UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER EXAMINATIONS 2005

STA 302 H1F / STA 1001 H1F

Duration - 3 hours

Aids Allowed: Calculator

LAST NAME: SOLUTIONS **FIRST NAME:** ____________________

STUDENT NUMBER: ____

- There are 17 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 14 and tables of the F distribution can be found on pages 15 and 16.
- Total marks: 95

1	2abc	2de	2f	3	4ab

4cde	5a	5bcde	6	7

1. Suppose we have $n = 102$ pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we fit the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Here are some summary statistics:

$$\begin{array}{cc} \overline{X} = 50 & \overline{Y} = 100 \\ \sum_{i=1}^n (X_i - \overline{X})^2 = 100 & \sum_{i=1}^n (Y_i - \overline{Y})^2 = 200 \\ \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = 100 & \text{SSR} = 100 \end{array}$$

(a) (5 marks) Complete the following ANOVA table:

Source	df	SS	MS	F
Regression	1	100	100	100
Error	100	100	1	
Total	101	200		

(b) (3 marks) Estimate the slope and give a 95% confidence interval for β_1.

$b_1 = 100/100 = 1$
$t_{100,.025} \doteq 2$
95% CI for β_1*:* $1 \pm 2(1)/\sqrt{100} = (0.8, 1.2)$

(c) (2 marks) Use the ANOVA table to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

Test statistic: $F_{obs} = 100$
Approximating $F_{1,100}$ *with* $F_{1,60}$ *distribution,* $p < 0.001$
Strong evidence that $\beta_1 \neq 0$

(d) (3 marks) Give a 90% prediction interval for a new observation at $X = 50$.

Since $\overline{X} = 50$, $\hat{Y}$ *at* $X = 50$ *is* $\overline{Y} = 100$
$t_{100,0.05} = 1.71$
Prediction interval: $100 \pm 1.671(1)\sqrt{1 + \frac{1}{102} + 0} = (98.3, 101.7)$

2. The data in this question were collected as part of a study of the adult female Dungeness crab. While planning fishing restrictions to control crab populations, biologists want to study the growth rate of crabs. The data are measurements of the widest part of the crabs' shells, in millimeters. Crabs molt regularly, casting off their old shells and growing new ones. Of particular interest is predicting the size of the shell before molting (variable name: `presize`) having observed the size of the shell after the crab molted (variable name: `postsize`).

   SAS output is given below for the regression of `postsize` on `presize` for 342 adult female crabs raised in a laboratory setting.

```
                          The REG Procedure
                 Number of Observations Read        342
                 Number of Observations Used        342

                       Descriptive Statistics
                                    Uncorrected                    Standard
Variable          Sum         Mean           SS     Variance      Deviation
Intercept   342.00000      1.00000    342.00000            0              0
postsize        49151    143.71696      7096984     97.13490        9.85570
presize         44133    129.04357      5736616    121.84434       11.03831

                          The REG Procedure
                            Model: MODEL1
                     Dependent Variable: presize

                        Analysis of Variance
                                  Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F
Model                    1         40192          40192    10072.0    <.0001
Error                  340    1356.76275        3.99048
Corrected Total        341         41549

             Root MSE              1.99762    R-Square     0.9673
             Dependent Mean      129.04357    Adj R-Sq     0.9672
             Coeff Var             1.54802

                        Parameter Estimates
                          Parameter       Standard
     Variable      DF      Estimate          Error     t Value     Pr > |t|
     Intercept      1     -29.26843        1.58114      -18.51       <.0001
     postsize       1       1.10155        0.01098      100.36       <.0001
```

Questions related to these data are on the next three pages.

Continued

(a) (7 marks) Complete the chart below.

