Penalized Regression

Nate Wells

Math 243: Stat Learning

October 11th, 2021

Outline

In today's class, we will. . .

- Investigate the relationship between coefficient size and variance in linear models
- Discuss penalized regression models as means of improving MSE of linear models

Section 1

Penalized Regression

Motivation

- Recall, for SLR, $\hat{\beta}_0, \hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Motivation

- Recall, for SLR, $\hat{\beta}_0, \hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Under the standard assumptions, the coefficients produced by least squares regression are unbiased.

Motivation

- Recall, for SLR, $\hat{\beta}_0, \hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- Under the standard assumptions, the coefficients produced by least squares regression are unbiased.
- That is, if the true relationship between Y and X is linear $Y = \beta_0 + \beta_1 X + \epsilon$, then

$$E[\hat{\beta}_0] = \beta_0 \qquad E[\hat{\beta}_1] = \beta_1$$

Motivation

- Recall, for SLR, $\hat{\beta}_0, \hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Under the standard assumptions, the coefficients produced by least squares regression are unbiased.
- That is, if the true relationship between Y and X is linear $Y = \beta_0 + \beta_1 X + \epsilon$, then

$$E[\hat{\beta}_0] = \beta_0 \qquad E[\hat{\beta}_1] = \beta_1$$

- Moreover, among all **unbiased** linear models, the least squares model has the lowest variance.

Motivation

- Recall, for SLR, $\hat{\beta}_0, \hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Under the standard assumptions, the coefficients produced by least squares regression are unbiased.
- That is, if the true relationship between Y and X is linear $Y = \beta_0 + \beta_1 X + \epsilon$, then

$$E[\hat{\beta}_0] = \beta_0 \qquad E[\hat{\beta}_1] = \beta_1$$

- Moreover, among all **unbiased** linear models, the least squares model has the lowest variance.
- Does this mean that the least squares model has the lowest MSE among all linear models?

Motivation

- Recall, for SLR, $\hat{\beta}_0, \hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Under the standard assumptions, the coefficients produced by least squares regression are unbiased.
- That is, if the true relationship between Y and X is linear $Y = \beta_0 + \beta_1 X + \epsilon$, then

$$E[\hat{\beta}_0] = \beta_0 \qquad E[\hat{\beta}_1] = \beta_1$$

- Moreover, among all **unbiased** linear models, the least squares model has the lowest variance.
- Does this mean that the least squares model has the lowest MSE among all linear models?
  - No! MSE is a combination of bias and variance.
  - It is possible that a small *increase* in bias can correspond to large *decrease* in variance.

Shrinking Coefficients

- Suppose the true relationship between Y and X_1, X_2 is given by
$$Y = 1 + X_1 + 5X_2 + \epsilon \quad \epsilon \sim N(0,1).$$
- Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ be the model coefficient estimates given by least squares regression. Which of the following models has higher variance in predictor estimates? Higher bias?

Shrinking Coefficients

- Suppose the true relationship between Y and X_1, X_2 is given by

$$Y = 1 + X_1 + 5X_2 + \epsilon \quad \epsilon \sim N(0,1).$$

- Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ be the model coefficient estimates given by least squares regression. Which of the following models has higher variance in predictor estimates? Higher bias?

$$\begin{aligned}\text{Model 1:} \quad \hat{y} =& \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \text{Model 2:} \quad \hat{y} =& \hat{\beta}_0 + 0.97 \cdot \hat{\beta}_1 x_1 + 0.98 \cdot \hat{\beta}_2 x_2\end{aligned}$$

Shrinking Coefficients

- Suppose the true relationship between Y and X_1, X_2 is given by

$$Y = 1 + X_1 + 5X_2 + \epsilon \quad \epsilon \sim N(0, 1).$$

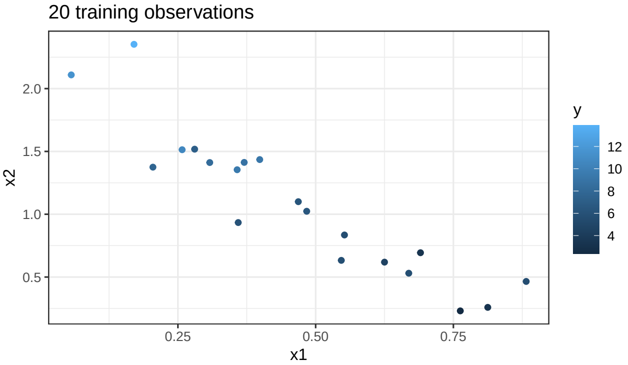
- Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ be the model coefficient estimates given by least squares regression. Which of the following models has higher variance in predictor estimates? Higher bias?

