UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER EXAMINATIONS 2006

STA 302 H1F / STA 1001 HF

Duration - 3 hours

Aids Allowed: Calculator

LAST NAME:________________________**FIRST NAME:**______________________

STUDENT NUMBER: ____

- There are 20 pages including this page.
- The last page is a table of formulae that may be useful. For all questions you can assume that the results on the formula page are known.
- Tables of the t distribution can be found on page 17 and tables of the F distribution can be found on pages 18 and 19.
- Total marks: 95

1,2	3	4	5	6	7a	7bc	7de

7fg	7hi	7j	8ab	8c	9ab	9cd

1. (5 marks) For each of the following models state whether its parameters can be estimated using standard linear regression techniques. If linear regression can be used, what are the independent and dependent variables that should be used if using `proc reg` in SAS?

   (a) $Y_i = \beta_0 + e^{\beta_1 X_i} + \epsilon_i$

   (b) $Y_i = \frac{1}{\beta_0 + \beta_1 X_i + \epsilon_i}$

   (c) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$ where β_2 is known to be 5.

2. (5 marks) A simple linear regression equation was fit to 10 observations. The following were calculated: $\sum_{i=1}^{10}(X_i - \overline{X})(Y_i - \overline{Y}) = 50$, $\sum_{i=1}^{10}(Y_i - \overline{Y})^2 = 100$, $R^2_{adj} = 0.37$. Complete the analysis of variance table with the column headings: Source of variation, Degrees of freedom, Sum of squares, Mean square, F ratio, and p-value.

3. Let b_0 and b_1 be the estimated intercept and slope from the simple linear regression of Y on X using n observations. Let c_1 and c_2 be non-zero constants.

   (a) (4 marks) Let b_0^* and b_1^* be the estimated intercept and slope from the simple linear regression of c_1Y on c_2X. Express b_0^* and b_1^* in terms of b_0, b_1, c_1, and c_2.

   (b) (4 marks) What is the effect of scaling X and Y to c_2X and c_1Y respectively on the test of $H_0: \beta_1 = 0$? Explain fully.

4. Consider the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ with non-random X. Assume that the usual assumptions hold.

   (a) (3 marks) Show that $b_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$ is an unbiased estimator of β_1.

   (b) (5 marks) Show that $E(\text{MSR}) = \sigma^2 + \beta_1^2 S_{XX}$ where $S_{XX} = \sum(X_i - \overline{X})^2$. Explain how this is related to the construction of the analysis of variance F-test.

Continued

5. Consider a model for regression through the origin $Y_i = \beta_1 X_i + \epsilon_i$ in which the ϵ's are independent and $\epsilon_i \sim N(0, \sigma^2 X_i)$. A possible solution to the non-constant variance is to use "weighted" least squares. In weighted least squares, the estimate of β_1 is obtained by minimizing

$$Q = \sum_{i=1}^{n} \frac{e_i^2}{\text{Var}(\epsilon_i)}$$

where e_i is the residual for the ith observation.

(a) (4 marks) Show that the resulting estimate of β_1 is

$$b_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$$

(b) (5 marks) Find the distribution of b_1 assuming that the X_i's are not random.

6. Consider the multiple linear regression model given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$. The vector $\mathbf{b} = (b_0, b_1, \ldots, b_k)'$ is the $(k+1) \times 1$ vector of least squares estimates and the vector $\mathbf{e}$ is the $n \times 1$ vector of residuals. Assume that $\mathbf{X}$ is not random.

   (a) (2 marks) What is $E(\mathbf{e})$? Justify fully.

   (b) (6 marks) The covariance matrix of $\mathbf{b}$ and $\mathbf{e}$ is defined as the $(k+1) \times n$ matrix given by

$$\mathbf{C} = E(\mathbf{b}\mathbf{e}') - E(\mathbf{b})E(\mathbf{e}').$$

   Show that $\mathbf{C}$ is a $(k+1) \times n$ matrix in which all the entries are zero.

