Lecture 15: May 23

Lecturer: Arash Amini *Scribe: Eric Chuu*

1.1 Additional Sum of Squares

Recall the following setup for generalized linear hypothesis, where $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$.

$$\begin{cases} H_0 : \boldsymbol{\mu} \in L_A(X) \\ H_a : \boldsymbol{\mu} \notin L_A(X), \boldsymbol{\mu} \in L(X) \end{cases}$$

We also have the following relationship between the residual of the full model and the restricted model.

$$\|e_A\|^2 = \|e\|^2 + \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}_A}\|^2 \tag{1.1}$$

Theorem. Suppose that H_0, as defined above, holds. Then $\mu \in L_A(X)$ and $A\boldsymbol{\beta} = \mathbf{0}$, where $A \in \mathbb{R}^{q \times p+1}$, where q is the number of restrictions, and p is the number of predictors. Then,

(a) $\frac{\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}_A}\|^2}{\sigma^2} \sim \chi_q^2$
(b) $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}_A}\|^2$ is independent of $\|e\|^2$

Proof. We can then use the relationship given in equation (1.1) and decompose the residuals,

$$\begin{aligned} \|\mathbf{e}_A\|^2 = \|\mathbf{e}\|^2 + \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}_A}\|^2 &\Rightarrow \|(I - H_A)\boldsymbol{\varepsilon}\|^2 = \|(I - H)\boldsymbol{\varepsilon}\|^2 + \|(H - H_A)\boldsymbol{\varepsilon}\|^2 \\ &\Rightarrow \frac{\|(I - H_A)\boldsymbol{\varepsilon}\|^2}{\sigma^2} = \frac{\|(I - H)\boldsymbol{\varepsilon}\|^2}{\sigma^2} + \frac{\|(H - H_A)\boldsymbol{\varepsilon}\|^2}{\sigma^2} \\ &\Rightarrow Q = Q_1 + Q_2 \end{aligned}$$

Using the simpler notation given in the last equality, we note that Q, Q_1, Q_2 are all quadratic forms in $\boldsymbol{\varepsilon}$. Further, we know that $Q \sim \chi^2_{\text{rank}(I - H_A)} \Rightarrow Q \sim \chi^2_{n-(p+1-q)}$. Similarly, $Q_1 \sim \chi^2_{n-p-1}$. It is clear that $Q_2 \geq 0$, so by Cochran's Theorem, we conclude that $Q_2 \sim \chi^2_{r_2}$, where $r_2 = n - (p + 1 - q) - (n - (p + 1)) = q$, so $Q_2 \sim \chi_q^2$ and $Q_2 \perp\!\!\!\perp Q_1$. □

As a corollary, it is easy to see that under H_0, we can form the following quantity

$$\hat{F} := \frac{\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}_A}\|^2/q}{\|\mathbf{e}\|^2/(n-p-1)} = \frac{\left(\|\mathbf{e}_A\|^2 - \|\mathbf{e}\|^2\right)/q}{\|\mathbf{e}\|^2/(n-p-1)} \sim F_{q,n-p-1} \tag{1.2}$$

1.1.1 Example

We consider the the following relationship,

$$y = \alpha z + \beta_1 x_1 + \cdots + \beta_5 x_5 + \epsilon \tag{1.3}$$

where $x_1, \dots, x_5$ are indicator variables. We consider the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 =: \gamma$$

In other words, we are testing to see if there are differences between the five groups. Under the null hypothesis, we can write the restricted model as

$$\begin{aligned} \mathbf{y} &= \alpha \mathbf{z} + \gamma \mathbf{x}_1 + \gamma \mathbf{x}_2 + \gamma \mathbf{x}_3 + \gamma \mathbf{x}_4 + \gamma \mathbf{x}_5 + \boldsymbol{\varepsilon} \\ &= \alpha \mathbf{z} + \gamma \cdot \mathbf{1} + \boldsymbol{\varepsilon} \end{aligned}$$

The second equality holds because the $\mathbf{x}_i$'s are indicator variables, adding them together gives us a vector of 1's. See pages 113, 119, and 120 for numerical calculations for the F statistic and calculating the degrees of freedom.

1.2 ANOVA

ANOVA is considered a special case of general linear hypothesis testing. Suppose we have the usual setup:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \beta_j x_j + \boldsymbol{\varepsilon}$$

Then we want to form the following test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Under this test, we are considering the model

$$\mathbf{y} = \beta_0 \cdot \mathbf{1} + \boldsymbol{\varepsilon}$$

We define a few quantities

$$\begin{aligned} \text{SSE} &= \|\mathbf{e}\|^2 \\ \text{SSR} &= \left\| \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}_A}^2 \right\| \\ \text{SST} &= \|\mathbf{e}_A\|^2 \end{aligned}$$

We can arrange these in a table to more easily read off the degrees of freedom:

Source	df	Sum of Squares	Mean Squares	F
Model (Regression)	p	$\text{SSR} = \hat{\boldsymbol{\beta}}' X' \mathbf{y} - n\bar{y}^2$	$\text{MSR} = \text{SSR}/p$	$\frac{\text{MSR}}{\text{MSE}}$
Residual (Error)	$n-p-1$	$\text{SSE} = S(\hat{\boldsymbol{\beta}}) = \|\mathbf{e}\|^2$	$\text{MSE} = \text{SSE}/(n-p-1)$	
Corrected Total	$n-1$	$\text{SST} = \|\mathbf{e}_A\|^2 = \sum (y_i - \bar{y})^2$		