UNIVERSITY OF NEW SOUTH WALES
SCHOOL OF MATHEMATICS AND STATISTICS

MATH2831/2931 Linear Models
Assignment Two

Note: This is a group assignment and is due on Friday's lecture in Week 8.

Names (Print): ____________________________________

I (We) declare that this assessment item is my (our) own work, except where acknowledged, and has not been submitted for academic credit elsewhere, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

- Reproduce this assessment item and provide a copy to another member of the University; and/or,
- Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

I (We) certify that I (We) have read and understood the University Rules in respect of Student Academic Misconduct.

Signed: ____________________________________

Date: ______________

Please follow the instructions below for completing the assignment (worth 10%).

1. You can work in groups of up to 3 people and submit a single (typeset) assignment (with the names of these 3 people). All members of the group will get the same marks.

2. You can split the work amongst the members of the group in any way you like, but free-riders should note that all of the assignment questions are examinable and potentially on the final exam. Thus, it is in everybody's interest to understand the assignment questions and be able to solve them.

3. Individuals who do not like working in groups are still free to submit an individual assignment.

4. To properly typeset the assignment, work with the package `http://www.latex-project.org/`. There is a large amount of information available on Latex at this URL including the installation guides for the compiler (miktex) and the editors (free of charge). You can use either Lyx `http://www.lyx.org/` or Texnic `http://www.texniccenter.org/`. Both are free latex editors that you can easily install and get started with. Once you have completed installation, you may use the latex template provided in the Assignment folder on Moodle.

Use the R-output given to answer questions 1,2,3.

1. The data is partly taken from the *Energy efficiency Data Set* on the *UCI Machine Learning Repository*[1] . The study performed energy analysis using 12 different building shapes. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. The response in this data set is the `Cooling_Load` and `Heating_Load`.

   To study the energy required in cooling the building, we fit the following multiple linear regression model using the predictors `Relative_Compactness`, `Surface_Area` and `Wall_Area`.

```
> summary.lm(cooling)

Call:
lm(formula = Cooling_Load ~ Relative_Compactness + Surface_Area +
    Wall_Area, data = energyMod)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6007  -3.1655  -0.7148   2.7230  11.4541

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.440e+03  1.835e+02  -7.847 4.37e-14 ***
Relative_Compactness  8.184e+02  1.039e+02   7.873 3.65e-14 ***
Surface_Area          1.358e+00  1.692e-01   8.028 1.25e-14 ***
Wall_Area            -1.017e-01  1.965e-02  -5.179 3.63e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.427 on 380 degrees of freedom
Multiple R-squared:  0.3508,Adjusted R-squared:  0.3457
F-statistic: 68.44 on 3 and 380 DF,  p-value: < 2.2e-16
```

   (a) State the value of the F statistic used to test the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$ versus $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$. State the conclusion of the test under a 1% level.

   (b) Construct an 95% confidence interval for the parameter β_3 associated with `Wall_Area`.

   (c) How many observations are there in this data set?

[1] URL: http://archive.ics.uci.edu/ml/datasets/Energy+efficiency

2. The ANOVA table of the fitted model is given below

```
> anova(cooling)
Analysis of Variance Table

Response: Cooling_Load
                      Df Sum Sq Mean Sq F value    Pr(>F)
Relative_Compactness   1 2260.5 2260.52 115.358 < 2.2e-16 ***
Surface_Area           1 1237.5 1237.47  63.150  2.20e-14 ***
Wall_Area              1  525.5  525.55  26.819  3.63e-07 ***
Residuals            380 7446.4   19.60
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
```

(a) Compute from the above table, the value of the F statistic used question 1 (a).

(b) Conduct the appropriate sequential F test under a 5% level of significance to test whether a model containing all the predictors is preferred over a model with `Relative_Compactness` as the predictors.

