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: SOLUTIONS **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.)

$$N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

   (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 $\text{Var}(\hat{Y}_h)$.

$$\begin{aligned} Var(b_0 + b_1 X_h) &= Var(b_0) + X_h^2 Var(b_1) + 2X_h Cov(b_0, b_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right) + X_h^2 \frac{\sigma^2}{S_{XX}} - 2\sigma^2 \frac{X_h \overline{X}}{S_{XX}} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{S_{XX}}\right) \end{aligned}$$

   (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 $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

$$\begin{aligned} At\ \overline{X},\ \hat{Y} &= b_0 + b_1 \overline{X} \\ &= (\overline{Y} - b_1 \overline{X}) + b_1 \overline{X} \\ &= \overline{Y} \end{aligned}$$

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

$$\begin{aligned} \mathit{Var}(\mathbf{b}) = \mathit{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathit{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

(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 $\mathbf{X}_h$.

$$\begin{aligned} \mathit{Var}\left(\hat{E(\mathbf{Y})}\right) &= \mathit{Var}(\mathbf{X}_h'\mathbf{b}) \\ &= \mathbf{X}_h'\,\mathit{Var}(\mathbf{b})\mathbf{X}_h \\ &= \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h \end{aligned}$$

(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?)

$$\frac{1}{n}\mathbf{X}'\mathbf{1} = \frac{1}{n}\begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_{11} & X_{21} & \ldots & X_{n1} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1k} & X_{2k} & \ldots & X_{nk} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{n}\begin{pmatrix} n \\ \sum X_{i1} \\ \vdots \\ \sum X_{ik} \end{pmatrix} = \begin{pmatrix} 1 \\ \overline{X}_1 \\ \vdots \\ \overline{X}_k \end{pmatrix}$$

At $\overline{X}_1, \ldots, \overline{X}_k$,

$$\begin{aligned} \hat{Y} &= \begin{pmatrix} 1 & \overline{X}_1 & \cdots & \overline{X}_k \end{pmatrix}\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} \\ &= \left(\frac{1}{n}\mathbf{X}'\mathbf{1}\right)'\mathbf{b} = \frac{1}{n}\mathbf{1}'\mathbf{X}\mathbf{b} = \frac{1}{n}\mathbf{1}'\hat{\mathbf{Y}} = \frac{1}{n}\sum \hat{Y}_i = \overline{Y} \end{aligned}$$

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?

   Plot would show decreasing variance since $\beta_1 > 0$. (Sketch of plot of residuals versus X is adequate.)

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

   Need to find the function $f(\mu)$ such that

$$f(\mu) \propto \int \mu \, d\mu \propto \mu^2$$

   so use Y^2.

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
                                          Uncorrected                   Standard
Variable           Sum          Mean           SS      Variance      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
                                     Sum of           Mean
Source                    DF        Squares         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
                               Parameter       Standard
     Variable      DF          Estimate          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) = 1

   (B) = 186

   (C) = $24210/6.21796 = 3893.57$

   (D) = $24210/25367 = 0.9544$

   (E) = $-0.23786/1.06445 = -0.22$

(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.

$r = \sqrt{0.9544} = 0.977$
The slope is the correlation times the ratio of the s.d. of Y *to the s.d. of* X*, i.e.,* $b_1 = r\frac{s_Y}{s_X}$.

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

Test statistic: $t_{obs} = \frac{1.00607-1}{0.01612} = 0.377$
Under the null hypothesis, this is an observation from a t-distribution with 186 degrees of freedom. (Approximate this from tables with a t_{120} *distribution.)*
$2(0.3) < p < 2(0.4)$
The data give no evidence that the slope differs from 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.

No. Prediction intervals are wider the further X *is from* $\overline{X}$.

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?

Expect 5% (of 188 which is 9) of the points to be outside the limits of the intervals. There are 13 points outside the prediction intervals, so perhaps there are some outliers or another problem such as non-normality. (But 13 versus 9 is not a large discrepancy from what is expected.)

(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?

Looking for: no curvature, constant variance, no obvious outliers and/or influential points.
The plot looks good.

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

The residuals are not normally distributed.
The distribution of the residuals has heavy tails.

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

The answer to part (c) is OK. The t-test for β_1 is robust against departures from normality.
Re part (d): Coverage of the prediction intervals is not 95% as prediction intervals are not robust against non-normality.

(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?

What is regression to the mean: In a test - re-test situation, we expect the slope to be less than 1 because extreme values regress to the mean.
It is not seen here since the slope is approximately 1.
Over time, life-expectancy is improving. So this is not a test - re-test situation.

