MSAN 691 Linear Regression Case Study

Dixin Yan, Yu Tian, Jade Yun, Zhengjie Xu

10/5/2017

Part I Exploratory Modeling

Task 1

For this part of the case study, our goal is to determine and explain what features of a house (the regressors) are most relevant in determining its expected sales price (the regressand). Since we are concerned with the explanatory power of the model, we need to validate assumptions for our final model.

Procedures:

1. Exploratory Data Analysis:

- Examine the structure of raw data: total of 1470 observations and 81 variables
- Identify the data type of variables: composed of 43 factor/character variables and 38 numeric variables
- Identify variables with NAs: total 17 variables with NA values
- Imputation of missing values

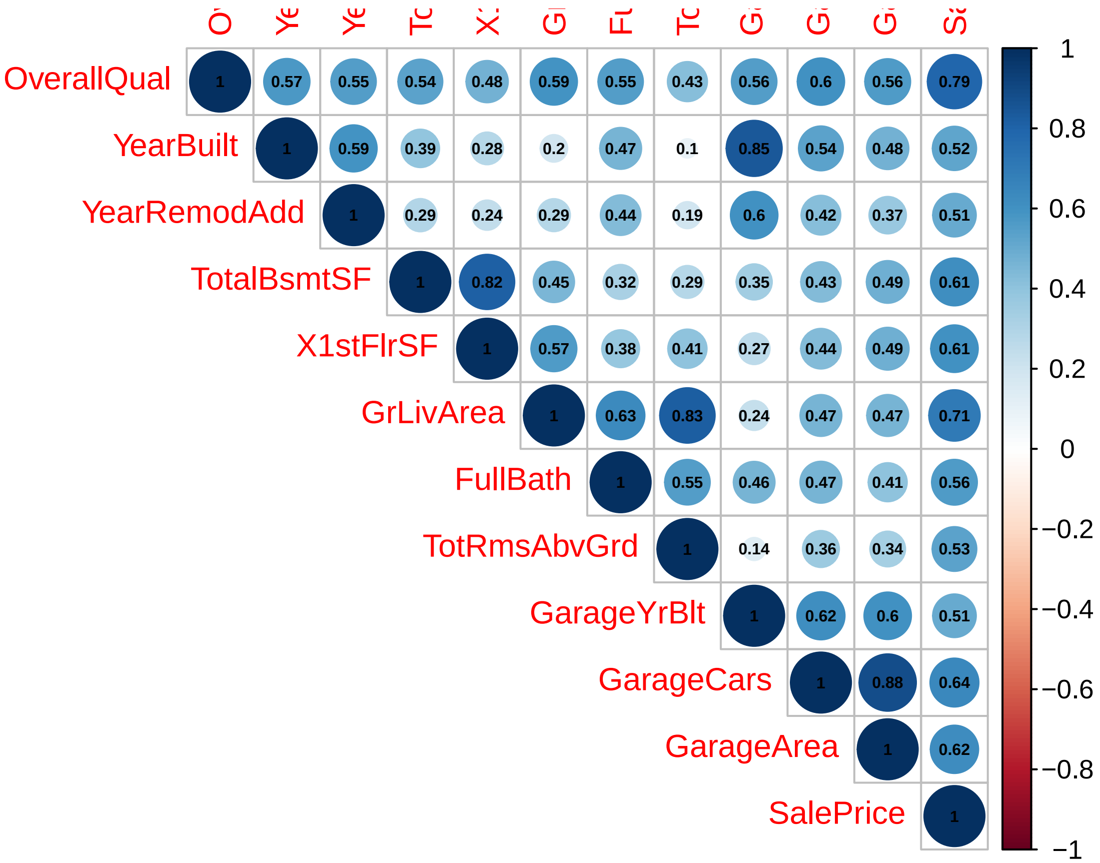
Missing Values:

We notice that R would interpret those missing values as the built-in NA values, which is not desirable as some have true meanings. It is important to determine whether the NA value is missing from the collection of data or it represents a property of the underlying variable. Below is the summary of the variables that contain NA values and ways of dealing with those values.

