Lesson 08 Regularization

Lusine Zilfimian

April 06 (Monday), 2020

Contents

- Quiz

Contents

Contents

- Quiz
- Linear Regression (Reminder)
- Problems

Contents

Contents

- Quiz
- Linear Regression (Reminder)
- Problems
- L2 Regularization – Ridge regression
- L1 Regularization – LASSO

Contents

- Quiz
- Linear Regression (Reminder)
- Problems
- L2 Regularization – Ridge regression
- L1 Regularization – LASSO
- Selecting the Tuning Parameter

Contents

- Quiz
- Linear Regression (Reminder)
- Problems
- L2 Regularization – Ridge regression
- L1 Regularization – LASSO
- Selecting the Tuning Parameter
- Elastic Net Regression

Last Lecture ReCap

- Bring an example of Poisson experiment.

Last Lecture ReCap

- Bring an example of Poisson experiment.
- Why we cannot model mean as a linear function of independent variable?

Last Lecture ReCap

- Bring an example of Poisson experiment.
- Why we cannot model mean as a linear function of independent variable?
- What is overdispersion, how to deal with in?

Last Lecture ReCap

- Bring an example of Poisson experiment.
- Why we cannot model mean as a linear function of independent variable?
- What is overdispersion, how to deal with in?
- How to check the goodness of fit in Poisson Regression?

Linear Regression (Reminder)

- Linear regression coefficients $\hat{\beta}_{OLS}$ are the values that minimize the following RSS:

Linear Regression (Reminder)

- Linear regression coefficients $\hat{\beta}_{OLS}$ are the values that minimize the following RSS:
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 \rightarrow min$

Linear Regression (Reminder)

- Linear regression coefficients $\hat{\beta}_{OLS}$ are the values that minimize the following RSS:
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 \rightarrow min$
- $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$

Linear Regression (Reminder)

- Linear regression coefficients $\hat{\beta}_{OLS}$ are the values that minimize the following RSS:
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 \rightarrow min$
- $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$
- $Bias(\hat{\beta}_{OLS}) = \mathbb{E}(\hat{\beta}_{OLS}) - \beta = 0$

Linear Regression (Reminder)

- Linear regression coefficients $\hat{\beta}_{OLS}$ are the values that minimize the following RSS:
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 \rightarrow min$
- $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$
- $Bias(\hat{\beta}_{OLS}) = \mathbb{E}(\hat{\beta}_{OLS}) - \beta = 0$
- $var(\hat{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1}$

Linear Regression (Reminder)

- Linear regression coefficients $\hat{\beta}_{OLS}$ are the values that minimize the following RSS:
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 \rightarrow min$
- $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$
- $Bias(\hat{\beta}_{OLS}) = \mathbb{E}(\hat{\beta}_{OLS}) - \beta = 0$
- $var(\hat{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1}$
- When we have a lot of observations we can be fairly confident that the Least Squares line accurately reflects the relationship between y and x.

Linear Regression: Assumptions (Reminder)

- In practice, some of the assumptions of linear regression are violated.

Linear Regression: Assumptions (Reminder)

- In practice, some of the assumptions of linear regression are violated.
- Regularization solves the problem caused by the violation of the following assumptions:

Linear Regression: Assumptions (Reminder)

- In practice, some of the assumptions of linear regression are violated.
- Regularization solves the problem caused by the violation of the following assumptions:
- Number of observations is much larger than the number of variables ($n >> p$)

Linear Regression: Assumptions (Reminder)

- In practice, some of the assumptions of linear regression are violated.
- Regularization solves the problem caused by the violation of the following assumptions:
- Number of observations is much larger than the number of variables ($n >> p$)
- Absence of multicollinearity.

Problems

- $n \leq p$, $n > p$ problems: there are not enough observations to fit data - more, slightly less, or equal number of variables than data points.

Problems

- $n \leq p$, $n > p$ problems: there are not enough observations to fit data - more, slightly less, or equal number of variables than data points.
- In the first case the OLS estimate for β coefficient is not unique.

Problems

- $n \leq p$, $n > p$ problems: there are not enough observations to fit data - more, slightly less, or equal number of variables than data points.
- In the first case the OLS estimate for β coefficient is not unique.
- In the second case there are overfitting.

Problems

- $n \leq p$, $n > p$ problems: there are not enough observations to fit data - more, slightly less, or equal number of variables than data points.
- In the first case the OLS estimate for β coefficient is not unique.
- In the second case there are overfitting.
- Multicollinearity problem:

Problems

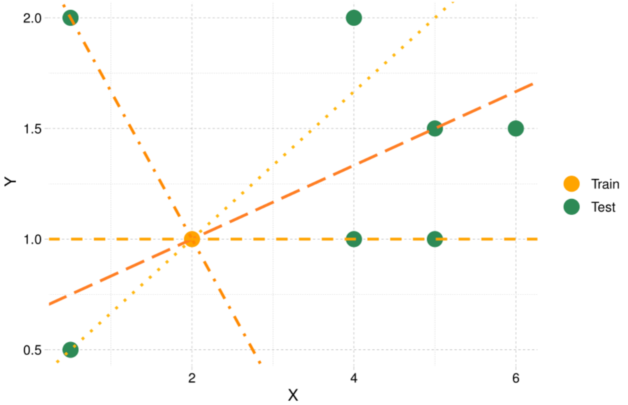
- $n \leq p, \; n > p$ problems: there are not enough observations to fit data - more, slightly less, or equal number of variables than data points.
- In the first case the OLS estimate for β coefficient is not unique.
- In the second case there are overfitting.
- Multicollinearity problem:
- With perfect collinearity $rank(X) < m \Rightarrow (X^T X)^{-1}$ does not exist.

Problems

- $n \leq p$, $n > p$ problems: there are not enough observations to fit data - more, slightly less, or equal number of variables than data points.
- In the first case the OLS estimate for β coefficient is not unique.
- In the second case there are overfitting.
- Multicollinearity problem:
- With perfect collinearity $rank(X) < m \Rightarrow (X^T X)^{-1}$ does not exist.
- With multicollinearity: $rank(X) = m$, but there are high correlation, thus standard error for $\hat{\beta}_j$ will be large.

