

part1

Cui Qingxuan

2024-12-01

Contents

1 Assignment 1	1
1.1 Quality of fit, prediction and model with linear regression.	1
1.2 Report the cost function	1
1.3 interpret LASSO Coefficient Paths Plot	1
1.4 model with only three features	2
1.5 fit Ridge regression and compare the plots from steps 3 and 4	3
1.6 Task 5	4
2 Assignment 3	5
2.1 Task 1	5
2.2 Task 2	5
2.3 Task 3	7
2.4 Task 4	7

1 Assignment 1

1.1 Quality of fit, prediction and model with linear regression.

The MSE of training data is: 0.005709117

The MSE of test data is: 722.4294

The model performs well on the training data (MSE = 0.0057), but has a large error on the test data (MSE = 722.4294), indicating that the model cannot be generalized and there is a serious overfitting phenomenon.

1.2 Report the cost function

The cost function for **LASSO regression** is defined as:

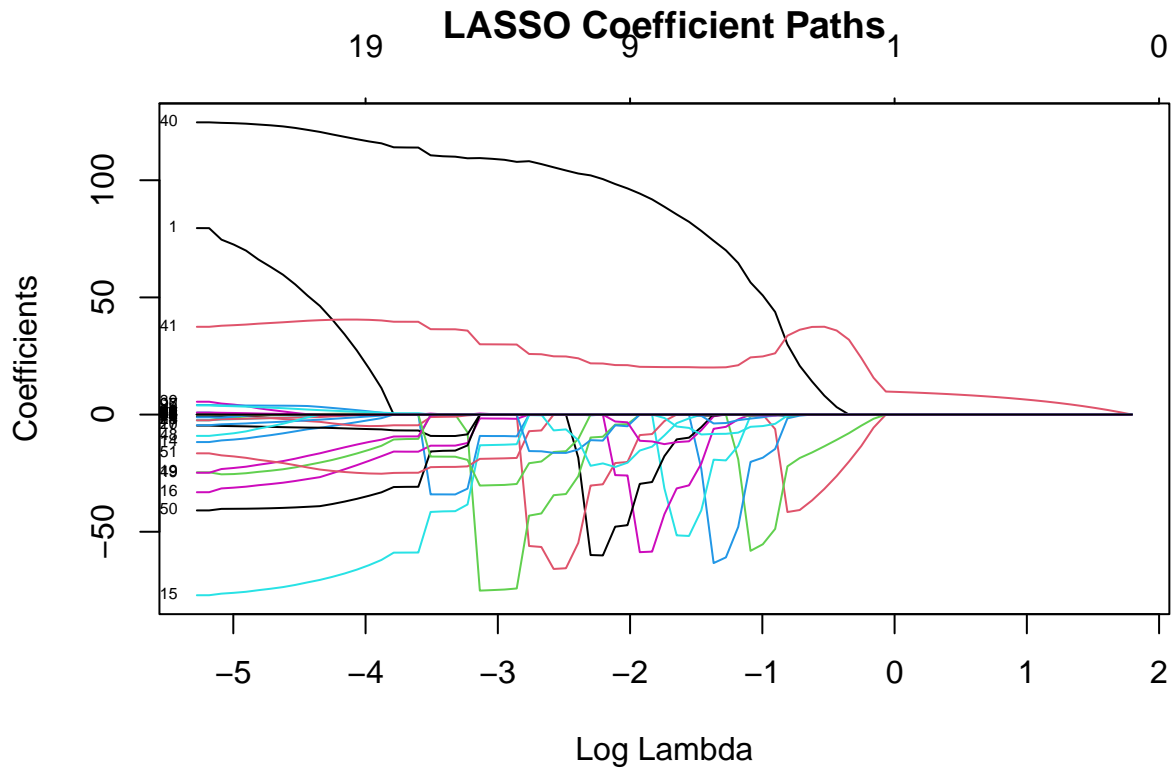
$$J(\beta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \mathbf{X}_i \cdot \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Where lambda controls the regularization strength.

1.3 interpret LASSO Coefficient Paths Plot

##

Coefficients for different lambda values:



1. This graph shows the path that the regression coefficient of each feature in the **LASSO regression model** changes with the regularization parameter $\log(\lambda)$: the horizontal axis is $\log(\lambda)$, and the vertical axis is the value of the regression coefficient for each feature.
2. With the increase of λ (from right to left), the regularization force is enhanced, and the coefficients of most features are compressed to 0, leaving only a few important features with non-zero coefficients, which realizes feature selection.

