BIOS663 Midterm Spring 2013
Thursday, March 7, 2013

Instructions: Please be as rigorous as possible in all of your answers and show all your work.

Please sign the honor code pledge and submit it with your report. Violation of the honor code below will be prosecuted (penalties may include failure of the course and expulsion from the university).

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

Name:

Signature:

Date:

1. *(20 points total)* MULTIPLE CHOICE QUESTIONS (Please circle the best answer).

   - *(5 points)* Which choice is not an appropriate description of $\hat{y}$ in a regression model?
     A. Estimated response
     B. Predicted response
     C. Estimated average response
     D. Observed response

     Solution: D

   - *(5 points)* Which of the following is the best way to determine whether or not there is a statistically significant linear relationship between two variables?
     A. Compute a regression line from a sample and see if the sample slope is 0.
     B. Compute the correlation coefficient and see if it is greater than 0.5 or less than 0.5.
     C. Conduct a test of the null hypothesis that the population slope is 0.
     D. Conduct a test of the null hypothesis that the population intercept is 0.

     Solution: C

   - *(5 points)* Which of the following case diagnostic measures is based on Y values only (and not X values)?
     A. Cooks distance
     B. Studentized residual
     C. Leverage
     D. None of the above

     Solution: D

   - *(5 points)* Which of the following is NOT true for the linear regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i (i = 1, \cdots, 100)$ where all 5 model assumptions hold.
     A. $\hat{\beta}_1 + \beta_2$ is a statistic
     B. $\hat{y}_i$ can be uniquely predicted from the above model
     C. β_1 may not be always estimable
     D. the residuals from the model are summed to 0

     Solution: A

2. *(40 points total)* You are working on a statistical consulting lab. One day, a client came with a gas consumption data. In this study, the client is interested in modeling the fuel efficiency of automobiles. A typical measure of fuel efficiency used by EPA and car manufactures is "gallons/100 miles". The client collected data on 100 cars. He measured two explanatory variables, x1=weight (in unit of 1000lb); and x2=number of cylinders. He also measured the fuel efficiency of each car (in "gallons/100 miles"). Let $\mathbf{X} = (\mathbf{J}_n, \mathbf{x}_1, \mathbf{x}_2)$ and the linear regression model considered is

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + error.$$

Potentially helpful results:

$$(X'X)^{-1} = \begin{bmatrix} 0.308 & -0.06 & -0.017 \\ -0.06 & 0.025 & -0.004 \\ -0.017 & -0.004 & 0.006 \end{bmatrix} \text{ and } X'y = \begin{bmatrix} 405 \\ 1402 \\ 2350 \end{bmatrix}.$$

(a) *(8 points)* A partial ANOVA table for testing the association of the three covariates with the response y is given below. Complete the table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	???	79	???	???	<0.001
Error	???	11	???		
Corrected Total	???	???			

Solution:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	79	39.5	349.6	<0.001
Error	97	11	0.113		
Corrected Total	99	90			

(b) *(6 points)* Fill in the cells with ??? in the following table.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	???	???	???	0.009
x1	???	???	???	<0.001
x2	???	???	???	<0.001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	0.67	0.187	3.58	0.009
x1	1.35	0.053	25.47	<0.001
x2	1.61	0.026	61.81	<0.001

(c) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = 1$.
Solution:
$t - test = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{1.35-1}{0.053} = 6.6 \sim t_{97}$ which is greater than 1.96, so reject the null hypothesis at $\alpha = 0.05$.

(d) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = \beta_2 = 1$.
Solution:
Let $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ & & \end{bmatrix}$ and $\boldsymbol{\theta}_0 = \begin{bmatrix} 1 \\ 1 \\ \ \end{bmatrix}$, then $H_0 : C\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

So $M = \mathbf{C}(X'X)^{-1}\mathbf{C}' = \begin{bmatrix} 0.025 & -0.004 \\ -0.004 & 0.006 \\ & \end{bmatrix}$ with $M^{-1} = \begin{bmatrix} 44.78 & 29.85 \\ 29.85 & 186.57 \\ & \end{bmatrix}$ and

$\hat{\boldsymbol{\theta}} = (1.35, 1.61)'$.

Thus

$F - test = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' M^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/2}{\hat{\sigma}^2} = \frac{43.83}{0.113} = 387.8 \sim F_{2,97}$

(e) *(6 points)* Find a 95% confidence interval of $\beta_1 + \beta_2$.

Solution:

Let $\theta = \beta_1 + \beta_2$ then $\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_2 = 2.96$.

$v\hat{a}r(\hat{\theta}) = (0, 1, 1) v\hat{a}r(\hat{\boldsymbol{\beta}})(0, 1, 1)' = 0.0026$ and $SE(\hat{\theta}) = \sqrt{0.0026} = 0.051$

95 % CI of θ is $2.96 \pm 1.96 * 0.051 = [2.86, 3.06]$

(f) *(8 points)* Now you decide to transform x1 and x2 to z1=x1 - 2 and z2=x2-4 where 2 and 4 refer the population minimal car weight and minimal number of cylinders. Refit data with the following linear model $y = \beta_0^* + \beta_1^* z1 + \beta_2^* z2 + error$. Please describe the meaning of β_0^* and fill in the following table:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	???	???	???	-
z1	???	???	???	???
z2	???	???	???	???