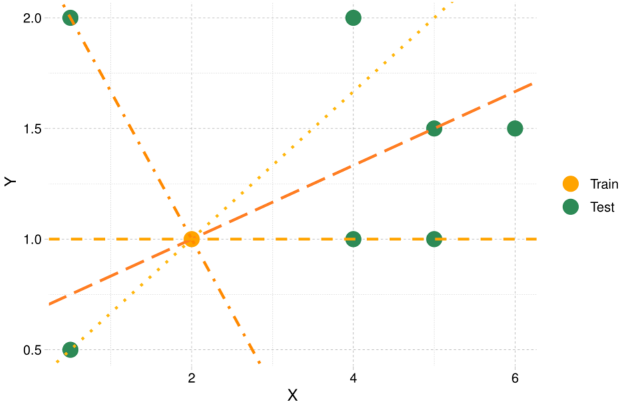
Problem with number of variables and observations

- Suppose train data consists of 1 observation and 1 variable



Problem with number of variables and observations

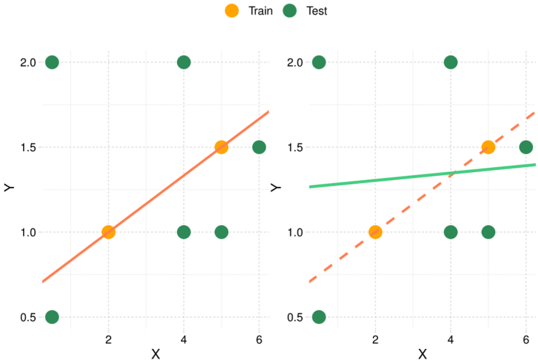
- Suppose train data consists of 1 observation and 1 variable



- All regressions has $RSS = 0$ for train data.

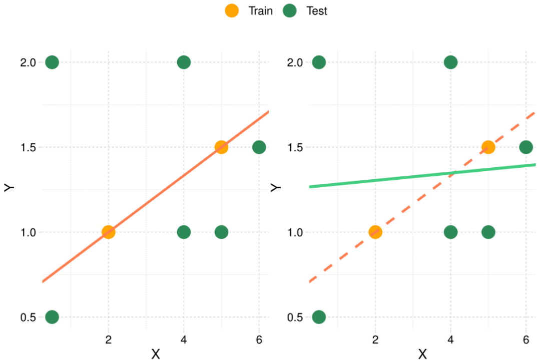
Problem with number of variables and observations

- Suppose train data consists of 2 two observations and 1 variable:



Problem with number of variables and observations

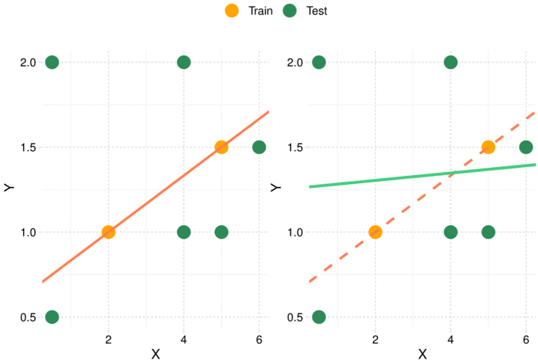
- Suppose train data consists of 2 two observations and 1 variable:



- Train $RSS = 0$, test set RSS is large.

Problem with number of variables and observations

- Suppose train data consists of 2 two observations and 1 variable:



- Train $RSS = 0$, test set RSS is large.
- The regression has high variance and zero bias.

The way of solving these problems

- **Subset Selection.** Identifying a subset of p predictors that we believe to be related to the response (using theory, significant tests, R^2, AIC, BIC etc.)

The way of solving these problems

- **Subset Selection.** Identifying a subset of p predictors that we believe to be related to the response (using theory, significant tests, R^2, AIC, BIC etc.)
- **Shrinkage (Regularization).** Fitting a model involving all p predictors and then shrinking coefficients towards zero relative to OLS estimates.

The way of solving these problems

- **Subset Selection.** Identifying a subset of p predictors that we believe to be related to the response (using theory, significant tests, R^2, AIC, BIC etc.)
- **Shrinkage (Regularization).** Fitting a model involving all p predictors and then shrinking coefficients towards zero relative to OLS estimates.
- **Dimension Reduction.** Projecting the p predictors into a m-dimensional subspace by computing m (m<p) linear combinations of p variables.

The idea of regularization

- Perform linear regression model, while shrinking the coefficients $\hat{\beta}$ toward 0.

The idea of regularization

- Perform linear regression model, while shrinking the coefficients $\hat{\beta}$ toward 0.
- Regularization introduces bias, but may significantly decrease the variance of the estimates.

The idea of regularization

- Perform linear regression model, while shrinking the coefficients $\hat{\beta}$ toward 0.
- Regularization introduces bias, but may significantly decrease the variance of the estimates.
- Regularization penalizes complex models.

The idea of regularization

- Perform linear regression model, while shrinking the coefficients $\hat{\beta}$ toward 0.
- Regularization introduces bias, but may significantly decrease the variance of the estimates.
- Regularization penalizes complex models.
- Regularization is the method of subset selection.

The idea of regularization

- Perform linear regression model, while shrinking the coefficients $\hat{\beta}$ toward 0.
- Regularization introduces bias, but may significantly decrease the variance of the estimates.
- Regularization penalizes complex models.
- Regularization is the method of subset selection.
- The main types of regularization are:

The idea of regularization

- Perform linear regression model, while shrinking the coefficients $\hat{\beta}$ toward 0.
- Regularization introduces bias, but may significantly decrease the variance of the estimates.
- Regularization penalizes complex models.
- Regularization is the method of subset selection.
- The main types of regularization are:
- L2 Regularization – Ridge regression

The idea of regularization

- Perform linear regression model, while shrinking the coefficients $\hat{\beta}$ toward 0.
- Regularization introduces bias, but may significantly decrease the variance of the estimates.
- Regularization penalizes complex models.
- Regularization is the method of subset selection.
- The main types of regularization are:
- L2 Regularization – Ridge regression
- L1 Regularization – LASSO regression

L2 Regularization – Ridge regression

- Ridge regression coefficients β_{ridge} are the values that minimize the following RSS:

L2 Regularization – Ridge regression

- Ridge regression coefficients β_{ridge} are the values that minimize the following RSS:
- $RSS_{ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\hat{\beta}_j^{\,2} \rightarrow min$

L2 Regularization – Ridge regression

- Ridge regression coefficients β_{ridge} are the values that minimize the following RSS:
- $RSS_{ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2} \rightarrow min$
- $RSS_{ridge} = RSS + \lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2} \rightarrow min$

