BIOS663 Final Exam Spring 2019
Thursday, May 2 noon - 2pm

Instructions: Please be as rigorous as possible in all of your answers and show all your work.

Remember that to report a test, you should provide H_0, the test statistic, the degrees of freedom, the p-value, the decision, and an interpretation of the decision in terms of the subject matter (if possible).

You may not consult with anyone except the instructor for clarification of questions. The work you present should be your work alone. Violation of the honor code will be prosecuted (penalties may include failure of the course and expulsion from the university). Please sign the honor code pledge and submit it with your report.

Honor Code Pledge: On my honor, I have neither given nor received aid on this examination.

Name:

Signature:

Date:

1. (20pts) Consider the model $\mathbf{y}_{6\times 1} = \mathbf{X}_{6\times 3}\boldsymbol{\beta}_{3\times 1} + \boldsymbol{\epsilon}_{6\times 1}$, where $\mathbf{y}$ is blood pressure of 6 individuals, $\mathbf{X}$ includes intercept (1st column of $\mathbf{X}$) and two covariates: age (2nd column of $\mathbf{X}$) and body weight (lbs) (3rd column of $\mathbf{X}$).

   (a) (10pts) Is each of the following statement correct or not? If it is not correct, please explain why it is wrong and try to correct it.

   (i) $\boldsymbol{\beta}$ are statistics.

   (ii) $\boldsymbol{\epsilon}$ are parameters.

   (iii) $\mathbf{y}$ is a random variable following multivariate normal distribution with mean value $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$.

   (iv) $\hat{\sigma}^2$ is a random variable.

   (v) ϵ_1 is independent with ϵ_2.

   (b) (4pts) What is the interpretation of β_0, β_1, and β_2, respectively. Is the interpretation of β_0 meaningful, if so, why? If not, how to fix this problem?

   (c) (6pts) Now suppose we know the 6 individuals are from two families. The first three are from one family and the next three are from the other family. In order to accommodate the correlations between individuals within one family, we decide to use a random effect model to study the relation between blood pressure versus age and weight.

(i) (3pts) If we use "unstructured" covariance structure, how many parameters of the covariance matrix of the 6 individuals need to be estimated? Write out the covariance matrix using concise notations (you just need to present the form of the matrix, but do not need to calculate the actual values of the matrix elements).

(ii) (3pts) If we used "compound symmetry" covariance structure, how many parameters of the covariance matrix of the 6 individuals need to be estimated? Write out the covariance matrix using concise notations.

2. *(30 points total)* A group of subjects was recruited to a nutritional study in a medical center at UNC. The data consist of their BMI (y = BMI), daily exercise time (exer = exercise (in hours)) and daily vegetable intake (veg = vegetable (in servings)). One of the objectives in this study is to estimate how the exercise and vegetable consumption affect BMI. To address the question, we consider the following model:

$$y = \beta_0 + \beta_1 exer + \beta_2 veg + \epsilon.$$

Let $\mathbf{X}$ be the associated design matrix of the above model, we have

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 3.7 & -0.5 & 0.1 \\ -0.5 & 0.17 & 0 \\ 0.1 & 0 & 0.05 \end{pmatrix}.$$

(a) (11 points) A partial ANOVA table is given below. Complete the tables with missing numbers in ( ).

```
The GLM Procedure
Dependent Variable: y
Sum of
Source           DF       Squares    Mean Square   F Value    Pr > F
Model           (   )     85.5432    (         )   (      )     -
Error           (   )     1101.48    (         )
Corrected Total  199      (      )
```

```
                        Standard
  Parameter   Estimate    Error         t Value   Pr > |t|
  Intercept   (30.1)     (       )    (       )      -
    exer      (-2.3)     (       )    (       )      -
    veg       (-0.6)     (       )    (       )      -
```

(b) (5 points) Conduct the following hypothesis test $H_0 : \beta_1 = 0$ & $\beta_2 = -1$.

(c) (5 points) Compute the 95% confidence interval of the expected BMI of those who exercise 2 hours and eat 6 servings of vegetables daily.