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	9.8	0.085	115.3	-
21	1.35	0.053	25.47	<0.001
22	1.61	0.026	61.81	<0.001

3. *(40 points total)* Consider the set of hypothetical data below $\mathbf{y}_{5\times1} = \mathbf{X}_{5\times3}\boldsymbol{\beta}_{3\times1} + \boldsymbol{\varepsilon}_{5\times1}$, where

$$\mathbf{y} = \begin{bmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 10 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 6 & 11 \\ 1 & 7 & 13 \\ 1 & 8 & 15 \\ 1 & 9 & 17 \\ 1 & 11 & 21 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

(a) *(5 points)* Is there a problem of multicollinearity in this regression? Prove or disprove that there exists no multicollinearity problem.

Solution: yes since the rank of the desing matrix is 2 instead of 3.

(b) *(5 points)* Can you compute OLS estimates of the three parameters and explain why.

Solution: No since the design matrix is not full rank.

(c) *(5 points)* Throwing out any redundant columns of the X matrix if necessary and re-express the model as $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ where $\mathbf{X}^*$ is full rank. Express $\boldsymbol{\beta}^*$ in terms of $\boldsymbol{\beta}$.

One solution is to let $\mathbf{X}^* = \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 11 \end{bmatrix}$ and $\boldsymbol{\beta}^* = \begin{bmatrix} \beta_0 - \beta_2 \\ \beta_1 + 2\beta_2 \end{bmatrix}$.

(d) *(5 points)* Suppose that there are two students in the Bios663 class whose names are Jim and Chris. Suppose further that they estimated the parameters β_0, β_1 and β_2 by trial and error. As a result, they got different answers, i.e., (-6, -10, 6) and (-10, -2, 2) respectively. And each of them argues that his answer is better. What do you think about these two answers? Which answer fits better to the data?

Solution: the two solutions are the same since they give the same estimated regression model.

(e) *(8 points)* Compute a 95% confidence interval for the mean response of individuals with $x_1 = 1$ and $x_2 = 1$. Do you think the model provides a good estimate for this mean response? Why?

Solution: SSE from the model is 0 and also we can check that the mean respose for $x_1 = 1$ and $x_2 = 1$ is estimable, so the 95% CI is $[-10, -10]$. The model does not provide a good estimate for this mean response since $x_1 = 1$ is far outside the range of the observed x_1 values.

(f) *(6 points)* Show as rigorously as possible whether $H_0 : \beta_0 - \beta_2 = 0 \ \&\beta_1 + 2\beta_2 = 2 \ \&2\beta_0 + \beta_1 = 2$ is testable. If not, can it be reduced to an equivalent testable hypothesis? If yes, present an equivalent testable hypothesis.

Solution: it is not testable but can be reduced to a testable hypothesis, such as $H_0 : \beta_0 - \beta_2 = 0 \ \& \beta_1 + 2\beta_2 = 2$.

(g) *(6 points)* Show as rigorously as possible whether $H_0 : \beta_0 + \beta_1 = 0$ is testable. If so, report your test.

Solution: it is not testable.

1. (10 pts) For a general linear regression problem with p covariates (including intercept and $p-1$ additional covariates) and sample size n, the regression model can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{e} \sim N(0, \sigma^2 I_{n\times n})$, and $\mathbf{X}$ is full rank.

   (a) (4 pts) What are the dimensions of matrix/vector of $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{e}$? What is the rank of $\mathbf{X}$? Please explain why $\text{cov}(\mathbf{e}) = \sigma^2 I_{n\times n}$ implies the assumptions of independence and homogeneity.

$y_{n\times 1}$ $X_{n\times p}$ $e_{n\times 1}$ rank$(x)=p$

$\text{Cov}(e) = \sigma^2 \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \Rightarrow \begin{cases} \text{independence since off } \text{diagonals are } 0 \\ \text{homogeneity since } Var(y_i)=\sigma^2 \text{ for all } i \end{cases}$

   (b) (4 pts) Derive the least squares estimates: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathsf{T}}\mathbf{y})$ by minimizing the least squares objective function, i.e., to minimize $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

See lecture note

   (c) (4pts) Calculate $E(\hat{\boldsymbol{\beta}})$ and $\text{cov}(\hat{\boldsymbol{\beta}})$.

