2009/09/17

Qualified Test of Statistical Methods

1. A data set consists of n observations on $\mathbf{X}_n$ and $\mathbf{y}_n$. The least squares estimator based on these n observations is $\mathbf{b}_n = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{y}_n$. Another observation, $\mathbf{x}_s$ and y_s, becomes available. Prove that the least squares estimator computed using this additional observation is

$$\mathbf{b}_{n,s} = \mathbf{b}_n + \frac{1}{1+\mathbf{x}_s'(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{x}_s}(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{x}_s(y_s - \mathbf{x}_s'\mathbf{b}_n). \quad (10\%)$$

2. Consider the general linear model $Y = X\beta + \varepsilon$, ε is distributed $N(0, \sigma^2 I)$, Y is a 16×1 random vector, and $\beta' = [\beta_0, \beta_1, \beta_2, \beta_3]$. The normal equations are

$$16\hat{\beta}_0 + 8\hat{\beta}_1 + 4\hat{\beta}_2 - 4\hat{\beta}_3 = 4$$

$$8\hat{\beta}_0 + 5\hat{\beta}_1 + 3\hat{\beta}_2 = 5$$

$$4\hat{\beta}_0 + 3\hat{\beta}_1 + 6\hat{\beta}_2 + 3\hat{\beta}_3 = 0$$

$$-4\hat{\beta}_0 + 3\hat{\beta}_2 + 7\hat{\beta}_3 = 5$$

and $y'y = 54$.

(a) Find point estimates of the following: (i) β (ii) σ^2 (iii) $4\beta_0 + 5\beta_1 + \beta_2 + 5\beta_3$ (iv) $8\beta_0 + 5\beta_1 + 9\beta_2 + 4\beta_3$.

(b) Test the hypothesis

$$H_0 : \begin{cases} 2\beta_1 + 4\beta_3 = 0 \\ 2\beta_2 + \beta_3 = 0 \end{cases} \quad \text{vs.} \quad H_a : \begin{cases} 2\beta_1 + 4\beta_3 \neq 0 \\ 2\beta_2 + \beta_3 \neq 0 \end{cases}.$$

Use $\alpha = 0.05$. (20%)

3. Consider the following linear statistical model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases}$$

where y_{ij} is the (ij) th observation, μ is a parameter common to all treatments called the overall mean, τ_i is a parameter unique to the i th treatment called the treatment effect, and ε_{ij} is a random error component. The

model errors are assumed to be normally and independently distributed random variables with mean zero and variance σ^2. The variance σ^2 is assumed to be constant for all levels of the factor. In the fixed effects model, the treatment effects τ_i are usually defined as deviations from the overall mean, so

$$\sum_{i=1}^{a} \tau_i = 0 .$$

Now, we could partition total variability into its component parts. That is,

$$SS_T = SS_{Treatments} + SS_E .$$

Equivalently,

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 ,$$

where $\bar{y}_{i.} = \sum_{j=1}^{n} y_{ij} \Big/ n$, $\bar{y}_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \Big/ N$, and $N = an$.

According the above statements, show the following results

(i). $E(MS_E) = \sigma^2$.

(ii). $E(MS_{Treatments}) = \sigma^2 + \dfrac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$. (20%)

4. An industrial engineer is studying the effect of five illumination levels on the occurrence of defects in an assembly operation. Because time may be a factor in the experiment, she has decided to run the experiment in five blocks, where each block is a day of the week. However, the department in which the experiment is conducted has four work stations and these stations represent a potential source of variability. The engineer decided to run a Youden square (incomplete Latin square designs) with 5 rows (days or blocks), 4 columns (work stations), and 5 treatments (the illumination levels). The coded data are shown in the following table.

Day	Work Station			
(Block)	1	2	3	4
1	$A = 3$	$B = 1$	$C = -2$	$D = 0$
2	$B = 0$	$C = 0$	$D = -1$	$E = 7$
3	$C = -1$	$D = 0$	$E = 5$	$A = 3$
4	$D = -1$	$E = 6$	$A = 4$	$B = 0$
5	$E = 5$	$A = 2$	$B = 1$	$C = -1$

The complete analysis of variance is shown in the following table. Show that

(i). Sum of Squares of Illumination level, adjusted = 120.37.

(ii). Sum of Squares of Days, adjusted=0.87.

(iii). Such design could be regarded as what kind of design?

(iv). What's your conclusion?

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Illumination level, adjusted	120.37	4	30.09	36.87[a]
Days, unadjusted	6.70	4	--	
Days, adjusted	(0.87)	(4)	0.22	
Work station	1.35	3	0.45	
Error	6.53	8	0.82	
Total	134.95	19		

[a] Significant at 1 percent. (20%)

5. Assume that the 4×1 vector X is distributed $N(\mu,\Sigma)$ and that a random sample of size $n=26$ was selected form this p.d.f. and the matrix $A=\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})'$ computed, where

$$A=\begin{bmatrix} 10 & 9 & -1 & -16 \\ 9 & 20 & -3 & -16 \\ -1 & -3 & 5 & 3 \\ -16 & -16 & 3 & 27 \end{bmatrix}.$$

(i). Find the M.L. estimate of $\rho_{13|(2,4)}$, correlation coefficient of X_1 and X_3 in the conditional pdf of $(X_1, X_3 \mid X_2, X_4)$.

(ii). Find the M.L. estimate of $\rho_{12|(3)}$.

(iii). Test $H_0:\rho_{23|(1)}=0$ vs. $H_a:\rho_{23|(1)}\neq 0$.

(iv). Find the M.L. estimate of $\rho^2_{1(2,3,4)}$, square of multiple correlation coefficient of X_1 and (X_2, X_3, X_4). (20%)

6. In the case of ridge regression,

(i). for the ridge regression estimator, which is found by solving the system of equations $(X'X+kI)b_R=X'y$ where $k\geq 0$, show that

$$\sum_{i=1}^{k}\frac{Varb_{i,R}}{\sigma^2}=\sum_{i=1}^{k}\frac{\lambda_i}{(\lambda_i+k)^2}$$

where the λ_i are the eigenvaules of the $(X^{*'}X^{*})$ matrix (correlation form).

(ii). show that the variance inflation factors, which are the diagonal elements of

$$(X^{*'}X^{*}+kI)^{-1}(X^{*'}X^{*})(X^{*'}X^{*}+kI)^{-1}$$

decrease with an increasing k. (10%)