UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER EXAMINATIONS 2011
STA 302 H1F / STA 1001 HF

Duration - 3 hours

Aids Allowed: Calculator

LAST NAME:______________________**FIRST NAME:**____________________

STUDENT NUMBER: ____

- There are 22 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 16 through 21 contain output from SAS that you will need to answer Questions 3 and 4.
- Total marks: 95

1ab	1c	2abcd	2ef	3ab	3cd	3ef(i,ii)

3f(iii,iv)	3g	4a	4bc(i,ii)	4c(iii)	5a	5b

1. The following questions are about simple linear regression. For all of them, assume that the values of the explanatory variable are under the control of the researcher (so they are not random).

   (a) (2 marks) What is wrong with each of the following equations related to the simple linear regression model?

      i. $Y_i = b_0 + b_1 x_i + e_i$

      ii. $\hat{Y}_i = b_0 + b_1 x_i + \hat{e}_i$

   (b) (8 marks) The assumptions of the simple linear regression model are:
   (1) The form of the model is appropriate.
   (2) The error terms have expectation 0.
   (3) The error terms have constant variance.
   (4) The error terms are uncorrelated.
   (5) The error terms are normally distributed.
   State which of these assumptions are required to carry out each of the following procedures:

      i. Find the least squares estimate of the slope and intercept.

      ii. Show that estimator of the slope is unbiased.

      iii. Derive the formula for the variance of the estimator of the slope.

      iv. Construct a confidence interval for the slope.

(Question 1 continued)

(c) (3 marks) The weighted least squares estimates are found by minimizing

$$\text{WRSS} = \sum_{i=1}^{n} \left[w_i (y_i - \hat{y}_i)^2 \right]$$

and the resulting weighted least squares estimates of the slope and intercept are

$$b_{0W} = \overline{y}_W - b_{1W}\overline{x}_W \quad \text{and} \quad b_{1W} = \frac{\sum w_i x_i y_i - \sum w_i \overline{x}_W \overline{y}_W}{\sum w_i x_i^2 - \sum w_i \overline{x}_W^2}$$

where $\overline{y}_W = \sum w_i y_i / \sum w_i$ and $\overline{x}_W = \sum w_i x_i / \sum w_i$.

Derive the formula for the weighted least squares estimate of the slope. You may assume that the formula for the weighted least squares estimate of the intercept is known.

2. Consider the multiple regression model with p predictor variables

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

(a) (2 marks) What is the distribution of $\mathbf{Y}|\mathbf{X}$?

(b) (3 marks) Derive the following result: $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

(c) (4 marks) Show that the matrix $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent.
(The only results you may assume are known are those on the formula sheet.)

(d) (2 marks) The residual sum of squares is $\mathrm{RSS} = \sum_{i=1}^{n} \hat{e}_i^2$. Show that $\mathrm{RSS} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$.

(Question 2 continued)

(e) (2 marks) Why is $\sum_{i=1}^{n} h_{ii} = p + 1$ where h_{ii} $(i = 1, \ldots, n)$ are the diagonal entries of the matrix $\mathbf{H}$?

(f) (4 marks) Show that $S^2 = \text{RSS}/(n - p - 1)$ is an unbiased estimator for σ^2.
For this question, you may use any of the following results if they are useful:
• $\text{E}(\mathbf{YY}' | \mathbf{X}) = \mathbf{X}\beta\beta'\mathbf{X}' + \sigma^2\mathbf{I}$ (shown in lecture)
• trace($\mathbf{AB}$) = trace($\mathbf{BA}$) assuming the products of the matrices can be calculated
• any of the facts you were required to show in parts (b) through (e) of this question

3. The data considered in this question are the same data considered in Assignment 1, plus another measurement taken at the same time. The goal of this analysis is to provide an accurate prediction of the amount of time to the next eruption of the Old Faithful geyser in Yellowstone National Park, so we need to predict the time interval from one eruption to the next eruption. The geyser was observed for a month in July 1995, and measurements were taken for every eruption that took place during the daytime. We will consider analyses using the following variables:
   - `interval` – The time (in minutes) from the previous eruption to the current eruption.
   - `duration` – The duration (in minutes) of the previous eruption.
   - `height` – The height of the previous eruption as estimated by a park ranger (in feet).