	Number of NAs	Data Type	NA_treatment
PoolQC	1453	character	No Pool
MiscFeature	1406	character	No MiscFeature
Alley	1369	character	No Alley
Fence	1179	character	No Fence
FireplaceQu	690	character	No Fireplace
GarageType	81	character	No Garage
GarageFinish	81	character	No Garage
GarageQual	81	character	No Garage
GarageCond	81	character	No Garage
BsmtExposure	38	character	No Basement
BsmtFinType2	38	character	No Basement
BsmtQual	37	character	No Basement
BsmtCond	37	character	No Basement
BsmtFinType1	37	character	No Basement
MasVnrType	8	character	No MasVnr
Electrical	1	character	take the mode
LotFrontage	259	integer	take the average
GarageYrBlt	81	integer	same as YrBlt
MasVnrArea	8	integer	0

Examine the correlation between variables:

There are 10 regressors exhibiting strong correlations with the regressand (`SalePrice`): `OverallQual`, `YearBuilt`, `YearRemodAdd`, `TotalBsmtSF`, `X1stFlrSF`, `GrLivArea`, `FullBath`, `TotRmsAbvGrd`, `GarageCars` and `GarageArea`. Please refer to the following graph for the visualization of the correlations between the regressand:



2. Variable Selection and Fit the Model

Build OLS model with all variables:

The adjusted R-squared value (0.92) and the F-statistics value (66.99) indicate that the OLS model can explain the variance in the dependent variable very well. Although that overall, all 81 regressors have significant effects on the response, when closely inspecting each regressor, the individual p-value of some are too large to have an significant effect on regressand. For example, the p-value of FireplaceQuFa (0.904715) indicates that its effect on regressand is minor. Thus, it is necessary for us to perform variable selection.

Check for Multicollinearity:

The variance inflation factor shows that there are variables that are affected by multicollinearity since some estimators have variance inflation factor greater than 10.

Use lasso to select variables:

To select important features, we run the LASSO model and leverage the feature that it would push some insignificant variables to zero.

The result shows that, for a factor, some levels are statistically significant while others are not. We can divide the variable into sub-variables corresponding to different levels. We decide to select the variable as long as it has at least one significant level.

After the above considerations, we have selected the following variables: `MSSubClass`, `MSZoning`, `LotArea`, `Street`, `LotShape`, `LandContour`, `Utilities`, `LotConfig`, `Neighborhood`, `Condition1`, `Condition2`, `BldgType`, `OverallQual`, `OverallCond`, `YearBuilt`, `YearRemodAdd`, `RoofStyle`, `RoofMatl`, `Exterior1st`, `Exterior2nd`, `MasVnrType`, `MasVnrArea`, `ExterQual`, `Foundation`, `BsmtQual`, `BsmtCond`, `BsmtExposure`, `BsmtFinType1`, `BsmtFinSF1`, `BsmtFinType2`, `TotalBsmtSF`, `Heating`, `LowQualFinSF`, `GrLivArea`, `BsmtFullBath`, `FullBath`, `BedroomAbvGr`, `KitchenAbvGr`, `KitchenQual,Functional`, `Fireplaces`, `GarageType`, `GarageFinish`, `GarageCars`, `GarageArea`, `GarageQual`, `WoodDeckSF`, `OpenPorchSF`, `ScreenPorch`, `PoolArea`, `PoolQC`, `SaleType`, `SaleCondition`.

Use selected variables to fit OLS model:

