Surname: Oyeboade
Name: Ikeoluwa Deborah
Student number:

STK 320: Categorical Data Analysis – Practical Assignment 1

ANOVA

(Completely randomised design)

QUESTION 1:

A botanist wants to determine the effect of microscopic worms on seedling growth. He prepares 16 identical planting pots and then introduces four sets of worm populations into them. There are four groups of pots with four pots in each group. The worm population group sizes are 0 (introduced into the first group of four pots), 500 (introduced into the second group of four pots), 1000 (introduced into the third group of four pots), and 4000 (introduced into the fourth group of four pots). Two weeks after planting, he measures the seedling growth in centimetres. The results are given in the table below.

Group	Growth
1	10.7; 9.0; 13.4; 9.2
2	11.1; 11.1; 8.9; 11.4
3	5.7; 5.1; 7.2; 4.8
4	4.7; 3.2; 6.5; 5.3

1. Identify the dependent as well as the explanatory or independent variable.

Dependent variable = Growth

Independent variable = Planting pots (4 levels)

2. Define the statistical ANOVA model.

The statistical model
E(y)=μ+α_i^A for i=1,2,3,4
Where α_i^A is the effect of the ith level of factor A (planting pot) and $\sum_{i=1}^{4} \alpha_i^A = 0$
And μ is the overall mean growth

3. Define the statistical GLM.

E(y)= β_0+$\beta_1^A x_1^A$+$\beta_2^A x_2^A$+$\beta_3^A x_3^A$

The β_0, β_1^A, β_2^A, β_3^A are the regression coefficients, while x_1^A, x_2^A, x_3^A are dummy variables indicating the presence or absence of an observation from a population.

Where μ=β_0 , α_1^A=β_1^A, α_2^A=β_2^A, α_3^A=β_3^A, α_4^A=$-(\beta_1^A + \beta_2^A + \beta_3^A)$

Surname: Oyeboade
Name: Ikeoluwa Deborah
Student number:

4. Define, in table form, the dummy variables to be used in the GLM analysis by completing the table below.

Group	x_1^A	x_2^A	x_3^A	**Score**
1	1	0	0	10.7
1	1	0	0	9.0
1	1	0	0	13.4
1	1	0	0	9.2
2	0	1	0	11.1
2	0	1	0	11.1
2	0	1	0	8.9
2	0	1	0	11.4
3	0	0	1	5.7
3	0	0	1	5.1
3	0	0	1	7.2
3	0	0	1	4.8
4	-1	-1	-1	4.7
4	-1	-1	-1	3.2
4	-1	-1	-1	6.5
4	-1	-1	-1	5.3

5. Write an appropriate SAS program (using either ANOVA or GLM specification) and create the output to answer the questions below.
   a. Interpret the value of R^2 (coefficient of determination).

The coefficient of determination of 0.816154 indicates that 82% of the variation in growth Can be explained by the type of planting pots used.

   b. Is the model significant? Make a conclusion.

Using hypothesis test

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: $\mu_i's\ are\ not\ all\ equal, at\ least\ 2\ population\ means\ differ$

The value of the test statistic is f=17.76

The p-value is 0.0001 < 0.05

At 5% level of significance, the null hypothesis of equal population means is rejected and we conclude that the mean of the planting pots is not the same i.e. the 4 planting pots are not all equally effective.

c. By using multiple comparisons, determine the best environment in which to study effectively.

Using the SAS output of the Scheffe's test, the group '500 worms' with mean 10.625 and '0 worms' with mean 10.575 are marked to be group A. Thus, we can say that these 2 groups do not differ significantly from each other. Groups '1000 worms' and '4000 worms' have the same scheffe grouping letter 'B'. Thus, we can say that they do not differ from each other.

Therefore, we can say that the best environment for optimal seedling growth is the second group named '500 worms' and the first group named '0 worms'. Since they do not differ significantly from each other.

