Multiple Linear Regression

The Hat Matrix and Ridge Regression

Math 392

Preamble

In the last lecture, we noted that the least squares estimate, $\hat{\beta} = (X'X)^{-1}X'Y$, is unique only if the matrix $X'X$ is invertible.

Example: invert a 2 by 3 matrix.

```
X <- matrix(c(1, 1, 1, 1, 3, 5),
            byrow = TRUE, nrow = 2)
X
```

```
##      [,1] [,2] [,3]
## [1,]    1    1    1
## [2,]    1    3    5
```

```
solve(t(X) %*% X)
```

```
## Error in solve.default(t(X) %*% X): system is computationally si
```

What if we transpose X so that it is tall rather than wide?

```
X <- t(X)
X
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    3
## [3,]    1    5
```

```
solve(t(X) %*% X)
```

```
##           [,1]   [,2]
## [1,]  1.458333 -0.375
## [2,] -0.375000  0.125
```

A Very Interesting Matrix: H

Recall:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$$

The "hat matrix", H,

$$H = X(X'X)^{-1}X'$$

because it puts a hat on the Y.

Properties of H

H is *symmetric*.

$$\begin{aligned} H' &= X(X'X)^{-1}X' \\ &= X''[(X'X)^{-1}]'X' \\ &= X[(X'X)']^{-1}X' \\ &= X(X'X)^{-1}X' \\ &= H \end{aligned}$$

H is *idempotent*.

$$\begin{aligned} H^2 &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' \\ &= H \end{aligned}$$

We can easily express the vector of *residuals*, $\hat{\epsilon}$.

$$\hat{\epsilon} = Y - \hat{Y} = Y - HY = (I - H)Y$$

Leverage

Let $h_{i,j}$ denote the $(i,j)^{th}$ element of H. Then we can express the fitted value of the i^{th} observation as

$$\hat{Y}_i = h_{i,i}Y_i + \sum_{j \neq i} h_{i,j}Y_j$$

Leverage, cont.

In simple linear regression,

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ &= \end{aligned}$$

For the i^{th} observation:

$$h_{i,i} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_X}$$

High leverage:

$$h_{i,i} > 2\,\mathrm{avg}(h_{i,i}) = 2\frac{p}{n}$$

Exercises

Using the hat matrix and its properties, find $E(\hat{\epsilon})$ and $Var(\hat{\epsilon})$.

$$
\begin{aligned}
E(\hat{\epsilon}|X) &= E((I-H)Y|X) \\
&= (I-H)E(Y|X) \\
&= (I-H)X\beta \\
&= x\beta - X(X'X)^{-1}X'X\beta \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\epsilon}|X) &= Var((I-H)Y) \\
&= (I-H)Var(Y)(I-H)' \\
&= (I-H)\sigma^2(I-H)' \\
&= \sigma^2(II' - HI' - IH' + HH') \\
&= \sigma^2(I - H - H + H) \\
&= \sigma^2(I-H)
\end{aligned}
$$

Ridge Regression

Consider an alternative estimator for β.

$$\hat{\beta}_{ridge} = \text{argmin}(RSS(\beta)) \quad \text{subject to} \quad c \geq \sum_{j=1}^{p-1} \beta_j^2; \quad c \geq 0$$

This is equivalent to minimizing the penalized RSS (for the scalar λ):

$$\begin{aligned} PRSS(\beta) &= (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta \\ &= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta + \lambda\beta'\beta \\ &= Y'Y - 2\beta'X'Y. \end{aligned}$$

To find the $\hat{\beta}_{ridge}$ that minimize this function, we take the derivative with respect to β, set to zero, and solve.

$$\begin{aligned} \frac{\partial\, PRSS}{\partial\, \beta} &= 0 - 2X'Y + 2X'X\beta + 2\lambda\beta = 0 \\ X'Y &= (X'X + \lambda I)\beta \\ \hat{\beta}_{ridge} &= (X'X + \lambda I)^{-1}X'Y \end{aligned}$$

Preamble revisited

Let's apply this technique to solve the problem of invertibility that we encountered in the preamble.

```r
X <- matrix(c(1, 1, 1, 1, 3, 5), byrow = TRUE, nrow = 2)
X
```

```
##      [,1] [,2] [,3]
## [1,]    1    1    1
## [2,]    1    3    5
```

```r
I <- diag(ncol(X))
I
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
```

```r
lambda <- .5
solve(t(X) %*% X + lambda * I)
```

```
##             [,1]       [,2]        [,3]
## [1,]  1.02890173 -0.4624277  0.04624277
## [2,] -0.46242775  1.3988439 -0.73988439
## [3,]  0.04624277 -0.7398844  0.47398844
```

Ridge on a tall matrix

```
X <- t(X)
X
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    3
## [3,]    1    5
```

```
I <- diag(ncol(X))
I
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

```r
solve(t(X) %*% X + lambda * I)
```

```
##             [,1]        [,2]
## [1,]  0.8208092 -0.20809249
## [2,] -0.2080925  0.08092486
```

Ridge on a tall matrix, cont.

