UNIVERSITY OF TORONTO SCARBOROUGH
Department of Computer and Mathematical Sciences
Sample Exam
Note: This is one of our past exams, In fact the only past exam with R. Before that we were using SAS. In almost every year, I change the material a little bit and so some of the questions or some parts of questions(a very few) are from material that we haven't discussed in this year.

STAC67H3 Regression Analysis

Duration: 3 hours

Last Name:________________________ First Name: ________________________

Student number: ________________________

Aids allowed:

- The textbook: Applied Regression Analysis by Kutner et al published by McGraw Hill
- Class notes
- A calculator (No phone calculators are allowed)

No other aids are allowed. For example you are not allowed to have any other textbook or past exams.

All your work must be presented clearly in order to get credit. Answer alone (even though correct) will only qualify for **ZERO** credit. Please show your work in the space provided; you may use the back of the pages, if necessary but you MUST remain organized.

t and F tables are attached at the end.

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

There are 32 pages including this page and statistical tables. Please check to see you have all the pages.

Good luck!!

Question:	1	2	3	4	5	6	7	8	9	Total
Points:	5	5	15	10	12	24	5	13	11	100
Score:										

1. (5 points) Suppose we wish to fit the model $y_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) + \varepsilon_i$ for a given data set with one dependent and one independent variables. Find the least squares estimates of β_0^* and β_1^*. How do they relate to b_0 and b_1, the least squares estimates for the SLR model we discussed in class. i.e. express your least squares estimates of β_0^* and β_1^* in terms of b_0, b_1 and $\bar{x}$. State clearly the quantity you minimize to obtain the least squares estimates and show your work clearly.

Solution: $y_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) + \varepsilon_i$. We minimize $Q = \sum_{i=1}^n (y_i - \beta_0^* - \beta_1^*(x_i - \bar{x}))^2$.
$\left.\frac{\partial Q}{\partial \beta_0^*}\right|_{\beta_0^*=b_0^*} = 0 \implies b_0^* = \bar{y} = \bar{y} - b_1\bar{x} + b_1\bar{x} = b_0 + b_1\bar{x}$.
$\frac{\partial Q}{\partial \beta_1^*} = -\sum_{i=1}^n (x_i - \bar{x})(y_i - \beta_0^* - \beta_1^*(x_i - \bar{x}))$

$$\begin{aligned}
\left.\frac{\partial Q}{\partial \beta_1^*}\right|_{\beta_1^*=b_1^*} = 0 \quad &\implies \quad \sum_{i=1}^n (x_i - \bar{x})y_i = b_1^* \sum_{i=1}^n (x_i - \bar{x})^2 \\
&\implies \quad b_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})y_1}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_1 - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = b_1 \quad \square
\end{aligned}$$

2. (5 points) A linear regression was run on a set of data with using an intercept and one independent variable. A part of the R output used in this regression analysis is given below:

```
> data=read.table("C:/Users/Mahinda/Desktop/slr.txt", header=1)
> fit <- lm(y ~ x, data=data)
> summary(fit)

Call:
lm(formula = y ~ x, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7990    35.1542   0.137    0.898
x             0.5947     0.4301   1.383    0.239

Residual standard error: 8.615 on 4 degrees of freedom
```

Complete the analysis of variance table using the given results.
Note:Your analysis of variance table should include SSE, $SSReg$, the degrees of freedom for each SS and the F- value. You don't need to calculate the p-value and you don't have to read F table.

Solution: $MSE = 8.615^2 = 74.22$ with $df = 4$ and so $SSE = 4 \times 74.22 = 296.88$ $F = \frac{MSReg}{MSE} = \frac{MSReg}{74.22} = t^2 = 1.383^2 = 1.91 \implies MSR = 1.91 \times 74.22 = 141.95$ with $df = 1$ and so, $SSREG = 141.95$

Here is an R output (Check ANOVA table)

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
x          1 141.93 141.931  1.9122 0.2389
Residuals  4 296.90  74.225
```

3. Researchers studied the relationship between the purity of oxygen produced in a chemical distillation process (y), and the percentage of hydrocarbons that are present(x) in the main condenser of the distillation unit. The purity is measured by an index (a percentage). Some useful R outputs from this study are given below:

```
> purity=read.table("C:/Users/Mahinda/Desktop/purity.txt", header=1)
> mean(purity$y)
[1] 92.1605
> fit <- lm(y ~ x, data=purity)
> coefficients(fit)
(Intercept)           x
   74.28331    14.94748
> anova(fit)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 152.13 152.127  128.86 1.227e-09 ***
Residuals 18  21.25   1.181