7. An experiment was conducted to evaluate the effects of certain variables on soil erosion. Plots of sloped farm land were subjected to 5 cm of simulated rain for 20 minutes. The response is the amount of soil lost (variable: `soillost`) in kg/ha. The predictor variables of interest are: slope of the plot (variable: `slope`) in percent, percent ground cover (variable: `cover`), and porosity index (variable: `porosity`, low values are more porous).

Here are the data used in the analysis that follows:

soillost	slope	porosity	cover
27.1	0.43	1.95	0.34
35.6	0.47	5.13	0.32
31.4	0.44	3.98	0.29
37.8	0.48	6.25	0.30
40.2	0.48	7.12	0.25
39.8	0.49	6.50	0.26
55.5	0.53	10.67	0.10
43.6	0.50	7.08	0.16
52.1	0.55	9.88	0.19
43.8	0.51	8.70	0.18
35.7	0.48	4.96	0.28
.	0.50	6.00	0.20

(a) (1 mark) The last observation is not used in the regression. What is the purpose of including it?

Here is some SAS output for the simple linear regression with `cover` as the independent variable.

```
                          Dependent Variable: soillost

                              Analysis of Variance
                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F
Model                     1       569.03996      569.03996      40.15    0.0001
Error                     9       127.54549       14.17172
Corrected Total          10       696.58545

             Root MSE               3.76453    R-Square     0.8169
             Dependent Mean        40.23636    Adj R-Sq     0.7966
             Coeff Var              9.35605

                              Parameter Estimates
                             Parameter       Standard
     Variable      DF         Estimate          Error    t Value    Pr > |t|
     Intercept      1         64.57028        4.00442      16.12      <.0001
     cover          1       -100.25209       15.82098      -6.34      0.0001
```

(b) (3 marks) Give 90% simultaneous confidence intervals for the slope and intercept of the regression of `soillost` on `cover`.

Here is some SAS output for the simple linear regression with `porosity` as the independent variable.

```
                         Dependent Variable: soillost

                             Analysis of Variance
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1      668.14475      668.14475     211.43    <.0001
Error                     9       28.44070        3.16008
Corrected Total          10      696.58545

             Root MSE              1.77766    R-Square     0.9592
             Dependent Mean       40.23636    Adj R-Sq     0.9546
             Coeff Var             4.41805

                             Parameter Estimates
                          Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|
     Intercept     1       19.29777        1.53651      12.56      <.0001
     porosity      1        3.18921        0.21933      14.54      <.0001
```

(c) (3 marks) For the 2 simple linear regressions, which variable (`cover` or `porosity`) do you think is a better predictor of `soillost`? Why? What else would you like to see to answer this question and why do you want to see it?

(Question 7 continued)

Here is some SAS output from the multiple linear regression with independent variables `slope`, `porosity`, and `cover`.

```
                    The REG Procedure
              Dependent Variable: soillost

                  Analysis of Variance
                              Sum of           Mean
Source              DF       Squares         Square    F Value    Pr > F
Model                3     680.68178      226.89393      99.87    <.0001
Error                7      15.90368        2.27195
Corrected Total     10     696.58545

             Root MSE              1.50730    R-Square     0.9772
             Dependent Mean       40.23636    Adj R-Sq     0.9674
             Coeff Var             3.74611

                       Parameter Estimates
                    Parameter       Standard
 Variable    DF      Estimate          Error    t Value    Pr > |t|      Type I SS
 Intercept    1      -1.59534       18.04351      -0.09      0.9320          17809
 slope        1      76.45678       44.29509       1.73      0.1280      640.42489
 porosity     1       1.57585        0.73126       2.15      0.0681       33.04967
 cover        1     -23.77054       13.34612      -1.78      0.1181        7.20722
```

(d) (3 marks) Using matrix form, state the multiple linear regression model that is being fit and the model assumptions.

(e) (2 marks) What are the hypotheses for the analysis of variance F-test and what do you conclude?

(f) (4 marks) Calculate the coefficient of partial correlation for `cover` given `slope` and `porosity` are in the model. Explain how to interpret your result.