   On pages 16 through 18 there is SAS output for the analysis of these data.

   (a) (4 marks) Prior to an earthquake that took place in 1983, the Yellowstone National Park predicted the time between eruptions of Old Faithful by the equation

   $$\texttt{interval} = 30 + 10\ \texttt{duration}.$$

   Is there evidence that the slope of the linear relationship between `interval` and `duration` has changed since then? Since you are not given statistical tables of any probability distributions, you will need to rely on facts that you know about any relevant distribution.

   (b) (2 marks) Does the fact that the measurements were taken (mostly) consecutively over time affect your answer to part (a)? Why or why not?

(Question 3 continued)

(c) So far, we have investigated predicting the time interval between eruptions using the duration of the last eruption as the explanatory variable. However, the physical properties of the geyser suggest that the height as well as the duration of eruption may be useful in predicting the time interval between eruptions. One way to incorporate height into the analysis is to define a new explanatory variable, `height_dur` = `height` × `duration` as a measure that tries, in some sense, to capture information about the volume of water that is expelled from the geyser during an eruption. In the SAS output on pages 16 and 17, MODEL 1 uses `duration` as the explanatory variable and MODEL 2 uses `height_dur` as the explanatory variable. Suppose our goal is to compare these two models for which gives better predictions of the interval to the next prediction.

i. (1 mark) Why is R^2 not an appropriate statistic to use for the comparison of these two models for these data?

ii. (2 marks) What would be a good choice of statistic to use to compare these two models? Why?

(d) (5 marks) For the simple linear regression with `height_dur` as the explanatory variable (MODEL 2) you are given a plot of the standardized residuals versus the predicted values and a normal quantile plot of the standardized residuals. The plots are on page 17. State clearly what you are looking for in each plot and what you conclude.

Continued

(Question 3 continued)

(e) (3 marks) For the simple linear regression with `height_dur` as the explanatory variable (MODEL 2) Cook's D for one observation is equal to 0.040 and, for the same observation, the standardized residual is equal to 4.382. (These are by far the largest values of both Cook's D and the standardized residual among all observations.) Explain fully why this point is unusual.

(f) Another way to incorporate **height** into the model is to use multiple regression with **height** and **duration** as explanatory variables. The SAS output for this model is given on page 18 (MODEL 3).

i. (3 marks) Give a practical interpretation of the estimated coefficient of **duration** in MODEL 3.

ii. (2 marks) What do you conclude from the t-test associated with the coefficient of **height** in MODEL 3?

Continued

(Question 3(f) continued)

iii. (3 marks) Based on what you know from the output for MODEL 3, draw a sketch of the added variable plot for height. Indicate clearly what is being plotted on each axis.

iv. (2 marks) Explain what is wrong with the following statement:
"Comparing MODELS 1 and 3, Adjusted R^2 is higher for MODEL 3, as expected, since MODEL 3 has more explanatory variables than MODEL 1. So Adjusted R^2 does not help with determining whether or not `height` *is a useful predictor of the mean of* `interval` *over and above* `duration`*."*

(Question 3 continued)

(g) Another multiple regression that may be of interest has `duration`, `height` and `height_dur` (as defined in part (c)) as predictor variables. The complete SAS output is not given for this model but it was fit to these data and the resulting equation is

$$\widehat{\texttt{interval}} = 40.097 + 12.880\,\texttt{duration} - 0.050\,\texttt{height} + 0.003\,\texttt{height\_dur}$$

i. (2 marks) For the regression with explanatory variables `duration`, `height` and `height_dur`, explain how a change in duration of one minute affects the estimate of the mean of the interval to the next eruption.

ii. (2 marks) For the regression with explanatory variables `duration`, `height` and `height_dur`, the variance inflation factors are 79.5 for `duration`, 10.1 for `height` and 85.8 for `height_dur`. Are such large values surprising? Why or why not?

4. In this question, we are using the same data that we used in Assignment 2 about the movies released in 2003 that had wide distribution. We are interested in whether better reviews lead to better box office earnings. For all of the analyses carried out here, the six movies with the largest box office earnings have been removed from the data.

