UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER EXAMINATIONS 2010
STA 302 H1F / STA 1001 HF

Duration - 3 hours

Aids Allowed: Calculator

LAST NAME: SOLUTIONS **FIRST NAME:**________________

STUDENT NUMBER: ______________________________

- There are 19 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 unless the question states otherwise.
- Pages 14 through 18 contain output from SAS that you will need to answer Question 5.
- Total marks: 85

1	2ab	2cd	3	4	5a

5b	5c	5d(i-iii)	5d(iv-vi)	6	7, 8

1. (10 marks) Beside each description, write the letter of the term from the list below that provides the best match.

(I) *I* What to include when the effect of X_1 on Y is different for different values of X_2.

(II) *C* The proportion of variation explained by the regression line.

(III) *F* An observed response minus its estimated mean according to some model.

(IV) *Q* A measure of how influential a particular observation is.

(V) *N* A method for estimating regression coefficients.

(VI) *B* A test comparing a model of interest to the model with only an intercept.

(VII) *S* A statistic used to identify problems of multicollinearity.

(VIII) *any of D,Z,T* A statistic for comparing models with different sets of explanatory variables.

(IX) *T* Another name for the estimate of σ^2 in regression analysis.

(X) *R* A measure of how unusual the x-values are for a particular observation.

(A) Analysis of Variance

(B) Analysis of Variance F-test

(C) R-squared

(D) Adjusted R-squared

(E) t-test

(F) Residual

(G) Standardized residual

(H) Fitted value

(I) Interaction

(J) Indicator

(K) Explanatory variable

(L) Response variable

(M) Outlier

(N) Least Squares

(O) Correlation

(P) Degrees of freedom

(Q) Cook's Distance

(R) Leverage

(S) Variance Inflation Factor

(T) Residual mean square

(U) Mean square of regression

(V) Residual sum of squares

(W) Regression sum of squares

(X) Total sum of squares

(Y) Variance

(Z) Extra sum of squares

2. Suppose that we believe that a response variable Y is related to a non-random explanatory variable x by the model $Y_i = \beta x_i + e_i$, $i = 1, \ldots, n$. That is, we believe that it is appropriate to use a model that goes through the origin. Assume that the following conditions hold:
   - The errors $e_1, \ldots, e_n$ have expectation 0.
   - The errors have common variance σ^2.
   - The errors are uncorrelated.

   (a) (3 marks) Show that the least squares estimator of β is

$$\hat{\beta} = \sum_{i=1}^{n} x_i Y_i \Bigg/ \sum_{i=1}^{n} x_i^2$$

The least square estimates minimize the residual sum of squares:

$$RSS = \sum (Y_i - \beta x_i)^2$$

$$\frac{\partial\, RSS}{\partial\, \beta} = -2 \sum x_i (Y_i - \beta x_i)$$

Setting equal to 0 and solving gives

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$$

   (b) (3 marks) Assuming that the model is correct, show that $\hat{\beta}$ is an unbiased estimator of β.

$$\begin{aligned} E(\hat{\beta}) &= \frac{x_i E(Y_i)}{\sum x_i^2} \\ &= \frac{\sum x_i (\beta x_i)}{\sum x_i^2} \\ &= \beta \end{aligned}$$

(Question 2 continued.)

(c) (2 marks) Find Var($\hat{\beta}$).

$$\begin{aligned} Var(\hat{\beta}) &= \frac{1}{(\sum x_i^2)^2} \sum x_i^2 \, Var(Y_i) \\ & \quad \textit{(since the } Y_i\textit{'s are assumed uncorrelated because the } e_i\textit{'s are assumed uncorrelated)} \\ &= \frac{1}{\sum x_i^2} \sigma^2 \end{aligned}$$

(d) (2 marks) Suppose that the model $Y_i = \beta x_i + e_i$ is correct, but the model $Y_i = \beta_0 + \beta_1 x_i + e_i$ is used. Show that Var($\hat{\beta_1}$) $\geq$ Var($\hat{\beta}$).

$$\begin{aligned} Var(\hat{\beta_1}) &= \frac{\sigma^2}{\sum x_i^2 - n\overline{x}^2} \geq \frac{\sigma^2}{\sum x_i^2} \\ & \quad \textit{since } \sum x_i^2 - n\overline{x}^2 \leq \sum x_i^2 \end{aligned}$$