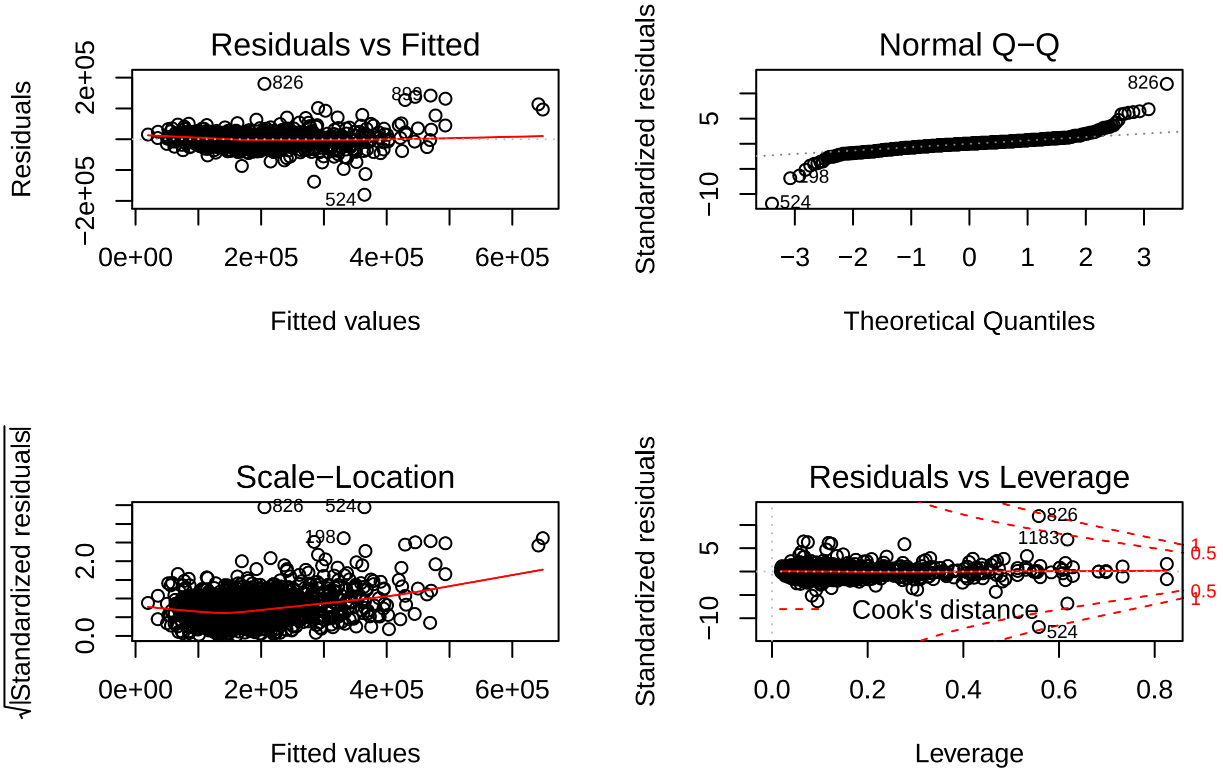
The result shows that after variable selection, almost every variable is statistically significant at level of 0.05. The overall F statistics is 83.74, the p-value is very small, and the Adjusted R-squared is 0.92. According to AIC criteria, our model has better performance after variable selection. Thus, we can say that the variables chosen are most relevant in determining the sale price.

From the size of the coefficients, we can tell that `RoofMaltl(roof material)`, `PoolQC(pool quality)`, `Roofstyle`, `Neighborhood` and `Condition` have the largest effects on `SalePrice`.

3. Validation of Model

A Quick Diagnostic Analysis

Please refer to the plot below.



Since residuals-versus-fitted plot shows approximately flat trend with equal vertical spreads, the linear model assumption holds.

Because the points on Q-Q plot does not show an approximate straight line, the error terms do not satisfy the normality assumption.

The Scale-Location does not follow a flat trend, which means that error terms do not follow homoscedasticity.

