UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER EXAMINATIONS 2009
STA 302 H1F / STA 1001 HF

Duration - 3 hours

Aids Allowed: Calculator

LAST NAME:________________________**FIRST NAME:**______________________

STUDENT NUMBER: ____

- There are 23 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.
- A table of the t distribution can be found on page 19 and tables of the F distribution can be found on pages 20, 21 and 22.
- Total marks: 90

1ab	1cde	2abcd	2ef	3abcde	3fg	3h(i,ii,iii)

3h(iv,v)	4ab	4cd	4e	5	6a	6bc

1. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$

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) Explain the method of least squares as used in simple linear regression. Use language that someone who has studied no statistics can understand.

(b) (5 marks) Derive the formula for b_1, the least squares estimate of β_1. (Your answer should be one of the expressions on the formula sheet. You may assume that the formula for b_0 is known.)

(Question 1 continued.)

(c) (4 marks) $\hat{Y}_i$ is the value of Y for the ith observation estimated from the fitted regression line. Show that $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$.

(d) (2 marks) The formula for h_{ij} is on the formula sheet. We are usually most interested in h_{ii} $(i = 1, \ldots, n)$, the values of h_{ij} when $i = j$. What do the h_{ii} measure? Why are they of interest?

(e) (2 marks) Show that $\sum_{i=1}^{n} h_{ii} = 2$.

2. A multiple linear regression model with dependent variable Y and k explanatory variables is fit to n observations $(X_{i1}, X_{i2}, \ldots, X_{ik}, Y_i)$, $i = 1, \ldots, n$. You may assume that the X's are not random.

   (a) (3 marks) State the multiple regression model in matrix terms, defining all matrices and vectors.

   (b) (5 marks) State all the assumptions of the multiple regression model that are necessary to estimate and carry out inference on the model parameters.

   (c) (2 marks) Which assumption is most critical? Why?

   (d) (2 marks) Which assumption is least critical? Why?

(Question 2 continued.)

(e) (3 marks) What is the variance-covariance matrix of the vector of the fitted values of Y?

(f) (2 marks) Show that the vector of least squares estimators $\mathbf{b}$ is unbiased for the vector of model parameters $\boldsymbol{\beta}$.

3. A company that publishes a newspaper in a mid-size American city wants to investigate the feasibility of introducing a Sunday edition of the paper. The current circulation (the average number of newspapers sold per day) of the company's weekday newspaper is 210,000. The goal of this analysis is to predict the Sunday circulation of a newspaper with a weekday circulation of 210,000.

   The data are circulations of 89 U.S. newspapers that publish both weekday and Sunday editions.

   Analysis was carried out on the natural logarithms of the circulations.

   Some output from SAS is below. `logSunCirc` is the natural logarithm of the Sunday circulation and `logWkdayCirc` is the natural logarithm of the weekday circulation.

```
                            Descriptive Statistics

                                         Uncorrected                  Standard
 Variable              Sum         Mean           SS     Variance    Deviation

 Intercept        89.00000      1.00000     89.00000            0            0
 logWkdayCirc   1101.17108     12.37271        13651      0.30582      0.55301
 logSunCirc     1126.23888     12.65437        14281      0.33017      0.57460

                               The REG Procedure
                        Dependent Variable: logSunCirc

                    Number of Observations Read          89
                    Number of Observations Used          89

                             Analysis of Variance
                                    Sum of           Mean
Source                     DF      Squares         Square     F Value    Pr > F
Model                       1     26.20543       26.20543      800.19    <.0001
Error                      87      2.84916        0.03275
Corrected Total            88     29.05458

             Root MSE              0.18097    R-Square     0.9019
             Dependent Mean       12.65437    Adj R-Sq     0.9008
             Coeff Var             1.43007

                              Parameter Estimates
                            Parameter       Standard
    Variable        DF       Estimate          Error     t Value    Pr > |t|
    Intercept        1        0.44511        0.43204        1.03      0.3057
    logWkdayCirc     1        0.98679        0.03488       28.29      <.0001
```

Questions about this output begin on the next page.

Continued

(Question 3 continued.)

(a) (1 mark) What are the null and alternative hypotheses for the test with test statistic 800.19?

(b) (1 mark) What percent of variability in the dependent variable is explained by its linear relationship with the independent variable?

