Carleton University – School of Mathematics and Statistics
STAT 2509 – Test 2 – **SOLUTION**

/30

1. [22.5 marks]

[3.5] (a) $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ $\alpha = 0.10$
H_a : *at least one of the* $\beta' s \neq 0$ **[1]**

test-statistics: F = 26.510

R.R. we reject H_0 if *p-value* < α **[1]** (or if $F > F_{\alpha;(k,n-(k+1))} = F_{0.10;(7,12)} = 2.28$)

Since *p-value* < 0.001 < 0.10 **[1/2]** (or F = 26.51 > 2.28), we do reject H_0 **[1/2]** and conclude that at 10% level of significance we have enough evidence to conclude that the model is useful, i.e. it can be used. **[1/2]**

[6] (b) $H_0: \beta_5 = \beta_6 = \beta_7 = 0$, **[1]** $\alpha = 0.10$
$H_a: \text{at least one } \beta \neq 0$

Full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_4 + \beta_6 x_2 x_4 + \beta_7 x_3 x_4 + \varepsilon$

Reduced Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ **[1/2]** if correct reduced model was used

test-statistics: **[1/2]** $F_{drop} = \frac{[SSE_r - SSE_f]/[df_{SSE_r} - df_{SSE_f}]}{MSE_f} = \frac{[SSE_r - SSE_f]/[n-5-(n-8)]}{SSE_f / n-8} =$

$$= \frac{(16522.327 - 12873.373)/(15-12)}{12873.373/12} = \frac{3648.954/3}{12873.373/12} = \frac{1216.318}{1072.781} = 1.133798977$$ **[1/2]**

[1/2] [1/2] (1/2 mark for each correct d.f.)

or equivalently

$$F_{part} = \frac{[SSR_f - SSR_r]/[df_{SSR_f} - df_{SSR_r}]}{MSE_f} = \frac{(199077.177 - 195428.223]/(7-4)}{12873.373/12}$$
$$= \frac{3648.954/3}{12873.373/12} = \frac{1216.318}{1072.781} = 1.133798977$$

R.R. we reject H_0 **if** F_{drop} (or F_{part}) $> F_{\alpha;(3,n-8)} = F_{0.10;(3,12)} = 2.61$ **[1]**

- **Since** F_{drop} (or F_{part}) $= 1.3379 \not> 2.61$ **[1/2], we do not reject** H_0 **[1/2] and conclude that at 10% level of significance there is not enough evidence that the interaction terms are needed. [1/2]**

[12] (c) Model we are using is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

- x_1 **(undergraduate degree GPA):** $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$ } $\alpha = 0.10$ **[1]**

t-test: t = 9.736
R.R.: we reject *Ho* if p-value < α **[1/2]** (or if $|t| > t_{\alpha/2;n-(k+1)} = t_{0.05;15} = 1.753$)

Since p-value < 0.001 < 0.10 **[1/2]** (or t = 9.736 > 1.753), we reject H_0 **[1/2]** and conclude that 'undergraduate degree GPA' affects the 'score on the entrance test'. **[1/2]**

- x_2 **(age):** $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$ } $\alpha = 0.10$ **[1]**

t-test: t = 0.210
R.R.: we reject *Ho* if p-value < α **[1/2]** (or if $|t| > t_{\alpha/2;n-(k+1)} = t_{0.05;15} = 1.753$)

Since p-value = 0.836 > 0.10 **[1/2]** (or t = 0.210 $\ngtr$ 1.753), we do not reject H_0 **[1/2]** and conclude that 'age' does not affect the 'score on the entrance test'. **[1/2]**

- x_3 **(years of volunteer experience in a health field):** $H_0 : \beta_3 = 0$, $H_a : \beta_3 \neq 0$ } $\alpha = 0.10$ **[1]**

t-test: t = 0.496
R.R.: we reject *Ho* if p-value < α **[1/2]** (or if $|t| > t_{\alpha/2;n-(k+1)} = t_{0.05;15} = 1.753$)

Since p-value = 0.627 > 0.10 **[1/2]** (or t = 0.496 $\ngtr$ 1.753), we do not reject H_0 **[1/2]** and conclude that 'years of volunteering' does not affect the 'score on the entrance test'. **[1/2]**

- x_4 **(undergraduate degree in health field):** $H_0 : \beta_4 = 0$, $H_a : \beta_4 \neq 0$ } $\alpha = 0.10$ **[1]**

t-test: t = 4.339
R.R.: we reject *Ho* if p-value < α **[1/2]** (or if $|t| > t_{\alpha/2;n-(k+1)} = t_{0.05;15} = 1.753$)

Since p-value < 0.001 < 0.10 **[1/2]** (or t = 4.339 > 1.753), we reject H_0 **[1/2]** and conclude that 'undergraduate degree in health field' affects the 'score on the entrance test'. **[1/2]**

[1] (d) the best model is: $\underline{\underline{y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \varepsilon}}$ **[1]**

2. [4.5 marks] Refers to Question 1.

