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?

$$\text{Model 1:} \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

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

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?

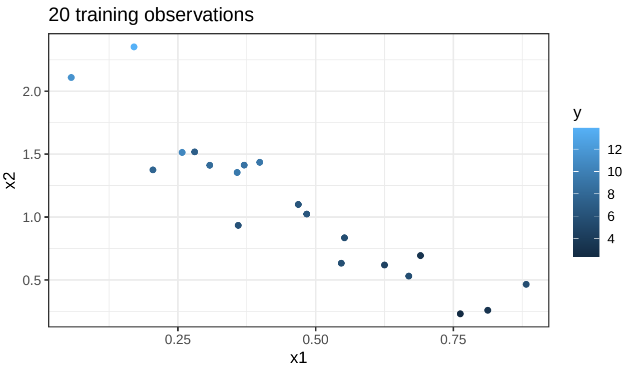
$$\text{Model 1:} \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

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

- Model 2 has higher bias, but lower variance.

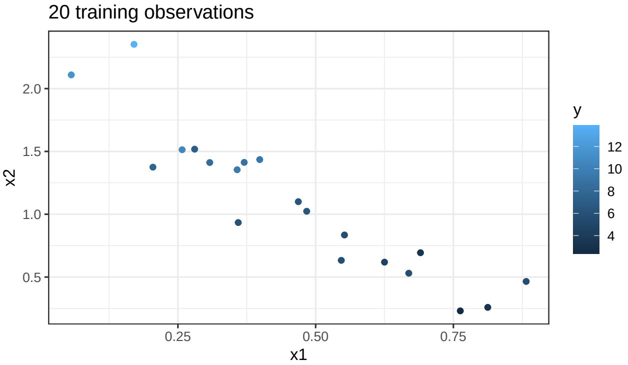
A Linear Model

- 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

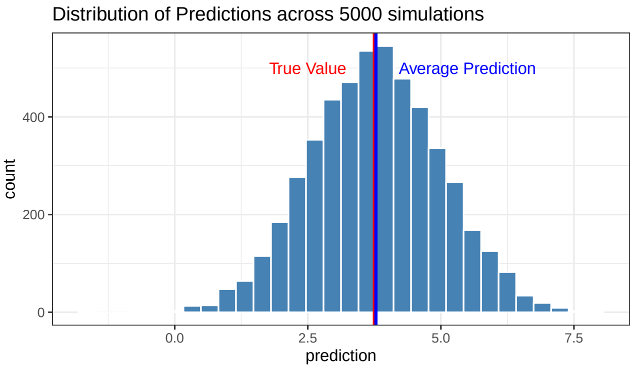
$$Y = -0.5 + 2.8X_1 + 5.8X_2 = -0.5 + 2.8 \cdot 0.25 + 5.8 \cdot 0.5 = 3.1$$

