BIOS663 Final Exam

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

Name:

Signature:

Date:

1. (28pts) Consider the model $\mathbf{y}_{8\times1} = \mathbf{X}_{8\times3}\boldsymbol{\beta}_{3\times1} + \boldsymbol{\epsilon}_{8\times1}$, where $\mathbf{y}$ is blood pressure of 8 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 $\mathbf{X}$). More specifically,

$$\mathbf{y} = \begin{bmatrix} 137 \\ 126 \\ 114 \\ 95 \\ 111 \\ 112 \\ 107 \\ 121 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 26 & 134 \\ 1 & 27 & 138 \\ 1 & 23 & 118 \\ 1 & 24 & 124 \\ 1 & 22 & 123 \\ 1 & 30 & 135 \\ 1 & 20 & 128 \\ 1 & 25 & 131 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \text{and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_8 \end{bmatrix} \sim N(0, \sigma^2\mathbf{I})$$

You should NOT run any software to answer the following questions. However, some computation by calculator maybe needed given the following potential helpful facts.

- The corrected total sum of squares of $\mathbf{y}$ is 1476.
- $(\mathbf{X}^T\mathbf{X})^{-1} =$

	intercept	age	weight
intercept	57.406	0.435	-0.528
age	0.435	0.028	-0.009
weight	-0.528	-0.009	0.006

- $\hat{\sigma}^2 = 145.37$.

(a) (5pts) 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 $\mathbf{0}_{8\times1}$ and variance $\sigma^2\mathbf{I}_{8\times8}$.

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

v. ϵ_1 is independent with ϵ_2.

(b) (3pts) Fill in the following t-table and please show your work on calculating the Standard Errors.

Parameter	Estimate	Standard Error	t value	Pr(>|t|)
(Intercept)	-22.0801			0.823
age	-0.1105			0.959
weight	1.0877			0.299

(c) (5pts) Test $\beta_0 = \beta_1 = \beta_2$ using GLH approach. Write out the contrast matrix $\mathbf{C}$, calculate test statistic and specify its null distribution and the corresponding degree of freedom. Though you do not need to calculate the p-value.

(d) (5pts) Test $\beta_1 = \beta_2 = 0$ using GLH approach. Write out the contrast matrix $\mathbf{C}$, calculate test statistic and specify its null distribution and the degree of freedom. Though you do not need to calculate the p-value.

(e) (5pts) Calculate the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$.

(f) (5pts) 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?

2. (20pts) Still use the data presented in problem 1. Suppose we are interested in the event of whether blood pressure is larger than 120. Let $\tilde{y}_i = 1$, if $y_i > 120$, and $\tilde{y}_i = 0$ otherwise. Here $i = 1, 2, ..., 8$ is the index of the 8 individuals. Let $p_i = Pr(y_i > 120)$.

(a) (5pts) Is p_i a parameter or a statistic? Given p_i, what the distribution of $\tilde{y}_i$? Calculate $\tilde{y}_i$'s expectation and variance.

(b) (5pts) Calculate the odds ratio of the event $y_i > 120$ vs. the event weight > 132.

(c) (5pts) Now we fit a logistic regression to study the relation $\tilde{\mathbf{y}}$ and age and weight. Please use the following regression coefficients estimates,

```
             Estimate Std. Error
(Intercept) -118.9085   135.9180
age           -0.7111     0.9114
weight         1.0373     1.1950
```

to estimate the probability that blood pressure is larger than 120 for an individual of age 30 and weight 133.

(d) (5pts) Please use the regression coefficient estimates in part (c) to calculate the odds ratio of the event $y_i > 120$ for person B vs. person A. They are of the same age, but B is 10 pounds heavier than A.

3. (12pts) Now suppose we know the 8 individuals are from two family. The first four are from one family and the next four 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.

(a) (4pts) If we use "unstructured" covariance structure, how many parameters of the covariance matrix of the 8 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).

(b) (4pts) If we used "compound symmetry" covariance structure, how many parameters of the covariance matrix of the 8 individuals need to be estimated? Write out the covariance matrix using concise notations.

(c) (2pts) Which covariance structure (unstructured or compound symmetric) should we use for this dataset and why?

(d) (2pts) Mixed model parameters can be estimated using either Maximum Likelihood (ML) method or Restricted maximum likelihood (REML) method. In order to compare a model with fixed effects of age and weight vs. the other model with only one fixed effect weight, should we use ML or REML method, and why? (Assume the same covariance structure is used both models.)

4. (25pts) We want to compare two drugs (denoted by A and B) for their effects of reducing cholesterol levels (LDL, in the unit of mg/dL). The following table shows the sample size for each combination of drug and dosage.

