UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER EXAMINATIONS 2007
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 16 and tables of the F distribution can be found on pages 17, 18 and 19.
- Total marks: 95

1	2	3	4a	4bcd(i)	4d(ii)	4ef

5a	5b	5cdef	5gh	6ab	6cd	7

1. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$i = 1, \ldots, n$ where the ϵ_i are independent and identically distributed $N(0, \sigma^2)$ random variables. Assume that the X_i are not random. Let b_0 and b_1 be the least squares estimates of β_0 and β_1 respectively.

   (a) (3 marks) What is the distribution of b_1? (Just state it. You do not have to derive anything for this part.)

   (b) (3 marks) Let X_h be a value of the predictor variable that is of interest. Show that $\text{Var}(b_0 + b_1 X_h) = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{XX}} \right)$. You may use any formulae on the formula sheet that you find useful except the formula for $\text{Var}(\hat{Y}_h)$.

   (c) (2 marks) Show that the estimated regression line goes through $(\overline{X}, \overline{Y})$ where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Continued

2. Consider the multiple regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is an $n \times (k+1)$ matrix, $\boldsymbol{\beta}$ is a vector of length $k+1$, and $\boldsymbol{\epsilon}$ is the length-n vector of errors with $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$. Let $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ be the vector of least squares estimates of $\boldsymbol{\beta}$. You may treat the independent variables as non-random.

(a) (2 marks) Show that $\mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

(b) (2 marks) Let $\mathbf{X}_h' = (1, X_{h1}, \ldots, X_{hk})$ be a combination of values of the independent variables that is of interest. Derive the formula for the variance of the estimated mean value of Y at $\mathbf{X}_h$.

(c) (4 marks) Show that the least squares hyperplane goes through $(\overline{X}_1, \ldots, \overline{X}_k, \overline{Y})$ where $\overline{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$. You may use the fact that $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$ without proof. (*Hint*: What is $\frac{1}{n}\mathbf{X}'\mathbf{1}$ where $\mathbf{1}$ is a vector of n 1's?)

Continued

3. Consider a simple linear regression analysis with β_0 and β_1 both positive. Suppose that the mean of the response variable is μ_Y and that its variance is proportional to $\frac{1}{\mu_Y^2}$.

   (a) (1 mark) How would this be seen in the plot of the residuals versus the independent variable?

   (b) (2 marks) What is the appropriate variance-stabilizing transformation?

4. The data considered in this question are life expectancies for 188 countries in 2000 (`le2000`) and 1998 (`le1998`). We are interested in how well the 1998 value can be used to predict the 2000 value. Analysis is given below for the countries for which life expectancy is available in both 1998 and 2000.

   Some output from SAS is below. Some numbers have been purposely replaced by letters.

The REG Procedure

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	188.00000	1.00000	188.00000	0	0
le1998	12229	65.04840	819402	127.90818	11.30965
le2000	12259	65.20532	824692	135.65013	11.64689

The REG Procedure
Dependent Variable: le2000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	(A)	24210	24210	(C)	<.0001
Error	(B)	1156.54003	6.21796		
Corrected Total	187	25367			

Root MSE	2.49358	R-Square	(D)
Dependent Mean	65.20532	Adj R-Sq	0.9542
Coeff Var	3.82420		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.23786	1.06445	(E)	0.8234
le1998	1	1.00607	0.01612	62.40	<.0001

(a) (5 marks) Find the values of the number that have been replaced by letters in the output.

