GROUND RULES:

- This exam contains 5 questions. Each question is worth 10 points. Therefore, this exam is worth 50 points.
- Print your name at the top of this page in the upper right hand corner.
- This is a closed-book and closed-notes exam. You may use a calculator if you wish. Show all of your work and explain all of your reasoning!
- Any discussion or otherwise inappropriate communication between examinees, as well as the appearance of any unnecessary material, will be dealt with severely.
- You have 55 minutes to complete this exam. GOOD LUCK!

HONOR PLEDGE FOR THIS EXAM:

After you have finished the exam, please read the following statement and sign your name below it.

> *I promise that I did not discuss any aspect of this exam with anyone other than the instructor, that I neither gave nor received any unauthorized assistance on this exam, and that the work presented herein is entirely my own.*

1. Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Recall that in this model the x_i's are treated as fixed constants (i.e., they are not random). Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote the least squares estimators. In class, we proved that $\widehat{\beta}_0 \sim \mathcal{N}(\beta_0, c_{00}\sigma^2)$ and $\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2)$, where

$$c_{00} = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{and} \quad c_{11} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

We also proved that

$$\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \sigma^2 \left[\frac{-\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right].$$

Let $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ denote the ith fitted value. Show that $V(\widehat{Y}_i) = \sigma^2 m_{ii}$, where

$$m_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

2. Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Recall that in this model the x_i's are treated as fixed constants (i.e., they are not random). Let $\widehat{\beta}_1$ denote the least squares estimator of β_1. In class, we proved that $\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2)$, where

$$c_{11} = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

We also stated, without proof, that

$$W = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2),$$

where

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2.$$

Use these two results to derive a $100(1-\alpha)$ percent confidence interval for β_1. Clearly explain all steps in your derivation.

3. Suppose that $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)'$ has a multivariate normal distribution; specifically,

$$\mathbf{Y} \sim \mathcal{N}_4 \left\{ \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 & 1 & 0 & -1 \\ 1 & 2 & -1 & 2 \\ 0 & -1 & 1 & 0 \\ -1 & 2 & 0 & 1 \end{pmatrix} \right\}.$$

(a) What is the distribution of $Y_1 - Y_4$? Be precise.

(b) Define

$$\mathbf{a} = \begin{pmatrix} -2 \\ 2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} -1 & 2 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Define $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Y}$. Find the distribution of $\mathbf{X}$.

4. Consider the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Recall that in this model $\mathbf{X}$ is a constant matrix (i.e., it is not random). Let $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denote the hat matrix. Let $\mathbf{I}$ denote the identity matrix that has the same dimensions as $\mathbf{M}$.

(a) Prove that $[(\mathbf{I} - \mathbf{M})\mathbf{Y}]'\mathbf{M}\mathbf{Y} = 0$.

(b) Find $E[(\mathbf{I} - \mathbf{M})\mathbf{Y}]$ and $V[(\mathbf{I} - \mathbf{M})\mathbf{Y}]$.

5. A researcher is interested in understanding how the body mass index (BMI) of grade school children, denoted by Y, is related to

$$\begin{aligned} x_1 &= \text{age of child} \\ x_2 &= \text{average calories eaten at breakfast} \\ x_3 &= \text{average exercise hours per day} \\ x_4 &= \text{gender } (1 = \text{F}; 0 = \text{M}). \end{aligned}$$

For a sample of $n = 20$ children, she considers the statistical model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i, \quad i = 1, 2, ..., 20,$$

where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$.
(a) This model can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. State the dimensions of $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$.
(b) Here is the ANOVA output from fitting this model in SAS with the available data (from $n = 20$ children).

Source	DF	SS	MS	F	Pr > F
Model	4	49.84	12.46	7.55	0.0015
Error	15	24.78	1.65		
Corrected Total	19	74.62			

The researcher believes that only age and exercise (x_1 and x_3) are related to BMI and therefore believes that the **reduced model**

$$Y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 20$, is adequate. Here is the ANOVA table from the reduced model, obtained using `SAS`.

Source	DF	SS	MS	F	Pr > F
Model	2	43.09	21.55	11.65	0.0007
Error	17	31.53	1.85		
Corrected Total	19	74.62			

Perform a level $\alpha = 0.05$ test to assess whether or not the reduced model is adequate for the data. State your hypotheses, show how your test statistic is computed, state the rejection region, and write your conclusion.

This is an extra page for Problem 5.