(c) (3 marks) Use the Bonferroni method to find simultaneous 90% confidence intervals for the slope and the intercept.

(d) (2 marks) What does it mean for the confidence intervals in part (c) to be "simultaneous"?

(e) (2 marks) Explain how to interpret the estimated slope in a practical way. Your answer should be in terms of the original circulations and not the log transformations of them.

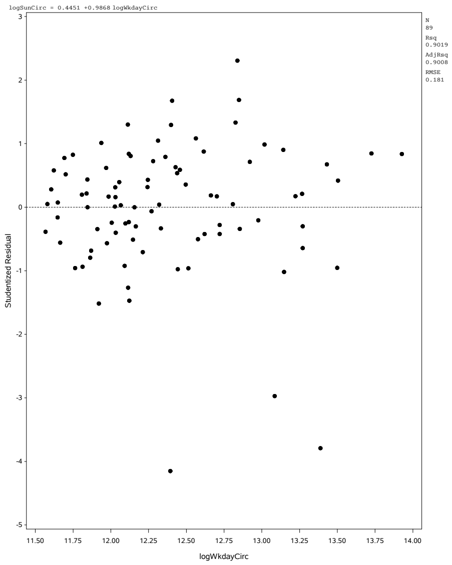
(Question 3 continued.)

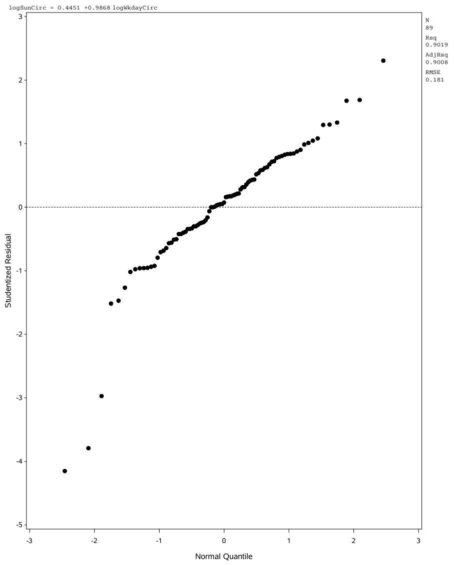
(f) (2 marks) In order to satisfy the usual regression model assumptions, it was necessary to take the log transform of both Sunday circulation and weekday circulation. Describe the features of the plot of Sunday circulation versus weekday circulation (that is, the scatterplot of the variables before transformation) that indicate that the log transformation of both of the variables is necessary?

(g) (5 marks) Calculate a 95% interval estimate of the Sunday circulation for the newspaper that is considering adding a Sunday newspaper. (Recall that its weekday circulation is 210,000.) (Note that you'll first need to calculate the interval for the log of the circulation, and then back-transform it for an interval estimate of the circulation.)

(Question 3 continued.)

(h) The plots below are a plot of the studentized residuals versus the explanatory variable and a normal quantile plot of the studentized residuals for the regression whose output is on page 6.





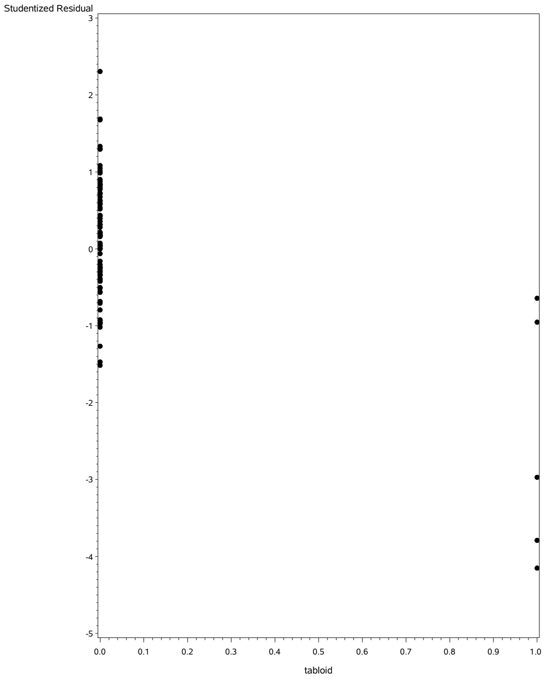
i. (1 mark) You are not given the plot of the studentized residuals versus the predicted values. Describe what it would look like.

ii. (4 marks) What are you looking for in the plot of the studentized residuals versus the explanatory variable? What do you conclude?