   The variables we will consider here are:
   • `score` – Critics' score, out of 100, calculated as a composite from a number of reviews.
   • `boxoffice` – The amount, in millions of dollars, that the movie earned at the box office.
   • `logboxoffice` – The natural log of `boxoffice`.
   • `ratingisG` – An indicator variable that is 1 if the movie is rated G and 0 otherwise.
   • `ratingisPG` – An indicator variable that is 1 if the movie is rated PG and 0 otherwise.
   • `ratingisPG13` – An indicator variable that is 1 if the movie is rated PG-13 and 0 otherwise.
   • `score_G` – The product of `score` and `ratingisG`.
   • `score_PG` – The product of `score` and `ratingisPG`.
   • `score_PG13` – The product of `score` and `ratingisPG13`.

   On pages 19 through 21 there is SAS output for the analysis of these data.

   (a) (2 marks) You are given the following plots:
   • On page 19: a scatterplot of `boxoffice` versus `score`, a plot of the standardized residuals versus the predicted values from the regression of `boxoffice` on `score`, a normal quantile plot of the standardized residuals from the regression of `boxoffice` on `score`.
   • On page 20: a scatterplot of `logboxoffice` versus `score`, a plot of the standardized residuals versus the predicted values from the regression of `logboxoffice` on `score`, a normal quantile plot of the standardized residuals from the regression of `logboxoffice` on `score`.

   Based on these plots, should we use the untransformed or log transformed values of box office earnings? Explain.

Continued

(Question 4 continued)

(b) (2 marks) Assuming that appropriate weights were known, would weighted least squares be appropriate for these data? Why or why not?

(c) Regardless of what you concluded in part (a) and whether or not this is the right choice of model, analysis is given using the log transformation of box office earnings on page 21. The following questions are related to this analysis.

i. (5 marks) Several numbers on the SAS output have been replaced by letters. What are the missing values?

(A) = ______________________

(B) = ______________________

(C) = ______________________

(D) = ______________________

(E) = ______________________

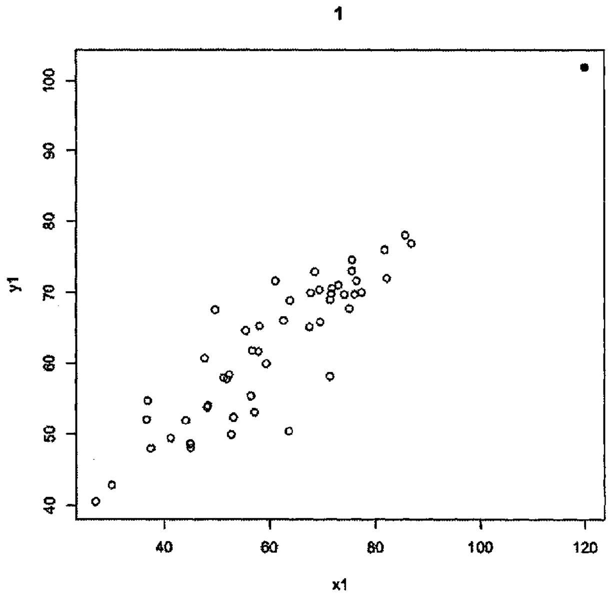
ii. (3 marks) For a movie with a critic's score of 50, estimate what it's box office earnings would be if (1) the movie is rated PG and (2) the movie is rated R. If you don't have enough information, indicate what you need in order to calculate the estimates.

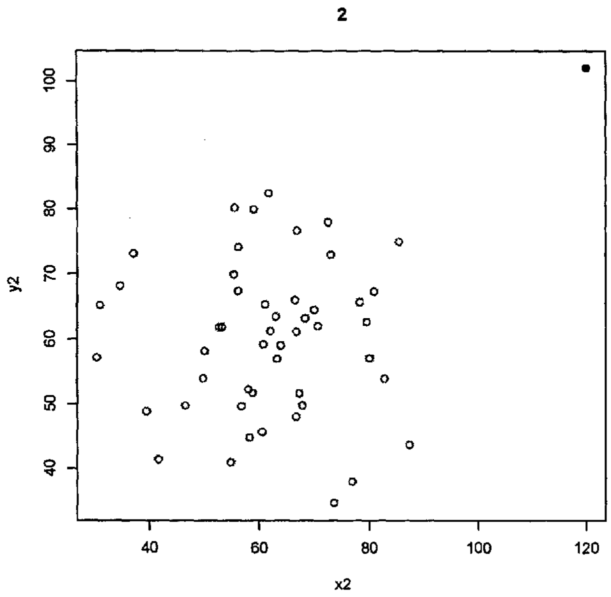
(Question 4(c) continued)

