# Set up

## Data cleansing:

* Use data contained in Housepricedata\_scaled as base features, and Original\_Data.xlsx as pool for additional features to select from
* Notice that data file has misaligned columns and also some headers has space in the name. We have removed spaces to align column names between two data files
* Categorical features that exhibits a natural order are mapped naively on a scale of 0 to length of possible values. For those who don’t have a natural order, they are mapped to numerical values as well instead of creating dummy variables. Notice that this is probably over simplification as we should have created dummy variables.

## Feature selection:

* Goal is to select two features in addition to *LotFrontage* and *LotShape*, i.e. a total of four additional features, to add to base features and train a model using either linear regression, ridge regression or lasso regression.
* Methodology: We will only look at features that would survive lasso regression if we use lasso on the full feature set, then select two features that minimizes MSE on validation set using ridge regression. From this process, we have determined that *MSSubClass* and *BsmtExposure* are significant in determining housing price. Hence our final feature selection is *MSSubClass*, *BsmtExposure, LotFrontage* and *LotShape.*

# Regression analysis part I

## Splitting data set:

* In this part, we split our training, validation and test set by simply slicing the original data and taking first 1800 as training set, next 600 as validation set and the rest as test set

## Regression results and observations

* When examining model fitting results for linear regression, we can observe that some features has absurdly large coefficients. This is likely due to “dummy variable trap” and it provides us an incentive to perform regularization
* We can observe that for ridge or lasso, above issue does not appear
* Indeed, for both ridge and lasso regression, with an increasing lambda value, we can observe a decrease in average magnitude of coefficients (refer to Appendix I)
* Ridge regression presents a lower MSE on validation set in general compare to lasso regression. MSE plots for ridge and lasso can be find in Appendix II.
* We chose to use ridge regression with a lambda value of 0.16 to perform final prediction on test set and achieved an MSE of 10.8%

# Regression analysis part II

## Splitting data set:

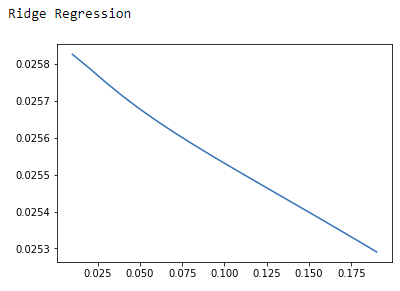
* For this part, we will randomly select 1800 data points for training set, 600 data points for validation set and the remaining goes to test set
* We have set a seed for randomization in order for consistent reproduction of results
* We also performed k-fold cross-validation with k=4 since we have 2400 data points in total for training + validation set, and 600 for validation set. Process for this is that we will segregate 2400 data points into 4 sets of 600 points, and each time use 3 of them for training and one of them for validation, and take the average on MSE and coefficients. Each of those 4 sets will only be used for validation exactly once. We opted to do this because we don’t have a large number of training data, and this way we can potentially reduce bias while preserving distribution.

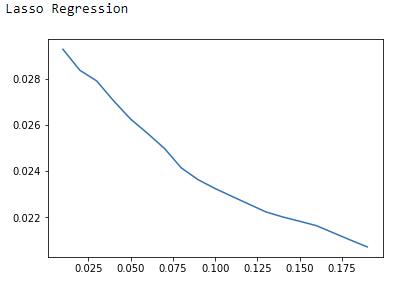
## Regression results and observations:

* Model fitting process is almost exactly the same as part I, except the fact that we are using k-fold cross-validation and taking averages on MSE and coefficients. (refer to Appendix III for MSE plots for ridge and lasso)
* Again, we used ridge regression and a lambda of 0.16 to perform final prediction on test set. We observed an MSE of 10.9% in this case with random data split and cross validation
* Notice that the MSE in part II is very close to part I. One conclusion could be that if our sample is already well shuffled, then randomly splitting data would not necessarily improve skill of our algorithm. However, in real application, we cannot guarantee our source data is always well shuffled, so we should still always perform random data splitting.

Appendix I

Average magnitude of coefficients vs Lambda level

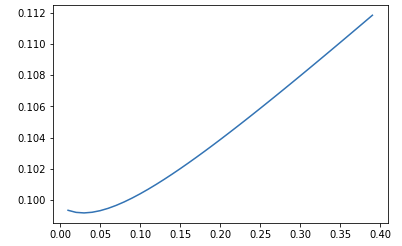




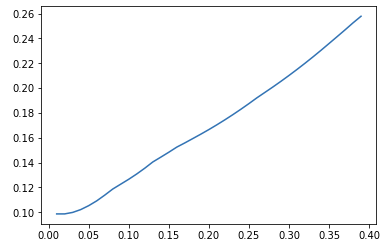
Appendix II

MSEs for ridge and lasso regression in Part I

Ridge regression



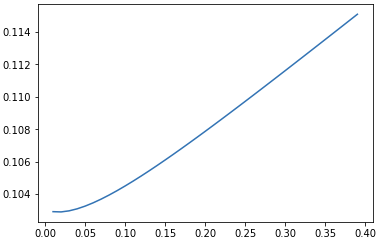
Lasso regression



Appendix III

MSEs for ridge and lasso regression in Part II

Ridge regression



Lasso regression

