MATH 4310 **Name_ _ _ _ _ _ _ _ _ _ _ _ _ _ _**

Test #2

Problem 1: A study examines the vital capacity measurements of 60 adult males classified by 4 different types of occupations and three age groups. The file **test21.csv** contains the values of vital capacity (VC) vs age group (AGE) and occupation (OCC).

- **i)** Apply the Anova test to answer the following questions: **(a)** does the vital capacity differ among individuals with different occupations, **(b)** does the vital capacity differences among individuals with different age groups, and **(c)** is there an interaction between age and occupation? Use $\alpha = 0.01$ for all the tests. You must state the hypothesis testing problem you are solving.
- **ii)** Use the Tukey's HSD procedure to test for significant differences among individual pairs of means for age group and occupation, if appropriate (you can ignore the interaction term in the Tukey's HSD procedure). Justify your conclusion.

(i) *We use the Anova to test the null hypothesis that there is no difference among the means of (a) individuals with different occupations, (b) different age and (b) that there is no interaction between age and occupation.*

```
> data21 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test21.csv")
> data21$AGE = factor(data21$AGE,levels=unique(data21$AGE))
> data21$OCC = factor(data21$OCC,levels=unique(data21$OCC))
> data21.model = aov(VC~AGE+OCC+AGE:OCC, data = data21)
> anova(data21.model)

Analysis of Variance Table

Response: VC
             Df  Sum Sq Mean Sq F value    Pr(>F)    
AGEGROUP      2 12.3088  6.1544 29.3817 4.652e-09 ***
OCC           3 19.7785  6.5928 31.4750 2.129e-11 ***
AGEGROUP:OCC  6  8.9489  1.4915  7.1205 1.825e-05 ***
Residuals    48 10.0542  0.2095
```

The p-value in the above shows that for each of the 3 cases p-value < 0.01. Thus, we reject the null hypothesis, and we conclude that: (a) vital capacity differs among individuals with different occupations, (b) vital capacity differs among individuals with different age groups, and (c) there is an interaction between age and occupation

(ii) *We run the Tukey's HSD procedure:*

```
> TukeyHSD(data21.model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = VC ~ AGE + OCC + AGE:OCC, data = data21)

$AGE
       diff         lwr       upr     p adj
2-1 -0.7395 -1.08952404 -0.389476 0.0000164
3-1  0.3465 -0.00352404  0.696524 0.0528802
3-2  1.0860  0.73597596  1.436024 0.0000000

$OCC
         diff         lwr       upr     p adj
b-a 0.2073333 -0.23743004 0.6520967 0.6045104
c-a 0.4613333  0.01656996 0.9060967 0.0393618
d-a 1.4940000  1.04923663 1.9387634 0.0000000
c-b 0.2540000 -0.19076337 0.6987634 0.4338300
d-b 1.2866667  0.84190330 1.7314300 0.0000000
```

```
d-c 1.0326667  0.58790330 1.4774300 0.0000008
```

For the age factor, we observe p-value < 0.01 only for the comparison 2-1 and 3-2. For the occupation factor, we observe p-value < 0.01 only for the comparison d-a, d-b and d-c. All the other comparisons are not statistically significant at level $\alpha = 0.01$.

Problem 2: An experiment was run on six pregnant women to evaluate the effect of labor on glucose production and utilization. Glucose concentrations were collected on the six subjects during four stages of labor: latent (A1) and active (A2) phases of cervical dilatation, fetal expulsion (B), and placental expulsion (C); data are stored in file **test22.csv**

- **i)** Apply the Anova test (with blocks) to answer the following question: (a) is there an effect of labor on glucose production and utilization? (b) Is the experimental design balanced or not? Use $\alpha = 0.01$ for all these tests. [Hint: the subject variable is the factor block]
- **ii)** Use the Tukey's HSD procedure to test for significant differences among the four stages of labor, if appropriate.

(i) *We apply the two-way anova with blocks. This is an additive model where the first term in the formula is the block factor.*

```
> data22 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test22.csv")
> data22$GROUP = factor(data22$GROUP,levels=unique(data22$GROUP))
> data22$SUBJ = factor(data22$SUBJ,levels=unique(data22$SUBJ))
> str(data22)
'data.frame':   24 obs. of  3 variables:
 $ GC   : num  3.6 3.53 4.02 4.9 4.06 3.97 4.4 3.7 4.8 5.33 ...
 $ GROUP: Factor w/ 4 levels "A1","A2","B",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ SUBJ : Factor w/ 6 levels "1","2","3","4",..: 1 2 3 4 5 6 1 2 3 4 ...
> table(data22$GROUP,data22$SUBJ)

     1  2  3  4  5  6
  A1 1  1  1  1  1  1
  A2 1  1  1  1  1  1
  B  1  1  1  1  1  1
  C  1  1  1  1  1  1
> data22.model = aov(GC~SUBJ+GROUP, data = data22)
> anova(data22.model)
Analysis of Variance Table

Response: GC
          Df Sum Sq Mean Sq F value    Pr(>F)
SUBJ       5 8.7735 1.75470   6.426 0.0022156 **
GROUP      3 8.3409 2.78030  10.182 0.0006583 ***
Residuals 15 4.0960 0.27306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table shows that the experimental design is balanced.
Since the p-value corresponding to the GROUP variable is less than 0.01, we conclude that there is a statistically significant effect of labor on glucose production and utilization.
NOTE: solution is the same using GC~SUBJ+GROUP or GC~GROUP+SUBJ (due to balanced design)

