**ISQA 8156/4156 Assignment 2 Due: By Monday, Oct. 30 2017, 5:30 PM**

**Name: Vineeth Reddy Ramasahayam**

**NOTES:**

1. Use statistics software (R, SPSS, etc.) for the calculations.
2. There are two required submissions:
   1. The hard copy of this answer file must be submitted in class. Only this hard copy of the answer file will be graded.
   2. For each problem in part 2, you have to submit your R/SPSS/Excel code in Canvas embedded in a Zip folder. If you use R, simply submit the R script(s) in the zip folder. If you use Excel, include the actual Excel files with the calculations. If you use SPSS, save the resulting output as Viewer File (.spv). Name your Zip Folder with your name, HW 2, and the course # (Example: LastName-HW2-ISQA 8156/4150).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Part 1: Understanding (circle correct answer, only one correct answer per question). 10 points**

1) In testing for the equality of k population means through ANOVA, the number of treatments is

**a. k.**

b. k - 1.

c. nT

d. nT - k

2) In a completely randomized experimental design involving five treatments, 13 observations were recorded for each of the five treatments (a total of 65 observations).

Also, the design provided the following information.

SSTR = 200 (Sum of Squares Due to Treatments)

SST = 800 (Total Sum of Squares)

The number of degrees of freedom corresponding to the error term (SSE) is

**a. 60.**

b. 59.

c. 5.

d. 4.

3) Consider the following information.

SSTR = 6750

SSE = 8000

H0: μ1 = μ2 = μ3 = μ4

Ha: At least one mean is different

The mean square due to treatments (MSTR) equals

a. 400.

b. 500.

c. 1687.5.

**d. 2250.**

4) Regression analysis is a statistical procedure for developing a mathematical equation that describes how

a. one independent and one or more dependent variables are related.

b. several independent and several dependent variables are related.

**c. one dependent and one or more independent variables are related**.

d. one dependent, one independent, and several error variables are related.

5) It is possible for the coefficient of determination (r2) to be

a. larger than 1.

**b. less than 1.**

c. less than -1.

d. equal to -1.

**GRAD STUDENTS ONLY:**

6) If two variables, x and y, have a strong linear relationship, then

1. **there may or may not be any causal relationship between x and y**
2. x causes y to happen
3. y causes x to happen
4. None of these alternatives is correct.

**Part 2: Application (60 points)**

**Problem I.** This problem refers to the case problem Sales Salaries in chapter 13 of the book (see attached PDF file for the problem description**). The data file is provided as csv**. (30 points)

1. The salary is your response variable. What are the factors in this experiment, and what are the factor levels (treatments)?

There are two factors in this experiment. They are **1. Experience** and **2. Position**

The Factor levels corresponding to Experience are **Low, Medium, High.**

The Factor levels corresponding to Position are **Inside, Outside.**

1. Ignore the experience level for now, and only consider Position and Salary. Conduct an ANOVA at the alpha = 0.05 level to see if the position type has an effect on the salary amount. Show the ANOVA table including F statistic and p value. Is the position significant?

H0 : µ1= µ2 µ1= Mean Salary of Inside Position

Ha : µ1≠ µ2  µ1= Mean Salary of outside Position

Analysis of Variance Table

Response: Salary

Df Sum Sq Mean Sq F value Pr(>F)

Position 1 9515793950 9515793950 251.54 < 2.2e-16 \*\*\*

Residuals 118 4463949042 37830077

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Here p value i.e. <2.2e-16 is much less than α i.e. 0.05

Therefore, We reject the null hypothesis

So, we can conclude that the position is significant

1. Ignore the position type now. Conduct an ANOVA at the alpha = 0.05 level to see if the experience level has an effect on the salary. What is your conclusion?

H0 : µ1= µ2 = µ3 µ1= Mean salary of high experience Personnel

Ha : µ1 ≠ µ2 ≠ µ3  µ2= Mean salary of medium experience Personnel

µ3= Mean salary of low experience Personnel

Analysis of Variance Table

Response: Salary

Df Sum Sq Mean Sq F value Pr(>F)

Experience 2 1.6681e+09 834050050 7.9261 0.0005913 \*\*\*

Residuals 117 1.2312e+10 105227717

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

p value i.e. 0.00059 < α i.e 0.05

Therefore, We reject the null hypothesis

So, we can conclude that Experience is significant.

1. Conduct a multiple comparison procedure (LSD test) for the ANOVA in part c. Use the LSD.test code that was provided in class for this. In R, you need to change the arguments of the LSD.test code for it to run, based on your own object and factor names. Which experience levels lead to significantly different salaries when compared with each other, at the **alpha = 0.01** level?