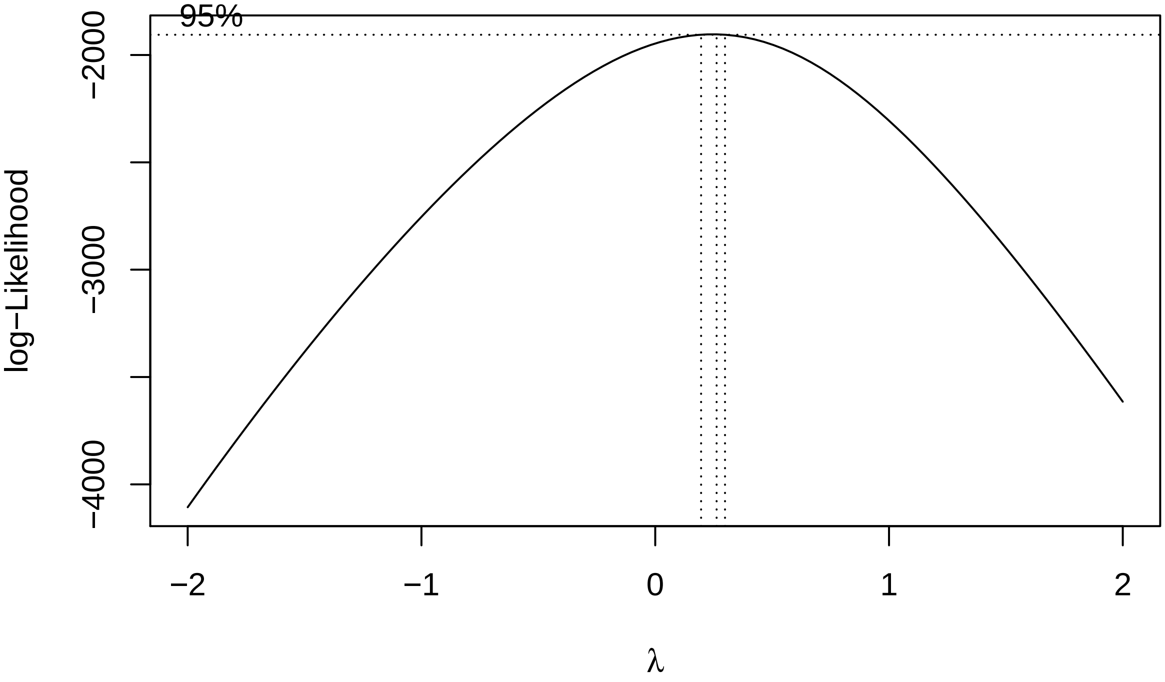
There are three points on residuals-versus-leverage plot falling outside of the red dashed lines, which indicates that these are the influential points. These three error points correspond to observation #524, #826 and #1183 respectively.

Formally test assumptions

The Shapiro-Wilk test indicates that normal distribution assumption of error terms is not satisfied. The non-constant variance test indicates that the variance of the error term is not constant. From the outlier test, we can see there are 10 outliers. Since the error terms are not normally distributed according to the Shapiro-Wilk test, the result of non-constant variance test and outlier test may be unreliable. It is essential to take this into consideration when we construct our model.

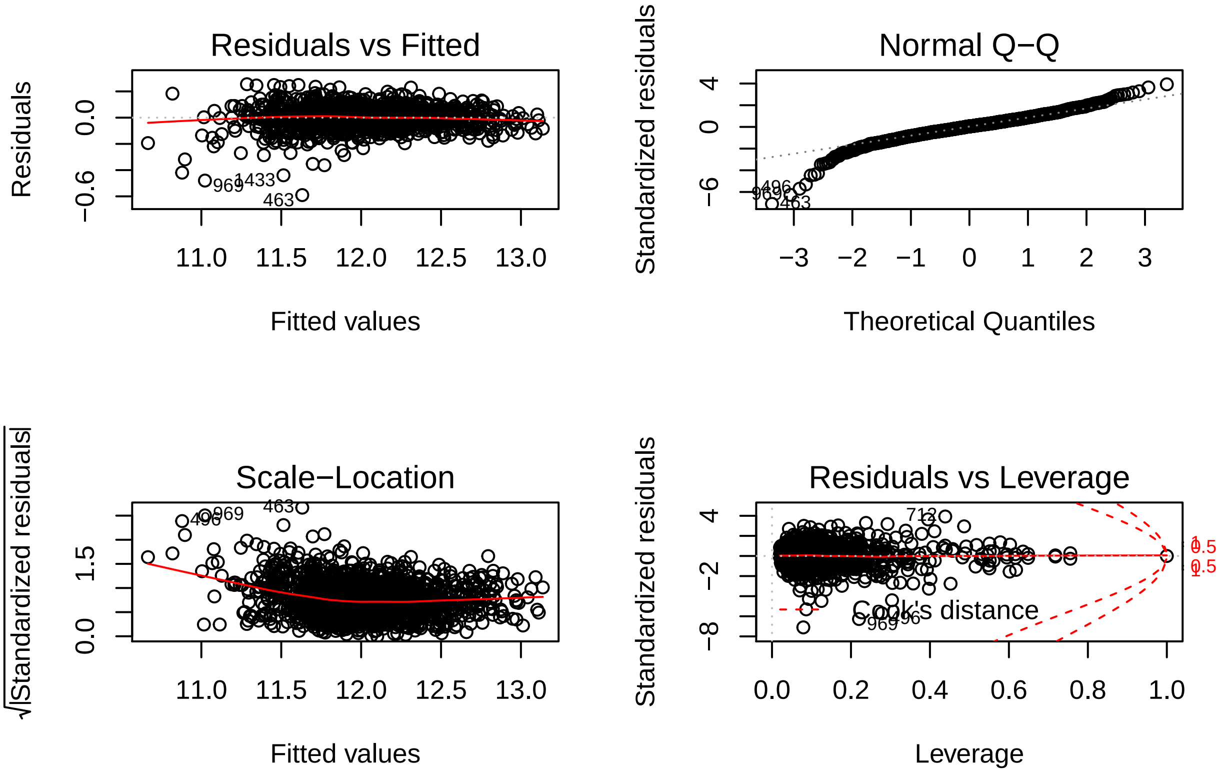
The above tests tell us that we need to apply transformation algorithms to improve the model performance. We choose to apply a box-cox transformation first, and further transform the independent variables that are highly skewed if the normality assumption still does not hold.