> x0=data.frame(x=1.5)
> predict(fit, x0, interval="confidence", level=0.95)
       fit      lwr     upr
1 96.70453 95.72077 97.6883
```

(a) (5 points) Calculate a 95% confidence interval for β_1, i.e. the regression coefficient of x. Show your work clearly.

> **Solution:** t-value for x is $\sqrt{128.86} = 11.35$ and so $s_{b_1} = 14.94748/11.35 = 1.317$.
> You can also get s_{b_1} from $s_{b_1} = \frac{s}{\sqrt{SS_{XX}}}$ where $s = \sqrt{MSE} = \sqrt{1.181} = 1.086738239$ and $SS_{XX} = \frac{SSR}{b_1^2} = \frac{152.13}{14.94748^2} = 0.6808930531$ and $s_{b_1} = \frac{1.086738239}{\sqrt{0.6808930531}} = 1.317$
> The CI for β_1 is $14.94748 \pm t_{18,0.975} \times s_{b_1} = 14.94748 \pm 2.101 \times 1.317$
> $= (12.18, 17.71)$ $\square$

(b) (5 points) Calculate 95% prediction interval for Y when $x = 1.5$. Show your work clearly.

> **Solution:** $\hat{y} = \frac{97.6883+95.72077}{2} = 96.704535$
>
> $$s_{\hat{Y}} = \frac{\frac{97.6883-95.72077}{2}}{t_{18,0.975}} = \frac{\frac{97.6883-95.72077}{2}}{2.101} = 0.468236554$$

Question 3 continues on the next page. . .

95% perdition interval is
$\hat{Y} \pm t_{18,0.975}\sqrt{MSE + s^2_{\hat{Y}}} = 96.704535 \pm 2.101 \times \sqrt{1.181 + 0.468236554^2}$
$= 96.704535 \pm 2.101 \times 1.183319682 = 96.704535 \pm 2.49 = ((94.22, 99.19)$.

Here is the R output

```
> predict(fit, x0, interval="prediction", level=0.95)
       fit      lwr      upr
1 96.70453 94.21886 99.19021
```

(c) (5 points) Calculate a 95% confidence interval for β_0, the y-intercept. Show your work clearly.

Solution: 95% CI for β_0 is given by $b_0 \pm t_{18,0.975} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{XX}}}$, $b_0 = \bar{y} - b_1\bar{x} \implies$
$\bar{x} = \frac{\bar{y}-b_0}{b_1} = \frac{92.1605-74.28331-}{14.94748} = 1.196$
$Reg = b_1^2 SS_{XX} \implies SS_{XX} = \frac{SSReg}{b_1^2} = \frac{152.13}{14.94748^2} = 0.6808930531$
$b_0 \pm t_{18,0.975} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{XX}}} = 74.28331 \pm 2.101 \times \sqrt{1.181}\sqrt{\frac{1}{20} + \frac{1.196^2}{0.6808930531}}$
$= 74.28331 \pm 2.101 \times 1.466551229 = 74.28331 \pm 3.35 = (70.935, 77.632)$
Here is the R output

```
> confint(fit, level=0.95)
               2.5 %   97.5 %
(Intercept) 70.93555 77.63108
x           12.18107 17.71389
```

You may continue your answer to question 3 on this page.

4. The following information (i.e. $(X'X)^{-1}$, $\mathbf{b}$, error sum of squares (SSE)) were obtained from a study of the relationship between plant dry weight (Y), measured in grams and two independent variables, percent soil organic matter (X_1) and kilograms of supplemental nitrogen per 1000 m^2 (X_2) based on a sample of n = 7 experimental fields . The regression model included an intercept.

$$(X'X)^{-1} = \begin{pmatrix} 1.7995972 & -0.0685472 & -0.2531648 \\ -0.0685472 & 0.0100774 & -0.0010661 \\ -0.2531648 & -0.0010661 & 0.0570789 \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} 51.5697 \\ 1.4974 \\ 6.7233 \end{pmatrix},$$
$SSE = 27.5808$.

(a) (5 points) Compute the Bonferroni confidence intervals for β_1 and β_2 using a joint confidence level 95%.

Solution: $s^2_{b_1} = MSE(X'X)^{-1}_{22} = \frac{27.5808}{7-3} \times 0.0100774 = 6.8952 \times 0.0100774 = 0.0694622448$ and $s_{b_1} = \sqrt{0.0694622448} = 0.26355691$
$s^2_{b_2} = MSE(X'X)^{-1}_{33} = \frac{27.5808}{7-3} \times 0.0570789 = 6.8952 \times 0.0570789 = 0.3935704313$ and $s_{b_2} = \sqrt{0.3935704313} = 0.6273519198$

(b) (5 points) Use a t-test to test the null hypothesis $H_0 : \beta_2 = 0.5\beta_1$ against the alternative $H_1 : \beta_2 > 0.5\beta_1$.

Solution: $s^2_{b_2-0.5b_1} = MSEc'(X'X)^{-1}c$ where $c' = \begin{pmatrix} 0 & -0.5 & 1 \end{pmatrix}$, $t = \frac{b_2-0.5b_1}{s_{b_2-0.5b_1}} \sim t_{df_{Error}} = t_{7-3}$.

Here is an R code with calculations