1.4 model with only three features

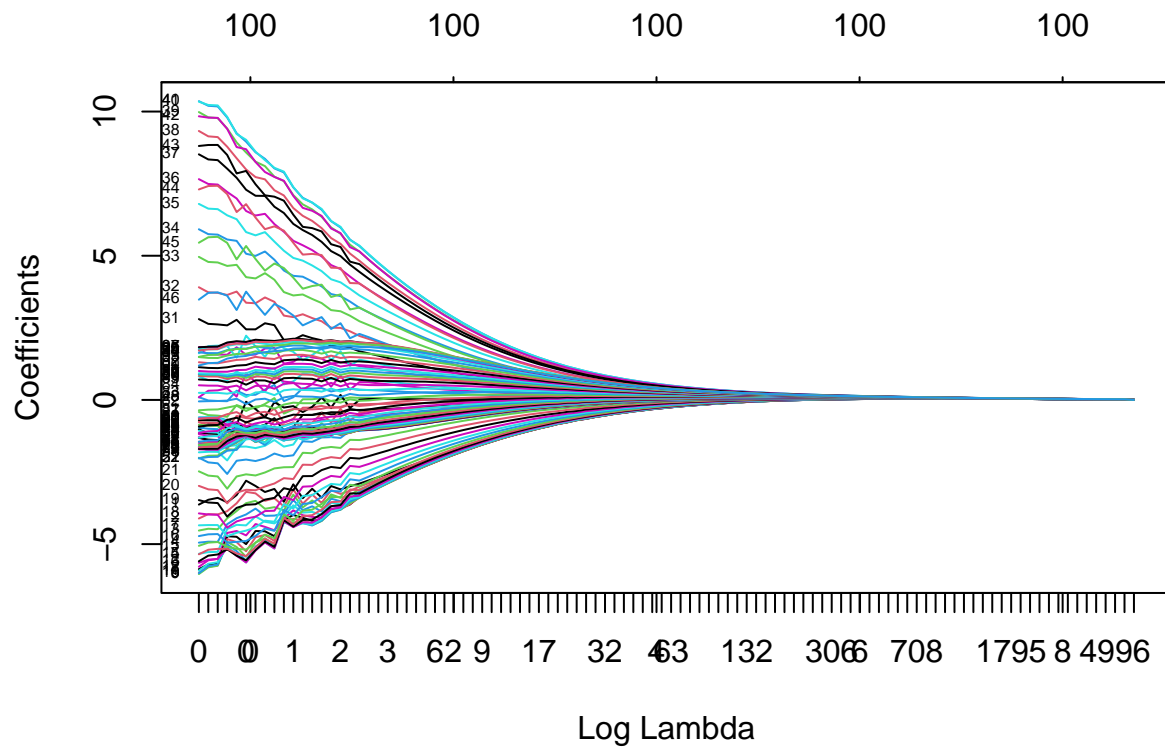
```
## Lambda with 3 non-zero coefficients: 0.8530452
```

```
## The remaining features are:
```

```
## Channel6 Channel7 Channel41
```

```
## -5.580149 -1.793787 15.696612
```

1.5 fit Ridge regression and compare the plots from steps 3 and 4



LASSO coefficient path diagram:

