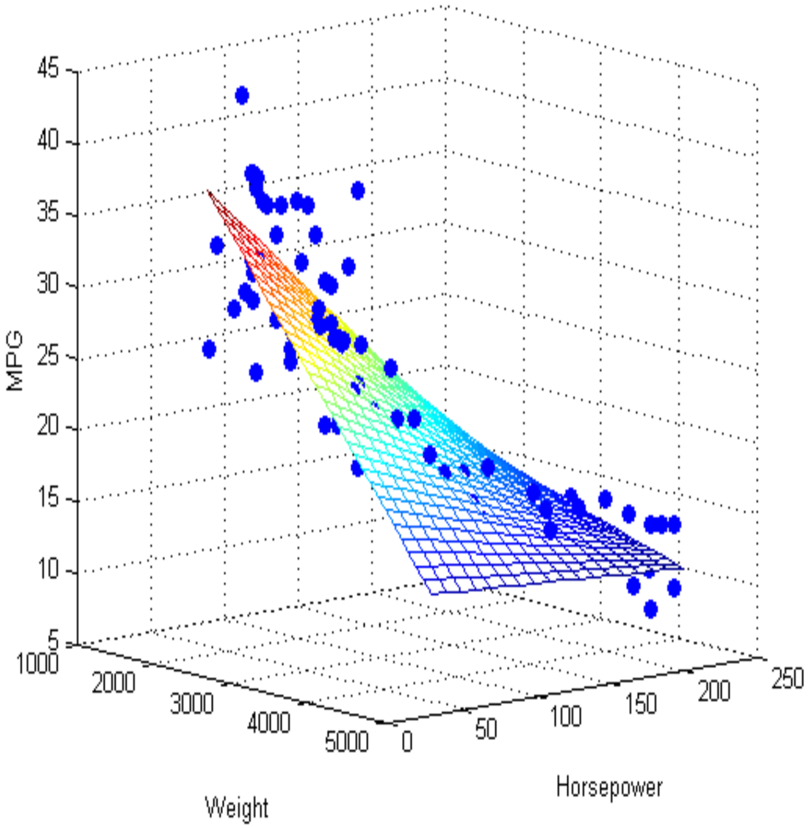
Ridge Regression Analysis on Correlates of Cars' Miles Per Gallon

