Erindale College – University of Toronto
Faculty of Arts and Science
December Examinations 1994
STA 302F
Duration – 3 hours
Aids allowed: Calculator

Note: Formula sheet, printout and scratch paper supplied. *Scratch work is NOT to be handed in.*

Name (Please print) ____________________

Signature ____________________________

Student Number ______________________

1. (2 points) Recall the "cars" data set, in which expected gas mileage (Y) was potentially a function of country (represented by dummy variables $x_1$ and $x_2$), weight ($x_3$) and length ($x_4$). We want to know whether the car's length helps predict gas mileage once we control for weight and country. Give the **C, β** and **h** matrices for testing $H_0$: $\mathbf{C}\boldsymbol{\beta} = \mathbf{h}$ to answer this question. Give the complete matrices in order to get credit.

**Continued on page 2**

2. The statistical model for multiple linear regression may be written $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

a. (1 pt) The matrix $\mathbf{Y}$ has _____ rows and _____ column(s).

b. (1 pt) The matrix $\mathbf{X}$ has _____ rows and _____ column(s).

c. (1 pt) The matrix $\boldsymbol{\beta}$ has _____ rows and _____ column(s).

d. (1 pt) The matrix $\boldsymbol{\varepsilon}$ has _____ rows and _____ column(s).

e. (1 pt) Is the matrix $\mathbf{Y}$ a fixed constant or a random variable?

f. (1 pt) Is the matrix $\mathbf{X}$ a fixed constant or a random variable?

g. (1 pt) Is the matrix $\boldsymbol{\beta}$ a fixed constant or a random variable?

h. (1 pt) Is the matrix $\boldsymbol{\varepsilon}$ a fixed constant or a random variable?

3. (2 pts) One or more matrices in the multiple regression model has a multivariate normal distribution. For EACH of the multivariate normal matrices in the model, give the mean and variance–covariance matrices. There is no need to show any work or cite anything from the formula sheet; just write down the answer or answers.

**Continued on page 3**

4. (8 points) For a multiple regression model with 4 independent variables, we want to test $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. Derive a nice, **simple** expression for SSE for the reduced model, something you could readily compute given a set of sample data. Hint: You will need to differentiate, but don't bother with second derivative testing. **Circle your final formula for SSE(R).**

**Continued on page 4**

5. Over a narrow but meaningful range, blood pressure is usually a linear function of emotional stress. A new drug has been developed; the drug does not necessarily reduce blood pressure, but in the range where the relationship of stress and blood pressure is linear, it theoretically should make blood pressure completely unresponsive to stress level. To test this possibility, a study is conducted in which volunteers are randomly assigned to either take the drug or a sugar pill that looks like the drug. Within each of these two conditions, subjects are randomly assigned to experience some level of stress; all stress levels are within the range where the relationship of blood pressure to stress is linear. The data are as follows: Y = blood pressure, $x_1$ = stress level, and $x_2$ = 1 if the subject took the drug and $x_2$ = 0 otherwise.

a. (2 points) What is E(Y) for the full model? Include the interaction.

b. (1 point) What is E(Y) for subjects who take the drug?

c. (1 point) What is E(Y) for subjects who take the sugar pill?

d. (4 points) What is the null hypothesis needed to test the main theoretical prediction of the study? The right answer is NOT what you might guess.

**Continued on page 5**

6. (10 points) One and only one of the following statements is true in general for the multiple linear regression model. Either

(a) $\mathbf{e} = \boldsymbol{\varepsilon}$

(b) $\mathbf{Y} = \hat{\mathbf{Y}}$

(c) $\hat{\mathbf{Y}}'\mathbf{e} = \mathbf{Y}'\mathbf{e}$

(d) $\hat{\mathbf{Y}}'\mathbf{e} = \mathbf{0}$

(e) $\mathbf{X}'\mathbf{e} = \mathbf{X}'\boldsymbol{\varepsilon}$

(f) $\mathbf{X}\mathbf{b} = \mathbf{X}\boldsymbol{\beta}$

Select one and only one statement and prove that it is true. *Please do any preliminary work on scratch paper and write only a clean answer on this page.* You will get no credit for saying why the false statements are false, so don't bother. State clearly which statement you have proved.

**Continued on page 6**

7. (8 pts) This question uses the cars data from your first several computer assignments. I will give you the beginning of a SAS data step. Remember: 1=US, 2=Japanese, 3=Other.

```
data auto;
    infile 'cars.dat';
    input country mpg weight length;
    if country = 1 then c1=1; else c1=0;
    if country = 3 then c2=1; else c2=0;
```

*We want to test whether, controlling for weight and length, the expected gas mileage for "Other" cars is equal to the average of the expected value for Japanese cars and the expected value US cars.* Give SAS statements to accomplish this. First, you will need at least one more SAS statement in the data step. Then you need a single `proc reg` containing two `model` statements, one for the full model and one for the reduced model Do it this way, and NOT with the general linear test approach (or you will get no points). The models should contain NO interaction or polynomial terms. Show some work below, and then **write the SAS statements you need to add, and circle them**.