```
Y <- c(3, 5, 6)
# Ridge
solve(t(X) %*% X + lambda * I) %*% t(X) %*% Y
```

```
##           [,1]
## [1,] 1.5028902
## [2,] 0.9710983
```

```
# OLS
solve(t(X) %*% X) %*% t(X) %*% Y
```

```
##          [,1]
## [1,] 2.416667
## [2,] 0.750000
```

What's wrong?

Excluding the intercept

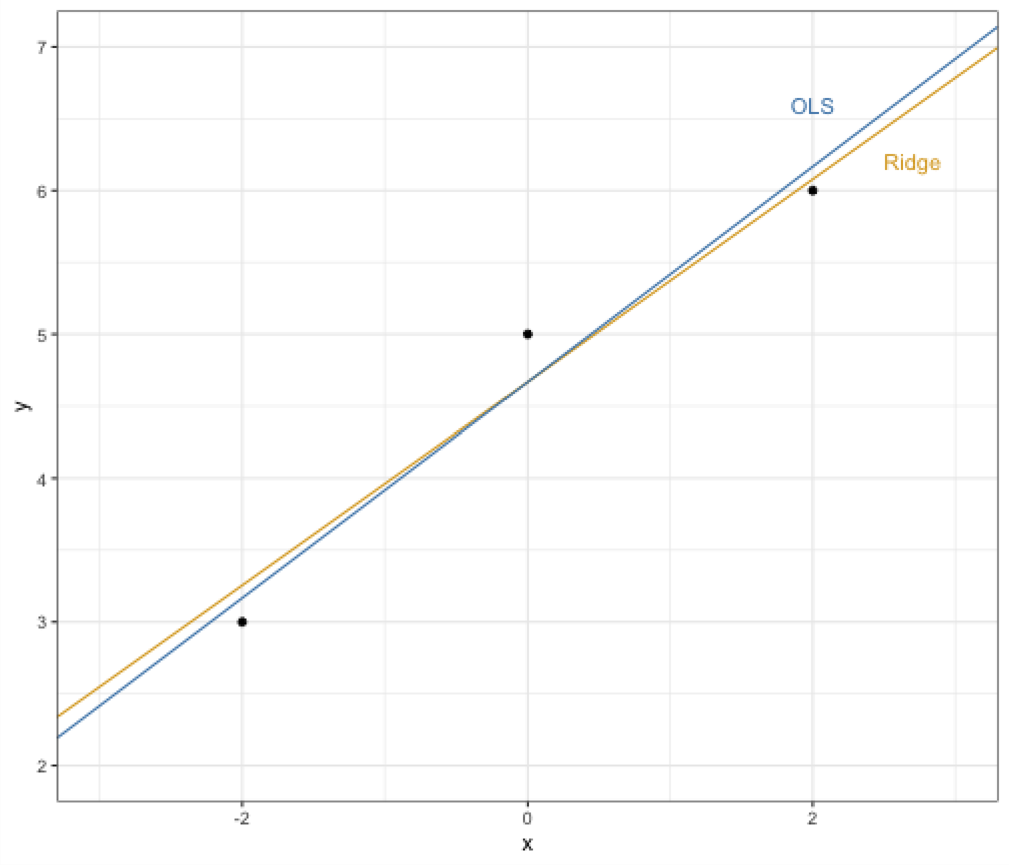
```
X_no_int <- X[, -1]
X_no_int <- X_no_int - mean(X_no_int)
I <- 1
# Ridge
solve(t(X_no_int) %*% X_no_int + lambda * I) %*%
  t(X_no_int) %*% Y
```

```
##           [,1]
## [1,] 0.7058824
```

```
# OLS
solve(t(X_no_int) %*% X_no_int) %*% t(X_no_int) %*% Y
```

```
##      [,1]
## [1,] 0.75
```

Ridge vs OLS



Properties of $\hat{\beta}_{ridge}$

$$
\begin{aligned}
E(\hat{\beta}_{ridge}|X) &= E((X'X + \lambda I)^{-1}X'Y|X) \\
&= (X'X + \lambda I)^{-1}X'E(Y|X) \\
&= (X'X + \lambda I)^{-1}X'X\beta
\end{aligned}
$$

Therefore the ridge estimates are biased for any $\lambda \neq 0$.

$$
\begin{aligned}
Var(\hat{\beta}_{ridge}|X) &= Var((X'X + \lambda I)^{-1}X'Y|X) \\
&= (X'X + \lambda I)^{-1}X'Var(Y|X)[(X'X + \lambda I)^{-1}X']' \\
&= (X'X + \lambda I)^{-1}X'\sigma^2 I[(X'X + \lambda I)^{-1}X']' \\
&= \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}
\end{aligned}
$$

Note that as $\lambda \to \infty$, $Var(\hat{\beta}_{ridge}) \to 0$.