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

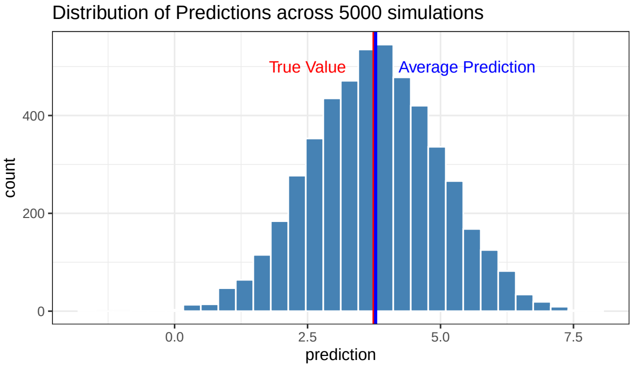
Simulation

```
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

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

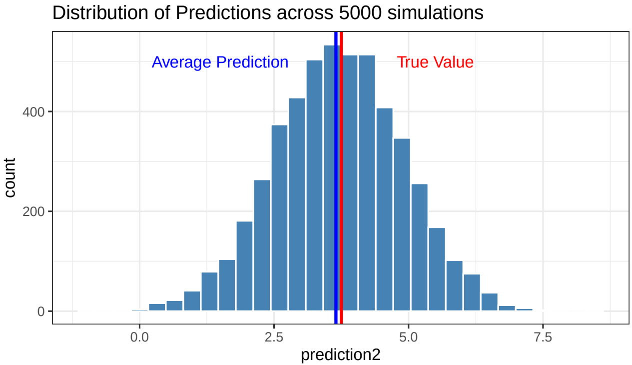
- And the prediction when $X_1 = 0.25$ and $X_2 = 0.5$ is

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

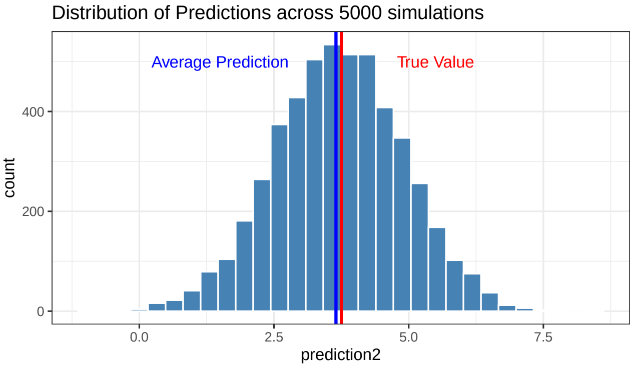
Simulation II

```r
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!

Section 2

Ridge Regression

Shrinkage Penalty

- There are some situations in which multiple linear regression has high MSE:

Shrinkage Penalty

- There are some situations in which multiple linear regression has high MSE:
  - Predictors are strongly correlated (high variance)
  - Many predictors relative to data size (high variance)
  - Model form is non-linear (high bias)

Shrinkage Penalty

- There are some situations in which multiple linear regression has high MSE:
  - Predictors are strongly correlated (high variance)
  - Many predictors relative to data size (high variance)
  - Model form is non-linear (high bias)
- To improve models in the first two cases, we reduce MSE by reducing variance at the cost slight increase in bias.

Shrinkage Penalty

- There are some situations in which multiple linear regression has high MSE:
  - Predictors are strongly correlated (high variance)
  - Many predictors relative to data size (high variance)
  - Model form is non-linear (high bias)
- To improve models in the first two cases, we reduce MSE by reducing variance at the cost slight increase in bias.
- In the presence of multicollinearity or over-fitting, least squares estimates tend to be too large.

Shrinkage Penalty

- There are some situations in which multiple linear regression has high MSE:
    - Predictors are strongly correlated (high variance)
    - Many predictors relative to data size (high variance)
    - Model form is non-linear (high bias)
- To improve models in the first two cases, we reduce MSE by reducing variance at the cost slight increase in bias.
- In the presence of multicollinearity or over-fitting, least squares estimates tend to be too large.
- To build a better model, we reduce the size of coefficients relative to least squares regression.

Ridge Regression

- Recall that least squares regression estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for
$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$
are obtained by finding the values of β that minimize
$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge Regression

- Recall that least squares regression estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for
$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$
are obtained by finding the values of β that minimize
$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$
- To perform **Ridge Regression**, we instead find coefficients β that minimize
$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 \qquad \text{where } \lambda \geq 0 \text{ is tuning parameter}$$

Ridge Regression

- Recall that least squares regression estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for
$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$
are obtained by finding the values of β that minimize
$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$
- To perform **Ridge Regression**, we instead find coefficients β that minimize
$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 \qquad \text{where } \lambda \geq 0 \text{ is tuning parameter}$$

Why?

Ridge Regression

- Recall that least squares regression estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ for
$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$
are obtained by finding the values of β that minimize
$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$
- To perform **Ridge Regression**, we instead find coefficients β that minimize
$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 \qquad \text{where } \lambda \geq 0 \text{ is tuning parameter}$$

Why?

