1. (25pts) A new drug "B" has been developed to reduce cholesterol level. It was claimed that the new drug is more effective than the old one named "A". In a large scale study, each of these two drugs is tested on 500 patients at 5 doses, with 100 patients per dose, and thus the total sample size is 1000.

   (a) (3pts) First consider the dose variable as a factor with 5 levels, and employ an additive model:

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

   Using reference cell coding, where $i = 1$, and α_1 models the effect of drug A (drug B is reference); $j = 1, 2, 3, 4$, such that β_j models the effect for dose j (dose 5 is reference); k=1,2, ..., 100, which are patient indices within one cell, and e_{ijk} indicates residual error. If we write this ANOVA model as a regression model: $y = \mathbf{X}b + e$, what is the dimension of y, $\mathbf{X}$, b and e, and for an ANOVA model, what kind of distribution e should follow?

   y: 1000×1

   X: 1000×6

   b: 6×1

   e: 1000×1

   $e \sim N(0, \sigma^2 I)$

   (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	$\mu+\alpha_1+\beta_3$
A	4	$\mu+\alpha_1+\beta_4$
A	5	$\mu+\alpha_1$
B	1	$\mu+\beta_1$
B	2	$\mu+\beta_2$
B	3	$\mu+\beta_3$
B	4	$\mu+\beta_4$
B	5	μ

(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	5	146250	29250	70.14	<.0001
Error	994	414498	417		
Corrected Total	999	560748			

(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. If we write this ANOVA model as a regression model: $y = \mathbf{X}b + e$, what is the dimension of y, $\mathbf{X}$, b and e

$y : 1000 \times 1$

$X : 1000 \times 10$

$b : 10 \times 1$

$e : 1000 \times 1$

(e) (4pts) 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	$\mu+\alpha_1+\beta_3+\gamma_{13}$
A	4	
A	5	$\mu+\alpha_1+\beta_4+\gamma_{14}$
B	1	$\mu+\alpha_1$
B	2	$\mu+\beta_1$
B	3	$\mu+\beta_2$
B	4	$\mu+\beta_3$
B	5	$\mu+\beta_4$
		μ

$$\gamma_{11} = \left(E[y|A,1] - E[y|A,5]\right) - \left(E[y|B,1] - E[y|B,5]\right)$$

γ_{11} is the difference between the difference of dose 1 and dose 5 given drug A and B, respectively.

(f) (4pts) Let μ_A and μ_B be the overall mean values of cholesterol level for drug A and B, respectively. Write down 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 down H_0 in terms of α_i, β_j and γ_{ij}.

$$\mu_A = \frac{(\mu+\alpha_1+\beta_1+\gamma_{11})+(\mu+\alpha_1+\beta_2+\gamma_{12})+(\mu+\alpha_1+\beta_3+\gamma_{13})+(\mu+\alpha_1+\beta_4+\gamma_{14})+(\mu+\alpha_1)}{5}$$

$$\mu_B = \frac{(\mu+\beta_1)+(\mu+\beta_2)+(\mu+\beta_3)+(\mu+\beta_4)+\mu}{5}$$

$$H_0: \mu_A=\mu_B \iff H_0: \alpha_1 + \frac{\gamma_{11}+\gamma_{12}+\gamma_{13}+\gamma_{14}}{5} = 0$$

(g) (3pts) Give an example that $\mu_A = \mu_B$, but the effect of drug A and B are not the same for all the doses.

suppose $\gamma_{11}=\gamma_{12}=\gamma_{13}=\gamma_{14}/2 = -\alpha_1 \neq 0$

then
$$\begin{aligned} E[y|A,4] &= \mu+\alpha_1+\beta_4+\gamma_{14} \\ &= \mu+\alpha_1+\beta_4-2\alpha_1 \\ &= \mu+\beta_4-\alpha_1 \neq E[y|B,4] \end{aligned}$$

but clearly $\mu_A=\mu_B$

2. (15pts) Following question 1, we consider to include interval type of variables.

(a) (4pts) Now if we model dose as a interval variable, with doses equals to 1, 2, 3, 4, and 5, and fit a model of cholesterol level with additive effect of dose and drug, but no interaction, fill the following ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	145560	72780	174.95	<.0001
Error	997	414752	416		
Corrected Total	999	560312			

(b) (1pts) Is the model in 2(a) an ANOVA model, an ANCOVA model, or a full model in each cell?