**Continued on page 7**

8. In an experiment designed to assess the effectiveness of training materials for computer users, a sample of business executives with no computer experience are timed while they try to set up a complete computer system out of the box. Half the executives are randomly assigned to set up Macintosh computers running System 7.5, and half are randomly assigned to set up IBM PCs running OS/2. Within each group, half the executives are randomly given a half–hour training video, and half are given an extra half–hour to look at the manuals. All participants have access to a full set of manuals while setting up the system. Time required to set up the system (well, actually natural log of time) is the dependent variable.

```
data frustrat;
     input computer video time;
     c_by_v = computer*video;
/*  Unfortunately, computer is coded 1=PC & 2=MAC, and video
    is coded 1=Yes & 2=No.  The Bozo who did this could have
    used if statements to create more convenient dummy
    variables, but he did not. */
proc reg;
      model time = computer video c_by_v;  /* Full model */
```

a. (4 pts) Fill in the table below, Putting E(Y) for the full model in each cell.

| | PC | Macintosh |
|---|---|---|
| Video | | |
| No Video | | |

**Continued on page 8**

We are still on the computer setup video example.

8b. (8 pts) In this two–by–two context, a good way to think about interaction is that the difference in expected setup time (video minus no video) may be different for Macs and PCs. Using the notation of your answer to part (a) of this question, state the null hypothesis for testing whether the difference in expected setup time (video minus no video) is different for the two types of computer (that is, the effectiveness of the video depends on type of computer). Guessing right without showing your work will get you half marks on this question.

**Continued on page 9**

9. (10 pts) Prove $\hat{Y}_h$ has a (univariate) normal distribution and give the parameters. In addition to common mathematical techniques (mostly matrix algebra) and the fact that a "multivariate normal" 1x1 vector is univariate normal, you are allowed to use ONLY what is on the formula sheet as a basis for your proof. When you do use something from the sheet, cite it.

**Continued on page 10**

The rest of the questions on this exam refer to the computer printout.

10. (4 pts) Please fill in the table below for the body fat data. You must get the whole row correct in order to get credit for each null hypothesis.

| Null Hypothesis | Value of F* or t* | p−value | Reject at $\alpha$=.05? Yes or no. |
|---|---|---|---|
| $\beta_1 = 0$ | | | |
| $\beta_2 = 0$ | | | |
| $\beta_3 = 0$ | | | |
| $\beta_1 = \beta_2 = \beta_3 = 0$ | | | |

11. (2 pts) How do the tolerance statistics help explain the answers to the preceding question?

12. (2 pts) We are still on the body fat data. For which observation is the vector of X values farthest from the average? Give an observation number from one to twenty, and the value of the appropriate diagnostic measure.

**Continued on page 11**

13. (2 pts) We are still on the body fat data. Which observation has the greatest influence on the value of $b_2$? Give an observation number from one to twenty, and the value of the appropriate diagnostic measure.

14. (2 pts) We are still on the body fat data. Which observation has the greatest influence on the value of the entire **b** vector? Give an observation number from one to twenty, and the value of the appropriate diagnostic measure.

15. (2 pts) Now we move to the **SMSA data**. For all the census tracts in the data (an SMSA is a census tract), what is the mean percentage of high school graduates?

16. (2 pts) We are still on the SMSA data. In the stepwise regression, what was the first variable that entered the model?

17. (2 pts) We are still on the SMSA data. The two automatic selection procedures appear to have arrived at the same model. What are the independent variables in this model?

18. (2 pts) We are still on the SMSA data. In the model described in the preceding question, what percentage of the variation in crime rate is explained by all the independent variables together? Don't correct for the number of variables.

19. (2 pts) We are still on the SMSA data. Of the 165 models with the same number of independent variables as the one that was automatically selected, what is the (uncorrected) $R^2$ value for the one explaining the <u>second</u> most variation in crime rate?

**Continued on page 12**

20. We are still on the SMSA data, using the model that was selected automatically.

a. (1 pt) Give an unbiased estimate of the population mean crime rate for SMSAs in the Southern region, provided that all other variables in the model are fixed at their mean levels. The answer is a single number.

b. (1 pt) Give a 95% confidence interval for your answer to (a). Do NOT do a Bonferroni correction! The answer is a pair of numbers.

21. (8 pts) We are still on the SMSA data, using the model that was selected automatically. Give a 95% PREDICTION interval for a single NEW observation with the same IV values as SMSA Number 1 (the first case in the file with real data). Hint: It's not on the printout but you can figure it out, and you do not need a table to get the exact $t(1-\alpha/2;n-p)$ value. Show your work.

**TOTAL MARKS = 100 POINTS**