Statistic	**Observed Value**
Slope of line	1.10155
Correlation between `postsize` and `presize`	$\sqrt{0.9673} = 0.9835$
Average change in `presize` for an increase of 10 mm in `postsize`	$1.10155(10)$ $= 11.015$ mm
Estimate of `presize` when `postsize` is 130 mm	$-29.27 + 1.10155(130)$ $= 113.93$
Estimated variance of the intercept	$(1.58114)^2 = 2.5$
Test statistic for test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$	100.36 or 10072.0
Estimate of σ^2	3.99048

(b) (2 marks) Assume the usual simple linear regression assumptions hold. What is the distribution of the values of `presize` for crabs with `postsize` of 130 mm?

$N(113.93, 3.99048)$ *when mean and variance are approximated from the data*

(c) (3 marks) Find the limits of the 95% Working-Hotelling confidence interval when `postsize` is 130 mm.

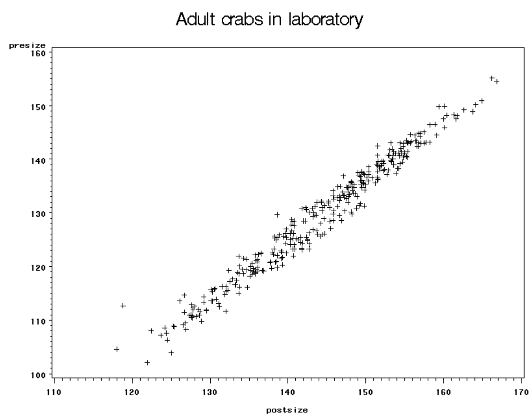
$W = \sqrt{2F_{2,340,.05}} \doteq 2.48$

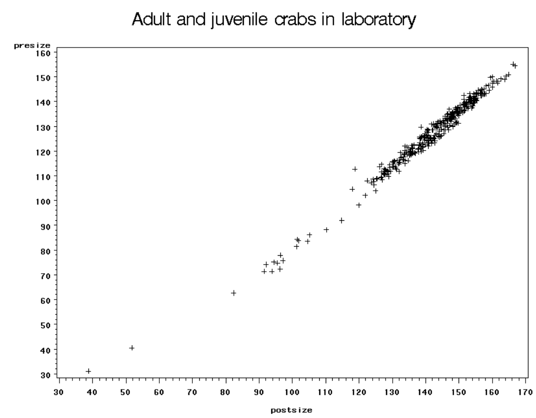
$113.93 \pm 2.48\sqrt{3.99048}\sqrt{\frac{1}{342} + \frac{(130-143.71696)^2}{7096984-342(143.71696)^2}} = (113.47, 114.39)$

(d) (2 marks) Explain the meaning of the Working-Hotelling confidence interval in part (c).

At least 95% of the time, for repressions on samples of the same size, the procedure gives a confidence interval that captures the line (the mean of Y) at all values of X in the range. The interval in (c) is an example.

(e) (4 marks) Two scatterplots are given below. The first is the plot of `presize` versus `postsize` for the 342 adult female crabs considered in the analysis above. The second plot includes these 342 crabs, as well as an additional 19 juvenile crabs.





How will adding the juvenile crabs to the regression affect the estimated slope and the value of R^2? Explain.

The points mostly follow the linear pattern of the other points except for 2 points on the far left which may cause the slope to be a little less.
R^2 will increase as there is more variation in the Y's (greater range) and the line explains this well.

Continued

(f) (3 marks) A quadratic model was also fit to the original data for the 342 adult females (new variable: `postsize2` is the square of `postsize`) and some resulting SAS output is given below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	40211	20106	5096.25	<.0001
Error	339	1337.42536	3.94521		
Corrected Total	341	41549			

Root MSE	1.98625	R-Square	0.9678
Dependent Mean	129.04357	Adj R-Sq	0.9676
Coeff Var	1.53921		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.40308	19.33811	0.69	0.4887
postsize	1	0.50049	0.27171	1.84	0.0663
postsize2	1	0.00211	0.00095144	2.21	0.0275

Find the coefficient of partial determination for the inclusion of the quadratic term in the model given the linear term, and interpret its meaning.

$$\frac{40211 - 40192}{1356.76} = 0.014$$

Measures the additional variability explained by the quadratic term given that the linear term is in the model.

