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,

$$\mathbf{y} = \begin{array}{c} \text{BP} \\ \downarrow \\ \begin{bmatrix} 137 \\ 126 \\ 114 \\ 95 \\ 111 \\ 112 \\ 107 \\ 121 \end{bmatrix} \end{array}, \quad \mathbf{X} = \begin{array}{ccc} \text{intercept} & \text{age} & \text{bwt} \\ \downarrow & \downarrow & \downarrow \\ \multicolumn{3}{c}{\begin{bmatrix} 1 & 26 & 134 \\ 1 & 27 & 138 \\ 1 & 23 & 118 \\ 1 & 24 & 124 \\ 1 & 22 & 123 \\ 1 & 30 & 135 \\ 1 & 20 & 128 \\ 1 & 25 & 131 \end{bmatrix}} \end{array}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \text{and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_8 \end{bmatrix} \sim N(0, \sigma^2\mathbf{I})$$

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.

Incorrect, β's are parameters that we can't observe. We use $\hat{\beta}$ to estimate them.

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

Incorrect. ε's are random errors.

iii. $\mathbf{y}$ is a random variable following multivariate normal distribution with mean value $\mathbf{0}_{8\times1}$ and variance $\sigma^2\mathbf{I}_{8\times8}$.

Incorrect, y is a random variable following multivariate normal distribution but the mean value $E(y) = X\beta$, not $E(\varepsilon)$, covariance $= \sigma^2 I_{8\times8}$

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

Correct. $\hat{\sigma}^2$ is the estimator of σ^2 and is a random variable.

v. ϵ_1 is independent with ϵ_2.

Correct. random errors are assumed to be independent of each other.

(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	91.3516	-0.2417	0.823
age	-0.1105	2.0175	-0.0548	0.959
weight	1.0877	0.9339	1.1647	0.299

$$Cov(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$
$$= \hat{\sigma}^2 (X'X)^{-1}$$
$$= 145.37 \begin{bmatrix} 57.406 & 0.435 & -0.528 \\ 0.435 & 0.028 & -0.009 \\ -0.528 & -0.009 & 0.006 \end{bmatrix}$$
$$= \begin{bmatrix} 8345.11 & 63.23595 & -76.7554 \\ 63.23595 & 4.07036 & -1.30833 \\ -76.7554 & -1.30833 & 0.87222 \end{bmatrix}$$

$Var(\hat{\beta}_0) = 8345.11$ $\quad Se(\hat{\beta}_0) = \sqrt{8345.11} = 91.3516$

$Var(\hat{\beta}_1) = 4.07036$ $\quad Se(\hat{\beta}_1) = \sqrt{4.07036} = 2.0175$

$Var(\hat{\beta}_2) = 0.87222$ $\quad Se(\hat{\beta}_2) = \sqrt{0.87222} = 0.9339$

$t_{\beta_0} = \frac{-22.0801 - 0}{91.3516} = -0.2417$

$t_{\beta_1} = \frac{-0.1105 - 0}{2.0175} = -0.0548$

$t_{\beta_2} = \frac{1.0877 - 0}{0.9339} = 1.1647$

(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.

$$\begin{matrix} \beta_0 = \beta_1 \\ \beta_1 = \beta_2 \end{matrix} \Rightarrow \begin{matrix} \beta_0 - \beta_1 = 0 \\ \beta_1 - \beta_2 = 0 \end{matrix} \Rightarrow C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

H_0: $\theta = \begin{bmatrix} \beta_0 - \beta_1 \\ \beta_1 - \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, corresponding contrast matrix $C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$

Because X is full rank, so θ is estimable.
Also since C is full rank, so θ is testable.

$$M_{2\times2} = C(X'X)^{-1}C' = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 57.406 & 0.435 & -0.528 \\ 0.435 & 0.028 & -0.009 \\ -0.528 & -0.009 & 0.006 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$
$$= \begin{bmatrix} 56.564 & 0.926 \\ 0.926 & 0.052 \end{bmatrix} \qquad M^{-1} = \begin{bmatrix} 0.025 & -0.444 \\ -0.444 & 27.144 \end{bmatrix}$$

3

$$\hat{\theta} = C\hat{\beta} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} -22.0801 \\ -0.1105 \\ 1.0877 \end{bmatrix} = \begin{bmatrix} -21.9696 \\ -1.1982 \end{bmatrix}$$

degree of freedom: 2, 5

$$F\text{-}obs = \frac{(\hat{\theta}-\theta_0)' M^{-1} (\hat{\theta}-\theta_0)/a}{\hat{\sigma}^2} = \frac{[-21.9696 \;\; -1.1982] \begin{bmatrix} 0.025 & -0.444 \\ -0.444 & 27.144 \end{bmatrix} \begin{bmatrix} -21.9696 \\ -1.1982 \end{bmatrix} / 2}{145.37} = 0.095$$

