MATH2831 ASSIGNMENT 2

Question 1 - z5232937 - Mustafa Dohadwalla

a) i)

```
Subset selection object
Call: regsubsets.formula(y ~ ., data = squid, nvmax = 4)
5 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
x5     FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4  x5
1  ( 1 ) " " " " " " " " "*"
2  ( 1 ) " " " " " " "*" "*"
3  ( 1 ) " " "*" " " "*" "*"
4  ( 1 ) " " "*" "*" "*" "*"
```

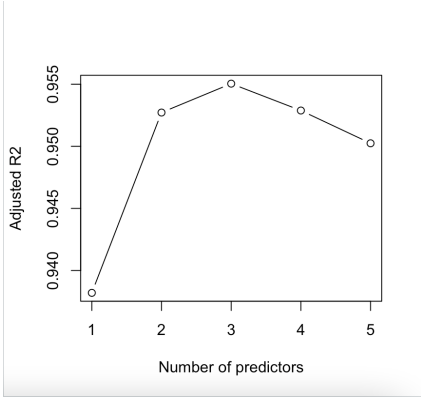
We can see that x5 is the preferred predictor for the chosen model of predictor size =1. We also can see that x1 is the worst predictor among the 5, not being in any model substets, with x3 the second worst. The best model obtained with 4 predictors contains x2, x3, x4 and x5.

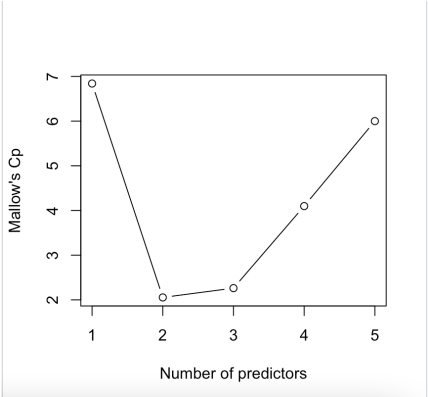
ii)

	x1	x2	x3	x4	x5	AdjR2	MallowCp	PRESS
1	0	0	0	0	1	0.9381995	6.841083	15.38393
2	0	0	0	1	1	0.9527197	2.054383	11.98158
3	0	1	0	1	1	0.9550458	2.262637	14.88133
4	0	1	1	1	1	0.9528821	4.098442	18.18457
5	1	1	1	1	1	0.9502433	6.000000	20.39306

The table above shows the adjusted R^2 and Cp values for the best subsets of each size.
Model with lowest adjusted R^2: Model 1, with x5 predictor
Model with lowest Cp value: Model 2, with x4 and x5 predictors
Model with lowest PRESS value: Model 2, with x4 and x5 predictors

Models with the lowest PRESS/Cp/Adjusted R^2 values are better models. Looking at the data it seems the model with 2 predictors (x4 and x5) is the best performing model of the best subset models.





b) i)

```
Start:  AIC=52.26
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x5    1    203.35  12.717 -8.058
+ x1    1    198.01  18.053 -0.351
+ x4    1    191.79  24.281  6.171
+ x3    1    189.72  26.349  7.969
+ x2    1    189.60  26.465  8.065
<none>               216.067 52.260

Step:  AIC=-8.06
y ~ x5

       Df Sum of Sq     RSS      AIC
+ x4    1    3.4744  9.2428 -13.0783
+ x1    1    2.4447 10.2725 -10.7547
<none>               12.7172  -8.0579
+ x2    1    0.7846 11.9326  -7.4589
+ x3    1    0.0071 12.7101  -6.0702

Step:  AIC=-13.08
y ~ x5 + x4

       Df Sum of Sq    RSS     AIC
+ x2    1   0.91727 8.3255 -13.378
<none>              9.2428 -13.078
+ x3    1   0.49055 8.7523 -12.278
+ x1    1   0.00038 9.2424 -11.079

Step:  AIC=-13.38
y ~ x5 + x4 + x2

       Df Sum of Sq    RSS     AIC
<none>              8.3255 -13.378
+ x3    1  0.084058 8.2415 -11.601
+ x1    1  0.041419 8.2841 -11.488

  Call:
  lm(formula = y ~ x5 + x4 + x2, data = squid)

  Coefficients:
  (Intercept)           x5           x4           x2
       -6.032       16.775        7.609       -3.208
```

Using forward model selection with the stepAIC() function on the model with just an intercept at the start. A lower AIC compared to another AIC of models on the same dataset indicates the model with a lower AIC is better.

