1. *(40 points total)* A group of subjects was recruited to a nutritional study in a medical center at UNC. The data consist of their BMI (y = BMI), daily exercise time (x1 = exercise (in hours)) and daily vegetable intake (x2 = vegetable (in servings)). One of the objectives in this study is to estimate how the exercise and vegetable consumption affect BMI. To address the question, we consider the following model:

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \epsilon$$

. Let **X** be the associated design matrix of the above model. The data is summarized below:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 200.0000 & 588.7676 & 1033.797 \\ 588.7676 & 2321.7635 & 2951.400 \\ 1033.7973 & 2951.3999 & 7138.232 \end{pmatrix}, \mathbf{X}'\mathbf{y} = \begin{pmatrix} 4647.273 \\ 13561.768 \\ 23709.514 \end{pmatrix},$$

$$\text{and } (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.037523205 & -0.005495959 & -0.003161934 \\ -0.005495959 & 0.001712864 & 0.00008774738 \\ -0.003161934 & 0.00008774738 & 0.0005617387 \end{pmatrix}.$$

- (8 points) A partial ANOVA table is given below. Complete the table.

```
The GLM Procedure
Dependent Variable: y
Sum of
Source            DF        Squares       Mean Square     F Value     Pr > F
Model              ? 2      85.543209       ? 42.77        ? 7.65        -
Error              ? 197    1101.480037     ? 5.59
Corrected Total   199      1187.023246
```

- (8 points) Compute the least square estimates of the model parameters and their standard errors. Conduct the tests for the significance of each parameter (i.e., $H0 : \beta_1 = 0$, and $H0 : \beta_2 = 0$).

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{pmatrix} 24.878 \\ -0.231 \\ -0.1858 \end{pmatrix} \qquad \widehat{Cov(\hat{\beta})} = \hat{\sigma}^2 (X'X)^{-1}$$

$$\hat{\sigma}^2 = MSE = 5.59$$

For $Ho: \beta_1 = 0$

$$T = \frac{\hat{\beta}_1}{se(\hat{\beta})} = \frac{-0.231}{\sqrt{0.0017128 \times 5.59}} = -2.36 \overset{Ho}{\sim} t_{197}$$

For $Ho: \beta_2 = 0$

$$T = \frac{-0.1858}{\sqrt{0.0005617 \times 5.59}} = -3.32 \overset{Ho}{\sim} t_{197}$$

both $|t| > 1.96$
So reject Ho at 0.05.

- (8 points) Compute the 95% confidence interval for the BMI of individuals who on average exercise 2 hours and eat 6 servings of vegetables daily.

$\beta^* = \beta_0 + 2\beta_1 + 6\beta_2$

$\Rightarrow \hat{\beta}^* = \hat{\beta}_0 + 2\hat{\beta}_1 + 6\hat{\beta}_2 = 23.3$

$\widehat{Var}(\hat{\beta}^*) = (1 \quad 2 \quad 6)\, Var(\hat{\beta}) \begin{pmatrix} 1 \\ 2 \\ 6 \end{pmatrix} = 0.03789$

so CI of β^* is $\hat{\beta}^* \pm 1.96\, se(\hat{\beta}^*)$

$= (22.92, \ 23.68)$

- (8 points) Test $H_0 : \beta_1 = 3\beta_2$.

Let $\theta = \beta_1 - 3\beta_2$ then $H_0 : \theta = 0$

$$t = \frac{\hat{\theta}}{se(\hat{\theta})} = \frac{(0 \ 1 \ -3)\hat{\beta}}{\sqrt{\hat{\sigma}^2 (0 \ 1 \ -3)(X'X)^{-1}\begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix}}}$$

$$= \frac{0.327}{0.1868} = 1.75 \sim t_{197}$$

cannot reject H_0 since $|t| < 1.96$

- (8 points) Next we center the exercise and vegetable consumptions at their means, which are 1 hour and 5 servings respectively and refit the data with the new transformed variables. Fill in the cells with ? in the following table.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	? 23.7174	? 0.2541	? 93.37	-
newx1	? -0.231	? 0.0979	? -2.36	-
newx2	? -0.186	? 0.0596	? -3.13	-

newX1 = X1 - 1

newX2 = X2 - 5

so if original model is

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

then new model should be

$y = \beta_0^* + \beta_1^* newx_1 + \beta_2^* newx_2 + e$

with $\beta_0^* = \beta_0 + \beta_1 + 5\beta_2$

$\beta_1^* = \beta_1$

$\beta_2^* = \beta_2$

2. *(40 points total)* This study investigates how the four dose levels of Vitamin C (1, 2, 3 and 4 mg) and two delivery methods (orange juice or ascorbic acid) affect the length of odontoblasts (teeth) in 800 guinea pigs. The study is balanced, so for each dose and delivery method combination, 100 pigs are assigned.

- (14 points) first consider the dose variable as categorial and employ an additive model using reference cell coding (where ascorbic acid and dosage 1mg are used as references respectively):

$$y_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \beta x_{4i} + e_i$$

where α_is refer the dosage effects and β refers the effect of delivery method. Describe dummy variables x_{1i}, x_{2i}, x_{3i} and x_{4i}, based on which write down the cell mean of each group in terms of μ, α_is and β in the following table.

