BIOS 6612 Homework 4: The General Linear Model
Solutions

1. Consider a hypothetical example where a normally distributed outcome Y is simulated from a binary exposure variable X for $n = 20$ subjects. We fit a simple linear regression model to this data and obtain the following results: The MSE is estimated as 0.876588; estimates and standard errors appear in the table below:

Coefficient	Estimate	Std. error	t value	p-value
(Intercept)	0.13220	0.29607	0.4465	0.6605
X	0.61664	0.41871	1.4727	0.1581

   (a) Write out the estimated data-generating model in subject (not matrix) form using the parameter estimates appearing in the table above. Include the distribution of the error term.

   (2 points) $Y_i = 0.13220 + 0.61664X_i + \epsilon_i, \epsilon_i \sim N(0, 0.876588)$

   (b) If X is coded as an indicator variable (e.g., '1' for Female and '0' for Male), you will essentially get the same model fit whether or not you put this variable into the CLASS statement (SAS) or use `as.factor()` (R). Thus, although gender is clearly not a continuous variable, we can treat it as such when fitting the model. Briefly describe why this is the case.

   (4 points) The conditional mean of Y_i given $X_i = 0$ can be written as β_0; the conditional mean given $X_i = 1$ can be written as $\beta_0 + \beta_1$. The full model, written as $\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$, is the same whether X_i is treated as continuous or categorical because the two levels of the categorical variable are 0 and 1.

   (c) Use the information in the table above to construct the ANOVA table for this model: **(5 points)**

	Df	Sum Sq	Mean Sq	F value	p-value
Model	1.0000	1.9012	1.9012	2.1689	0.1581
Residuals	18.0000	15.7786	0.8766		

   The p-value for the ANOVA table is the same as for the t test above. The F statistic is equal to $1.4727^2 = 2.1689$. We can use this and the MSE to get the mean squares for the regression: $MSR = F \times MSE = 2.1689 \times 0.8766 = 1.9012$.

The df for the model is 1 since there is one covariate, and for the residuals it is $n-2=18$ since there are two parameters (intercept plus the X coefficient) in the model. The sums of squares are just the mean squares multiplied by degrees of freedom.

2. For n independent subjects, consider a normally distributed outcome $\mathbf{Y}$ and a group variable with 4 levels (i.e., 4 groups). We will look at three models:

   - Model 1: $\mathbb{E}(Y_i|\texttt{group}_i) = \beta_0 + \beta_1 \texttt{group}_i$, where $\texttt{group}_i = 0, 1, 2, 3$ for the 4 groups.
   - Model 2: $\mathbb{E}(Y_i|\texttt{group}_i) = \alpha_0 + \alpha_1\, \mathbb{1}(\texttt{group}_i = 1) + \alpha_2\, \mathbb{1}(\texttt{group}_i = 2) + \alpha_3\, \mathbb{1}(\texttt{group}_i = 3)$; recall that $\mathbb{1}(\cdot)$ is the indicator function, equal to 1 if $\cdot$ is true and 0 otherwise.
   - Model 3: $\mathbb{E}(Y_i|\texttt{group}_i) = \gamma_0 + \gamma_1\, \mathbb{1}(\texttt{group}_i = 0) + \gamma_2\, \mathbb{1}(\texttt{group}_i = 1) + \gamma_3\, \mathbb{1}(\texttt{group}_i = 2) + \gamma_4\, \mathbb{1}(\texttt{group}_i = 3)$.

   In matrix form, **Model 1** may be written as $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ where

$$\mathbb{E}(\mathbf{Y}) = \mathbb{E}\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

   for the $n \times 1$ vector $\mathbf{Y}$. Also for **Model 1**, in matrix form

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & \texttt{group}_1 \\ 1 & \texttt{group}_2 \\ \vdots & \vdots \\ 1 & \texttt{group}_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

   for the $n \times 2$ matrix $\mathbf{X}$ and the 2×1 vector $\boldsymbol{\beta}$.