ANCOVA model

(c) (4pts) Compare the model using dose as a categorical variable and the model using dose as a interval variable 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.

$$Ho: \beta_1 = 4\beta_4 \ \& \ \beta_2 = 3\beta_4 \ \& \ \beta_3 = 2\beta_4$$

$$\Longleftrightarrow Ho: \beta_1 - 4\beta_4 = 0 \ \& \ \beta_2 - 3\beta_4 = 0 \ \& \ \beta_3 - 2\beta_4 = 0$$

$$F\text{-test} = \frac{[SSE(R) - SSE(F)]/3}{SSE[F]/df_E}$$

$$= \frac{(414752 - 414498)/3}{414498/994} = \frac{84.67}{417} = 0.2 \sim F_{3,994}$$

Now we introduce another interval variable "age", and obtained the following output.

```
Dependent Variable: LDL   LDL cholesterol, mg/dL

                                          Sum of
      Source                   DF        Squares     Mean Square    F Value    Pr > F

      Model                     5    177854.2865      35570.8573      92.38    <.0001

      Error                   994    382744.6685        385.0550

      Corrected Total         999    560598.9550


                  R-Square     Coeff Var      Root MSE      LDL Mean

                  0.317258      15.32973      19.62282      128.0050


      Source                   DF     Type I SS     Mean Square    F Value    Pr > F

      drug                      1   123876.9000     123876.9000     321.71    <.0001
      dose                      1    21681.7710      21681.7710      56.31    <.0001
      age                       1    26676.8663      26676.8663      69.28    <.0001
      drug*dose                 1     2526.5193       2526.5193       6.56    0.0106
      drug*age                  1     3092.2299       3092.2299       8.03    0.0047
```

```
Source                    DF     Type III SS     Mean Square    F Value    Pr > F

drug                       1      188.590646      188.590646       0.49    0.4842
dose                       1     4596.699264     4596.699264      11.94    0.0006
age                        1     6103.211364     6103.211364      15.85    <.0001
drug*dose                  1     2443.995325     2443.995325       6.35    0.0119
drug*age                   1     3092.229910     3092.229910       8.03    0.0047
```

```
                                          Standard
Parameter              Estimate              Error    t Value    Pr > |t|

Intercept           98.82355769         3.53245943      27.98      <.0001
drug                 3.55006409         5.07268042       0.70      0.4842
dose                 2.14467577         0.62072607       3.46      0.0006
age                  0.29584942         0.07431096       3.98      <.0001
drug*dose            2.21124424         0.87770366       2.52      0.0119
drug*age             0.30090703         0.10618369       2.83      0.0047
```

(d) (2pts) Write down the fitted model based on the above output. Is dose treated as categorical or interval variable? Is this model an ANOVA model, an ANCOVA model, or a full model in each cell?

$$\hat{y} = 98.82 + 3.55\, I\{drug=A\} + 2.145 \cdot dose + 0.296 \cdot age + 2.211\, I\{drug=A\} \cdot dose + 0.3\, I\{drug=A\} \cdot age$$

In this model, dose is treated as an interval variable. This model is a full model in each cell.

(e) (2pts) Write down the fitted model when drug B is used (the reference level for variable drug), using cholesterol level as response, and using age ~~drug~~ and dose as covariates.

$$\hat{y} = 98.82 + 2.145\, dose + 0.296\, age$$

(f) (2pts) Write down the fitted model when drug A is used, using cholesterol level as response, and using ~~drug~~ age and dose as covariates

$$\hat{y} = 102.37 + 4.356\, dose + 0.596\, age$$

(b) (5pts) The result in the previous logistic regression suggest weight is not important, we tried to fit the following smaller model.

```
                         Model Fit Statistics

                                                 Intercept
                                  Intercept            and
               Criterion               Only     Covariates

               AIC                  541.990        287.592
               SC                   545.981        303.557
               -2 Log L             539.990        279.592

              Testing Global Null Hypothesis: BETA=0

      Test                 Chi-Square       DF     Pr > ChiSq

      Likelihood Ratio       260.3980        3         <.0001
      Score                  200.4918        3         <.0001
      Wald                    80.9970        3         <.0001


                       Type 3 Analysis of Effects

                                          Wald
        Effect               DF     Chi-Square     Pr > ChiSq

        strain                1         1.6216         0.2029
        activity              1        37.1344         <.0001
        activity*strain       1         8.3173         0.0039


                  Analysis of Maximum Likelihood Estimates

                                                 Standard          Wald
Parameter                 DF     Estimate        Error     Chi-Square     Pr > ChiSq

Intercept                  1      -7.2628       1.1901        37.2428         <.0001
strain            B6       1       1.5156       1.1901         1.6216         0.2029
activity                   1       1.7819       0.2924        37.1344         <.0001
activity*strain B6         1      -0.8433       0.2924         8.3173         0.0039
```

Compared these two models in part (a) and (b) by a likelihood ratio test. Write down H_0, test-statistic, degree of freedom and the distribution of the test statistic when H_0 is true.

H_0: $\beta_{weight} = \beta_{weight*strain} = 0$

LRT [redacted] $-2LR(Reduced)$
$+2LR(full)$

$= 279.59 - 279.38$

$= 0.21 \overset{H_0}{\sim} \chi^2_2$

3. (20 pts) In a mouse study, we are interested in tumor occurrences of 400 mice from two strains: 200 mice from B6 and 200 mice from Cast. Mice from one strain all share the same genetic background. This is a regression problem with one response, tumor occurrence, and three predictors: mouse strain (a binary variable), body weight (a continuous/interval variable), and activity index (an continuous/interval variable).