```
> TukeyHSD(data22.model, which = "GROUP")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GC ~ SUBJ + GROUP, data = data22)
```

```
$GROUP
            diff          lwr      upr     p adj
A2-A1 0.6666667 -0.20287041 1.536204 0.1653989
B-A1  1.3366667  0.46712959 2.206204 0.0024454
C-A1  1.4816667  0.61212959 2.351204 0.0009660
B-A2  0.6700000 -0.19953708 1.539537 0.1624015
C-A2  0.8150000 -0.05453708 1.684537 0.0699315
C-B   0.1450000 -0.72453708 1.014537 0.9622295
```

The Tukey test shows that there is a statistically significant difference (p_adj < 0.01) between the stages B-A1 and C-A1.

The differences between the other stages are not statistically significant.

Problem3: Most fractures in older people are caused by the combination of weak bones and falls. A study conducted on a cohort of 169 elderly patients aims to predict fracture, using AGE, SEX, BMI (body mass index) and BMD (bone density mass) as main effects. See file: **Test23.csv**

(a) Compute and write the multiple logistic regression equation to predict fracture from AGE, BMI, BMD and SEX (round to 2 decimal digits).

(b) Compute the odds ratios for all coefficients of the multiple logistic regression equation.

(c) Test the null hypothesis H0: $\beta i = 0$ vs H1: $\beta i \neq 0$ for i=1,2,3,4 at significance level 0.05

(d) Compute the 95% confidence interval of all coefficients βi for i=1,2,3,4.

```
> data23 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test23.csv")
> str(data23)
'data.frame':   169 obs. of  9 variables:
 $ id         : int  469 8724 6736 24180 17072 3806 17106 23834 2454 2088 ...
 $ fracture   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age        : num  57.1 75.7 70.8 78.2 54.2 ...
 $ sex        : chr  "F" "F" "M" "F" ...
 $ fracture.1: chr  "no fracture" "no fracture" "no fracture" "no fracture" ...
 $ weight_kg  : num  64 78 73 60 55 65 77 59 64 72 ...
 $ height_cm  : num  156 162 170 148 161 ...
 $ bmd        : num  0.879 0.795 0.907 0.711 0.791 ...
 $ bmi        : num  26.5 29.7 25.1 27.4 21.2 ...
> data23$sex <- factor(data21$sex)
> mylogit <- glm(fracture ~ age+bmi+bmd+sex, family = "binomial", data = data21)
> summary(mylogit)

Call:
glm(formula = fracture ~ age + bmi + bmd + sex, family = "binomial",
    data = data21)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3753  -0.5039  -0.1985   0.3924   2.5843

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   9.79488    2.69720   3.631 0.000282 ***
age           0.01844    0.02094   0.881 0.378540
bmi          -0.05131    0.06013  -0.853 0.393537
bmd         -15.11747    2.80337  -5.393 6.94e-08 ***
sexM          0.84599    0.51249   1.651 0.098792 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 205.27  on 168  degrees of freedom
Residual deviance: 110.86  on 164  degrees of freedom

> exp(coefficients(mylogit))
 (Intercept)          age          bmi          bmd         sexM
1.794160e+04 1.018611e+00 9.499859e-01 2.719989e-07 2.330290e+00
> exp(confint(mylogit, level=0.95))
                   2.5 %       97.5 %
(Intercept) 1.308845e+02 5.716835e+06
age         9.776261e-01 1.062057e+00
bmi         8.425579e-01 1.069855e+00
bmd         5.960754e-10 3.890696e-05
sexM        8.722987e-01 6.619108e+00
```

i) We write the equation of the multiple logistic regression equation

Ln(p/(1-p) = 9.79 +0.02 age– 0.05 bmi -15.12 bmd +0.85 sex

ii) The odds ratios are

Exp(β1) = 1.018611, *Exp(β2)* = *0.*9499859, *Exp(β3)* = 2.719989 e-7, *Exp(β4)* = 2.330290

iii) Test the hypothesis about $\beta_1 = 0$ versus β_1 different from 0 at significance levels $\alpha = 0.05$.

The tables shows that only for the bmd (coefficient i=3) the p-value is less than 0.05, while the p-value is above 0,05 for the other coefficients of the regression model (i=1,2.4) We conclude that, at level α = 0.05 we do reject H0 for bmi but not for the other factors.

.

iv) We compute the approximate 95\% confidence intervals:

```
age         9.776261e-01 1.062057e+00
bmi         8.425579e-01 1.069855e+00
bmd         5.960754e-10 3.890696e-05
sex        8.722987e-01 6.619108e+00
```