$$\begin{aligned}
\text{Model 1:} \quad \hat{y} =& \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\
\text{Model 2:} \quad \hat{y} =& \hat{\beta}_0 + 0.97 \cdot \hat{\beta}_1 x_1 + 0.98 \cdot \hat{\beta}_2 x_2
\end{aligned}$$

- Model 2 has higher bias, but lower variance.

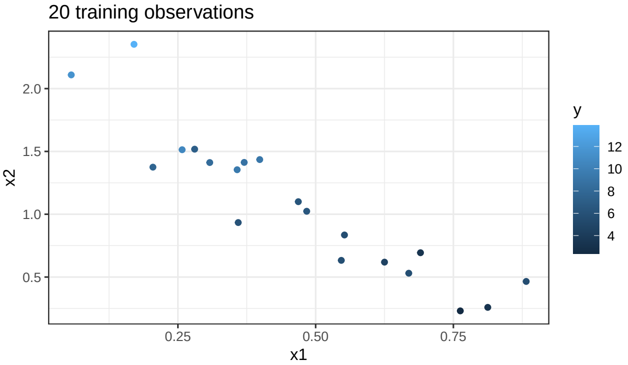
A Linear Model

- Consider the following training data for the model: $Y = 1 + X_1 + 5X_2 + \epsilon \quad \epsilon \sim N(0,1)$



A Linear Model

- Consider the following training data for the model: $Y = 1 + X_1 + 5X_2 + \epsilon \quad \epsilon \sim N(0,1)$



- What are some likely problems with the MLR model?

Bias-Variance in Least Squares

- Using least squares, the model estimates are

$$\hat{Y} = -0.5 + 2.8X_1 + 5.8X_2$$

Bias-Variance in Least Squares

- Using least squares, the model estimates are

$$\hat{Y} = -0.5 + 2.8X_1 + 5.8X_2$$

- Let's consider variance and bias for estimate Y when $X_1 = 0.25$ and $X_2 = .5$.

Bias-Variance in Least Squares

- Using least squares, the model estimates are

$$\hat{Y} = -0.5 + 2.8X_1 + 5.8X_2$$

- Let's consider variance and bias for estimate Y when $X_1 = 0.25$ and $X_2 = .5$.
  - Using the true model, the expected value of Y is

$$Y = 1 + X_1 + 5 \cdot X_2 = 1 + 0.25 + 5 \cdot 0.5 = 3.75$$

Bias-Variance in Least Squares

- Using least squares, the model estimates are

$$\hat{Y} = -0.5 + 2.8X_1 + 5.8X_2$$

- Let's consider variance and bias for estimate Y when $X_1 = 0.25$ and $X_2 = .5$.
    - Using the true model, the expected value of Y is

    $$Y = 1 + X_1 + 5 \cdot X_2 = 1 + 0.25 + 5 \cdot 0.5 = 3.75$$

    - Using the least squares model from training data, the predicted value of Y is

    $$Y = -0.5 + 2.8X_1 + 5.8X_2 = -0.5 + 2.8 \cdot 0.25 + 5.8 \cdot 0.5 = 3.1$$

Bias-Variance in Least Squares

- Using least squares, the model estimates are

$$\hat{Y} = -0.5 + 2.8X_1 + 5.8X_2$$

- Let's consider variance and bias for estimate Y when $X_1 = 0.25$ and $X_2 = .5$.
  - Using the true model, the expected value of Y is

$$Y = 1 + X_1 + 5 \cdot X_2 = 1 + 0.25 + 5 \cdot 0.5 = 3.75$$

  - Using the least squares model from training data, the predicted value of Y is

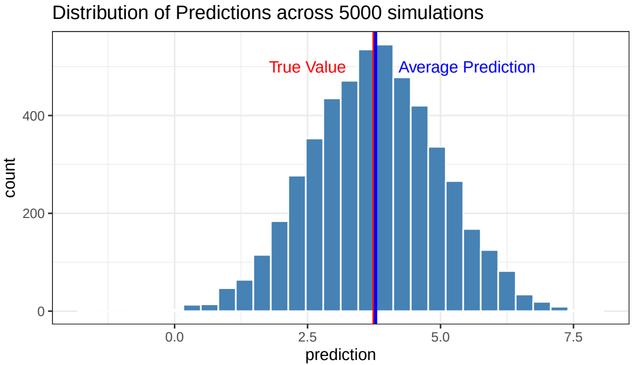
$$Y = -0.5 + 2.8X_1 + 5.8X_2 = -0.5 + 2.8 \cdot 0.25 + 5.8 \cdot 0.5 = 3.1$$