Box-Cox Transformation



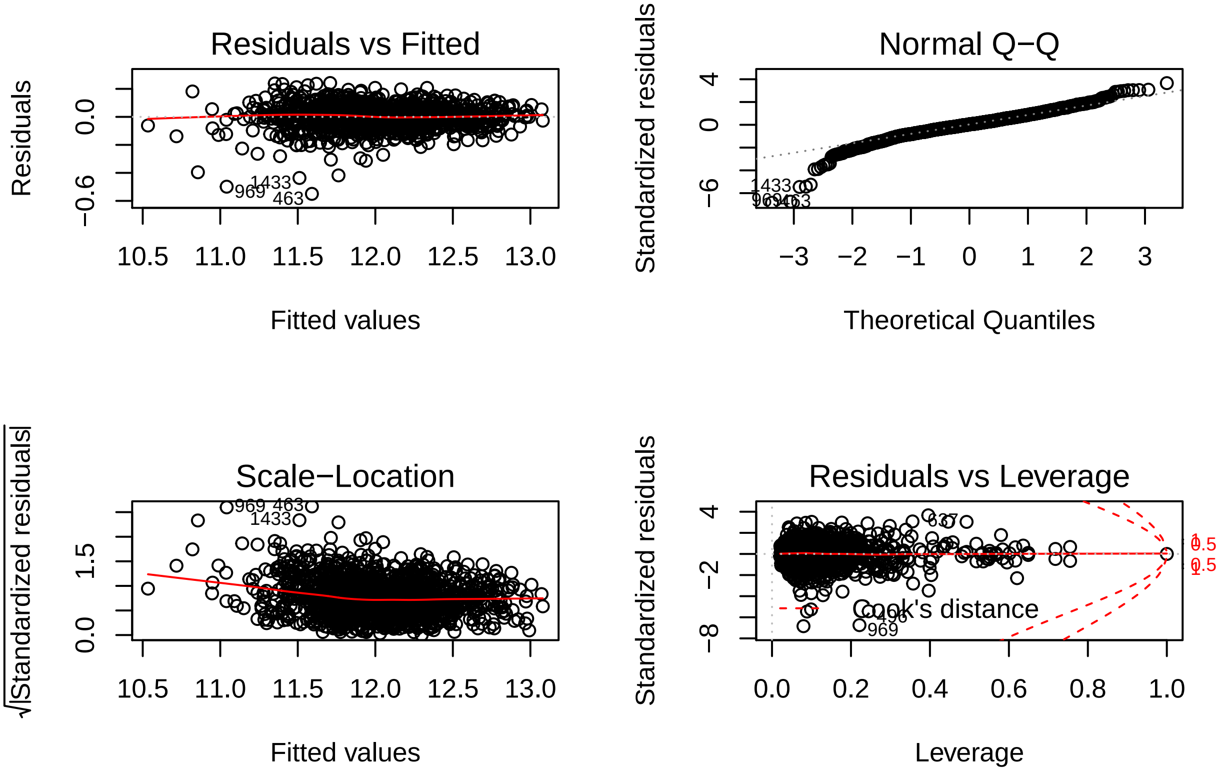
After running a Box-Cox test in R, we get a lambda approximating to 0 as shown in the Box-Cox plot above, which suggests a log transformation on the regressand values.

