STAT 43000/MA 53001 **Applied Statistics** **Spring 2023**

Test 2-Part II **Name:**

Due: April 24, 2023.

Please provide the complete solutions of all problems.

Q.N. 1) (15 points) The dataset "lrb.dta" is provided with this assignment. All married Southern Baptists between the ages of 20 to 25 are in the data file. The variables included in this dataset are

happymar: respondent's marital happiness
church: Church attendance
Gender (female):
educ: years of education, etc.

a) Import the data in R to determine its dimension and print first five observations.

```
> library(haven)
Warning message:
package 'haven' was built under R version 4.2.3
> lrb <- read_dta("C:\\Users\\Zhang\\Downloads\\lrb.dta")
> dim(lrb)
[1] 61  7
> head(lrb, 5)
# A tibble: 5 × 7
  happymar          church                female     educ    cheduc  educx cheducx
  <dbl+lbl>         <dbl+lbl>             <dbl+lbl>  <dbl+l>  <dbl>  <dbl>   <dbl>
1 0 [not as happy] 0 [rarely in church] 0 [male]     9           0 -3.28    0
2 1 [very happy]   0 [rarely in church] 1 [female] 10           0 -2.28    0
3 1 [very happy]   1 [often in church]  1 [female] 16          16  3.72    3.72
4 1 [very happy]   1 [often in church]  1 [female] 11          11 -1.28   -1.28
5 1 [very happy]   1 [often in church]  1 [female] 12          12 -0.279  -0.279
>
```

b) Create a binary variable for happymar by using respondent's marital happiness (1 = Very Happy, 0 = Otherwise).

```
lrb$happymar_bin <- ifelse(lrb$happymar == 1, 1, 0)
```

c) Run a simple logistic regression of happymar using years of educ and state the model equation.

```
> logistic_model <- glm(happymar_bin ~ educ, data = lrb, family = binomial(link = "logit"))
> summary(logistic_model)

Call:
glm(formula = happymar_bin ~ educ, family = binomial(link = "logit"),
    data = lrb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9371  -1.1585   0.5766   0.9582   1.4537

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.4782     2.6750  -2.422  0.01545 *
educ          0.5849     0.2235   2.617  0.00887 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 79.763  on 60  degrees of freedom
Residual deviance: 69.602  on 59  degrees of freedom
AIC: 73.602

Number of Fisher Scoring iterations: 4
```

Π=[1+exp(6.48-0.58x)]^-1

d) Test for significance of year of education to determine the marital happiness.

The p-value 0.008871295 for edu is smaller than 0.05, which means years of education is not a significant predictor of marital happiness

e) Update the model in (c) using the variable female (1= female, 0 = male) to predict the probability of marital happiness of a female with 16 years of education

```
> logistic_model_updated <- glm(happymar_bin ~ educ + female, data = lrb, family = binomial)
> predict_data <- data.frame(educ = 16, female = 1)
> predicted_prob <- predict(logistic_model_updated, newdata = predict_data, type = "response")
> predicted_prob
        1
0.9743252
```

Probability=0.97

Q.N. 2) (**5 points**) Given four treatments, a sample size of five observations of the response variable per treatment, TSS = 400 and SSE = 100, determine the analysis of variance table and test the null hypothesis of no treatment differences, on average, support your conclusion.

Here is a performing of a one-way ANOVA to test the null hypothesis of no treatment differences.

Number of treatments (k) = 4
Number of observations per treatment (n) = 5
Total number of observations (N) = 4 * 5 = 20
Total Sum of Squares (TSS) = 400

Sum of Squares due to Error (SSE) = 100

Sum of Squares due to Treatments (SST)= TSS - SSE = 400 - 100 = 300

Mean Squares due to Treatments (MST) = SST / (k - 1) = 300 / (4 - 1) = 100

Mean Squares due to Error (MSE) = SSE / (N - k) = 100 / (20 - 4) = 6.67

F = MST / MSE = 100 / 6.67 = 14.99

Degrees of freedom for treatments (DF_treatments) = k - 1 = 4 - 1 = 3

Degrees of freedom for error (DF_error) = N - k = 20 - 4 = 16

After checking the F table,

Since the calculated F-statistic 14.99 is greater than the critical F-value 3.238872, we can reject the null hypothesis and conclude that there are significant treatment differences, on average.

```
> alpha <- 0.05
> critical_F <- qf(1 - alpha, 3, 16)
> critical_F
[1] 3.238872
```

Q. N. 3) (10 points) A chemist wishes to test the effect of four chemical agents on the strength of a particular type of cloth. Because there might be variability from one bolt to another, the chemist decides to use a **randomized block design,** with the bolts of cloths considered as blocks. She selects five bolts and applies all four chemicals in random order to each bolt. The resulting tensile strengths are as follows.