3. A multiple linear regression model with dependent variable Y and 3 explanatory variables was fit to 15 observations. The residual sum of squares was found to be 22.0 and it was also found that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0.6 \\ 0.3 & 6.0 & 0.5 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.7 \\ 0.6 & 0.4 & 0.7 & 3.0 \end{bmatrix}$$

(a) (1 mark) What degrees of freedom would be used when finding a confidence interval for β_1?

$$n - (p + 1) = 15 - 4 = 11$$

(b) (1 mark) What is the estimate of the error variance?

$$\frac{22}{11} = 2$$

(c) (1 mark) What is the estimated variance of the estimator of β_2?

$$2(0.2) = 0.4$$

4. Consider the multiple regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

(a) (3 marks) Show that $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{e}$.

$$\begin{aligned}
\hat{\mathbf{e}} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \\
&= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{He} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{e}
\end{aligned}$$

(b) (1 mark) Why is $\mathrm{E}(\mathbf{ee}') = \mathrm{Var}(\mathbf{e})$?

Var$(\mathbf{e}) = E(\mathbf{ee}') - E(\mathbf{e})\,(E(\mathbf{e}))'$ *(from the formula sheet)*
and $E(\mathbf{e}) = \mathbf{0}$

(c) (4 marks) Show that $\mathbf{I} - \mathbf{H}$ is idempotent and symmetric.

Idempotent:

$$\begin{aligned}
(\mathbf{I} - \mathbf{H})^2 &= \mathbf{I} - 2\mathbf{H} + \mathbf{H}^2 \\
&= \mathbf{I} - 2\mathbf{H} + \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\
&= \mathbf{I} - 2\mathbf{H} + \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\
&= \mathbf{I} - 2\mathbf{H} + \mathbf{H} = \mathbf{I} - 2\mathbf{H}
\end{aligned}$$

Symmetric:

$$\begin{aligned}
(\mathbf{I} - \mathbf{H})' &= \mathbf{I}' - \mathbf{H}' \\
&= \mathbf{I} - \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{I} - \mathbf{H}
\end{aligned}$$

(d) (3 marks) Show that $\mathrm{Var}(\hat{\mathbf{e}}|\mathbf{X}) = \sigma^2(\mathbf{I} - \mathbf{H})$.

$$\begin{aligned}
Var(\hat{\mathbf{e}}|\mathbf{X}) &= Var\left((\mathbf{I} - \mathbf{H})\mathbf{e}|\mathbf{X}\right) \quad \textit{from (a)} \\
&= (\mathbf{I} - \mathbf{H})\,Var(\mathbf{e})(\mathbf{I} - \mathbf{H})' \\
&= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' \\
&= \sigma^2(\mathbf{I} - \mathbf{H}) \quad \textit{using (c)}
\end{aligned}$$

5. The data considered in this question are the same data considered in Assignment 1, taken from a 2007 *Wall Street Journal* article on the decline of U.S. house prices. The data are indicators of the real-estate market in 28 U.S. cities. The variables considered in this question are:
Response variable:
• `PriceChange` – The percent change in average price of a home from one year ago.
Explanatory variables:
• `LoansOverdue` – The percentage of mortgage loans that are 30 days or more overdue.
• `InventoryChange` – The percent change in housing inventory from one year ago. A positive value indicates that more houses are on the market.
• `EmployOutlook` – A character variable that classifies the projected growth in the number of jobs as one of Strong, Average, or Weak. (An observation that had an employment outlook of Very Weak in the original data has been re-classified as Weak.)
• `iEmployOutIsWeak` – An indicator variable that is 1 if `EmployOutlook` is Weak and 0 otherwise.
• `iEmployOutIsAverage` – An indicator variable that is 1 if `EmployOutlook` is Average and 0 otherwise.
• `iEmpWeak_LoansOD` – The product of `iEmployOutIsWeak` and `LoansOverdue`.
• `iEmpAvg_LoansOD` – The product of `iEmployOutIsAverage` and `LoansOverdue`.

On pages 14 through 18 there is SAS output for the analysis of these data. The questions below relate to the SAS output.