iii. (6 marks) Is a parallel lines model appropriate for these data? Carry out an appropriate statistical test. Indicate:
(I) The null and alternative hypotheses
(II) The value of the test statistic
(III) The distribution of the test statistic under the null hypothesis
(IV) What you can say about the p-value
(V) A practical conclusion.
Since no tables of probability distributions have been provided, you will need to rely on what you know about the relevant probability distribution.

5. Below are two scatterplots. For each we are interested in the relationship between a response (on the vertical axis) and an explanatory variable (on the horizontal axis). In each plot, one point in the upper right has been drawn with a solid black dot.

   Questions about these plots are below and on the next page.





(a) (6 marks) For each plot, state which of the following are true of the point drawn with a solid black dot.
(1) It is a leverage point.
(2) It is an influential point.
(3) It is an outlier that indicates model failure.

True of point in the first plot:

True of point in the second plot:

(Question 5 continued)

(b) (6 marks) For each plot, state the effect of removing the point drawn with a solid black dot on the following quantities. That is, state whether the quantity will be approximately the same, smaller, or larger when the point is removed.
(1) the estimate of the slope
(2) the standard error of the slope
(3) R^2

When the point is removed in the first plot:

When the point is removed in the second plot:

(c) (4 marks) Explain how each plot illustrates the fact that high values of R^2 are insufficient to determine the adequacy of the regression line.

First plot:

Second plot:

The SAS output on pages 16 to 18 is relevant to question 3.

MODEL 1

Simple linear regression with **duration** as the explanatory variable

---

```
                              The REG Procedure

          Number of Observations Read                          343
          Number of Observations Used                          263
          Number of Observations with Missing Values            80


                            Descriptive Statistics

                                        Uncorrected                     Standard
Variable           Sum          Mean             SS       Variance     Deviation
Intercept    263.00000       1.00000      263.00000              0             0
duration     874.91667       3.32668     3274.46416        1.38892       1.17852
interval         20394      77.54373        1656660      287.14980      16.94549


                         Dependent Variable: interval

                             Analysis of Variance

                                      Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       1          64229          64229    1523.31    <.0001
Error                     261          11005       42.16368
Corrected Total           262          75233


          Root MSE               6.49336    R-Square     0.8537
          Dependent Mean        77.54373    Adj R-Sq     0.8532
          Coeff Var              8.37380


                             Parameter Estimates
                          Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|

     Intercept     1       33.34745        1.20108      27.76      <.0001
     duration      1       13.28540        0.34039      39.03      <.0001
```

(SAS output for question 3 continues on the next page)

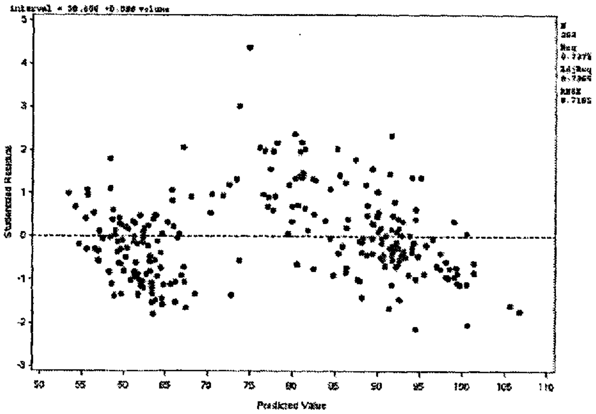
Continued

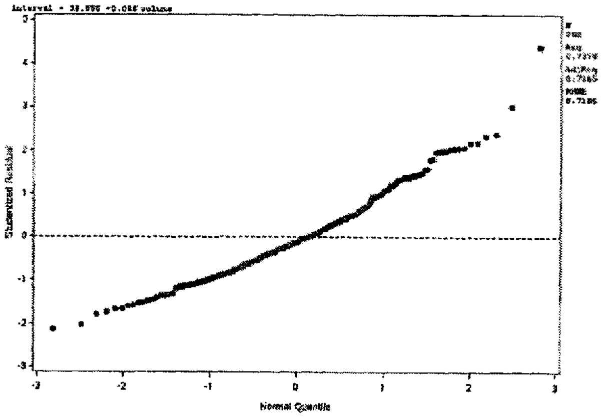
(SAS output for question 3 continued)

