Statistical Learning Homework 1

Yizhen (Jeremy) Dai, S2395479

1. Tasks

This report is the result of the first Statistical Learning homework, which goal is to apply linear regression on the Boston Housing Data[1], circa 1978. The dataset was originally taken from the StatLib library at Carnegie Mellon University. The linear regression model without regularization, which is optimized based on least square method, is used as our baseline. Then three ways are adopted as an effort to improve prediction accuracy: subset search method, ridge regression and lasso regression. After training the models, the four models are evaluated on the test dataset and their performance are compared.

2. Introduction

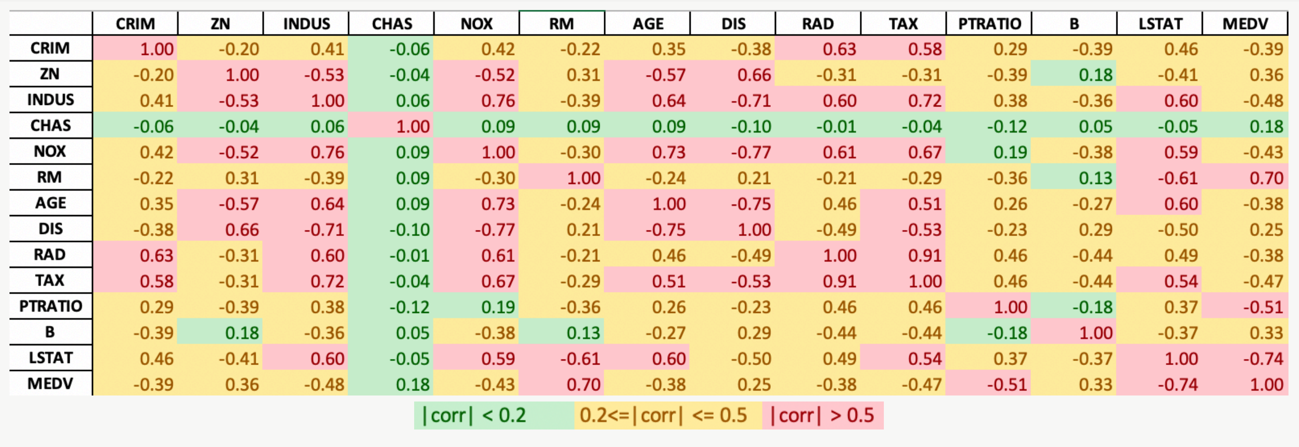
2.1 Explorative Data Analysis

The available dataset consists 506 data points with following fourteen variables:

1.	CRIM	per capita crime rate by town
2.	ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
3.	INDUS	proportion of non-retail business acres per town
4.	CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5.	NOX	nitric oxides concentration (parts per 10 million)
6.	RM	average number of rooms per dwelling
7.	AGE	proportion of owner-occupied units built prior to 1940
8.	DIS	weighted distances to five Boston employment centres
9.	RAD	index of accessibility to radial highways
10.	TAX	full-value property-tax rate per $10,000
11.	PTRATIO	pupil-teacher ratio by town
12.	B	1000*(Bk - 0.63)^2 where Bk is the black proportion by town
13.	LSTAT	% of lower status of the population
14.	MEDV	median value of owner-occupied homes in $1000's

The 'MEDV' variable, which is the median value of owner-occupied homes in the unit of $1000, is the response variable in the linear models. The rest thirteen rest variables will be treated as features in the linear models. Twelve of these variables are continuous attributes. However, 'CHAS' is a binary-valued variable, which will be treated as a categorical variable.

Table 1: The correlation table for all the variables



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1.00	-0.20	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	-0.39	0.46	-0.39
ZN	-0.20	1.00	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	0.18	-0.41	0.36
INDUS	0.41	-0.53	1.00	0.06	0.76	-0.39	0.64	-0.71	0.60	0.72	0.38	-0.36	0.60	-0.48
CHAS	-0.06	-0.04	0.06	1.00	0.09	0.09	0.09	-0.10	-0.01	-0.04	-0.12	0.05	-0.05	0.18
NOX	0.42	-0.52	0.76	0.09	1.00	-0.30	0.73	-0.77	0.61	0.67	0.19	-0.38	0.59	-0.43
RM	-0.22	0.31	-0.39	0.09	-0.30	1.00	-0.24	0.21	-0.21	-0.29	-0.36	0.13	-0.61	0.70
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24	1.00	-0.75	0.46	0.51	0.26	-0.27	0.60	-0.38
DIS	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75	1.00	-0.49	-0.53	-0.23	0.29	-0.50	0.25
RAD	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49	1.00	0.91	0.46	-0.44	0.49	-0.38
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1.00	0.46	-0.44	0.54	-0.47
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1.00	-0.18	0.37	-0.51
B	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1.00	-0.37	0.33
LSTAT	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	-0.37	1.00	-0.74
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	1.00

🟩 |corr| < 0.2 🟨 0.2<=|corr| <= 0.5 🟥 |corr| > 0.5

[1] Dataset source: http://www.timvanerven.nl/wp-content/uploads/teaching/statlearn2016/housing-p.data

Table 1 shows the pairwise correlation between all the variables. ‘MEDV’ is highly correlated with ‘RM’ and ‘LSTAT’ with a correlation of 0.70 and -0.74 respectively, which is in line with the intuition that the housing prices are related with socio-economic status and the house size.

Table 2: The variances for all the explanatory variables

Variable	Variance
CRIM	73.99
ZN	543.94
INDUS	47.06
CHAS	0.06
NOX	0.01
RM	0.49
AGE	792.36
DIS	4.43
RAD	75.82
TAX	28404.76
PTRATIO	4.69
B	8334.75
LSTAT	50.99

The variance of the explanatory variables differ from the smallest of 0.01 to as big as 28404.76. With with huge difference between feature variance, the variables features need to be normalized before lasso and ridge regression are applied.

