BIOS663 Midterm Exam Spring 2019
March 6, 2019.

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

Please sign the honor code pledge and submit it with your report. Violation of the honor code below will be prosecuted (penalties may include failure of the course and expulsion from the university).

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

Name:

Signature:

Date:

1. *(20 points total)* Suppose $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim N(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 & 0.6 \\ 0 & 1 & 0.5 \\ 0.6 & 0.5 & 1 \end{bmatrix}$

   - (7 points) Derive the distribution of $2x_1 + x_2 - x_3$.
     Solution: Let $c = (2, 1, -1)$, then $var(2x_1+x_2-x_3) = c\mathbf{\Sigma}c^{'} = 2.6$ thus $2x_1+x_2-x_3$ follows a normal distribution with mean 0 and variance 2.6.
   - (7 points) Calculate $Cov(x_1 - x_2, 2x_2 + x_3)$.
     Solution: Let $c1 = (1, -1, 0)$ and $c2 = (0, 2, 1)$, then $Cov(x_1 - x_2, 2x_2 + x_3) = c1\mathbf{\Sigma}c2^{'} = -1.9$.
   - (6 points) Prove or dis-prove (with details) that $\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 0 & 2 \\ 4 & 1 & -2 \end{bmatrix}$
     has linearly independent columns.
     Solution: Since $\|\mathbf{A}\| = 9 \neq 0$, the rank of $\mathbf{A}$ is thus full rank, and $\mathbf{A}$ has linearly independent columns.

2. *(40 points total)* Consider the model $\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 5 \\ 6 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} -2 \\ -1 \\ 1 \\ 0 \\ 3 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} -1 \\ -1 \\ 3 \\ -1 \\ -2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

   with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$. Potentially helpful facts:
   A: the corrected *total* sum of squares is 17.2,
   B: a generalized inverse (if $\mathbf{X}$ is full rank, this inverse is unique) of $\mathbf{X}'\mathbf{X}$ is
   $(\mathbf{X}'\mathbf{X})^{-} = \begin{bmatrix} 0.21 & -0.02 & 0.02 \\ -0.02 & 0.07 & -0.02 \\ 0.02 & -0.02 & 0.08 \end{bmatrix}$; $\mathbf{X}'y = \begin{bmatrix} 17 \\ 16 \\ -5 \end{bmatrix}$ and
   C: 97.5 percentiles of student t-distributions:

df	1	2	3	4	5
	12.706	4.303	3.182	2.776	2.571

   - *(8 points)* Compute the least square estimates of the model parameters and their standard errors.

     Solution: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = (3.15, 0.88, -0.38)^{'}$; $\hat{y} = \mathbf{X}\hat{\beta} = (1.77, 2.65, 2.89, 3.53, 6.17)'$ and $\hat{\epsilon} = y - \hat{y} = (0.23, -1.65, 0.11, 1.47, -0.17)^{'}$. Thus $\hat{\sigma}^2 = 2.49$ and $\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, leading to $se(\hat{\beta}_0) = \sqrt{0.52} = 0.72, se(\hat{\beta}_1) = \sqrt{0.17} = 0.41$, and $se(\hat{\beta}_2) = \sqrt{0.2} = 0.45$.

- *(8 points)* Compute the 95% prediction interval for a subject with $x_1 = 1$ and $x_2 = 2$.

  Solution: $\hat{y} = (1, 1, 2)\hat{\beta} = 3.27$ and $var(\hat{y}) = (1, 1, 2)\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}(1, 1, 2)' + \hat{\sigma}^2 = 3.9$ a 95% prediction interval is $3.27 \pm 4.3\sqrt{3.9} = (-5.3, 11.7)$.

- *(8 points)* Calculate the corrected ${R^2}_c$, interpret its value, and test the hypothesis that its corresponding population value is zero, that is, $H_0 : \rho_c^2 = 0$.