See lecture note

2. (8pts) Now assume $\mathbf{e} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{cov}(\mathbf{e})$ and $\boldsymbol{\Sigma}$ is positive definite. Given the eigen-value decomposition of $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}'$, where $\boldsymbol{\Gamma}$ is a diagonal matrix and $\mathbf{V}$ is an orthonormal matrix such as $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = I_{n\times n}$, we define $\boldsymbol{\Sigma}^{-1/2} = \mathbf{V}\boldsymbol{\Gamma}^{-1/2}\mathbf{V}'$. Let $\tilde{\mathbf{y}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$, $\tilde{\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$, and $\tilde{\mathbf{e}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{e}$. We consider a linear regression problem $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$.

by matrix multiplication for diagonal matrix

(a) (2 pts) Please show $\text{cov}(\tilde{\mathbf{e}}) = I_{n\times n}$.

$$\text{cov}(\tilde{e}) = \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}}$$

$$= V\Gamma^{-\frac{1}{2}}\underline{V'V}\Gamma\underline{V'V}\Gamma^{-\frac{1}{2}}V'$$

$$= \underline{V\Gamma^{-\frac{1}{2}}\Gamma\Gamma^{-\frac{1}{2}}}V' = VV' = I$$

(b) (2 pts) For a linear regression problem $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, the formula for least squares estimates is: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y})$. Please use this formula to calculate the least squares estimates of regression coefficients $\hat{\boldsymbol{\alpha}}$ for the regression model $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$, in terms of $\mathbf{X}$, $\mathbf{y}$ and $\boldsymbol{\Sigma}$.

$$\hat{\alpha} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$$

$$= (X^T\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}X)^{-1} X^T\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}Y$$

$$= (X^T\Sigma^{-1}X)^{-1} X^T\Sigma^{-1}Y$$

(c) (4pts) Calculate $E(\hat{\boldsymbol{\alpha}})$ and $\text{cov}(\hat{\boldsymbol{\alpha}})$.

$$E(\hat{\alpha}) = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X\beta = \beta$$

$$\text{cov}(\hat{\alpha}) = (X^T\Sigma^{-1}X)^{-1}X^T\underline{\Sigma^{-1}\Sigma}\,\Sigma^{-1}X (X^T\Sigma^{-1}X)^{-1}$$

$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$

$$= (X^T\Sigma^{-1}X)^{-1}$$

3. (20pts) We are interested in data collected by the Environmental Protection Agency (EPA) at the Health Effects Research Laboratory at UNC: Chapel Hill. One hundred seventy young adult males received a battery of pulmonary function tests. Fit a model with average forced vital capacity (FVC) (in ml) as the outcome and height, weight, body mass index (BMI=$\frac{\text{weight (kg)}}{(\text{height (m)})^2}$), body surface area, age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity as predictors.

   (a) (5pts) To assess for possible co-linearity in the covariates, we perform PCA on the correlation matrix of this data. As shown in the following output, the 10-th eigen-value is very small, which means a particular

Eigenvalue decomposition of the Correlation Matrix

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.98457157	0.95976513	0.2985	0.2985
2	2.02480644	0.36097943	0.2025	0.5009
3	1.66382702	0.63279205	0.1664	0.6673
4	1.03103496	0.06376972	0.1031	0.7704
5	0.96726525	0.20929379	0.0967	0.8672
6	0.75797146	0.20748316	0.0758	0.9429
7	0.55048830	0.53278371	0.0550	0.9980
8	0.01770459	0.01584173	0.0018	0.9998
9	0.00186286	0.00139531	0.0002	1.0000
10	0.00046755		0.0000	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4
height	Height (cm)	0.429342	0.036906	0.384653	-.240983
weight	Weight (kg)	0.562292	-.092679	-.095943	0.026718
bmi		0.340930	-.147689	-.443854	0.239531
area	Body Surface Area (M**2)	0.566969	-.047500	0.092208	-.079656
age	Age (years)	0.084799	-.102998	-.198442	-.321636
avtrel	Average Treadmill Elevation (deg)	-.116240	-.026373	0.497176	0.182497
avtrsp	Average Speed of Treadmill (mph)	0.144346	0.094273	0.571216	0.156073
temp	Air Temperature (deg C)	0.084996	0.675223	-.123262	0.099538
barm	Barometric Pressure (mmHg)	0.070861	-.175834	-.013681	0.832373
hum	Relative Humidity %	0.089569	0.677455	-.095377	0.116735

Eigenvectors

	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
height	-.120483	-.361837	0.222180	0.018428	0.521355	0.375921
weight	-.012319	0.158716	0.071437	0.004011	-.639418	0.475397
bmi	0.092604	0.520638	-.117929	0.006137	0.557078	0.060451
area	-.061800	-.035176	0.137407	-.017784	-.093401	-.792758
age	0.856166	-.289687	-.149123	-.013509	-.001675	-.003181
avtrel	0.442067	0.455127	0.550162	0.007661	-.000498	-.001474
avtrsp	0.112098	0.166298	-.760874	0.019013	-.010800	0.001191
temp	0.096489	-.017527	0.055784	0.706187	-.007909	-.015994
barm	0.121960	-.502455	0.060375	0.005988	0.003385	-.001295
hum	0.087996	-.009950	0.046292	-.707073	0.009011	0.017050

linear combination of the covariates has small variance. Which linear combination it is? Explain why is it possible that this combination has small variance? Could this PCA captures co-linearity between intercept and other covariates? and why?