LSD High to Low H0 : µlow = µHigh

Ha : µlow ≠ µHigh

LSD High to Medium H0 : µMedium = µHigh

Ha : µMedium ≠ µHigh

LSD Low to Medium H0 : µlow = µMedium

Ha : µlow ≠ µMedium

$statistics

MSerror Df Mean CV t.value LSD

105227717 117 64925.47 15.79974 2.618504 6006.249

$parameters

test p.ajusted name.t ntr alpha

Fisher-LSD none Experience 3 0.01

$means

Salary std r LCL UCL Min Max Q25 Q50 Q75

High 66338.68 9699.513 40 62091.62 70585.73 51027 79081 57770.75 65441.0 76975.00

Low 59819.62 6005.055 40 55572.57 64066.68 48621 71345 55446.75 60410.0 63813.25

Medium 68618.12 13621.377 40 64371.07 72865.18 51246 88730 54646.50 69615.5 81827.25

$comparison

difference pvalue signif. LCL UCL

High - Low 6519.05 0.0053 \*\* 512.8009 12525.299

High - Medium -2279.45 0.3224 -8285.6991 3726.799

Low - Medium -8798.50 0.0002 \*\*\* -14804.7491 -2792.251

For High – Low : ≥ LSD Reject H0

The mean salaries of High Experience and low Experience Personnel are not Equal

For High – Medium : ≤ LSD Cannot Reject H0

The mean salaries of High Experience and Medium Experience Personnel are not significantly different

For Low – Medium : ≥ LSD Reject H0

The mean salaries of Low Experience and Medium Experience Personnel are not Equal

From the LSD table we can see that there is significant decrease in salaries from medium to low experience personnel.

1. **BONUS (5 points)**: In class we talked about comparison-wise vs experiment-wise Type 1 error. Run the LSD test (see part d) again, now including the Bonferroni adjustment. Consult the specification of the LSD test (?LSD.test) on how to do this. Using the **alpha = 0.01** level, what changes in your conclusion when you compare the results with part d?

$statistics

MSerror Df Mean CV t.value MSD

105227717 117 64925.47 15.79974 2.996665 6873.665

$parameters

test p.ajusted name.t ntr alpha

Fisher-LSD bonferroni Experience 3 0.01

$means

Salary std r LCL UCL Min Max Q25 Q50 Q75

High 66338.68 9699.513 40 62091.62 70585.73 51027 79081 57770.75 65441.0 76975.00

Low 59819.62 6005.055 40 55572.57 64066.68 48621 71345 55446.75 60410.0 63813.25

Medium 68618.12 13621.377 40 64371.07 72865.18 51246 88730 54646.50 69615.5 81827.25

$comparison

difference pvalue signif. LCL UCL

High - Low 6519.05 0.0159 \* -354.6146 13392.715

High - Medium -2279.45 0.9672 -9153.1146 4594.215

Low - Medium -8798.50 0.0006 \*\*\* -15672.1646 -1924.835

The significance difference between high and low salary personnel has been decreased compared to part d. This is because we are using MSD instead of LSD. The control limits have also been changed.

1. Finally, consider both factors (position type and experience level) in a factorial design. Conduct an ANOVA to see if factor 1, factor 2, and the interaction of the factors influence the salary amount. Show the ANOVA table. What is your conclusion?

H01 : There is no significant difference between the means of Position i.e.

µinside= µoutside

H02 : There is no significant difference between the mean years of Experience i.e.

µlow = µmedium = µHigh

Analysis of Variance Table

Response: Salary

Df Sum Sq Mean Sq F value Pr(>F)

Position 1 9515793950 9515793950 751.360 < 2.2e-16 \*\*\*

Experience 2 1668100099 834050050 65.856 < 2.2e-16 \*\*\*

Position:Experience 2 1352066184 676033092 53.379 < 2.2e-16 \*\*\*

Residuals 114 1443782758 12664761

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Here P value < 0.05 i.e. α

Therefore, we reject our Null Hypothesis

There is a significant difference between the means of position and between the means of Experience.

