Lec 11

Administrative:

- Assignment 2 out now (due June 24 @ 11:59 PM).
- Midterm 1 results (Mean: 71.6%, Median: 75.6%)
  - Out of 39 (instead of 40)
  - ★ Email Steve Van Doormaal (svandoor@uwaterloo.ca) for concerns about grading. He will pass your concern to the TAs who graded the question.

Last class:

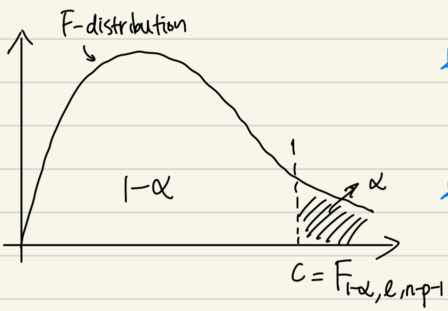
- t-test can be used to test for a single parameter
- F-test can be used to test for a single parameter and more -

F-test for General Linear Hypothesis ($H_0: A\vec{\beta} = \vec{0}$ vs. $H_a: A\vec{\beta} \neq \vec{0}$)

expect $\|\hat{\vec{\mu}} - \hat{\vec{\mu}}_A\|^2 \approx 0$.

Under H_0, F-statistic:

$$F = \frac{\big[\overbrace{SS_A(res) - SS(res)}^{\|\hat{\vec{\mu}} - \hat{\vec{\mu}}_A\|^2}\big]/\ell}{\underbrace{SS(res)/n-p-1}_{\hat{\sigma}^2 \text{ (in full model)}}} \sim F_{\ell, n-p-1}$$



★ Reject $H_0: A\vec{\beta} = 0$ at α-level if $F > c$, where c is $1-\alpha$ quantile of $F_{\ell, n-p-1}$ dist^n.

★ Reject $H_0: A\vec{\beta} = 0$ at α-level if p-value $= P(Y \geq F) < \alpha$

↳ $Y \sim F_{\ell, n-p-1}$

Recall :

Summaries in ANOVA table

Source	SS	df	Mean Squares (MS)	F
Regression	SS(reg)	p	$MS(reg) = SS(reg)/p$	$MS(reg)/MS(res)$
Residual	SS(res)	$n-p-1$	$MS(res) = SS(res)/n-p-1$	
Total	SS(tot)	$n-1$		

Recall that $F = \dfrac{[SS_A(res) - SS(res)]/\ell}{SS(res)/n-p-1} \sim F_{\ell, n-p-1}$ under H_0.

Suppose that $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs. H_a: at least one of $\beta_1, \cdots, \beta_p$ is not 0.

$H_0: \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \vec{0}$

From before, $A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{p \times (p+1)}$

Thus, we have p constraints ($\ell = p$)

$\therefore$ In this scenario, we test for the overall significance of our model (to see if any predictors are associated w/ the outcome).

If H_0 is true, then reduced model is $Y_i = \beta_0 + \varepsilon_i$ ($Y_i \overset{iid}{\sim} N(\beta_0, \sigma^2)$).

We can fit the reduced model using LS that aims to minimize $\sum_{i=1}^{n} (y_i - \beta_0)^2$. We will find that $\hat{\beta}_0 = \bar{y}$ (LS estimate).

$$\therefore SS_A(res) = \sum_{i=1}^{n} e_{i,A}^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_{i,A})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS(tot).$$

Thus, $F = \dfrac{[SS(tot) - SS(res)]/p}{SS(res)/(n-p-1)} = \dfrac{SS(reg)/p}{SS(res)/n-p-1} = \dfrac{MS(reg)}{MS(res)}$

This is the F-statistic from the ANOVA table!

- The ANOVA table allows us to test for overall significance of our model.

Interaction Effects

Suppose that we have x_1 and x_2. The interaction term between the 2 covariates is $x_1 x_2$. , e.g.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

- β_1 and β_2 are main effects.
- β_3 is the interaction effect between x_1 and x_2.

General Interpretation

- Mean response at x_1 and x_2:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \quad \text{—— (1)}$$

- Mean response at x_1+1 and x_2:

$$E(Y) = \beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \beta_3 (x_1+1) x_2$$
$$= (\beta_0 + \beta_1) + \beta_1 x_1 + (\beta_2 + \beta_3) x_2 + \beta_3 x_1 x_2 \quad \text{—— (2)}$$

- Take the difference (2) − (1):

$$\beta_1 + \beta_3 x_2$$

→ This is the change in mean response as x_1 increases by 1 unit holding x_2 constant (now depends on x_2 via β_3 : the interaction effect).

→ This implies that the association between x_1 and the response now also depends on x_2 (via β_3)

e.g. Lung function data:

$$E(Y) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 \underbrace{x_1 x_2}_{\text{age}*\text{sex}}$$

($E(Y)$ ↑ FEV; x_1 ↑ age; x_2 ↑ sex)

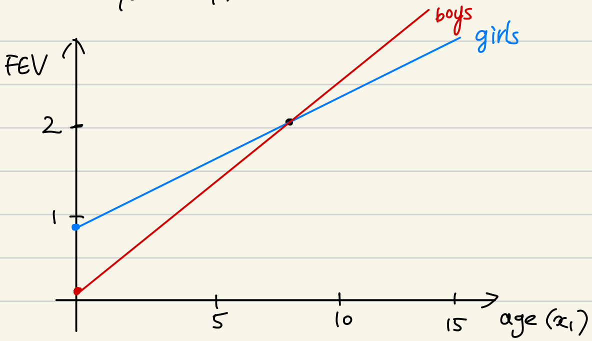
- For girls ($x_2 = 0$) with age equals x_1, mean FEV is
  $E(Y) = \alpha_0 + \alpha_1 x_1$ (α_0: intercept, α_1: slope)
- For boys ($x_2 = 1$) with age equals x_1, mean FEV is
  $E(Y) = \alpha_0 + \alpha_1 x_1 + \alpha_2 + \alpha_3 x_1$
  $= \underbrace{(\alpha_0 + \alpha_2)}_{\text{intercept}} + \underbrace{(\alpha_1 + \alpha_3)}_{\text{slope}} x_1$

- α_2: the difference in intercept between boys and girls, holding age constant
- α_3: " " in slope " " , " "

From R, the fitted model is $E(Y) = 0.926 + 0.156x_1 - 0.852x_2 + 0.117x_1x_2$
(e.g. lm (FEV~age + sex + age:sex, data= ...))

Girls: $E(Y) = 0.926 + 0.156x_1$

Boys: $E(Y) = 0.074 + 0.273x_1$



Q: Is the relationship between FEV and age the same for boys and girls?

A: Test for α_3 e.g. $H_0: \alpha_3 = 0$ vs. $H_a: \alpha_3 \neq 0$

Hierarchical principle

- Generally, we want to have a hierarchically well-formulated models:
  - → If there is pairwise interaction term, include main effects (if higher order interactions, include lower order interactions).
    - e.g. If we had $x_1 x_2$, we should also include x_1 and x_2 in model.
  - → otherwise, we might end up with inappropriate interpretations/implications.

Eg. $E(Y) = \alpha_0 + \alpha_3 \underset{\text{age}}{x_1} \underset{\text{sex}}{x_2}$

Mean FEV for girls ($x_2=0$) w/ x_1:

$$E(Y) = \alpha_0$$

Mean FEV for girls ($x_2=0$) w/ x_1+1:

$$E(Y) = \alpha_0$$

→ suddenly the mean FEV for girls is the same regardless of age.