MODEL 2

Simple linear regression with `height_dur` as the explanatory variable

```
                              The REG Procedure

          Number of Observations Read                          343
          Number of Observations Used                          252
          Number of Observations with Missing Values            91

                         Dependent Variable: interval

                             Analysis of Variance
                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F
Model                     1           53401          53401     702.54    <.0001
Error                   250           19003       76.01177
Corrected Total         251           72404

          Root MSE              8.71847    R-Square     0.7375
          Dependent Mean       77.72619    Adj R-Sq     0.7365
          Coeff Var            11.21690

                             Parameter Estimates
                            Parameter       Standard
     Variable     DF         Estimate          Error    t Value    Pr > |t|
     Intercept     1         38.65631        1.57302      24.57      <.0001
     height_dur    1          0.08798        0.00332      26.51      <.0001
```





(SAS output for question 3 continues on the next page)

MODEL 3

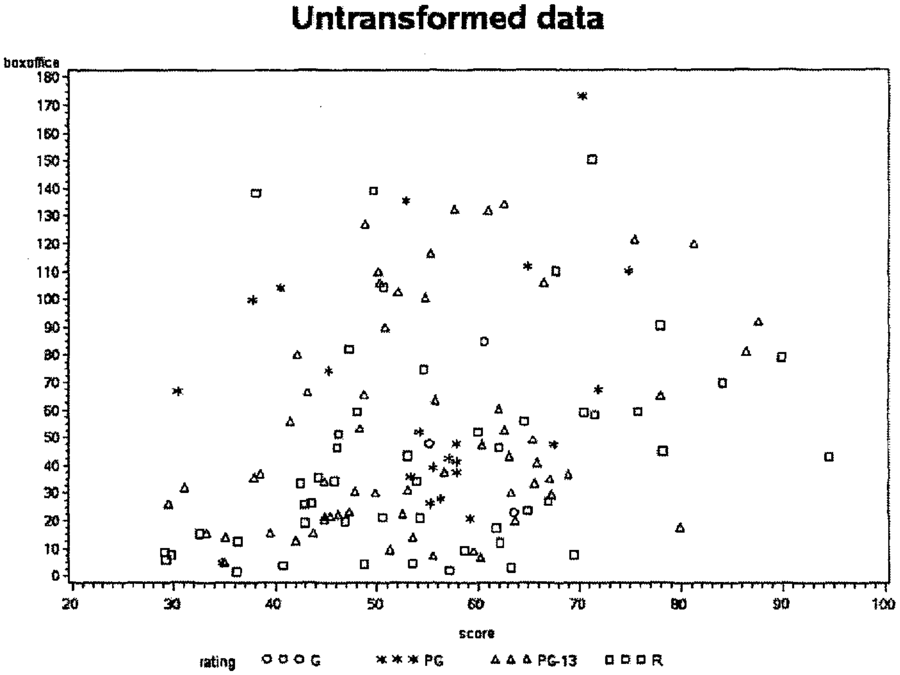
Multiple regression with **duration** and **height** as the explanatory variables