- The term $\lambda \sum_{i=1}^{p} \beta_i^2$ is the **shrinkage penalty**, and is small when the β are small.

Ridge Regression

- Recall that least squares regression estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for

$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

are obtained by finding the values of β that minimize

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- To perform **Ridge Regression**, we instead find coefficients β that minimize

$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 \qquad \text{where } \lambda \geq 0 \text{ is tuning parameter}$$

Why?

- The term $\lambda \sum_{i=1}^{p} \beta_i^2$ is the **shrinkage penalty**, and is small when the β are small.
- With a shrinkage penalty, the algorithm prefers models with lower coefficients.

Ridge Regression

- Recall that least squares regression estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for

$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

  are obtained by finding the values of β that minimize

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- To perform **Ridge Regression**, we instead find coefficients β that minimize

$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 \qquad \text{where } \lambda \geq 0 \text{ is tuning parameter}$$

Why?

- The term $\lambda \sum_{i=1}^{p} \beta_i^2$ is the **shrinkage penalty**, and is small when the β are small.
- With a shrinkage penalty, the algorithm prefers models with lower coefficients.
- This tends to reduce variance, at the cost of increased bias.

Effects of the Tuning Parameter

- **Goal:** Find β which minimize $\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2$

Effects of the Tuning Parameter

- **Goal:** Find β which minimize $\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2$
- What will happen to β_i as $\lambda \to \infty$? As $\lambda \to 0$?

Effects of the Tuning Parameter

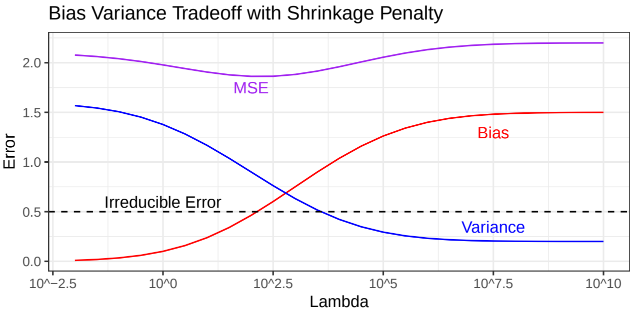
- **Goal:** Find β which minimize $\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2$
- What will happen to β_i as $\lambda \to \infty$? As $\lambda \to 0$?
- What will happen to β_0 as $\lambda \to \infty$? As $\lambda \to 0$?

Effects of the Tuning Parameter

- **Goal:** Find β which minimize $\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2$
- What will happen to β_i as $\lambda \to \infty$? As $\lambda \to 0$?
- What will happen to β_0 as $\lambda \to \infty$? As $\lambda \to 0$?
- What happens to MSE as $\lambda \to 0$ or $\lambda \to \infty$?

Effects of the Tuning Parameter

- **Goal:** Find β which minimize $\text{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2$
- What will happen to β_i as $\lambda \to \infty$? As $\lambda \to 0$?
- What will happen to β_0 as $\lambda \to \infty$? As $\lambda \to 0$?
- What happens to MSE as $\lambda \to 0$ or $\lambda \to \infty$?



Simulation

- Consider a linear model with 9 predictors and 100 observations.

$$y = 10 + 1x_1 + 2x_2 \cdots + 8x_8 + 9x_9 + \epsilon \quad \epsilon \sim N(0, 4)$$

Simulation

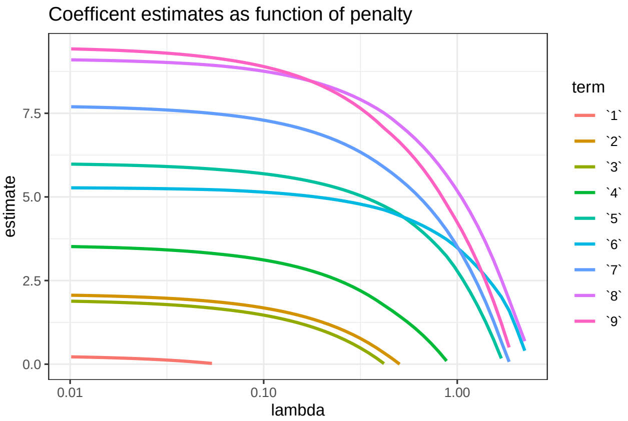
- Consider a linear model with 9 predictors and 100 observations.