3. Assume we are fitting the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $n \times 1$ vectors, $\mathbf{X}$ is a $n \times (k+1)$ matrix, and $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector. Recall that the least squares estimate of $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\text{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Assume the Gauss-Markov conditions apply.

(a) (2 marks) State the Gauss-Markov conditions for this model.

$$E(\boldsymbol{\epsilon}) = \mathbf{0}$$

$$\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

(b) (2 marks) Show that $\mathbf{b}$ is unbiased for $\boldsymbol{\beta}$.

$$\begin{aligned} E(\mathbf{b}) &= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} \end{aligned}$$

(c) (4 marks) The hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Show that $\mathbf{H}$ is symmetric and idempotent.

Symmetric: $\mathbf{H}' = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$
Idempotent: $\mathbf{H}^2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$

(d) (4 marks) Suppose we express the jth independent variable X_j in centimeters instead of meters (all other variables don't change). (There are 100 centimeters in a meter.) What will happen to b_j and the variance of b_j?

$(j+1)$th column of $\mathbf{X}$ matrix will be $100\times$ its previous values
$b_j = (j+1)$th component of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ will change by a factor of $\frac{1}{100^2}(100) = \frac{1}{100}$ (increase in X_j of 100 cm corresponds to a change of 1 m)
$Var(b_j) = (j+1)$th diagonal entry of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ will change by a factor of $\frac{1}{100^2}$

(e) (3 marks) What would happen to b_j and the variance of b_j if all of the values $X_{1j}, \ldots, X_{nj}$ of the jth independent variable were replaced by numbers that were nearly constant?

The $(j+1)$th column of $\mathbf{X}$ would have strong correlation with the 1st column of $\mathbf{X}$ (which is all 1's) so the corresponding entry of $(\mathbf{X}'\mathbf{X})^{-1}$ would be large so both b_j and $Var(b_j)$ would be large.

4. Data are available for 67 construction crews on the number of lost days of work due to injury over a period of one year. We are interested in understanding whether the number of lost days per person (variable name: `lostdays_pp`, the average number of lost days per person on the crew) is related to the size of the work crew (variable name: `size`, the number of people on the crew) and the experience of the foreman in charge of the crew (variable name: `f_exp`, measured in years).

   Some output from SAS is given below.

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

               Number of Observations Read          67
               Number of Observations Used          67

                       Analysis of Variance
                                  Sum of           Mean
Source                   DF      Squares         Square    F Value    Pr > F
Model                     2    215.73625      107.86812      11.37    <.0001
Error                    64    607.32882        9.48951
Corrected Total          66    823.06507

             Root MSE              3.08051    R-Square     0.2621
             Dependent Mean        3.54168    Adj R-Sq     0.2391
             Coeff Var            86.97865