(a) ANALYSIS 1 (page 14) was carried out using only observations having `EmployOutlook` either Strong or Weak. (That is, cities with Average employment outlook were removed from the data for this analysis only.) The questions in part (a) relate to ANALYSIS 1.

i. (2 marks) What is the estimated difference in the mean of percent change in average price of a home between cities with Strong and cities with Weak employment outlook?

5.645% with cities with weak outlook having the smaller (negative) mean percent change.

ii. (2 marks) Can you conclude that there is a difference in the mean of percent change in average price of a home between cities with Strong and cities with Weak employment outlook? Justify your answer.

There is only weak evidence that the mean of percent change in price differs between cities with strong and weak outlook (p=0.0867).

(b) ANALYSIS 2 (page 15) was carried out using all of the available data. It is a simple linear regression using `LoansOverdue` as the explanatory variable. The questions in part (b) relate to ANALYSIS 2.

i. (4 marks) Four numbers in the SAS output have been replaced by letters. What are they?

(A) = 94.14588

(B) = 0.0437

(C) = 337.0752

(D) = −2.179

ii. (2 marks) R^2 is only 22%. As a consequence, can we conclude that there is not a linear relationship between `PriceChange` and `LoansOverdue`? Explain.

No. There is moderate evidence of a linear relationship ($p = 0.0437$ for the slope).

iii. (5 marks) On page 15 you are given a plot of the standardized residuals versus the predicted values and a normal quantile plot of the standardized residuals for this analysis. What are you looking for in each plot and what do you conclude?

From plot of residuals versus predicted values:
Looking for non-linearity, outliers, influential points, constant variance. There are no concerns with these so the linear model is appropriate and it is reasonable to assume that the variance of the error terms is constant. (It is possible that you might have some concern about one large positive residual which means that point is a potential outlier.)

In normal quantile plot:
Looking for a straight line. The curvature at the ends indicates that the distribution of the residuals has lighter tails than a normal distribution.

(c) ANALYSIS 3 (page 16) was carried out on all of the available data. It is a multiple regression using `LoansOverdue` and `InventoryChange` as explanatory variables. The questions in part (c) relate to ANALYSIS 3.

i. (1 mark) Write down the model that is being fit. Do not use matrix form.

$$\texttt{PriceChange} = \beta_0 + \beta_1\texttt{LoansOverdue} + \beta_2\texttt{InventoryChange} + e$$

ii. (3 marks) What do you conclude from the t-tests for the coefficients for `LoansOverdue` and `InventoryChange`?

There is no evidence that inventory change affects price change over and above loans overdue ($p = 0.2464$).
There is some evidence that the number of loans overdue affects price change over and above inventory change ($p = 0.0422$).

iii. (2 marks) On page 16 there are two added variable plots; the first is for `LoansOverdue` and the second is for `InventoryChange`. For the first of these plots, explain what is being plotted.

The residuals from the regression of price change on inventory change versus the residuals from the regression of loans overdue on inventory change.

iv. (2 marks) Explain how the added variable plots are related to your conclusions to the t-tests considered in part ii. (of part (c)).

The first plot shows a (not very strong) linear relationship showing that there is a relationship between price change and loans overdue over and above inventory change.
The second plot shows no relationship between price change and inventory change over and above loans overdue.

(d) ANALYSIS 4 (page 17) was carried out on all of the available data. It is a multiple regression using `LoansOverdue`, `iEmployOutIsWeak`, and `iEmployOutisAverage` as explanatory variables. ANALYSIS 5 (page 18) uses the same data and explanatory variables as ANALYSIS 4, but includes the additional explanatory variables `iEmpWeak_LoansOD` and `iEmpAvg_LoansOD`. The questions in part (d) relate to ANALYSES 4 and 5.

i. (2 marks) Explain the purpose of including the explanatory variables that are in the model in ANALYSIS 5 but are not in the model in ANALYSIS 4.

We can consider if the slope of the relationship between price change and loans overdue differs among the 3 categories of employment outlook. (Allows non-parallel lines.)

ii. (4 marks) Carry out one statistical test to determine whether both of the extra terms in the model of ANALYSIS 5 (that are not in the model of ANALYSIS 4) should be excluded from the model. (You have not been given any tables for probability distributions. However, you should be able to make a conclusion without tables based on what you know about the relevant probability distribution.)