At the start the AIC without any predictors is 52.56. Forward selection then chooses the model with just the x5 predictor (AIC = -8.06), as it has a lower AIC than the model with just an intercept, and the lowest AIC among models with just one predictor. This becomes the new current model.

Forward selection then chooses the model with x4 and x5 (AIC = -13.08). This model has a lower AIC compared to the model with just the x5 predictor, and has the lowest AIC compared to all other models of it's subset size. Thus, this is the new current model.

Forward selection then chooses the model with x2, x4 and x5 (AIC = -13.38). It follows the same technique described above and sets this model as the current model. It then checks if any of the models with 4 predictors has a better AIC than the current model. This is not the case, so the current model remains and becomes the final model. Coefficient estimates are shown above.

ii)

```
Start:  AIC=-9.74
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq     RSS      AIC
- x1    1    0.0504  8.2415 -11.6010
- x3    1    0.0930  8.2841 -11.4875
- x2    1    0.5456  8.7367 -10.3173
<none>               8.1911  -9.7359
- x4    1    1.7195  9.9105  -7.5438
- x5    1    7.5738 15.7649   2.6683

Step:  AIC=-11.6
y ~ x2 + x3 + x4 + x5

       Df Sum of Sq     RSS      AIC
- x3    1    0.0841  8.3255 -13.3777
- x2    1    0.5108  8.7523 -12.2781
<none>               8.2415 -11.6010
- x4    1    3.4287 11.6702  -5.9482
- x5    1    9.7049 17.9464   3.5196

Step:  AIC=-13.38
y ~ x2 + x4 + x5

       Df Sum of Sq     RSS      AIC
<none>               8.3255 -13.3777
- x2    1    0.9173  9.2428 -13.0783
- x4    1    3.6071 11.9326  -7.4589
- x5    1   14.1029 22.4284   6.4243

Call:
lm(formula = y ~ x2 + x4 + x5, data = squid)

Coefficients:
(Intercept)           x2           x4           x5
     -6.032       -3.208        7.609       16.775
```

Using backward model selection with the stepAIC() function, our initial model is the model with all the predictors (AIC = -9.74).

Backward selection then selects the model without the x1 predictor (AIC = -11.6), as this model has a lower AIC than the current model, and has the lowest AIC of all the models of subset size = 4. This model is now the current model.

Backward selection then selects the model without x3 (AIC = -13.38), for the same reasons as described above. This model is now the current model, with subset size = 3.

Backward selection then attempts to find a model with an AIC better than the current model with subset size = 2. None of the models for 2 predictors has a better AIC than the current model, so the current model with x2, x4 and x5 predictors is the final model. Coefficient estimates are shown in the report.

We do get the model from backward selection as well as forward selection.

The model with just x4, has AIC of 6.71 as seen in the report for i) in step 1 (choosing model with 1 predictor, but this value has no relevance without comparing it to other similar AIC values.

iii) Final model with stepwise selection:

```
Coefficients:
(Intercept)           x1           x5           x4           x2
    -6.1602       0.7551      16.2045       6.9658      -3.3389
```

y^ = -6.1602 + 0.7751*x1 + 16.02045*x5 + 0.9658*x4 -3.3389*x2

It's clear to see that the final model is different from the ones obtained by both forward and backward selection. The reason for this is because the starting model had the x1 predictor, which means some paths are not explored by this algorithm, which means it never finds a model where the AIC is lower than the final model. In this case it doesn't find any model with a better AIC without the x1 predictor, which means it's final model will differ from that of the backward/forward selection algorithms.