**Problem II**. This problem refers to the case problem Bukeye Creek Amusement Park in chapter 14 (look at the PDF file). **The data is provided as csv file**. Please answer following questions: (30 points)

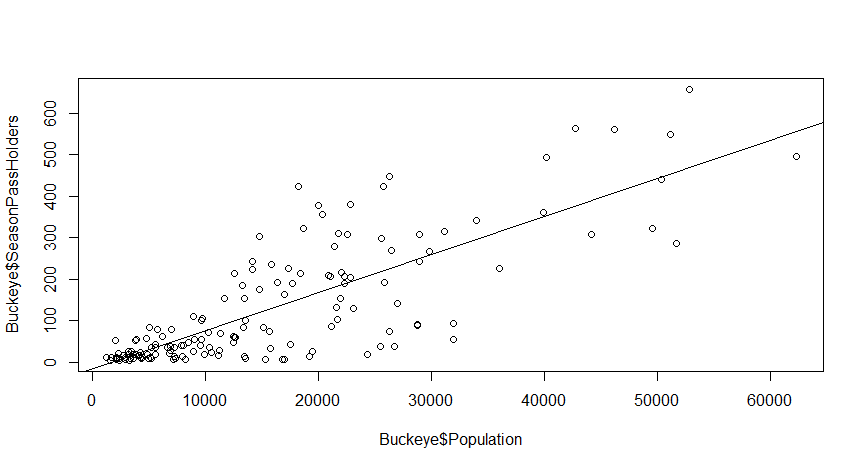
1. You’re interested to see if there is a (linear) relationship between the population of a Zip Code and the number of season passes. Specifically, you want to see if the number of season passes can be predicted by the population of a Zip code. What is your independent variable, and what is your dependent variable?

Independent Variable : Population of a ZIP Code

Dependent Variable : Number of Season Passes Sold

We want to predict the number of season passes sold by the population of a zip code. So, I concluded that Population is independent and No of season passes sold as dependent variable

1. Plot the population against the number of season tickets. Show the plot. What do you observe?



From the plot we can observe that it is a positive linear relationship since the slope is positive. Number of season pass holders appear to increase with increase in the population.

1. To statistically analyze the relationship, run a simple linear regression for the data. What is the resulting regression equation? Are the parameters b0 and b1 statistically significant? If yes, at what level (p value)?

Regression Equation: y= -16.258 + 0.0091 x

H0 : β1 = 0

Ha : β1 ≠ 0

Assuming α= 0.05

t = 16.1802 ( calculation in R )

Here , p value = < 2.2e-16 i.e. between 0 and 0.001 ( calculation in R )

So p < α

We reject Null Hypothesis that β1 = 0

Therefore, the parameters b0 and b1 are statistically significant.

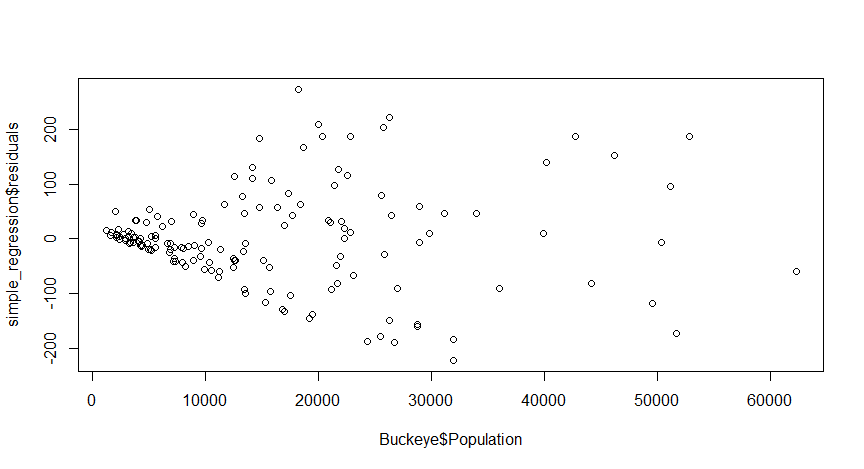
1. Does the regression model provide a good fit for the data (i.e., what is the R2 value)?

R2 = SSR/SST = 0. 6374 (calculation in R)

The regression relationship is strong ; 63.74 % of the variability in the number of season passes sold can be explained by the linear relationship between the Population of a Zip Code and number of season passes sold

\*Reference taken from lecture slides

1. **GRAD STUDENTS ONLY**: Now, test if the variance of the error term is constant. For this, plot the population values against the residuals of the regression model. Show the plot. What do you observe? Is the assumption of constant variance valid?



The value of residuals increases with increase in the population. The model does not fit consistently at al values of population. So, the assumption of constant variance is not valid.

1. What other data (variables) besides the population size might be useful to predict the number of season pass holders for a given Zip Code?

The age group of People residing at a particular Zip Code might be useful to predict the number of season pass holders for a given Zip code. Because young people may be more interested in visiting the amusement frequently than older people.

The distance of the particular Zip Code might also be useful to predict number of season pass holders. People residing near the part might frequently attend compared to people staying at 50th mile Zip Code.