	Bolt				
Chemical	1	2	3	4	5
I	73	68	74	71	67
II	73	67	75	72	70
III	75	68	78	73	68
IV	73	71	75	75	69

a) Analyze the data from the experiment at $\alpha = 0.05$ and draw appropriate conclusions.

```
> data <- matrix(c(73, 68, 74, 71, 67,
+                  73, 67, 75, 72, 70,
+                  75, 68, 78, 73, 68,
+                  73, 71, 75, 75, 69),
+                nrow = 4, byrow = TRUE)
> colnames(data) <- c("Bolt1", "Bolt2", "Bolt3", "Bolt4", "Bolt5")
> rownames(data) <- c("Chemical_I", "Chemical_II", "Chemical_III", "Chemical_IV")
> # Perform two-way ANOVA with blocks and treatments
> blocks <- factor(rep(1:5, each = 4))
> treatments <- factor(rep(1:4, times = 5))
> response <- c(data)
> # Fit the ANOVA model
> anova_model <- aov(response ~ treatments + blocks)
> summary(anova_model)
            Df Sum Sq Mean Sq F value   Pr(>F)
treatments   3  12.95    4.32   2.376    0.121
blocks       4 157.00   39.25  21.606 2.06e-05 ***
Residuals   12  21.80    1.82
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

The p-value for treatments is 0.121. we cannot reject the null hypothesis, which means there is no significant difference on average, among the effects of the four chemicals on the cloth strength.

b) Test for model assumption to justify your model.

```
> shapiro_test <- shapiro.test(residuals(anova_model))
> shapiro_test$p.value
[1] 0.04053571
```

p-value 0.04<0.05, which means we reject the null hypothesis of normality of residuals.

Q.N. 4) (10 points) The effective life of insulating fluids at an accelerated load of 35 KV is being studied. Test data have been obtained for four types of fluids. The result from a completely randomized experiment is as follows:

Type I	Type II	Type III	Type IV
17.6	16.9	21.4	19.3
18.9	15.3	23.6	21.1
16.3	18.6	19.4	16.9
17.4	17.1	18.5	17.5
20.1	19.5	20.5	18.3
21.6	20.3	22.3	19.8

a) Is there any indication that the fluids differ at $\alpha = 0.1$. What about at $\alpha = 0.05$?

```
> anova_model <- aov(Life ~ Fluid_Type, data = data_long)
> summary(anova_model)
             Df Sum Sq Mean Sq F value Pr(>F)
Fluid_Type    3  30.16   10.05   3.047 0.0525 .
Residuals    20  65.99    3.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

For $\alpha = 0.1$, there is a difference.
For $\alpha = 0.05$, there isn't a difference.

b) Use the Tukey's HSD test to identify the fluid types that are different (if any).

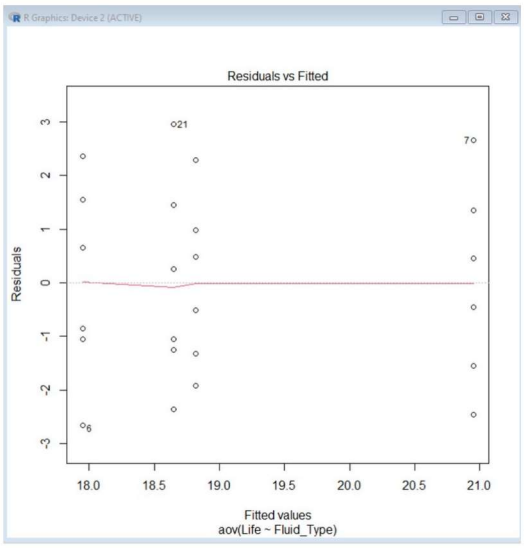
```
> summary(tukey_test)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Life ~ Fluid_Type, data = data_long)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Type_II - Type_I == 0   -0.7000     1.0488  -0.667   0.9081
Type_III - Type_I == 0   2.3000     1.0488   2.193   0.1594
Type_IV - Type_I == 0    0.1667     1.0488   0.159   0.9985
Type_III - Type_II == 0  3.0000     1.0488   2.861   0.0443 *
Type_IV - Type_II == 0   0.8667     1.0488   0.826   0.8413
Type_IV - Type_III == 0 -2.1333     1.0488  -2.034   0.2091
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

The significant difference exists between type Ⅲ and type Ⅱ

c) Which fluid would you select, given that the objective is long life?

