//90

STAT 2509B
Test#2
SOLUTION

1. **Agency revenues.** An economic consultant was retained by a large employment agency in a metropolitan area to develop a regression model for predicting monthly agency revenues (y). She decided that three economic indicators for the area were potentially useful as independent variables, namely, average weekly overtime hours of production workers in manufacturing (x_1), number of job vacancies in manufacturing (x_2), and index of help wanted advertising in newspapers (x_3). Monthly observations on agency revenues and the three independent variables were obtained for the past 25 months. The ANOVA table for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ is as follows:

Source	d.f.	SS	MS
Regression	3	5409.89	1803.30
Error	21	16.35	0.78
Total	24	5426.24	

The consultant decided to screen the independent variables to determine the best set for predicting agency revenues. The regression and the error sums of squares for all possible regression models were found to be as follows:

Independent variables in the model	R^2	MSE	$d.f._{SSE}$
x_1	0.5474582	106.76522	23
x_2	0.6735511	77.016957	23
x_3	0.6605937	80.073913	23
x_1, x_2	0.9442634	13.747273	22
x_1, x_3	0.9969315	0.7568181	22
x_2, x_3	0.6894829	76.588182	22
x_1, x_2, x_3	0.9969868	0.7785714	21

[6] (a) Determine the subset of variables that is selected as best using **max R^2 criterion**. Show your steps.

(1) $R^2 = \frac{SSR}{TSS}$, **the set $\{X_1, X_3\}$ is selected as the best one. (Please note that the full model gives the highest R^2, however we prefer the second highest one, other than the full model).** (1)

[5] (b) Determine the subset of variables that is selected as best using **min MSE criterion**. Show your steps.

The best model is determined by the set $\{X_1, X_3\}$ (since the *min MSE* and *max R^2* are equivalent).

[10] (c) Determine the subset of variables that is selected as best using **Mallows C_p criterion**. Show your steps.

Full model has 4 parameters (including β_0), we will select as the best model the one whose C_p is as close to p as possible.

$$C_p = \frac{SSE_p}{MSE(X_1, X_2, X_3)} - (n - 2p) \quad , n = 25 \text{ (since } d.f._{TSS} = (n\text{-}1) = 24)$$

- **when p = 2 (i.e. one-variable models):**

for X_1: $C_p = \frac{SSE(X_1)}{MSE(X_1, X_2, X_3)} - (25 - 2(2)) = \frac{2455.60}{0.7785714} - 21 = \underline{\mathbf{3132.9818}}$

for X_2: $C_p = \frac{SSE(X_2)}{MSE(X_1, X_2, X_3)} - (25 - 2(2)) = \frac{1771.39}{0.7785714} - 21 = \underline{\mathbf{2254.1799}}$

for X_3: $C_p = \frac{SSE(X_3)}{MSE(X_1, X_2, X_3)} - (25 - 2(2)) = \frac{1841.70}{0.7785714} - 21 = \underline{\mathbf{2344.4863}}$

- **when p = 3 (i.e. two-variable models):**

for X_1, X_2: $C_p = \frac{SSE(X_1, X_2)}{MSE(X_1, X_2, X_3)} - (25 - 2(3)) = \frac{302.44}{0.7785714} - 19 = \underline{\mathbf{369.45506}}$

for X_1, X_3: $C_p = \frac{SSE(X_1, X_3)}{MSE(X_1, X_2, X_3)} - (25 - 2(3)) = \frac{16.65}{0.7785714} - 19 = \underline{\mathbf{2.3853219}}$

for X_2, X_3: $C_p = \frac{SSE(X_2, X_3)}{MSE(X_1, X_2, X_3)} - (25 - 2(3)) = \frac{1684.94}{0.7785714} - 19 = \underline{\mathbf{2145.1432}}$