(A) = __________

(B) = __________

(C) = __________

(D) = __________

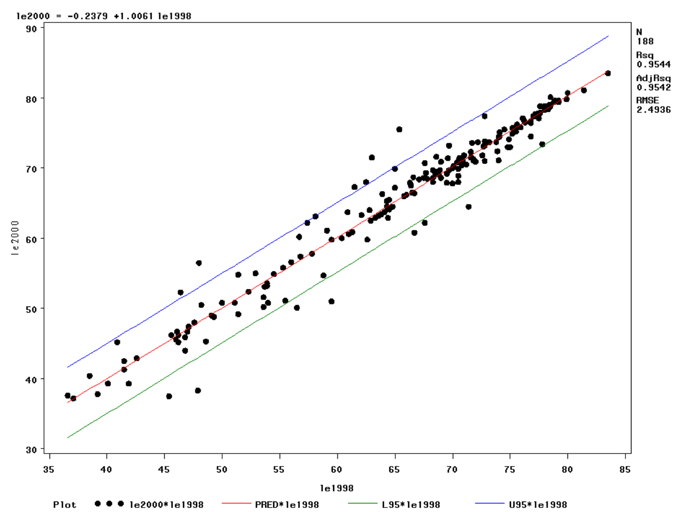
(E) = __________

Continued

(b) (2 marks) What is the value of the correlation between life expectancy in 1998 and life expectancy in 2000? Explain how it is related to the slope of the regression line.

(c) (4 marks) Carry out a two-sided hypothesis test to test whether the slope of the line is 1.

(d) The plot below shows the data, the fitted line, and lines joining the limits of 95% prediction intervals at each point on the line. Questions about the plot come after it and continue on the next page.

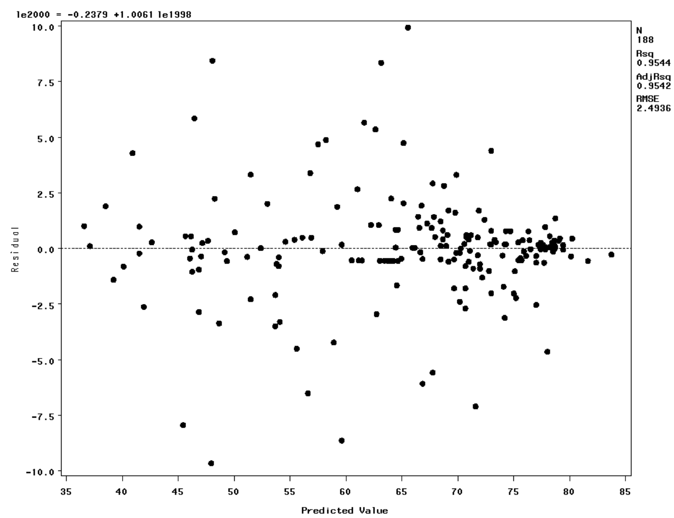


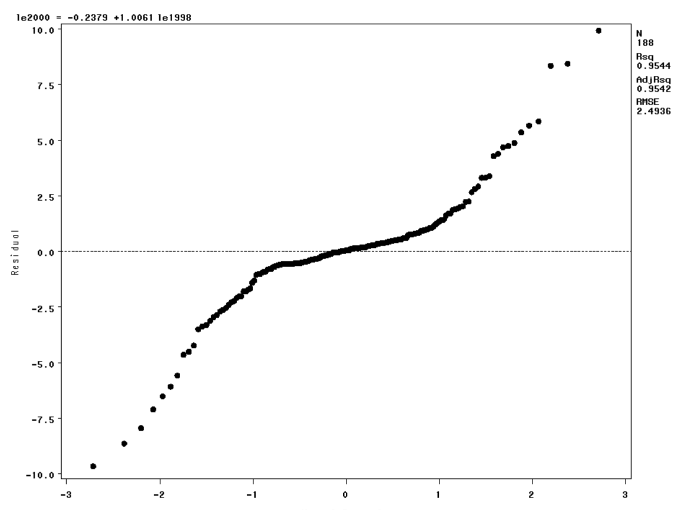
i. (2 marks) Are the lines in the plot parallel? Explain fully.

Continued

ii. (2 marks) There are several points that do not fall within the prediction intervals. Is this evidence that they are outliers? Why or why not?

(e) Below are a plot of the residuals versus predicted values and a normal quantile plot of the residuals for these data.