2.2 Assumptions for Linear Regression

Before further analysis, the whole dataset is used ONLY to check if the assumptions for linear regression are true for this Boston housing price dataset. The models that will actually be built will NOT be trained on the whole dataset. This way, there will be no data leakage issue.

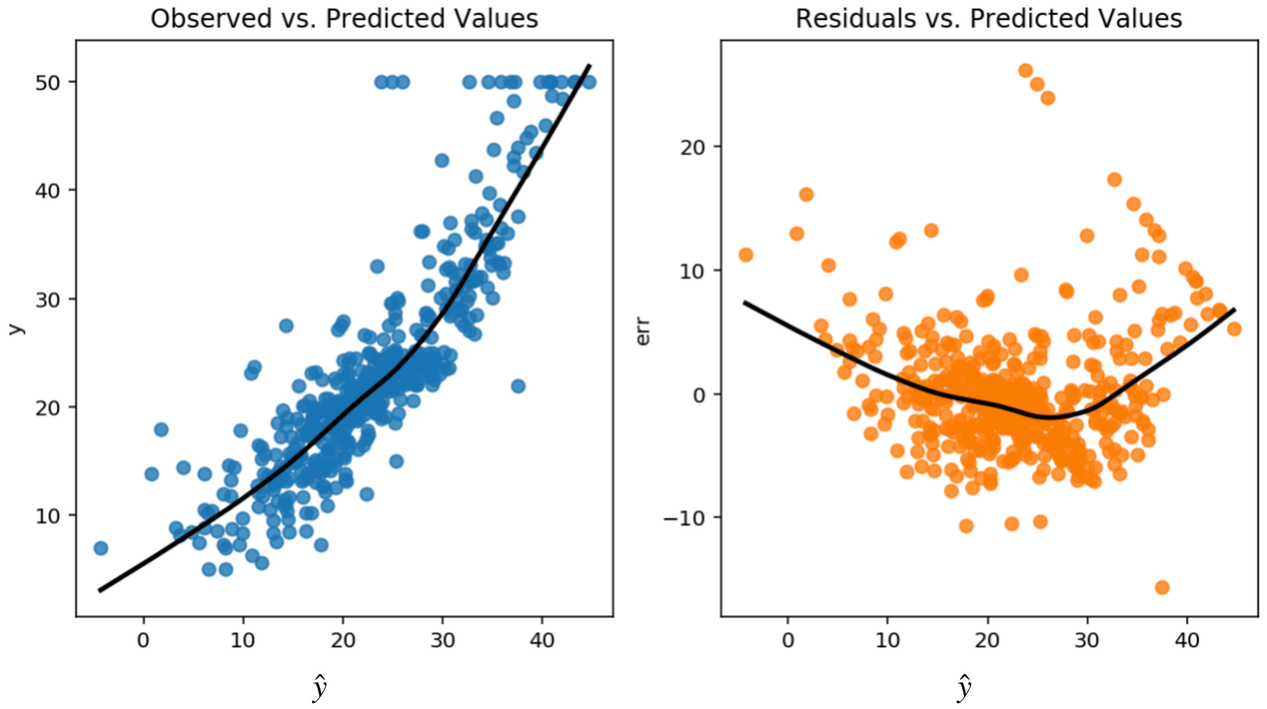


Figure 1: The scatter plots of observed values vs predicted values (left) and the residuals vs. predicted values (right) with fitted regression line

To meet the assumptions of linear regressions, the residual errors should be normally distributed with a common variance, and the features should be independent from each other (no multicollinearity).

Figure 1shows that the errors do not have equal variance. Also, the data might be manually cut, with the upper bound on possible values of response variable equal to 50. This may result lower prediction precious of the linear models.

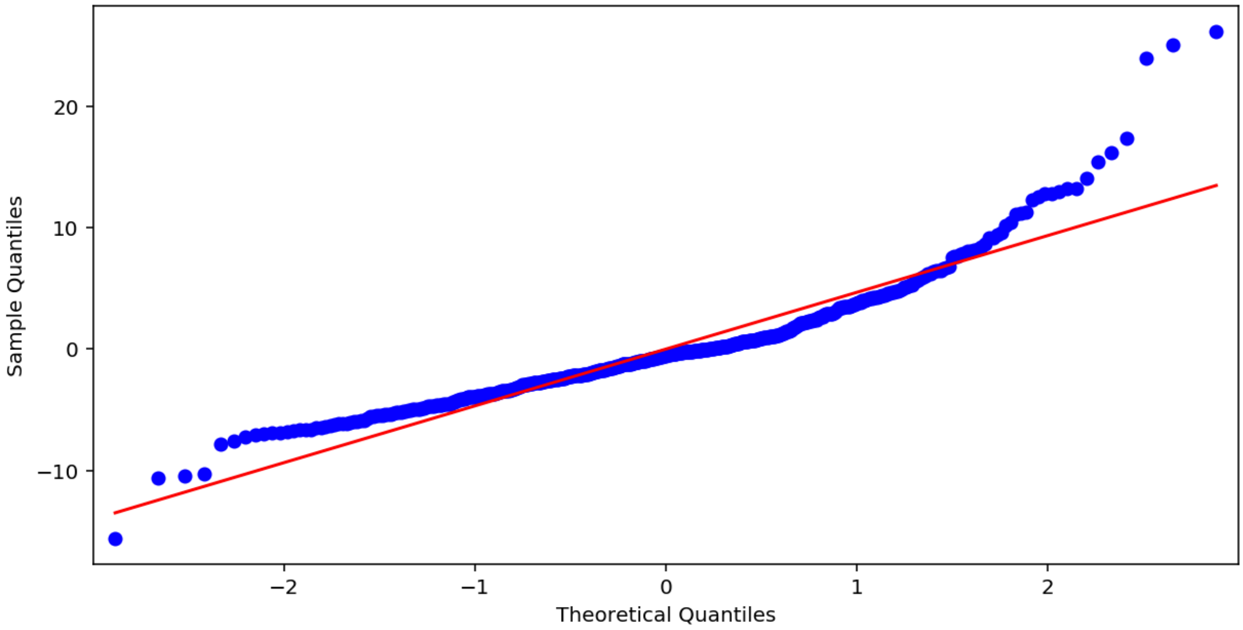


Figure 2: The Q-Q plot of the residual errors

Figure 2 shows that the residual errors are not normally distributed with many outliers to the right. It is consistent with the previous observation that the data might be manually cut.

