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:______________________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) _____ What to include when the effect of X_1 on Y is different for different values of X_2.

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

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

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

(V) ______ A method for estimating regression coefficients.

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

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

(VIII) ______ A statistic for comparing models with different sets of explanatory variables.

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

(X) ______ A measure of now 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

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

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

(Question 2 continued.)

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

(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 $\text{Var}(\hat{\beta}_1) > \text{Var}(\hat{\beta})$.

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?

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

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

Continued

4. Consider the multiple regression model

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

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

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

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

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

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?

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.

(Question 5 continued.)

(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) = ____________

(B) = ____________

(C) = ____________

(D) = ____________

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.

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?

(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.

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

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.

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)).

Continued

(Question 5 continued.)

(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.

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.)

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.

(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?

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?

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

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.

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

   (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?

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?

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.

The SAS output on pages 14 to 18 is relevant to question 5.

ANALYSIS 1:

Analysis on Data using only Cities with Employment Outlook that is either Strong or Weak

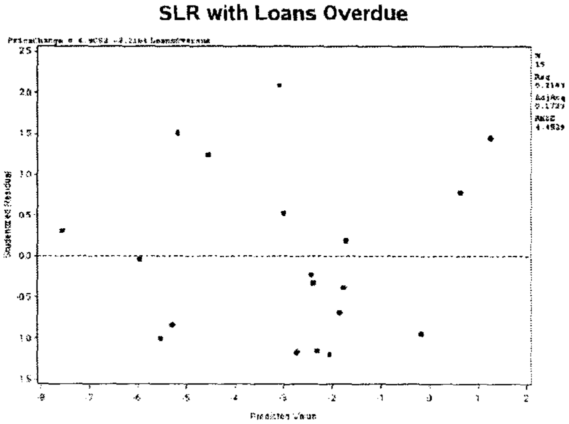
```
                          The REG Procedure
                    Dependent Variable: PriceChange

            Number of Observations Read                         15
            Number of Observations Used                          9
            Number of Observations with Missing Values           6

                        Analysis of Variance
                                 Sum of           Mean
Source                   DF     Squares         Square    F Value    Pr > F
Model                     1    70.81339       70.81339       3.97    0.0867
Error                     7   125.01550       17.85936
Corrected Total           8   195.82889

             Root MSE              4.22603    R-Square     0.3616
             Dependent Mean       -2.01111    Adj R-Sq     0.2704
             Coeff Var          -210.13425

                         Parameter Estimates
                                Parameter       Standard
  Variable             DF        Estimate          Error    t Value    Pr > |t|
  Intercept             1         1.12500        2.11302       0.53      0.6109
  iEmployOutIsWeak      1        -5.64500        2.83491      -1.99      0.0867
```

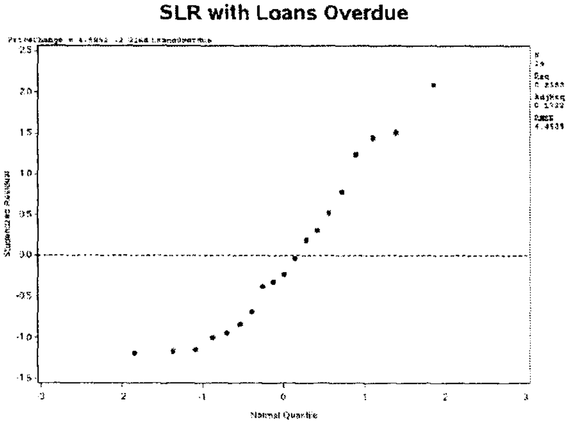
(SAS output for Question 5 continued.)