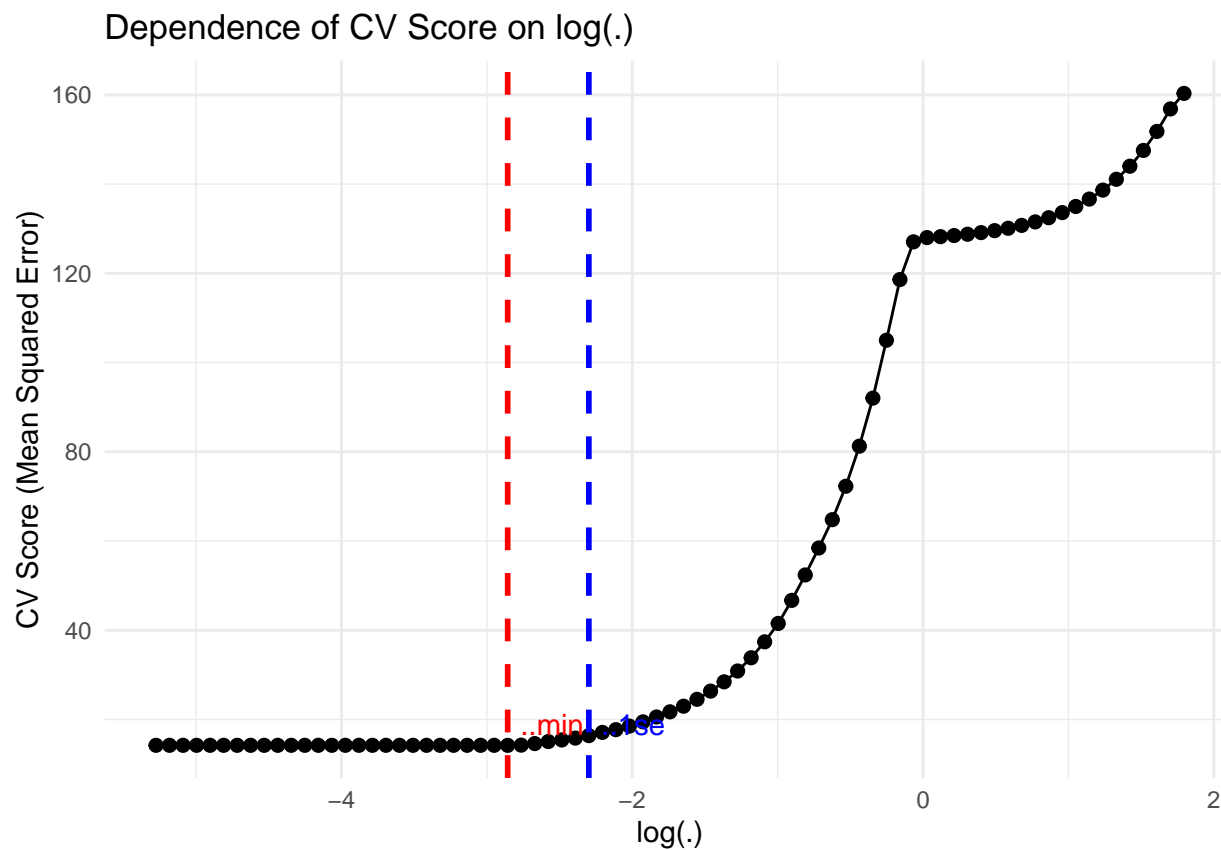
As λ increases, the coefficients of many features are compressed to 0, and the model becomes sparse.

Ridge coefficient path map:

As λ increases, the coefficients of all features gradually shrink, but they do not become zero.