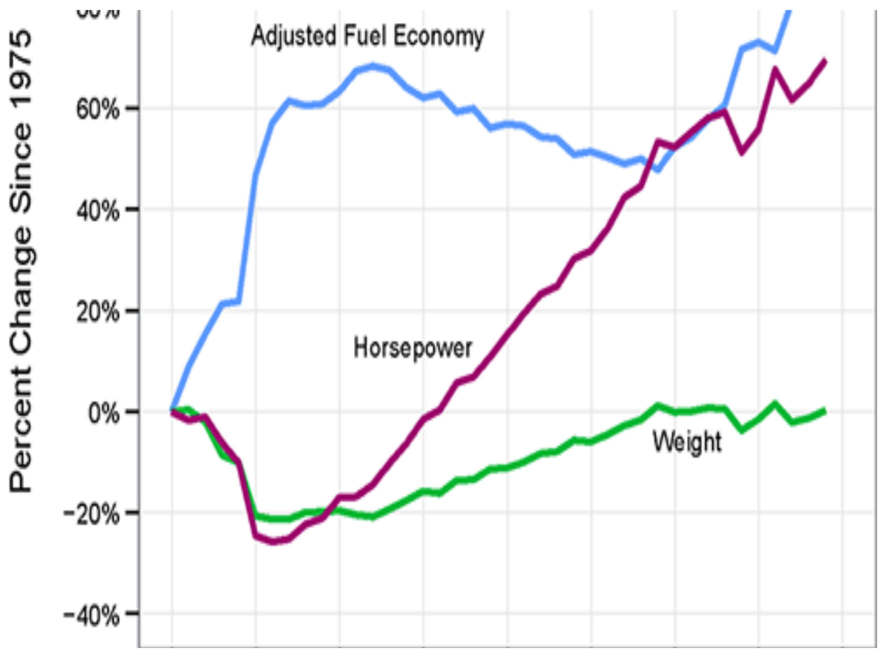


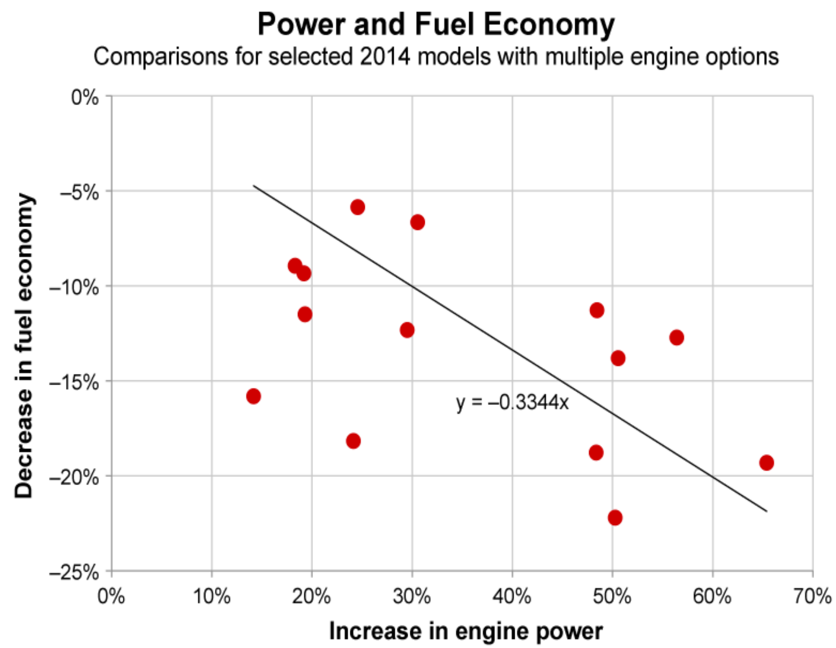
Jiannan(Jeffrey) Zhang
University of Texas at Austin

Introduction

- Concerns about the fuel supply and fuel efficiency have been growing a lot in recent years
- My study focuses on using ridge regression to analyze the correlations between car's MPG and seven predictor variables

