Chapter 6 Additional Sum of Squares and Testing Subsets of Regression Coefficients

- Testing subsets of regression coefficients

Can several predictor variables be eliminated from a model simultaneously?

Model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon \Leftrightarrow E(Y) = \beta_0 + \beta_1 X + \cdots + \beta_{p-1} X_{p-1}$

$H_0: \beta_1 = \beta_2 = 0$ vs. H_a: at least one of β_1, β_2 is not 0. Equivalently,

$H_0: E(Y) = \beta_0 + \beta_3 X_3 + \cdots + \beta_{p-1} X_{p-1}$ (reduced model) vs.
$H_a: E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$ (full model)

Procedure:

(1) Fit Y on $X_3, X_4, \cdots, X_{p-1}$ $\rightarrow SS_{reg}(red)$.
(2) Fit Y on $X_1, X_2, X_3, \cdots, X_{p-1}$ $\rightarrow SS_{reg}(full)$.

Additional sum of squares (SS) due to adding X_1, X_2:

$$
\begin{aligned}
SS\left(X_1, X_2 \middle| X_3, X_4, \cdots, X_{p-1}\right) &= SS_{reg}(full) - SS_{reg}(red) \\
&= [TSS - RSS(full)] - [TSS - RSS(red)] \\
&= RSS(red) - RSS(full)
\end{aligned}
$$

— The sum of squares of X_1, X_2 given $X_3, X_4, \cdots, X_{p-1}$, which measures the contribution of X_1, X_2 to the regression sum of squares given $X_3, X_4, \cdots, X_{p-1}$.

Test statistic: $F = \frac{\left[SS_{reg}(full) - SS_{reg}(red)\right]/2}{RSS(full)/(n-p)}$

In general, want to test

$H_0: \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_q} = 0$ vs. H_a: at least one of $\beta_{i_1}, \cdots, \beta_{i_q}$ is not 0.

$$F = \frac{\left[SS_{reg}(full) - SS_{reg}(red)\right]/a}{RSS(full)/b},$$

where $a = df\ of\ \left[SS_{reg}(full) - SS_{reg}(red)\right],\ b = df\ of\ RSS(full) = n - p.$

Note: If $\boldsymbol{X}'\boldsymbol{X}(full)$ is nonsingular,
a = df of $\left[SS_{reg}(full) - SS_{reg}(red)\right]$ = $[df\ of\ SS_{reg}(full)] - [df\ of\ SS_{reg}(red)]$
= (# of predictor variables in the full model) − (# of predictor variables in the reduced model)
= $(p-1) - [(p-1) - q]$ = q = # of predictor variables eliminated.

If $\varepsilon_1, \cdots, \varepsilon_n$ are $iid\ N(0, \sigma^2)$ and H_0 is true, $F \sim F_{a,b} = F_{q,n-p}$.

Reject H_0 if p-value = $P(F_{q,n-p} \geq F_{obs}) \leq \alpha$ or $F_{obs} \geq F_{q,n-p}(1-\alpha)$.

Note: Testing overall linear relationship and testing a coefficient are special cases of testing subsets of regression coefficients.

<u>Additional (Extra) Sum of Squares Principle</u>: Assess the importance of q predictor variables $X_{i_1}, \cdots, X_{i_q}$ in a multiple regression model by the additional SS they account for, after all other predictor variables have been accounted for, i.e.,

$$
\begin{aligned}
SS\left(X_{i_1}, \cdots, X_{i_q} \middle| \text{all other predictor variables}\right) &= SS_{reg}(full) - SS_{reg}(red) \\
&= RSS(red) - RSS(full),
\end{aligned}
$$

where the full model is the model with all predictor variables involved and the reduced model is the model with $X_{i_1}, \cdots, X_{i_q}$removed.

Note: Essentially, $SS\left(X_{i_1}, \cdots, X_{i_q} \middle| \text{all other predictor variables}\right)$ is the variation in Y explained by $X_{i_1}, \cdots, X_{i_q}$ given all other variables in the model, which measures the contribution of $X_{i_1}, \cdots, X_{i_q}$ to the regression sum of squares given all other predictor variables in the model.

- Sequential Sums of Squares

SS_{reg} can be decomposed into $(p-1)$ sum of squares, each with 1 df corresponding to the $(p-1)$ predictor variables. However, the decomposition is not unique. Different orders of the predictor variables yield different decompositions.

Source of variation	df	Sequential SS
X_1	1	$SS(X_1)$
$X_2 \| X_1$	1	$SS(X_2 \| X_1)$
$X_3 \| X_1, X_2$	1	$SS(X_3 \| X_1, X_2)$
$\vdots$	$\vdots$	$\vdots$
$X_{p-1} \| X_1, X_2, \cdots, X_{p-2}$	1	$SS(X_{p-1} \| X_1, X_2, \cdots, X_{p-2})$

Notes: (1) $\left(\frac{SS(X_1)}{TSS} \times 100\right)$ % of the variation in Y is explained by the regression using X_1 alone.

(2) The contribution of X_i to SS_{reg} given that $X_1, X_2, \cdots, X_{i-1}$ are already in the model is $SS(X_i | X_1, X_2, \cdots, X_{i-1})$, which accounts for $\left(\frac{SS(X_i | X_1, X_2, \cdots, X_{i-1})}{TSS} \times 100\right)$ % of the variation in Y.

Example 6.1: For the data in Example 5.3,

(1) How useful is the regression using X_1 alone? What does X_2 contribute, given that X_1 is already in the regression?

(2) Test to determine whether we can eliminate X_1, X_2, X_3, X_4, X_6, X_8, and X_9 simultaneously using α = 0.05.