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

iv. (2 marks) Studentized rather than raw residuals were plotted. (The “raw” residuals are $e_i = Y_i - \hat{Y}_i$.) What are the advantages of looking at plots of studentized residuals rather than raw residuals?

v. (2 marks) Some of the newspapers in the dataset are tabloids. Below is a plot of the studentized residuals versus an indicator variable which is 1 if the newspaper is a tabloid and 0 otherwise. What additional information do you learn from this plot?



4. Which aspects of a professional golfer's play are most important in determining the amount of prize money he will earn? In order to answer this question, data on the top 196 professional golfers in 2006 were collected. The aspects of a golfer's play we will consider are listed below. For each aspect, it is noted whether a high or low value indicates that the golfer is performing well.

- `daccuracy`: Driving Accuracy is the percent of time the golfer hit the fairway from the tee. High values are good.
- `gir`: Greens in Regulation is the percent of time the golfer hit the green in the number of shots allotted for this. High values are good.
- `puttavg`: Putting Average is the average number of putts to score on holes where the green is hit in regulation. Low values are good.
- `birdies`: The percent of time that the golfer makes a birdie after hitting the green in regulation. High values are good.
- `sandsaves`: The percent of time a golfer was able to get out of a sand bunker. High values are good.
- `scrambling`: The percent of time a golfer misses the green in regulation but recovers. High values are good.
- `nputts`: The average number of putts per round. Low values are good.

The analysis has been carried out using the natural logarithm of prize money (`logprizemoney`) as the dependent variable. Some output from SAS is below.

```
                         The REG Procedure
        Number of Observations Read                      196
        Number of Observations Used                      193
        Number of Observations with Missing Values         3

                              Correlation
 Variable            daccuracy             gir         puttavg          birdies
 daccuracy              1.0000          0.4114         -0.0176          -0.2637
 gir                    0.4114          1.0000          0.0710           0.0130
 puttavg               -0.0176          0.0710          1.0000          -0.7662
 birdies               -0.2637          0.0130         -0.7662           1.0000
 sandsaves              0.0348         -0.0804         -0.2651           0.1305
 scrambling             0.3859          0.1830         -0.1939          -0.0335
 nputts                 0.0702          0.4935          0.7941          -0.5060
 logprizemoney          0.1667          0.4936         -0.4215           0.4590

                             Correlation
Variable            sandsaves       scrambling          nputts    logprizemoney
daccuracy              0.0348           0.3859          0.0702           0.1667
gir                   -0.0804           0.1830          0.4935           0.4936
puttavg               -0.2651          -0.1939          0.7941          -0.4215
birdies                0.1305          -0.0335         -0.5060           0.4590
sandsaves              1.0000           0.5058         -0.4207           0.2457
scrambling             0.5058           1.0000         -0.4079           0.3519
nputts                -0.4207          -0.4079          1.0000          -0.1745
logprizemoney          0.2457           0.3519         -0.1745           1.0000
```

(SAS output for this question continues on the next page.)

Continued

(Question 4 continued.)

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

            Number of Observations Read                         196
            Number of Observations Used                         193
            Number of Observations with Missing Values            3

                             Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     7       97.41905       13.91701      31.31    <.0001
Error                   185       82.23911        0.44454
Corrected Total         192      179.65817

             Root MSE              0.66674    R-Square     0.5422
             Dependent Mean       10.39148    Adj R-Sq       XXX
             Coeff Var             6.41617