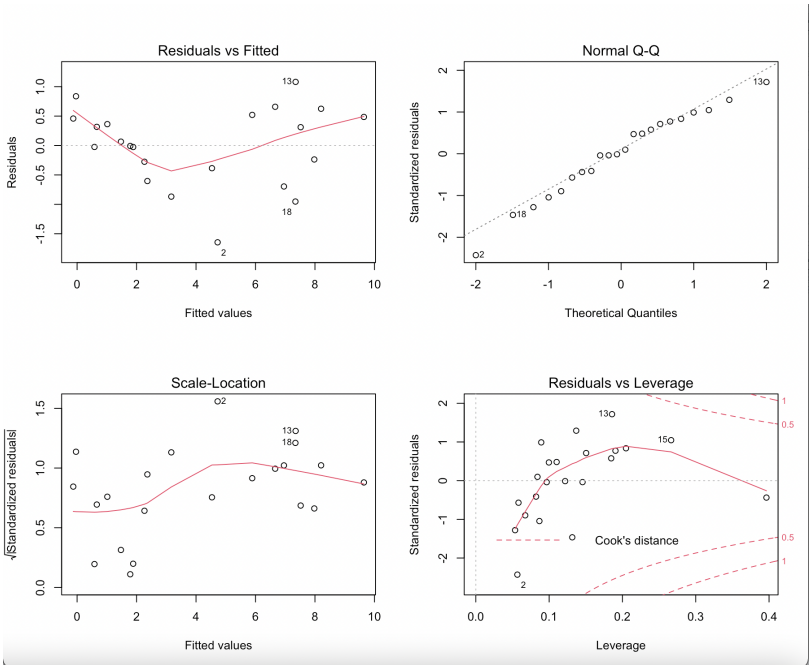
c) i)

```
Call:
lm(formula = y ~ x4 + x5, data = squid)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64530 -0.35808  0.02891  0.47855  1.08110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3951     0.6107 -10.472 2.49e-09 ***
x4            4.5769     1.7126   2.672   0.0151 *
x5           14.3951     2.5890   5.560 2.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6975 on 19 degrees of freedom
Multiple R-squared:  0.9572,    Adjusted R-squared:  0.9527
F-statistic: 212.6 on 2 and 19 DF,  p-value: 9.921e-14
```

The two models in contention are the forward/backward model from b) i) (let's call this Model 1 for this answer), and the best model from a) i) (Model 2 in that answer, let's also call this Model 2 here).

The optimal model I would recommend is Model 2, with the estimates as shown above. AIC forward/backward selection showed that Model 1 is best. As explained before, AIC stepwise does not select the best model possible, and thus the model with 4 predictors is not in contention for optimal model.

The AIC for Model 1 is -13.38, while the AIC for the optimal model is -13.08. So although the AIC for the optimal model is slightly worse, it is a better model in terms of PRESS, Cp and Adjusted R^2, with Model 2 having a significantly better PRESS statistic than Model 1.



We assume that the errors are independent and uncorrelated, which is based on the assumption that the sample taken was a random sample.
Looking at the residuals vs fitted plot, we can see that there is no fan shape, hence the variance of the errors is constant.
There is a U shape on the residual vs fitted plot, indicating the mean of the response is is not a linear combination of the predictors
Looking at the qq plot, we can see that the errors and hence the response is normally distributed, as there is no deviation from the straight line.

Looking at the Scale-Location plot and Residuals vs Leverage, we can see that observation 2 has a high standardized residual, thus an outlier, and also falls outside Cook's distance, indicating it's a point of influence.

ii)

```
Call:
lm(formula = y ~ x4 + x5, data = squid)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64530 -0.35808  0.02891  0.47855  1.08110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3951     0.6107 -10.472 2.49e-09 ***
x4            4.5769     1.7126   2.672   0.0151 *
x5           14.3951     2.5890   5.560 2.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6975 on 19 degrees of freedom
Multiple R-squared:  0.9572,    Adjusted R-squared:  0.9527
F-statistic: 212.6 on 2 and 19 DF,  p-value: 9.921e-14
```

The summary output for the model is shown above. Both the predictor variables are significant at an alpha of 0.05, according to the partial t tests, with the x5 predictor being very significant, with its p-value 2.31e^-05 <<< 0.001, which shows x5 significantly differs from 0 in the presence of x4, while x4's p-value is 0.0151 < 0.05, indicating it also differs from 0 in the presence of x5.