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: ●

   - *(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: ●

   - *(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: ●

   - *(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: ●

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					
Error					
Corrected Total					

(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)				
x1				
x2				

(c) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = 1$.

Solution:

(d) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = \beta_2 = 1$.

Solution:

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: [redacted]

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

Solution: [redacted]

(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 $\boldsymbol{\beta}$.

[redacted]

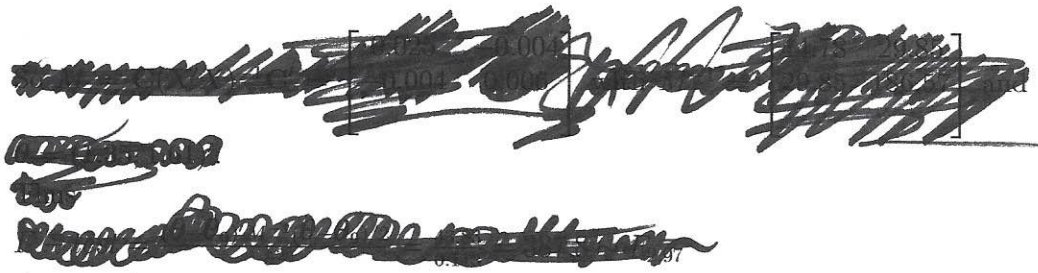
(d) *(5 points)* Suppose that there are two students in the Bios663 class whose names are Jim and Chris. Suppose further that they estimated the parameters β_0, β_1 and β_2 by trial and error. As a result, they got different answers, i.e., (-6, -10, 6) and (-10, -2, 2) respectively. And each of them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

[redacted]

(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?

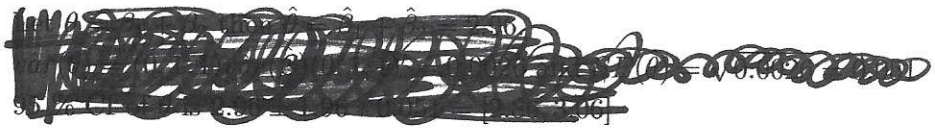
[redacted]

(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.



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

Solution:



(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)				
				

[redacted]

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

Solution: [redacted]