                             Parameter Estimates

                         Parameter     Standard                       Variance
  Variable       DF       Estimate        Error  t Value  Pr > |t|   Inflation
  Intercept       1        0.61671      7.83849     0.08    0.9374           0
  daccuracy       1       -0.00413      0.01186    -0.35    0.7280     1.78632
  gir             1        0.19557      0.04440     4.40    <.0001     6.27168
  puttavg         1       -1.01815      6.96608    -0.15    0.8840    12.86093
  birdies         1        0.15287      0.04092     3.74    0.0002     3.48775
  sandsaves       1        0.01621      0.00997     1.63    0.1057     1.47982
  scrambling      1        0.04953      0.03199     1.55    0.1232     4.30293
  nputts          1       -0.30758      0.47911    -0.64    0.5217    19.45000
```

(a) (3 marks) Calculate adjusted R^2. In multiple regression, why is it preferred over R^2?

(b) (2 marks) Give a practical interpretation for the coefficient of `birdies`.

(Question 4 continued.)

(c) (2 marks) The p-value for the t-test for the coefficient of `birdies` is 0.0002. What do you conclude from this?

(d) (2 marks) Give 2 indications from the SAS output above that there are problems with multicollinearity.

(e) The t-tests for the coefficients of 5 of the predictor variables (`daccuracy`, `puttavg`, `sandsaves`, `scrambling` and `nputts`) have high p-values. These 5 predictors were removed from the model and the data were re-fit to the reduced model giving the following SAS output:

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

          Number of Observations Read                       196
          Number of Observations Used                       193
          Number of Observations with Missing Values          3

                       Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     2       80.58561       40.29280      77.27    <.0001
Error                   190       99.07256        0.52143
Corrected Total         192      179.65817

           Root MSE              0.72210    R-Square     0.4485
           Dependent Mean       10.39148    Adj R-Sq     0.4427
           Coeff Var             6.94900

                       Parameter Estimates

                         Parameter       Standard
   Variable        DF     Estimate          Error    t Value    Pr > |t|
   Intercept        1     -6.72601        1.42201      -4.73      <.0001
   gir              1      0.17385        0.01920       9.05      <.0001
   birdies          1      0.19940        0.02373       8.40      <.0001
```

(Questions continue on the next page.)

(Question 4 continued.)

i. (2 marks) Was it a good idea to remove the 5 predictors with high p-values and fit this reduced model? Why or why not?

ii. (4 marks) Carry out an hypothesis to test simultaneously whether all of the variables removed from the original model have coefficients equal to 0.

5. In Assignment 1, we examined whether there is a relationship between an NFL kicker's field-goal percentage one year and the previous year. For each of 19 kickers, we have data for four consecutive years. The analysis of these data below has as the dependent variable the percentage of field goals scored in one year (`FG`) and as independent variables: the percentage of field goals scored in the previous year (`prevFG`) and 18 indicator variables for the football players. For example, the indicator variable `AV` is 1 if the observation is for the kicker with initials A.V. and 0 if the observation is for another kicker. The regression below fits 19 separate lines, one for each player. In this regression, the lines are parallel.

```
                               The REG Procedure
                             Dependent Variable: FG

                      Number of Observations Read          76
                      Number of Observations Used          76

                              Analysis of Variance
                                      Sum of           Mean
Source                   DF          Squares         Square    F Value    Pr > F
Model                    19       2339.66699      123.14037       3.19    0.0004
Error                    56       2160.95656       38.58851
Corrected Total          75       4500.62355

             Root MSE              6.21197    R-Square     0.5199
             Dependent Mean       82.25921    Adj R-Sq     0.3569
             Coeff Var             7.55170

                              Parameter Estimates
                             Parameter       Standard
     Variable     DF          Estimate          Error     t Value     Pr > |t|
     Intercept     1         116.31347        9.32238       12.48       <.0001
     prevFG        1          -0.50370        0.11276       -4.47       <.0001
     AV            1          10.37368        4.45141        2.33       0.0234
     DA            1           5.72740        4.41591        1.30       0.2000
     JE            1           7.35703        4.39774        1.67       0.0999
     JaH           1          12.49090        4.47697        2.79       0.0072
     JR            1           2.07814        4.41828        0.47       0.6399
     JW            1          12.68387        4.44053        2.86       0.0060
     JC            1           4.39629        4.40068        1.00       0.3221
     JoH           1           1.88722        4.39253        0.43       0.6691
     KB            1          -2.98610        4.40553       -0.68       0.5007
     MS            1          19.10997        4.51993        4.23       <.0001
     MV            1          15.26923        4.49769        3.39       0.0013
     NR            1           3.75369        4.42014        0.85       0.3994
     OM            1          -2.66278        4.39253       -0.61       0.5468
     PD            1          13.92609        4.46388        3.12       0.0029
     RiL           1           5.50629        4.39670        1.25       0.2156
     RyL           1           8.14221        4.42371        1.84       0.0710
     SJ            1           6.39740        4.40287        1.45       0.1518
     SG            1          12.50869        4.43971        2.82       0.0067
```

Questions begin on the next page.