Introduction

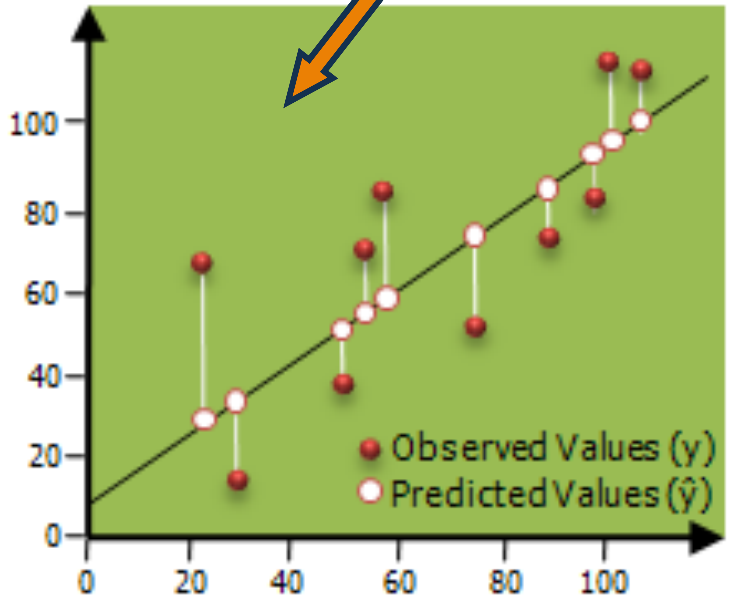




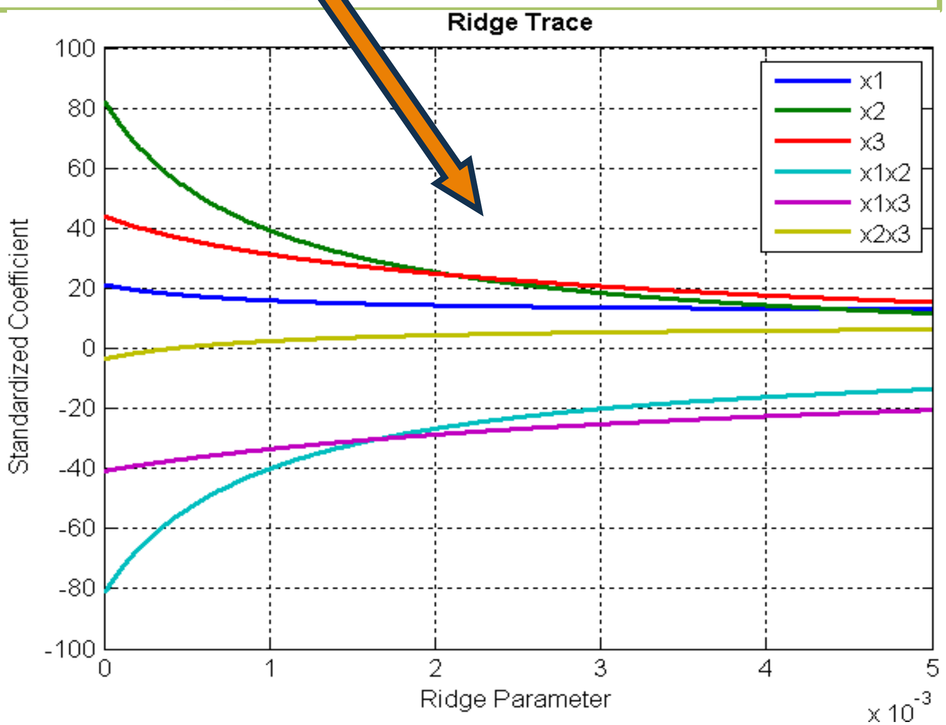
There are many factors affecting MPG such as horsepower (HP), weight, and year of the car.

Introduction

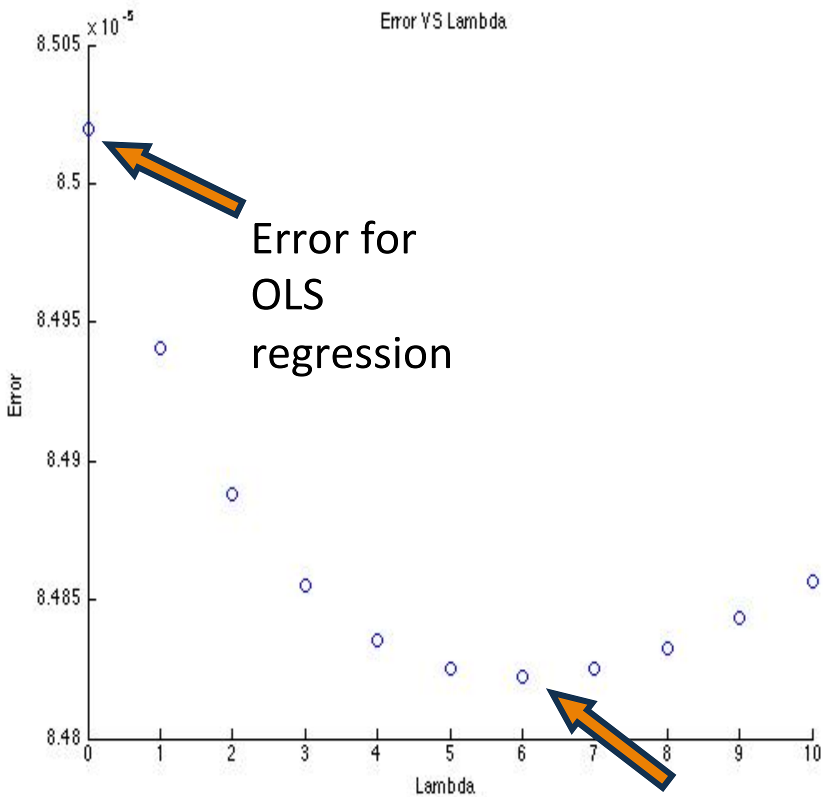
The general way to do the regression is to use Ordinary Least-Squares (OLS) Regression.



Why ridge regression ?

