2.6: Multiple Inference and Multicollinearity

Dr. Bean - Stat 5100

1 Why Multiple Inference?

We already have tools to test:

- Individual coefficients: t-tests
- *All* coefficients: model F-test

What if we want to consider the singnificance of a subset of the X predictor variables? (More than one, but not all of them).

(Individual) Why might we be interested in a "subset" F test?

We may wish to know if a group of predictors have a singificance influence on the response variable, *after accounting for* another set of variables that are already in the model.

Example: Bodyfat Dataset (Handout 2.6.1)

Y = body, X_1 = triceps, X_2 = thigh, X_3 = midarm

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

We've looked at the model F-test ($H_0 : \beta_1 = \beta_2 = \beta_3 = 0$)
– also individual t-tests ($H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$, $H_0 : \beta_3 = 0$)
– what about subset tests?

Consider $H_0 : \beta_2 = \beta_3 = 0$ – how to test this?

- basically, compare model fit with and without this assumption (H_0)
- Notation: $SSE(X_1, X_2, X_3) = SS_{error}$ when model has predictors X_1, X_2, and X_3
  – represents amount variation in Y left unexplained by model
- Assuming $H_0 : \beta_2 = \beta_3 = 0$ is true, fit "reduced" model (only predictor X_1) and calculate $SSE(X_1)$
- Note that $SSE(X_1) > SSE(X_1, X_2, X_3)$

    – ALWAYS true, as a "worthless" X variable won't ever increase the SSE, but may reduce it slightly by chance.
    – NOT true of validation error (more discussion in Module 4).

  – then define "extra sum of squares"

$$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$$

  Note: this represents amount variation in Y accounted for by X_2 & X_3 when X_1 already in model

- Define

$$MSR(X_2, X_3|X_1) = \frac{SSR(X_2, X_3|X_1)}{2}$$

  – think of this as the mean square reduction

- Build test statistic for $H_0 : \beta_2 = \beta_3 = 0$

$$\begin{aligned} F^* &= \frac{MSR(X_2, X_3|X_1)}{MSE(X_1, X_2, X_3)} \\ &= \frac{SSR(X_2, X_3|X_1)/(2)}{SSE(X_1, X_2, X_3)/(16)} \end{aligned}$$

- When $H_0 : \beta_2 = \beta_3 = 0$ is true, $F^* \sim F_{2,16}$

General test of any # of β_k's:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon \\ H_0 &: \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0 \\ p &= \text{\# of } \beta\text{'s in full model (incl. intercept)} \\ q &= \text{\# of } \beta\text{'s in reduced model (incl. intercept)} \\ p - q &= \text{\# of } \beta\text{'s being tested in } H_0 \\ F^* &= \frac{[(\text{SSE in reduced model}) - (\text{SSE in full model})]/(p-q)}{[\text{SSE in full model}]/(n-p)} \end{aligned}$$

Under H_0, $F^* \sim F_{p-q,n-p}$

Recall the t-statistic from test of individual predictor ($H_0 : \beta_k = 0$)?

$$t^* = \frac{b_k}{s\{b_k\}}$$

– if only have one predictor in model then $(t^*)^2 \sim F_{1,n-p}$

SSR also called sequential sums of squares or Type I SS; example in SAS:

- $SSR(X_1) \approx 352.27$
- $SSR(X_2|X_1) \approx 33.17$
- $SSR(X_3|X_1, X_2) \approx 11.55$

(Individual) True or False (and explain): Because the Type I SS associated with X_1 is greatest, it means that X_1 is the most significant coefficient in the model.

FALSE The first of the Type I SS will often be the largest because no other predictors have yet been accounted for. This is why order matters in the Type I SS calculation.

Related concept: "Coefficients of Partial Determination"

- what proportion of [previously unexplained] variation in Y can be explained by addition of predictor X_k to model

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)}$$

  - $SSR(X_3|X_1, X_2)$ - reduction in SSE that occurs when X_3 is added to the model when X_1 and X_2 are already in the model.
  - $SSE(X_1, X_2)$ - amount of unexplained variation in Y when X_1 and X_2 are in the model.

- example in SAS:
  - $R^2_{Y1} \approx 0.711$
  - $R^2_{Y2|1} \approx 0.232$
  - $R^2_{Y3|12} \approx 0.105$

(Draw box and fill in the first 71% of the big box, then fill in 23% of the little box that remains, finally fill in 10% of the even smaller box that remains.

<u>Textbook sections 7.6 and 10.5</u>

In bodyfat example (full model), compare model F-test to individual predictor t-tests