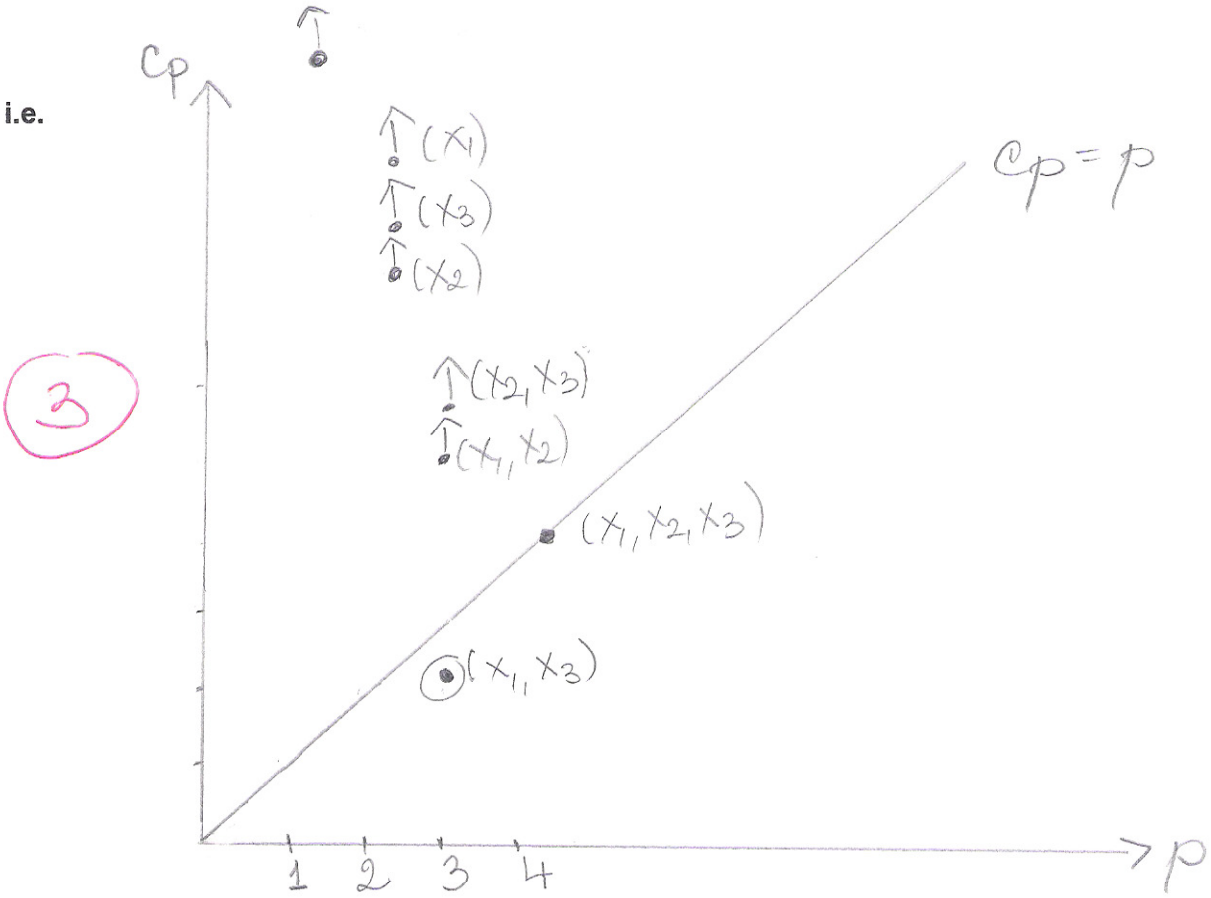
- **when p = 4 (i.e. three-variable model, i.e. the full model):**

for X_1, X_2, X_3: $C_p = \frac{SSE(X_1, X_2, X_3)}{MSE(X_1, X_2, X_3)} - (25 - 2(4)) = \frac{16.35}{0.7785714} - 17 = \underline{\mathbf{4}}$

- **when p = 1 (i.e. no variables in the model, only β_0):**

$$C_p = \frac{TSS}{MSE(X_1, X_2, X_3)} - (25 - 2(1)) = \frac{5426.24}{0.7785714} - 23 = \underline{\mathbf{6946.4828}}$$

i.e.



∴ **the best set is given by $\{X_1, X_3\}$, since its C_p is closest to p (other than the full model). However, since in this case the full model's C_p is exactly equal to p, we may consider the full model as the best model, as well.**

2. Peak blood level data (in mg/ml) were obtained for 20 patients for a single dose of a drug product. In addition to the peak blood level, the patient's weight (in lbs) and the amount of drug (in mg) were recorded.

