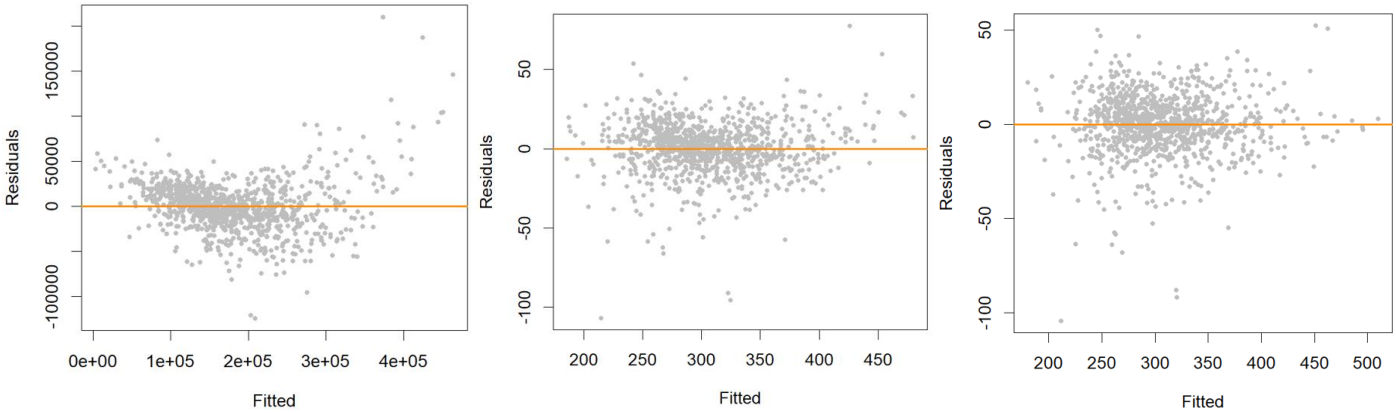
	Model 1: Original	**Model 2: Box-Cox**	**Model 3: Interaction terms**
LOOCV RMSE	$28,330	$23,919	$22,316
AIC	23,302	8,465	8,340
BIC	23,366	8,529	8,500
Adjusted R-squared	0.87	0.89	0.90

The Box-Cox transformation decreased the LOOCV RMSE, indicating it decreased the test set prediction error of the model. Furthermore, it decreased the AIC and BIC, and increased the adjusted R-squared, indicating it improved the goodness of fit.

Adding the higher-order terms slightly decreased the LOOCV RMSE. Furthermore, it decreased the AIC and BIC slightly, and increased the adjusted R-squared. This indicates that it significantly improved the goodness of fit (enough to overcome the effect of increasing the no. of predictors).

Model diagnostics

A **residual plot** is used to identify any patterns in a model's residuals ($y - \hat{y}$).



The leftmost residual plot indicates that model 1 violated the linearity assumption, because the residuals were not centered around 0 for each fitted value. The middle residual plot indicates that the Box-Cox transformation fixed this violation.

The **Breusch-Pagan test** is used to check the equal variance assumption:

$H_0 : Var(\varepsilon)$ is constant for all $\hat{y}$

$H_a : Var(\varepsilon)$ varies depending on $\hat{y}$

The **Shapiro-Wilk test** is used to check the normality assumption:

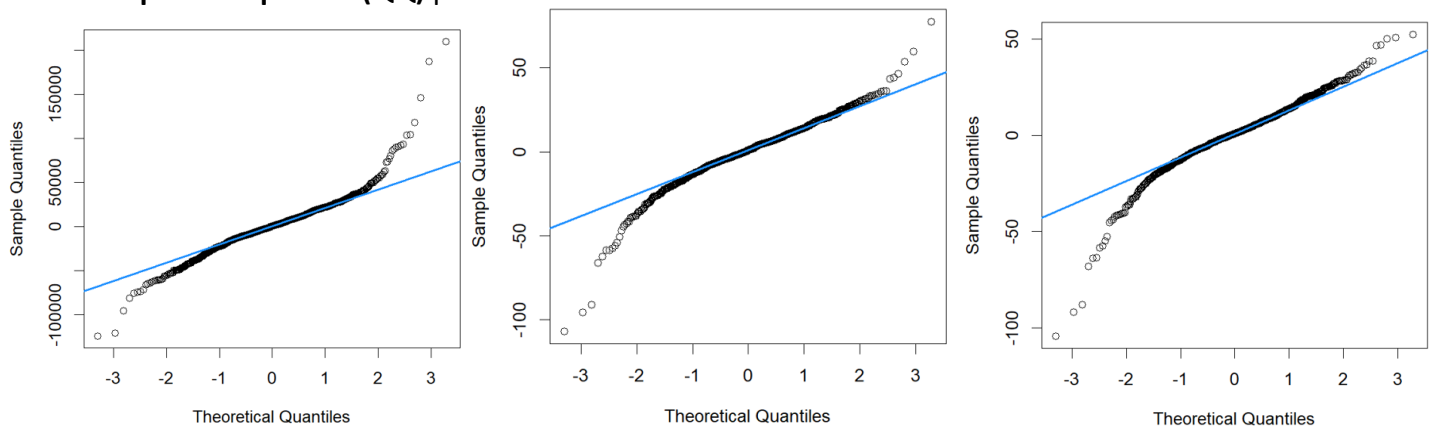
$H_0 : \varepsilon$ is normally distributed