```
> #R code for testing a linear combination of betas
> c <- c(0, -0.5, 1)
> c
[1]  0.0 -0.5  1.0
> xpxinv <- matrix(c(1.7995972,  -0.0685472 , -0.2531648 , -0.0685472 ,
+ 0.0100774  ,-0.0010661, -0.2531648 , -0.0010661 , 0.0570789),  nrow=3,
+ ncol=3, byrow = T)
> xpxinv
           [,1]       [,2]       [,3]
[1,]  1.7995972 -0.0685472 -0.2531648
[2,] -0.0685472  0.0100774 -0.0010661
[3,] -0.2531648 -0.0010661  0.0570789
> MSE = 6.8952
> s_sq = MSE*t(c)%*%xpxinv%*%c
> s_sq
          [,1]
[1,] 0.4182928
```

```
> s = sqrt(s_sq)
> s
          [,1]
[1,] 0.6467556
> b1 = 1.4974
> b2 = 6.7233
> lc = -0.5*b1 + b2
> lc
[1] 5.9746
> t=lc/s
> t
         [,1]
[1,] 9.237802
```

Question 4 continues on the next page. . .

You may continue your answer to question 4 on this page.

5. You are given the following matrices computed for a regression analysis.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 9 & 136 & 269 & 260 \\ 136 & 2114 & 4176 & 3583 \\ 269 & 4176 & 8257 & 7104 \\ 260 & 3583 & 7104 & 12276 \end{pmatrix}, \ \mathbf{X}'\mathbf{Y} = \begin{pmatrix} 45 \\ 648 \\ 1,283 \\ 1,821 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 9.610932 & 0.0085878 & -0.2791475 & -0.0445217 \\ 0.0085878 & 0.5099641 & -0.2588636 & 0.0007765 \\ -0.2791475 & -0.2588636 & 0.1395 & 0.0007396 \\ -0.0445217 & 0.0007765 & 0.0007396 & 0.0003698 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} -1.163461 \\ 0.135270 \\ 0.019950 \\ 0.121954 \end{pmatrix}, \ \mathbf{Y}'\mathbf{Y} = 285$$

(a) (8 points) Use the preceding results to complete the analysis of variance table. Note:Your analysis of variance table should include SSE, $SSReg$, the degrees of freedom for each SS and the F- value. You don't need to calculate the p-value (and you don't have to read F table)

Solution: $SST = \sum_i y_i^2 - n\bar{y}^2 = 285 - 9 \times \left(\frac{45}{9}\right)^2 = 60$, $SSE = Y'(I-H)Y = Y'Y - b'X'Xb = Y'Y - b'X'Y$

Here is an R code with the calculations

```
> b <- c(-1.163461 , 0.135270 ,0.019950 , 0.121954 )
> b
[1] -1.163461  0.135270  0.019950  0.121954
> ypy = 285
> SST = 60
>
> xpy = c(45, 648, 1283, 1821)
> xpy
[1]   45  648 1283 1821
> SSE = ypy - t(b)%*%xpy
> SSE
         [,1]
[1,] 2.026701
> SSReg = SST - SSE
> SSReg
        [,1]
[1,] 57.9733
```

Question 5 continues on the next page. . .