QUESTION 2:

Somnolent Sam is interested in the effects of sleep deprivation on memory function. He randomly assigns each of 25 participants to one of four groups. Seven subjects take a test of memory function after being awake for eight hours (no sleep deprivation). Seven subjects take the test after they have been awake for 18 hours (mild sleep deprivation). Five subjects take the test after they have been awake for 24 hours (moderate sleep deprivation). Six subjects take the test after they have been awake for 48 hours (severe sleep deprivation). Higher scores on the test of memory function indicate higher levels of performance.

Sam predicts that an overall effect of sleep deprivation on memory function will be observed. He predicts that people with at least some sleep deprivation will perform significantly more poorly than people who have not been sleeping deprived. He also predicts that people who have been deprived of sleep for 24 hours will perform significantly more poorly than those deprived of sleep for 18 hours. He also predicts

Surname: Oyeboade
Name: Ikeoluwa Deborah
Student number:

that people who have been deprived of sleep for 48 hours will perform significantly more poorly than those deprived of sleep for 24 hours.

The subjects' scores are as follows:

Group	**Test scores**
1: No deprivation	25; 19; 18; 21; 24; 25; 21
2: 18 hours deprivation	14; 14; 12; 13; 16; 14; 17
3: 24 hours deprivation	14; 13; 15; 11; 13
4: 48 hours deprivation	7; 9; 6; 11; 5; 10

1. Identify the dependent as well as the explanatory or independent variable.

Dependent variable = Memory function score

Independent variable = Sleep deprivation (4 levels)

2. Define the statistical ANOVA model.

The statistical model
E(y)=μ+α_i^A for i=1,2,3,4
Where α_i^A is the effect of the ith level of factor A (sleep deprived level) and $\sum_{i=1}^{4}\alpha_i^A = 0$
And μ is the overall mean test scores

3. Define the GLM with reference to the dummy variables.

E(y)= β_0+$\beta_1^A x_1^A$+$\beta_2^A x_2^A$+$\beta_3^A x_3^A$

The β_0, β_1^A, β_2^A, β_3^A are the regression coefficients, while x_1^A, x_2^A, x_3^A are dummy variables as presented in the table below

Group	x_1^A	x_2^A	x_3^A
1	1	0	0
2	0	1	0
3	0	0	1
4	-1	-1	-1

4. Write a SAS program and create the output to compare the four GROUPS. (Use the GLM procedure and the Scheffé-statement.)

```
proc format;
value gr 1 = 'No Deprivation'
         2 = '18 Hours Deprivation'
```

Surname: Oyeboade
Name: Ikeoluwa Deborah
Student number:

```
             3 = '24 Hours Deprivation'
             4 = '48 Hours Deprivation'
             ;
data sleep;
input group memory @@;
cards;
1 25 1 19 1 18 1 21 1 24 1 25 1 21
2 14 2 14 2 12 2 13 2 16 2 14 2 17
3 14 3 13 3 15 3 11 3 13
4 7  4 9  4 6  4 11 4 5  4 10
;

proc glm;
class group;
model memory = group;
means group/ scheffe lines cldiff;
format group gr.;
    run;
```

5. Modify your program in Question 2.4 to satisfy the following criteria: Use the same data step, define dummy variables and use the GLM procedure without a class statement.

```
proc format;
 value gr 1 = 'No Deprivation'
             2 = '18 Hours Deprivation'
             3 = '24 Hours Deprivation'
             4 = '48 Hours Deprivation'
             ;
data sleep;
input group memory @@;
x1a=0; x2a=0; x3a=0;
if group=1 then x1a=1;
if group=2 then x2a=1;
if group=3 then x3a=1;
if group=4 then do;
                x1a=-1;
                        x2a=-1;
                        x3a=-1;
end;
cards;
1 25 1 19 1 18 1 21 1 24 1 25 1 21
2 14 2 14 2 12 2 13 2 16 2 14 2 17
3 14 3 13 3 15 3 11 3 13
4 7  4 9  4 6  4 11 4 5  4 10
;
proc glm;
model memory = x1a x2a x3a;
title2 'use proc glm to find model parameters no class statement';
run;
```

6. CONSIDER ONLY THE OUTPUT OF QUESTION 2.4 TO ANSWER THE FOLLOWING QUESTIONS:

   a. Use the critical value approach to test at a 5% level of significance whether the mean scores of the four sleep deprivation groups differ.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_A: \mu_i's\ are\ not\ all\ equal, at\ least\ 2\ population\ means\ differ$

The test statistic is F = 43.48

The critical value is $F_{3,21,0.05}$= 3.07247

At 5% level of significance, the null hypothesis of equal population means is rejected and we conclude that the mean of the four sleep deprivation groups is significantly different.

b. Use the multiple comparisons results to determine if Somnolent Sam's prediction about sleep deprivation was correct.