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

H_0: $\theta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, corresponding contrast matrix $C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Because X is full rank, C is full rank, so θ is testable.

$$M_{2\times2} = C(X'X)^{-1}C' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 57.406 & 0.435 & -0.528 \\ 0.435 & 0.028 & -0.009 \\ -0.528 & -0.009 & 0.006 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.028 & -0.009 \\ -0.009 & 0.006 \end{bmatrix}$$

$$M^{-1} = \begin{bmatrix} 68.966 & 103.448 \\ 103.448 & 321.839 \end{bmatrix} \qquad \hat{\theta} = C\hat{\beta} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -22.801 \\ -0.1105 \\ 1.0877 \end{bmatrix} = \begin{bmatrix} -0.1105 \\ 1.0877 \end{bmatrix}$$

$$F_{obs} = \frac{(\hat{\theta}-\theta_0)' M^{-1} (\hat{\theta}-\theta_0)/a}{\hat{\sigma}^2} = \frac{\begin{bmatrix} -0.1105 & 1.0877 \end{bmatrix} \begin{bmatrix} 68.966 & 103.448 \\ 103.448 & 321.839 \end{bmatrix} \begin{bmatrix} -0.1105 \\ 1.0877 \end{bmatrix} / 2}{145.37} = \frac{356.789/2}{145.37} = 1.23$$

Df: 2, 5

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

$$\text{Correlation} = \frac{Cov(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_0) Var(\hat{\beta}_1)}} = \frac{63.23595}{\sqrt{8345.11 \times 4.07036}} = 0.3431$$

(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?

β_0 — the expected blood pressure when age and body weight the value zero.

β_1 — the expected increase in blood pressure for one unit increase in age.

β_2 — the expected increase in blood pressure for one unit increase in body weight.

The interpretation of β_0 is not meaningful because of no biological meaning for BP with age = 0, body weight = 0. To fix the problem, we can center the age variable and weight variable by subtracting the average of age and body weight from each observation respectively. In doing so, the intercept β_0 will be the expected blood pressure when age is at the observed average and body weight is at the observed average value.

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.

p_i is a parameter

$$\hat{y}_i = \begin{cases} 1, & Pr = p_i; \\ 0, & Pr = 1-p_i \end{cases}$$

Given p_i, $\hat{y}_i \sim \text{Bernoulli}(p_i)$

$$E(\hat{y}_i) = p_i$$

$$Var(\hat{y}_i) = p_i(1-p_i)$$

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

For $y_i > 120$, $\frac{p_1}{1-p_1} = \frac{3/8}{5/8} = \frac{3}{5} = 0.60$

For weight > 132, $\frac{p_0}{1-p_0} = \frac{3/8}{5/8} = \frac{3}{5} = 0.60$

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{0.60}{0.60} = 1$$

(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

I count this as correct, But what I meant is

	3	5	
weight >132	2	1	3
≤132	1	4	5
	>120	≤120	

$$\text{odd ratio} = \frac{\frac{2/3}{1-2/3}}{(1/5)/(4/5)} = 8$$

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

$$p = \frac{\exp(\beta_0 + \beta_1 \, age + \beta_2 \, wt)}{1 + \exp(\beta_0 + \beta_1 \, age + \beta_2 \, wt)} = \frac{\exp(-118.9085 + (-0.7111)\times 30 + 1.0373\times 133)}{1 + \exp(-118.9085 + (-0.7111)\times 30 + 1.0373 \times 133)}$$

$$= 0.093$$

(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.

$$\log(odds_B) = \hat\beta_0 + \hat\beta_1 \times age_B + \hat\beta_2 \times wt_B$$

$$\log(odds_A) = \hat\beta_0 + \hat\beta_1 \times age_A + \hat\beta_2 \times wt_A, \quad age_B = age_A, \quad wt_B = 10 + wt_A$$

$$\log(OR_{B vs A}) = \log(odds_B) - \log(odds_A) = \hat\beta_2 (wt_B - wt_A)$$

$$= \hat\beta_2 \times 10$$

$$OR_{B vs A} = e^{10\hat\beta_2} = e^{10(1.0373)} = 31984$$

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.

$Y_{ij} = X_{ij}\beta + b_i + \varepsilon_{ij}$

two families $i = 2$

four from a family, $j = 4$

(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).

unstructured covariance matrix in one family:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix}_{4\times 4} \quad 6$$

$\frac{4\times(4+1)}{2} = 10$ unique elements need to be estimated

For all individuals in the study

$$CoV = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} & 0 & 0 & 0 & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} & 0 & 0 & 0 & 0 \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ 0 & 0 & 0 & 0 & \sigma_{13} & \sigma_{23} & \sigma_{22} & \sigma_{24} \end{bmatrix}$$

(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.