Test $H_0: \ \beta_4 = \beta_5 = 0$ *versus* $H_a: $ *at least one* $\neq 0$ *where* β_4 *and* β_5 *are the coefficients of* `iEmpWeak_LoansOD` *and* `iEmpAvg_LoansOD` *respectively.*
Test statistic: $\frac{(237.80794-227.23465)/2}{17.47959} = 0.302$
Under H_0*, this is an observation from an* $F(2, 13)$ *distribution with p-value being the area to the right of 0.302 which is a large area.*
So the data are consistent with zero coefficient for both terms and we can exclude them from the model.

iii. (2 marks) R^2 is higher in ANALYSIS 5 than ANALYSIS 4, while adjusted R^2 is higher in ANALYSIS 4 than ANALYSIS 5. Explain, in practical terms, why this happened.

R^2 *always increases with more terms in the model but Adjusted* R^2 *only increases if the terms are useful predictors and result in a reduction in MSE.*

Continued

(Question 5 part (d) continued.)

iv. (2 marks) For ANALYSIS 4, what do you conclude from the analysis of variance F-test? Is your conclusion consistent with the t-tests for the coefficients of the explanatory variables? Why or why not?

There is some evidence that not all of the coefficients of the explanatory variables ore 0 ($p = 0.0267$).
It is consistent with the t-tests, all of which give some evidence that the relevant coefficient is not 0 given that the other variables are in the model.

v. (3 marks) For ANALYSIS 5, what do you conclude from the t-test for the coefficient of `LoansOverdue`? Does this conclusion contradict the t-test for the coefficient of `LoansOverdue` in ANALYSIS 4? Why or why not?

Give the other variables are in the model, the data are consistent with a coefficient of `LoansOverdue` *of 0 ($p = 0.1520$).*
It does not contradict the t-test for analysis 4 since we have added terms related to `LoansOverdue` *(interaction terms invovling it) and the t-test is for a model that includes all other explanatory variables.*

vi. (2 marks) Explain how the Variance Inflation Factors given in ANALYSIS 5 support your answer to part v. (of part (d)).

The high values of VIF indicate that there is multicollinearity. So the explanatory variables are related to each other and thus there is no evidence that the coefficients of the explanatory variables are different from 0 given that the other variables are in the model.

Continued

6. (a) (4 marks) For each scenario, sketch a scatterplot that shows the given situation.

i. A simple linear regression that includes a point with high leverage and low influence.

(The answer is a sketch of a scatterplot of Y versus x with a linear pattern and one point that is far from the other points in the x direction but that still follows the linear pattern of the other points.)

ii. A simple linear regression that includes a point with high leverage and high influence.

(The answer is a sketch of a scatterplot of Y versus x with a linear pattern in most points except for one point that is far from the other points in the x direction and that does not follow the linear pattern of the other points.)

(b) (2 marks) A regression is carried out and a point is identified as having high leverage and low influence. Why should you be concerned about the presence of that point?

Such points increase R^2 and decrease the standard errors. As a result we get an overly optimistic view of the fit of the regression line and the precision of the estimates.

Continued

7. (4 marks) Suppose that a simple linear regression model has been fit to n observations. Suppose that the distribution of the explanatory variable appears to be normal, while the distribution of the response variable is highly right-skewed. A plot of the residuals versus the explanatory variable produces a pattern with a parabolic shape with increasing variance. It is suggested that we should consider carrying out a regression using a quadratic model, that is, a model of the form $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$. Is this suggestion appropriate? Why or why not?

Adding the quadratic term may address the parabolic shape, but will not address the increasing variance. We should try a transformation of Y to stabilize the variance. Note that transforming Y will also change the shape of its relationship with x so the polynomial model in x may not be appropriate any longer.

8. (3 marks) A multiple linear regression model was fit in order to examine the effects of gestational period (X_1, measured in days) and litter size (X_2) on brain weight (Y, measured in g) after controlling for body size (X_3, measured in kg). The fitted regression was

$$\widehat{\log(Y)} = 0.85 + 0.42\log(X_1) - 0.31\log(X_2) + 0.58\log(X_3).$$

Explain carefully how to interpret the coefficient estimated by 0.42 in practical terms.

Changing gestational period (X_1) by a factor of k results in changing brain weight (Y) by a factor of $k^{0.42}$ assuming litter size (X_2) and body size (X_3) are held constant.