L2 Regularization – Ridge regression

- Ridge regression coefficients β_{ridge} are the values that minimize the following RSS:
- $RSS_{ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\hat{\beta}_j^{\,2} \rightarrow min$
- $RSS_{ridge} = RSS + \lambda \sum_{j=1}^{p}\hat{\beta}_j^{\,2} \rightarrow min$
- $\lambda \sum_{j=1}^{p}\hat{\beta}_j^{\,2}$ is a shrinkage penalty

L2 Regularization – Ridge regression

- Ridge regression coefficients β_{ridge} are the values that minimize the following RSS:
- $RSS_{ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\hat{\beta}_j^{\,2} \rightarrow min$
- $RSS_{ridge} = RSS + \lambda \sum_{j=1}^{p}\hat{\beta}_j^{\,2} \rightarrow min$
- $\lambda \sum_{j=1}^{p}\hat{\beta}_j^{\,2}$ is a shrinkage penalty
- $\lambda \geq 0$ is the **tuning** parameter

L2 Regularization – Ridge regression

- The shrinkage penalty is applied to coefficients of x, but not to the intercept:

L2 Regularization – Ridge regression

- The shrinkage penalty is applied to coefficients of x, but not to the intercept:
- $RSS_{ridge} = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \rightarrow min$

L2 Regularization – Ridge regression

- The shrinkage penalty is applied to coefficients of x, but not to the intercept:
- $RSS_{ridge} = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^{\,2} \rightarrow min$
- Ridge shrinks the estimated association of each variable with the response.

L2 Regularization – Ridge regression

- The shrinkage penalty is applied to coefficients of x, but not to the intercept:
- $RSS_{ridge} = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \rightarrow min$
- Ridge shrinks the estimated association of each variable with the response.
- There is no need to shrink the intercept, which is simply a measure of the mean value of the response when $x_{i1} = x_{i2} = ... = x_{ip} = 0$.

L2 Regularization – Ridge regression

- The **main idea** of regularization is to find new line that does not fit the Train data as well (to introduce a small amount of bias) in order to have less variance.

L2 Regularization – Ridge regression

- The **main idea** of regularization is to find new line that does not fit the Train data as well (to introduce a small amount of bias) in order to have less variance.
- $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$

L2 Regularization – Ridge regression

- The **main idea** of regularization is to find new line that does not fit the Train data as well (to introduce a small amount of bias) in order to have less variance.
- $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$
- $\lambda \to 0, \hat{\beta}_{ridge} \to \hat{\beta}_{OLS}$;

L2 Regularization – Ridge regression

- The **main idea** of regularization is to find new line that does not fit the Train data as well (to introduce a small amount of bias) in order to have less variance.
- $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$
- $\lambda \to 0, \hat{\beta}_{ridge} \to \hat{\beta}_{OLS}$;
- $\lambda \to \infty, \hat{\beta}_{ridge} \to 0$

L2 Regularization – Ridge regression

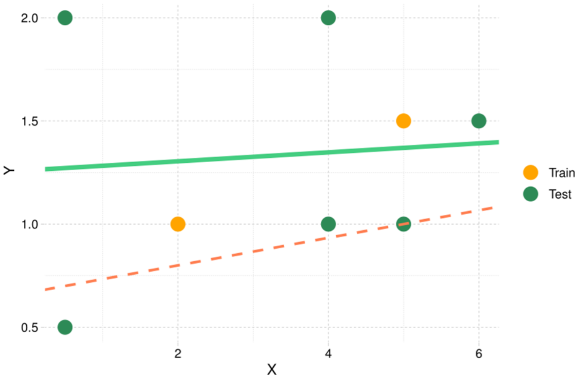
- The **main idea** of regularization is to find new line that does not fit the Train data as well (to introduce a small amount of bias) in order to have less variance.
- $\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^TY$
- $\lambda \to 0, \hat{\beta}_{ridge} \to \hat{\beta}_{OLS}$;
- $\lambda \to \infty, \hat{\beta}_{ridge} \to 0$
- $Bias(\hat{\beta}_{ridge}) = \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$;

L2 Regularization – Ridge regression

- The **main idea** of regularization is to find new line that does not fit the Train data as well (to introduce a small amount of bias) in order to have less variance.
- $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$
- $\lambda \to 0, \hat{\beta}_{ridge} \to \hat{\beta}_{OLS}$;
- $\lambda \to \infty, \hat{\beta}_{ridge} \to 0$
- $Bias(\hat{\beta}_{ridge}) = \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}$;
- $\lambda \uparrow \Rightarrow Var \downarrow$

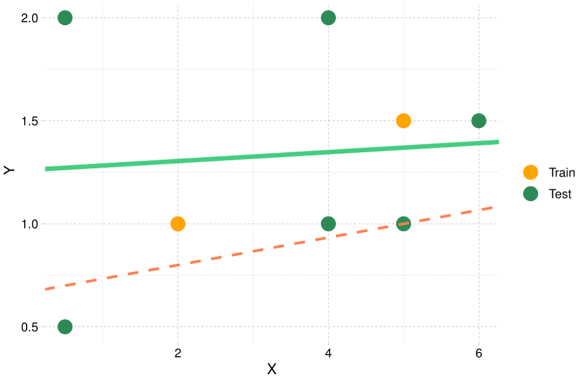
Decrease in the slope: solving the problem of n=p+1

- $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$



Decrease in the slope: solving the problem of n=p+1

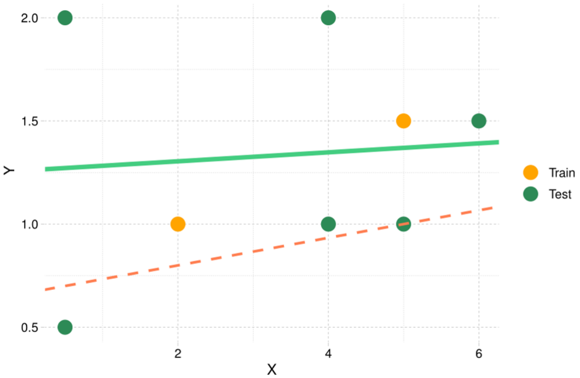
- $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$



- Now we have another slope (**smaller**) and intercept.