Independent variables *in the model*	*SSR*	*SSE*	*d.f.* $_{SSE}$	*MSE*	R^2	*Cp*
no *X*'s						174.422
X_1	173136.755	38813.795	18	2156.322	0.8170	19.2376
X_2	35749.465	176201.085	18	9788.949	0.1690	143.9664
X_3	790.594	211159.956	18	11731.109	0.0040	175.7043
X_4	69031.250	142919.300	18	7939.961	0.3260	113.7511
X_1, X_2	173311.462	38639.088	17	2272.888	0.8180	21.07899
X_1, X_3	174492.432	37458.118	17	2203.419	0.8230	20.00683
X_1, X_4	195114.282	16836.268	17	990.369	0.9210	1.285022
X_2, X_3	36856.436	175094.114	17	10299.654	0.1740	144.9614
X_2, X_4	91023.541	120927.009	17	7113.353	0.4290	95.78514
X_3, X_4	69060.555	142889.995	17	8405.294	0.3260	115.7245
X_1, X_2, X_3	174687.460	37263.090	16	2328.943	0.8240	21.82977
X_1, X_2, X_4	195157.003	16793.547	16	1049.597	0.9210	3.246237
X_1, X_3, X_4	195379.568	16570.982	16	1035.686	0.9220	3.044178
X_2, X_3, X_4	91026.972	120923.578	16	7557.724	0.4290	97.78202
X_1, X_2, X_3, X_4	195428.223	16522.327	15	1101.488	0.9220	5.000006

[1] (a)

Using $\max R^2$, the set [1/2] $\{X_1, X_3, X_4\}$ (or $\{X_1, X_2, X_4\}$) is selected as the best one. But because the set $\{X_1, X_4\}$ has R^2 very close to 0.9220 (0.9210) and only has 2 variables, we could also select the model with X1 and X4. (Please note that the full model gives the highest R^2, however we prefer the second highest [1/2] one, other than the full model).

Note: we would accept either of the 2 models as the best model

[1] (b)

The best model is determined by the set [1/2] $\{X_1, X_4\}$ (since the *min MSE* and *max R^2* should give the same answer). [1/2]

[2.5] (c) Determine the subset of variables that is selected as best using **Mallows C_p criterion**. Give reason for your answer.

- **We will select as the best-fitting model, the one with the smallest $|C_p - p|$.**
- **Hence,**

Independent variables in the model	*p*	*Cp*	$\lvert C_p - p \rvert$
no *X*'s	1	174.422	173.422
X_1	2	19.2376	17.2376
X_2	2	143.9664	140.9664
X_3	2	175.7043	173.7043
X_4	2	113.7511	111.7511
X_1, X_2	3	21.07899	18.07899
X_1, X_3	3	20.00683	17.00683
X_1, X_4	3	1.285022	1.714978
X_2, X_3	3	144.9614	141.9614
X_2, X_4	3	95.78514	92.78514
X_3, X_4	3	115.7245	112.7245
X_1, X_2, X_3	4	21.82977	17.82977
X_1, X_2, X_4	4	3.246237	0.753763
X_1, X_3, X_4	4	3.044178	0.955822
X_2, X_3, X_4	4	97.78202	93.78202
X_1, X_2, X_3, X_4	5	5.000006	0.000006

[1] mark if $\lvert C_p - p \rvert$ or graph were used

or using graph

The best model is determined by the set [1/2] $\{X_1, X_2, X_4\}$ (since the *Cp is closest to p* [1/2]), where *p* =4 [1/2] (other than a full model).

3. [2 marks]

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

[2] where

x_1 = distance between locations, $x_2 = \begin{cases} 1, & \text{if the vehicle is a truck} \\ 0, & \text{if the vehicle is a car} \end{cases}$

<u>if truck:</u> $y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \varepsilon$,

or $y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon$ **[1/2]**

<u>if car:</u> $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \varepsilon$,

or $y = \beta_0 + \beta_1 x_1 + \varepsilon$ **[1/2]**

$\beta_2 = (\beta_0 + \beta_2) - \beta_0$ = difference in y-intercepts between the lines for truck and car models **[1/2]**

$\beta_3 = (\beta_1 + \beta_3) - \beta_1$ = difference in slopes of the lines for truck and car models **[1/2]**

4. [1 mark]

MSR will be an unbiased estimator of σ^2, if $E(MSR) = \sigma^2$, i.e. if $H_0 : \beta_1 = \beta_2 = 0$ [1].