Here is some more of the SAS output for the multiple regression above.

```
                    Pearson Correlation Coefficients
                       Prob > |r| under H0: Rho=0

                soillost          slope       porosity          cover
soillost         1.00000        0.95884        0.97937       -0.90382
                                 <.0001         <.0001         0.0001

slope            0.95884        1.00000        0.93660       -0.82085
                  <.0001                        <.0001         0.0011

porosity         0.97937        0.93660        1.00000       -0.85255
                  <.0001         <.0001                        0.0004

cover           -0.90382       -0.82085       -0.85255        1.00000
                  0.0001         0.0011         0.0004
```

(g) (4 marks) Given this additional output is there any indication of multicollinearity? Is multicollinearity indicated in the output that proceeded the correlations (on page 9)? Why or why not?

Continued

(Question 7 continued)

The following SAS output was also obtained for the multiple regression above.

```
                         The REG Procedure
                   Dependent Variable: soillost

                         Output Statistics
    Dependent Predicted    Std Error
Obs  Variable       Value Mean Predict     95% CL Mean       95% CL Predict
  1   27.1000     26.2720       1.0275   23.8422  28.7018   21.9584  30.5856
  2   35.6000     34.8169       0.7501   33.0432  36.5905   30.8357  38.7980
  3   31.4000     31.4241       0.9078   29.2775  33.5707   27.2634  35.5848
  4   37.8000     37.8218       0.8249   35.8712  39.7724   33.7588  41.8848
  5   40.2000     40.3813       0.8699   38.3243  42.4384   36.2661  44.4965
  6   39.8000     39.9312       0.5173   38.7080  41.1543   36.1629  43.6994
  7   55.5000     53.3640       1.1012   50.7600  55.9680   48.9499  57.7781
  8   43.6000     43.9868       1.0964   41.3942  46.5793   39.5794  48.3941
  9   52.1000     51.5089       1.1946   48.6842  54.3336   46.9611  56.0567
 10   43.8000     46.8288       0.6498   45.2922  48.3654   42.9475  50.7101
 11   35.7000     36.2644       0.8224   34.3196  38.2092   32.2041  40.3246
 12         .     41.3340       1.3244   38.2024  44.4657   36.5895  46.0786
```

(h) (3 marks) The output on this page above includes 2 sets of intervals. Explain the difference between the 2 intervals for the first observation.

(i) (3 marks) Suppose you are interested in putting a 95% interval for the mean value of `soillost` for the entire range of the data. How should the first 11 intervals for the mean from the output above be used?

(j) (4 marks) Sketch typical residual plots that illustrate each of the following conditions. Clearly indicate what you are plotting.

i. The error variance increases with porosity index.

ii. There is a non-linear relationship with ground cover.

8. The data considered in this question are oxygen uptake amounts for salamanders. Of interest is whether or not oxygen uptake can be explained by relative head width of the salamander. Measurements were taken on 10 salamanders; 4 of the salamanders were lungless, and 6 were lunged. Some output from SAS follows. `lung_head` is an interaction term.

```
                              The REG Procedure
                         Dependent Variable: oxyup

                            Analysis of Variance
                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F
Model                     3      1677.00423      559.00141      47.18    0.0001
Error                     6        71.09577       11.84929
Corrected Total           9      1748.10000

             Root MSE              3.44228    R-Square     0.9593
             Dependent Mean       69.30000    Adj R-Sq     0.9390
             Coeff Var             4.96722

                            Parameter Estimates
                      Parameter      Standard
 Variable    DF        Estimate         Error   t Value   Pr > |t|      Type I SS
 Intercept    1        76.66667      10.69130      7.17     0.0004          48025
 head         1        33.33333      69.19244      0.48     0.6471     1056.13333
 lung         1        37.49814      12.98215      2.89     0.0278      423.23068
 lung_head    1      -325.98367      79.81866     -4.08     0.0065      197.64021
```

(a) (2 marks) Lung is a categorical variable. How does SAS `proc reg` deal with categorical variables when used as predictor variables? Give an example of how `lung` might be coded.

(b) (3 marks) Do the data give evidence that relative head width affects oxygen uptake? Explain.

(c) (4 marks) Carry out an hypothesis test to see if whether or not a salamander has lungs affects the relationship between oxygen uptake and relative head width.

