MATH2831 Linear Models

Assignment

Note: This assignment is due by 11:59pm Monday 16 November (week 10)

Please follow the instructions below for completing the assignment, it's worth 20% of your final mark.

- This assignment must be completed individually.
- Your assignment **must be submitted as a pdf file**. It may be typed or handwritten, then converted into **one** pdf file. You must include the completed coversheet in your assignment.
- You must sign and date your submitted assignment, and include your name and zID below.

I declare that this assessment item is my 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 certify that I have read and understood the University Rules in respect of Student Academic Misconduct.

Md Faiyam Islam z5258151

Student's full name and zID

Signed: F Islam Date: 16/11/20

MATH2831 – Assignment

1) I) a)

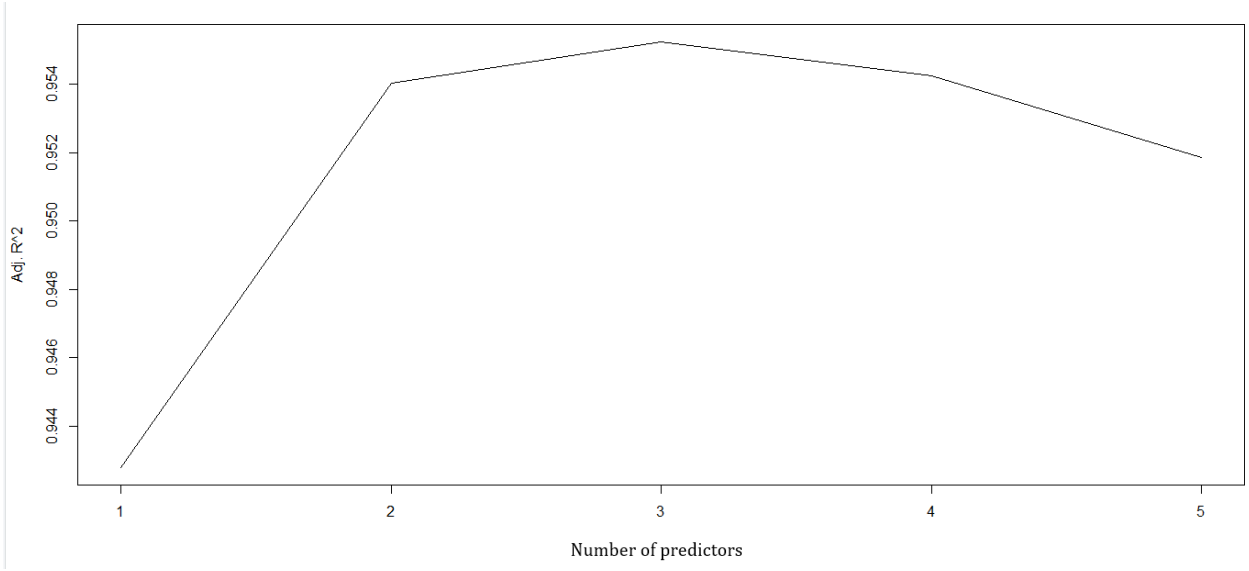
```
Subset selection object
Call: regsubsets.formula(y ~ ., data = squid)
5 Variables  (and intercept)
   Forced in Forced out
x1      FALSE      FALSE
x2      FALSE      FALSE
x3      FALSE      FALSE
x4      FALSE      FALSE
x5      FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: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

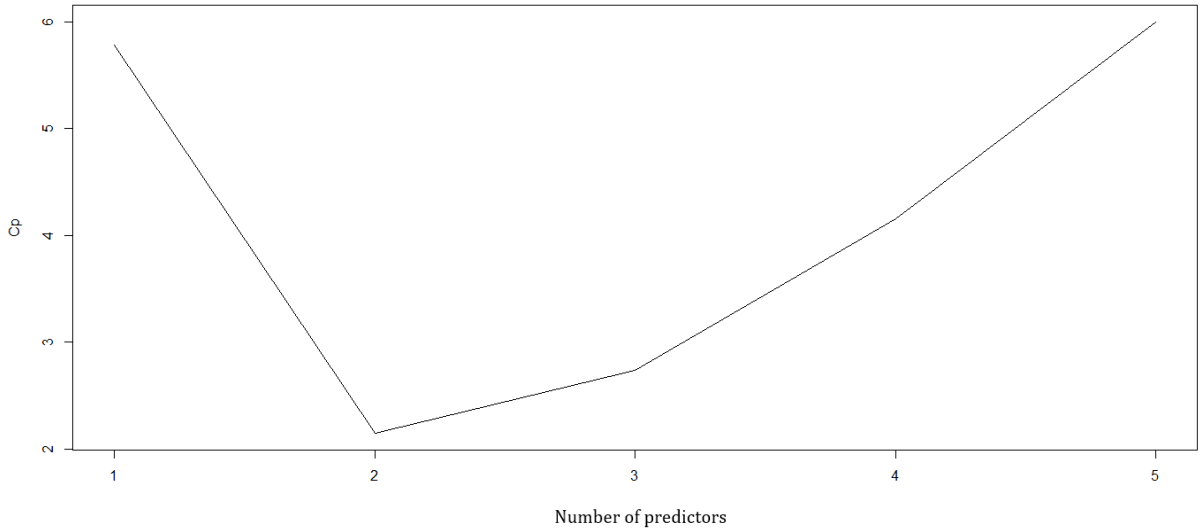
Using the regsubsets() function, the summary output shows 5 models. The model which includes all the predictors is obsolete because the model will contain a high variance in prediction. Other subsets include a model only containing x_5 which may be subjected to bias. The models containing x_5 and x_4 , as well as the model containing x_5 , x_4 and x_2 as predictors have better predictor performance however the best model in this case is the model containing the four predictors x_5 , x_4 x_2 and x_1, which contains the best balance between bias/variance trade off and goodness of fit/complexity trade off.

b)

x1	x2	x3	x4	x5	PRESS	AdjR2	Cp
0	0	0	0	1	14.1914	0.9428	5.7794
0	0	0	1	1	11.9225	0.9540	2.1456
0	1	0	1	1	15.0982	0.9552	2.7354
1	1	0	1	1	15.9452	0.9543	4.1582
1	1	1	1	1	19.4225	0.9512	6.0000

There is no overall winner, but based on PRESS statistic and Mallow's Cp we obtain the model containing the predictors x_4 (notch to wing length) and x_5 (width in inches) to be the best model because they contain the lowest PRESS (11.9225) and Cp (2.1456), but not the highest adjusted R^2 (0.9540). In terms of the adjusted R^2 the best model in terms of this statistic would be the model containing the predictors x_2(wing length in inches), x_4 and x_5 (0.9552).


Adj. R^2
0.944
0.946
0.948
0.950
0.952
0.954
1
2
3
4
5
Number of predictors


Cp
2
3
4
5
6
1
2
3
4
5
Number of predictors

II) a) i) This forward selection procedure begins with the null model which is y~1 (model containing only the intercept) and attempts to add each of the predictors to the scope and compute the AIC. If no predictors were added, then we start with an AIC of 52.25. In the next step, the predictor x5 was added since it contains the lowest AIC value from all the predictors, thus obtaining an AIC value of -9.77. Our current model becomes y~x5. Following the next step, the predictor x4 contains the lowest AIC value which makes sense for us to include this predictor. Thus, our current model becomes y~x5+x4. However, in the next step we encounter that x4 still remains the predictor with the lowest AIC value, thus we stop the forward selection procedure here.