Questions about these plots are on the next page.

i. (4 marks) What are you looking for in the plot of the residuals versus the predicted values? What do you conclude?

ii. (2 marks) What additional information do you learn from the normal quantile plot of the residuals?

iii. (2 marks) How does the information learned from these plots affect your answers to parts (c) and (d) of this question?

(f) (3 marks) In this analysis, we are considering the relationship between the same variable measured at two time points. A possible criticism of the analysis is that we are witnessing an example of "regression to the mean". Explain what is meant by "regression to the mean". Does it appear to hold for these data? Why or why not?

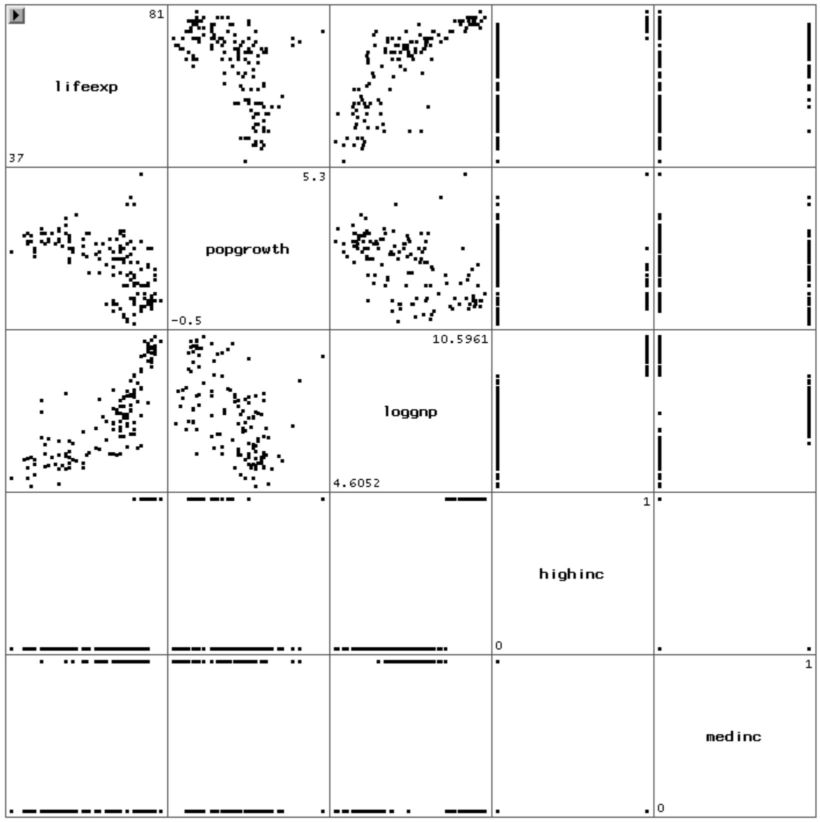
5. The data for this question were obtained from the World Bank web site. The purpose of the analysis is to examine the relationship between life expectancy (here we used the value in 1998) and the variables that are considered to be its important predictors. The variables included in the analysis are:
`lifeexp` – life expectancy at birth (in years) in 1998
`popgrowth` – average annual rate of population growth between 1980 and 1998 as a percentage
`logGNP` – natural logarithm of per capita gross national product, the total income that residents of the country earned in 1998
`income` – countries were classified as having High, Medium, or Low income economies

Output from SAS is given below. (There are fewer observations in this analysis than in the analysis in Question 4 because not all data were available on all countries.)

```
                          The REG Procedure
                     Dependent Variable: lifeexp

                        Analysis of Variance
                                 Sum of           Mean
Source                 DF       Squares         Square    F Value    Pr > F
Model                   4         12142     3035.55580      71.00    <.0001
Error                 131    5600.83564       42.75447
Corrected Total       135         17743

             Root MSE              6.53869    R-Square     0.6843
             Dependent Mean       65.14706    Adj R-Sq     0.6747
             Coeff Var            10.03682

                         Parameter Estimates
                     Parameter      Standard
 Variable     DF      Estimate         Error   t Value   Pr > |t|      Type I SS
 Intercept     1      33.37781       6.44914      5.18     <.0001         577203
 popgrowth     1      -2.17461       0.58807     -3.70     0.0003     6107.56364
 logGNP        1       4.60458       1.01913      4.52     <.0001     5626.46482
 highinc       1       0.15869       4.38806      0.04     0.9712      295.88890
 medinc        1       3.62977       2.23959      1.62     0.1075      112.30584
```

(a) (3 marks) The two variables labelled `highinc` and `medinc` are indicator variables, replacing the `income` variable described above. Describe how they are coded. Why is there no variable in the SAS output to indicate the countries that have a Low income economy?