```
> n = 9
> n
[1] 9
> p = 4
> p
[1] 4
> df_Reg = p-1
> MSReg = SSReg/df_Reg
> df_Reg
[1] 3
> MSReg
         [,1]
[1,] 19.32443
>
> df_Error = n-p
> MSE = SSE/df_Error
> df_Error
[1] 5
> MSE
          [,1]
[1,] 0.4053402
>
> F=MSReg/MSE
> F
        [,1]
[1,] 47.6746
```

(b) (4 points) Calculate a 95% confidence interval for β_1, the coefficient of X_1.

Solution: $s_{b_1^2} = MSE(\mathbf{X}'\mathbf{X})^{-1}_{22} = 0.4053402 \times 0.5099641 = 0.2067089503$
$s_{b_1} = \sqrt{0.2067089503} = 0.45465256$, $t_{0.025,5} = 2.571$ and the confidence interval for β_1 is $0.135270 \pm 2.571 \times 0.45465256 = 0.135270 \pm 1.168911732$
$= (-1.033641732, 1.304181732)$ $\square$

Question 5 continues on the next page. . .

You may continue your answer to question 5 on this page.

6. The R output shown below was obtained from a regression analysis of a dependent variable Y on four independent variables x_1, x_2, x_3 and x_4.

```
> data=read.table("C:/Users/Mahinda/Desktop/typesSS.txt", header=1)
> library(car)
> fit <- lm(Y ~ x1 + x2 + x3 + x4, data=data)
> summary(fit)

Call: lm(formula = Y ~ x1 + x2 + x3 + x4, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.0911    49.9367   1.644   0.1122
x1           -0.4758     0.2696  -1.765   0.0894 .
x2           -0.1073     0.1609  -0.667   0.5109
x3           -0.3443     0.4941  -0.697   0.4921
x4            1.7633     1.8158   0.971   0.3405
---

> anova(fit)
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x1         1   46.90   46.90  0.9847 0.330182
x2         1    0.06    0.06  0.0012 0.972236
x3         1  411.45  411.45  8.6387 0.006819 **
x4         1   44.91   44.91  0.9430 0.340451
Residuals 26 1238.35   47.63

> vif(fit)
      x1       x2       x3       x4
1.243544 1.131751 4.363584 3.997464

> Anova(lm(Y ~ x1 + x2 + x3 +x4, data=data), type="III")
Anova Table (Type III tests)

Response: Y
             Sum Sq Df F value  Pr(>F)
(Intercept)  128.71  1  2.7024 0.11224
x1           148.32  1  3.1140 0.08937 .
x2            21.17  1  0.4444 0.51087
x3            23.13  1  0.4857 0.49205
x4            44.91  1  0.9430 0.34045
Residuals   1238.35 26
```

Question 6 continues on the next page. . .

(a) (5 points) Calculate the value of the F-statistic for testing the null hypothesis $H_0: \beta_1 = \beta_2 = 0$ against the alternative $H_a:$ not all β_k(k=1, 2) equal to zero, in the regression model with **only two predictors**, x_1 and x_2 (i.e. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$)

Solution: Using Type I SS, $SSR(x_1, x_2) = 46.90 + 0.06 = 46.96$ and $MSR = \frac{46.96}{2} = 23.48$ and $MSE = \frac{1238.35+411.45+44.91}{31-3} = \frac{1694.71}{28} = 60.52535714$ and $F = \frac{23.48}{60.52535714} = 0.3879$

Here is an R output to check these calculations:

```
> fit12 <- lm(Y ~ x1 + x2, data=data)
> summary(fit12)

Call:
lm(formula = Y ~ x1 + x2, data = data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.369847  21.209690   3.035  0.00515 **
x1          -0.237883   0.280304  -0.849  0.40327
x2           0.005468   0.175395   0.031  0.97535
---
Residual standard error: 7.78 on 28 degrees of freedom
Multiple R-squared: 0.02696,    Adjusted R-squared: -0.04254
F-statistic: 0.3879 on 2 and 28 DF,  p-value: 0.682

> anova(fit12)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value Pr(>F)
x1         1   46.90  46.901  0.7749 0.3862
x2         1    0.06   0.059  0.0010 0.9754
Residuals 28 1694.72  60.526
```

(b) (5 points) Calculate the value of the F-statistic for testing the null hypothesis $H_0: \gamma_2 = \gamma_3 = \gamma_4 = 0$ against the alternative $H_a:$ not all γ_k(k=2, 3, 4) equal to zero, in the regression model with **only three predictors**, x_2, x_3 and x_4 (i.e. $Y_i = \gamma_0 + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \varepsilon_i'$)

Solution: $SST = 46.9 + .06 + 411.45 + 44.91 + 1238.35 = 1741.67$ (still using Type I SS)
$SSE = 1238.35 + 148.32 = 1386.67$ (Using Type III SS)
$SSReg = SST - SSE = 1741.67 - 1386.67 = 335$

Question 6 continues on the next page. . .

$F = \frac{335/3}{1386.67/(31-4)} = \frac{118.333}{51.358} = 2.304$

Here is an R output to check these calculations:

```
> fit234 <- lm(Y ~ x2 + x3 + x4, data=data)
> summary(fit234)