$H_a : \varepsilon$ is not normally distributed

	Model 1: Original	**Model 2: Box-Cox**	**Model 3: Interaction terms**
Breusch-Pagan	p < 2.2e-16	p = 1.4e-06	p = 6.34e-07
Shapiro-Wilk	p < 2.2e-16	p < 2.2e-16	p < 2.2e-16

The Box-Cox transformation improved the violation of the equal variance assumption (increased the p-value of the Breusch-Pagan test) but was unable to fix the violation (the p-value was still smaller than 0.05).

A **normal quantile-quantile(QQ)** plot is used to determine if a variable follows a normal distribution.



Both the Box-Cox transformation (middle image) and higher-order terms (right image) were unable to fix the violation of the normality assumption (the p-values of the Shapiro-Wilk test were all smaller than 0.05, none of the normal QQ plots followed a straight line).

Discussion

Our final linear regression model was:

Predicted House Price =
− 5300 + 0.0057 **area** + 0.0011 **Lot.Area**
− 21 **NeighborhoodNridght** + 35 **Overall.Qual** + 4.8 **Year.Built**
+ 0.29 **Year.Remod.Add** + **7.9 Mas.Vnr.TypeStone** − 5.3 **Exter.QualTA**
+ 0.014 **Bsmt.Fin.SF.1** + 0.014 **Total.Bsmt.SF** + 5.8 **Garage.Cars**
− 0.0012 [**Year.Built**]^2 −5.4 [**Overall.Qual**]^2 + 0.30 [**Overall.Qual**]^3
+ 0.013 (**area** × **NeighborhoodNridgHt**) + 0.0042(**area** × **Overall.Qual**)

Interpreting the model:

1. Interpreting coefficients. For every unit increase in "area", the predicted house price increases by $(0.0057 + 0.013 × NeighborhoodNridgHt), given that the values of all the other predictors are held constant.

2. Valid prediction range. Linear regression cannot be used for extrapolation, so predictions that are outside of the range of house prices in the training set are invalid (valid predictions must be between $12,789 and $611,657).

3. Confidence intervals for coefficients. The (1 − α)% confidence interval for each coefficient is given by:

$$\mathrm{CI}(\beta_j, \alpha) = \hat{\beta}_\mathrm{j} \pm (\mathrm{t}_{\alpha/2,\mathrm{df}}) \left(\mathrm{SE}(\hat{\beta}_\mathrm{j})\right)$$

For example, when α = 0.05, there is a 95% probability that CI($\widehat{\beta_J}$) includes the population value β_j. $\widehat{\beta_J}$ is the least squares estimate of β_j. $\widehat{\beta_J}$ is the best linear unbiased estimator (smallest variance) of β_j (the Gauss Markov theorem).

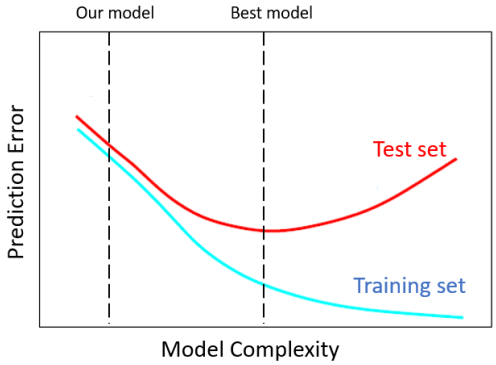
Limitations of the model:

1. High prediction error. Our model's house price test set predictions are, on average, wrong by $15,769 (LOOCV mean absolute error [MAE]). This is 11% of the mean house price, so we don't recommend using the model in real life. Meanwhile, the training set MAE was $15,441. Since both the LOOCV MAE and training set MAE were large, our model is underfitting the data.

2. Assumption violations. Our model violated the normality and equal variance assumptions of linear regression, so we cannot trust its coefficients and their standard errors.

Suggestions:

1. Increase model complexity. Our model is underfitting the data so we can increase the model complexity by adding more predictors and adding more interaction terms. As we increase the model complexity, the test set error will decrease. However, beyond a certain model complexity, the test set error will start to increase, because the model will start overfitting the training data (fitting to the noise in the training data). This U-shaped relationship between test set error and model complexity is called the **bias-variance tradeoff**. To fix overfitting, we could use regularization techniques like lasso regression or ridge regression to decrease the model complexity.



Appendix (R Code)

Refer to the other pdf file.