approximatly 0.4 height + 0.5 weight ~~weight~~
— 0.8 area

No intercept effect has been removed from correlation matrix since it remove mean values

(b) (4pts) Consider a linear regression model with all the covariates. Let $\beta = (\beta_0, \beta_1, ..., \beta_{10})^T$ be the intercept and the regression coefficients for height, weight, bmi, area, age, avtrel, avtrsp, temp, barm, and hum, respectively. Test the hypothesis: H_0: $\beta_1 = \beta_2 = 2\beta_4$ using general linear hypothesis. Please write down C and θ_0 so that the test can be written $C\beta = \theta_0$, and please write down the formula of test-statistic while denoting the data matrix for intercept and the 10 covariates by $\mathbf{X}$, and denoting the residual variance of this linear regression model by $\hat{\sigma}^2$.

$$C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & -2 & 0 & \cdots & 0 \end{pmatrix} \qquad \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$F = \frac{(\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)/2}{\hat{\sigma}^2}$$

$$M = C(X'X)^{-1}C'$$

(c) (7pts) After a few rounds of testing, we decide to have final model without area, temp, hum, and barm.

i. (2pts) Based on the following ANOVA table, what is the R^2? Please show your calculation and you may round those numbers to simplify the calculation.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	49984918	8330820	20.81	<.0001
Error	163	65242013	400258		
Corrected Total	169	115226931			

Root MSE	632.65927	R-Square	
Dependent Mean	5335.43235	Adj R-Sq	0.4130
Coeff Var	11.85769		

$$R^2 \approx \frac{50}{115}$$

ii. (2pts) Based on the following t-table, if we test whether the regression coefficient for age is 0 by added last test, what is the value of F-statistic, and what are the degrees of freedom?

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	4899.55068	13356	0.37	0.7142
height	Height (cm)	1	-32.70393	75.89989	-0.43	0.6671
weight	Weight (kg)	1	119.42970	92.38958	1.29	0.1980
bmi		1	-286.38260	297.80536	-0.96	0.3377
age	Age (years)	1	27.86381	13.61020	2.05	0.0422
avtrel	Average Treadmill Elevation (deg)	1	51.50263	37.65313	1.37	0.1733
avtrsp	Average Speed of Treadmill (mph)	1	755.16631	379.56118	1.99	0.0483

$$F = (2.05)^2$$

$$df = (1, 163)$$

iii. (3pts) Based on this reduced model with 6 covariates, which characteristics are associated with the best (largest) FVC?

Shorter heavier, low bmi, older, higher Avtrel and higher avtrsp

(d) (4pts) In the diagnosis of this model, we detect a few data points as outliers based on either leverage or cook's distance. Please explain what are the difference of leverage and cook's distance.

high leverage means outlier in X

large cook's distance means high influence on regression

4. (20pts) Consider a linear regression problem to study the association between the physical activity of 12 mice vs. environment (0 for standard environment and 1 for enriched one) and dosage of a drug (with dosage 0, 1, and 2).

observation	activity	environment	drug
1	102	0	0
2	97	0	0
3	102	0	1
4	82	0	1
5	108	0	2
6	111	0	2
7	95	1	0
8	100	1	0
9	106	1	1
10	110	1	1
11	118	1	2
12	116	1	2

(a) (4pts) First consider a linear model with two covaraites:

$$E(\texttt{activity}) = b_0 + b_1\texttt{environment} + b_2\texttt{dose}$$

If we write the above model by a matrix form: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, what are the meanings of $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{e}$, and what are their dimensions?

$y_{12\times1}$ $\quad X_{12\times3}$ $\quad e_{12\times1}$

response $\quad$ covariate $\quad$ error

(b) (4pts) Please calculate the correlation between two variables: environment and drug. For added in-order test, would the p-values for environment and drug remain the same for two orders: environment followed by drug; and drug followed by environment?

$$\mathrm{Cor}(\underset{X}{\underset{\uparrow}{en}}, \underset{Y}{\underset{\uparrow}{drug}}) = E(XY) - E(X)\,E(Y)$$

$$= \frac{\sum x_i y_i}{12} - \frac{\sum x_i}{12}\,\frac{\sum y_i}{12} = 0$$

(c) (8pts) Given the following regression coefficient estimates and type III ANOVA table.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	92.958	4.000	23.240	2.41e-09
enviorment	7.167	4.276	1.676	0.1281
drug	7.375	2.619	2.816	0.0202

Response: activity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
enviorment	1	154.08	154.08	2.8088	0.12806
drug	1	435.12	435.12	7.9321	0.02017
Residuals	9	493.71	54.86		

Please test the null hypothesis $H_0: b_1 = b_2 = 0$ using (1) general linear hypothesis testing and (2) comparison of the sum squares of two models. Write down your test statistic, its asymptotic distribution and the degree of freedom. You should plug in the numbers into your formula of test statistic but do not need to calculate it. If you need $(\mathbf{X}'\mathbf{X})^{-1}$. Simply use $(\mathbf{X}'\mathbf{X})^{-1}$ rather than the actual numbers.