We fit the model again with log transformation on `SalePrice`. The diagnostics plot below shows that the problem with non-constant variance is less significant and the non-normality is sort of relieved at the upper tail (as seen from the top right plot in the chart below). However, error terms are still not normally distributed. Therefore, we further transform all skewed independent variables and then fit the model again.



Transform skewed regressor

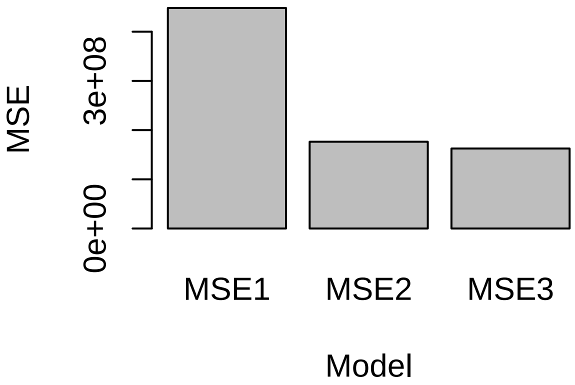
For numeric variables that are highly skewed, we choose to transform excessively skewed features with log(x+1). We get the diagnostics plots below.



Yet, after all the efforts we tried (box-cox transformation, independent variables transformation, outliers deletion), the normal and constant variance assumptions over error terms still do not hold.