(b) (5 marks) Here are the pairwise scatterplots of the variables used in the above analysis. Which are useful? What do they tell you?



Continued

(c) (2 marks) The log of gross national product was taken before any analysis was carried out. Why would this be done?

(d) (1 mark) What is the fitted regression equation?

(e) (1 mark) What is the estimated standard deviation for the distribution of the points about the regression hyperplane?

(f) (6 marks) Interpret the practical meaning of the coefficients of each of the following variables:

i. `popgrowth`

ii. `logGNP`

iii. `highinc`

Continued

(g) (3 marks) What hypothesis is being tested with the test that has a p-value of 0.0003. What do you conclude?

(h) (5 marks) Since the p-values associated with the coefficients for `highinc` and `medinc` are both large, it could be concluded that the income classification has no statistically significant effect on predicting life expectancy and could be removed from the model. Is this a valid conclusion? Why or why not? Support your answer with appropriate hypothesis test(s).

6. The following questions require short answers.

(a) (3 marks) In a simple linear regression analysis, explain the differences between σ^2 and s^2 and $\text{Var}(b_1)$.

(b) Suppose in a multiple regression analysis, it is of interest to compare a model with 3 independent variables to a model with the same response and these same 3 independent variables plus 2 additional independent variables.

i. (2 marks) Explain why the model with 5 predictor variables will have higher R^2.

ii. (2 marks) Explain why the partial F-test for the coefficients of the 2 additional predictor variables is equivalent to testing that the increase in R^2 is statistically significant.

iii. (1 mark) Show that the ranking of the competing models using adjusted R^2 is equivalent to using s^2.

(c) (4 marks) What is multicollinearity and why is it important?

(d) (2 marks) State 2 criticisms of automated variable selection techniques such as stepwise regression.

Continued

7. (10 marks) In question 5, the relationship between life expectancy and some economic variables was considered. Suppose now we are interested in building a regression model that examines the relationship between life expectancy and literacy rate, after controlling for the effects of the economic variables. Also, suppose that each country has been classified as developing or developed and we are interested in whether the relationship with literacy rate differs between these two classifications of countries. Describe how you would carry out a regression analysis to examine these interests. Describe what plots you would look at and what statistics you would consider and what information you would want to learn from each plot or statistic.

Continued

Continued

Continued

Continued

Continued

Simple regression formulae

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

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

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

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

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

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

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

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

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

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

$\hat{X}_h \pm \frac{t_{n-2,1-\alpha/2}}{|b_1|} *$ appropriate s.e.
(valid approximation if $\frac{t^2s^2}{b_1^2\sum(X_i-\overline{X})^2}$ is small)

Working-Hotelling coefficient:
$W = \sqrt{2F_{2,n-2;1-\alpha}}$

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

$$S_{XX} = \sum_{i=1}^n(X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2$$

Regression in matrix terms

$$\text{Var}(\mathbf{Y}) = \text{E}[(\mathbf{Y} - \text{E}\mathbf{Y})(\mathbf{Y} - \text{E}\mathbf{Y})'] = \text{E}(\mathbf{Y}\mathbf{Y}') - (\text{E}\mathbf{Y})(\text{E}\mathbf{Y})'$$

$$\text{Var}(\mathbf{AY}) = \mathbf{A}\text{Var}(\mathbf{Y})\mathbf{A}'$$

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

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

$$\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{HY}$$

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

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

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

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

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

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

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

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

Total pages 20
Total marks 95