(d) (9 points) Next we center *exer* to its mean which is 1 hour and scale *veg* by its average which is 5, and obtain new variables $new.exer = exer - 1$ and $new.veg = veg/5$, respectively. We then refit the data with the transformed predictors. Fill in the missing numbers inside ( ) in the following table.

```
                              Standard
 Parameter    Estimate         Error          t Value    Pr > |t|
Intercept     (       )    (       )     (       )          -
new.exer      (       )    (       )     (       )          -
new.veg       (       )    (       )     (       )          -
```

3. *(34 points total)* The abundance of Circulating Tumor Cells (CTC) is a new type of biomarker for cancer growth. In this study, we assess the association between the abundance of CTC and two drugs (drug A and B). Each of these two drugs is tested on 300 patients at 3 doses (1 to 3), with 100 patients per dose, and thus the total sample size is 600.

(a) (10 points) First consider the dose variable as a factor with 3 levels, and consider the following two way ANOVA model:

$$y = \alpha_0 + \alpha_1 x_1 + \sum_{j=2}^{3} \beta_j x_j + \sum_{j=2}^{3} \gamma_j (x_1 x_j) + e \qquad (1)$$

where x_1 is the indicator of drug B, and x_j is the indicator of dose j, and $x_1 x_j$ is the product of x_1 and x_j. Please write the cell mean for each combination of drug and dose in terms of α_0, α_1, β_i, and/or γ_j, and interpret the meaning of γ_2.

Drug	Dose	Mean
A	1	
A	2	
A	3	
B	1	
B	2	
B	3	

(b) (6 points) If we write the above model as a matrix form $\mathbf{y} = \mathbf{X}\boldsymbol{\eta} + \mathbf{e}$, using the data collected in this study, what is the dimension of $\mathbf{y}$, $\mathbf{X}$, and $\boldsymbol{\eta}$? please specify C and $\boldsymbol{\theta}_0$ for hypothesis testing of (1) dose level 2 and 3 cannot be distinguished, (2) there is no interaction between drug and dose. For each hypothesis testing, please explicitly specify your null hypothesis, the asymptotic distribution of your test statistic and its degree of freedom. You do not need to actually calculate the test statistic.

(c) (6 points) Next we consider CTC abundance vs. drug and age.

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 age + \alpha_3 (x_1 age) + e \tag{2}$$

Please write down the relation between CTC abundance and age for drug A and drug B separately, in terms of $_0$, α_1, α_2 and α_3. Can we compare this model with the model in question (a) by a likelihood ratio test and why?

(d) (6 points) Next consider the additive model with interaction terms removed:

$$y = \alpha_0 + \alpha_1 x_1 + \sum_{j=2}^{3} \beta_j x_j + e \tag{3}$$

Let μ_A and μ_B be the overall means of the two drugs. Express μ_A and μ_B by the parameters in the above model. Derive $\mathbf{C}$ for testing $H_0 : \mu_A = 2\mu_B$.

(e) (6 points) Finally treat the dose variable as a continuous variable and fit the following model

$$y = \mu_0 + \mu_1 x_1 + \mu_2 x + \epsilon \tag{4}$$

where x is the dose level. Compare the model using dose as a categorical variable (part (d)) and the model treating dose as a continuous variable (part (e)) by F-test. Please write down H_0, and give the degree of freedom of the corresponding F-distribution when H_0 is true.

4. *(16 points total)* In a mouse study, we are interested in tumor occurrences of 600 mice from three strains: 200 mice from B6, 200 mice from Cast, and 200 mice from PWK. This is a regression problem with one response, tumor occurrence, and one predictor: mouse strain (a categorical variable). In a simplified situation, we record 1 if a mouse has at least one tumor and 0 otherwise. Then tumor occurrence is a binary variable, and the results of a logistic regression is shown below:

```
                     Estimate    Std. Error  z value Pr(>|z|)
intercept             0.7538      0.3032     2.486    0.0129
I(strain=B6)         -0.5126      0.4160     -1.232   0.2179
I(strain=PWK)        -3.1401      0.4657     -6.742   <0.001
```

(a) (8 points) What is the odds ratio of the tumor occurrence for a B6 mouse vs a CAST mouse? Do you have enough information to construct a 95% confidence interval of the odds ratio? If yes, construct the CI. If not, explain why.

(b) (8 points) What is the estimated odds ratio of the tumor occurrence for a B6 mouse vs a PWK mouse? Do you have enough information to construct a 95% confidence interval of the odds ratio? If yes, construct the CI. If not, explain why.