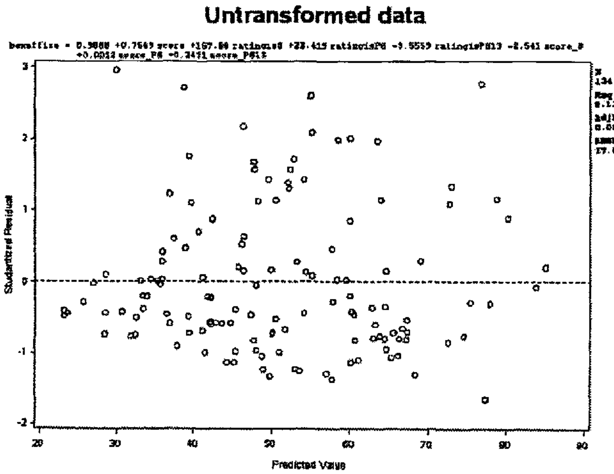
```
                             The REG Procedure
                               Model: MODEL1
                        Dependent Variable: interval

           Number of Observations Read                          343
           Number of Observations Used                          252
           Number of Observations with Missing Values            91


                            Analysis of Variance

                                      Sum of           Mean
Source                      DF       Squares         Square    F Value    Pr > F

Model                        2         62278          31139     765.73    <.0001
Error                      249         10126       40.66596
Corrected Total            251         72404


            Root MSE              6.37699    R-Square     0.8601
            Dependent Mean       77.72619    Adj R-Sq     0.8590
            Coeff Var             8.20442


                            Parameter Estimates

                       Parameter       Standard
 Variable     DF        Estimate          Error    t Value    Pr > |t|

 Intercept     1        38.71956        3.71178      10.43      <.0001
 duration      1        13.27599        0.34071      38.97      <.0001
 height        1        -0.04001        0.02589      -1.55      0.1235
```

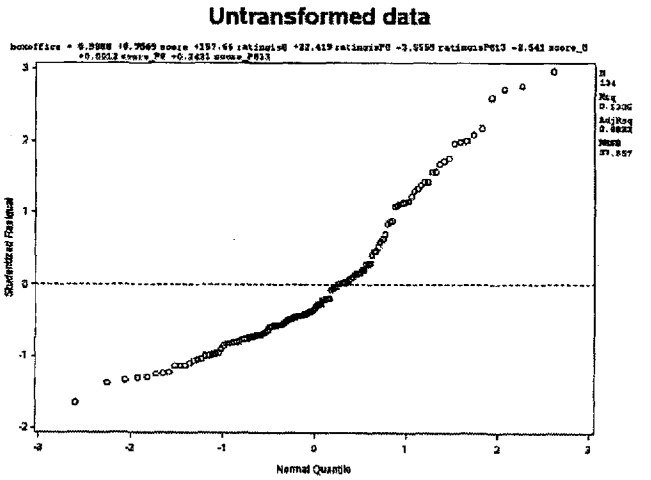
Continued

The SAS output on pages 19 to 21 is relevant to question 4.