Table 2: The VIFs for all the explanatory variables

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
vif	1.8	2.3	4.0	1.1	4.4	1.9	3.1	4.0	7.5	9.0	1.8	1.3	2.9

If there was no multicollinearity between variables, the VIF would be 1 for all the variables. Table 2 shows the VIFs for 'TAX' and 'RAD' are 9.01 and 7.48 respectively, indicating they are correlated to other features. Apply ridge or lasso methods may help reduce the influence of the multicollinearity.

Based on above discussion, the assumptions of linear regression are not met for the Boston housing price dataset. The linear analysis cannot be conducted on the fitted linear model. That means the p-values from the t-test, F-test are not valid.In this study, we will not work on the assumptions, and therefore will not interprete the coefficients of the variables and their corresponding statistical significance.

2.3 Train/Test Data Split

The dataset will be divided into two parts:

- Test set: 156 data points
- Training set: the first 350 data points

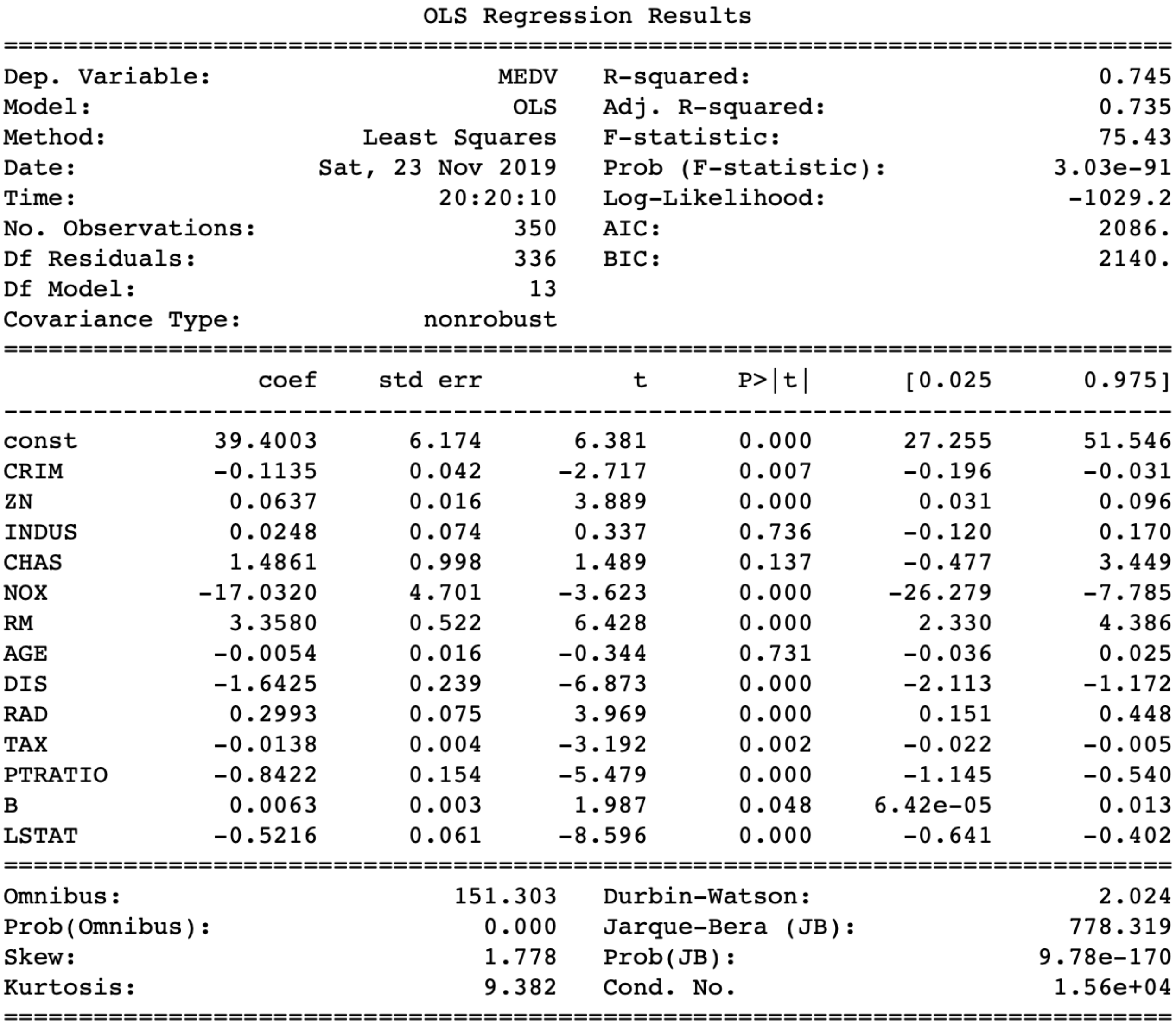
To compare different models, the models are evaluated using test data. Around 30% of the dataset is kept as the test set to give a high confidence in the overall performance of our models.

