

UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER 2003 EXAMINATIONS

STA 302 H1F / 1001 H1F

Duration - 3 hours

Aids Allowed: Calculator

NAME: _____

STUDENT NUMBER: _____

- There are 22 pages including this page.
- The last page is a table of formulae that may be useful.
- Tables of the t distribution can be found on page 20 and tables of the F distribution can be found on page 21.
- Total marks: 90

1a	1bc	1de	1f	2ab	2cd	3ab

3cd	4ab	4cde	4fg	5abc	5de	6

1. Expenditures on the criminal justice system is an area of continually rising cost in the U.S. This question examines the relationship between the total number of police employed in an American state, and the total spending on the state's criminal justice system (in millions of dollars US) for the 50 American states. The base 10 logarithm of each variable was taken before fitting the model. The variables are named `logpol` and `logexp`. Some output from SAS is given below. Seven of the numbers have been replaced by upper case letters.

The REG Procedure					
Descriptive Statistics					
Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	50.00000	1.00000	50.00000	0	0
logexp	129.26739	2.58535	334.94405	0.01516	0.12313
logpol	195.86603	3.91732	778.15035	0.22205	0.47122

The REG Procedure					
Model: MODEL1					
Dependent Variable: logpol					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	(A)	1.09909	1.09909	(C)	0.0245
Error	(B)	9.78122	(D)		
Corrected Total	49	10.88031			
	Root MSE	0.45141	R-Square	(E)	
	Dependent Mean	3.91732	Adj R-Sq	0.0823	
	Coeff Var	11.52356			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.77272	1.35553	(F)	0.5713
logexp	1	1.21632	0.52373	2.32	(G)

- (a) (7 marks) Find the 7 missing values (A through G) in the SAS output.

(b) (3 marks) Construct simultaneous 90% confidence intervals for the slope and intercept of the regression line.

(c) (4 marks) Carry out an hypothesis test to determine whether or not the data give evidence that the coefficient of $\log \text{exp}$ is greater than 1.

(d) (5 marks) The District of Columbia (not one of the 50 states but a separate region of the U.S.) spends 1,217,000,000 (1,217 million) dollars on its criminal justice system. Predict how many police officers it has. Construct a 99% interval for your value. Express your answer as a count of the number of police.

(e) (2 marks) On the next page there is a scatterplot of the logged data and a plot of the residuals versus the predicted values for the fitted regression above. Using information from the plots, give two reasons why you may not trust your prediction in (b).

- (f) (8 marks) The points for the states Alaska and Texas are identified in both plots on the previous page. For each of these states, indicate what would happen to the numbers below if the state was removed from the analysis. If a number changes, indicate how. The numbers you should consider are:

Parameter Estimates, RootMSE, R-Square

2. For the multiple linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, the least squares estimates are $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and the residuals are $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$. Assume the Gauss-Markov conditions hold.

(a) (2 marks) Show that \mathbf{b} is an unbiased estimator of β .

(b) (5 marks) Show $\text{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

(c) (5 marks) Show $\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$ and $\text{Cov}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(d) (3 marks) Find $\text{Cov}(\hat{\mathbf{Y}})$ in terms of the matrix \mathbf{H} , where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}$.

3. The data for this question are the proportion of male births (variable name: pmale) in Canada and the United States for the years 1970 through 1990 (variable name: year). Regressions were carried out separately for the two countries. Output from SAS is given below. Questions begin on the next page.

Canadian data

The REG Procedure
Model: MODEL1
Dependent Variable: pmale

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	0.00000952	0.00000952	16.13	0.0007	
Error	19	0.00001121	5.898319E-7			
Corrected Total	20	0.00002072				
Root MSE		0.00076801	R-Square	0.4592		
Dependent Mean		0.51367	Adj R-Sq	0.4307		
Coeff Var		0.14951				

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.73379	0.05480	13.39	<.0001
year	1	-0.00011117	0.00002768	-4.02	0.0007

USA data

The REG Procedure
Model: MODEL1
Dependent Variable: pmale

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	0.00000227	0.00000227	33.40	<.0001	
Error	19	0.00000129	6.793985E-8			
Corrected Total	20	0.00000356				
Root MSE		0.00026065	R-Square	0.6374		
Dependent Mean		0.51260	Adj R-Sq	0.6183		
Coeff Var		0.05085				

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.62009	0.01860	33.34	<.0001
year	1	-0.00005429	0.00000939	-5.78	<.0001

(a) (4 marks) Are the proportions of male births on the decline in Canada and the U.S.? What can you conclude from these regressions to answer this question?

(b) (3 marks) Explain why the United States has the larger t statistic for the test of $H_0 : \beta_1 = 0$ even though its slope is closer to zero. Give both a statistical explanation and suggest a practical reason why this happened.

(c) (4 marks) Give an equation for a single linear model from whose fit both of the regression equations from the output above can be obtained. Be sure to define all of your variables and explain how to test whether the change in the male birth rate differs between the two countries.

(d) (4 marks) On the next page there are residual plots for the regression of `pma1e` on `year` for Canada. What additional information about the data is provided by these plots? How does this affect your answer to part (a) for Canada?