```
Start:  AIC=52.25
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x5    1    204.16  11.767 -9.766
+ x1    1    199.15  16.779 -1.960
+ x3    1    197.35  18.573  0.275
+ x4    1    191.25  24.674  6.524
+ x2    1    190.28  25.645  7.373
<none>              215.925 52.246

Step:  AIC=-9.77
y ~ x5

       Df Sum of Sq     RSS      AIC
+ x4    1   2.78787  8.9793 -13.7148
+ x1    1   2.58453  9.1826 -13.2221
<none>              11.7671  -9.7661
+ x2    1   0.67408 11.0931  -9.0639
+ x3    1   0.46899 11.2981  -8.6609

Step:  AIC=-13.71
y ~ x5 + x4

       Df Sum of Sq    RSS     AIC
<none>              8.9793 -13.715
+ x2    1   0.69781 8.2814 -13.495
+ x1    1   0.16420 8.8151 -12.121
+ x3    1   0.02499 8.9543 -11.776
```

ii) Our final model including the coefficients becomes:

$$y = 15.016x5 + 4.1542x4$$

We get the same model. Essentially, we are obtaining a model with the same predictors chosen that is, x_4 and x_5 when we are looking at PRESS statistic and Cp. However, the difference is when we interpret adjusted R^2 and found in (I) that the model contains predictors x_2, x_4 and x_5. This is different to the model we obtain from forward selection. Also, the methods utilised to obtain the best possible model is different.

b) i) This backward elimination method starts with the full model which is y~x1+x2+x3+x4+x5 and attempts to remove each of the predictors then recalculate the AIC and obtain the best model. If all the predictors were kept then the AIC becomes -10.48. In the next step, the predictor x3 is removed because it contains the lowest AIC value and since the overall AIC drops to -12.27, our current model becomes y~x1+x2+x4+x5. In the following step, we remove the predictor x1, as it contains the lowest AIC of -13.495, making the overall AIC value even lower to -13.49 of the overall model, thus our current model becomes y~x2+x4+x5. Next, the predictor x2 is removed as it is still lower than the AIC of the null model (-10.48), thus we obtain an even lower AIC value of -13.7148 which brings us to our current model which is y~x4+x5. However, this is where the selection procedure stops because the AIC values of x4 is -9.7661 and x4 is 6.5236 which is both greater than the AIC of the null model. **The AIC value with the model containing x4 as the predictor is 6.52536 as seen from the output below.**

```
Start:  AIC=-10.48
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq     RSS      AIC
- x3    1    0.0783  7.9958 -12.2668
- x1    1    0.2987  8.2163 -11.6684
<none>               7.9175 -10.4832
- x2    1    0.8688  8.7863 -10.1927
- x4    1    0.9827  8.9002  -9.9093
- x5    1    4.3522 12.2697  -2.8460