9. Short answers are required for each of the following questions.

   (a) (4 marks) What is the effect of putting additional predictor variables in the model on each of the following?

      i. R^2

      ii. SSE

      iii. the estimated standard deviation of the errors

   (b) (3 marks) In polynomial regression, it is often recommended that quadratic and cubic terms be "centred" before being fit. What does this mean and why is this often done?

Continued

(c) (3 marks) Explain the practical difference between saying two predictor variables are *independent* and saying that two variables are *interacting*. Can the variables be both?

(d) (3 marks) A job training program was made available. Because more people wanted to enroll in it than the number of spaces available, the job skills of the potential participants were evaluated and the people with fewer job skills were enrolled in the training program. The following model was used to test the effectiveness of the job training program: $\log(\texttt{wage}) = \beta_0 + \beta_1 I_{\textit{[training]}} + \beta_2\,\texttt{education} + \beta_3\,\texttt{experience} + \epsilon$, where `wage` is the hourly rate of the job in which a person was employed 6 months after the training program ended, `education` is the number of years of education, `experience` is the number of years of job experience before the start of the training program, and $I_{\textit{[training]}}$ is an indicator variable with the value 1 for people who enrolled in the training program and 0 for people who did not take the training. Explain how the estimate of β_1 should be interpreted.

Continued

Continued

Continued

Simple regression formulae

$$b_1 = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2} = \frac{\sum X_iY_i - n\overline{X}\overline{Y}}{\sum(X_i-\overline{X})^2}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

$$\text{Var}(b_1) = \frac{\sigma^2}{\sum(X_i-\overline{X})^2}$$

$$\text{Var}(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i-\overline{X})^2}\right)$$

$$\text{Cov}(b_0, b_1) = -\frac{\sigma^2\overline{X}}{\sum(X_i-\overline{X})^2}$$

$$\text{SSTO} = \sum(Y_i - \overline{Y})^2$$

$$\text{SSE} = \sum(Y_i - \hat{Y}_i)^2$$

$$\text{SSR} = b_1^2\sum(X_i - \overline{X})^2 = \sum(\hat{Y}_i - \overline{Y})^2$$

$$\sigma^2\{\hat{Y}_h\} = \text{Var}(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h-\overline{X})^2}{\sum(X_i-\overline{X})^2}\right)$$

$$\sigma^2\{\text{pred}\} = \text{Var}(Y_h - \hat{Y}_h) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h-\overline{X})^2}{\sum(X_i-\overline{X})^2}\right)$$

$$\hat{X}_h \pm \frac{t_{n-2,1-\alpha/2}}{|b_1|} * \text{appropriate s.e.}$$

(valid approximation if $\frac{t^2s^2}{b_1^2\sum(X_i-\overline{X})^2}$ is small)

Working-Hotelling coefficient:

$$W = \sqrt{2F_{2,n-2;1-\alpha}}$$

$$r = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}$$

Regression in matrix terms

$$\text{Cov}(\mathbf{X}) = \text{E}[(\mathbf{X} - \text{E}\mathbf{X})(\mathbf{X} - \text{E}\mathbf{X})'] = \text{E}(\mathbf{X}\mathbf{X}') - (\text{E}\mathbf{X})(\text{E}\mathbf{X})'$$

$$\text{Cov}(\mathbf{A}\mathbf{X}) = \mathbf{A}\text{Cov}(\mathbf{X})\mathbf{A}'$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\text{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{Y}$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\text{SSR} = \mathbf{Y}'(\mathbf{H} - \tfrac{1}{n}\mathbf{J})\mathbf{Y}$$

$$\text{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\text{SSTO} = \mathbf{Y}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})\mathbf{Y}$$

$$\sigma^2\{\hat{Y}_h\} = \text{Var}(\hat{Y}_h) = \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$$

$$\sigma^2\{\text{pred}\} = \text{Var}(Y_h - \hat{Y}_h) = \sigma^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

$$R^2_{\text{adj}} = 1 - (n-1)\frac{MSE}{SSTO}$$

Total pages 20
Total marks 95