Decrease in the slope: solving the problem of n=p+1

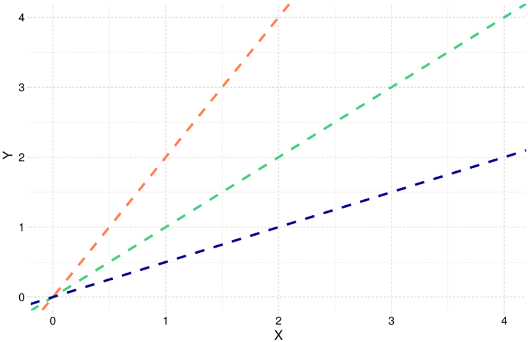
- $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$



- Now we have another slope (**smaller**) and intercept.
- Shrinking the coefficient estimates can significantly reduce their

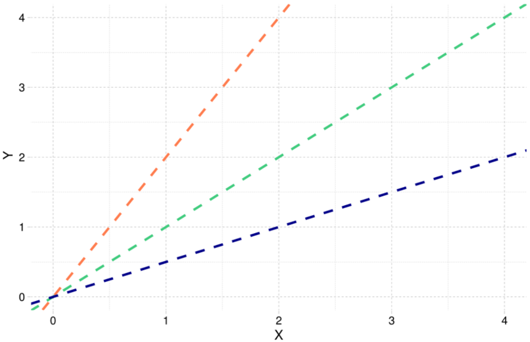
Decrease in the slope is important

- Smaller the slope less is the sensitivity to x (variables)



Decrease in the slope is important

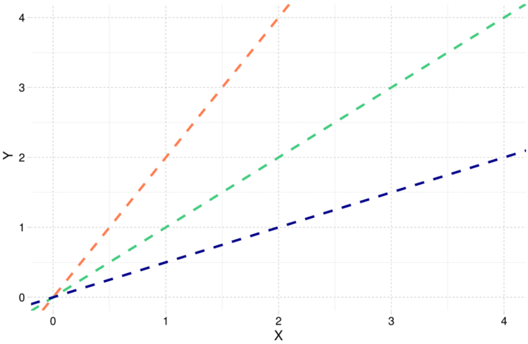
- Smaller the slope less is the sensitivity to x (variables)



- $\lambda \to \infty, \hat{\beta}_{ridge} \to 0$

Decrease in the slope is important

- Smaller the slope less is the sensitivity to x (variables)



- $\lambda \to \infty, \hat{\beta}_{ridge} \to 0$
- Less and less sensitive to x variable

Decrease in the slope: example

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

```
data.frame(y = c(10,20,30), x = c(1,1,2))
```

```
##    y x
## 1 10 1
## 2 20 1
## 3 30 2
```

Decrease in the slope: example

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

```
data.frame(y = c(10,20,30), x = c(1,1,2))
```

```
##     y x
## 1 10 1
## 2 20 1
## 3 30 2
```

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\hat{\beta}\sum_{i=1}^{n} y_i x_i + \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 \rightarrow min$

Decrease in the slope: example

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

```
data.frame(y = c(10,20,30), x = c(1,1,2))
```

```
##    y x
## 1 10 1
## 2 20 1
## 3 30 2
```

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\hat{\beta}\sum_{i=1}^{n} y_i x_i + \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 \to min$
- $RSS' = -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}\sum_{i=1}^{n} x_i^2 = 0$

Decrease in the slope: example

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

```r
data.frame(y = c(10,20,30), x = c(1,1,2))
```

```
##    y x
## 1 10 1
## 2 20 1
## 3 30 2
```

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\hat{\beta}\sum_{i=1}^{n} y_i x_i + \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 \rightarrow min$
- $RSS' = -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}\sum_{i=1}^{n} x_i^2 = 0$
- $\hat{\beta_{OLS}} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{10+20+60}{1^2+1^2+2^2} = \frac{90}{6} = 15$

Decrease in the slope: example

- $RSS_{ridge} = RSS + \lambda \hat{\beta}^2 \rightarrow min$

Decrease in the slope: example

- $RSS_{ridge} = RSS + \lambda\hat{\beta}^2 \rightarrow min$
- $RSS'_{ridge} = -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}\sum_{i=1}^{n} x_i^2 + 2\lambda\hat{\beta} = 0$

Decrease in the slope: example

- $RSS_{ridge} = RSS + \lambda\hat{\beta}^2 \rightarrow min$
- $RSS'_{ridge} = -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}\sum_{i=1}^{n} x_i^2 + 2\lambda\hat{\beta} = 0$
- $\hat{\beta_{ridge}} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2 + \lambda} = \frac{90}{6+\lambda}$

Decrease in the slope: example

- $RSS_{ridge} = RSS + \lambda\hat{\beta}^2 \rightarrow min$
- $RSS'_{ridge} = -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}\sum_{i=1}^{n} x_i^2 + 2\lambda\hat{\beta} = 0$
- $\hat{\beta}_{ridge} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2 + \lambda} = \frac{90}{6+\lambda}$
- Suppose $\lambda = 240$

Decrease in the slope: example

- $RSS_{ridge} = RSS + \lambda\hat{\beta}^2 \rightarrow min$
- $RSS'_{ridge} = -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}\sum_{i=1}^{n} x_i^2 + 2\lambda\hat{\beta} = 0$
- $\hat{\beta}_{ridge} = \dfrac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2 + \lambda} = \frac{90}{6+\lambda}$
- Suppose $\lambda = 240$
- $\hat{\beta}_{ridge}\frac{90}{6+240} = \frac{90}{6+246} = 0.37 < 15$

Solving the problem of non-existing coefficient

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

```
data.frame(y = c(10,20,30), x = c(0,0,0))
```

```
##    y x
## 1 10 0
## 2 20 0
## 3 30 0
```

Solving the problem of non-existing coefficient

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

```
data.frame(y = c(10,20,30), x = c(0,0,0))
```

```
##    y x
## 1 10 0
## 2 20 0
## 3 30 0
```

- $\hat{\beta_{OLS}} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{10+20+60}{0^2+0^2+0^2} = \frac{90}{0}$

Solving the problem of non-existing coefficient

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

```
data.frame(y = c(10,20,30), x = c(0,0,0))
```

```
##    y x
## 1 10 0
## 2 20 0
## 3 30 0
```

- $\hat{\beta_{OLS}} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{10 + 20 + 60}{0^2 + 0^2 + 0^2} = \frac{90}{0}$
- $\hat{\beta_{ridge}} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2 + \lambda} = \frac{90}{0+\lambda}$

Solving the problem of non-existing coefficient

- Suppose we have $y_i = \beta x_i + \varepsilon_i$ and we have the following data:

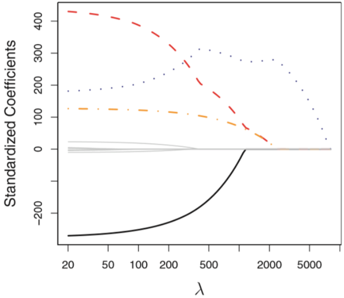
```r
data.frame(y = c(10,20,30), x = c(0,0,0))
```

```
##    y x
## 1 10 0
## 2 20 0
## 3 30 0
```

- $\hat{\beta_{OLS}} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{10 + 20 + 60}{0^2 + 0^2 + 0^2} = \frac{90}{0}$
- $\hat{\beta_{ridge}} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2 + \lambda} = \frac{90}{0+\lambda}$
- $rank(X) < m \Rightarrow (X^T X)^{-1}$ does **NOT** exist, but $(X^T X + \lambda I)^{-1}$ does exist.