2.4 Linear Regression Model without Regularization

The linear regression model has the form $Y = \beta_0 + \beta X + \epsilon$. β_0 is the intercept; β is the coefficients vector; X is input matrix. The least squares method (LS) will be used as the baseline estimation. LS derives the β_0 and β by minimizing the residual sum of squares (RSS) of the linear model.

$$RSS = \sum_{k=1}^{n} (y_i - \hat{y}_i)^2$$

Table 3: The summary table for the linear regression without regularization



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.745
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     75.43
Date:                Sat, 23 Nov 2019   Prob (F-statistic):           3.03e-91
Time:                        20:20:10   Log-Likelihood:                -1029.2
No. Observations:                 350   AIC:                             2086.
Df Residuals:                     336   BIC:                             2140.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         39.4003      6.174      6.381      0.000      27.255      51.546
CRIM          -0.1135      0.042     -2.717      0.007      -0.196      -0.031
ZN             0.0637      0.016      3.889      0.000       0.031       0.096
INDUS          0.0248      0.074      0.337      0.736      -0.120       0.170
CHAS           1.4861      0.998      1.489      0.137      -0.477       3.449
NOX          -17.0320      4.701     -3.623      0.000     -26.279      -7.785
RM             3.3580      0.522      6.428      0.000       2.330       4.386
AGE           -0.0054      0.016     -0.344      0.731      -0.036       0.025
DIS           -1.6425      0.239     -6.873      0.000      -2.113      -1.172
RAD            0.2993      0.075      3.969      0.000       0.151       0.448
TAX           -0.0138      0.004     -3.192      0.002      -0.022      -0.005
PTRATIO       -0.8422      0.154     -5.479      0.000      -1.145      -0.540
B              0.0063      0.003      1.987      0.048    6.42e-05       0.013
LSTAT         -0.5216      0.061     -8.596      0.000      -0.641      -0.402
==============================================================================
Omnibus:                      151.303   Durbin-Watson:                   2.024
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              778.319
Skew:                           1.778   Prob(JB):                    9.78e-170
Kurtosis:                       9.382   Cond. No.                     1.56e+04
==============================================================================
```

Table 3 shows the estimated coefficients, their standard errors, and statistical significance (P>|t|) . As illustrated in the part 2.2, statistical significance will not be interpreted because the t-test is not valid when the assumptions for linear regression is not met.

The R-squared statistic is 0.745 for the training dataset, which is higher 0.719 of the test dataset. There is a chance of overfitting. It may be due to that the least squares estimates over generalize the training dataset and have large variance. The following regularization techniques will be used in the linear model and evaluated if they can reduce model complexity and improve the prediction precision

- Best subset serach
- Shrinkage methods
  - Ridge regression
  - Lasso regression

3. Methodology

3.1 Cross Validation Setting

There are only 300 data points in the training dataset. With limited data, cross validation is applied to optimize the parameters. The training dataset is divided into 5 folds. During each round, the training dataset will be further broke down into the two sets:

- Cross validation - Training set: 80% of 350 data points
- Cross validation - Development set: 20% of 350 data points

The hyper parameters (number of features to use in the linear model, regularization term lambda) are optimized using Cross validation datasets. We will select the model with the least average prediction error using Cross validation - Development set.

3.2 Evaluation Metrics

The prediction error will be evaluated using the mean squared error (MSE). MSE measures the average squared difference between the estimated values and the actual value.

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(y_i - \hat{y}_i)^2 = \frac{RSS}{n}$$

4. Cross Validation and regularization

4.1 Least Square Model

The least square model is evaluated using cross validation method. As mentioned above, RSS is the residual sum of squares of the model.

Table 3: The cross validation error and training error of the linear model

Dataset	RSS
Cross Validation Error	8612.92
Training Error	7341.97

As shown in the Table 3, the Sum of RSS for each fold (Cross Validation Error) is 8612.92, which is bigger than 7341.97, the RSS of the linear model trained on the full training dataset (Training Error). It shows that the cross validation models produce higher bias since the linear model trained on N- N/K examples might have performed better if trained on the whole training dataset.

4.2 Best-subset search

The best-subset search allows us to eliminate the unimportant features from the model. It finds for each subset of all the features that gives smallest MSE using LS method.

Based on the cross validation, the average MSE drops at first when the number of feature increases, and increases again when number of features is bigger than 10. It reaches its minimum value 24.28 on the cross validation dataset when ten features ('CRIM', 'ZN', 'NOX', 'RM', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B' and 'LSTAT'). These features are included in the best-subset model. Also, the variance between different sub sets of features for each number of feature scenario generally decreases when the number gets bigger.

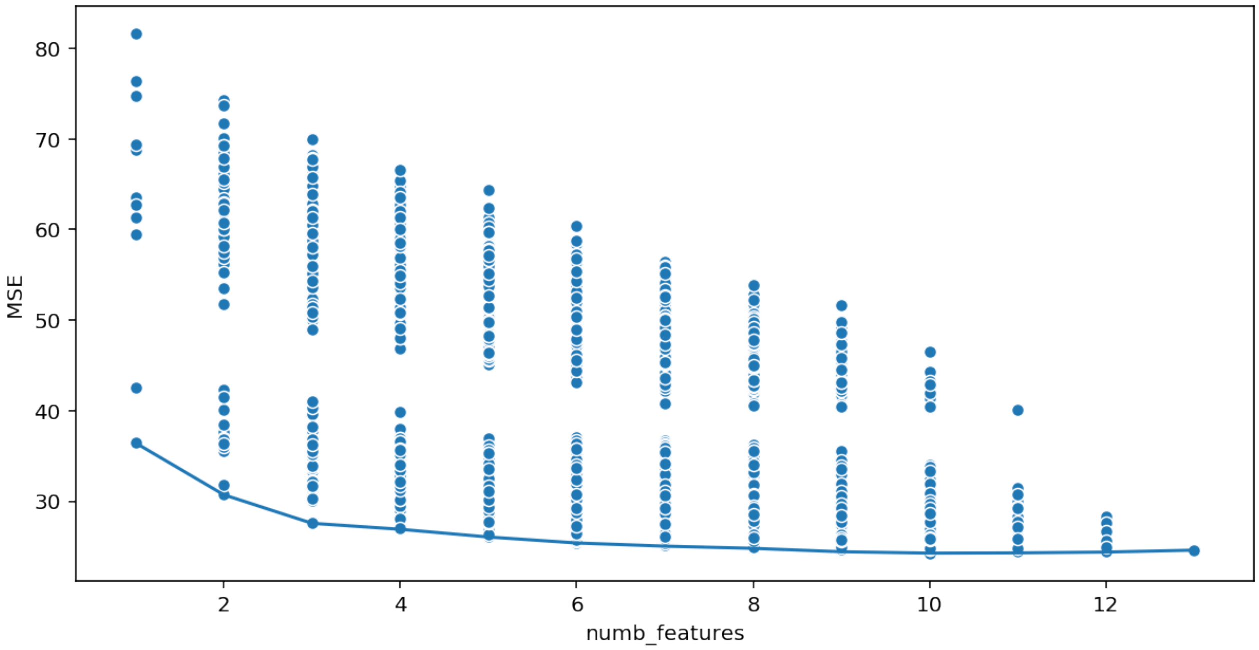


Figure 3: the average MSE of the cross validation dataset for all possible subset models

The variables selected for different numbers of features are shown (marked in yellow cells) in the Table 4. We can see that 'LSTAT' is selected in every model while 'Age' is eliminated in every scenario except the full model. This is consistent with the covariance table we shown before. 'LSTAT' has a higher correlation coefficient (-0.74) than age (-0.38).

Table 4: The variables selected for different numbers of features

# of features	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
1	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	1	0	0	0	0	0	0	1
3	0	0	0	0	0	1	0	0	0	0	1	0	1
4	0	0	0	0	0	1	0	1	0	0	1	0	1
5	0	0	0	0	1	1	0	1	0	0	1	0	1
6	0	1	0	0	1	1	0	1	0	0	1	0	1
7	0	1	0	0	1	1	0	1	0	0	1	1	1
8	1	1	0	0	0	1	0	1	1	1	1	0	1
9	1	1	0	0	1	1	0	1	1	1	1	0	1
10	1	1	0	0	1	1	0	1	1	1	1	1	1
11	1	1	0	1	1	1	0	1	1	1	1	1	1
12	1	1	1	1	1	1	0	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	1

4.3 Ridge Regression

In ridge regression, the cost function is altered by adding a penalty equivalent to the square of the size of the regression coefficients. The λ is the regularization parameter, controlling the weight of the penalty.