Blood (y)	Dose(x_1)	Weight(x_2)	Blood (y)	Dose(x_1)	Weight(x_2)
300	1	120	270	4	190
250	1	135	240	4	195
210	1	150	340	8	150
150	1	128	330	8	160
210	2	150	180	8	200
230	2	160	320	8	140
350	2	135	270	16	195
270	2	180	290	16	170
380	4	132	315	16	161
330	4	148	350	16	145

[6] (a) State all assumptions which are necessary for the statistical inference.

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, **$n = 20$**

Assumptions: (i) X_1, X_2 are observed without error

(ii) y's (or ε's) are independently distributed with mean $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (or $E(\varepsilon) = 0$)

(iii) variance of y's (or ε's) is constant, σ^2 for all X_1, X_2

(iv) $y \sim N(E(y), \sigma^2)$ for any value of X_1, X_2 (or $\varepsilon \sim N(0, \sigma^2)$ for any value of X_1, X_2).

[8] (b) Use matrices to compute the estimates of the population parameters β_0, β_1, β_2 and hence obtain the fitted least squares prediction line.

$$(X^TX)^{-1}X^TY = \begin{bmatrix} 2.3713723973 & 0.0122919033 & -0.015251795 \\ 0.0122919033 & 0.0019086088 & -0.000153469 \\ -0.015251795 & -0.000153469 & 0.0001030744 \end{bmatrix} * \begin{bmatrix} 5585 \\ 36870 \\ 869715 \end{bmatrix} =$$

$$= \begin{bmatrix} 432.6024 \\ 5.546395 \\ -1.19433 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \hat{\beta}$$

$\therefore$ **the least squares fitted regression line is given by: $\hat{Y} = X\hat{\beta}$, i.e.**

$\hat{y}$ **= 432.6024 + 5.546395 x_1 -1.19433 x_2**

[21] (c) Set up the ANOVA table and hence test for the significance of the model. Use $\alpha = 0.10$.

$$TSS = Y^TY - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 1633325 - \frac{(5585)^2}{20}$$ **= 1 633 325 – 1 559 611=**

= 73 713.75

$$SSR = \hat{\beta}^T (X^T Y) - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \begin{bmatrix} 432.6024 & 5.546395 & -1.19433 \end{bmatrix} * \begin{bmatrix} 5585 \\ 36870 \\ 869715 \end{bmatrix} - \frac{(5585)^2}{20} =$$

= 1 581 857 – 1 559 611 = 22 246.24

$$SSE = TSS - SSR = \mathbf{51\ 467.51}$$

$$MSR = \frac{SSR}{k} = \frac{22246.24}{2} = \mathbf{11\ 123.12}$$

$$MSE = \frac{SSE}{n-(k+1)} = \frac{51467.51}{17} = \mathbf{3\ 027.5}$$

$$F = \frac{MSR}{MSE} = \mathbf{3.674028}$$

Source	*d.f.*	*SS*	MS	*F*
Regression	**2**	**22 246.24**	**11 123.12**	**3.674028**
Error	**17**	**51 467.51**	**3 027.5**	
Total	**19**	**73 713.75**		

$H_0 : \beta_1 = \beta_2 = 0$ $\quad \alpha = 0.10$

H_a : *at least one of the* $\beta's \neq 0$

test-statistics: $F = \frac{MSR}{MSE} = \mathbf{3.674028}$

R.R: **we reject** H_0 **if** $F > F_{\alpha(k,n-(k+1))} = F_{0.10(2,17)} = \mathbf{2.64}$

Since *F* = 3.67 > 2.64, we reject H_0 and conclude that at 10% level of significance there is an evidence to say that a linear relationship between the blood level and at least one of the dose and weight exists. I.e. The MLR model is significant.

[7] (d) Test whether x_2 term (i.e. whether the patient's weight) contributes to the given model. Use t-test with α = 0.10.

$H_0 : \beta_2 = 0$ $\quad \alpha = 0.10 \Rightarrow \alpha/2 = 0.05$

$H_a : \beta_2 \neq 0$

test-statistics: $t = \frac{\hat{\beta}_2}{\sqrt{v_{22} MSE}} = \frac{-1.19433}{\sqrt{(0.0001030744)(3027.5)}} = \mathbf{-2.13799}$

↑ last diagonal element of $(X^T X)^{-1}$

R.R: we reject H_0 if $t < -t_{\alpha/2;n-(k+1)} = -t_{0.05;17} =$ **- 1.740**

or $t > t_{\alpha/2;n-(k+1)} = t_{0.05;17} =$ **1.740**

Since t = -2.13799 < -1.740, we reject H_0 and conclude that at 10% level of significance there is an evidence to say that the X_2 term (i.e. the patient's weight) contributes to the model.