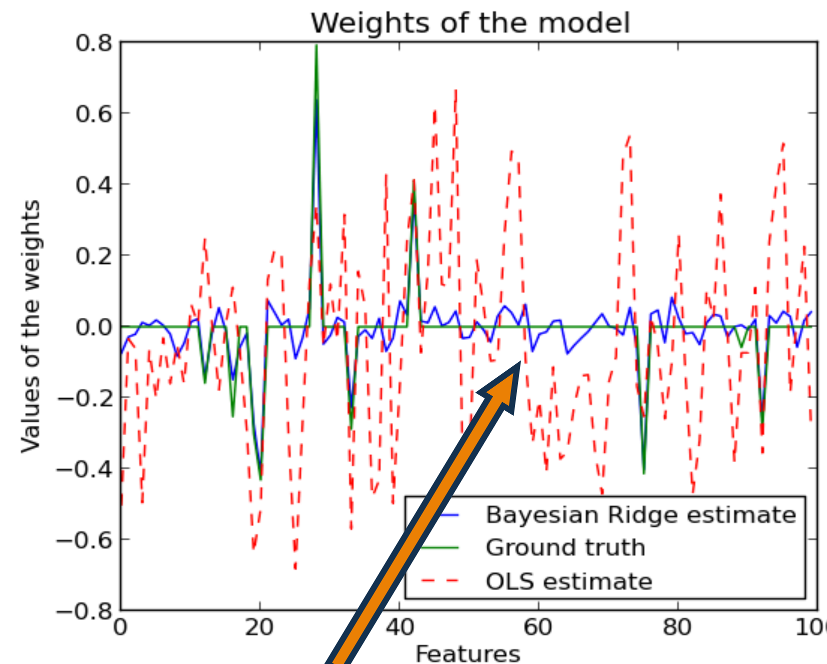


Main methods: 1. Ridge regression



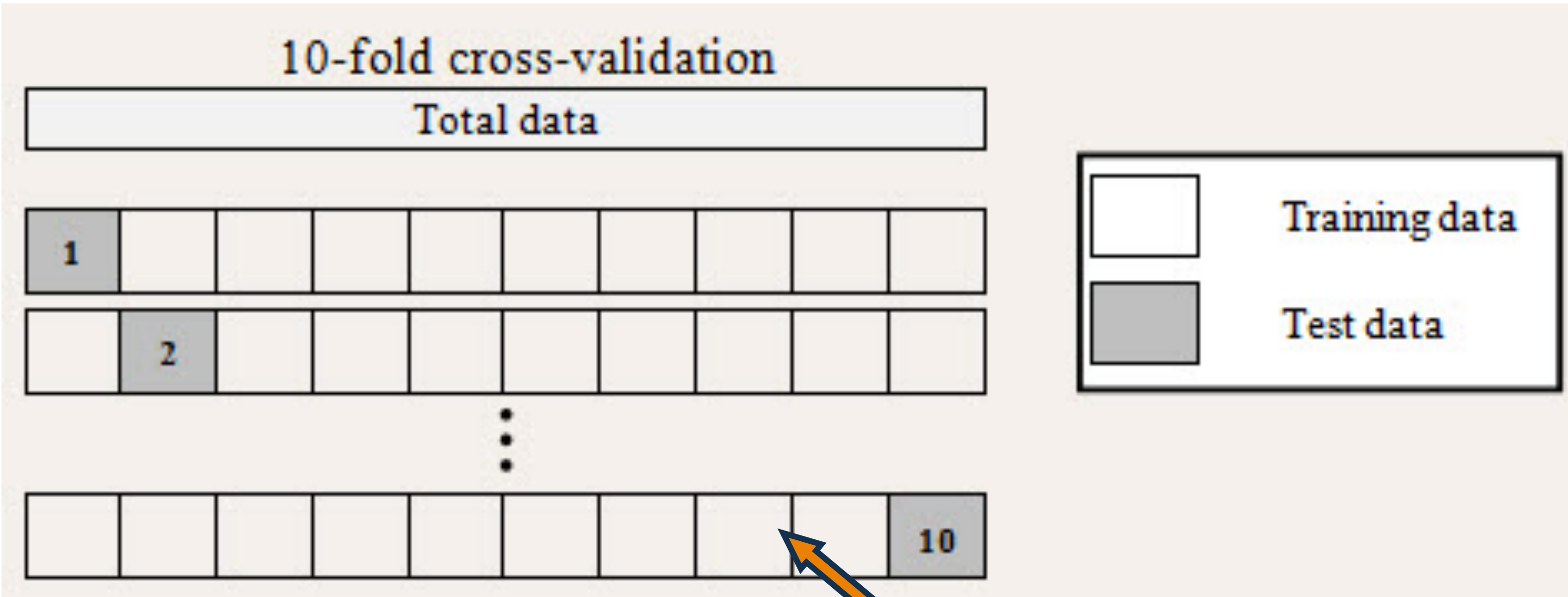
Error for OLS regression

Error for ridge regression is smaller!



The size of weights is controlled. Therefore model is more stable. In ridge regression, my goal is find the **best ridge parameter (Lambda)** that minimizes the ERROR

Main methods: 2. Cross-validation



- I used ten-fold cross-validation
- Every time I use nine folds to train data and one fold to test data

Used to limit problems like overfitting

Results

- Model I (data set with six features : number of cylinders, engine displacement, HP, weight, acceleration time, origin country)

Table 1. Sample weights for model I (K = Intercept, Num = Number)

Variables	K	Num of Cylinders	Engine displacem	Horsepo wer	Weight	Accelera tion time	Origin