$$y = 10 + 1x_1 + 2x_2 \cdots + 8x_8 + 9x_9 + \epsilon \quad \epsilon \sim N(0, 4)$$

```
##
## Call:
## lm(formula = y ~ ., data = sim_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5148 -1.5155 -0.0932  1.8054  5.1007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6034     1.3023   0.463   0.6443
## `1`           0.2653     0.8831   0.300   0.7645
## `2`           2.1047     0.8005   2.629   0.0101 *
## `3`           1.9316     0.7766   2.487   0.0147 *
## `4`           3.5635     0.8133   4.382 3.18e-05 ***
## `5`           6.0143     0.7925   7.589 2.84e-11 ***
## `6`           5.2844     0.7810   6.766 1.30e-09 ***
## `7`           7.7421     0.8657   8.944 4.51e-14 ***
## `8`           9.1352     0.7466  12.236  < 2e-16 ***
## `9`           9.4859     0.8046  11.789  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.244 on 90 degrees of freedom
## Multiple R-squared:  0.8437, Adjusted R-squared:  0.828
## F-statistic: 53.97 on 9 and 90 DF,  p-value: < 2.2e-16
```

Simulation

- What happens to the size of coefficients as λ gets larger?



Effect of Scale

- Suppose $\hat{y} = 1 + 0.01x_1 + 20x_2$ is the best fitting linear model for Y using X_1 and X_2, and that both are statistically significant.

Effect of Scale

- Suppose $\hat{y} = 1 + 0.01x_1 + 20x_2$ is the best fitting linear model for Y using X_1 and X_2, and that both are statistically significant.
  - Are we justified in saying that X_2 is a more important predictor than X_1?

Effect of Scale

- Suppose $\hat{y} = 1 + 0.01x_1 + 20x_2$ is the best fitting linear model for Y using X_1 and X_2, and that both are statistically significant.
  - Are we justified in saying that X_2 is a more important predictor than X_1?
  - What if $\mathrm{sd}(x_1) = 10000$ and $\mathrm{sd}(x_2) = .1$?
- Suppose we first standardize X_1 and X_2 by subtracting off their means and dividing by their standard deviations:

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \qquad Z_2 = \frac{X_2 - \mu_2}{\sigma_2}$$

Effect of Scale

- Suppose $\hat{y} = 1 + 0.01x_1 + 20x_2$ is the best fitting linear model for Y using X_1 and X_2, and that both are statistically significant.
  - Are we justified in saying that X_2 is a more important predictor than X_1?
  - What if $\mathrm{sd}(x_1) = 10000$ and $\mathrm{sd}(x_2) = .1$?
- Suppose we first standardize X_1 and X_2 by subtracting off their means and dividing by their standard deviations:

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \qquad Z_2 = \frac{X_2 - \mu_2}{\sigma_2}$$

- If we build a model and find $\hat{y} = 1 + 0.01z_1 + 20z_2$, where Z_1 and Z_2 are standardized, are we now justified in saying that Z_2 is more important than Z_1?

Effect of Scale

- Suppose $\hat{y} = 1 + 0.01x_1 + 20x_2$ is the best fitting linear model for Y using X_1 and X_2, and that both are statistically significant.
    - Are we justified in saying that X_2 is a more important predictor than X_1?
    - What if $\mathrm{sd}(x_1) = 10000$ and $\mathrm{sd}(x_2) = .1$?
- Suppose we first standardize X_1 and X_2 by subtracting off their means and dividing by their standard deviations:

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \qquad Z_2 = \frac{X_2 - \mu_2}{\sigma_2}$$

- If we build a model and find $\hat{y} = 1 + 0.01z_1 + 20z_2$, where Z_1 and Z_2 are standardized, are we now justified in saying that Z_2 is more important than Z_1?
    - Assuming both are statistically significant, we are probably justified.