Drug	Dose	Sample Size (n_{ij})	i (drug index)	j (dose index)
	1	100	1	1
A	2	100	1	2
	3	100	1	3
	1	100	2	1
B	2	100	2	2
	3	100	2	3

(a) (3pts) First consider the dose variable as a categorical variable with 3 levels, and employ an additive model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk},$$

where i=1, j=1,2, k=1, 2., ..., n_{ij}. We use reference cell coding with drug B and dose 3 as reference. Therefore α_1 models the effect of drug A (drug B is reference), β_j models the effect for dose j (j=1 or 2) (dose 3 is reference); and e_{ijk} (k=1, 2., ..., n_{ij}) indicates residual error. If we write this ANOVA model as a regression model: $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, what is the dimension of $\mathbf{y}$, $\mathbf{X}$, $\mathbf{b}$ and $\mathbf{e}$, and for an ANOVA model, what kind of distribution we usually assume $\mathbf{e}$ should follow?

(b) (4pts) For the model specified in part (a), write the cell mean for each combination of drug and dose in terms of μ, α_i and β_j.

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

(c) (3pts) For the model specified in part (a), fill the following ANOVA table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	---	-----	22428.3	----	<.0001
Error	---	-----	392.2		
Corrected Total	---	-----			

(d) (3pts) If we model the interaction between dose and drug, the model can be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where γ_{ij} indicates interaction effects. Write the cell mean for each combination of drug and dose in terms of μ, α_i, β_j and γ_{ij}. Explain the meaning of interaction effect γ_{11} by comparing the table in question (b) and the table in this question.

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

(e) (2pts) Now if we model dose as a interval variable, with doses equals to 1, 2, 3 and fit a model of LDL with main effects of dose and drug, but no interaction, fill the following ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	---	-----	33601.8	-----	<.0001
Error	---	-----	391.7		
Corrected Total	---	-----			

(f) (3pts) Compare the model using dose as a categorical variable (part (c)) and the model using dose as a interval variable (part (f)) by F-test. Please write down H_0, calculate F-Statistic, and give the degree of freedom of the corresponding F-distribution when H_0 is true. Though you do not need to calculate the p-value.

(g) (4pts) Let μ_A and μ_B be the overall mean values of LDL for drug A and B, respectively. Write μ_A and μ_B in terms of α_i, β_j and γ_{ij}. If we want to test $H_0 : \mu_A = \mu_B$, write H_0 in terms of α_i, β_j and γ_{ij}, the contrast matrix, and the degrees of freedom.

(h) (3pts) If the design is unbalanced, with sample size shown in the following table. Test $H_0 : \mu_A = \mu_B$. Write H_0 in terms of α_i, β_j and γ_{ij}, the contrast matrix, and the degrees of freedom.

Drug	Dose	Sample Size (n_{ij})	i (drug index)	j (dose index)
A	1	100	1	1
	2	100	1	2
	3	50	1	3
B	1	100	2	1
	2	100	2	2
	3	50	2	3

5. (15pts) Still using the data of Problem 4 (with balanced design of 100 samples in each cell). Now we introduce another interval variable "age" and the interaction between drug and dose, fit a model using the following SAS code

```
proc glm;
class drug;
model LDL= age dose drug drug*dose/ solution;
run;
```

and obtained the following output.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	86791.9439	21697.9860	60.26	<.0001
Error	595	214247.6617	360.0801		
Corrected Total	599	301039.6056			

R-Square	Coeff Var	Root MSE	LDL Mean
0.288307	15.19667	18.97578	124.8680

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	21309.98280	21309.98280	59.18	<.0001
dose	1	6664.73548	6664.73548	18.51	<.0001
drug	1	58218.74750	58218.74750	161.68	<.0001
dose*drug	1	598.47814	598.47814	1.66	0.1978

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	19205.20980	19205.20980	53.34	<.0001
dose	1	6683.65025	6683.65025	18.56	<.0001
drug	1	4691.53105	4691.53105	13.03	0.0003
dose*drug	1	598.47814	598.47814	1.66	0.1978

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	104.9204188	B	3.94321568	26.61	<.0001
age	0.4855570		0.06648601	7.30	<.0001
dose	5.3125392	B	1.34185926	3.96	<.0001
drug 0	-14.8100813	B	4.10298291	-3.61	0.0003
drug 1	0.0000000	B	.	.	.
dose*drug 0	-2.4479126	B	1.89876546	-1.29	0.1978
dose*drug 1	0.0000000	B	.	.	.

(a) (3pts) Write down the fitted model based on the above output. Is dose treated as categorical or interval variable?

(b) (2pts) Why is the regression coefficient estimate for "drug 1" is 0 without estimate for standard error? Note the numerical value of drug is 0 for drug A and 1 for drug B.

(c) (3pts) Briefly explain what is the difference between Type I SS and Type III SS. Why the Type I SS of age is larger than the Type III SS of age, but the Type I SS of dose*drug is the same as the Type III SS of dose*drug?

(d) (4pts) Write the contrast matrix to estimate the average LDL level when drug A is used for an individual of age 40. Similarly, Write the contrast matrix to estimate the average LDL level when drug B is used for an individual of age 40.

(e) (3pts) Write the contrast matrix to test the hypothesis that the average LDL level for the individuals of age 40 taking drug A is different from the average LDL level for the individuals of age 40 taking drug B. Write the formula to calculate the test-statistic and what is the degree of freedom of this test?