Weights (best lambda)	0.0873	-0.1021	0.0474	-0.2682	-0.5452	-0.0365	0.1033
Weights	0.0842	-0.0576	-0.0204	-0.2365	-0.4796	-0.0675	0.1021
Weights	0.0843	-0.0565	-0.0160	-0.2383	-0.4864	-0.0676	0.1025

Results

- Model II (data set with all continuous variables : engine displacement, HP, weight, acceleration time)

Table 2. Sample weights for model II (K = Intercept)

Variables	K	Engine displacement	Horsepower	Weight	Acceleration time
Weights (best lambda)	0.0919	-0.0965	-0.2072	-0.5733	-0.0369
Weights	0.0923	-0.1243	-0.2068	-0.5246	-0.0657
Weights	0.0936	-0.0841	-0.2312	-0.5807	-0.0408

Results

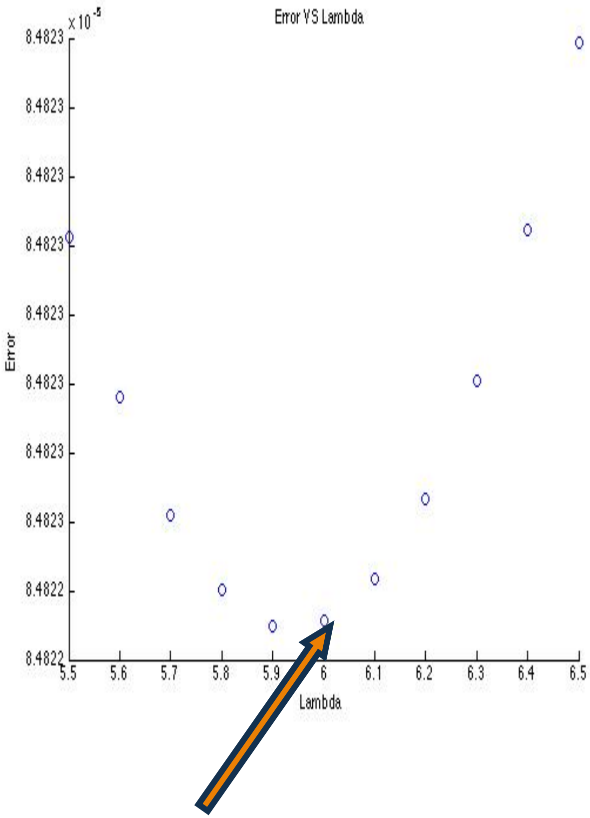
- Model III (data set with all discrete variables : number of cylinders, origin country, years)

Table 3. Sample weights for model III (K = intercept)