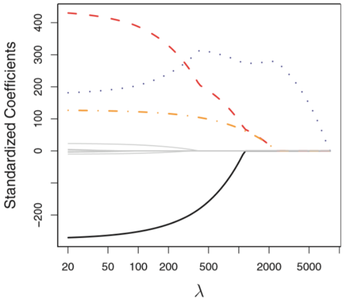
Coefficient as a function of lambda

- Ridge regression will produce a different set of coefficients for each value of λ



Coefficient as a function of lambda

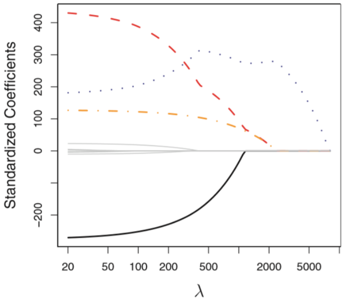
- Ridge regression will produce a different set of coefficients for each value of λ



- As λ increases, the ridge coefficient estimates shrink towards zero. When λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the null model that contains no predictors.

Coefficient as a function of lambda

- Ridge regression will produce a different set of coefficients for each value of λ



- As λ increases, the ridge coefficient estimates shrink towards zero. When λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the null model that contains no predictors.

Norms and measurement

- l_2 - norms:

Norms and measurement

- l_2 - norms:
- $||\hat{\beta}||_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_j^2}$

Norms and measurement

- l_2 - norms:
- $||\hat{\beta}||_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_j^{\,2}}$
- $||\hat{\beta}_{ridge}(\lambda)||_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_{j\,ridge}^{\,2}}$

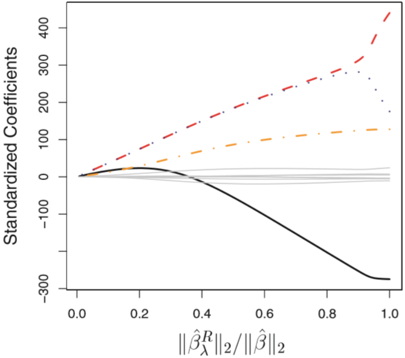
Norms and measurement

- l_2 - norms:
- $||\hat{\beta}||_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_j^{\,2}}$
- $||\hat{\beta}_{ridge}(\lambda)||_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_{j\,ridge}^{\,2}}$
- $0 < ||\hat{\beta}_{ridge}(\lambda)||_2 / ||\hat{\beta}||_2 \leq 1$

Norms and measurement

- l_2 - norms:
- $||\hat{\beta}||_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_j^{\,2}}$
- $||\hat{\beta}_{ridge}(\lambda)||_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_{j\,ridge}^{\,2}}$
- $0 < ||\hat{\beta}_{ridge}(\lambda)||_2 / ||\hat{\beta}||_2 \leq 1$
- The amount that the ridge regression coefficient estimates have been shrunken towards zero is the 2 norm of the ridge regression coefficient estimates divided by the 2 norm of the least squares estimates.

Norms and measurement



- $\lambda \uparrow ||\hat{\beta}_{ridge}(\lambda)||_2 \downarrow \Rightarrow ||\hat{\beta}_{ridge}(\lambda)||_2/||\hat{\beta}||_2 \downarrow$

Standardizing the predictors

- $\hat{\beta}_{OLS}$ is **scale invariant**

Standardizing the predictors

- $\hat{\beta}_{OLS}$ is **scale invariant**
- $\hat{\beta}_{ridge}$ is **NOT** scale invariant

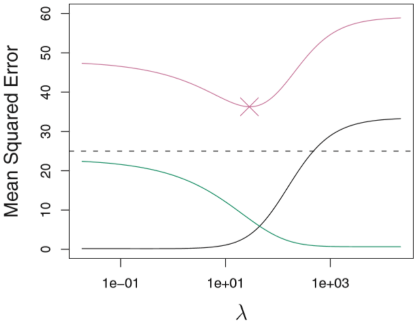
Standardizing the predictors

- $\hat{\beta}_{OLS}$ is **scale invariant**
- $\hat{\beta}_{ridge}$ is **NOT** scale invariant
- The solution is applying standardized predictors

Standardizing the predictors

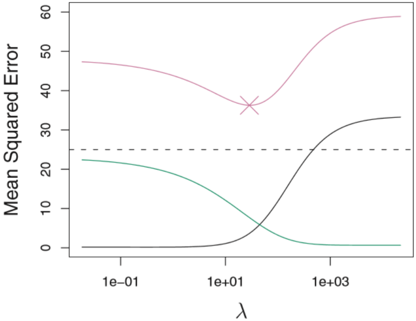
- $\hat{\beta}_{OLS}$ is **scale invariant**
- $\hat{\beta}_{ridge}$ is **NOT** scale invariant
- The solution is applying standardized predictors
- $\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum(x_{ij}-\bar{x}_j)^2}}$

Bias-variance trade-off as a function of lambda



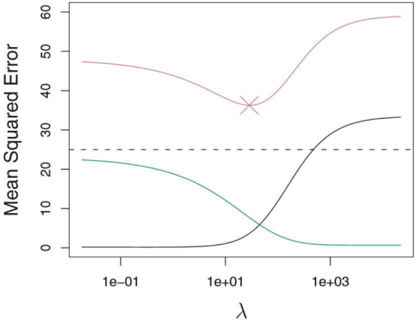
- $\lambda \uparrow \Rightarrow Bias \uparrow \ and \ Var \downarrow$

Bias-variance trade-off as a function of lambda



- $\lambda \uparrow \Rightarrow$ *Bias* $\uparrow$ *and Var* $\downarrow$
- Up to about $\lambda = 10$, the variance, plotted in green, decreases rapidly, with very little increase in bias, plotted in black.