Call:
lm(formula = Y ~ x2 + x3 + x4, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.18416   43.52234   0.785    0.439
x2          -0.02099    0.15923  -0.132    0.896
x3          -0.06651    0.48631  -0.137    0.892
x4           2.27131    1.86166   1.220    0.233

Residual standard error: 7.166 on 27 degrees of freedom
Multiple R-squared: 0.2038,     Adjusted R-squared: 0.1154
F-statistic: 2.304 on 3 and 27 DF,  p-value: 0.09942

> anova(fit234)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value  Pr(>F)
x2         1    3.37   3.368  0.0656 0.79983
x3         1  275.19 275.195  5.3584 0.02847 *
x4         1   76.45  76.447  1.4885 0.23300
Residuals 27 1386.67  51.358
```

(c) (3 points) Calculate the coefficient of partial determination between Y and x_3 given x_1 and x_2. Interpret your result.

Solution: $R^2_{Y3|12} = \frac{SSR(x_3|x_1,x_2)}{SSE(x_1,x_2)} = \frac{411.45}{1238.35+411.45+44.91} = \frac{411.45}{1694.71} = 0.243$

Useful R outputs (Check $SSE(x_1, x_2)$):

```
> fit12 <- lm(Y ~ x1 + x2, data=data)

> anova(fit12)
Analysis of Variance Table

Response: Y
```

Question 6 continues on the next page...

```
          Df  Sum Sq Mean Sq F value Pr(>F)
x1         1   46.90  46.901  0.7749 0.3862
x2         1    0.06   0.059  0.0010 0.9754
Residuals 28 1694.72  60.526
```

(d) (3 points) Calculate the coefficient of partial determination between Y and x_3 given x_1, x_2 and x_4. Interpret your result.

Solution: $R^2_{Y3|124} = \frac{SSR(x_3|x_1,x_2,x_4)}{SSE(x_1,x_2,x_4)} = \frac{23.13}{1238.35+23.13} = \frac{23.13}{1261.48} = 0.018$

Useful R outputs (Check $SSE(x_1, x_2, x_4)$):

```
> fit124 <- lm(Y ~ x1 + x2 + x4, data=data)

> anova(fit124)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x1         1   46.90   46.90  1.0038 0.325277
x2         1    0.06    0.06  0.0013 0.971958
x4         1  433.23  433.23  9.2726 0.005143 **
Residuals 27 1261.48   46.72
```

(e) (2 points) Consider the initial model, i.e. the model for Y on x_1, x_2, x_3 and x_4. Does the R output indicate any evidence of multicollinearity? What particular value (or values) in the R output supports your answer?

Solution: All VIF values are less than 10 and so no indication of multicollinearity.

(f) (6 points) Perform an F-test to test whether there is a regression relation between x_4 and the remaining predictors i.e. x_1, x_2 and x_3. Test at $\alpha = 0.05$.

Solution: $VIF(x_4) = \frac{1}{1-R^2(x_4 \text{ on } x_1,x_2,x_3)}$. From R output $VIF(x_4) = 3.997464$ and so $1 - R^2(x_4 \text{ on } x_1, x_2, x_3) = \frac{1}{3.997464} = 0.2501586006$ and $R^2(x_4 \text{ on } x_1, x_2, x_3) = 0.7498413994$.
$F = \frac{R^2/(3)}{(1-R^2)/(31-4)} = 26.97717599$ and compare this with $F^3_{(31-4),0.05}$ $\square$.

Question 6 continues on the next page...

You may continue your answer to question 6 on this page.

Question 6 continues on the next page...

You may continue your answer to question 6 on this page.

7. (5 points) The R output shown below was obtained from an investigation of unusual observations in a regression analysis of a dependent variable Y on three independent variables x_1, x_2 and x_3.