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**
  - That is, scaling a predictor by a value c just leads to scaling the estimate by $1/c$.

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**
  - That is, scaling a predictor by a value c just leads to scaling the estimate by $1/c$.
  - The predicted value is the same, regardless of scale.

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**
  - That is, scaling a predictor by a value c just leads to scaling the estimate by $1/c$.
  - The predicted value is the same, regardless of scale.
  - Therefore, rescaling predictors *does not* change the fit of the model (RSS is the same)

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**
  - That is, scaling a predictor by a value c just leads to scaling the estimate by $1/c$.
  - The predicted value is the same, regardless of scale.
  - Therefore, rescaling predictors *does not* change the fit of the model (RSS is the same)
  - Suppose $y = 1 + 0.01x_1 + 20x_2$, $\sigma_1 = 10000, \sigma_2 = 0.1$, and both x_1, x_2 have mean 0.
  - After rescaling, $z_1 = \frac{x_1}{10000}, z_2 = \frac{x_2}{0.1}$ and the linear model is
$$y = 100z_1 + 2z_2$$
- However, for Ridge Regression, coefficient estimates can change depending on scale.

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**
  - That is, scaling a predictor by a value c just leads to scaling the estimate by $1/c$.
  - The predicted value is the same, regardless of scale.
  - Therefore, rescaling predictors *does not* change the fit of the model (RSS is the same)
  - Suppose $y = 1 + 0.01x_1 + 20x_2$, $\sigma_1 = 10000, \sigma_2 = 0.1$, and both x_1, x_2 have mean 0.
  - After rescaling, $z_1 = \frac{x_1}{10000}, z_2 = \frac{x_2}{0.1}$ and the linear model is

$$y = 100z_1 + 2z_2$$

- However, for Ridge Regression, coefficient estimates can change depending on scale.
  - Recall the shrinkage penalty is $\lambda \sum_{i=1}^{2} \beta_i^2 = \lambda(0.01^2 + 20^2)$

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**
  - That is, scaling a predictor by a value c just leads to scaling the estimate by $1/c$.
  - The predicted value is the same, regardless of scale.
  - Therefore, rescaling predictors *does not* change the fit of the model (RSS is the same)
  - Suppose $y = 1 + 0.01x_1 + 20x_2$, $\sigma_1 = 10000, \sigma_2 = 0.1$, and both x_1, x_2 have mean 0.
  - After rescaling, $z_1 = \frac{x_1}{10000}, z_2 = \frac{x_2}{0.1}$ and the linear model is

$$y = 100z_1 + 2z_2$$

- However, for Ridge Regression, coefficient estimates can change depending on scale.
  - Recall the shrinkage penalty is $\lambda \sum_{i=1}^{2} \beta_i^2 = \lambda(0.01^2 + 20^2)$
  - Which models will ridge regression favor?

Scale

- The coefficients in the least squares regression equation are **scale-equivalent**
  - That is, scaling a predictor by a value c just leads to scaling the estimate by $1/c$.
  - The predicted value is the same, regardless of scale.
  - Therefore, rescaling predictors *does not* change the fit of the model (RSS is the same)
  - Suppose $y = 1 + 0.01x_1 + 20x_2$, $\sigma_1 = 10000, \sigma_2 = 0.1$, and both x_1, x_2 have mean 0.
  - After rescaling, $z_1 = \frac{x_1}{10000}, z_2 = \frac{x_2}{0.1}$ and the linear model is
$$y = 100z_1 + 2z_2$$
- However, for Ridge Regression, coefficient estimates can change depending on scale.
  - Recall the shrinkage penalty is $\lambda \sum_{i=1}^{2} \beta_i^2 = \lambda(0.01^2 + 20^2)$
  - Which models will ridge regression favor?
- Ridge regression is most effective if predictors are standardized first.