ANALYSIS 2:

Simple Linear Regression with LoansOverdue as the Predictor Variable

```
                              The REG Procedure
                        Dependent Variable: PriceChange

          Number of Observations Read                             28
          Number of Observations Used                             19
          Number of Observations with Missing Values               9

                             Analysis of Variance
                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F
Model                      1      94.14588            (A)       4.75       (B)
Error                     17           (C)       19.82795
Corrected Total           18     431.22105

           Root MSE              4.45286    R-Square     0.2183
           Dependent Mean       -2.93158    Adj R-Sq     0.1723
           Coeff Var          -151.89285

                              Parameter Estimates
                             Parameter       Standard
    Variable         DF       Estimate          Error    t Value    Pr > |t|
    Intercept         1        4.90522        3.73874       1.31      0.2070
    LoansOverdue      1       -2.21642        1.01716        (D)      0.0437
```





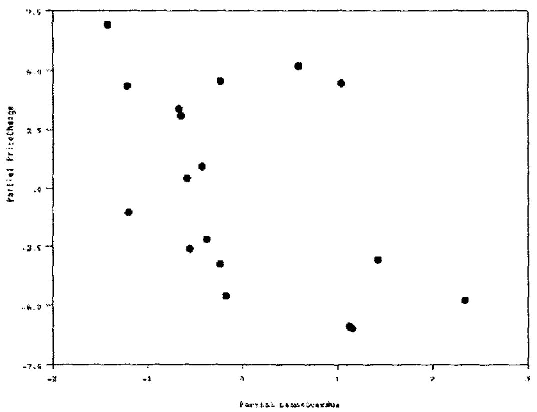
Continued

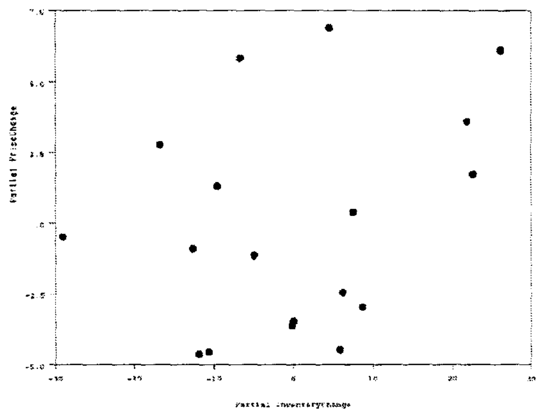
(SAS output for Question 5 continued.)

ANALYSIS 3:

Multiple Regression with LoansOverdue and InventoryChange as Explanatory Variables

```
                              The REG Procedure
                        Dependent Variable: PriceChange

          Number of Observations Read                                28
          Number of Observations Used                                18
          Number of Observations with Missing Values                 10

                             Analysis of Variance
                                      Sum of           Mean
Source                   DF          Squares         Square    F Value    Pr > F
Model                     2        118.98895       59.49448       3.91    0.0429
Error                    15        228.02716       15.20181
Corrected Total          17        347.01611

           Root MSE              3.89895    R-Square     0.3429
           Dependent Mean       -3.42778    Adj R-Sq     0.2553
           Coeff Var          -113.74570

                              Parameter Estimates
                            Parameter       Standard
 Variable          DF        Estimate          Error    t Value    Pr > |t|
 Intercept          1         2.65873        3.62085       0.73      0.4741
 LoansOverdue       1        -2.02190        0.91033      -2.22      0.0422
 InventoryChange    1         0.07833        0.06494       1.21      0.2464
```





Continued

(SAS output for Question 5 continued.)

ANALYSIS 4:

Multiple Regression with `LoansOverdue`, `iEmployOutIsWeak`, and `iEmployOutIsAverage`

```
                               The REG Procedure
                        Dependent Variable: PriceChange

           Number of Observations Read                         28
           Number of Observations Used                         19
           Number of Observations with Missing Values           9

                             Analysis of Variance
                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     3      193.41312       64.47104       4.07    0.0267
Error                    15      237.80794       15.85386
Corrected Total          18      431.22105

            Root MSE              3.98169    R-Square     0.4485
            Dependent Mean       -2.93158    Adj R-Sq     0.3382
            Coeff Var          -135.82070

                             Parameter Estimates
                                  Parameter       Standard
Variable                 DF        Estimate          Error    t Value    Pr > |t|
Intercept                 1         9.87078        3.89616       2.53      0.0229
iEmployOutIsWeak          1        -6.12724        2.67738      -2.29      0.0370
iEmployOutIsAverage       1        -5.26246        2.36003      -2.23      0.0415
LoansOverdue              1        -2.38142        0.91194      -2.61      0.0196
```

Continued

(SAS output for Question 5 continued.)

