Name __________________________

Student Number __________________

Test 2 STA302F94

Aids Allowed: Calculator; formula sheet supplied

1. (10 pts) State the multiple linear regression model (with normal distribution assumption) in matrix notation. Be sure to specify the dimensions of each matrix (3 pts) as well as whether each constant or matrix of constants is known or unknown (3 pts). One of the terms on the right side of the equality is random. Be sure to state its distribution and parameters (2 pts).

2. (3 pts) For the multiple linear regression model, show that **b** is an unbiased estimator of $\boldsymbol{\beta}$.

3. (10 pts) For the <u>multiple</u> linear regression model, show that $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$. If you show this for simple regression instead, it is worth 2 points.

4. (12 pts) For the multiple linear regression model, show that the residual vector **e** has a multivariate normal distribution and give its mean and variance-covariance matrix. Leaving the mean and variance-covariance matrices unsimplified earns half marks.

Questions 5 through 12 refer to the following example. The registrar's office at Erindale decides to test the effectiveness of an orientation programme for new students. A random sample of entering students is selected, and students in the sample are randomly assigned to one of two conditions. They either are invited to the orientation (where they get baseball caps that say Erindale, free food and lots of advice) or they are not invited to the orientation. After one year the grade point average (GPA) of each student is recorded. A data file is assembled that has the values of three variables for each student: High school GPA (call it x_1), a variable x_2 that equals one of the student was offered the orientation programme and zero otherwise, and university GPA after 1 year (call it Y). Throughout, assume that the relationship between High School GPA and university GPA is linear.

5. (2 pts) What is the dependent variable?

6. (3 pts) For a model in which there is *no interaction* between orientation program and high school GPA, what is $E(Y_i)$?

7. (5 pts) For the model in which there is *no interaction* between orientation program and high school GPA, suppose we hold high school GPA constant a fixed (but unspecified) level. What is the difference in expected university GPA between students who get the program and those who do not?

8. (4 pts) For a model in which there is *no interaction* between orientation program and high school GPA, the following question is of primary interest: Once we control for high school grade point average, is there a difference between the university GPAs of students who get the orientation program and those who do not? **State the null hypothesis that should be tested to answer this question**. Use symbols and not words, or you will get no credit.

9. (5 pts) We are still considering the model in which there is *no interaction* between orientation program and high school GPA. Suppose $\beta_2<0$. What does that tell you about the effect of the program on academic performance?

10. (4 pts) We are still on the study assessing the effectiveness of the orientation program. Now consider a model in which the difference in University GPA between students who get the orientation program and those who do not <u>depends</u> on high school GPA. What is $E(Y_i)$ for this model?

11. (4 pts) Suppose you want to test whether the difference in University GPA between students who get the orientation program and those who do not depends on high school GPA. Using your notation from the preceding question, **State the null hypothesis**. Use symbols and not words, or you will get no credit.

12. (8 pts) Draw a set of X and Y axies with High School GPA on the X axis and University GPA on the Y axis. Suppose the orientation program helps students on the average regardless of High School GPA, but it helps good students less. Illustrate this situation by drawing two regression lines on your axies, one for those who get the program and one for those who do not. Label the lines clearly.

Now we are done with the orientation program example.

13. (6 pts) Consider a model in which $Y_i = \text{Exp}\{\beta_o + \beta_1 \cos(x_{i,1}) - \beta_2 \sqrt{x_{i,2}} + \varepsilon_i\}$ for i = 1, ..., 4. Convert this nonlinear model to a linear one and give the complete **X** matrix for the converted model. All you have to do is write down the **X** matrix, not **Y**.

14. (5 pts) Suppose we add a variable to a multiple regression model. *Write "Possible" or "Impossible" beside each statement* (I mean <u>mathematically</u> possible. You must get all three correct in order to get any credit. No explanation or proof is necessary.)

a) SSE can go up.

b) SSE can go down.

c) SSE can remain the same.

15. (5 pts) Write $SSR(X_2|X_1,X_3)$ in terms of SSE quantities (see formula sheet).

16. (5 pts) Show $SSR(X_1,X_2,X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1,X_2)$

17. For a large number of brands of chewing gum, we have sales in dollars (Y), advertising expenditures in dollars (X_1), distribution in percent of stores carrying the brand (X_2), and price per stick in cents (X_3). Assume all relationships are linear. Your answer to each item below should be SSTO or an SSE or SSR quantity.

a. (3 pts) Total amount of variation in the dependent variable.

b. (3 pts) With no other variables in the equation, this is how much of the variation in sales is explained by advertising expenditure.

c. (3 pts) Once we control for distribution and price, this is how much additional variation in sales is explained by advertising expenditure.

Total = 100 Points