Use of Extra Sums of Squares for Testing Hypotheses about

Context: Assume we have $Y, \underset{\sim}{X}_1, \dots, \underset{\sim}{X}_{p-1}$, we're assuming

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

* Testing whether multiple β_2, β_3 are jointly equal to 0

Null, H_0: $\beta_2 = \beta_3 = 0$
Alternative, H_a: $\beta_2 \neq 0$ or $\beta_3 \neq 0$

Compute

$$F^* = \left(\frac{SSR(X_2, X_3 \mid X_1)}{2}\right) \div \left(\frac{SSE(X_1, X_2, X_3)}{n-4}\right)$$

Decicision Rule for a level α-test will be:

* If $F^* \leq F(1-\alpha; 2, n-4)$ conclude H_0
* If $F^* > F(1-\alpha; 2, n-4)$ conclude H_a

p-value given by $Pr(F \geq F^*)$, where $F \sim F_{2, n-4}$

Use of Extra Sums of Squares for Testing Hypotheses about β_k

Context: Assume we have $Y, \underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_{p-1}$, we're assuming

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} X_{ik}\beta_k + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

* Testing whether multiple β_k, $k \in S$ are jointly equal to 0

S ← set of q indices

Null, H_0: $\beta_k = 0$ for all $k \in S$

Alternative, H_a: $\beta_k \neq 0$ for at least one $k \in S$

Compute

$$F^* = \left(\frac{SSR(X_k, k \in S \mid X_m, m \notin S)}{q}\right) \div \left(\frac{SSE(X_1, \ldots, X_{p-1})}{n-p}\right)$$

Decision Rule for a level α-test will be:

* If $F^* \leq F(1-\alpha; q, n-p)$ conclude H_0
* If $F^* > F(1-\alpha; q, n-p)$ conclude H_a

p-value given by $Pr(F \geq F^*)$, where $F \sim F_{q, n-p}$

Use of Extra Sums of Squares for Testing Hypotheses about β_k

Context: Assume we have $\underset{\sim}{Y}, \underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_{p-1}$, we're assuming

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} X_{ik}\beta_k + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

*Testing whether multiple $\beta_k = \beta_\ell$, $\ell \neq k$ are jointly equal to 0

Null, H_0: $\beta_k = \beta_\ell$
Alternative, H_a: $\beta_k \neq \beta_\ell$

Under the null, we assume a reduced model

$$Y_i = \beta_0 + \beta_k(X_{ik} + X_{i\ell}) + \sum_{m=1,\, m\neq k \text{ and } m\neq \ell}^{p-1} \beta_m X_{im} + \varepsilon_i,$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Test this null by fitting original model, computing the SSE(full) and fitting the reduced model, computing SSE(reduced)

Construct our $F^* = \left(\dfrac{SSE(reduced) - SSE(full)}{1}\right) \Big/ \dfrac{SSE(full)}{n-p}$

Perform a level α test with decision rule based on $F(1-\alpha, 1, n-p)$

if $F^* \leq F(1-\alpha, 1, n-p)$ conclude H_0 if $F^* > F(1-\alpha, 1, n-p)$ conclude H_a

* Standardizing the multiple regression model
* Correlated predictors

Multiple lin. reg. model

$$y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Standardized multiple regression model is

$$\frac{y_i - \bar{Y}}{S_y} = \gamma_0 + \sum_{k=1}^{p-1} \gamma_k \left(\frac{X_{ik} - \bar{X}_k}{S_k} \right) + \delta_i, \quad \delta_i \overset{iid}{\sim} N(0, \omega^2)$$

$$S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad S_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2}, \quad \bar{X}_k = \frac{1}{n} \sum_{i=1}^{n} X_{ik}$$

$$\tilde{y}_i = \gamma_0 + \sum_{k=1}^{p-1} \gamma_k \tilde{X}_{ik} + \delta_i, \quad \delta_i \overset{iid}{\sim} N(0, \omega^2)$$

$$\frac{y_i - \bar{y}}{s_y} = \gamma_0 + \sum_{k=1}^{p-1} \gamma_k \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) + \delta_i, \quad \delta_i \overset{iid}{\sim} N(0, \omega^2)$$

Same as saying

$$y_i - \bar{y} = s_y \gamma_0 + \sum_{k=1}^{p-1} s_y \gamma_k \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) + s_y \delta_i, \quad \delta_i \overset{iid}{\sim} N(0, \omega^2)$$

Same as saying

$$y_i = \bar{y} + s_y \gamma_0 + \sum_{k=1}^{p-1} s_y \gamma_k \left(\frac{x_{ik} - \bar{x}_k}{s_k} \right) + s_y \delta_i, \quad \delta_i \overset{iid}{\sim} N(0, \omega^2)$$

$$y_i = \bar{y} + s_y \gamma_0 - \left(\sum_{k=1}^{p-1} \left(\frac{s_y}{s_k} \right) \gamma_k \bar{x}_k \right) + \sum_{k=1}^{p-1} \left(\frac{s_y}{s_k} \right) \gamma_k x_{ik} + s_y \delta_i, \quad \delta_i \overset{iid}{\sim} N(0, \omega^2)$$

$$\Longrightarrow \quad \beta_k = \frac{s_y}{s_k} \gamma_k, \quad \text{for } k \geq 1$$

$$\Longrightarrow \quad \beta_0 = \bar{y} - \sum_{k=1}^{p-1} \beta_k \bar{x}_k + \underbrace{\cancel{s_y \gamma_0}}_{=0}$$

when we have subtracted off mean of y_i's and x_{ik}'s, we don't need an intercept, so $\gamma_0 = 0$

Relate standardized ($g_1, \ldots, g_{p-1}$) to non-standardized ($b_0, \ldots, b_{p-1}$) regression coefficient estimates

$$b_k = \left(\frac{s_Y}{s_k}\right) g_k \quad \text{for } k \geq 1$$

and

$$b_0 = \bar{Y} - \sum_{k=1}^{p-1} b_k \bar{X}_k$$

Why do we care about standardizing?

* Numerically, solving for the regression coefficients of the standardized model is more stable

* Interpretation! Can be helpful if we want to compare estimated regression coefficients

let's consider a problem where Y measured in dollars
X_1 is measured in thousands of dollars
X_2 is measured in cents

let's consider a problem where Y measured in dollars
X_1 is measured in thousands of dollars
X_2 is measured in cents

We assume

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

we get estimates $b_0 = 200$
$b_1 = 20{,}000$
$b_2 = 0.2$

what is the interpretation of b_1?

average change in Y in dollars when X_1 increases by ~~one unit~~ one thousand dollars, holding X_1 constant

what is the interpretation of b_2?

average change in Y in dollars when X_2 increases by ~~one unit~~ one cent, holding X_2 constant

What if we assumed the standardized model, and got estimates g_1 and g_2 of the standardized regression coefficients?

Interpretation of g_1 would be

average change in Y in standard deviations from $\bar{Y}$ when X_1 is increased by one standard deviation from its mean, holding X_2 constant