$\theta = C\,b$

(1) GLH $\quad C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$F = \frac{(\theta - \theta_0)' M^{-1} (\theta - \theta_0) / 2}{\hat{\sigma}^2} \qquad df = (2, 9)$$

$$M = C(X'X)^{-1} C'$$

(2) Model 1 $\quad y \sim \beta_0$

Model 2 $\quad y \sim \beta_0 + \beta_1\, en + \beta_2\, drug$

$$F = \frac{(SS_2 - SS_1)/2}{SSE_2 / 9} = \frac{(154.08 + 435.12)/2}{493.71 / 9} \qquad df = (2, 9)$$

Model $y \sim \beta_0 + \beta_1 \text{drug}$ $R_1^2 = \frac{453}{154+453+494}$

Mode 2 $y \sim \beta_0 + \beta_1 \text{en} + \beta_2 \text{drug}$ $\rightarrow R_2^2 = \frac{453+154}{154+453+494}$

(d) (4pts) What is the R^2 of a smaller model with intercept and drug? What is the R^2 of a larger model with intercept, environment, and drug? Feel free to use approximations in your calculation. Then if we double the sample size from 12 to 24, while assuming the R^2 of these two models remain the same, what would be the F-statistic to test the null hypothesis that the regression coefficient for environment is 0.

$$F = \frac{(CSS_2 - CSS_1)/1}{SSE_2/(n-p)}$$

SSY = sum squares of Y

$$= \frac{(CSS_2 - CSS_1)}{(SSY - CSS_2)(n-p)}$$

$$= \frac{R_2^2 - R_1^2}{(1 - R_2^2)/(n-p)}$$

$$\frac{F_{new}}{F_{old}} = \frac{n_{new} - p}{n_{old} - p} = \frac{24-3}{12-3} = \frac{21}{9}$$

BIOS663 Midterm Exam Spring 2019
March 6, 2019.

Instructions: Please be as rigorous as possible in all of your answers and show all your work.

Please sign the honor code pledge and submit it with your report. Violation of the honor code below will be prosecuted (penalties may include failure of the course and expulsion from the university).

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

Name:

Signature:

Date:

1. *(20 points total)* Suppose $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim N(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 & 0.6 \\ 0 & 1 & 0.5 \\ 0.6 & 0.5 & 1 \end{bmatrix}$

   - (7 points) Derive the distribution of $2x_1 + x_2 - x_3$.

     Solution: Let $c = (2, 1, -1)$, then $var(2x_1+x_2-x_3) = c\mathbf{\Sigma}c^{'} = 2.6$ thus $2x_1+x_2-x_3$ follows a normal distribution with mean 0 and variance 2.6.

   - (7 points) Calculate $Cov(x_1 - x_2, 2x_2 + x_3)$.

     Solution: Let $c1 = (1, -1, 0)$ and $c2 = (0, 2, 1)$, then $Cov(x_1 - x_2, 2x_2 + x_3) = c1\mathbf{\Sigma}c2^{'} = -1.9$.

   - (6 points) Prove or dis-prove (with details) that $\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 0 & 2 \\ 4 & 1 & -2 \end{bmatrix}$ has linearly independent columns.

     Solution: Since $\|\mathbf{A}\| = 9 \neq 0$, the rank of $\mathbf{A}$ is thus full rank, and $\mathbf{A}$ has linearly independent columns.

2. *(40 points total)* Consider the model $\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 5 \\ 6 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} -2 \\ -1 \\ 1 \\ 0 \\ 3 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} -1 \\ -1 \\ 3 \\ -1 \\ -2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Potentially helpful facts:
A: the corrected *total* sum of squares is 17.2,
B: a generalized inverse (if $\mathbf{X}$ is full rank, this inverse is unique) of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^{-} = \begin{bmatrix} 0.21 & -0.02 & 0.02 \\ -0.02 & 0.07 & -0.02 \\ 0.02 & -0.02 & 0.08 \end{bmatrix}; \ \mathbf{X}'y = \begin{bmatrix} 17 \\ 16 \\ -5 \end{bmatrix} \text{ and}$$

C: 97.5 percentiles of student t-distributions:

df	1	2	3	4	5
	12.706	4.303	3.182	2.776	2.571

- *(8 points)* Compute the least square estimates of the model parameters and their standard errors.

  Solution: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = (3.15, 0.88, -0.38)^{'}$; $\hat{y} = \mathbf{X}\hat{\beta} = (1.77, 2.65, 2.89, 3.53, 6.17)'$ and $\hat{\epsilon} = y - \hat{y} = (0.23, -1.65, 0.11, 1.47, -0.17)^{'}$. Thus $\hat{\sigma}^2 = 2.49$ and $\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, leading to $se(\hat{\beta}_0) = \sqrt{0.52} = 0.72, se(\hat{\beta}_1) = \sqrt{0.17} = 0.41$, and $se(\hat{\beta}_2) = \sqrt{0.2} = 0.45$.

