Full Name:

Midterm Exam
BIOSTAT 705 Spring 2023
March 9, 2023 (1:45 - 3:00 pm EST)

This is a closed book/notes exam, but you're allowed one-page as a formulae sheet. Use $\alpha = 0.05$ throughout unless the problem specifies otherwise. Be sure to give complete answers, state the hypothesis (H_0), and report p-value. Stat Tables will be provided. Show ALL work for partial credit!

1. For each question below circle either True or False (NO need to give reasons).

   a) (2 pts) Mallow's C_p for the full model is equal to number of predictors in the model? (T or F)

   b) (2 pts) If n is large, say $n > 30$, BIC imposes higher penalty as a variable selection criterion than AIC? (T or F)

   c) (3 pts) (i) β_1 is statistic? (T or F), (ii) β_1 is a parameter? (T or F),
   (iii) β_1 is unknown? (T or F)

   d) (1 pt) $\hat{\beta}_1 = \beta_1$? (T or F)

   e) (2 pts) If a new predictor added to a regression model, then error degrees of freedom and SSE will increase? (T or F)

   f) (2 pts) Adding a predictor to a linear regression model, R^2 and adjusted-R^2 will increase? (T or F)

   g) (2 pts) Parameters in polynomial linear regression model, can be estimated using least-squares method? (T or F)

   h) (1 pt) The "hat" matrix [$H = X(X'X)^{-1}X'$] is idempotent and its rank equal to the sum of its diagonal values? (T or F)

   i) (1 pt) LASSO regression will produce the least-squares estimates when tuning parameter $\lambda = 0$? (T or F)

   j) (2 pts) In a lack-of-fit setting, adding x^3 term to a quadratic model, pure-error df will remain the same? (T or F)

   k) (2 pts) In a lack-of-fit setting, adding x^3 term to a quadratic model, then SS(LoF) will decrease by SSR(x^3)? (T or F)

2. A linear regression model was fitted on 35 observations, with a dependent variable y and 3 predictors x_1, x_2 and x_3. Suppose a researcher fit 3 models below:

   Model 1: $y = \beta_0 + \beta_1 x_1 + \epsilon$ Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
   Model 3: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon$

   Below are the ANOVA tables for the 3 models:

Model 1

	df	SS
Regression		1448
Residual		152
Total		1600

Model 2

	df	SS
Regression		1569
Residual		31
Total		1600

Model 3

	df	SS
Regression		1571
Residual		29
Total		1600

   a) (5 pts) In the ANOVA tables above, complete the df column for each model?

   b) (10 pts) Test whether the interaction terms in model 3 are significant or not? State the null hypothesis, report the p-value and state your conclusion.

   c) (2 pts) Based your conclusion in part (b) above, is model 3 considered additive or multiplicative?

3. Consider the following outputs on several models using data with 40 observations:

variable(s) in model	SSR	SSE	F	R^2
X_1	2.4	404.2	0.2	0.006
X_2	172.5	234.1	24.3	0.424
X_3	323.4	83.2	128.4	0.796
X_1, X_2	173.1	233.5	11.9	0.426
X_1, X_3	329.7	76.9	68.6	0.811
X_2, X_3	327.2	79.4	65.9	0.805
X_1, X_2, X_3	333.7	72.9	47.3	0.822

a) (8 pts) What proportion of the variability of Y can be explained by the addition of X_3 to the model that has X_1 and X_2 in it?

b) (15 pts) Perform variable selection on the information given above using the forward selection to find the 'best' model? Support your answer with calculations as appropriate. (Use significance level for a variable to enter the model at 0.10).

4. A study was designed to assess the effect of two drugs on blood pressure(BP). Let y = BP after administering a drug, x = BP before administering a drug and $z = 1$ for drug A and $z = 0$ for drug B.

   Consider the model $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon$ and the following ANOVA tables.

Model	source	df	SS
y on x	regression	1	450
	error	18	240
y on x, z	regression	2	520
	error	17	170
y on x, z, xz	regression	3	530
	error	16	160

   a) (5 pts) Write two separate regression models (one for each drug).

   b) (8 pts) State the appropriate hypothesis if the two regression lines are coinciding? Perform the test, report p-value and state your conclusion.

   c) (7 pts) Are the two regression lines parallel? State the appropriate hypothesis, perform the test, report p-value and state your conclusion.

5. Assume you are asked to perform a stepwise variable selection procedure on a data set (n=40) containing 6 independent variables $X_1, X_2, \ldots, X_6$ and a continuous response Y. Unfortunately, you have been given only part of the output. The results are given below for Step 2 and include the partial F-statistics for each variable assuming X_2 and X_5 are in the model.

Variables in the Model

Variable	Coefficient	Partial F to Remove
Intercept	-10.5	
X2	-1.7	5.3
X5	3.9	11.8

Variables Not in the Model

Variable	Partial F to Enter
X1	1.1
X3	1.4
X4	1.2
X6	6.6

a) (5 pts) What is the estimated regression model at this step?

b) (7 pts) Using significance level to remove at 0.05, which variable (if any), will leave the model in Step 3 of the stepwise selection procedure? Support your answer.

c) (8 pts) Suppose no variable leaves the model in part (b), using significance level to enter at 0.05, which variable (if any), would enter the model in Step 4 of the stepwise selection procedure? Support your answer.