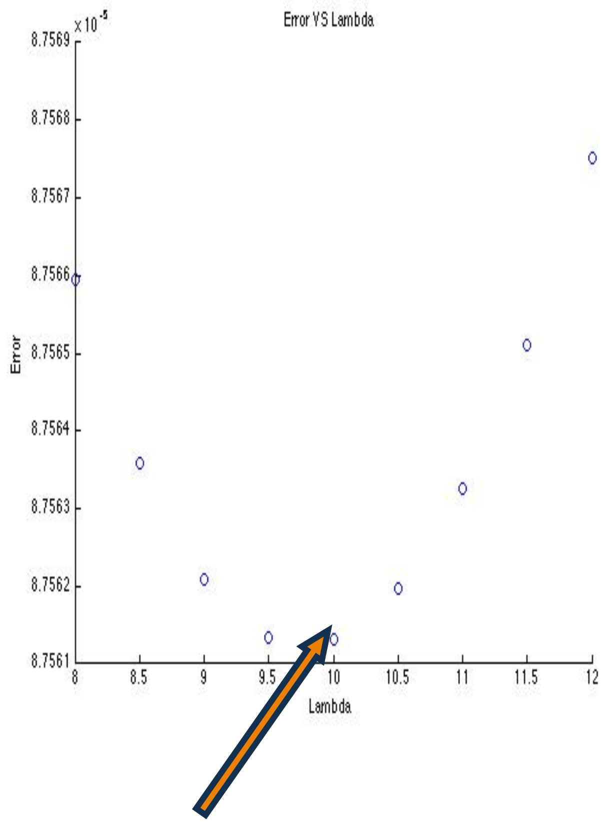
Variables	K	Number of Cylinders	Origin	Years
Weights (best lambda)	-0.0468	-0.5582	2.2883	0.1355
Weights	-0.0342	-0.5868	2.0849	0.1027
Weights	-0.0460	-0.5780	2.2930	0.1314

Results

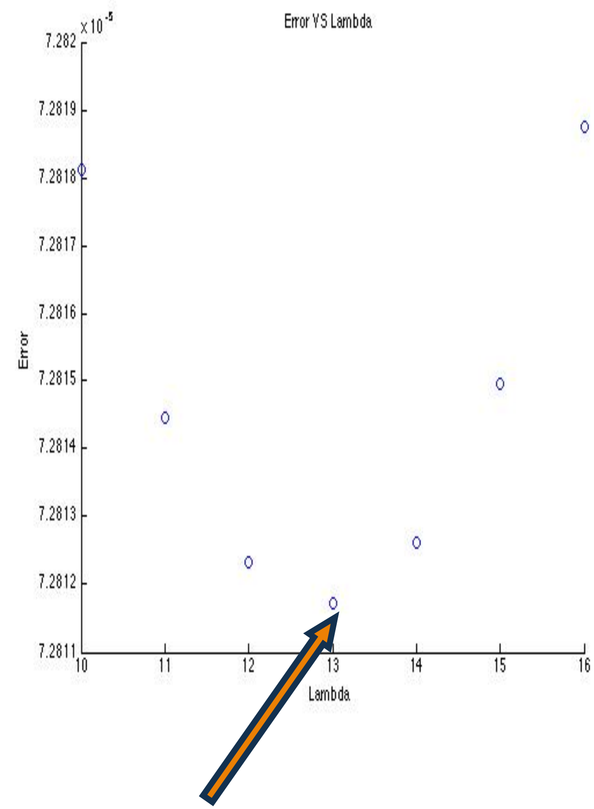
- Plots of Error VS Lambda for three models



Model I



Model II



Model III

Discussion

Table 1. Sample weights for model I (K = Intercept, Num = Number)

