STA 302 / 1002

Note Title 1/16/2012

Reminder: All about SAS (except programming and interpreting output) on Thursday (Jan 19)

SAS: GLM procedure
"General Linear Model"

Alternative to proc reg
- doesn't have all the regression features of proc reg (e.g. variable selection techniques)
- does have lots of other features; better for dealing

with categorical predictor variables

Output: ANOVA table as in proc reg

Then Type I SS (sequential Sum of Squares)
- we won't use in this course

then Type III SS - contribution to regression sum of squares over and above everything else in model

From solution option: estimates of β's and t-tests for tests $H_0: \beta_j = 0$

Comparing output from proc ttest, reg, glm:

From output:

t test

Means: $\bar{x}_{other} = 29.49$

$\bar{x}_{spock} = 14.62$

diff = 14.87

Test stat for testing equal means (assuming equal variances)

$t_{obs} = 5.67$

Distn under H_0: t_{44}

p-value $< .0001$

reg with $x_i = I_{spock,i}$

estimates of parameters

$b_0 = 29.49$

$b_1 = -14.87$

b_0 is est $E(Y_i \mid x_i = 0)$

$b_0 + b_1$ is est $E(Y_i \mid x_i = 1)$

Test $H_0: \beta_1 = 0$

equivalent to both groups have same mean

Test stat: $t_{obs} = -5.67$

Distn under H_0: t_{44}

p-value $< .0001$

glm

estimates of parameters

$b_0 = 14.62$

$b_1 = 14.87$

b_0 is est $E(Y_i \mid$ ith obs is spock's judge)

$b_0 + b_1$ is est $E(Y_i \mid$ ith obs is other judge)

Test $H_0: \beta_1 = 0$

Test stat: 5.67

Distn under H_0: t_{44}

p-value: $< .0001$

Model being fit by proc glm:

$$Y_i = \beta_0 + \beta_1 I_{other,i} + e_i$$

Type III SS F-test $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

Test stat: $F_{obs} = 32.15 \; (= 5.67^2)$

Distn under H_0: $F_{1,44}$

p-value: $< .0001$

Why does proc GLM give NOTE: "The X'X matrix has been found to be singular ..."

Review of Regression in Matrix Terms

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$

$i = 1, \ldots, n$ = # of observations

Matrix form $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

where $\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, $\underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

$$\underset{\sim}{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{np} \end{pmatrix}_{n \times (p+1)}$$

Least squares estimates of β's:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$X'X$ has dimension $(p+1)\times(p+1)$

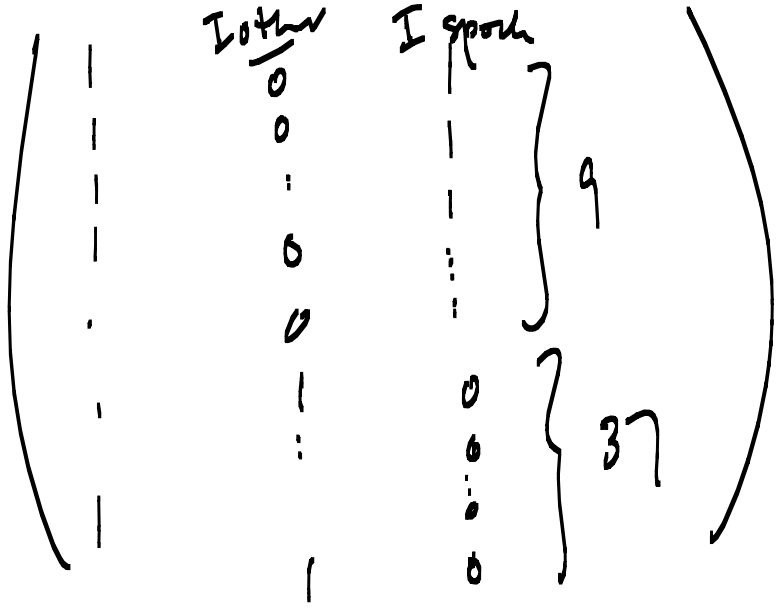
Need $X'X$ to be full rank to be invertible

rank $X'X$ = rank X

Need X to be rank $p+1$; need columns of X to be linearly independent

Class statement in proc glm creates dummy variable for all categories of the variable. So if created

I_{speech} and $I_{other} = \begin{cases} 1 & \text{if other judge} \\ 0 & \text{if Speech's judge} \end{cases}$

$\Rightarrow$ $X =$ 

(linearly dept columns)

Prof glm dropped the last column

For a categorical variable with G categories, need G-1 indicator variables in the model

Answer to 1st question: Does Spock's judge differ from other judges?