- *(8 points)* Compute the 95% prediction interval for a subject with $x_1 = 1$ and $x_2 = 2$.

  Solution: $\hat{y} = (1,1,2)\hat{\beta} = 3.27$ and $var(\hat{y}) = (1,1,2)\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}(1,1,2)' + \hat{\sigma}^2 = 3.9$ a 95% prediction interval is $3.27 \pm 4.3\sqrt{3.9} = (-5.3, 11.7)$.

- *(8 points)* Calculate the corrected $R^2{}_c$, interpret its value, and test the hypothesis that its corresponding population value is zero, that is, $H_0 : \rho_c^2 = 0$.

  Solution: Since $\bar{y} = 3.4$, we have $CSS(regression) = \sum_i \hat{y}_i^2 - 5\bar{y}^2 = 11.2$ $CSS(total) = \sum_i y_i^2 - 5\bar{y}^2 = 17.2$ and corrected $R^2{}_c = 11.2/17.2 = 0.65$.
  $H_0 : \rho_c^2 = 0$ is equivalent to $H_0 : \beta_1 = \beta_2 = 0$. Thus we have $F - test = \frac{11.2/2}{\hat{\sigma}^2} = 2.25 \sim F_{2,2}$.

- *(6 points)* Consider the following hypothesis test: $E(y \mid \text{covariates of individual 5}) = 2E(y \mid \text{covariates of individual 1})$. Give $\mathbf{C}$, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ that are associated with the hypothesis test. Show as rigorously as possible whether your $\boldsymbol{\theta}$ is testable. If so, test the hypothesis.

  Solution: $\beta_0 + 3\beta_1 - 2\beta_2 = 2(\beta_0 - 2\beta_1 - \beta_2)$ which is $\beta_0 - 7\beta_1 = 0$. Let $\mathbf{C} = (1, -7, 0)$, then $\theta = \beta_0 - 7\beta_1$. For $H_0 : \theta = 0$, the associated t-test $= \hat{\theta}/se(\hat{\theta}) = -3.01/3.124 = -0.96$. Since $\| -0.96\| < 4.3$, the hypothesis is not statistically significant given type I error of 0.05.

- *(5 points)* Show as rigorously as possible whether $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \\ \beta_0 + 2\beta_1 - 3\beta_2 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \\ 2 \end{pmatrix}$ is testable. If so, test the hypothesis. If not, can you construct an equivalent test that is testable? If yes, perform the equivalent test. If not, explain why.

  Solution: Let $\mathbf{C} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -2 \\ 1 & 2 & -3 \end{bmatrix}$, then $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$ which is estimable since $\mathbf{X}$ is full rank.

  The test is $H_0 : \boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$. Since $\mathbf{C}$ is not full rank, the test is not testable. Further, the test cannot be reduced to a testable hypothesis, since the three equations conflict with each other.

- *(5 points)* Show as rigorously as possible whether $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \\ \beta_0 + 2\beta_1 - 3\beta_2 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \\ 0 \end{pmatrix}$ is testable. If so, test the hypothesis. If not, can you construct an equivalent test that is testable? If yes, perform the equivalent test. If not, explain why.

Solution: Follow the above question, with the same $\mathbf{C}$ and , the test now is $H_0 : \boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ which again is not testable. However, we can reduce the above test to $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \end{pmatrix}$ or $\begin{pmatrix} \beta_0 + \beta_1 - \beta_2 \\ \beta_1 - 2\beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ with $\mathbf{C} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -2 \end{bmatrix}$ and $\hat{\boldsymbol{\theta}} = (4.41, 1.64)'$

F-test $= \frac{\hat{\boldsymbol{\beta}}'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1})\hat{\boldsymbol{\theta}}/2}{\hat{\sigma}^2} = 34.2/2.49 = 13.7 \sim F_{2,2}$.

3. *(40 points total)* An investigator at UNC conducted a survey of Chapel Hill residents both before and after construction of a new exercise trail. Before the trail was constructed, she determined the baseline physical activity levels of a number of Chapel Hill residents. After construction of the trail, she interviewed the same group of residents about their physical activity levels (after construction of the trail) along with their gender and age.

   Short descriptions of the variables of interest are provided below.

   - post: Average physical activity, measured in hrs per day, after construction of the trail.
   - pre: Physical activity, measured in hrs per day, before construction of the trail (baseline).
   - age: Age of each participant.
   - gender: Gender of each participant (Male =0 and Female=1).

