//90

STAT 2509C
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:

(4) (4)

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) (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).**

[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.

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

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

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

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

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

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

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

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

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

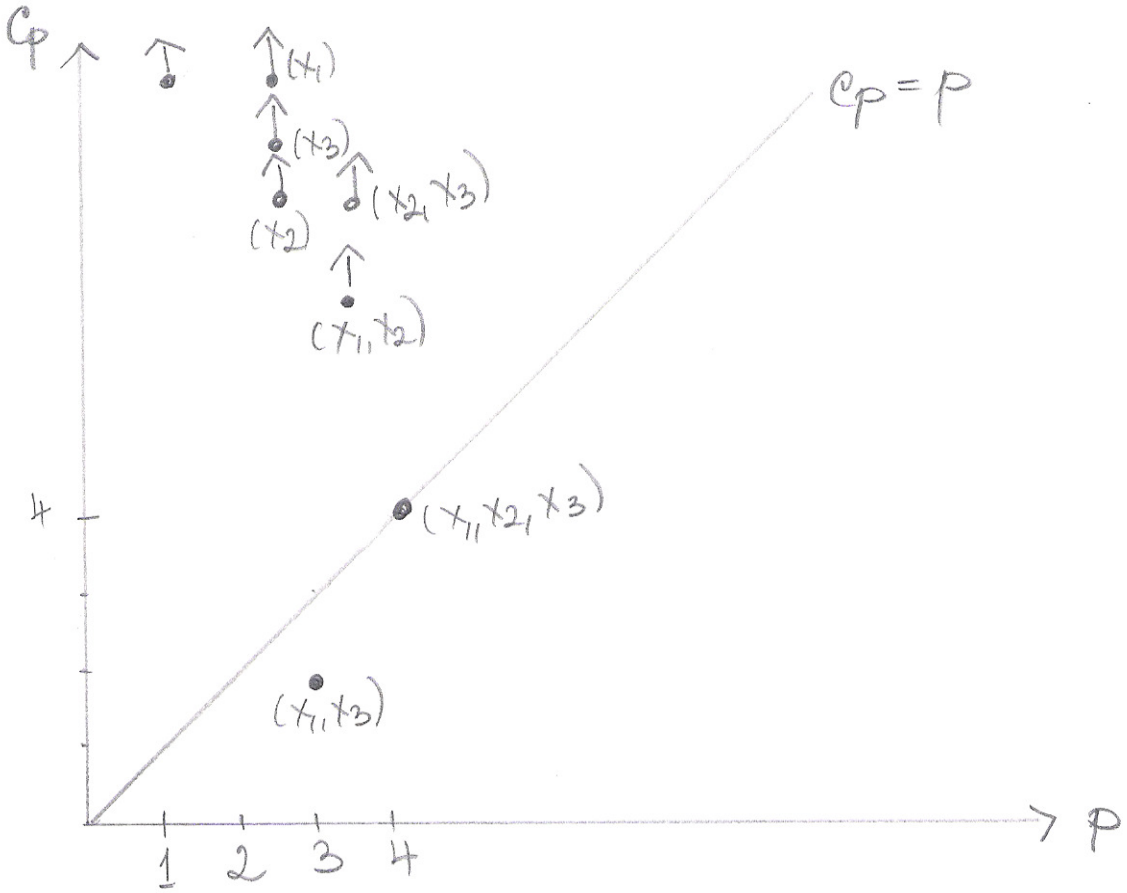
- **when p = 4 (i.e. three-variable model, i.e. the full model):**

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

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

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

i.e.



3

1

∴ 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. The personnel department of a large industrial corporation would like to develop a model to predict the weekly salary (in dollars) based upon the length of employment (in months) and age (in years) of its managerial employees. A random sample of 16 managerial employees is selected with the results displayed below:

Employee	Weekly salary (y)	Length of employ. (x_1)	Age (x_2)	Employee	Weekly salary (y)	Length of employ. (x_1)	Age (x_2)
1	839	330	46	9	752	352	55
2	946	569	65	10	729	256	61
3	870	375	57	11	656	87	28
4	718	113	47	12	874	337	51
5	802	215	41	13	606	42	28
6	812	343	59	14	729	129	37
7	748	252	45	15	728	216	46
8	791	348	57	16	792	327	56

[6] (a) Write down a multiple linear regression model and state all assumptions which are necessary for the statistical inference.

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, **$n$ = 16**

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.

$$(X^TX)^{-1}X^TY = \begin{bmatrix} 2.105742517 & 0.00325754930 & -0.0599101724 \\ 0.00325754930 & 0.00001401163 & -0.0001440882 \\ -0.0599101724 & -0.0001440882 & 0.0020241913 \end{bmatrix} * \begin{bmatrix} 12392 \\ 3475901 \\ 613509 \end{bmatrix} =$$

$$= \begin{bmatrix} 661.8502 \\ 0.671182 \\ -1.38359 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \hat{\beta}$$

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

$\hat{y}$ **= 661.8502 + 0.671182** x_1 **– 1.38359** x_2

[21] (c) Set up the ANOVA table and hence test for the significance of the model. Use $\alpha = 0.05$.

$$TSS = Y^TY - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 9706076 - \frac{(12392)^2}{16} = \textbf{9 706 076 – 9 597 604 =}$$

=108 472

$$SSR = \hat{\beta}^T (X^T Y) - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = [661.8502 \quad 0.671182 \quad -1.38359] * \begin{bmatrix} 12392 \\ 3475901 \\ 613509 \end{bmatrix} - \frac{(12392)^2}{16} =$$

