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. Recall that matrices cannot be inverted if one of the columns can be expressed as a linear combination of the others. We run into this issue when working with a design matrix X that has more columns than rows.

An an example, observe what happens when we try to 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 singular: reciprocal condition number =
```

If we *could* have come up with least squares estimates in this setting, that would correspond to having fit a plane to two points in 3D space. There are in fact many planes that would achieve the same RSS, so our solution would not be unique.

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
```

This allows us to proceed and will lead to a unique $\hat{\beta}$ because the new $(X'X)^{-1}$ matrix is full rank. This corresponds to fitting a 3D plane through three observations in 3D: a much more reasonable task.

A Very Interesting Matrix: H

Recall:

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

The matrix that we pre-multiply Y by, which is purely a function of the matrix X, has useful properties, so it was given a name, the "hat matrix", H,

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

because it puts a hat on the Y.

Properties of H

H is *symmetric*.

$$\begin{aligned} H' &= [X(X'X)^{-1}X']' && (1)\\ &= X''[(X'X)^{-1}]'X' && (2)\\ &= X[(X'X)']^{-1}X' && (3)\\ &= X(X'X)^{-1}X' = H && (4) \end{aligned}$$

H is *idempotent*.

$$\begin{aligned} H^2 &= X(X'X)^{-1}X'X(X'X)^{-1}X' && (5)\\ &= X(X'X)^{-1}X' = H && (6) \end{aligned}$$

The hat matrix, which contains all of the information in the X needed to find the fitted values, also allows us to 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

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

This demonstrates that the each fitted value can be expressed as a weighted sum of the elements Y. The entries of the hat matrix prescribe what those weights should be. The diagonal of H, which contains the $h_{i,i}$ captures the degree to which the i^{th} observation is able to draw the fitted value (and therefore the line) to itself. These elements are called the *leverages*.

In simple linear regression,

$$H = X(X'X)^{-1}X' \tag{7}$$

$$= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \tag{8}$$

After some tedious matrix algebra, you arrive at the following expression for the leverage:

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

This shows that observations that are farther from the mean will have a fitted $\hat{Y}_i$ very close to the observed Y_i. Under certain assumptions regarding the distribution of the X (and therefore the distribution of the leverages), the a rule of thumb was developed that if

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

Exercises

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

$$E(\hat{\epsilon}|X) = E((I-H)Y|X) \tag{9}$$
$$= (I-H)E(Y|X) \tag{10}$$
$$= (I-H)X\beta \tag{11}$$
$$= x\beta - X(X'X)^{-1}X'X\beta \tag{12}$$
$$= 0 \tag{13}$$

$$Var(\hat{\epsilon}|X) = Var((I-H)Y) \tag{14}$$
$$= (I-H)Var(Y)(I-H)' \tag{15}$$
$$= (I-H)\sigma^2(I-H)' \tag{16}$$
$$= \sigma^2(II' - HI' - IH' + HH') \tag{17}$$
$$= \sigma^2(I - H - H + H) \tag{18}$$
$$= \sigma^2(I-H) \tag{19}$$

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):

$$PRSS(\beta) = (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta \tag{20}$$
$$= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta + \lambda\beta'\beta \tag{21}$$
$$= Y'Y - 2\beta'X'Y + \lambda\beta'\beta + \beta'X'X\beta. \tag{22}$$

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

$$\frac{\partial\, PRSS}{\partial\, \beta} = 0 - 2X'Y + 2X'X\beta + 2\lambda\beta = 0 \tag{23}$$
$$X'Y = (X'X + \lambda I)\beta \tag{24}$$
$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'Y \tag{25}$$

This approach to estimating β was originally devised as a fix for situations when $X'X$ was non-invertible/singular due to multicollinearity of the predictors.