- But how will the predicted value change if we repeat across 5000 simulations from the model?

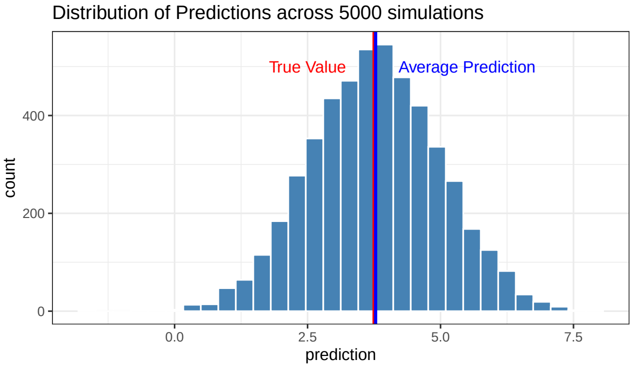
Simulation

```
set.seed(1011)
test_point <- data.frame(x1 = 0.25, x2 = .5)

trials<-5000
prediction <- rep(NA, trials)
for (i in 1:trials){
  e<- rnorm(20,0,1)
  y<- 1 + x1 + 5*x2 + e
  sim_data <- data.frame(x1,x2,y)
  mod <- lm(y ~ x1 + x2, data = sim_data)
  prediction[i] <- predict(mod, test_point)
}

simulation <- data.frame(trial_num = 1:trials, prediction)
```

Prediction Distribution



Prediction Distribution




```
simulation %>% summarize(
  mean = mean(prediction), variance = var(prediction))
```

```
##        mean variance
## 1 3.772056 1.480935
```

A Shrunken Model

- Now suppose we use the model algorithm

$$\hat{y} = \hat{\beta}_0 + 0.97 \cdot \hat{\beta}_1 x_1 + 0.98 \cdot \hat{\beta}_2 x_2$$

- Since $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are unbiased, then the expected prediction for Y when $X_1 = 0.25$ and $X_2 = 0.5$ is

$$E[\hat{y}] = \beta_1 + 0.97 \cdot \beta_1 x_1 + 0.98 \cdot \beta_2 x_2 = 1 + 0.97 \cdot 0.25 + 0.98 \cdot 5 \cdot 0.5 = 3.69$$

A Shrunken Model

- Now suppose we use the model algorithm

$$\hat{y} = \hat{\beta}_0 + 0.97 \cdot \hat{\beta}_1 x_1 + 0.98 \cdot \hat{\beta}_2 x_2$$

- Since $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are unbiased, then the expected prediction for Y when $X_1 = 0.25$ and $X_2 = 0.5$ is

$$E[\hat{y}] = \beta_1 + 0.97 \cdot \beta_1 x_1 + 0.98 \cdot \beta_2 x_2 = 1 + 0.97 \cdot 0.25 + 0.98 \cdot 5 \cdot 0.5 = 3.69$$

- Based on the first simulation, the model estimate is

$$\hat{Y} = -0.5 + 0.97 \cdot 2.8X_1 + 0.98 \cdot 5.8X_2 = -0.5 + 2.71X_1 + 5.68X_2$$

A Shrunken Model

- Now suppose we use the model algorithm

$$\hat{y} = \hat{\beta}_0 + 0.97 \cdot \hat{\beta}_1 x_1 + 0.98 \cdot \hat{\beta}_2 x_2$$

- Since $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are unbiased, then the expected prediction for Y when $X_1 = 0.25$ and $X_2 = 0.5$ is

$$E[\hat{y}] = \beta_1 + 0.97 \cdot \beta_1 x_1 + 0.98 \cdot \beta_2 x_2 = 1 + 0.97 \cdot 0.25 + 0.98 \cdot 5 \cdot 0.5 = 3.69$$