ANALYSIS 5:

Multiple Regression with LoansOverdue, iEmployOutIsWeak, iEmployOutIsAverage, iEmpWeak_LoansOD and iEmpAvg_LoansOD

```
                                  The REG Procedure
                           Dependent Variable: PriceChange

            Number of Observations Read                          28
            Number of Observations Used                          19
            Number of Observations with Missing Values            9

                              Analysis of Variance
                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F
Model                     5       203.98640       40.79728       2.33    0.1014
Error                    13       227.23465       17.47959
Corrected Total          18       431.22105

             Root MSE              4.18086    R-Square     0.4730
             Dependent Mean       -2.93158    Adj R-Sq     0.2704
             Coeff Var          -142.61461

                              Parameter Estimates
                                  Parameter       Standard                                Variance
Variable                 DF        Estimate          Error    t Value    Pr > |t|        Inflation
Intercept                 1        11.68757        7.24891       1.61      0.1309                0
iEmployOutIsWeak          1        -1.73422       12.69454      -0.14      0.8934         33.96623
iEmployOutIsAverage       1        -8.81329        8.48728      -1.04      0.3180         19.52066
LoansOverdue              1        -2.87613        1.88998      -1.52      0.1520          3.91633
iEmpWeak_LoansOD          1        -1.29487        3.50733      -0.37      0.7179         32.62860
iEmpAvg_LoansOD           1         0.98817        2.23799       0.44      0.6661         20.24662
```

Continued

Simple regression formulae

$$b_1 = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} = \frac{\sum x_iy_i-n\overline{x}\overline{y}}{\sum(x_i-\overline{x})^2}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

$$\mathrm{Var}(\hat{\beta}_1|X) = \frac{\sigma^2}{\sum(x_i-\overline{x})^2}$$

$$\mathrm{Var}(\hat{\beta}_0|X) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i-\overline{x})^2}\right)$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1|X) = -\frac{\sigma^2\overline{x}}{\sum(x_i-\overline{x})^2}$$

$$\mathrm{SST} = \sum(y_i - \overline{y})^2$$

$$\mathrm{RSS} = \sum(y_i - \hat{y}_i)^2$$

$$\mathrm{SSReg} = b_1^2\sum(x_i - \overline{x})^2 = \sum(\hat{y}_i - \overline{y})^2$$

$$\mathrm{Var}(\hat{y}|X = x^*) = \sigma^2\left(\frac{1}{n} + \frac{(x^*-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)$$

$$\mathrm{Var}(Y - \hat{y}|X = x^*) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^*-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)$$

$$r = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}}$$

$$\mathrm{SXX} = \sum(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2$$

$$h_{ij} = \frac{1}{n} + \frac{(x_i-\overline{x})(x_j-\overline{x})}{\mathrm{SXX}} \quad \left(h_{ii} > \frac{4}{n}\right)$$

$$\mathrm{DFBETAS}_{ik} = \frac{b_k-b_{k(i)}}{s.e.(b_k)} \quad \left(> 1 \text{ or } \frac{2}{\sqrt{n}}\right)$$

$$\mathrm{DFFITS}_i = \frac{\hat{y}_i-\hat{y}_{i(i)}}{s.e.(\hat{y}_i)} \quad \left(> 1 \text{ or } 2\sqrt{\frac{2}{n}}\right)$$

$$D_i = \frac{\sum(\hat{y}_{j(i)}-\hat{y}_j)^2}{2S^2} \quad \left(> \frac{4}{n-2}\right)$$

Regression in matrix terms

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

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

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

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

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

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

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

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

$$\mathrm{RSS} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

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

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

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}$$