

## 2.6: Multiple Inference and Multicollinearity

Dr. Bean - Stat 5100

### 1 Why Multiple Inference?

We already have tools to test the significance of model coefficients:

- Individual coefficients: t-tests ( $H_0 : \beta_k = 0$ )
- All coefficients: model F-test ( $H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ )

What if we want to consider the significance 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?**

### 2 Subset Testing

**Example: Bodyfat Dataset (Handout 2.6.1)**

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

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

**(Groups) Consider  $H_0 : \beta_2 = \beta_3 = 0$ . How would you describe this hypothesis in an English sentence?**

**How to do this:** See how much better the full model is (using tricep, thigh, and midarm) compared to the reduced one (using only triceps).

- 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 the full 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  $\beta_k$ 's:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1} + \epsilon$$

$$H_0 : \beta_q = \beta_{q+1} = \dots = \beta_{p-1} = 0$$

$$p = \# \text{ of } \beta\text{'s in full model (incl. intercept)}$$

$$q = \# \text{ of } \beta\text{'s in reduced model (incl. intercept)}$$

$$p - q = \# \text{ of } \beta\text{'s being tested in } H_0$$

$$F^* = \frac{[(SSE \text{ in reduced model}) - (SSE \text{ in full model})]/(p - q)}{[SSE \text{ in full model}]/(n - p)}$$

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:

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

(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.

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

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

- $SSR(X_3|X_1, X_2)$  - 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:
  - $R_{Y1}^2 \approx$
  - $R_{Y2|1}^2 \approx$
  - $R_{Y3|12}^2 \approx$

### 3 Multicollinearity

Textbook sections 7.6 and 10.5

The model F test says that the coefficients *collectively* are highly significant, but *none* of the individual variables are significant.

This is a symptom of **multicollinearity** (i.e. collinearity):

- Two X variables share a strong linear relationship *with each other* (independent of Y)
- One X variable is a near linear combination of two or more X variables

## Problems with Multicollinearity:

- $\beta_k$  hard to interpret as it no longer makes sense to “hold all other predictor variables constant.”
- The variance of  $b_k$  will be very large (inflated) as our estimates are starting to become non-unique  $\rightarrow$  makes inference of  $\beta_k$  difficult if not impossible.
  - Could make estimate of  $b_k$  counter-intuitive (example: getting a negative estimate of  $b_k$  despite knowing that X and Y are positively correlated)).
- Contradictory results between individual t-tests and model F tests (or subset F tests).

## NOT Problems with Multicollinearity:

- Multicollinearity does NOT affect a model’s predictive ability.

### 3.1 Standardizing Variables

One way to better understand multicollinearity is by standardizing variables.

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left( \frac{Y_i - \bar{Y}}{\text{SD of } Y} \right) \quad , \quad X_{ik}^* = \frac{1}{\sqrt{n-1}} \left( \frac{X_{ik} - \bar{X}_k}{\text{SD of } X_k} \right)$$

– sometimes called “correlation transformation” because

$$\text{Corr}(X_k, Y) = \sum_i X_{ik}^* Y_i^*$$

If all variables have been standardized, then consider matrix approach (with no Intercept column in matrix  $X^*$ ):

$$\begin{aligned} Y^* &= X^* \beta^* + \varepsilon \\ b^* &= (X^{*'} X^*)^{-1} X^{*'} Y^* \\ \text{Cov}(b^*) &= (X^{*'} X^*)^{-1} \sigma^2 \end{aligned}$$

To un-standardize regression coefficient estimates:

$$\begin{aligned} b_k &= \left( \frac{\text{SD of } Y}{\text{SD of } X_k} \right) \cdot b_k^* \\ b_0 &= \bar{Y} - \sum_{k=1}^{p-1} b_k \bar{X}_k \end{aligned}$$

Relevance to multicollinearity:

- the correlation matrix among the [original] predictor variables is  $X^{*'} X^*$
- the “closer”  $X_j$  and  $X_h$  are, the larger will be the  $j^{th}$  and  $h^{th}$  diagonal elements of  $\text{Cov}(b^*)$ , so the estimated variance is higher for  $b_j$  and  $b_h$

- We can use the correlation matrix to obtain a set of **condition indices** as obtained from the **eigenvalues** of the matrix.

While standardizing helps to better mathematically understand the effect of multicollinearity, it is not necessary to standardize to detect multicollinearity.

**(Groups) What other advantages (besides help with multicollinearity) might standardizing our variables provide us?**

## 3.2 Ways to Diagnose Multicollinearity

### 3.2.1 Condition Index/Principal Components

- Recall from linear algebra:  $\lambda$  is an **eigenvalue** of a symmetric, square matrix  $A$  iff there exists a vector  $x$  (the **eigenvector** for  $\lambda$ ) such that  $Ax = \lambda x$ .
- Let  $\lambda_1, \dots, \lambda_k$  be the eigenvalues of  $X^{*'}X^*$ , and let

$$\text{Condition Index}_i = \left( \frac{\lambda_{\max}}{\lambda_i} \right)^{1/2}$$

- Each condition index is associated with a **principal component**
  - Each principal component is a linear combination of the original predictor variables. Each principal component shares no correlation with any other principal component (i.e.  $\text{cor}(PC_1, PC_2) = 0$ ).

$$\begin{aligned} PC_1 &= a_1 X_1^* + \dots + a_{p-1} X_{p-1}^* \\ PC_2 &= c_1 X_1^* + \dots + c_{p-1} X_{p-1}^* \\ &\vdots \end{aligned}$$

- Each principal component explains some percentage of the variation in the original predictors.

**IF** the condition index is high (more than 10 or so) **AND** the associated principal component explains a high proportion of the variance (usually more than 50% variability) in two or more predictor variables, then we have potentially problematic multicollinearity.

### 3.2.2 Variance Inflation Factor (VIF)

- Let  $R_k^2$  be the coefficient of multiple determination (the  $R^2$  value) when predictor  $X_k^*$  is regressed on the other predictors
  - This is a measure of how much of the variance of  $X_k^*$  is explained by the other  $X$  variables.
- Define  $VIF_k = (1 - R_k^2)^{-1}$ , for  $k = 1, \dots, p - 1$  as the “Variance Inflation Factor” for  $b_k$  (the estimate of  $\beta_k$ )

**IF** the largest VIF is much more than 10 **OR** the average VIF is much more than 1, then we have evidence of potentially problematic multicollinearity.

We usually use a combination of the VIF and condition index to assess multicollinearity.

### 3.2.3 Important things to remember about standardization

- Relative magnitude of  $b_k^*$  estimates not meaningful if predictors are on different scales
- Standardization most common when predictors  $X_1, \dots, X_{p-1}$  have very different scales
- $\beta_k^*$  is expected change in Y for every SD (not unit) increase in predictor  $X_k$ , while all other predictors are held constant
- Standardizing has:
  - no effect on VIF
  - marginal effect on proportions of variance in Condition Index output
  - possibly substantial effect on magnitude of Condition Indexes
- Recommendations:
  - Standardize if either:
    - \* desire common scale of  $b_k^*$  estimates
    - \* need uncorrelated, higher-order predictors

### 3.3 Multicollinearity Summary

Three ways to diagnose multicollinearity:

- 1.
- 2.
- 3.

Possible remedial measures for multicollinearity:

- Collect more data

- Choose a subset of predictor variables
- Ridge regression
- Latent root regression – use Principal Components as predictors (may lack interpretability)

$$PC = a_1X_1 + a_2X_2 + \dots + a_{p-1}X_{p-1}$$