                       Parameter Estimates
                             Parameter       Standard
    Variable       DF        Estimate          Error    t Value    Pr > |t|
    Intercept       1         0.20769        1.28855       0.16      0.8725
    size            1         0.71201        0.14981       4.75      <.0001
    f_exp           1        -0.26702        0.08234      -3.24      0.0019
```

   (a) (2 marks) Explain the meaning of the coefficient of `size` for a non-statistician who is putting together construction crews for future jobs.

   For foremen with the same experience, increasing size of crew by 1 worker increases the number of lost days by 0.712 on average.

   (b) (3 marks) What are the null and alternative hypotheses of the analysis of variance F test? What do you conclude? Relate your answer to lost days for construction crews.

   $H_0 : \beta_1 = \beta_2 = 0$ *versus* H_a : *at least one of* β_1, β_2 *not zero*
   $p < 0.0001$ *so strong evidence that both aren't 0*
   At least one of foreman experience and size of crew help explain lost days in a staatistically significant way.

(c) (2 marks) In addition to the model in the SAS output above, a model was fit including an additional term which is the product of `size` and `f_exp` (variable name: `fexpsize`). What is the purpose of including this additional term?

It allows investigation of whether or not the effect of size of crew on days lost varies with foreman's experience (or vice versa).

(d) (2 marks) Output from SAS for the model in part (c) is given below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	241.74285	80.58095	8.73	<.0001
Error	63	581.32221	9.22734		
Corrected Total	66	823.06507			

Root MSE	3.03765	R-Square	0.2937
Dependent Mean	3.54168	Adj R-Sq	0.2601
Coeff Var	85.76871		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.75326	3.53921	1.63	0.1090
size	1	0.14136	0.37063	0.38	0.7042
f_exp	1	-0.62886	0.23032	-2.73	0.0082
fexpsize	1	0.03466	0.02064	1.68	0.0981

Based on this model, does size of construction crew have an effect on the number of lost days per person? Explain.

Yes. There is weak evidence of an interaction indicating that the effect of foreman experience on lost days depends on size.

(e) (2 marks) Which of the two models fit to these data do you prefer? Justify your answer with appropriate statistics.

The second model. Adjusted R^2 is higher (or MSE is lower).

5. An experimenter wished to compare three different drug products (labelled A, B, and C) for combatting a virus. Four different dosages (0.2, 0.4, 0.8, and 1.0 μg) of each of the drugs were compared. Each of the 12 treatment combinations (3 drug products times 4 dosages) were applied to a culture of the virus and the rates of reduction in the number of cells of the virus were recorded.

   Some output from SAS is below.

```
                              The REG Procedure
                          Dependent Variable: rate
                    Number of Observations Read          12
                    Number of Observations Used          12

                            Analysis of Variance
                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F
Model                     5       129.19267       25.83853      12.00    0.0044
Error                     6        12.91400        2.15233
Corrected Total          11       142.10667

             Root MSE              1.46708    R-Square     0.9091
             Dependent Mean        7.16667    Adj R-Sq     0.8334
             Coeff Var            20.47093