- Based on the first simulation, the model estimate is

$$\hat{Y} = -0.5 + 0.97 \cdot 2.8 X_1 + 0.98 \cdot 5.8 X_2 = -0.5 + 2.71 X_1 + 5.68 X_2$$

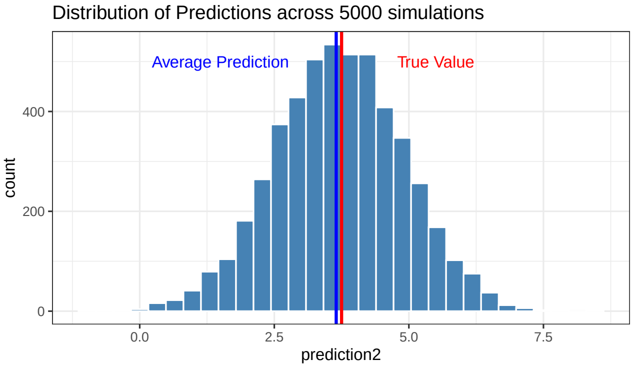
- And the prediction when $X_1 = 0.25$ and $X_2 = 0.5$ is

$$\hat{y} = -0.5 + 2.71 X_1 + 5.68 X_2 = -0.5 + 2.71 \cdot 0.25 + 5.68 \cdot 0.5 = 3.525$$

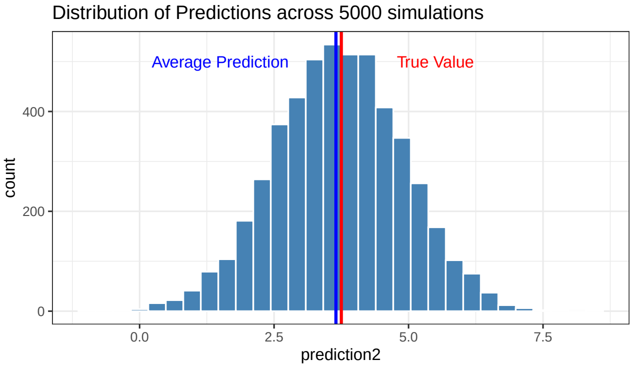
Simulation II

```
set.seed(1001)

trials<-5000
prediction2 <- rep(NA, trials)
for (i in 1:trials){
  e<- rnorm(20,0,1)
  y<- 1 + x1 + 5*x2 + e
  sim_data <- data.frame(x1,x2,y)
  mod <- lm(y ~ x1 + x2, data = sim_data)
  b0 <- 1*coef(mod)[1]
  b1 <- .97*coef(mod)[2]
  b2 <- .98*coef(mod)[3]
  prediction2[i] <- b0 + b1*0.25 + b2*0.5
}

simulation2 <- data.frame(trial_num = 1:trials, prediction2)
```

Prediction Distribution



Prediction Distribution




```
simulation2 %>% summarize(
  mean = mean(prediction2), variance = var(prediction2))
```

```
##      mean variance
## 1 3.70387 1.434099
```

Model Comparison

- True relationship: $Y = 1 + X_1 + 5X_2 + \epsilon$

Model Comparison

- True relationship: $Y = 1 + X_1 + 5X_2 + \epsilon$
- Model 1: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

```
##       mean variance avg_error
## 1 3.772056 1.480935  1.481125
```

Model Comparison

- True relationship: $Y = 1 + X_1 + 5X_2 + \epsilon$
- Model 1: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

```
##       mean variance avg_error
## 1 3.772056 1.480935  1.481125
```

- Model 2: $\hat{y} = \hat{\beta}_0 + 0.97 \cdot \hat{\beta}_1 x_1 + 0.98 \cdot \hat{\beta}_2 x_2$

```
##      mean variance avg_error
## 1 3.70387 1.434099  1.435941
```

Model Comparison

- True relationship: $Y = 1 + X_1 + 5X_2 + \epsilon$
- Model 1: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

```
##       mean variance avg_error
## 1 3.772056 1.480935  1.481125
```

- Model 2: $\hat{y} = \hat{\beta}_0 + 0.97 \cdot \hat{\beta}_1 x_1 + 0.98 \cdot \hat{\beta}_2 x_2$

```
##      mean variance avg_error
## 1 3.70387 1.434099  1.435941
```

- It looks like the model with smaller coefficients actually performed better!