   The investigator fit the following model, with data centered as indicated, to the physical activity data: $post = \beta_0 + \beta_1 pre + \beta_2 age + \beta_3 gender + error$. Let the design matrix of the model be $\mathbf{X}$, then the inverse of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.047 & -0.013 & 0 & -0.023 \\ -0.013 & 0.011 & 0 & 0 \\ 0 & 0 & 0.0000051 & 0 \\ -0.023 & 0 & 0 & 0.04 \end{bmatrix}.$$

Selected SAS output is also provided below.

```
                          The GLM Procedure

Dependent Variable: post
***Table One***
                                      Sum of
Source                     DF        Squares    Mean Square   F Value     Pr > F

Model                      ???        644.58        ???         ???        <.0001
Error                      ???        67.24         ???
Corrected Total             99         ???



***Table Two***
Standard Parameter         Estimate          Error        t Value    Pr > |t|
Intercept                  0.75               ???          ???        <0.0001
pre                        1.15               ???          ???        <0.0001
age                        0.052              0.0019       ???        <0.0001
gender                     ???                ???          6.43       <0.0001
```

Based on this output, answer the following questions:

- *(10 points)* Fill in the cells with ??? in Table One. What are the degrees of freedom associated with the F test?

```
                          The GLM Procedure

Dependent Variable: post
***Table One***
                                      Sum of
Source                     DF        Squares    Mean Square   F Value     Pr > F

Model                      (3)        644.58    (214.86)      (306.9)     <.0001
Error                      (96)        67.24     (0.7)
Corrected Total             99        (711.82)
```

- *(3 points)* Estimate σ^2.
  Solution: $\hat{\sigma}^2 = 0.7$.

- *(5 points)* Report a F-test of the hypothesis that the prior physical activity levels are unrelated to the post-construction physical activity, after adjusting effects of age and gender. Give the nested models implicitly being compared when one conducts this F-test.

Compare full model $post = \beta_0 + \beta_1 pre + \beta_2 age + \beta_3 gender + error$ with $post = \beta_0 + \beta_2 age + \beta_3 gender + error$ or testing $H_0 : \beta_1 = 0$.
t-test $= \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 1.15/\sqrt{\hat{\sigma}^2 * 0.011} = 13.1 \sim t(96) \approx N(0,1)$ thus significant at $\alpha = 0.05$. Thus F-test $= (t - test)^2 = 13.1^2 = 171.6 \sim F_{1,96}$.

- *(7 points)* Fill in the cells with ??? in Table Two.

```
 ***Table Two***
Standard Parameter         Estimate          Error         t Value    Pr > |t|
Intercept                  0.75              (0.181)       (4.14)     <0.0001
pre                        1.15              (0.088)       (13.1)     <0.0001
age                        0.052             0.0019        (27.4)     <0.0001
gender                     (1.07)            (0.167)       6.43       <0.0001
```

- *(4 points)* Test $H_0 : \beta_1 = 1$.
  Solution: t-test $= \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = 0.15/\sqrt{\hat{\sigma}^2 * 0.011} = 1.71 \sim t_{96} \approx N(0,1)$. Since $1.71 < 1.96$, the test is not significant at $\alpha = 0.05$.

- *(8 points)* What is the interpretation of the intercept in the aforementioned regression model? To make β_0 more interpretable, the investigator decides to rescale the variable age by its mean which is 40, and refit the following regression model:
  $post = \beta_0 + \beta_1 pre + \beta_2 newage + \beta_3 gender + error$
  where $newage = age - 40$. Fill in the cells with ??? in Table Three.

```
***Table Three ***
Standard Parameter         Estimate          Error         t Value    Pr > |t|
Intercept                  2.83               0.196        14.4     <0.0001
pre                        1.15               0.088        13.1     <0.0001
newage                     0.052              0.0019       27.4     <0.000
gender                     1.07               0.167        6.43     <0.0001
```

Solution: The intercept is the expected post construction physical activity per day for a male resident with age 0 and average physical activity per day of 0 hours before the construction.

(3 points) Explain the assumption of homogeneity in the context of this experiment. Is it possible to assess the validity of this assumption from the summary statistics given? If so, how?

Solution: Homogeneity means that variability of the random error is constant across all subjects. No way to assess this assumption without the residuals, or any way to compute them.

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: 1000×1

X: 1000×10

b: 10×1

e: 1000×1

(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	$\mu+\alpha_1+\beta_4+\gamma_{14}$
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

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

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.

$H_0: \beta_1 = 4\beta_4$ & $\beta_2 = 3\beta_4$ & $\beta_3 = 2\beta_4$

$\Leftrightarrow H_0: \beta_1 - 4\beta_4 = 0$ & $\beta_2 - 3\beta_4 = 0$ & $\beta_3 - 2\beta_4 = 0$

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

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

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

Dependent Variable: LDL LDL cholesterol, mg/dL

Source	DF	Sum of 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

Parameter	Estimate	Standard 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

Criterion	Intercept Only	Intercept and 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

Effect	DF	Wald 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

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.2628	1.1901	37.2428	<.0001
strain	B6	1	1.5155	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] $- 2 LR(\text{Reduced}) + 2 LR(\text{full})$