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?

$$\texttt{highinc} = \begin{cases} 1 & \textit{if country has high income economy} \\ 0 & \textit{otherwise} \end{cases}$$

$$\texttt{medinc} = \begin{cases} 1 & \textit{if country has medium income economy} \\ 0 & \textit{otherwise} \end{cases}$$

There is no indicator variable for low income because having all 3 results in a non-invertible $\mathbf{X}'\mathbf{X}$ *matrix. OR It is redundant to have 3 indicator variables as low income countries have 0 for both other indicator variables.*

(Question 5 continued.)

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

Half of the plots of an independent variable versus another independent variable are redundant.

The independent variable versus dependent variable plots are useless.

In the dependent variable versus independent variable plots we see:

- *Countries with high income economies have greater life expectancies.*
- *This might not be true for medium versus other economies.*
- *The relationship between* `loggnp` *and* `popgrowth` *may be non-linear and there are a few strange points.*

In the plots of an independent variable versus another independent variable we see that there is a (weak) relationship between `popgrowth` *and* `loggnp` *so it may be difficult to interpret each of their individual effects on* `lifeexp`*.*

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

Gross national product has a right-skewed distribution.
It allows you to better see its relationship with life expectancy.

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

$$\widehat{\texttt{lifeexp}} = 33.34 - 2.17\,\texttt{popgrowth} + 4.60\,\texttt{logGNP} + 0.16\,\texttt{highinc} + 3.63\,\texttt{medinc}$$

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

6.54 (square root of MSE)

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

i. `popgrowth`

For a country with the same income and GNP, as population growth increases 1%, life expectancy decreases by 2.17 years on average.

ii. `logGNP`

For a country with the same income and population growth, a k-fold change in GNP is associated with a change in mean life expectancy of $4.60\log(k)$.

iii. `highinc`

For countries with the same GNP and population growth, a high income country has a life expectancy of 0.16 years more, on average, than a low income country.

(Question 5 continued.)

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

We are testing the null hypothesis that the coefficient of `popgrowth` *is 0, given that the other predictor variables are in the model, versus the alternative that it is different from 0.*
We conclude that there is strong evidence that it is not 0.

(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).

No. t-tests assume that all other variables are in the model, so we should never remove more than one at a time.

Test for whether we can remove both variables from model (testing that the coefficients for both are simultaneously 0):
$H_0 : \beta_3 = \beta_4 = 0$ *versus* H_a*: at least one of* β_3*,* β_4 *is not zero*
Test statistic: $F_{obs} = \frac{(296+112)/2}{42.75} = 4.8$
Under the null hypothesis, this is an observation from an $F_{2,131}$ *distribution. Approximating the p-value from an* $F_{2,120}$ *distribution gives* $p \doteq 0.01$.
So we have strong evidence that at least one of the coefficients is not zero.

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 $\mathrm{Var}(b_1)$.

σ^2 is an unobserved model parameter while s^2 it its estimate from the data. σ^2 is the error variance while $Var(b_1)$ is a function of σ^2 since b_1 is a function of random variables with variance σ^2.

(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.

SSE is minimized over a larger dimensional space so it is smaller. SSTO is constant. So $R^2 = 1 - \frac{SSE}{SSTO}$ is larger.

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.

$R^2 = \frac{SSR}{SSTO}$

SSTO is constant. A change in SSR changes the numerator of the partial F-test test statistic.

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

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

SSTO is constant so adjusted R^2 increases if MSE decreases.

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

What it is: correlation among the predictor variables in multiple regression.

Why it is important:
- results in numerically unstable estimates
- makes it impossible to sort out which predictor variables are important
- large standard errors (inefficient estimates)
- coefficients may have wrong sign

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

Any 2 of:
- Different procedures give different final models.
- Small changes in the data may result in completely different models.
- Variables that seem important on one set of data may not be significant on another.
- They are not robust against multicollinearity.
- The techniques typically try to obtain a single model, but many may be valid.

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.

Dependent variable: life expectancy
Independent variables:
- the economic variables
- literacy rate
- an indicator variable that is 1 if the country is developed, 0 if developing
- the interaction between literacy rate and the indicator variable

Plots:
- life expectancy versus literacy rate to examine the nature of the relationship and whether a linear model is appropriate
- literacy rate versus each of the economic variables because we want to see see the effect of literacy rate over and above the effects of the economic variables so are concerned about possible multicollinearity (could also examine the pairwise correlations)

Residual plots to examine model assumptions:
- versus predicted values (for constant variance and possible outliers and influential points)
- versus the independent variables (for adequacy of linear model)
- normal quantile plot (for normality)

Carry out a partial F-test to evaluate whether or not the relationship is the same for developed and developing countries.

To evaluate the effect of literacy rate over and above economic factors, look at its partial correlation or consider the hypothesis test for whether or not its coefficient is 0, or could carry out a partial F-test for the significance of both the coefficient of literacy rate and the interaction term.