                            Parameter Estimates
                    Parameter      Standard
Variable     DF      Estimate         Error    t Value    Pr > |t|      Type I SS
Intercept     1       2.11000       1.57327       1.34      0.2284      616.33333
dose          1       3.65000       2.31966       1.57      0.1667       61.06133
dose_drugA    1       7.55000       3.28050       2.30      0.0610       44.00014
dose_drugB    1       2.90000       3.28050       0.88      0.4107       23.00000
drugA         1       0.72000       2.22494       0.32      0.7572        0.00419
drugB         1       1.61000       2.22494       0.72      0.4965        1.12700
```

   (a) (4 marks) Write the model that was fit in the SAS output above, defining all variables.

$$Y = \beta_0 + \beta_1 \texttt{dose} + \beta_2 \texttt{dose\_drugA} + \beta_3 \texttt{dose\_drugB} + \beta_4 \texttt{drugA} + \beta_5 \texttt{drugB} + \epsilon$$

   where
   Y = *rate of reduction in cell growth*
   `dose` = *dosage (continuous)*
   `drugA` = *1 if drug A and 0 otherwise*
   `drugB` = *1 if drug B and 0 otherwise*
   `dose_drugA` *and* `dose_drugB` *are products of* `dose` *and* `drugA` / `drug B`

(b) (1 mark) What is the estimated relationship between `rate` and `dose` for drug C?

$$\hat{Y} = 2.11 + 3.65\texttt{dose}$$

(c) (2 marks) Explain why it would seem reasonable to assume that the three linear models for the relationships between `rate` and `dose` for the three drugs have a common intercept. Show how to change the regression model from part (a) to reflect this.

Effect of a dose of 0 shouldn't depend on drug.

$$Y = \beta_0 + \beta_1\texttt{dose} + \beta_2\texttt{dose\_drugA} + \beta_3\texttt{dose\_drugB} + \epsilon$$

(d) (4 marks) Carry out an appropriate statistical test to test the assumption of equal intercepts.

$H_0 : \beta_4 = \beta_5 = 0$ *versus* $H_a :$ *at least one of* β_4, β_5 *non-zero*
Test statistic: $\frac{(0.00419+1.127)/2}{12.914/6} = 0.262$ *has* $F_{2,6}$ *distribution under* H_0.
$p > 0.5$ *so no evidence that the intercepts are different.*

(e) (3 marks) Can you also test whether the relationship between `rate` and `dose` is the same for all three drugs from the given output? If yes, briefly explain how (although you do not need to actually carry out the test). If not, explain what additional information you need.

Want to test:
$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ *versus* $H_a :$ *at least one non-zero*
Do another partial F*-test like the above except the extra SS is for 4 variables in the numerator.*
OK to do with given Type I SS since dose is given first.

6. (10 marks, 2 for each part) Sketch an example of a residual plot that would result from a regression for each of the following situations. Indicate what you are plotting on your axes.

   (a) The data has one large influential outlier.

   Plot of residuals versus predicted values with a point far out by itself in the horizontal direction.

   (b) The data has one large non-influential outlier.

   Plot of residuals versus predicted values with a point with a large outlier near the horizontal centre of the data.

   (c) A log transformation of Y is appropriate.

   Plot of residuals versus predicted values that shows curvature and increasing variance.

   (d) The distribution of the residuals is right-skewed.

   Plot of ordered residuals versus normal quantiles that is curved.

   (e) A simple linear regression of Y on X was carried out but a model with both a X term and an X^2 term is appropriate.

   Plot of residuals versus X that shows non-monotone curvature.

7. For each of the following questions a short answer is required.

(a) (2 marks) Is adjusted R^2 always less than R^2? Explain.

Yes. $R^2 = 1 - SSE/SSTO$ *and Adj* $R^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SSTO}$
$\frac{n-1}{n-k-1}$ *is always greater than 1 (unless there are no predictor variables in the model).*

(b) (2 marks) In a simple linear regression, suppose the goal is to get a good estimate of the slope. What is the advantage of increasing the standard deviation of the X's?

S_{XX} *increases so* $Var(b_1) = \sigma^2/S_{XX}$ *decreases*

(c) (2 marks) A linear regression is carried out and the plot of Y versus the predicted values is increasing. What should be done? Why?

Nothing. We expect a relationship.

(d) (3 marks) Explain the difference between a prediction interval and a calibration interval and how you would decide which to use.

A prediction interval is used for an estimated value of Y *at a given value of* X.
A calibration interval is used for an estimated value of X *corresponding to a given value of* Y.
You always regress Y *on* X *with dependent / independent variable based on which variable depends on the other for the context, and then use a prediction or calibration interval depending on what you are estimating.*

(e) (2 marks) In examining a normal quantile plot of residuals, why are the values on the extreme left and right often of more interest than the values in the centre?

For quantiles in confidence intervals and for p-values we need values that are in the tails of the distribution.

Simple regression formulae

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

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

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

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

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

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

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

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

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

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

$$\hat{X}_h \pm \frac{t_{n-2,1-\alpha/2}}{|b_1|} * \text{appropriate s.e.}$$
(valid approximation if $\frac{t^2s^2}{b_1^2\sum(X_i-\overline{X})^2}$ is small)

Working-Hotelling coefficient:
$$W = \sqrt{2F_{2,n-2;1-\alpha}}$$

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

Regression in matrix terms

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

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

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

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

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

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

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

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

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

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

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

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

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

$$C_p = \frac{SSE_p}{MSE_P} - (n-2p)$$

$$PRESS_p = \sum(Y_i - \hat{Y}_{i(i)})^2$$

Total pages 17
Total marks 95