Preamble revisited

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

```
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
```

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

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

```
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
```

Adding the positive elements down the diagonal succeeds in making $X'X$ invertible; this allows us to find a unique estimate of β. Thinking geometrically, we're still aiming to fit a plane through 2 observations in three dimensional space, but now we have an additional element in our loss function that prefers some orientations over others. Specifically, the orientations where the total squared magnitude of the coefficients is as small as possible, that is, plane closest to the horizontal.

To see this effect, we can use ridge regression to estimate the coefficients when we're working with the transpose of X as our design 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
```

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

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

We continue to find our estimates $\hat{\beta}_{ridge}$ and compare them to $\hat{\beta}_{OLS}$ (note that this requires a vector Y).

```
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
```

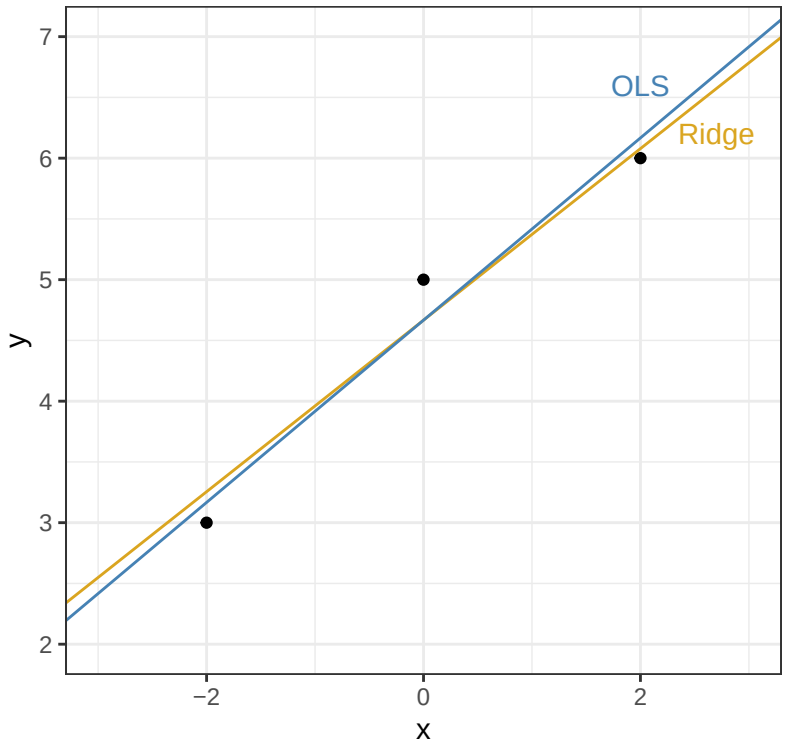
This is an unexpected result - the ridge estimate of β_1 is actually *higher* than the OLS estimate. From a practical standpoint, it doesn't make sense to include β_0 in the penalty because it simply accounts for the magnitude of the Y when the X takes values of zero. Therefore it is standard to perform the ridge regression without an intercept, but after subtracing the column means from each column of X. This allows shrinkage to perform as expected and leaves $\hat{\beta}_0 = \bar{y}$.

```
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
```



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

$$
\begin{aligned}
E(\hat{\beta}_{ridge}|X) &= E((X'X + \lambda I)^{-1}X'Y|X) && (26)\\
&= (X'X + \lambda I)^{-1}X'E(Y|X) && (27)\\
&= (X'X + \lambda I)^{-1}X'X\beta && (28)
\end{aligned}
$$

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

$$
\begin{aligned}
Var(\hat{\beta}_{ridge}|X) &= Var((X'X + \lambda I)^{-1}X'Y|X) && (29)\\
&= (X'X + \lambda I)^{-1}X'Var(Y|X)[(X'X + \lambda I)^{-1}X']' && (30)\\
&= (X'X + \lambda I)^{-1}X'\sigma^2 I[(X'X + \lambda I)^{-1}X']' && (31)\\
&= \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1} && (32)
\end{aligned}
$$

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