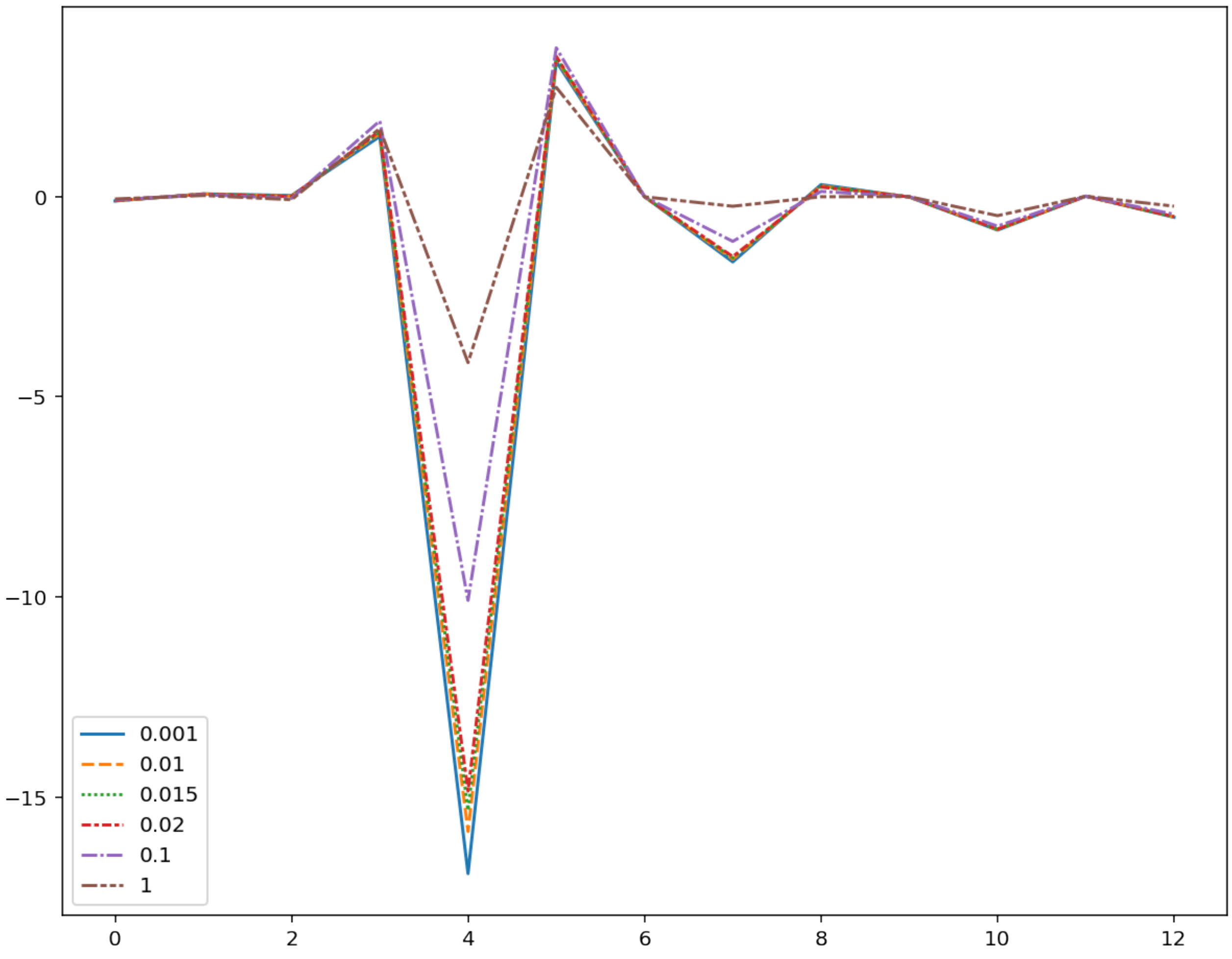


Figure 4: The variable coefficients of Ridge model with different λ

As shown in Figure 4, for λ= 0.0001, coefficients for Ridge regression and linear regression show close resemblance. As λ increases, Ridge model shrinks the coefficients to simply the model.

Table 4: The MSEs for different regularization term lambda

λ	MSE (normalized features)	MSE(unnormalized features)
0.001	24.59	24.61
0.010	24.47	24.59
0.015	24.43	24.58
0.020	24.40	24.58
0.100	24.50	24.48
1.000	31.01	24.30

Table 4 shows that Ridge regression works better after feature normalization when the regularization term (λ) is smaller than 1. This is due to the huge difference in feature variances as mentioned in part 2.1. Also, when the features are normalized, MSE varies from 24.40 to 31.01. When the features are not normalized, MSE only varies from 24.30 to 24.61. It shows the ridge regression does not adjust our model effectively to actually influence the evaluation metric. Normalized features are used in the following Ridge models.

MSE reach its minimum when λ equals 0.02. When λ is small than 0.02, Ridge does may smooth the output of the function enough to reduce overfitting. When λ is bigger than 0.02, Ridge may smooth out the function too much and cause under-fitting.

4.4 Lasso Regression

The cost function for least absolute shrinkage and selection operator (Lasso) regression is similar to that of Ridge regression. The only difference is that absolute magnitudes are taken into consideration instead of the square of the coefficients. The cost function is altered by adding a penalty equivalent to the absolute size of the coefficients. The λ is the regularization parameter, controlling the weight of the penalty.

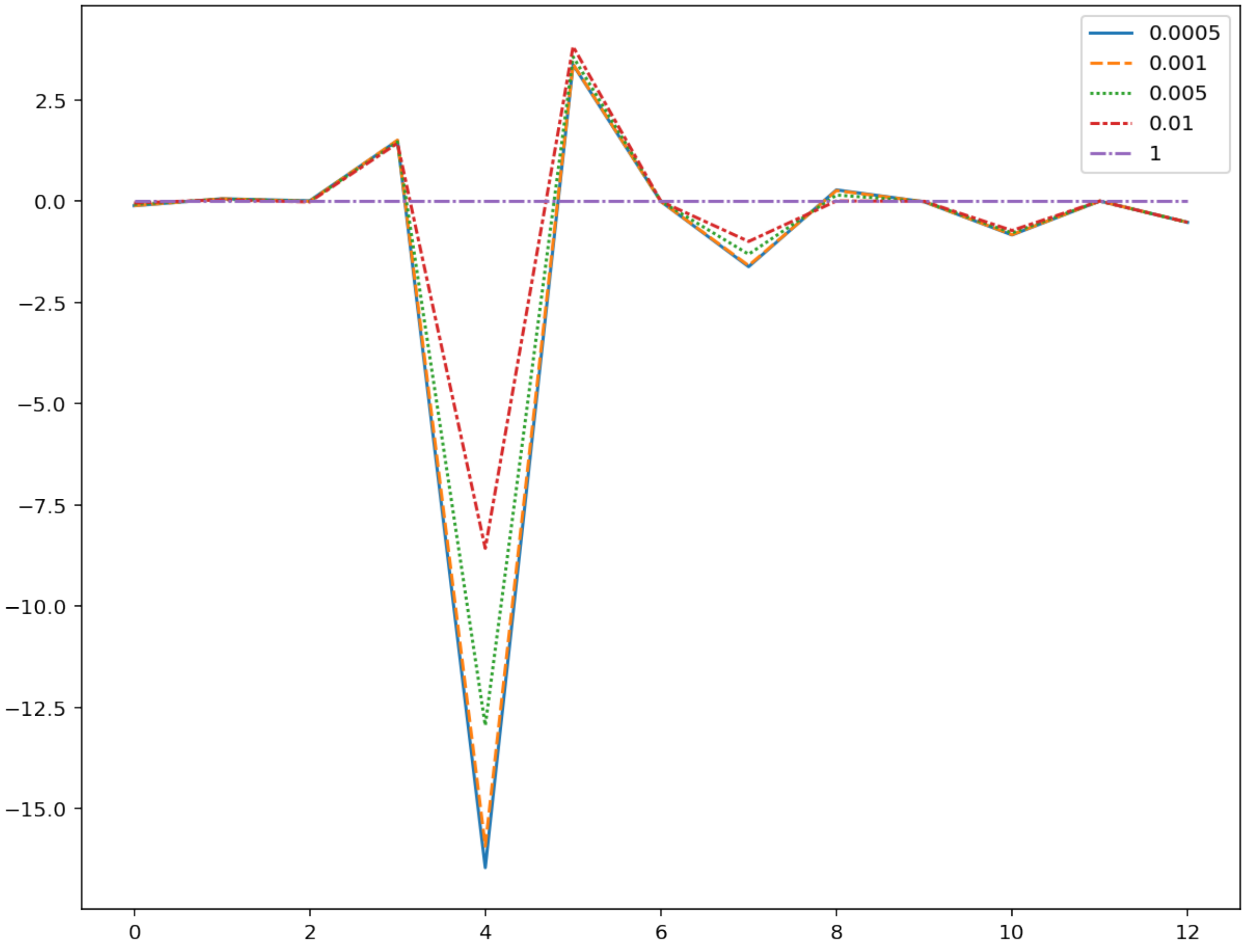


Figure 5: The variable coefficients of Lasso model with different λ

Just like Ridge regression cost function, as shown in Figure 5, for λ= 0.0005, coefficients for Lasso regression and linear regression show close resemblance. As λ increases, Ridge model shrinks the coefficients to simply the model. When λ is very big (equals to 1 in this case), Ridge model shrined every coefficient to 0.

It shows that, unlike Ridge, Lasso can lead to zero coefficients. In such cases, some features are completely discarded from the model.

Table 5: The MSEs for different regularization term lambda

λ	MSE (normalized features)	MSE(unnormalized features)
0.0005	24.58	24.60
0.0010	24.55	24.60
0.0050	24.73	24.57
0.0100	25.56	24.56
1.0000	83.01	28.03

Similar to ridge regression, lasso regression works better after feature normalization when the regularization term (λ) is smaller than 0.005.

In the study, on the cross validation datasets, MSE for ridge regression varies from 24.55 to 83.01 when the features are normalized. However, When the features are not normalized, MSE only varies from 24.57 to 28.03. With regularization not too big(<0.01 in this case), Lasso regression does not adjust our model effectively to actually influence the evaluation metric when the features are not normalized. Normalized features are used in the following Lasso models.

5. Model Comparison

5.1 Accuracy between Datasets

The MSE on the training dataset, Test dataset and Cross Validation dataset are shown in Table 6.

Table 6: The MSEs for different models and datasets

Model	MSE (Train)	MSE (Test)	MES (Cross Validation)
least squares	20.98	25.15	24.61
best-subset	21.14	26.00	24.28