[6] (e) Find the values of the coefficient of determination, r^2, and the adjusted r^2 and interpret their meanings in this problem.

$$r^2 = \frac{SSR}{TSS} = \frac{22246.24}{73713.75} = 0.301792 \cong \underline{30.2\%}$$

i.e. approximately 30.2% of the total variation in the data is explained by the regr. line (and 69.8% of the variation is due to error).

$$r^2_{adj} = 1 - \frac{SSE/n-(k+1)}{TSS/n-1} = 1 - \frac{MSE}{TSS/n-1} = 1 - \frac{3027.5}{73713.75/19} = 1 - 0.780349663 = 0.21965 \cong \underline{22\%}$$

Since r^2_{adj} is approx. 22% (very low) we can conclude that the full model is not very good. Probably some other terms are needed (or maybe X_1 is not needed).

3. A linear model relating y (number of vehicle sales per month) to independent variables x_1 (price per gallon), x_2 (interest rate) and x_3 is

[4] $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \varepsilon,$$

where the independent variable (vehicle model) is defined by the dummy variable

$$x_3 = \begin{cases} 1, & if \quad standard \\ 0, & if \quad luxury \end{cases}$$

Interpret the meanings of the parameters β_3 and β_5 .

if standard model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(1) + \beta_4 x_1(1) + \beta_5 x_2(1) + \varepsilon$

or $y = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) x_1 + (\beta_2 + \beta_5) x_2 + \varepsilon$

if luxury model: $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3(0)+\beta_4x_1(0)+\beta_5x_2(0)+\varepsilon$

or $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$ (1)

$\therefore\ \beta_3=(\beta_0+\beta_3)-\beta_0=$ **difference in y-intercepts between the lines for standard and luxury vehicle models** (1)

$\beta_5=(\beta_2+\beta_5)-\beta_2=$ **difference in partial slopes of the lines for standard and luxury vehicle models, holding x_1 constant** (1)

4. In a study of grade school children, ages (x_1), heights (x_2), weights (x_3), and scores(y) on a physical fitness test were recorded.

[5] (a) Find the multiple linear regression equation relating the scores to the ages, heights and weights of the children using SAS output (next page).

(1) (1) (1) (1)

$$\hat{y}=-146.886525+4.640368x_1+4.173852x_2-0.475050x_3$$ (1)

where X_1 = age
X_2 = height
X_3 = weight
Y = scores

[2] (b) Based on the SAS output, define your full model and your reduced model.

full model: $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$ (1)

reduced model: $y=\beta_0+\beta_2x_2+\beta_3x_3+\varepsilon$ (1)

[10] (c) Test whether the age contributes to the model. Use SAS output provided on the next page. Use α = 0.05. (Use partial F-test).

$H_0:\beta_1=0$
$H_a:\beta_1\neq0$ (1)

$\alpha=0.05$

SSR_f = 3602.40747 (d.f. = 3)
SSE_f = 4360.39253 (d.f. = 16)

SSR_r = 3446.34060 (d.f. = 2)
SSE_r = 4516.45940 (d.f. = 17)

test-statistics :

$$F_{part} = \frac{[SSR_f - SSR_r]/[df_{SSR_f} - df_{SSR_r}]}{SSE_f / df_{SSE_f}} = \frac{(3602.40747 - 3446.34060)/(3-2)}{4360.39253/16} =$$

$$= \frac{156.0669/1}{272.5245} = \underline{\mathbf{0.572671}}$$

or equivalently,

$$F_{drop} = \frac{[SSE_r - SSE_f]/[df_{SSE_r} - df_{SSE_f}]}{SSE_f / df_{SSE_f}} = \frac{(4516.45940 - 4360.39253)/(17-16)}{4360.39253/16} =$$

$$= \frac{156.0669/1}{272.5245} = \underline{\mathbf{0.572671}}$$

R.R: **we reject H_0 if $F_{part} > F_{\alpha(1,16)} = F_{0.05(1,16)} = 4.49$**

Since F_{part} = 0.572671 < 4.49, we do not reject H_0 and conclude that at 5% level of significance there is not enough evidence to say that the X_1 term (i.e. the age) contributes to the score of children.