Lecture 32.

Analysis of Variance Approach to Test significance of regression.

$\underbrace{\text{Analysis of Variance Approach}}_{\text{(ANOVA)}}$

- Assume a two-sided test $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$.
- Recall from last time that

$$SS_E = SS_T - SS_R$$

$$\sum_{i=1}^{n} (\underbrace{y_i - \hat{y}_i}_{e_i})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{= \hat{\beta}_1 S_{xy}}$$

Let

$$F_0 = \frac{MS_R}{MS_E} := \frac{SS_R/1}{SS_E/(n-2)} = \frac{SS_R}{\hat{\sigma}^2}$$

MS_R ← mean square (regression)

MS_E ← mean square (error)

Rmk: the $SS_R/1$ is weird, but we write it like this because more generally, a mean square is computed by dividing a sum of squares by the degrees of freedom.

Assuming $H_0: \beta_1 = 0$, F_0 follows a $F_{1,n-2}$ distribution

* This is a new distribution!! * "F-distribution with 1 deg. of freedom in the numerator, and n-2 degrees of freedom in the denominator"

Idea: F_0 is large implies that more of the variability in Y is explained by the regression which means more evidence of a linear relationship.

- Given an observation $F_0 = f_0$ (based on data) and given a significance level α, we should reject H_0 if $f_0 > \underbrace{f_{\alpha,1,n-2}}_{\text{find } \alpha \text{ in f-table, column 1, row } n-2.}$

Notice that

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \;\Rightarrow\; T_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} = \frac{SS_R}{SS_E/n-2} = F_0.$$

- So (two-sided) t-test and f-test are equivalent.
- (for 1-sided, use the t-test).

~~11.5, 11.6 later if time~~

11.7 Adequacy of the Regression Model.

- we have made many assumptions in order to apply the regression model.

- we assume

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where $\varepsilon = N(0,1)$

↳ i.e. we are assuming ε is normally distributed

↳ the model is actually linear (i.e. no higher order terms in x).

- How would we examine the normality of ε?
- Here are two techniques based on residual analysis.

In this context, we call the terms

$$e_i = y_i - \hat{y}_i \quad \text{(the "error" terms)}$$

residuals

① Probability plots. we expect the residuals to be uniformly distributed, so (Reminder!)

let

$$e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$$

be a reordering of the residuals in increasing order.

Let c_i be such that

$$P(Z \leq c_i) = \frac{i - 0.5}{n}.$$

Then the points $\{(e_{(i)}, c_i)\}$ should be roughly along a straight line. (use the fat pencil).

② Standardize the residuals:

- If we are correct in our assumption that ε is normally distributed with mean 0 and variance σ^2, then the sample $\{e_1, \ldots, e_n\}$ of residuals can be "standardized" to

$$d_i = \frac{e_i}{\hat{\sigma}^2}.$$

- we should find that 95% of the d_i's will be in the interval $(-2, 2)$ (since $P(-2 \leq Z \leq 2) \approx 0.95$.

- Another way to measure the adequacy is via $\underline{R^2}$ (the coefficient of determination).

- we define

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Since $SS_T = SS_R + SS_E$, $0 \le R^2 \le 1$.

- Having large R^2 is usually evidence of adequacy of the linear regression model.

- R^2 close to 1 means that the regression model explains much of the variability in the data.

- note that one should use R^2 cautiously.
  - ↳ R^2 can be artificially inflated in various ways.
  - ↳ <u>ex.</u> using higher order regression

technique: given any set of n points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ there is a polynomial of degree $(n-1)$ that goes through every point (and so gives $R^2 = 1$).

↳ this is an example of "overfitting" data, and usually is not useful/not good for predictions.