  Solution: Since $\bar{y} = 3.4$, we have $CSS(regression) = \sum_i \hat{y}_i^2 - 5\bar{y}^2 = 11.2$ $CSS(total) = \sum_i y_i^2 - 5\bar{y}^2 = 17.2$ and corrected ${R^2}_c = 11.2/17.2 = 0.65$.
  $H_0 : \rho_c^2 = 0$ is equivalent to $H_0 : \beta_1 = \beta_2 = 0$. Thus we have $F - test = \frac{11.2/2}{\hat{\sigma}^2} = 2.25 \sim F_{2,2}$.

- *(6 points)* Consider the following hypothesis test: $E(y \mid \text{covariates of individual 5}) = 2E(y \mid \text{covariates of individual 1})$. Give $\mathbf{C}$, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ that are associated with the hypothesis test. Show as rigorously as possible whether your $\boldsymbol{\theta}$ is testable. If so, test the hypothesis.

  Solution: $\beta_0 + 3\beta_1 - 2\beta_2 = 2(\beta_0 - 2\beta_1 - \beta_2)$ which is $\beta_0 - 7\beta_1 = 0$. Let $\mathbf{C} = (1, -7, 0)$, then $\theta = \beta_0 - 7\beta_1$. For $H_0 : \theta = 0$, the associated t-test $= \hat{\theta}/se(\hat{\theta}) = -3.01/3.124 = -0.96$. Since $\| - 0.96\| < 4.3$, the hypothesis is not statistically significant given type I error of 0.05.

- *(5 points)* Show as rigorously as possible whether $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \\ \beta_0 + 2\beta_1 - 3\beta_2 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \\ 2 \end{pmatrix}$ is testable. If so, test the hypothesis. If not, can you construct an equivalent test that is testable? If yes, perform the equivalent test. If not, explain why.

  Solution: Let $\mathbf{C} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -2 \\ 1 & 2 & -3 \end{bmatrix}$, then $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$ which is estimable since $\mathbf{X}$ is full rank.
  The test is $H_0 : \boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$. Since $\mathbf{C}$ is not full rank, the test is not testable. Further, the test cannot be reduced to a testable hypothesis, since the three equations conflict with each other.

- *(5 points)* Show as rigorously as possible whether $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \\ \beta_0 + 2\beta_1 - 3\beta_2 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \\ 0 \end{pmatrix}$ is testable. If so, test the hypothesis. If not, can you construct an equivalent test that is testable? If yes, perform the equivalent test. If not, explain why.

Solution: Follow the above question, with the same $\mathbf{C}$ and , the test now is $H_0 : \boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ which again is not testable. However, we can reduce the above test to $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \end{pmatrix}$ or $\begin{pmatrix} \beta_0 + \beta_1 - \beta_2 \\ \beta_1 - 2\beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ with $\mathbf{C} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -2 \end{bmatrix}$ and $\hat{\boldsymbol{\theta}} = (4.41, 1.64)'$

F-test $= \frac{\hat{\boldsymbol{\beta}}'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1})\hat{\boldsymbol{\theta}}/2}{\hat{\sigma}^2} = 34.2/2.49 = 13.7 \sim F_{2,2}$.

3. *(40 points total)* An investigator at UNC conducted a survey of Chapel Hill residents both before and after construction of a new exercise trail. Before the trail was constructed, she determined the baseline physical activity levels of a number of Chapel Hill residents. After construction of the trail, she interviewed the same group of residents about their physical activity levels (after construction of the trail) along with their gender and age.

   Short descriptions of the variables of interest are provided below.

   - post: Average physical activity, measured in hrs per day, after construction of the trail.
   - pre: Physical activity, measured in hrs per day, before construction of the trail (baseline).
   - age: Age of each participant.
   - gender: Gender of each participant (Male =0 and Female=1).