We have very strong evidence ($p < .0001$) that the mean % women differs between Spock's judge's venires and the other judge's venires. On average, Spock's has 14.6% women, and the other judges have 29.5% women.

2nd question: Is there evidence of any differences among other 6 judges?

H_0: $\mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$

vs

H_a: at least one of these means is not equal to another

where μ_A is mean % women on venires for judge A, etc.

Linear model:

$$Y_i = \beta_0 + \beta_1 I_{A,i} + \beta_2 I_{B,i} + \beta_3 I_{C,i} + \beta_4 I_{D,i} + \beta_5 I_{E,i} + e_i$$

where $I_{A,i} = \begin{cases} 1 & \text{if } i\text{th observation is judge A} \\ 0 & \text{otherwise} \end{cases}$

etc.

In general, have G groups, want to test whether they all have the same mean, need $G-1$ indicator variables

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1 & \text{if } i^{th} \text{ obs}^n \text{ in } 1^{st} \text{ group} \\ \beta_0 + \beta_2 & \text{if } i^{th} \text{ obs}^n \text{ in } 2^{nd} \text{ group} \\ \vdots & \vdots \\ \beta_0 + \beta_{G-1} & \text{if } i^{th} \text{ obs}^n \text{ in } G-1^{st} \text{ group} \\ \beta_0 & \text{if } i^{th} \text{ obs}^n \text{ in } G^{th} \text{ group} \end{cases}$$

To test whether the means of all G groups are equal

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_{G-1} = 0$$

Can show least squares estimates of $\beta_0, \ldots, \beta_5$ are

$b_0 = \bar{y}_6 \quad (= \bar{y}_G)$ average of observations in G^{th} group

$b_1 = \bar{y}_1 - \bar{y}_6 \quad (= \bar{y}_1 - \bar{y}_G)$ ($G = 6$ for our example)

$\hat{y}_i$ = predicted value of y_i

$$= \begin{cases} b_0 + b_1 & i^{th} \text{ observation in } 1^{st} \text{ group} \\ b_0 + b_2 & " \quad " \quad " \ 2^{nd} " \\ \vdots & \\ b_0 & " \quad " \quad " \ G^{th} \text{ group} \end{cases}$$

$$= \begin{cases} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_4 \end{cases}$$

For our example Test $H_0: \beta_1 = \ldots = \beta_r = 0$

vs $H_a:$ at least one not zero

Analysis of variance F-test:

$F_{obs} = 1.22$

p-value = .3235

F_{obs} is an observation from F distribution with 5, 31 distribution

5 → # of judges being tested

31 → df error

No evidence of differences in means among other judges

Sums and squares and Analysis of Variance Table

Total SS (Corrected SS)

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2 \qquad N = n_1 + n_2 + \cdots + n_G$$

$df = N - 1$

Decomposition $= \underbrace{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}_{SSReg} + \underbrace{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}_{RSS}$

In 1-way analysis of variance: (predictor are indicator variables that classify the observations one way)

$\hat{y}_i$ = mean of observations for group, g_i, in which the i^{th} observation belongs.

$SSReg$ (Model SS) $= \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$, $\hat{y}_i$ is one of

$df = G - 1$

g is an index for group $g = 1, 2, \ldots, G$

$$= \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2$$

$\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_G$

RSS (Residual or error SS) $= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

$$= \sum_{g=1}^{G} \sum_{(g)} (y_i - \bar{y}_g)^2$$

$\sum_{(g)}$ indicates summation over observations in group g

$df = n - (G-1) - 1$

$= n - G$

ANOVA Table for 1-way classification

Source	df	SS	MS	F-Stat
Group	G-1	SSReg	SSReg/G-1	SSReg/G-1 / MSE
Error	N-G	RSS	RSS/N-G = MSE	
Total	N-1	SST		

MS = "Mean Square"

In 1-way analysis of variance

SSReg often called "between group" SS

RSS often called "within group" SS

Idea of F-test: If between group variation is large compared to within group variation, there is evidence of differences in means among the groups