```
> data=read.table("C:/Users/Mahinda/Desktop/outliers.txt", header=1)
> fit <- lm(Y ~ x1 + x2 + x3, data=data)
> X <- model.matrix(fit)
> data$hii=hat(X)
> data$cookD <- cooks.distance(fit)
> p <- 4
> n <- 20
> qf(0.5, p, n-p)
[1] 0.875787
> data
   Row   x1   x2   x3    Y        hii        cookD
1    1 19.5 43.1 29.1  5.0 0.34120920 1.328961e+00
2    2 24.7 49.8 28.2 22.8 0.15653638 2.708477e-02
3    3 30.7 51.9 37.0 18.7 0.44042770 9.293256e-02
4    4 29.8 54.3 31.1 20.1 0.11242972 2.627835e-02
5    5 19.1 42.2 30.9 12.9 0.36109984 4.534338e-02
6    6 25.6 53.9 23.7 21.7 0.13151364 4.101559e-03
7    7 31.4 58.5 27.6 27.1 0.19433721 3.766692e-03
8    8 27.9 52.1 30.6 25.4 0.16418081 4.374498e-02
9    9 22.1 49.9 23.2 21.3 0.19278940 9.851165e-03
10  10 25.5 53.5 24.8 19.3 0.24051819 2.433832e-02
11  11 31.1 56.6 30.0 25.4 0.13935816 9.027553e-04
12  12 30.4 56.7 28.3 27.2 0.10929380 8.404170e-03
13  13 18.7 46.5 23.0 11.7 0.21357666 8.256439e-02
14  14 19.7 44.2 28.6 17.8 0.18808377 1.034024e-01
15  15 14.6 42.7 21.3 12.8 0.34830629 1.062918e-02
16  16 29.5 54.4 30.1 23.9 0.11439069 8.554424e-07
17  17 27.7 55.3 25.7 22.6 0.12532943 2.500710e-03
18  18 30.2 58.6 24.6 25.4 0.22828343 3.298842e-02
19  19 22.7 48.2 27.1 25.0 0.13235798 1.381248e-01
20  20 25.2 51.0 27.5 21.1 0.06597771 2.996277e-04
```

Identify all unusual observations based on the methods we have discussed in class. Explain precisely how you identified them.

Solution: The 1st observation has Cook's distance greater then 0.875787 ($F(0.5, p, n-p)$ and so is an unusual observation.
For leverages, the critical value (by the rule of thumb) is $\frac{2p}{n} = \frac{2\times 4}{20} = 0.4$. Observation 3 has $h_{ii} > 0.4$ and so is a high leverage value and so unusual.

You may continue your answer to question 7 on this page.

8. The R output shown below was obtained from an investigation to select a suitable subset of variables from a collection of four variables x_1, x_2, x_3 and x_4 for a regression analysis.

```
> data=read.table("C:/Users/Mahinda/Desktop/stepwise.txt", header=1)
> fit <- lm(Y ~ x1+x2+x3+x4, data=data)
> anova(fit)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 2395.9  2395.9 142.620 1.480e-10 ***
x2         1 1807.0  1807.0 107.565 1.708e-09 ***
x3         1 4254.5  4254.5 253.259 8.045e-13 ***
x4         1  260.7   260.7  15.521   0.00081 ***
Residuals 20  336.0    16.8
---
> #Variable selection
> library(leaps)
> X <- model.matrix(fit)[,-1]
> Cp.leaps <- leaps(X, data$Y, method='Cp')
> Cp.leaps
$which
      1     2     3     4
1 FALSE FALSE  TRUE FALSE
1 FALSE FALSE FALSE  TRUE
1  TRUE FALSE FALSE FALSE
1 FALSE  TRUE FALSE FALSE
2  TRUE FALSE  TRUE FALSE
2 FALSE FALSE  TRUE  TRUE
2  TRUE FALSE FALSE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE  TRUE FALSE FALSE
3  TRUE FALSE  TRUE  TRUE
3  TRUE  TRUE  TRUE FALSE
3 FALSE  TRUE  TRUE  TRUE
3  TRUE  TRUE FALSE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5
```

Question 8 continues on the next page...

```
$Cp
 [1]  84.246496 110.597414 375.344689 384.832454  17.112978  47.153985
 [7]  80.565307  85.519650  97.797790 269.780029   3.727399  18.521465
[13]  48.231020  66.346500   5.000000
```

(a) (3 points) What subset of variables would you select based on Mallow's C_p method? Give reasons for your answer.

> **Solution:** The model with independent variables x_1, x_3 and x_4 have $p = 4$ and $C_p = 3.727399 \approx p$ and so is a reasonable modle based on Mallow's C_p method.

(b) (5 points) Calculate the value of $R^2_{Adjusted}$ for the simple linear regression model for Y on x_1 only.

> **Solution:** $SSTot = 2395.9 + 1807.0 + 4254.5 + 260.7 + 336 = 9054.1$
> The value of C_p for this model is 375.344689. and $SSR(X_1) = 2395.9$ and $SSE(X_1) = SST - SSR(X_1) = 9054.1 - 2395.9 = 6658.2$ and so $R^2_{Adj} = 1 - \frac{MSE}{SSTot/(n-1)} = 1 - \frac{6658.2/(25-2)}{9054.1/(25-1)} = 0.2326$ $\square$
>
> Here is an R output (Check $R^2_{Adjusted}$)

