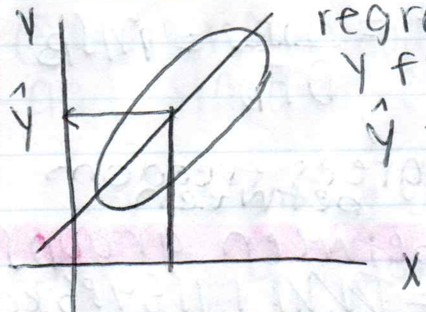
11/29/18
Wave Moretto

Lecture #18: ANOVA



regression line for predicting y from x

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{SE}(\hat{y})$ = uncertainty in using $\hat{y}$ to predict y

$\hat{SE}(\hat{y}) = s_{y|x}$ = RMSE = residual SD

L-261

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + e_i$

multiple linear regression (k=1 → simple linear regression)

1-way ANOVA

- Fisher (1925)
- L-275 case study
- Data values have 2 subscripts ex) y_{15} (group, trees)

- Basic model: $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma$
- null: $\mu_1 = \mu_2 = \dots = \mu_I = \mu$
- Alt: not so

- favor alt / reject null if distance between actual data and expected data (if null true) is large
- if null is true $\bar{y}_i$ would be close to overall $\bar{y}$

$$\frac{n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + \dots n_I(\bar{y}_I - \bar{y})^2}{n_1 + n_2 + \dots + n_I - 1}$$

$n_1(\bar{y}_1-\bar{y})^2 + \dots + n_I(\bar{y}_I-\bar{y})^2 = \sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y})^2 = SS_B$ sum of squares

$$\frac{1}{I-1}\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y})^2 = \frac{SS_B}{DF_B} = MS_B$$

degrees freedom between

MS_B = mean square between groups

$= \frac{3.346455 kg^2}{3} = 1.115485 kg^2$

- Can make SS_B bigger or smaller by changing units
- Need estimate of noise in kg 2

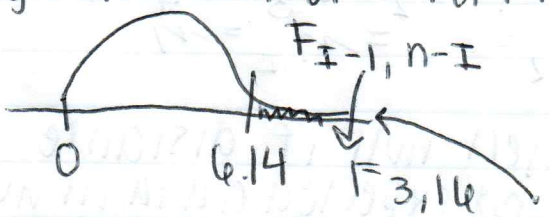
$s_1^2 \quad s_2^2 \quad \ldots \quad s_I^2$

$$\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_I-1)s_I^2}{(n_1-1) + \ldots + (n_I-1)}$$

$\hookrightarrow = n - I$

$\frac{\text{signal}}{\text{noise}} = \frac{MS_B}{MS_W} = F\text{ ratio} = 6.14127$ in this case

long run hist of F ratio if null true



stat sig!

$p = 0.0056 = 0.6\%$

reject null if $p \leq 5\%$ $\leftrightarrow$ stat sig ✓

- Which groups differ?
- Multiple comparisons: Many methods developed, look at simplest, Bonferroni for pairwise comparisons

① Decide # comparisons to make = k
② Decide level confidence = $100(1-\alpha)\%$

③ pairwise comparisons in form $(\bar{y}_i - \bar{y}_j)$ and standard error:

$$\hat{SE}\,(\bar{y}_i - \bar{y}_j) = \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $\hat{\sigma}$ = root mean squared error $= \sqrt{MS_W}$ ($\hat{\sigma}^2 = MS_W$ = pooled variance estimate)

~~④ ignore multiple comparisons prob~~

④ use a bigger t number to compensate for multiple comparisons being made (this is from Bonferroni) so make intervals wider!

$$t_{n-I}^{1-\frac{\alpha}{k}}$$

- always be given t number in problems bc given t values in table aren't enough