ridge	21.06	25.14	24.40
lasso	21.00	25.16	24.55

Based on Table 6, MSEs of cross validation datasets and test dataset are lower than those of training datasets for all models. This is because the models are trained on the training dataset. After learning the patterns of the training dataset, the models will generally perform worse on the new dataset.

It is surprising to see that the MSEs for test dataset are lower than those of cross validation datasets for all models. The MSE for test dataset is based on the models trained on the whole training dataset. The MSE for cross validation dataset is based on the models trained on the subsets (N-N/K). Ideally, the model trained on the bigger dataset will have better generalization ability. But the parameters (lambda, number of features) are chosen based on the training dataset. They may not perform as well as on the test dataset.

Another reason might be that the outliers in the train/test data split process. As mentioned in introduction part, the response variable might be manually cut from 50. The response variable is equal to 50 for 2.9% of the training data, and 3.8% of the test dataset. This may cause higher prediction error in the test dataset than validation datasets.

5.2 Accuracy between Models

Table 7 shows the different between linear regression models with without regularization on different datasets.

Table 7: The MSE differences for different models and datasets

Model	MSE (Train)	MSE (Test)	MES (Cross Validation)
Best-subset compared to LS	0.16	0.85	-0.33
Ridge compared to LS	0.08	-0.01	-0.21
Lasso compared to LS	0.02	0.01	-0.06

Based on Table 7, the models with regularization perform worse on the training dataset since the models are forced to be simpler and cannot perform as well as the full model. However, the simpler models all perform better on the cross validation dataset. The added regularization term avoids minimized RSS too carefully, which reduces the variance and the overfitting problem.

The best-subset dataset performs best on the cross-validation datasets but worst on the test dataset. By retaining only 10 non-zero features and discarding the rest three, best-subset model might have lower prediction error than the full model if it does reduce the overfitting problem. However, features are either retained or discarded in the best-subset model. It will also has high variance and increase the MSE on the test dataset compared to the full model. Reduce this under-fitting by reducing alpha and increasing number of iterations.

Shrinkage methods (Ridge and Lasso) are more continuous than the best-subset model. Therefore, it does not suffer as much from high variability. We can see that only Ridge and Lasso models both perform better than the best-subset model.

Surprisingly, none of the simpler models performs much better n the test data sets than the full model. Only Ridge model performs slightly better than the full model for the test dataset. It decreases MSE slightly by 0.01.

5.3 Coefficient of Parameters

Table 7 shows the the variable coefficients of different models. As mentioned in part 2.1, the variance differs significantly for different variables. To make coefficient more comparable, the variable coefficients are multiplied by the corresponding standard deviation, which are shown in Table 8 and Figure 6.

Table 7: The variable coefficients of different models

Variable	Least Square	Best-subset	Ridge	Lasso
CRIM	-0.1135	-0.1182	-0.1034	-0.1069
ZN	0.0637	0.0645	0.0570	0.0605