```
> fit <- lm(Y ~ x1, data=data)
> summary(fit)

Call:
lm(formula = Y ~ x1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-42.391 -11.670   0.531  11.842  27.407

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.3216    18.0099   2.294  0.03123 *
x1            0.4922     0.1711   2.877  0.00852 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 17.01 on 23 degrees of freedom
Multiple R-squared: 0.2646,     Adjusted R-squared: 0.2326
F-statistic: 8.276 on 1 and 23 DF,  p-value: 0.008517
```

(c) (5 points) Consider the two models:
Model 1: Y on x_1, x_2, x_3, x_4

Question 8 continues on the next page...

Model 2: Y on x_1, x_3, x_4
Which of these two models is the better model according to the Akaike's information criterion (AIC)? Support your answer with appropriate calculations.

Solution: For Model 1, $SSE = 336$ and $AIC = n \ln SSE_p - n \ln n + 2p = 25 \ln(336) - 25 \ln 25 + 2 \times 5 = 74.96$

For Model 2,
$C_p = \frac{SSE_p}{MSE_P} - (n - 2p) = \frac{SSE}{16.8} - (25 - 2 \times 4) = 3.727399$ (from the R output) $\Longrightarrow$ $SSE = 348.2203$ and $AIC = n \ln SSE_p - n \ln n + 2p = 25 \ln(348.2203) - 25 \ln 25 + 2 \times 4 = 73.85.$
Model 2 has smaller AIC and so is the better model according this method.

Here is an R output (Check AIC's)

```
> null=lm(Y~1, data=data)
> full=lm(Y~., data=data)
> step(full, scope=list(lower=null, upper=full), direction="both")
Start:  AIC=74.95
Y ~ x1 + x2 + x3 + x4

       Df Sum of Sq     RSS     AIC
- x2    1     12.22  348.20  73.847
<none>               335.98  74.954
- x4    1    260.74  596.72  87.314
- x1    1    759.83 1095.81 102.509
- x3    1   1064.15 1400.13 108.636
```

Question 8 continues on the next page. . .

You may continue your answer to question 8 on this page.

9. (a) (5 points) The data matrix X in a regression model is given by

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} \mathbf{x_{(1)}} & \mathbf{x_{(2)}} & \mathbf{x_{(3)}} \end{pmatrix}$$

where $\mathbf{x_{(j)}} = \begin{pmatrix} x_{1j} & x_{2j} & \dots & x_{nj} \end{pmatrix}'$ denotes the j^{th} column of the data matrix X. Let X^* be the data matrix obtained by multiplying the third column of X by a constant k ($k \neq 0$). i.e. $X^* = \begin{pmatrix} \mathbf{x}^*_{\mathbf{(1)}} & \mathbf{x}^*_{\mathbf{(2)}} & \mathbf{x}^*_{\mathbf{(3)}} \end{pmatrix}$ where $\mathbf{x}^*_{\mathbf{(1)}} = \mathbf{x_{(1)}}$, $\mathbf{x}^*_{\mathbf{(2)}} = \mathbf{x_{(2)}}$ and $\mathbf{x}^*_{\mathbf{(3)}} = k\mathbf{x_{(3)}}$. Prove that X and X^* have the same hat matrix. Be precise and give reasons for all your steps.

Solution:

$$X^* = X \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & k \end{pmatrix} = XA$$

A is symmetric and invertible.
$H_{X^*} = X^*((X^*)'X^*)^{-1}(X^*)' = XAA^{-1}(X'X)^{-1}A^{-1}AX = X(X'X)^{-1}X' = H_X$.

(b) (6 points) The R output shown below was obtained from a regression study of a dependent variable Y and three independent variables x_1, x_2 and x_3. The R code generates another variable $newx_3 = 2 \times x_3$.

```
> data=read.table("C:/Users/Mahinda/Desktop/trans.txt", header=1)
> data$newx3=data$x3*2
> data
    Y x1 x2  x3 newx3
1  48 50 51 2.3   4.6
2  57 36 46 2.3   4.6
3  66 40 48 2.2   4.4
4  70 41 44 1.8   3.6
5  89 28 43 1.8   3.6
6  36 49 54 2.9   5.8
7  46 42 50 2.2   4.4
8  54 45 48 2.4   4.8
9  26 52 62 2.9   5.8
10 77 29 50 2.1   4.2
> fit <- lm(Y ~ x1+x2+x3, data=data)
> summary(fit)