Plots for untransformed data





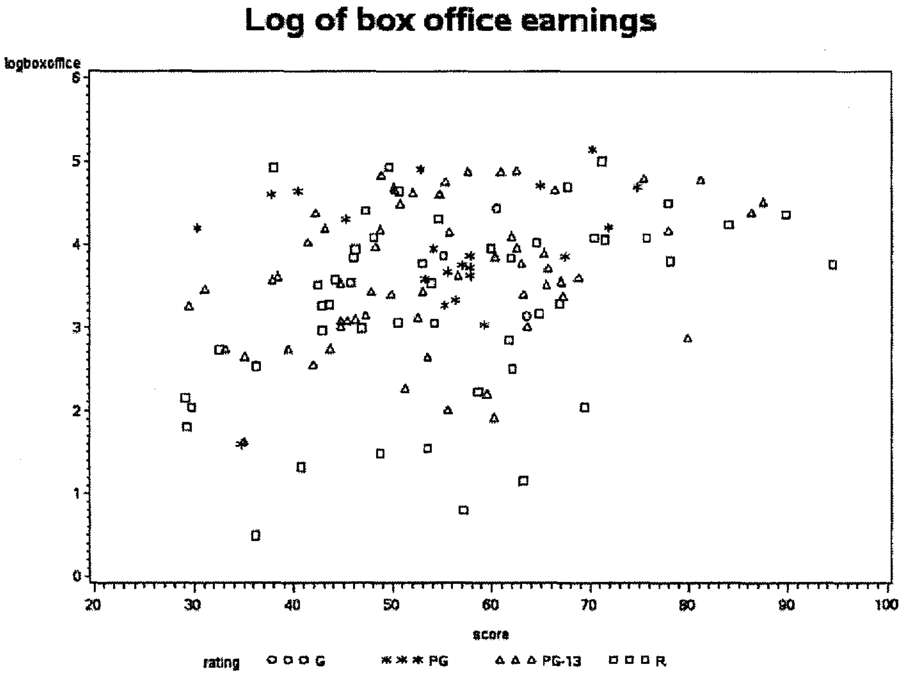


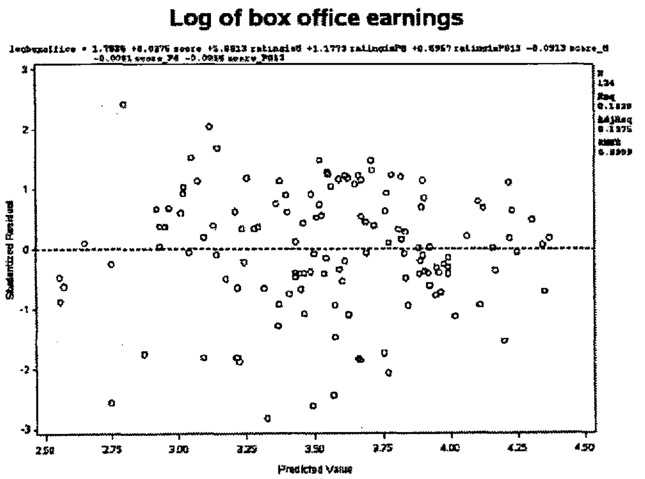
(SAS output for question 4 continues on the next page)

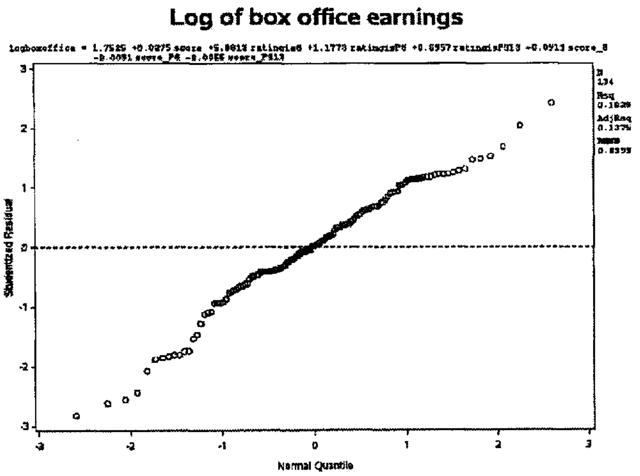
Continued

(SAS output for question 4 continued)

Plots using log of box office earnings







(SAS output for question 4 continues on the next page)

Continued

(SAS output for question 4 continued)

Regression model using log of box office earnings

```
                          The REG Procedure
                           Model: MODEL1
                    Dependent Variable: logboxoffice

              Number of Observations Read          134
              Number of Observations Used          134


                         Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                    (A)      22.84489            (D)        (E)    0.0005
Error                    (B)     102.04126        0.80985
Corrected Total          (C)     124.88615


       Root MSE              0.89992    R-Square     0.1829
       Dependent Mean        3.55928    Adj R-Sq     0.1375
       Coeff Var            25.28367


                         Parameter Estimates

                     Parameter     Standard
  Variable      DF    Estimate        Error  t Value  Pr > |t|    Type I SS

  Intercept      1     1.75254      0.47699     3.67    0.0004   1697.57823
  score          1     0.02746      0.00830     3.31    0.0012     14.38005
  ratingisG      1     5.88130      8.96721     0.66    0.5131      0.06480
  ratingisPG     1     1.17734      1.06271     1.11    0.2700      3.69510
  ratingisPG13   1     0.69570      0.68359     1.02    0.3108      4.14771
  score_G        1    -0.09129      0.14972    -0.61    0.5431      0.27977
  score_PG       1    -0.00912      0.01881    -0.48    0.6287      0.10601
  score_PG13     1    -0.00554      0.01203    -0.46    0.6462      0.17144
```

Continued

Simple regression formulae

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Regression in matrix terms

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

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

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

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

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

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

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

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

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

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

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

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

Total pages 22
Total marks 95