   (a) (5pts) In a simplified situation, we record 1 if a mouse has at least one tumor and 0 otherwise. Then tumor occurrence is a binary variable, and the results of a logistic regression is shown below:

```
                             Model Fit Statistics
                                                  Intercept
                                   Intercept            and
                    Criterion           Only     Covariates

                    AIC              541.990        291.381
                    SC               545.981        315.330
                    -2 Log L         539.990        279.381

                    Testing Global Null Hypothesis: BETA=0

        Test                   Chi-Square      DF     Pr > ChiSq

        Likelihood Ratio         260.6084       5         <.0001
        Score                    200.6430       5         <.0001
        Wald                      80.9250       5         <.0001

                    Analysis of Maximum Likelihood Estimates

                                                Standard          Wald
Parameter              DF     Estimate           Error    Chi-Square    Pr > ChiSq

Intercept               1      -7.2166          1.7079       17.8538        <.0001
strain          B6      1       1.9460          1.7079        1.2983        0.2545
weight                  1     -0.00269          0.0606        0.0020        0.9646
activity                1       1.7828          0.2931       36.9886        <.0001
weight*strain   B6      1      -0.0217          0.0606        0.1281        0.7205
activity*strain B6      1      -0.8427          0.2931        8.2642        0.0040
```

   Please write down the fitted model in the form of $E(y_i) = f(\hat{\beta})$ based on the above SAS output, where $\hat{\beta}$ are the regression coefficient estimates. What is $Var(y_i)$?

$$E(y_i) = f(\hat{\beta}) = \hat{p} = \frac{\exp\{g(\hat{\beta})\}}{1+\exp\{g(\hat{\beta})\}}$$

Where
$$g(\hat{\beta}) = -7.22 + 1.95\, I\{\text{strain} = B6\} - 0.0027\, \text{Weight} + 1.78\, \text{activity} - 0.0217\, I\{\text{strain} = B6\} \cdot \text{weight} - 0.8427\, I\{\text{Strain} = B6\} \cdot \text{activity}$$

$$Var(y_i) = \hat{p}(1-\hat{p}) = \frac{\exp\{g(\hat{\beta})\}}{(1+\exp\{g(\hat{\beta})\})^2}$$

In a follow-up study, we took 20 mice with tumor (10 from strain B6 and 10 from Cast) and 20 mice without tumor (10 B6 + 10 Cast), and measure the expression of a gene that is important in tumor progression at three tissues of each mouse: left forebrain, left hindbrain, and right whole brain. We have altogether (20+20)*3 = 120 measurements of gene expression.

(c) (2pts) Please describe the structure of the 120*120 covariance matrix of these 120 observations. How many elements of this matrix are expected to be 0?

block dragonal

120 × 120 − 3 × 3 × 40

= 14040 elements are expected to be 0

(d) (2pts) Here are the results of one mixed effect model, what kind of covariance structure are assumed for three expression measurements per mouse?

```
                 Estimated R Matrix for mouseID 1

              Row         Col1          Col2          Col3

               1        2.1015        0.6881        0.6881
               2        0.6881        2.1015        0.6881
               3        0.6881        0.6881        2.1015


                         Fit Statistics

              -2 Res Log Likelihood             417.4
              AIC (smaller is better)           421.4
              AICC (smaller is better)          421.5
              BIC (smaller is better)           424.8


                 Null Model Likelihood Ratio Test

                    DF     Chi-Square      Pr > ChiSq

                     1          11.11          0.0009


                   Type 3 Tests of Fixed Effects

                           Num       Den
          Effect            DF        DF    F Value    Pr > F

          tumor              1        37       4.43    0.0421
          strain             1        37      22.02    <.0001
```

why type I SS and type III SS in the following output are the same.

```
Dependent Variable: expression

                                           Sum of
         Source                 DF        Squares    Mean Square   F Value   Pr > F

         Model                   2     91.9864678     45.9932339     22.26   <.0001

         Error                 117    241.7472351      2.0662157

         Corrected Total       119    333.7337030


                    R-Square     Coeff Var      Root MSE    expression Mean

                    0.275628      158.7405      1.437434           0.905524


         Source                 DF      Type I SS    Mean Square   F Value   Pr > F

         tumor                   1    15.41973044    15.41973044      7.46   0.0073
         strain                  1    76.56673739    76.56673739     37.06   <.0001


         Source                 DF    Type III SS    Mean Square   F Value   Pr > F

         tumor                   1    15.41973044    15.41973044      7.46   0.0073
         strain                  1    76.56673739    76.56673739     37.06   <.0001
```

① the model violates the independence assumption!

② The pvalue gets small due to the independence assumption violation [redacted]

③ since the ~~expe~~ data is balanced & design matrix corresponding to tumor & strain are orthogonal.