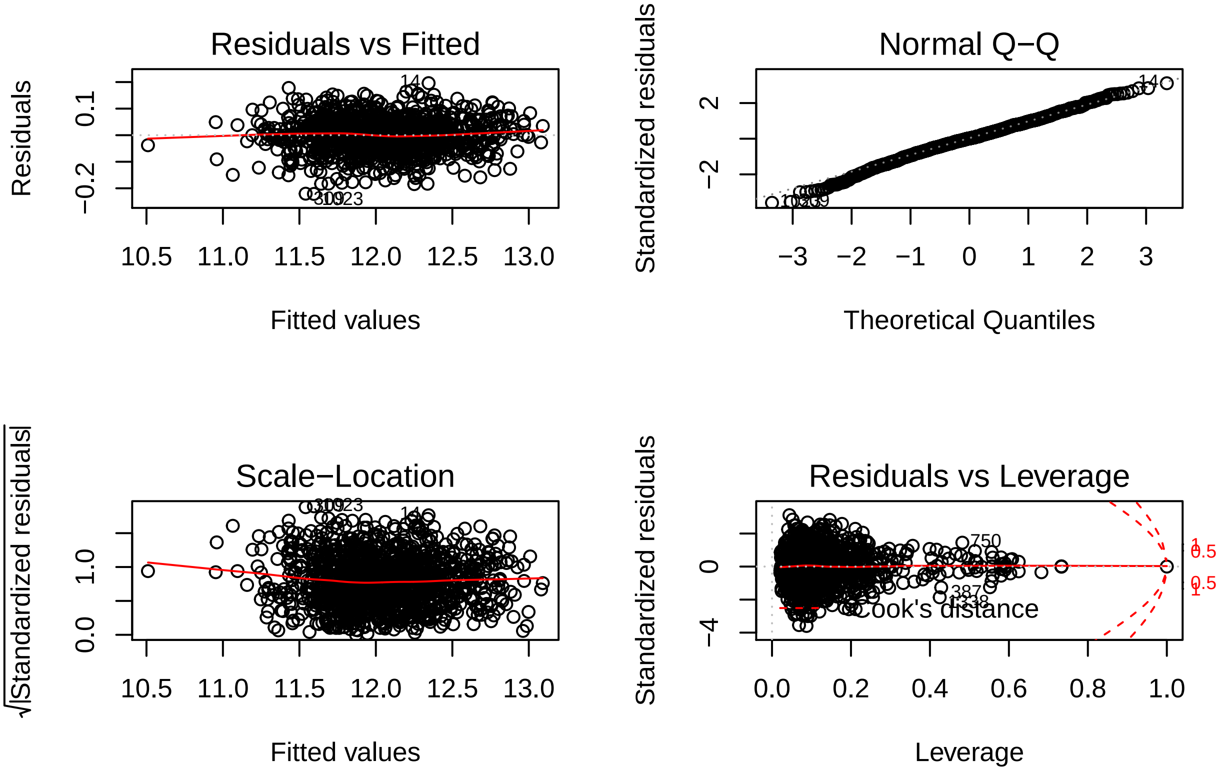
Compare the model accuracy

Fortunately, the model acquired has the lowest MSE after all transformations (as indicated by the plot below), which means that we successfully optimized our model to our best.



Other Consideration

We also tried to further remove influential points identified by the Cook's distance.



From the graph above, we see that the normality and constant variance assumptions for error terms are satisfied. However, the downside is that we have to drop 190 observations from the original dataset. This method is not appropriate because it is like we are altering data in order to fit the model. We choose to not deal with the influential points.

Task 2

Predict the maximum sale price for Morty

We use the Morty's data to fit the model to acquire the predicted sales price for his house. The upper bound of the prediction interval is used to obtain the maximum price, which is $182,501.9.

Aspects he can change to increase sales price

RoofMatlWdShngl	1.7945557
RoofMatlRoll	1.7124424
RoofMatlCompShg	1.6489273
RoofMatlWdShake	1.5757613
BsmtQualNo Basement	0.6566383
GrLivArea	0.4303601
MSZoningFV	0.3789280
MSZoningRH	0.3771031
MSZoningRL	0.3619275
Condition2PosA	0.3459462
MSZoningRM	0.3261318
SaleTypeOth	0.2578693
HeatingWall	0.2155665
HeatingGasW	0.1703366
HeatingOthW	0.1573460
RoofMatlMembran	0.1559984
HeatingGasA	0.1528611
SaleConditionAlloca	0.1420830
SaleTypeCWD	0.1380977
FoundationStone	0.1309674

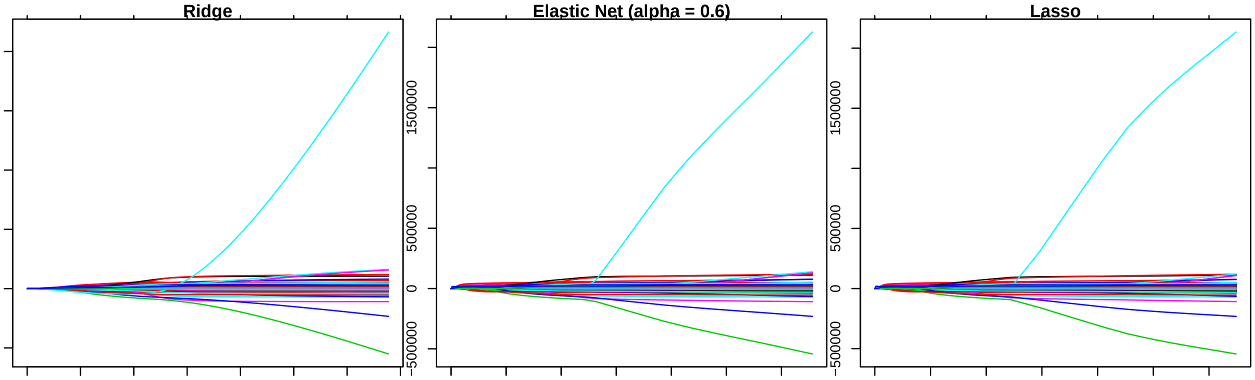
The table above presents the top 20 coefficients ordered by size after running our model. We discovered that `RoofMatl`(roof material), `BsmtQual` (basement quality), and `GrLivArea`(above ground living area) had the greatest impact on sales price. We recommend Morty to upgrade his basement quality to excellent, replace Roof Material from Compshg to Membran, and enlarge the above ground living area. These changes will increase the sales price of his house.