   The investigator fit the following model, with data centered as indicated, to the physical activity data: $post = \beta_0 + \beta_1 pre + \beta_2 age + \beta_3 gender + error$. Let the design matrix of the model be $\mathbf{X}$, then the inverse of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.047 & -0.013 & 0 & -0.023 \\ -0.013 & 0.011 & 0 & 0 \\ 0 & 0 & 0.0000051 & 0 \\ -0.023 & 0 & 0 & 0.04 \end{bmatrix}.$$

Selected SAS output is also provided below.

```
                        The GLM Procedure

Dependent Variable: post
***Table One***
                                   Sum of
Source                      DF        Squares    Mean Square   F Value     Pr > F

Model                       ???        644.58        ???         ???         <.0001
Error                       ???        67.24         ???
Corrected Total              99         ???



***Table Two***
Standard Parameter          Estimate           Error        t Value    Pr > |t|
Intercept                   0.75                ???          ???        <0.0001
pre                         1.15                ???          ???        <0.0001
age                         0.052               0.0019       ???        <0.0001
gender                      ???                 ???          6.43       <0.0001
```

Based on this output, answer the following questions:

- *(10 points)* Fill in the cells with ??? in Table One. What are the degrees of freedom associated with the F test?

```
                        The GLM Procedure

Dependent Variable: post
***Table One***
                                   Sum of
Source                      DF        Squares    Mean Square   F Value     Pr > F

Model                       (3)        644.58    (214.86)     (306.9)     <.0001
Error                       (96)        67.24     (0.7)
Corrected Total              99        (711.82)
```

- *(3 points)* Estimate σ^2.
  Solution: $\hat{\sigma}^2 = 0.7$.

- *(5 points)* Report a F-test of the hypothesis that the prior physical activity levels are unrelated to the post-construction physical activity, after adjusting effects of age and gender. Give the nested models implicitly being compared when one conducts this F-test.

Compare full model $post = \beta_0 + \beta_1 pre + \beta_2 age + \beta_3 gender + error$ with $post = \beta_0 + \beta_2 age + \beta_3 gender + error$ or testing $H_0 : \beta_1 = 0$.
t-test $= \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 1.15/\sqrt{\hat{\sigma}^2 * 0.011} = 13.1 \sim t(96) \approx N(0,1)$ thus significant at $\alpha = 0.05$. Thus F-test $= (t - test)^2 = 13.1^2 = 171.6 \sim F_{1,96}$.

- *(7 points)* Fill in the cells with ??? in Table Two.

Table Two

Standard Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	0.75	(0.181)	(4.14)	<0.0001
pre	1.15	(0.088)	(13.1)	<0.0001
age	0.052	0.0019	(27.4)	<0.0001
gender	(1.07)	(0.167)	6.43	<0.0001

- *(4 points)* Test $H_0 : \beta_1 = 1$.
  Solution: t-test $= \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = 0.15/\sqrt{\hat{\sigma}^2 * 0.011} = 1.71 \sim t_{96} \approx N(0,1)$. Since $1.71 < 1.96$, the test is not significant at $\alpha = 0.05$.
- *(8 points)* What is the interpretation of the intercept in the aforementioned regression model? To make β_0 more interpretable, the investigator decides to rescale the variable age by its mean which is 40, and refit the following regression model:
  $post = \beta_0 + \beta_1 pre + \beta_2 newage + \beta_3 gender + error$
  where $newage = age - 40$. Fill in the cells with ??? in Table Three.

***Table Three ***

Standard Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	2.83	0.196	14.4	<0.0001
pre	1.15	0.088	13.1	<0.0001
newage	0.052	0.0019	27.4	<0.000
gender	1.07	0.167	6.43	<0.0001

Solution: The intercept is the expected post construction physical activity per day for a male resident with age 0 and average physical activity per day of 0 hours before the construction.

(3 points) Explain the assumption of homogeneity in the context of this experiment. Is it possible to assess the validity of this assumption from the summary statistics given? If so, how?

Solution: Homogeneity means that variability of the random error is constant across all subjects. No way to assess this assumption without the residuals, or any way to compute them.