1.6 Task 5

1.6.1 dependence of the CV score on $\log(\lambda)$



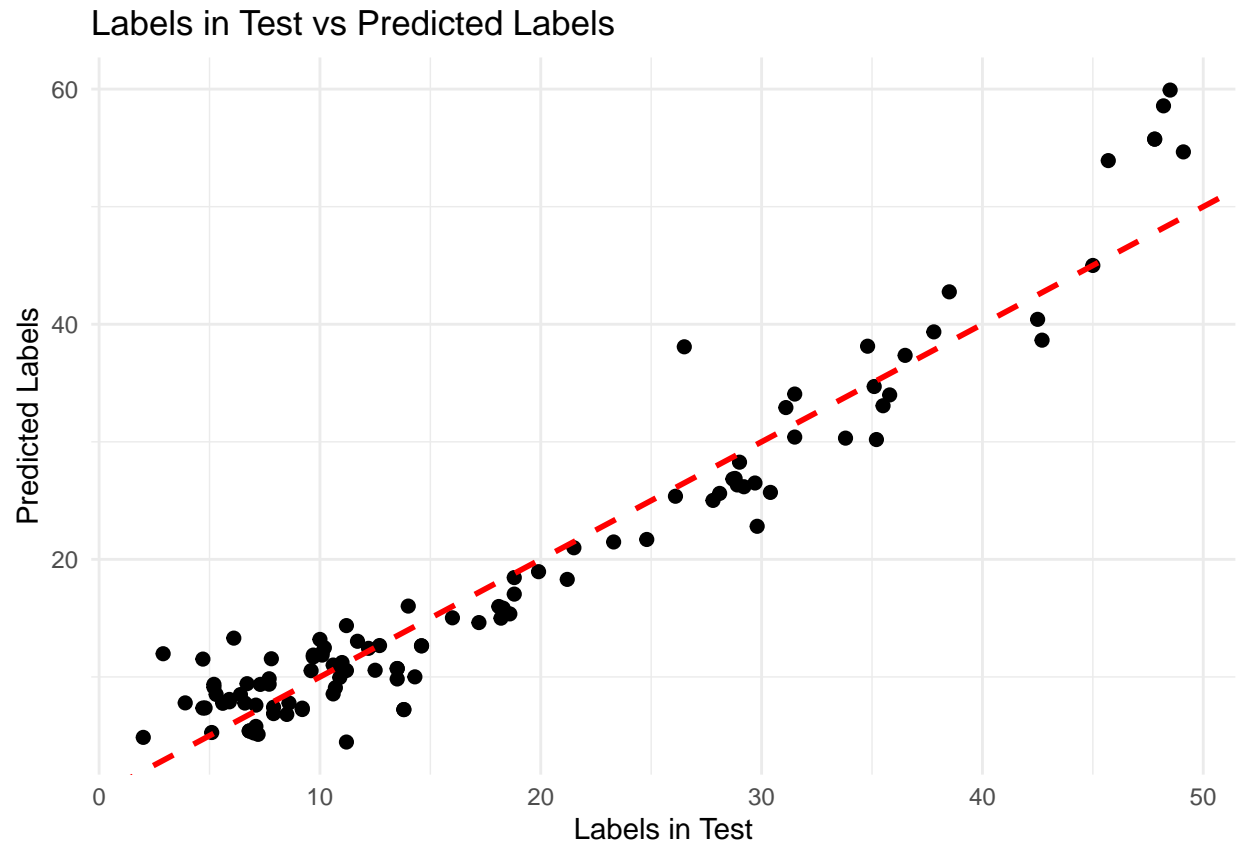
(comment)

1.6.2 Optimal λ

```
## the optimal lambda: 0.05744535
## the number of variables: 8
## log_lambda = -4, mse = 13.48375
## lambda = 0.05744535 mse = 13.67339
```

The difference in MSE values is very small, approximately 0.19, which is unlikely to be statistically significant.

1.6.3 Scatter plot of the original test versus predicted test values



The model performs well overall, as most predictions are close to the actual values.

The scatterplot shows that the model captures the general trend of the data, with no significant systematic errors.

But a few outliers at higher values suggest the model may struggle with extreme cases or higher variability in predictions for larger labels.

2 Assignment 3

2.1 Task 1

```
## At least 2182 components to obtain a 95% of variance in the data.  
## The proportion of variance of explained by each of the two principal components:  
## The first components: 0.2501699  
## The second components: 0.1693597
```

2.2 Task 2

2.2.1 Trace plot of the first principle component.

2.2.2 Top 5 contributing features

```
## Top 5 features that contributed mostly to the first principle component: medFamInc medIncome PctKid
```

medFamInc:median family income (differs from household income for non-family households)