Call:
lm(formula = Y ~ x1 + x2 + x3, data = data)
```

Question 9 continues on the next page. . .

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 175.5249    21.3345   8.227 0.000174 ***
x1           -1.1713     0.3885  -3.015 0.023556 *
x2           -0.5117     0.7986  -0.641 0.545374
x3          -19.6453    12.3606  -1.589 0.163083
---

> anova(fit)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 2626.73 2626.73 61.6380 0.0002258 ***
x2         1  296.83  296.83  6.9654 0.0385813 *
x3         1  107.65  107.65  2.5260 0.1630829
Residuals  6  255.69   42.62
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> newfit <- lm(Y ~ x1+x2+newx3, data=data)
> summary(newfit)

Call:
lm(formula = Y ~ x1 + x2 + newx3, data = data)

Coefficients:
            Estimate Std. Error
(Intercept)        A          E
x1                 B          -
x2                 C          -
newx3              D          F
---
```

In the R output some of the values have been deleted and some values have been replaced by letters A, B, C, D, E and F. Give the values of A, B, C, D, E and F.

Solution:

```
> newfit <- lm(Y ~ x1+x2+newx3, data=data)
> summary(newfit)

Call:
lm(formula = Y ~ x1 + x2 + newx3, data = data)

Residuals:
```

Question 9 continues on the next page...

```
     Min       1Q   Median       3Q      Max
-11.5263   0.1525   2.3012   2.7879   5.1077

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 175.5249    21.3345   8.227 0.000174 ***
x1           -1.1713     0.3885  -3.015 0.023556 *
x2           -0.5117     0.7986  -0.641 0.545374
newx3        -9.8226     6.1803  -1.589 0.163083
```

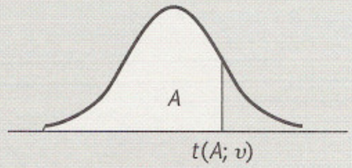
Question 9 continues on the next page. . .

You may continue your answer to question 9 on this page.

END OF EXAM

TABLE B.2
Percentiles of the *t* Distribution.

Entry is $t(A;\nu)$ where $P\{t(\nu) \leq t(A;\nu)\} = A$



	A						
ν	.60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

TABLE B.4 (*continued*) **Percentiles of the *F* Distribution.**

Den. df	*A*	Numerator df 1	2	3	4	5	6	7	8	9
8	.50	0.499	0.757	0.860	0.915	0.948	0.971	0.988	1.00	1.01
	.90	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.95	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.975	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.99	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.995	14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34
	.999	25.4	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.8
9	.50	0.494	0.749	0.852	0.906	0.939	0.962	0.978	0.990	1.00
	.90	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.95	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.975	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.99	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.995	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54
	.999	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	10.1
10	.50	0.490	0.743	0.845	0.899	0.932	0.954	0.971	0.983	0.992
	.90	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.95	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.975	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.99	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.995	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
	.999	21.0	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.96
12	.50	0.484	0.735	0.835	0.888	0.921	0.943	0.959	0.972	0.981
	.90	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.95	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.975	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.99	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.995	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
	.999	18.6	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.48
15	.50	0.478	0.726	0.826	0.878	0.911	0.933	0.949	0.960	0.970
	.90	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.95	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.975	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.99	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.995	10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54
	.999	16.6	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.26
20	.50	0.472	0.718	0.816	0.868	0.900	0.922	0.938	0.950	0.959
	.90	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.95	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.975	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.99	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.995	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
	.999	14.8	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
24	.50	0.469	0.714	0.812	0.863	0.895	0.917	0.932	0.944	0.953
	.90	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.95	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.975	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.99	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.995	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
	.999	14.0	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80

TABLE B.4 (*continued*) **Percentiles of the F Distribution.**

Den. df	A	Numerator df 1	2	3	4	5	6	7	8	9
30	.50	0.466	0.709	0.807	0.858	0.890	0.912	0.927	0.939	0.948
	.90	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.95	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.975	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.99	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.995	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45
	.999	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
60	.50	0.461	0.701	0.798	0.849	0.880	0.901	0.917	0.928	0.937
	.90	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.95	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.975	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.99	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.995	8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01
	.999	12.0	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
120	.50	0.458	0.697	0.793	0.844	0.875	0.896	0.912	0.923	0.932
	.90	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	.95	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
	.975	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
	.99	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	.995	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81
	.999	11.4	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38
∞	.50	0.455	0.693	0.789	0.839	0.870	0.891	0.907	0.918	0.927
	.90	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63
	.95	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88
	.975	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11
	.99	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41
	.995	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62
	.999	10.8	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10