Using compound symmetry covariance structure, we need to estimate 2 parameters: σ_b^2 and σ_w^2. For one family:

$$CS = \begin{bmatrix} \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \end{bmatrix}$$

For all 8 individuals:

$$CS = \begin{bmatrix} \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & 0 & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ 0 & 0 & 0 & 0 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 \\ 0 & 0 & 0 & 0 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 \\ 0 & 0 & 0 & 0 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 \end{bmatrix} \quad 8\times 8$$

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

Because of the small sample size in this dataset, we should use compound symmetric covariance matrix because it has fewer parameters than unstructured. If assumption for compound symmetry is not valid, we may need to force a compound symmetry structure with appropriate methods.

(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.)

We should use ML to compare the two models because the likelihood obtained for models with different fixed effects are not comparable when REML is used to estimate the models. REML maximizes the likelihood of the observed residuals, so different degrees of freedom btw two models, thus they're not comparable.

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)
A	1	100	1	1
	2	100	1	2
	3	100	1	3
B	1	100	2	1
	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:

$$y_{ijk} = \mu + \underset{\text{drug}}{\alpha_i} + \underset{\text{dose}}{\beta_j} + e_{ijk},$$

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: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, 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?

y = drugA dose1 dose2

$y_{600\times1}$, $X_{100\times4}$, $b_{4\times1}$, $e_{600\times1}$.

e follows a Gaussian distribution within cell.

100x4? —1

should be 600x4

(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	$\mu+\alpha_1+\beta_1$
A	2	$\mu+\alpha_1+\beta_2$
A	3	$\mu+\alpha_1$
B	1	$\mu+\beta_1$
B	2	$\mu+\beta_2$
B	3	μ

y=drugA dose1 dose2

(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	3	67284.9	22428.3	57.186	<.0001
Error	596	233751.2	392.2		
Corrected Total	599	301036.1			

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

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

y=drugA dose1 dose2 doseAd1 doseAd2

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	$\mu+\alpha_1+\beta_1+\gamma_{11}$
A	2	$\mu+\alpha_1+\beta_2+\gamma_{12}$
A	3	$\mu+\alpha_1$
B	1	$\mu+\beta_1$
B	2	$\mu+\beta_2$
B	3	μ

γ_{11} — the difference in drug effect for dose 1 versus dose 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

y = drugA dose

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	67203.6	33601.8	85.785	<.0001
Error	597	233844.9	391.7		
Corrected Total	599	301048.5			

~~or just say γ_{11} is the difference of drug effect at dose 1~~

~~categorical y = $y = \mu + \alpha_1 + \beta_1 + \beta_2$~~

$H_0: \beta_2 = 2\beta_1$

(f) (3pts) Compare the model using dose as a categorical variable (part (c)) and the model using dose as a interval variable (part (e)) 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.

categorical

$y = \mu + \alpha_1$ drug $+ \beta_1 (dose=1) + \beta_2 (dose=2)$

We can view the model using dose as an interval variable as a model nested in the categorical dose parameterization model.

$H_0: \beta_2 = 0$

$$F_{obs} = \frac{\frac{SSE(I) - SSE(C)}{df(I) - df(C)}}{SSE(C)/df(C)} = \frac{\frac{233844.9 - 233751.2}{597 - 596}}{233751.2/596} = 0.2389$$

numerial/interval

$df = 1, 596$

$y = \mu + \alpha_1$ drug $+ \beta_3$ dose

(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.

dose categorical

$$\mu_A: \frac{(\mu + \alpha_1 + \beta_1 + \gamma_{11}) + (\mu + \alpha_1 + \beta_2 + \gamma_{12}) + (\mu + \alpha_1)}{3} = \mu + \alpha_1 + \frac{\beta_1 + \beta_2 + \gamma_{11} + \gamma_{12}}{3}$$

$$\mu_B: \frac{(\mu + \beta_1) + (\mu + \beta_2) + \mu}{3} = \mu + \frac{\beta_1 + \beta_2}{3}$$

$$H_0: \mu_A = \mu_B \Rightarrow \mu + \alpha_1 + \frac{\beta_1 + \beta_2 + \gamma_{11} + \gamma_{12}}{3} = \mu + \frac{\beta_1 + \beta_2}{3}$$

$$\alpha + \frac{\gamma_{11} + \gamma_{12}}{3} = 0$$

dose=1	dose=2	
$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1$ ~~+~~ β_2	categorial
$\mu + \alpha_1 + \beta_3$	$\mu + \alpha_1 + 2\beta_3$	interval