medIncome:median household income

PctKids2Par: percentage of kids in family housing with two parents

pctWInvInc:percentage of households with investment / rent income in 1989

PctPopUnderPov:percentage of people under the poverty level

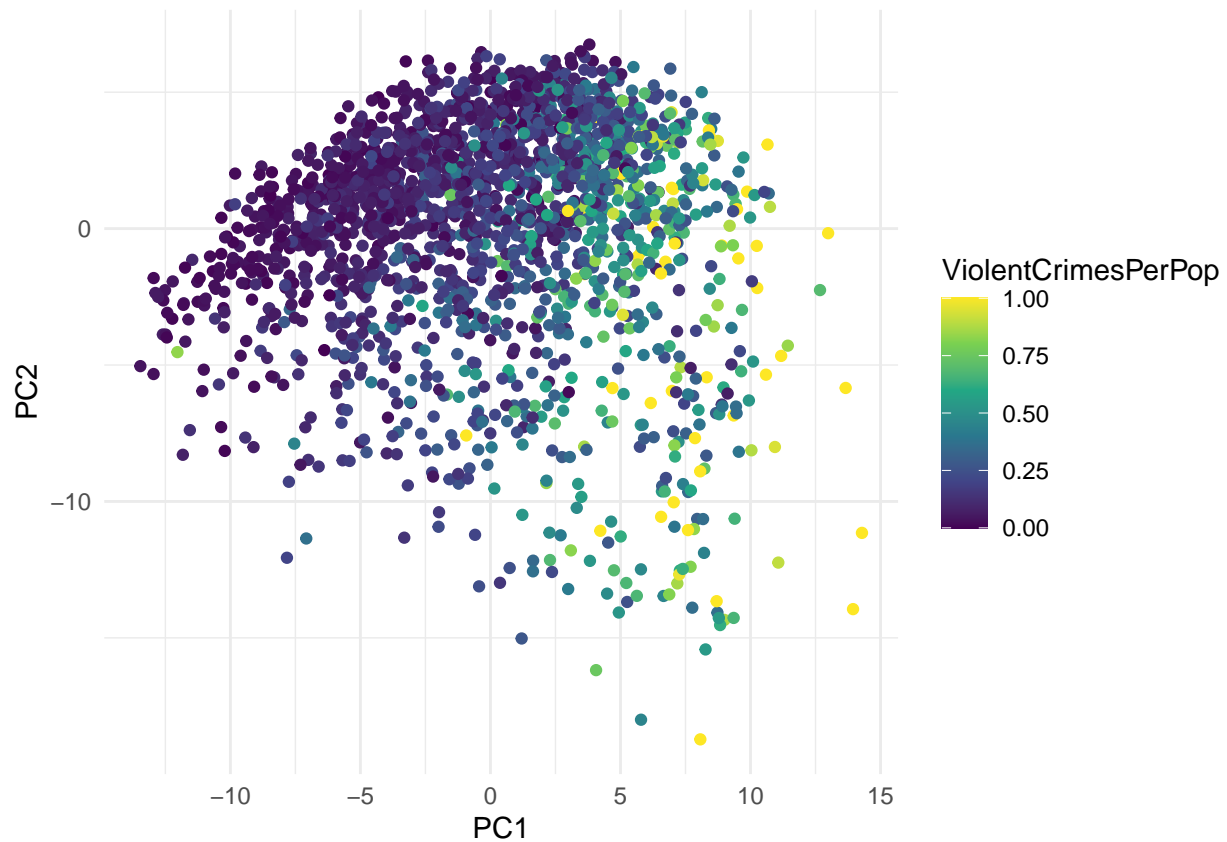
Conclusion:

These five features mainly describes the following aspects:

1.Median family income; 2.Median household income; 3.Percentage of kids in family housing with two parents; 4.Percentage of households with investment / rent income; 5.Percentage of people under the poverty level.

All five features are related to family property and household wealth levels. The logical relationship to crime levels is that families with fewer financial resources or more children (which increases financial budget) may be more likely to resort to illegal activities to obtain money. Financial stress within household may be a significant factor contributing to crime level.

2.2.3 plot of the PC scores in the coordinates (PC1, PC2)



1. The score of the data point on the first principal component (PC1) is significantly positively correlated with the violent crime rate ("ViolentCrimesPerPop"), which tends to increase when PC1 increases (yellow).
2. The second principal component (PC2) has a weak ability to distinguish the violent crime rate, and the distribution of data points in this direction is relatively symmetrical, and there is no obvious pattern.

3. The overall arc distribution of the data indicates that there may be a nonlinear relationship, and further analysis of the characteristic contribution of PC1 is needed to understand the key factors affecting the violent crime rate.

2.3 Task 3

After training the linear regression model.

```
## MSE for the training data is : 0.2752071
```

```
## MSE for the test data is : 0.4248011
```

```
## The R^2 value for the model is: 0.7245166
```

1. The MSE of the model on the training data is 0.275, and the MSE on the test data is 0.425. The test error is slightly higher but generally close, indicating that the model is stable and the degree of overfitting is low.
2. The coefficient of determination R^2 is 0.725, indicating that the model can explain 72.5% of the variance of violent crime rate, but 27.5% of the variation is still unexplained.
3. A slight increase in the test error may indicate the need to optimize features or introduce regularization methods to further improve the generalization ability of the model.

2.4 Task 4