4. An experiment was carried out with the goal of constructing a model for total oxygen demand in dairy wastes as a function of five laboratory measurements. Data were collected on samples kept in suspension in water in a laboratory for 220 days. Although all observations were taken on the same sample over time, assume that they are independent. The measured variables are:

Y log of oxygen demand (demand measured in mg oxygen per minute)

X_1 biological oxygen demand, mg/liter

X_2 total Kjeldahl nitrogen, mg/liter

X_3 total solids, mg/liter

X_4 total volatile solids, mg/liter

X_5 chemical oxygen demand, mg/liter

A regression of Y on all the X_j 's was run and part of the SAS output is shown below. A second regression of Y on X_3 and X_5 was also run, and part of the output from that regression is given on the next page.

The REG Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4.10838	0.82168	11.99	0.0001
Error	14	0.95953	0.06854		
Corrected Total	19	5.06792			

Root MSE	0.26180	R-Square	0.8107
Dependent Mean	0.11920	Adj R-Sq	0.7430
Coeff Var	219.62908		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	-2.15615	0.91349	-2.36	0.0333	0.28417
x1	1	-0.00000901	0.00051835	-0.02	0.9864	3.04562
x2	1	0.00132	0.00126	1.04	0.3153	0.44135
x3	1	0.00012780	0.00007690	1.66	0.1188	0.33049
x4	1	0.00790	0.01400	0.56	0.5815	0.03806
x5	1	0.00014165	0.00007375	1.92	0.0754	0.25286

The REG Procedure
 Model: MODEL1
 Dependent Variable: y

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.98933	1.99467	31.44	<.0001
Error	17	1.07859	0.06345		
Corrected Total	19	5.06792			
		Root MSE	0.25189	R-Square	0.7872
		Dependent Mean	0.11920	Adj R-Sq	0.7621
		Coeff Var	211.31322		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1.37148	0.19627	-6.99	<.0001
x3	1	0.00014918	0.00005473	2.73	0.0144
x5	1	0.00014188	0.00005302	2.68	0.0160

(a) (1 mark) Test the hypothesis that the coefficient of x3 is zero using the output from the first regression.

(b) (1 mark) Test the hypothesis that the coefficient of x3 is zero using the output from the second regression.

(c) (2 marks) Explain why there is a difference in your answers to parts (a) and (b).

(d) (3 marks) State the null and alternative hypotheses for the Analysis of Variance F test for the first regression. What conclusion do you draw from its p -value? (For the conclusion, do not say whether or not you reject the null hypothesis but rather say what the test tells you about the linear model.)

(e) (3 marks) Use the output from the first regression to test the joint hypothesis $\beta_4 = 0, \beta_5 = 0$.

(f) (4 marks) Which model do you prefer? Justify your choice.

(g) (2 marks) What residual plots would you like to see to check whether it is reasonable to “treat the observations as independent”?

5. For each of the following questions, give brief answers (one or two sentences). Answers without explanation will not receive any marks.

(a) (2 marks) In a simple regression of weight on height for a sample of adult males, the estimated intercept is 5 kg. Interpret this value for someone who has not taken any statistics courses.

(b) (2 marks) In simple linear regression, why can an R^2 value close to 1 not be used as evidence that the model is appropriate?

(c) (2 marks) Suppose that the variance of the estimated slope in the simple regression of Y on X_1 is 10. Suppose that X_2 is added to the model, and that X_2 is uncorrelated with X_1 . Will the variance of the coefficient of X_1 still be 10?

(d) (2 marks) A regression analysis was carried out with response variable sales of a product (Y) and two predictor variables: the amount spent on print advertisements (X_1) and the amount spent on television advertisements (X_2) for the product. The fitted equation was $\hat{Y} = -2.35 + 2.36X_1 + 4.18X_2 - .35X_1X_2$. The test for whether the coefficient of the interaction term is zero had p -value less than 0.0001. Explain what this test means in practical terms for the company executive who has never studied statistics.

(e) (2 marks) Explain why we might prefer to use adjusted R^2 rather than R^2 when comparing two models.

6. (5 marks) A large real estate firm in Toronto has been keeping records on selling prices for single family dwellings. They have also recorded numerous other features of the houses that sold, including square footage, number of rooms, property taxes, type of heating, lot size, area of city, existence of finished basement, etc. An agent for this firm hopes to use these data to show that the rate of change in house prices over the past seven years differs depending on area of the city. Describe how you would help the agent.

Simple regression formulae

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

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

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

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

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

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

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

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

$$\sigma^2\{\hat{Y}^h\} = \text{Var}(\hat{Y}^h)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(X^h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right)$$

$$\sigma^2\{\text{pred}\} = \text{Var}(Y^h - \hat{Y}^h)$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(X^h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right)$$

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

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

Working-Hotelling coefficient:

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

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

$$= \text{E}(\mathbf{X}\mathbf{X}') + (\text{E}\mathbf{X})(\text{E}\mathbf{X})'$$

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

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

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

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

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

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

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

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

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

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

$$C_p = \frac{\text{SSE}_p}{\text{MSE}_p} - (n-2p)$$

$$\text{PRESS}_p = \sum(Y_i - \hat{Y}_{i(i)})^2$$