Bias-variance trade-off as a function of lambda



- $\lambda \uparrow \Rightarrow Bias \uparrow$ *and* $Var \downarrow$
- Up to about $\lambda = 10$, the variance, plotted in green, decreases rapidly, with very little increase in bias, plotted in black.
- The **MSE** drops considerably as λ increases from 0 to 10.

The LASSO

The disadvantage of Ridge

- Ridge regression will include all p predictors in the final model.

The LASSO

The disadvantage of Ridge

- Ridge regression will include all p predictors in the final model.
- The penalty $\lambda \sum_{j=1}^{p} \hat{\beta}_j^{\,2}$ will shrink all of the coefficients towards zero (reduce the magnitudes of the coefficients), but it will not set any of them exactly to zero for any real number of λ.

The LASSO

The disadvantage of Ridge

- Ridge regression will include all p predictors in the final model.
- The penalty $\lambda \sum_{j=1}^{p} \hat{\beta}_j^{\,2}$ will shrink all of the coefficients towards zero (reduce the magnitudes of the coefficients), but it will not set any of them exactly to zero for any real number of λ.
- The Lasso is an alternative to ridge regression overcoming this disadvantage.

The LASSO

The disadvantage of Ridge

- Ridge regression will include all p predictors in the final model.
- The penalty $\lambda \sum_{j=1}^{p} \hat{\beta}_j^2$ will shrink all of the coefficients towards zero (reduce the magnitudes of the coefficients), but it will not set any of them exactly to zero for any real number of λ.
- The Lasso is an alternative to ridge regression overcoming this disadvantage.
- Lasso Regression is similar to Ridge Regression with the difference in Penalty term.

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:
- $RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:
- $RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$
- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:
- $RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| \rightarrow min$
- $RSS_{LASSO} = RSS + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| \rightarrow min$
- $\lambda\sum_{j=1}^{p}|\hat{\beta}_j|$ is a shrinkage penalty

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:
- $RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$
- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$
- $\lambda \sum_{j=1}^{p} |\hat{\beta}_j|$ is a shrinkage penalty
- $\lambda \geq 0$ is the **tuning** parameter

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:
- $RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\hat{\beta}_j| \rightarrow min$
- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p}|\hat{\beta}_j| \rightarrow min$
- $\lambda \sum_{j=1}^{p}|\hat{\beta}_j|$ is a shrinkage penalty
- $\lambda \geq 0$ is the **tuning** parameter
- Like Ridge Regression:

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:
- $RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| \to min$
- $RSS_{LASSO} = RSS + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| \to min$
- $\lambda\sum_{j=1}^{p}|\hat{\beta}_j|$ is a shrinkage penalty
- $\lambda \geq 0$ is the **tuning** parameter
- Like Ridge Regression:
- $\lambda \to 0, \hat{\beta}_{LASSO} \to \hat{\beta}_{OLS}$;

Least Absolute Shrinkage and Selection Operator

- Lasso regression coefficients β_{LASSO} are the values that minimize the following RSS:
- $RSS_{LASSO} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$
- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$
- $\lambda \sum_{j=1}^{p} |\hat{\beta}_j|$ is a shrinkage penalty
- $\lambda \geq 0$ is the **tuning** parameter
- Like Ridge Regression:
- $\lambda \rightarrow 0, \hat{\beta}_{LASSO} \rightarrow \hat{\beta}_{OLS}$;
- $\lambda \rightarrow \infty, \hat{\beta}_{LASSO} \rightarrow 0$

Norms

- l_1- norms:

Norms

- l_1- norms:
- $||\hat{\beta}||_1 = \sum_{j=1}^{p} |\hat{\beta}_j|$

Norms

- l_1- norms:
- $||\hat{\beta}||_1 = \sum_{j=1}^{p} |\hat{\beta}_j|$
- $||\hat{\beta}_{LASSO}(\lambda)||_1 = \sum_{j=1}^{p} |\hat{\beta}_{j\,LASSO}|$

Norms

- l_1- norms:
- $||\hat{\beta}||_1 = \sum_{j=1}^{p} |\hat{\beta}_j|$
- $||\hat{\beta}_{LASSO}(\lambda)||_1 = \sum_{j=1}^{p} |\hat{\beta}_{j\,LASSO}|$
- $0 < ||\hat{\beta}_{LASSO}(\lambda)||_1 / ||\hat{\beta}||_1 \leq 1$

Norms

- l_1- norms:
- $||\hat{\beta}||_1 = \sum_{j=1}^{p} |\hat{\beta}_j|$
- $||\hat{\beta}_{LASSO}(\lambda)||_1 = \sum_{j=1}^{p} |\hat{\beta}_{j\,LASSO}|$
- $0 < ||\hat{\beta}_{LASSO}(\lambda)||_1 / ||\hat{\beta}||_1 \leq 1$
- The amount that the ridge regression coefficient estimates have been shrunken towards zero is the 1 norm of the lasso regression coefficient estimates divided by the 1 norm of the least squares estimates.

Another Formulation for Ridge Regression and the Lasso

- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$

Another Formulation for Ridge Regression and the Lasso

- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$
- For every value of λ, there is some s such that the equation above and below will give the same lasso coefficient estimates.

Another Formulation for Ridge Regression and the Lasso

- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \rightarrow min$
- For every value of λ, there is some s such that the equation above and below will give the same lasso coefficient estimates.
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 \rightarrow min$

Another Formulation for Ridge Regression and the Lasso

- $RSS_{LASSO} = RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \to min$
- For every value of λ, there is some s such that the equation above and below will give the same lasso coefficient estimates.
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 \to min$
- $\sum_{j=1}^{p} |\hat{\beta}_j| \leq s$

Another Formulation for Ridge Regression and the Lasso

- $RSS_{Ridge} = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \rightarrow min$

Another Formulation for Ridge Regression and the Lasso

- $RSS_{Ridge} = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^{\,2} \rightarrow min$
- For every value of λ, there is some s such that the equation above and below will give the same lasso coefficient estimates.