$x_{1i} = \begin{cases} 1 & \text{if the dose level is } 2 \\ 0 & \text{if the dose level} \neq 2 \end{cases}$ $x_{3i} = \begin{cases} 1 & \text{if dose level} = 3 \\ 0 & \text{otherwise} \end{cases}$

$x_{2i} = \begin{cases} 1 & \text{if dose level} = 3 \\ 0 & \text{otherwise} \end{cases}$ $x_{4i} = \begin{cases} 1 & \text{if delivering method is orange juice} \\ 0 & \text{otherwise} \end{cases}$

Delivery method	Dose	Mean
Orange juice	1	$\mu + \beta$
Orange juice	2	$\mu + \alpha_1 + \beta$
Orange juice	3	$\mu + \alpha_2 + \beta$
Orange juice	4	$\mu + \alpha_3 + \beta$
Ascorbic acid	1	μ
Ascorbic acid	2	$\mu + \alpha_1$
Ascorbic acid	3	$\mu + \alpha_2$
Ascorbic acid	4	$\mu + \alpha_3$

- (7 points) If we add the interaction terms between the delivery methods and dosage into the above model and express the new model in matrix notation $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$, what are the dimensions of $\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}$ and $\mathbf{e}$?

$y_{800 \times 1}$; $X : 800 \times 8$; $\theta : 8 \times 1$

$e : 800 \times 1$

- (12 points) Let μ_{orange} and $\mu_{ascorbic}$ be the overall means of the two delivery methods. Write down μ_{orange} and $\mu_{ascorbic}$ for the models with and without interaction terms. Derive the two **C** matrices for testing $H_0 : \mu_{orange} = 2\mu_{ascorbic}$ under the two models.

Without interaction : $\mu_{orange} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} + \beta$

$\mu_{ascorbic} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4}$

So $H_0 : \mu_{orange} = 2\mu_{ascorbic} \iff \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} - \beta = 0$

i.e. $C = (1 \ \frac{1}{4} \ \frac{1}{4} \ \frac{1}{4} \ -1)$ for $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta)^T$

With interaction: $\mu_{orang} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} + \beta + \frac{\gamma_{11}+\gamma_{12}+\gamma_{13}}{4}$

$\mu_{ascorbic} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4}$

H_0. $\mu_{orange} = 2\mu_{ascorbic} \iff \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} - \beta - \frac{\gamma_{11}+\gamma_{12}+\gamma_{13}}{4} = 0$

i.e. $C = (1, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, -1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4})^T$ for $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta, \gamma_{11}, \gamma_{12}, \gamma_{13})$

- (7 points) Next, treat the Vitamin C dosage as a continuous variable and fit a model with additive effects of the delivery method and vitamin C level, with no interaction. Is this model nested within Model

$$y_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \beta x_{4i} + e_i?$$

If yes, write down H_0 for comparing the two models and derive **C** matrix. What are the degrees of freedom of the corresponding F test under H_0?

yes When letting $\alpha_2 = 2\alpha_1$ and $\alpha_3 = 3\alpha_1$ we basically assume that the dosage level as a continuous variable.

So the C matrix assuming $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta)^T$

$$C = \begin{bmatrix} 0 & -2 & 1 & 0 & 0 \\ 0 & -3 & 0 & 1 & 0 \end{bmatrix}$$

the degrees of freedom of F ~~is~~ are 2, 795.

3. *(20 points total)* Table below lists a data set derived from a study on the relationship between incubation temperature for hatching turtle eggs and gender of baby turtles.

Temperature	Male	Female	Total
low	10	40	50
medium	28	22	50
high	34	16	50

To study how the incubation temperature affects the sex of baby turtles, we fit the logistic regression model

$$logit(p) = \mu + \beta_1 I(temperature = low) + \beta_2 I(temperature = medium)$$

where p is the probability of hatching a male turtle, and get the following output:

```
                        Estimate    Std. Error  z value Pr(>|z|)
intercept                 0.7538     0.3032     2.486    0.0129
I(temperature=low)       -2.1401     0.4657    -4.595    4.33e-06
I(temperature=medium)    -0.5126     0.4160    -1.232    0.2179
```

- (10 points) Estimate the probability that a male turtle hatches from an egg incubated at medium temperature.

$$\log \frac{\hat{p}}{1-\hat{p}} = 0.7538 - 0.5126$$

$$\Rightarrow \hat{p} = 0.56$$

- (5 points) What is the estimate of the odds ratio of low vs high temperatures and construct a 95% confidence interval for this odds ratio.

$$\log(OR) = \frac{\log\left(\frac{\hat{p}_{low}}{1-\hat{p}_{low}}\right)}{\left(\frac{\hat{p}_{high}}{1-\hat{p}_{high}}\right)} = (\hat{\mu}+\hat{\beta}_1) - (\hat{\mu})$$

$$= \hat{\beta}_1 = -2.1401$$

So CI of OR is

$$[\exp(\hat{\beta}_1 - 1.96 \times 0.4657), \exp(\beta_1 + 1.96 \times 0.4657)]$$

$$= [0.0472, 0.293]$$

- (5 points) What is the estimate of the odds ratio of low vs medium temperatures. Do you have enough information to construct a 95% confidence interval for this odds ratio? If yes, construct the CI. If not, explain why.

point estimate

$$\hat{OR} = \exp\{\hat{\mu} + \hat{\beta}_1 - \hat{\mu} - \hat{\beta}_2\}$$

$$= \exp\{\hat{\beta}_1 - \hat{\beta}_2\} = \exp\{-2.1401 + 0.5726\}$$

$$= 0.196$$

to get confidence interval, we need $Cov(\hat{\beta}_1, \hat{\beta}_2)$ which is not available to us. So cannot get the CI.