Math 455 Class 6 February 2

Deadline for HW 2 Extended until tonight.

Reminder: handwritten and scanned pdf is OK until we start doing R-based homework from the textbook of Faraway. ~~This~~ HW 3 will also be from Wackerly, a review of hypothesis testing and p-values.

The point of HW 2: in regression, we get an output with coefficient estimates, e.g. $\hat{\beta}_0, \hat{\beta}_1$. and p-values associated to these.

A p-value is defined in the context of a statistical test. The test is based on a statistic, and in order to compute the <u>level</u> of a test (which goes into the def'n of p-value) we need to know the distribution of the statistic.

How would we know that in regression?

We have the standard assumptions of regression.

Which say that the errors $\varepsilon_i$ are iid normal $N(0, \sigma^2)$. The "statistics" are cooked up out of these errors and therefore have the t- or F-distributions by the type of arguments you used in HW 2.

To see these arguments in more detail, see Wackerly 11.4, and sections following.

Example: $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2}$$

The $X_i$ are not random, so $E[\hat{\beta}_1]$ cane be written:

$$E[\hat{\beta}_1] = \frac{1}{\sum (X_i - \bar{X})^2} \sum (X_i - \bar{X}) E[Y_i]$$

But $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (model assumption)

and Thus $E[Y_i] = \beta_0 + \beta_1 X_i + E[\varepsilon_i]$

Nowe, ~~see~~ using the null hypothesis ($\beta_1 = 0$) and model assumption $E[\varepsilon_i] = 0$.

This simplifies further:

(This is the sort of thing that goes on in 11.4.)

$$E[\hat{\beta}_1] = \frac{1}{\sum (x_i - \bar{x})^2} \underbrace{\cancel{\sum (x_i - \bar{x})}}_{\to 0} \beta_0$$

$$+ \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) \beta_1 x_i$$

$$+ \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) \underbrace{\cancel{E[\varepsilon_i]}}_{\to 0}$$

$$E[\hat{\beta}_1] = \frac{1}{\sum (x_i - \bar{x})^2} \beta_1 \cdot \sum (x_i - \bar{x}) x_i$$

$$= \beta_1 \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x})(x_i - \bar{x}) = \beta_1$$

(This proves that $\hat{\beta}_1$ is an unbiased estimator of the regression parameter $\beta_1$).

---

What does least-squares try to do?

Minimize the sum of squared residuals.

Let's imagine that we have, for a particular choice of $\beta_0, \beta_1$. 5 residuals:

$$1, -1, 3, 2, 100$$

What is our sum of squares?

$$10{,}000 + \text{small numbers.}$$

Imagine that we could reduce the 100 to 90 at the cost of increasing the other residuals by 10 each. Then our sum of squares is

$$11, 9, 13, 12, 90$$

$$SSR = 11^2 + 9^2 + 13^2 + 12^2 + 8{,}100.$$

Notice that this is smaller. We reduced 10,000 by 1,900 and the increases in the others are swamped by the $-1900$.

This means: minimizing sum of squares means we really want to reduce BIG errors.

Points that produce a BIG error will have a BIG influence on the regression line.

Another way to say this: least-squares is sensitive to outliers.

There are 4 standard graphs associated to a linear regression (obtained with plot(lmod) in R). One of the purposes of these plots is to see if this outlier effect is happening.

[Another purpose is to check if the 4 assumptions are satisfied.]

---

More about linear regression as orthogonal projection. (From Faraway p. 33)

If we have a response $y$ and predictor variables $x_1, \dots, x_p$ and we are trying to estimate $\beta_0, \dots, \beta_p$ in the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

We think of this as a vector equation.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} \cdots + \beta_p \begin{pmatrix} x_{1p} \\ \vdots \\ x_{np} \end{pmatrix} + \varepsilon$$

$$y = X\beta + \varepsilon \quad \text{where}$$

$$X = \begin{bmatrix} 1 & x_{11} & & x_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & & x_{np} \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

We can't solve $y = X\beta$ (with 0 error)

We look for the "best" (least-squares) $\hat{\beta}$,

such that $\hat{y} = X\hat{\beta}$ is the orthogonal projection of $y$ on the column space of $X$.

The solution is $\hat{y} = X(X^T X)^{-1} X^T y$.

and $\hat{\beta} = (X^T X)^{-1} X^T y$.

$H = X(X^T X)^{-1} X^T$ is called the "hat matrix."

and is the orthogonal projection of $\mathbb{R}^n$ (where $y$ lives) onto the column space of $X$.

Q: Why is it called the "hat matrix"?

A: $Hy = \hat{y}$, so $H$ "puts a hat" on $y$.

Q: Why do we care about the hat matrix?

A: Remember we want to check for outliers/influential points. One way of doing that is to look at entries $h_{ii}$ in the hat matrix. (There are also other reasons.) (More in Ch. 6).

---

Q: Why is this $X(X^T X)^{-1} X^T$ the formula for orthogonal projection? Can't this be made simpler somehow? For example, Why cant we write:

$$X(X^T X)^{-1} X^T = X X^{-1} (X^T)^{-1} X^T$$

$$= I \cdot I = I \,.?$$

A: $X$ is an $n \times (p+1)$ matrix. It is not square. There is no $X^{-1}$.

Q: Why can we assume $(X^T X)^{-1}$ exists?

A: If it doesn't this reflects linear dependence in the columns of $X$. We exclude such cases from our treatment of regression.

Techniques to detect this: 7.3 Collinearity.

Q: OK, why is the formula $H = X(X^T X)^{-1} X^T$?

A: $y = \hat{y} + e$ where the error is perp. to the column space of $X$.

This means $X^T e = 0$.

We want $\hat{y} = X\hat{\beta}$ $\qquad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$.

$$e = y - \hat{y} = y - X\hat{\beta}$$

and $$X^T e = X^T(y - X\hat{\beta}) = 0$$

Thus $$X^T y = X^T X \hat{\beta}.$$

Assuming (as before) that $X^T X$ is invertible

$$\hat{\beta} = (X^T X)^{-1} X^T y. \quad \text{and} \quad \hat{y} = X\hat{\beta}$$

so $$\hat{y} = X(X^T X)^{-1} X^T y = Hy.$$

Q: If $e_1, \ldots, e_{p+1}$ is an orthonormal basis of the column space of $X$, we learned that the orthogonal projection of $y$ is

$$\hat{y} = (y \cdot e_1) e_1 + \cdots + (y \cdot e_{p+1}) e_{p+1}.$$

Why can't we use this simpler formula?

A: We have no guarantee that the columns of $X$ are orthonormal. Maybe if we were good with experimental design and got to choose all the columns of $X$, we could.

But in general we don't get to choose, e.g. the blood pressure / weight of our patients.