= 9 685 762 – 9 597 604 = 88 158.46

$$SSE = TSS - SSR = \mathbf{20\ 313.54}$$

$$MSR = \frac{SSR}{k} = \frac{88158.46}{2} = \mathbf{44\ 079.23}$$

$$MSE = \frac{SSE}{n-(k+1)} = \frac{20313.54}{13} = \mathbf{1\ 562.58}$$

$$F = \frac{MSR}{MSE} = \mathbf{28.20927}$$

Source	*d.f.*	*SS*	*MS*	*F*
Regression	**2**	**88 158.46**	**44 079.23**	**28.2093**
Error	**13**	**20 313.54**	**1 562.58**	
Total	**15**	**108 472**		

$H_0 : \beta_1 = \beta_2 = 0$ $\quad \alpha = 0.05$

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

test-statistics: $F = \frac{MSR}{MSE} = \mathbf{28.2093}$

R.R: **we reject** H_0 **if** $F > F_{\alpha(k,n-(k+1))} = F_{0.05(2,13)} = \mathbf{3.81}$

Since *F* = 28.2093 > 3.81, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that a linear relationship between *y* and at least one of the *x*'s exists (i.e. linear model is significant).

[7] (d) Test whether x_2 term (i.e. whether the age) contributes to the given model. Use t-test with $\alpha = 0.05$.

$H_0 : \beta_2 = 0$ $\quad \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$H_a : \beta_2 \neq 0$

test-statistics: $t = \frac{\hat{\beta}_2}{\sqrt{v_{22} MSE}} = \frac{-1.38359}{\sqrt{(0.0020241913)(1562.58)}} = \mathbf{-0.77797}$

↑ 3rd diagonal element of $(X^T X)^{-1}$

R.R: we reject H_0 if $t < -t_{\alpha/2;n-(k+1)} = -t_{0.025;13} = -\ 2.160$

or $t > t_{\alpha/2;n-(k+1)} = t_{0.025;13} = 2.160$

Since $t = -0.77797 \not< -2.160$, we do not reject H_0 and conclude that at 5% level of significance there is not enough evidence to say that the X_2 term (i.e. the age) 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.

$$r^2 = \frac{SSR}{TSS} = \frac{88158.46}{108472} = 0.81273 \cong \underline{81.30\%}$$

i.e. approximately 81.30% of the total variation in the data is explained by the regression line (and 18.70% is due to error).

$$r^2_{adj} = 1 - \frac{SSE/n-(k+1)}{TSS/n-1} = 1 - \frac{MSE}{TSS/n-1} = 1 - \frac{1562.58}{108472/15} = 1 - 0.216081 = 0.783919 \cong \underline{78.40\%}$$

Since r^2_{adj} is 78.40% (that is slightly lower than r^2) and since we know that X_2 is not needed in the model, we may conclude that the model is not very good.

3. A linear model relating y to independent variables x_1 and x_2 is

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

where the independent variable truck type is defined by the dummy variable

$$x_2 = \begin{cases} 1, & \text{if type } A \\ 0, & \text{if type } B \end{cases}$$

Interpret the meanings of the parameters β_2 and β_3 .

if type A: $y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \varepsilon$

or $y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1 + \varepsilon$ (1)

if type B: $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \varepsilon$

or $y = \beta_0 + \beta_1 x_1 + \varepsilon$ (1)

$\therefore\ \beta_2 = (\beta_0 + \beta_2) - \beta_0 =$ **difference in y-intercepts between the lines for type A & B** (1)

$\beta_3 = (\beta_1 + \beta_3) - \beta_1 =$ **difference in slopes of the lines for type A & B** (1)

4. When studying the amount of heat evolved during curing in calories per gram of cement, the following variables were considered: (x_1) tricalcium aluminate, (x_2) tricalcium silicate, (x_3) tetracalcium alumino ferrite and (x_4) dicalcium silicate in percent weight of the clinkers from which the cement was made.

[5] (a) Find the multiple linear regression equation relating the amount of heat evolved during curing of cement to the independent variables x_1, x_2, x_3 and x_4 using SAS output (next page).

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

$$\hat{y} = 62.405369 + 1.551103x_1 + 0.510168x_2 + 0.101909x_3 - 0.144061x_4$$

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

full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ (1)

reduced model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ (1)

[10] (c) Test whether variables x_3 and x_4 contribute to the model. Use SAS output provided on the next page. Use α = 0.05. (Use partial F-test).

$H_0 : \beta_3 = \beta_4 = 0$ $\quad \alpha = 0.05$

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

SSR_f = 2 667.89944 (d.f. = 4)
SSE_f = 47.86364 (d.f. = 8)

SSR_r = 2 657.85859 (d.f. = 2)
SSE_r = 57.90448 (d.f. = 10)

test-statistics :

$$F_{part} = \frac{[SSR_f - SSR_r]/[df_{SSR_f} - df_{SSR_r}]}{SSE_f / df_{SSE_f}} = \frac{(2667.89944 - 2657.85859)/(4-2)}{47.86364/8} =$$

$$= \frac{10.04085/2}{47.86364/8} = \frac{5.020425}{5.982955} = \underline{\mathbf{0.839121}}$$

or equivalently,

$$F_{drop} = \frac{[SSE_r - SSE_f]/[df_{SSE_r} - df_{SSE_f}]}{SSE_f / df_{SSE_f}} = \frac{(57.90448 - 47.86364)/(10-8)}{47.86364/8} =$$

$$= \frac{10.04085/2}{47.86364/8} = \frac{5.020425}{5.982955} = \underline{\mathbf{0.839121}}$$

R.R: **we reject H_0 if $F_{part} > F_{\alpha(2,8)} = F_{0.05(2,8)} =$ 4.46**

Since F_{part} = 0.839121 $\ngtr$ 4.46, we do not reject H_0 and conclude that at 5% level of significance there is not enough evidence to say that variables x_3 and x_4 contribute to the model.