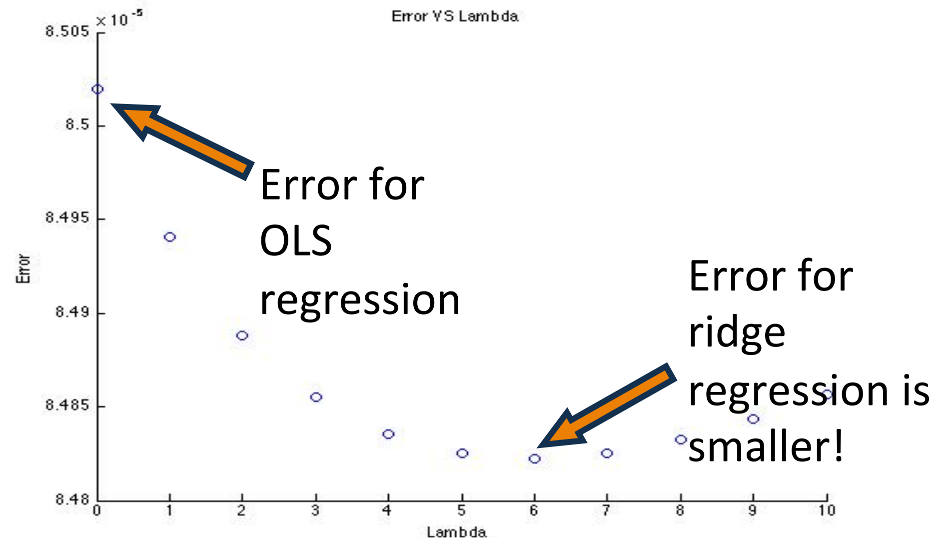
Variables	K	Num of Cylinders	Engine displacem	Horsepo wer	Weight	Accelera tion time	Origin
Weights (best lambda)	0.0873	-0.1021	0.0474	-0.2682	-0.5452	-0.0365	0.1033
Weights	0.0842	-0.0576	-0.0204	-0.2365	-0.4796	-0.0675	0.1021
Weights	0.0843	-0.0565	-0.0160	-0.2383	-0.4864	-0.0676	0.1025

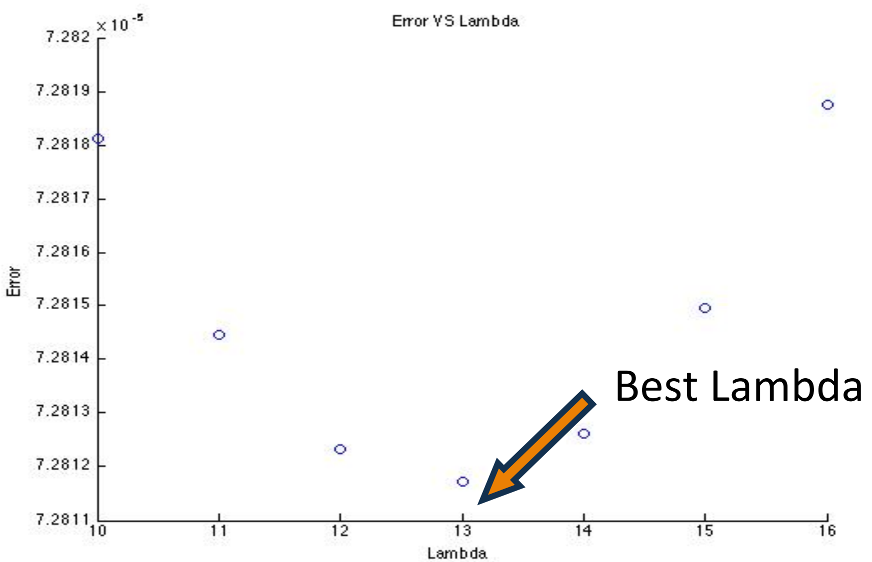
Table 3. Sample weights for model III (K = intercept)

Variables	K	Number of Cylinders	Origin	Years
Weights (best lambda)	-0.0468	-0.5582	2.2883	0.1355
Weights	-0.0342	-0.5868	2.0849	0.1027
Weights	-0.0460	-0.5780	2.2930	0.1314

- Weight of the car and horsepower are best two predictors for the MPG.
- Five variables that are negative correlated with MPG. They are number of cylinders, engine displacement, HP, weight of the vehicle and acceleration time.
- Two variables that are positively correlated with MPG. They are year of the car and the country origin. Japanese cars usually have the highest MPG, followed by Germany and U.S.

Discussion





- Ridge regression is better than OLS regression in this study
- By controlling the size of coefficients, ridge regression is a good way to battle multicollinearly for a large set of predictor variables in a regression model.

Acknowledgements

Jiannan(Jeffrey) Zhang   University of Texas at Austin