Math 455 Class 7 February 5

Decoding Regression Output

In the regression output there are p-values

p-values are defined in the context of a statistical test.

A statistical test has a null hypothesis H_0 and an alternative hypothesis H_a.

The p-values in the $Pr(>|t|)$ column are ~~the~~ for tests of

$$H_0: \beta_i = 0$$

$$\text{vs. } H_a: \beta_i \neq 0$$

The p-value associated to the F-test.

This test has

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

(In other words all your predictors are useless.)

$$H_a: \text{at least one } \beta_i \text{ is not } 0$$

(At least one explanatory variable has something to do with Y)

Note that the F-test is usually very weak in the sense that

"At least one X_i has some association with a change in Y" is a weak statement.

In the case of only one predictor the t-test (for that predictor) and the F-test have the same H_0, H_a.

The p-values will be the same and the F-stat will be $(t\text{-stat})^2$.

(This is only for the one-predictor case.)

See Faraway p. 3~~4~~9 for an example.

What if we want to test

$$H_0: \beta_1 = 0.5 \quad \text{v.s.} \quad H_a: \beta_1 \neq 0.5.$$

We can do this by finding $\dfrac{\hat{\beta}_1 - 0.5}{s.e.(\hat{\beta}_1)}$.

this (under H_0) has the t-dist ;

just like under $H_0: \beta_1 = 0$

$$\frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)}$$ has the t-dist.

Again see Faraway p. 39.

Outliers in Regression Analysis:

Suppose our residuals are $e_1, \dots, e_5 = 1, 1, 1, 1, 100$

What should we be doing to minimize

$$SSR = 1^2 + 1^2 + 1^2 + 1^2 + 100^2 .$$

100 is by far the most important residual.

if $100 \searrow 90$ $\qquad 100^2 \searrow 8100.$ $\qquad \downarrow 1900$

if $1 \nearrow 11$ $\qquad$ ~~A~~ $1^2 \nearrow 121$ $\qquad \uparrow 120$

Conclusion: regression is strongly influenced by outliers.

More discussion of this in Ch. 6.

There are standard measures of this and graphs automatically provided to check for such points.