```
> mean_life <- aggregate(data_long$Life, by = list(data_long$Fluid_Type), mean)
> colnames(mean_life) <- c("Fluid_Type", "Mean_Life")
> print(mean_life)
  Fluid_Type Mean_Life
1     Type_I  18.65000
2    Type_II  17.95000
3   Type_III  20.95000
4    Type_IV  18.81667
> # Identify the fluid type with the highest mean life
> max_life <- mean_life[which.max(mean_life$Mean_Life), ]
> print(max_life)
  Fluid_Type Mean_Life
3   Type_III     20.95
> |
```

d) Analyze the residuals from this experiment. Are the basic analysis of variance assumptions satisfied?



the p-value for the Shapiro-Wilk test is greater than 0.05, we can assume that the residuals are normally distributed. If the p-value for the Levene's test is greater than the chosen significance level, we can assume that the variances are homogeneous across groups.

```
> plot(anova_model, which = 1)
> shapiro_test <- shapiro.test(residuals(anova_model))
> levene_test <- leveneTest(Life ~ Fluid_Type, data = data_long)
> shapiro_test$p.value
[1] 0.3759872
> levene_test[1, 3]
[1] 0.9367602
>
```

e) Use the Kruskal-Wallis test and compare the results with the one you got from above steps.

```
> kruskal_test <- kruskal.test(Life ~ Fluid_Type, data = data_long)
> summary(kruskal_test)
           Length Class  Mode
statistic  1      -none- numeric
parameter  1      -none- numeric
p.value    1      -none- numeric
method     1      -none- character
data.name  1      -none- character
> kruskal_test$statistic
Kruskal-Wallis chi-squared
                  6.217703
> kruskal_test$p.value
[1] 0.1014857
```

Pvalue for Kruskal is smaller, which may caused by the non-normal distributed data, means that the Kruskal-Wallis test is more appropriate.

Q.N. 5) (10 points) An aluminum master alloy manufacturer produces grain refiners in ingot form. The company produces the product in four furnaces (A,B,C,D). Each furnace is known to

have its own unique operating characteristics, so any experiment run in the foundry that involves more than one furnace will consider furnaces as a nuisance variable. The process engineers suspect that stirring rate affects the grain size of the product. Each furnace can be run at four different stirring rates. A randomized block design is run for a particular refiner, and the resulting grain size data are as follows.

	Furnace			
Stirring Rate (rpm)	A	B	C	D
5	8	4	5	6
10	14	5	6	9
15	14	6	9	2
20	17	9	3	6

a) Is there any evidence that stirring rate affects grain size? Perform the RCBD using Furnace as a blocking factor to investigate it.

```
> data <- data.frame(
+   A = c(8, 14, 14, 17),
+   B = c(4, 5, 6, 9),
+   C = c(5, 6, 9, 3),
+   D = c(6, 9, 2, 6)
+ )
> library(tidyr)
> library(dplyr)
> data_long <- data %>%
+   pivot_longer(
+     everything(),
+     names_to = "Furnace",
+     values_to = "Grain_Size"
+   ) %>%
+   mutate(
+     Furnace = as.factor(Furnace),
+     Stirring_Rate = factor(rep(c(5, 10, 15, 20), 4))
+   )
> # Perform RCBD using Furnace as a blocking factor
> RCBD_model <- aov(Grain_Size ~ Stirring_Rate + Furnace, data = data_long)
> summary(RCBD_model)
              Df Sum Sq Mean Sq F value Pr(>F)
Stirring_Rate  3  165.2   55.06   6.591  0.007 **
Residuals     12  100.2    8.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

p-value for stirring_rate 0.007 is much more smaller than 0.05, which means stirring rate affects grain size significantly.

b) Suppose that the observation for Furnace C at the Stirring rate 10 rpm is missing (The highlighted value 6). If you would like to estimate this value by minimizing SSE, please estimate its value.

average: 5 rpm = (8+4+5+6)/4 = 5.75
10 rpm = (14+5+9)/3 = 9.33
15 rpm = (14+6+9+2)/4 = 7.75
20 rpm = (17+9+3+6)/4 = 8.75

} for each stirring rate

Furnance A = (8+14+10+17)/4 = 13.25
B = 6
D = 5.75

} for each furnance exclude C.

A, B, D = (13.25 + 6 + 5.75)/3 = 8.33

estimate missing value = 9.33 - 8.33 = 1