Another Formulation for Ridge Regression and the Lasso

- $RSS_{Ridge} = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \rightarrow min$
- For every value of λ, there is some s such that the equation above and below will give the same lasso coefficient estimates.
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 \rightarrow min$

Another Formulation for Ridge Regression and the Lasso

- $RSS_{Ridge} = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \to min$
- For every value of λ, there is some s such that the equation above and below will give the same lasso coefficient estimates.
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 \to min$
- $\sum_{j=1}^{p} \hat{\beta}_j^2 \le s$

Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$

Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$
- **LASSO**

Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$
- **LASSO**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$

Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$
- **LASSO**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$
- $|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$

Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$
- **LASSO**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$
- $|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$
- **Ridge**

Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$
- **LASSO**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$
- $|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$
- **Ridge**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$

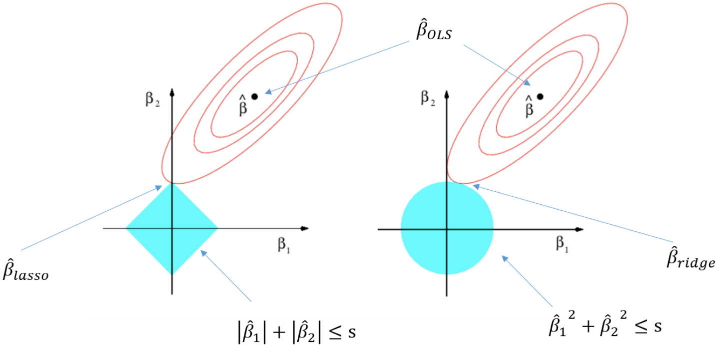
Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$
- **LASSO**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$
- $|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$
- **Ridge**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$
- $\hat{\beta}_1^{\,2} + \hat{\beta}_2^{\,2} \leq s$

Another Formulation for Ridge Regression and the Lasso

- Suppose $p = 2$
- **LASSO**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$
- $|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$
- **Ridge**
- $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \rightarrow min$
- $\hat{\beta}_1^{\,2} + \hat{\beta}_2^{\,2} \leq s$
- If s is sufficiently large, then the constraint regions will contain $\hat{\beta}$, and so the ridge regression and lasso estimates will be the same as the least squares estimates.

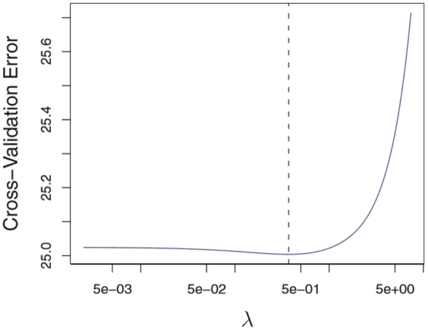
Another Formulation for Ridge Regression and the Lasso



- Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero.

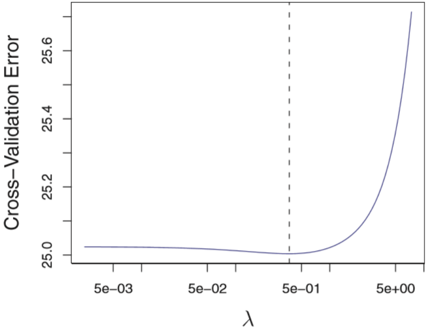
Selecting the Tuning Parameter

- We then select the tuning parameter value for which the cross-validation error is smallest.



Selecting the Tuning Parameter

- We then select the tuning parameter value for which the cross-validation error is smallest.



- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1:m$:
- for k in $1:k$:
- keep fold k as hold-out data

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:
- keep fold k as hold-out data
- use the remaining folds and $\lambda = \lambda_m$ to estimate $\hat{\beta}_{ridge}(\hat{\beta}_{LASSO})$

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:
- keep fold k as hold-out data
- use the remaining folds and $\lambda = \lambda_m$ to estimate $\hat{\beta}_{ridge}(\hat{\beta}_{LASSO})$
- predict hold-out data $y_{test,k} = x_{test,k}\hat{\beta}_{ridge}$

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:
- keep fold k as hold-out data
- use the remaining folds and $\lambda = \lambda_m$ to estimate $\hat{\beta}_{ridge}(\hat{\beta}_{LASSO})$
- predict hold-out data $y_{test,k} = x_{test,k}\hat{\beta}_{ridge}$
- compute a sum of squared residuals: $RSS_k = \sum(y - y_{test,k})^2$

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:
- keep fold k as hold-out data
- use the remaining folds and $\lambda = \lambda_m$ to estimate $\hat{\beta}_{ridge}(\hat{\beta}_{LASSO})$
- predict hold-out data $y_{test,k} = x_{test,k}\hat{\beta}_{ridge}$
- compute a sum of squared residuals: $RSS_k = \sum(y - y_{test,k})^2$
- end for k

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:
- keep fold k as hold-out data
- use the remaining folds and $\lambda = \lambda_m$ to estimate $\hat{\beta}_{ridge}(\hat{\beta}_{LASSO})$
- predict hold-out data $y_{test,k} = x_{test,k}\hat{\beta}_{ridge}$
- compute a sum of squared residuals: $RSS_k = \sum(y - y_{test,k})^2$
- end for k
- average RSS over the folds: $RSS_m = \frac{1}{k}\sum RSS_k$

Cross validation for lambda

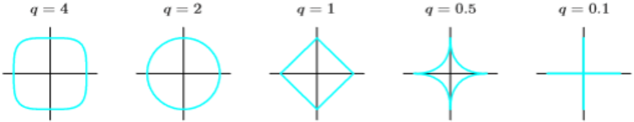
- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:
- keep fold k as hold-out data
- use the remaining folds and $\lambda = \lambda_m$ to estimate $\hat{\beta}_{ridge}(\hat{\beta}_{LASSO})$
- predict hold-out data $y_{test,k} = x_{test,k}\hat{\beta}_{ridge}$
- compute a sum of squared residuals: $RSS_k = \sum(y - y_{test,k})^2$
- end for k
- average RSS over the folds: $RSS_m = \frac{1}{k}\sum RSS_k$
- end for p

Cross validation for lambda

- We should choose a set of m values of λ to test, split the dataset into k folds, and follow this algorithm:
- for p in $1 : m$:
- for k in $1 : k$:
- keep fold k as hold-out data
- use the remaining folds and $\lambda = \lambda_m$ to estimate $\hat{\beta}_{ridge}(\hat{\beta}_{LASSO})$
- predict hold-out data $y_{test,k} = x_{test,k}\hat{\beta}_{ridge}$
- compute a sum of squared residuals: $RSS_k = \sum(y - y_{test,k})^2$
- end for k
- average RSS over the folds: $RSS_m = \frac{1}{k}\sum RSS_k$
- end for p
- Oprimal value: λ_m which correspond to min RSS_m

