MATH 244 (L2) Applied Statistics

Quiz 4

Name _____ Student ID _____ Tutorial section ____

Time allowed : 45 minutes

1. (*12 marks*) Subjects in a study of the impact of child abuse on IQ are selected from families receiving public assistance. They are classified into 4 groups according to the abuse level. The summary statistics of their IQ scores are given in the following table.

Abuse Level	Sample size	Mean	Standard Deviation
1	16	87.81	3.7
2	16	80.06	4.1
3	16	79.56	3.9
4	16	74.31	4.4

(a) Construct the ANOVA table and test whether there is treatment effect. Use $\alpha = 0.05$.

$$\bar{\bar{Y}} = \frac{16\times 87.81+16\times 80.06+16\times 79.56+16\times 74.31}{64} = 80.435$$

$$SS_A = \sum_{i=1}^{4} n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 = 16\left[(87.81-80.435)^2 + \cdots + (74.31-80.435)^2\right] = 1485$$

$$SS_E = \sum_{i=1}^{4} (n_i - 1)S_i^2 = 15\left(3.7^2 + 4.1^2 + 3.9^2 + 4.4^2\right) = 976.05$$

ANOVA table :

Source	SS	d.f.	MS	F-ratio
Treatment	1485	3	495	30.43
Error	976.05	60	16.2675	
Total	2461.05	63		

Test $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ vs H_1 : not H_0.

$F(3,60,0.05) = 2.76 < 30.43$

Reject H_0 at $\alpha = 0.05$.

(b) Find the 90% confidence interval for each of the pairwise differences between the four treatment effects.

90% C.I. for $\alpha_1 - \alpha_2$:
$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{60,0.05}\sqrt{MS_E\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$
$$= (87.81 - 80.06) \pm (1.645)\sqrt{(16.2675)\left(\frac{1}{16}+\frac{1}{16}\right)}$$
$$= 7.75 \pm 2.3457 = [5.4043, 10.0957] \qquad (\alpha_1 > \alpha_2)$$

90% C.I. for $\alpha_1 - \alpha_3$:
$$(87.81 - 79.56) \pm 2.3457$$
$$= 8.25 \pm 2.3457 = [5.9043, 10.5957] \qquad (\alpha_1 > \alpha_3)$$

90% C.I. for $\alpha_1 - \alpha_4$:
$$(87.81 - 74.31) \pm 2.3457$$
$$= 13.5 \pm 2.3457 = [11.1543, 15.8457] \qquad (\alpha_1 > \alpha_4)$$

90% C.I. for $\alpha_2 - \alpha_3$:
$$(80.06 - 79.56) \pm 2.3457$$
$$= 0.5 \pm 2.3457 = [-1.8457, 2.8457] \qquad \text{(not significant)}$$

90% C.I. for $\alpha_2 - \alpha_4$:
$$(80.06 - 74.31) \pm 2.3457$$
$$= 5.75 \pm 2.3457 = [3.4043, 8.0957] \qquad (\alpha_2 > \alpha_4)$$

90% C.I. for $\alpha_3 - \alpha_4$:
$$(79.56 - 74.31) \pm 2.3457$$
$$= 5.25 \pm 2.3457 = [2.9043, 7.5957] \qquad (\alpha_3 > \alpha_4)$$

(c) What conclusion can be drawn from the results in part (b)? Do you have 90% confidence that this conclusion is correct? Explain briefly.

We can conclude that $\alpha_1 > (\alpha_2 \ \alpha_3) > \alpha_4$, i.e. the mean IQ score decreases with the increase in abuse level. However, the mean IQ scores of the children in second group and third group are not significantly different.

This conclusion was drawn based on the six comparisons. For each of these comparisons we have only 90% confidence that it is correct. Hence the overall confidence that all these comparisons are correct is less than 90%. Therefore the overall confidence of this conclusion is less than 90%.

(d) After doing the above analyses, the researcher had written down the following conclusion in his report: *"The low IQ scores of these children were caused by child abuse. Higher level of abuse will definitely result in lower IQ of a child."* Comment on his statements.

The researcher was doing a sample survey rather than an experiment. The children were classified they to the four groups according to the abuse level. Hence the researcher had no control on the abuse level of the subjects. Therefore although we found some relationship between abuse level and IQ score, we cannot conclude that this is a causal relationship. Maybe low IQ score is the cause of abuse, or maybe there are other causes affecting both the abuse level and IQ score. Hence he had over-interpreted his findings in his statements.

2. (*13 marks*) Suppose that a random sample of 8 families had the following annual incomes and savings tabulated below.

Family	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Income (in $10000)	22	19	17	27	24	34	16	29
Saving (in $10000)	2.1	2.0	1.6	3.3	2.5	3.8	1.3	3.4

(a) Fit a regression line of saving on income. Please show all your steps.

$$\overline{X} = 23.5, \quad \overline{Y} = 2.5, \quad \sum_{i=1}^{8} X_i^2 = 4692, \quad \sum_{i=1}^{8} Y_i^2 = 55.8, \quad \sum_{i=1}^{8} X_i Y_i = 509.1$$

$$S_{xx} = \sum_{i=1}^{8} X_i^2 - n\overline{X}^2 = 4692 - (8)(23.5)^2 = 274$$

$$S_{yy} = \sum_{i=1}^{8} Y_i^2 - n\overline{Y}^2 = 55.8 - (8)(2.5)^2 = 5.8$$

$$S_{xy} = \sum_{i=1}^{8} X_i Y_i - n\overline{X}\overline{Y} = 509.1 - (8)(23.5)(2.5) = 39.1$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{39.1}{274} = 0.1427 \quad , \quad a = \overline{Y} - b\overline{X} = 2.5 - (0.1427)(23.5) = -0.8535$$

Fitted regression line : $\hat{Y} = -0.8535 + 0.1427X$

(b) Construct the ANOVA table.

$$SS_R = \frac{S_{xy}^2}{S_{xx}} = \frac{39.1^2}{274} = 5.5796$$

$$SS_E = S_{yy} - SS_R = 5.8 - 5.5796 = 0.2204$$

ANOVA table :

Source	**SS**	**d.f.**	**MS**	**F-ratio**
Regression	5.5796	1	5.5796	151.91
Error	0.2204	6	0.03673	
Total	5.8	7		

(c) Find the coefficient of determination.

$$R^2 = \frac{SS_R}{SS_T} = \frac{5.5796}{5.8} = 96.2\%$$

(d) Find the 90% prediction interval for the annual saving of a family with annual income \$300000.

$X_0 = 30$, $\widehat{Y}_0 = -0.8535 + (0.1427)(30) = 3.4275$

90% P.I. for Y_0 :

$$\widehat{Y}_0 \pm t_{6,0.05}\sqrt{MS_E\left(1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{S_{xx}}\right)}$$

$$= 3.4275 \pm (1.943)\sqrt{(0.03673)\left(1 + \frac{1}{8} + \frac{(30 - 23.5)^2}{274}\right)}$$

$$= 3.3804 \pm 0.4212 = [2.9592\,,\, 3.8016]$$

(e) Using the regression line obtained in (a), find the prediction of the saving of a family without any income (i.e. annual income = \$0). Is this prediction reasonable? Explain briefly.

$X_0 = 0$, $\widehat{Y}_0 = -0.8535 + (0.1427)(0) = -0.8535$

The predicted saving will be –\$8535. This prediction is not reasonable because $X_0 = 0$ is far beyond the range of the observed values of *X*. The fitted regression line would not be adequate for modelling the relationship between *X* and *Y* outside this range.