INDUS	0.0248	0.0000	-0.0048	0.0000
CHAS	1.4861	0.0000	1.6363	1.5089
NOX	-17.0320	-16.6651	-14.8435	-15.9263
RM	3.3580	3.3305	3.4987	3.3864
AGE	-0.0054	0.0000	-0.0062	-0.0041
DIS	-1.6425	-1.6609	-1.5009	-1.5882
RAD	0.2993	0.3038	0.2348	0.2639
TAX	-0.0138	-0.0138	-0.0107	-0.0120
PTRATIO	-0.8422	-0.8446	-0.8136	-0.8245
B	0.0063	0.0064	0.0064	0.0061
LSTAT	-0.5216	-0.5283	-0.5022	-0.5207

Table 8: The variable coefficients times the corresponding standard deviation

Variable	Least Square	Best-subset	Ridge	Lasso
CRIM	-1.0306	-1.0731	-0.9384	-0.9701
ZN	1.4846	1.5037	1.3279	1.4101
INDUS	0.1658	0.0000	-0.0320	0.0000
CHAS	0.3652	0.0000	0.4021	0.3708
NOX	-2.0183	-1.9748	-1.7590	-1.8873
RM	2.3828	2.3632	2.4826	2.4029
AGE	-0.1523	0.0000	-0.1759	-0.1149
DIS	-3.5185	-3.5580	-3.2151	-3.4022
RAD	2.5750	2.6137	2.0202	2.2704
TAX	-2.3010	-2.3051	-1.7894	-2.0074
PTRATIO	-1.7618	-1.7667	-1.7020	-1.7246
B	0.5604	0.5668	0.5695	0.5451
LSTAT	-3.4521	-3.4967	-3.3238	-3.4465

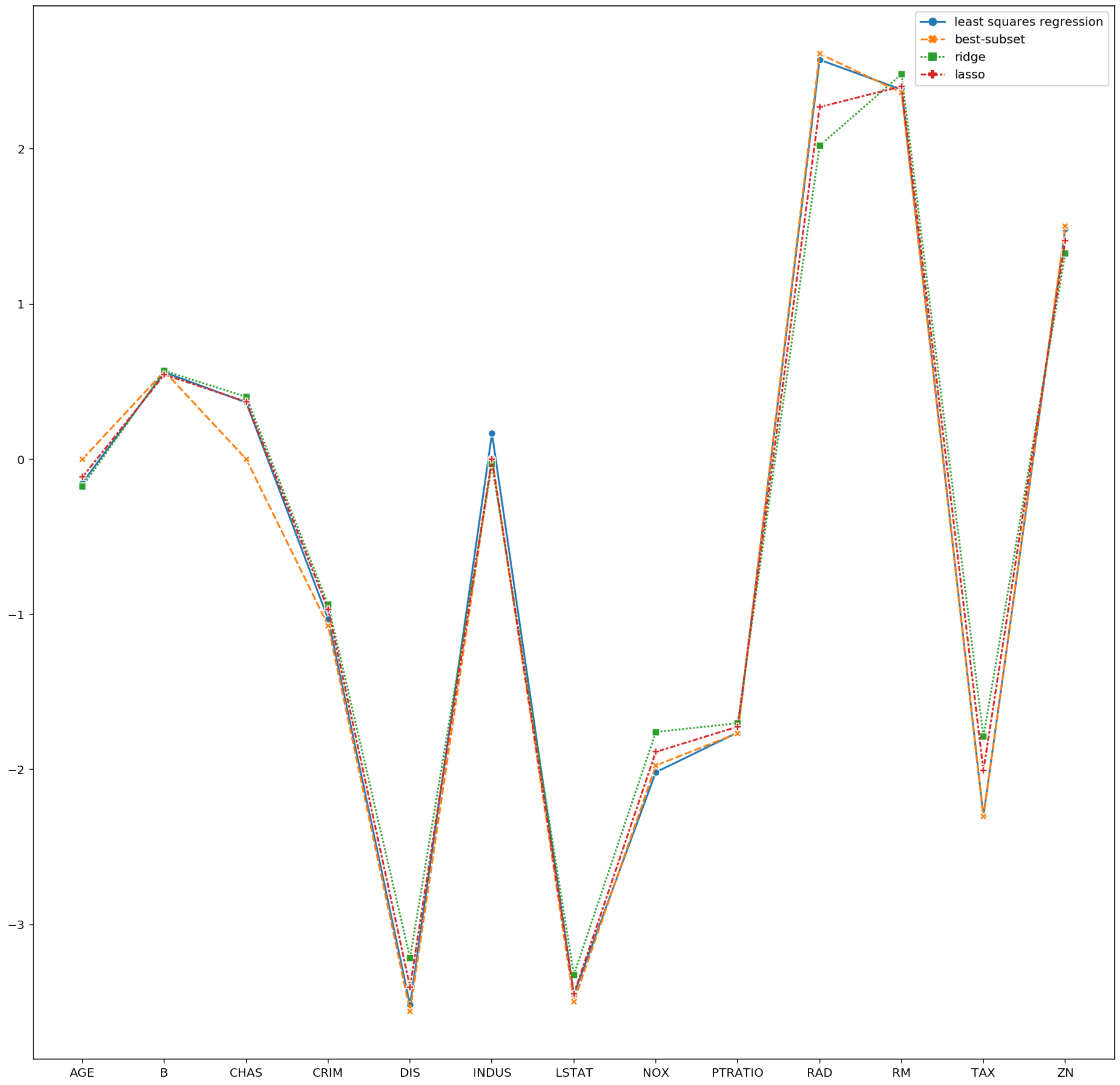


Figure 6: The variable coefficients times the corresponding standard deviation

The tables and the figure show both Ridge and Lasso decreases the magnitude of the coefficients to reduce overfitting. Lasso eliminates variable 'INDUS' in the model. Compared to Ridge, Lasso regression can also select features. Ridge regression, however, only reduces the variable coefficients close to 0, but not 0.

Both Ridge and Lasso methods help reduce the multi-collinearity, the magnitude of 'RAD' and 'TAX' are significantly reduced. This is consistent with the observation in the introduction part that their VIFs are higher than other variables.

5.4 The ability to predict housing prices

Table 9 shows the R-squared (R^2) statistic for different models on the test dataset. It represents the proportion of the variance that can be explained by variables in the regression models.

Table 9: the R-squared (R^2) statistic

Model	R2 (Test)
Least squares regression	0.719
Best-subset	0.709
Ridge	0.719
Lasso	0.718

We may use the model but with following concerns:

- Only around 70% of the variance in the response variable can be explained by the models.
- The feature 'B' may incorporate racial prejudice. We cannot rely on such data to increase the existing discrimination.
- We may need a piecewise model to reflect the upper limit of the response variable.

6. Conclusion

We have perform four linear models on the training, validation and test datasets. Generally, the regularization helps reduce the overfitting; and the simpler models perform better on the validation datasets compared to the full model. The Ridge performs better in our case than Best-subset and Lasso because the following two methods eliminate certain variables in the models.

However, the models do not perform that well on the test dataset.

- One reason is that the response variable might be manually cut, with the upper bound on possible values of response variable equal to 50. 2.9% 3.8%
- Also, there are only 13 features in the dataset but 506 data points. Overfitting might not be a significant issue in our case. If there is not a overfitting problem, the regularization cannot help reduce the prediction error of the model.
- The dataset size is limited.

Another concern is the incorporation of the variable with racial discrimination. 'Trash in, trash out.' We cannot blindly trust in our data. Otherwise, the models we build for social justice, hiring process and so on may augment the human error.