$= 279.59 - 279.38$

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

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

Criterion	Intercept Only	Intercept and 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

Parameter		DF	Estimate	Standard Error	Wald 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
$$g(\hat{\beta}) = -7.22 + 1.95\, I\{strain = B6\} - 0.0027\, weight + 1.78\, activity - 0.0217\, I\{strain = B6\} \cdot weight - 0.8427\, I\{strain = B6\} \cdot activity$$

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

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 hind-brain, 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

$120 \times 120 - 3 \times 3 \times 40$

$= 14040$ 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

Effect	Num DF	Den 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
```

Source	DF	Sum of 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	1[illegible].7435	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.

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 hind-brain, 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?

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

Compound Symmetry

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

Effect	Num DF	Den DF	F Value	Pr > F
tumor	1	37	4.43	0.0421
strain	1	37	22.02	<.0001

unstructured

(e) (3pts) Here are the results of the other mixed effect model, what kind of covariance structure are assumed for the three expression measurements per mouse in this model? Compare this model with previous one by a Likelihood Ratio test, write down test statistic, degree of freedom and the distribution of the test statistic when Null hypothesis is correct.

```
                    The Mixed Procedure

            Estimated R Matrix for mouseID 1

        Row        Col1          Col2          Col3

         1       2.4998        1.3469        0.1251
         2       1.3469        1.9588        0.5887
         3       0.1251        0.5887        1.8423

                      Fit Statistics

        -2 Res Log Likelihood           404.3
        AIC (smaller is better)         416.3
        AICC (smaller is better)        417.1
        BIC (smaller is better)         426.5

            Null Model Likelihood Ratio Test

              DF     Chi-Square       Pr > ChiSq

               5          24.21           0.0002

               Type 3 Tests of Fixed Effects

                       Num      Den
    Effect              DF       DF     F Value     Pr > F

    tumor                1       37        4.22     0.0471
    strain               1       37       23.26     <.0001
```

$LRT = 417.4 - 404.3$

$= 13.1 \sim \chi^2_4$

(f) (3pts) Someone ignored the fact that these mouse are not independent and did a fixed effect linear regression. Compared the following results with the results from question (e), explain (i) which assumption of general linear regression is violated, (ii) why we see smaller p-values in the fixed effect linear model? (iii) Give a reasonable guess

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 $\mathbf{X}$ 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 $H_0 : \theta = 0$

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

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

cannot reject Ho 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	? 0.05[illegible]6	? -3.[illegible]3	-

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.

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

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

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. $C\theta = (1 \ \frac{1}{4} \ \frac{1}{4} \ \frac{1}{4} \ -1)$ 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 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.

$$\log(OR) = \log\left(\frac{\left(\frac{\hat{p}_{low}}{1-\hat{p}_{low}}\right)}{\left(\frac{\hat{p}_{high}}{1-\hat{p}_{high}}\right)}\right) = (\hat{\mu}+\hat{\beta}_1) - (\hat{\mu}) = \hat{\beta}_1 = -2.1401$$

So CI of OR is

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

$$= [0.0472, 0.293]$$

- (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

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

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

$$= 0.196$$

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

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,

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

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 $E(\epsilon)$, 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

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

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

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

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

$t_{\hat{\beta}_0} = \frac{-22.0801 - 0}{91.3516} = -0.2417$

$t_{\hat{\beta}_1} = \frac{-0.1105 - 0}{2.0175} = -0.0548$

$t_{\hat{\beta}_2} = \frac{1.0877 - 0}{0.9339} = 1.1647$

(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.

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

$$M^{-1} = \begin{bmatrix} 0.025 & -0.444 \\ -0.444 & 27.144 \end{bmatrix}$$

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

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.

$$M_{2\times 2} = C(X'X)^{-1}C' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 57.906 & 0.435 & -0.528 \\ 0.435 & 0.028 & -0.009 \\ -0.528 & -0.009 & 0.006 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.028 & -0.009 \\ -0.009 & 0.006 \end{bmatrix}$$

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

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

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

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

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

(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	

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

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:

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

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

y = drugA dose1 dose2

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

e follows a Gaussian distribution within cell.

100×4 ? —1

should be 600×4

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

$H_0: \beta_2 = 2\beta_1$

categorical

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

numerial/interval

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

(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.

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

$H_0: \cancel{\beta_2 = 0}$

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

$df = 1, 596$

(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

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

$$= \mu + \alpha_1 + \frac{\beta_1 + \beta_2 + \gamma_{11} + \gamma_{12}}{3}$$

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

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

$\alpha_1 + \frac{\gamma_{11} + \gamma_{12}}{3} = 0$

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

df: 1, 594

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

(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)
	1	100	1	1
A	2	100	1	2
	3	50	1	3
	1	100	2	1
B	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_is?

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

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

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 for individual 40:

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

drug A: drug 0

drug B: drug 1

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?

$H_0: \mu_1 = \mu_2$

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

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

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

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

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

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

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

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

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

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

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

df: 1, 595