The no deprivation group has the highest the mean memory score. The 18 and 24 hours sleep deprivation are in the same group, we can say that their average memory score do not differ significantly from each other. The 48-sleep deprivation group has the lowest average memory score.

Sam's predication was quite correct, the part that isn't correct is that the people who have been deprived of sleep for 24 hours will perform significantly more poorly than those deprived of sleep for 18 hours. We see that this is incorrect from the scheffe grouping as they are both in the same group, indicating that they are not significantly different from each other. Apart from that Sam's predication is quite correct.

c. The mean memory scores for each sleep intervention are given in the Scheffé-output. The overall mean is calculated by the unbiased formula for $\bar{y}$ on page 6 of the class notes. Obtain the estimated parameters of the ANOVA model from the computer output (check whether the condition of the model is satisfied). NOTE: the SCORE mean in the output is calculated by the weighted formula on p. 5.

$\hat{\mu}$= $\bar{y}$= $\frac{21.857+14.286+13.200+8.000}{4}$=14.33575

\$\hat{\alpha}_1= 21.857-14.33575 = 7.52125$

\$\hat{\alpha}_2 = 14.286-14.33575 = -0.04975$

\$\hat{\alpha}_3 = 13.2-14.33575 = -1.13575$

\$\hat{\alpha}_4$=8-14.33575= -6.33575

And $\sum_{i=1}^{4} \alpha_i^A$ = $7.52125-0.04975-1.13575 -6.33575 = 0$

Surname: Oyeboade
Name: Ikeoluwa Deborah
Student number:

7. CONSIDER ONLY THE OUTPUT OF QUESTION 2.5 TO ANSWER THE FOLLOWING QUESTIONS:

a. Interpret the value of the coefficient of determination.

The coefficient of determination is 0.861340. This indicates that 86% of the variation in the memory score can be explained by the different levels of sleep deprivation.

b. Is the model significant? Make a conclusion.

Using hypothesis test

$H_0\text{: } \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_A\text{: } \mu_i' s\ are\ not\ all\ equal, at\ least\ 2\ population\ means\ differ$

The value of the test statistic is f= 43.48

The p-value is <0.0001 which is < 0.05

At 5% level of significance, the null hypothesis of equal population means is rejected and we conclude that the mean of the memory scores depends on the number of hours a person is sleep deprived for.

c. Write down the regression coefficients of the GLM.

$\beta_0 = 14.33571429$

$\beta_1^A = 7.52142857,$

\\β_2^A = -0.05000000

\\β_3^A=-1.13571429

d. Determine the estimated parameters of the ANOVA model.

$\hat{\mu}$= 14.33571429

$\backslash\hat{\alpha}_1$= 7.52142857

$\backslash\hat{\alpha}_2 =$ -0.05000000

$\backslash\hat{\alpha}_3 = -1.13571429$

$\backslash\hat{\alpha}_4$=-(7.52142857+(-0.05000000) +(-1.13571429)) =- 6.33571428

e. Interpret all the effects in the ANOVA model.

The mean over-all effect is 14.33571429 that is, on average the memory score is 14.33571429

On average the memory score for no deprivation is 14.33571429+7.52142857=21.85714286

On average the memory score for 18 hours sleep deprivation is 0.05 lower than the overall effect. That is 14.33571429 -0.05000000=14.28571429

On average the memory score for 24 hours of sleep deprivation is 1.13571429 lower than the overall effect. That is 14.33571429 -1.13571429 =13.2

On average the memory score for 48 hours sleep deprivation is 6.33571428 lower than overall effect. That is 14.33571429 - 6.33571428 = 8. 00000001