(Question 5 continued.)

(a) (2 marks) What is the fitted regression line for the kicker with initials S.G.?

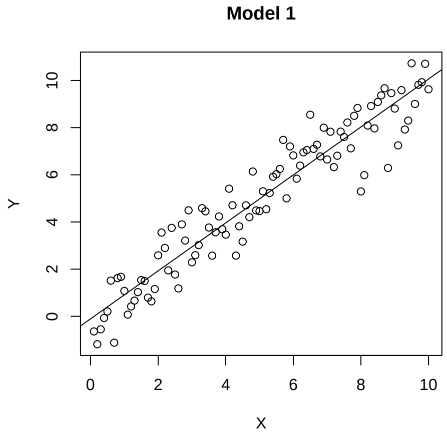
(b) (2 marks) If 19 indicator variables had been included in the model, one for each of the 19 kickers, SAS would have deleted one. Why?

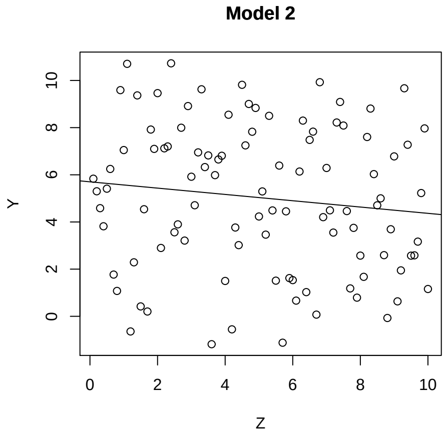
(c) (1 mark) Before fitting the parallel lines model as above, it should first be checked whether or not the data are consistent with parallel lines for each kicker. How should the model fit above be adjusted so that a different slope (and intercept) is estimated for each kicker?

(d) (3 marks) A statistical test can be carried out on the model in part (c) to determine whether it was reasonable to model the 19 lines as parallel. Indicate the type of test that is appropriate, the null and alternative hypotheses, and the distribution of the test statistic under the null hypothesis.

6. The following questions require short answers.

   (a) (3 marks) Two alternative straight line regression models have been proposed for Y. In the first model, Y is a linear function of X, while in the second model Y is a linear function of Z. The plots below show scatterplots with the fitted regression lines of Y versus X (first plot) and Y versus Z (second plot).





   Which one of the following statements is true? Give a detailed reason to support your choice.

   i. SSE for model 1 is greater for SSE for model 2, while SSR for model 1 is greater than SSR for model 2.

   ii. SSE for model 1 is less than SSE for model 2, while SSR for model 1 is less than SSR for model 2.

   iii. SSE for model 1 is greater than SSE for model 2, while SSR for model 1 is less than SSR for model 2.

   iv. SSE for model 1 is less than SSE for model 2, while SSR for model 1 is greater than SSR for model 2.

(Question 6 continued.)

(b) (3 marks) Suppose a multiple regression model with 5 predictor variables is fit to some data. The analysis of variance F-test is statistically significant (p-value< 0.01) but the t-tests for the coefficients of the predictor variables are all not statistically significant (all 5 p-values are > 0.10). What do you conclude? Explain.

(c) (1 mark) Explain the purpose of using centering in polynomial regression.

Continued

Continued

Continued

Continued

Simple regression formulae

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

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

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

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

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

$$\mathrm{SSTO} = \sum (Y_i - \overline{Y})^2$$

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

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

$$\sigma^2\{\hat{Y}_h\} = \mathrm{Var}(\hat{Y}_h) = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)$$

$$\sigma^2\{\mathrm{pred}\} = \mathrm{Var}(Y_h - \hat{Y}_h) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)$$

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

$$S_{XX} = \sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2$$

$$h_{ij} = \frac{1}{n} + \frac{(X_i - \overline{X})(X_j - \overline{X})}{S_{XX}}$$

Regression in matrix terms

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

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

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

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

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

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

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

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

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

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

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

$$\sigma^2\{\mathrm{pred}\} = \mathrm{Var}(Y_h - \hat{Y}_h) = \sigma^2 (1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

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

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

Total pages 23
Total marks 90