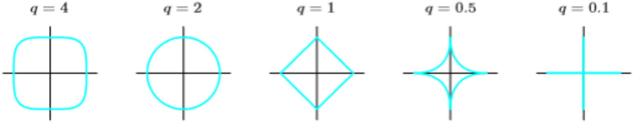
Other types of regularization

- For regularization different type of norms can be used:



Other types of regularization

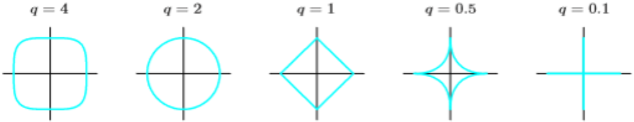
- For regularization different type of norms can be used:



- $||\hat{\beta}||_p = \{\sum_{j=1}^{p} |\hat{\beta}_j|^q\}^{\frac{1}{q}}$

Other types of regularization

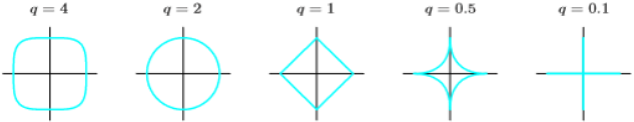
- For regularization different type of norms can be used:



- $||\hat{\beta}||_p = \{\sum_{j=1}^{p} |\hat{\beta}_j|^q\}^{\frac{1}{q}}$
- Or combination of different norms.

Other types of regularization

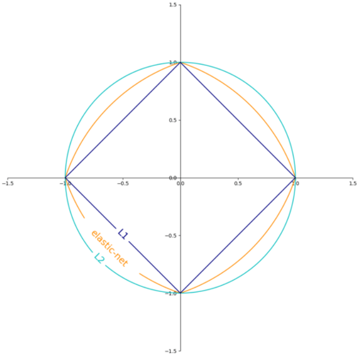
- For regularization different type of norms can be used:



- $||\hat{\beta}||_p = \{\sum_{j=1}^{p} |\hat{\beta}_j|^q\}^{\frac{1}{q}}$
- Or combination of different norms.
- The most popular combination is **Elastic Net Regression**.

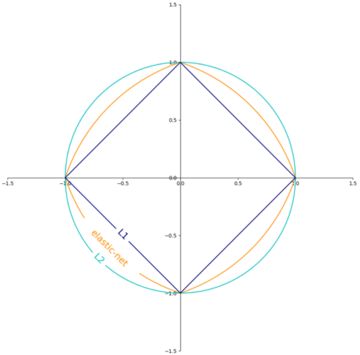
Regularization: Elastic Net

- Elastic Net – combines both methods



Regularization: Elastic Net

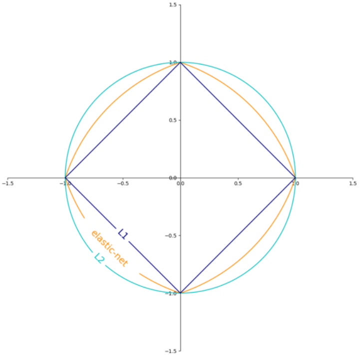
- Elastic Net – combines both methods



- Lasso equates coefficients for non important variables to zero

Regularization: Elastic Net

- Elastic Net – combines both methods



- Lasso equates coefficients for non important variables to zero
- Ridge regression shrinks the coefficients close to zero, but nor exactly to zero.

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:
- $RSS_{en} =$
$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda(\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2} + (1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|) \rightarrow min$

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:
- $RSS_{en} =$
  $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda(\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2} + (1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|) \rightarrow min$
- $(1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|$ is the penalty for **Lasso**

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:
- $RSS_{en} =$ $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda(\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2} + (1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|) \rightarrow min$
- $(1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|$ is the penalty for **Lasso**
- $\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2}$ is the penalty for **Ridge**

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:
- $RSS_{en} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda(\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2} + (1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|) \rightarrow min$
- $(1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|$ is the penalty for **Lasso**
- $\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2}$ is the penalty for **Ridge**
- $\lambda = 0 \Rightarrow OLS$

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:
- $RSS_{en} =$
  $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 + \lambda(\alpha \sum_{j=1}^{p} \hat{\beta}_j^{\,2} + (1-\alpha) \sum_{j=1}^{p} |\hat{\beta}_j|) \rightarrow min$
- $(1-\alpha) \sum_{j=1}^{p} |\hat{\beta}_j|$ is the penalty for **Lasso**
- $\alpha \sum_{j=1}^{p} \hat{\beta}_j^{\,2}$ is the penalty for **Ridge**
- $\lambda = 0 \Rightarrow OLS$
- $\alpha = 1 \Rightarrow Ridge$

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:
- $RSS_{en} =$ $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda(\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2} + (1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|) \to min$
- $(1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|$ is the penalty for **Lasso**
- $\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2}$ is the penalty for **Ridge**
- $\lambda = 0 \Rightarrow OLS$
- $\alpha = 1 \Rightarrow Ridge$
- $\alpha = 0 \Rightarrow LASSO$

Regularization: Elastic Net

- Elastic Net Regression coefficients β_{en} are the values that minimize the following RSS:
- $RSS_{en} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda(\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2} + (1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|) \rightarrow min$
- $(1-\alpha)\sum_{j=1}^{p}|\hat{\beta}_j|$ is the penalty for **Lasso**
- $\alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2}$ is the penalty for **Ridge**
- $\lambda = 0 \Rightarrow OLS$
- $\alpha = 1 \Rightarrow Ridge$
- $\alpha = 0 \Rightarrow LASSO$
- $\{\alpha, \lambda\} \neq 0 \Rightarrow$ *Elastic Net*

Regularization: Summary

- **Ridge regression** is useful when all variables need to be incorporated in the model according to domain knowledge.

Regularization: Summary

- **Ridge regression** is useful when all variables need to be incorporated in the model according to domain knowledge.
- **Lasso regression** is useful for subset selection, because only the most significant variables are kept in the final model.

Regularization: Summary

- **Ridge regression** is useful when all variables need to be incorporated in the model according to domain knowledge.
- **Lasso regression** is useful for subset selection, because only the most significant variables are kept in the final model.
- **Elastic Net regression** is useful if the knowledge about data is not available.