3. Forward selection method was applied on to the dataset (In this case, we use also `Heating_Load` as an predictor).

```
null<-lm(Cooling_Load~1,data = energyMod)
> full<-lm(Cooling_Load~.,data = energyMod)
> step(null,scope= list(lower = null, upper = full),
direction ='forward',test = 'F', k = 0.001)


Start:  AIC=1304.39
Cooling_Load ~ 1

                        Df Sum of Sq     RSS     AIC   F value  Pr(>F)
+ Heating_Load           1    8828.8  2641.1  740.48 1276.9559 < 2e-16 ***
+ Wall_Area              1    2554.1  8915.9 1207.66  109.4286 < 2e-16 ***
+ Surface_Area           1    2476.5  8993.4 1210.99  105.1927 < 2e-16 ***
+ Relative_Compactness   1    2260.5  9209.4 1220.10   93.7649 < 2e-16 ***
+ Roof_Area              1      95.0 11374.9 1301.19    3.1913 0.07482 .
<none>                               11469.9 1304.39
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Step:  AIC=740.48
```

```
Cooling_Load ~ Heating_Load

                       Df Sum of Sq    RSS    AIC F value   Pr(>F)
+ Relative_Compactness  1    49.495 2591.6 733.21  7.2763 0.007297 **
+ Surface_Area          1    48.548 2592.6 733.35  7.1346 0.007885 **
+ Roof_Area             1    44.888 2596.2 733.89  6.5874 0.010652 *
+ Wall_Area             1     7.618 2633.5 739.37  1.1021 0.294469
<none>                              2641.1 740.48
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Step:  AIC=733.21
Cooling_Load ~ Heating_Load + Relative_Compactness

               Df Sum of Sq    RSS    AIC F value Pr(>F)
+ Roof_Area     1    9.3064 2582.3 731.83  1.3695 0.2426
+ Wall_Area     1    8.0475 2583.6 732.02  1.1836 0.2773
+ Surface_Area  1    0.5994 2591.0 733.12  0.0879 0.7670
<none>                      2591.6 733.21

Step:  AIC=731.83
Cooling_Load ~ Heating_Load + Relative_Compactness + Roof_Area

               Df Sum of Sq    RSS    AIC F value Pr(>F)
+ Surface_Area  1    16.422 2565.9 729.38  2.4256 0.1202
+ Wall_Area     1    16.422 2565.9 729.38  2.4256 0.1202
<none>                      2582.3 731.83

Step:  AIC=729.38
Cooling_Load ~ Heating_Load + Relative_Compactness + Roof_Area +
    Surface_Area

       Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>              2565.9 729.38

Call:
lm(formula = Cooling_Load ~ Heating_Load + Relative_Compactness +
    Roof_Area + Surface_Area, data = energyMod)

Coefficients:
         (Intercept)          Heating_Load  Relative_Compactness
          -173.36976               0.76691             100.15325
           Roof_Area          Surface_Area
             0.04591               0.15281
```

(a) Perform the forward selection procedure using the sequential F-test as the selection rule. (i) What is the final model under a 5% level? (ii) What about under a 10% level?

(b) What is the sum of square regression for the model with all the predictors except `Wall_Area`.

(c) Conduct an appropriate hypothesis test to test under a 5% level, whether the model including only `Heating_Load` and `Relative_Compactness` is preferred over the model including `Heating_Load`, `Relative_Compactness`, `Roof_Area` and `Surface_Area`.

4. Observations (x_i, y_i) for $i = 1, \dots, n$ are made according to the model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where $x_i, \dots, x_n$ are fixed constants and ϵ_i are i.i.d $\mathcal{N}(0, \sigma^2)$. Suppose the model is then re-parametrized as

$$y_i = \alpha' + \beta'(x_i - \bar{x}) + \epsilon_i$$

let $\hat{\alpha}$ and $\hat{\beta}$ denotes the MLEs of α and β, respectively, and $\hat{\alpha}'$ and $\hat{\beta}'$ denote the MLEs of α' and β', respectively

(a) Show that $\hat{\beta}' = \hat{\beta}$

(b) Show that $\hat{\alpha}' \neq \hat{\alpha}$

(c) Show that $\hat{\alpha}'$ and $\hat{\beta}'$ are uncorrelated.

5. If X is the design matrix, show that

$$SS_{reg} = y^T(X(X^TX)^{-1}X^T - X_2(X_2^TX_2)^{-1}X_2^T)y.$$

where X_2 is a $n \times 1$-matrix with entries equal to one.

(Hint: if $\hat{y}$ denotes the $n \times 1$ vector of fitted values, use the fact that

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{y}_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

and show that

$$\sum_{i=1}^{n}\hat{y}_i^2 = \hat{y}^T\hat{y} = y^TX(X^TX)^{-1}X^Ty.$$

6. **(MATH2931 only)** The **Sherman-Morrison** formula, a special case of *matrix blockwise inversion*, states that for a matrix A and vectors u and v of appropriate size

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}, \quad v^TA^{-1}u \neq -1 .$$

Prove the Sherman-Morrison formula and verify the formula with a numerical example in which A is a 2×2 matrix.

7. **(MATH2931 only)** Let $H = X(X^TX)^{-1}X^T$ be the hat matrix corresponding to the $n \times p$ design matrix with predictors

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n-1,1} & x_{n-1,2} & \cdots & x_{n-1,p-1} \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p-1} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

where $x_n = (1, x_{n,1}, \ldots, x_{n,p-1})^T$. Assuming a full rank linear model, we can write the fitted values as $\hat{y} = Hy$. Let $\hat{y}_{-i}$ be the fitted values for a model fitted by omitting the i-th observation y_i

(a) Prove that for the i-th PRESS residual $y_i - \hat{y}_{i,-i}$, we have

$$y_i - \hat{y}_{i,-i} = \frac{y_i - \hat{y}_i}{1 - H_{ii}} .$$

(b) In no more than three sentences, explain the significance of the PRESS statistic $\sum_{i=1}^n (y_i - \hat{y}_{i,-i})^2$ and the identity above.

8. **(MATH2931 only)** Assume that $\beta_0, \ldots, \beta_k = 0$, derive the distribution of $\frac{SS_{reg}}{\sigma^2}$.