10

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

df: 1, 594

(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.

dose categorical

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

$$\mu_A = \frac{100(\mu+\alpha_1+\beta_1+\gamma_{11}) + 100(\mu+\alpha_1+\beta_2+\gamma_{12}) + 50(\mu+\alpha_1)}{250}$$

$$= \frac{250\mu + 250\alpha_1 + 100\gamma_{11} + 100\gamma_{12} + 100\beta_1 + 100\beta_2}{250}$$

$$= \mu + \alpha_1 + \frac{2}{5}(\beta_1+\beta_2+\gamma_{11}+\gamma_{12})$$

$$\mu_B = \frac{100(\mu+\beta_1) + 100(\mu+\beta_2) + 50(\mu)}{250}$$

$$= \frac{250\mu + 100(\beta_1+\beta_2)}{250}$$

$$= \mu + \frac{2}{5}(\beta_1+\beta_2)$$

H_0: $\mu_A = \mu_B \Rightarrow \mu+\alpha_1+\frac{2}{5}(\beta_1+\beta_2+\gamma_{11}+\gamma_{12}) = \mu + \frac{2}{5}(\beta_1+\beta_2)$

$$\Rightarrow \alpha_1 + \frac{2}{5}(\gamma_{11}+\gamma_{12}) = 0$$

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & \frac{2}{5} & \frac{2}{5} \end{bmatrix}$$

df 1, 494

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	.	.	.

plug in the values, of β_i's ?

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

$$LDL = \beta_0 + \beta_1 \times age + \beta_2 \times dose + \beta_3 \times drug + \beta_4 \times drug \times dose + \varepsilon$$

Dose ~~was treated~~ as continuous here since only drug was used in the class statement.

(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.

Because drug 1 (drug B) was used as the reference group and embedded in the intercept.

(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?

Type I SS are from added-in-order tests, and they are mutually exclusive and together exhaustive pieces of the model SS. The sizes of Type I SS for a covariate depends on the order the covariate is added to the model, except when all predictors are uncorrelated.

Type III SS are from added-last tests, and they are SS for each variable if it was entered last in the model. The size of Type III SS tells how much variance being explained by this variable after accounting for all other variables. Here Age was added first in the model, so its Type I SS is much larger than its Type III error.

The variable added last into the model in added-in-order test is equivalent to the added-last test of this variable since SS from these two tests are SS explained by this variable beyond the other variables. This is the reason why for dose*drug the Type I SS is the same as the Type III SS.

(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.

$LDL = \beta_0 + \beta_1 age + \beta_2 dose + \beta_3 drug + \beta_4 drug\text{-}dose + \varepsilon$

drug A: drug 0

drug B: drug 1

drug A for individual 40:

$$[1 \quad 40 \quad \overline{dose} \quad 1 \quad \overline{dose}]$$

drug B for individual 40: because drug B was the reference.

$$[1 \quad 40 \quad \overline{dose} \quad 0 \quad 0]$$

where $\overline{dose}$ = grand mean of the dose variable

(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?

$H_0: \mu_1 = \mu_2$

$\theta = \mu_1 - \mu_2 = 0$

$$\theta = \mu_1 - \mu_2 = \cancel{\beta_0} + \cancel{\beta_1(40)} + \cancel{\beta_2(\overline{dose})} + \beta_3(1) + \beta_4(\overline{dose}) - [\cancel{\beta_0} + \cancel{\beta_1(40)} + \cancel{\beta_2(\overline{dose})} + \beta_3(0) + \beta_4(0)]$$

$$= \beta_3 + \beta_4(\overline{dose}) = [0 \quad 0 \quad 0 \quad 1 \quad \overline{dose}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = C\beta$$

$C = [0 \quad 0 \quad 0 \quad 1 \quad \overline{dose}]$

$\hat{\theta} - \theta = C\hat{\beta} - 0 = C\hat{\beta}$

$Var(\hat{\theta}) = M\sigma^2$

$Var(\hat{\theta}) = C\,Var(\hat{\beta})C'$

$M = \frac{C\,Var(\hat{\beta})C'}{\sigma^2}$

$$F_{obs} = \frac{(\hat{\theta}-\theta)' M^{-1} (\hat{\theta}-\theta)/1}{MSE} = \frac{(C\hat{\beta})^2 / [Var(\hat{\theta})/\sigma^2]}{MSE}$$

$$= \frac{(C\hat{\beta})^2 / [C\,Var(\hat{\beta})C'/\sigma^2]}{360.0801}$$

$$= \frac{(C\hat{\beta})^2}{C\,Var(\hat{\beta})C'}$$

df: 1, 595