Part II Predictive Modeling

For this part of the case study, we need to come up with a predictive model that would best capture the sales price of a new house in the market. The models we considered include OLS, Ridge regression, LASSO regression, and Elastic Net regression. Since we are solely concerned with the prediction accuracy of the model, we do not validate the assumptions for each regression model. We choose the best model which gives the least MSPE for the testing data.

To begin with, we impute the missing values in the same way as in Part I. For each regression model of Ridge, LASSO, and Elastic Net, we search over the same grid of lambda, ranging from 10^{-2} to 10^{10}, for the best lambda respectively. We split the dataset into train set (50%) and test set (50%) by random sampling. Furthermore, to ensure comparability, we use the same train set and test set for each model.

Please refer to the following graphs for the different shrinking effects of the three models.



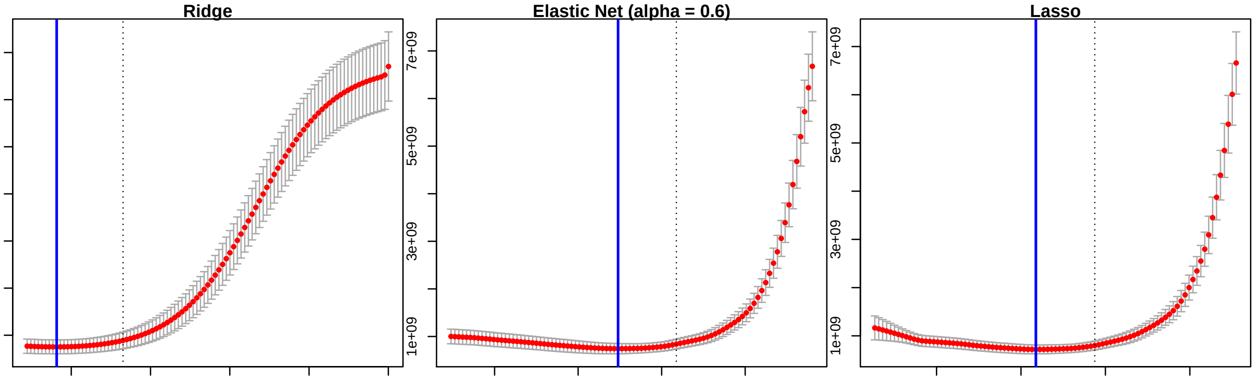
We see that LASSO reduces a number of estimators to zero, while Ridge shrinks the estimators but none will ultimately reach exact zero. Elastic Net has a combination of effects of both Ridge and LASSO.

The following describes the steps taken for fitting the Ridge regression:

- Create lambda grid
- Split dataset into train (50%) and test (50%) by random sampling
- Fit Ridge regression model (glmnet) on the train set of data with alpha = 0
- Cross-validate the model with cv.glmnet and choose the best lambda
- Predict the response on test set of data using the lambda chosen from step 4
- Calculate the MSPE from the prediction

For Elastic Net and Lasso, the changes are in different values of alpha. For Lasso we use alpha = 1, and for elastic net we use alpha = 0.2, 0.4, 0.6, 0.8, respectively.

The following graphs depict the best lambda selection process for each model. The best lambda corresponds to the lowest average MSPE within the cross-validation test sets.



For OLS method, we use the same train set to fit a Linear Regression model and get the MSPE by predicting the test set.

Comparing the results of MSPE from all models as shown below, Ridge has the least value and therefore we determine Ridge is the best model in this case.

Table 1: MSPE for Different Models

Model	MSPE
OLS	12,458,934,868
Ridge	1,530,415,655
Elastic.Net_0.2	1,617,495,572
Elastic.Net_0.4	1,628,006,047
Elastic.Net_0.6	1,636,572,023
Elastic.Net_0.8	1,639,512,301
Lasso	1,646,053,963

In the end, after choosing the best model (Ridge), we refit the entire dataset to get the final model.