   Use the information given for Model 1's matrix form as a guide as you answer the following questions.

   (a) Is $\mathbf{X}$ in Model 1 a full-rank model? Why or why not?

   (2 points) Yes, because the two columns are not linearly dependent on one another.

   (b) For **Model 2**, write $\mathbf{X}\boldsymbol{\alpha}$ in matrix form, give the number of columns of $\mathbf{X}$, and state whether or not Model 2 is full rank.

   (4 points)

$$\mathbf{X}\boldsymbol{\alpha} = \begin{pmatrix} 1 & \mathbb{1}(\texttt{group}_1 = 1) & \mathbb{1}(\texttt{group}_1 = 2) & \mathbb{1}(\texttt{group}_1 = 3) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbb{1}(\texttt{group}_n = 1) & \mathbb{1}(\texttt{group}_n = 2) & \mathbb{1}(\texttt{group}_n = 3) \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

   $\mathbf{X}$ has 4 columns and the model is full rank.

(c) For **Model 3**, write $\mathbf{X}\boldsymbol{\gamma}$ in matrix form, give the number of columns of $\mathbf{X}$, and state whether or not Model 3 is full rank.

(4 points)

$$\mathbf{X}\boldsymbol{\gamma} = \begin{pmatrix} 1 & \mathbb{1}(\texttt{group}_1 = 0) & \mathbb{1}(\texttt{group}_1 = 1) & \mathbb{1}(\texttt{group}_1 = 2) & \mathbb{1}(\texttt{group}_1 = 3) \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \mathbb{1}(\texttt{group}_n = 0) & \mathbb{1}(\texttt{group}_n = 1) & \mathbb{1}(\texttt{group}_n = 2) & \mathbb{1}(\texttt{group}_n = 3) \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix}$$

$\mathbf{X}$ has 5 columns and the model is not full rank since column 5 is equal to 1 minus the sum of columns 2, 3, and 4 (that column can only have an entry equal to 1 if there are all 0 entries for that row in columns 2–4).

(d) Using a regression coefficient vector of $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots)^T$, write out a group-level means model for $\mathbb{E}(Y_i|\texttt{group}_i)$ **in subject (not matrix) form**. How are these parameters related to those in Models 2 and 3?

(4 points) The subject-level model is

$$\mathbb{E}(Y_i|\texttt{group}_i) = \mu_0\,\mathbb{1}(\texttt{group}_i = 0) + \mu_1\,\mathbb{1}(\texttt{group}_i = 1) + \mu_2\,\mathbb{1}(\texttt{group}_i = 2) + \mu_3\,\mathbb{1}(\texttt{group}_i = 3).$$

The parameters of this model are related to those in model 2 as follows: $\mu_0 = \alpha_0, \mu_1 = \alpha_0 + \alpha_1, \mu_2 = \alpha_0 + \alpha_2, \mu_3 = \alpha_0 + \alpha_3$. If we drop the intercept γ_0 from model 3, then $\mu_j = \gamma_{j+1}, j = 0, 1, 2, 3$.

(e) Suppose the group variable is unequally spaced such that group 0 contains subjects who smoked no cigarettes per day, group 1 contains subjects who smoked 1 cigarette per day, group 2 contains subjects who smoked 20 cigarettes per day, and group 3 contains subjects who smoked 100 cigarettes per day. Should group be treated as a continuous variable with $\texttt{group}_i = 0, 1, 2, 3$ or with indicator variables? Justify your answer.

(2 points) Group should be treated as indicator variable since the unequal spacing for group is not acknowledged by the continuous variable (i.e., $\texttt{group}_i = 0, 1, 2, 3$). Alternatively, we could include it as a continuous variable if we modified it to be the number of cigarettes instead of the arbitrary $\texttt{group}_i = 0, 1, 2, 3$ coding. However, the issue with this approach is what to do with people who would fall between these levels (e.g., those who smoke more than 1 but fewer than 20 cigarettes per day).