Math 455 Class 10 February 11

Last time: prediction [Ch. 4] vs.

Explanation [Ch. 5]

Prediction means we have data points.

$(Y_i, X_{1i}, X_{2i}, \dots X_{p-1,i})$ $i=1, \dots, n$

response predictors.

We estimate $\beta_0, \dots, \beta_{p-1}$ in the model

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + \varepsilon$$

We then get new data point $i = n+1$.

We try to estimate Y_{n+1} by.

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{1,n+1} + \dots + \hat{\beta}_{p-1} X_{p-1,n+1}.$$

Why might we try to estimate Y_{n+1} when we already have the data point

$(Y_{n+1}, X_{1,n+1}, \dots, X_{p-1,n+1})$?

In real-life situations, we ~~would~~ may have

ONLY $(X_{1,n+1}, \dots X_{p-1,n+1})$ and NOT Y_{n+1}

Two cases of prediction: [Faraway p. 51-52]

(1) prediction of a single response (given predictors)

(2) prediction of the average response (many responses with exactly the same predictor values).

Also discussed in Wackerly, Sections 11.7 & 11.13.

Note that our "point estimate" (single best guess) is the same in cases (1) & (2).

The difference is the uncertainty (standard error).

Conceptually: Why the difference?

Suppose $Y_1, \dots, Y_n$ i.i.d. with mean μ, var σ^2.

What are the mean & variance of $\bar{Y}$?

$$\left[\bar{Y} = \frac{1}{n}(Y_1 + \dots + Y_n)\right] \qquad \mu \qquad \frac{\sigma^2}{n}.$$

Even if we are averaging $Y_{n+1}, \ldots, Y_{n+100}$

and trying to predict the average, uncertainty

in $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$ does not go away.

There is a matrix $X^T X$ that keeps showing up.

For example: in the "hat matrix" $H = X(X^T X)^{-1} X^T$.

our estimated $\hat{Y}$ is HY. $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Or in Wackerly p. 616: $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$

where $(c_{ij}) = (X^T X)^{-1}$.

What is this $(X^T X)$ matrix?

Let's consider the 1-variable case. $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$.

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$= \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}$$

What about 2 variables?

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ \vdots & x_{12} & x_{22} \\ \vdots & & \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$$

$$X^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & & x_{1n} \\ x_{21} & x_{22} & & x_{2n} \end{pmatrix}$$

$$X^T X = \begin{pmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{pmatrix}$$

This reminds us of the variance-covariance matrix: The variance-covariance matrix of (X_1, X_2) is

$$\begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{pmatrix}$$

It is not precisely the variance-covariance matrix because we haven't subtracted means.

$$Cov(X_1, X_2) = \frac{1}{n} \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2).$$

(or divided by n).

If we compute $X^T Y$ we are computing something very close to the ~~vari~~ covariances of each X_i with the response Y.

The upshot of this: almost everything in the regression output can be recovered from the ~~cov~~ variance-covariance matrix of Y and all the predictors, plus the means of Y and all the predictors plus n. (the number of data points.)