Step:  AIC=-12.27
y ~ x1 + x2 + x4 + x5

       Df Sum of Sq     RSS     AIC
- x1    1    0.2856  8.2814 -13.495
<none>               7.9958 -12.267
- x2    1    0.8193  8.8151 -12.121
- x4    1    0.9869  8.9827 -11.706
- x5    1    8.6436 16.6394   1.856

Step:  AIC=-13.49
y ~ x2 + x4 + x5

       Df Sum of Sq     RSS      AIC
- x2    1    0.6978  8.9793 -13.7148
<none>               8.2814 -13.4946
- x4    1    2.8116 11.0931  -9.0639
- x5    1   14.1791 22.4606   6.4558

Step:  AIC=-13.71
y ~ x4 + x5

       Df Sum of Sq     RSS      AIC
<none>               8.9793 -13.7148
- x4    1    2.7879 11.7671  -9.7661
- x5    1   15.6948 24.6740   6.5236
```

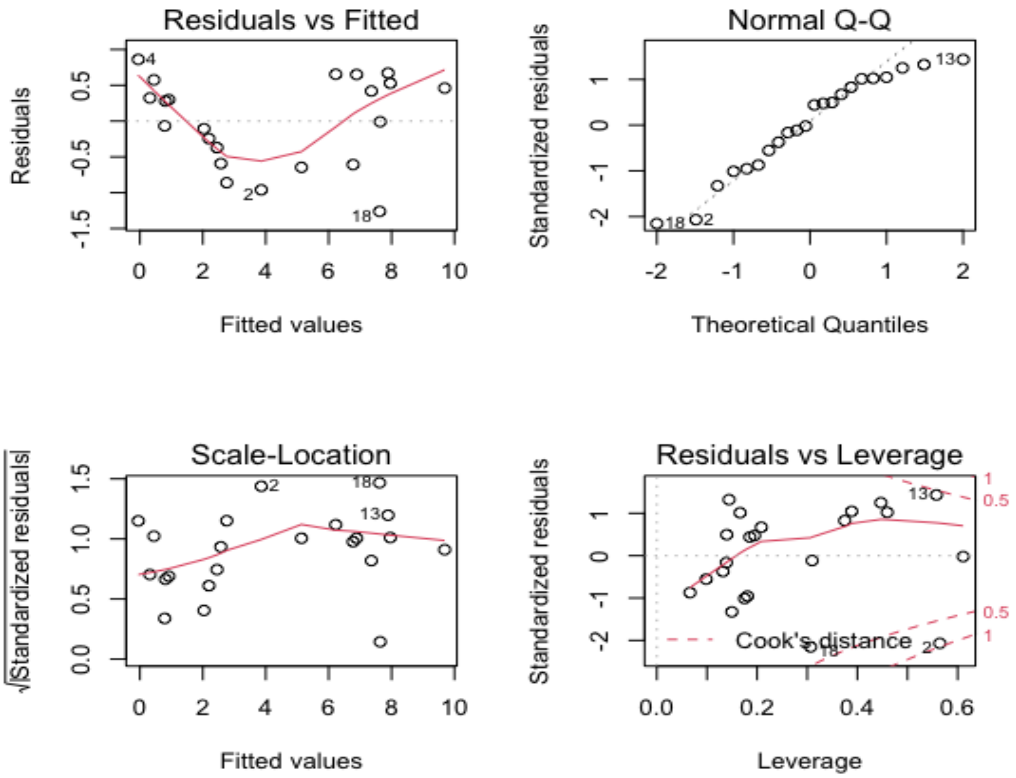
ii) Our final model including the coefficients becomes:

$y = 15.016x5 + 4.1542x4$. Yes, we obtain the same model using backward elimination as to forward selection.

c)

The final model is y~x1+x5, i.e. the model containing the predictors x_1 and x_5, whereas part a) and b), the final model is y~x4+x5, i.e. the model containing the predictors x_4 and x_5. The obvious differentiation from this model compared to a) and b) is that we start from the predictor x1 and choose the predictor which contains the lowest AIC value, which happens to be x5. However, unlike the methods used in a) and b), we do not pick x4 because we chose x1 as a starting point despite it containing the largest AIC value. In comparison to the methods in parts a) and b) we can see that the AIC value when using the stepwise method is larger, -13.22 compared to -13.71. This is largely because the predictor x1 contains a much larger AIC value compared to x4.

III) a)



To determine any violation, present on the model analysis, we need to check each assumption of the general linear model:

1. **The errors ϵ_i and thus the responses $y_i, i = 1, \dots, n$ are uncorrelated.**

The squid data provided only contains 22 observations which is a small sample size, thus further investigation is required. However, we shall assume that this assumption holds true for a simple random sample that has been taken for the given observations.

2. **The mean of the response y_i is a linear combination of the predictors $x_{i1}, \dots, x_{ik}$.**

From the plot of residual vs fitted values shows a prominent u shape, which indicates that the errors contain a non-zero mean. This leads towards a violation of assumption 2. This not only entails that the mean of the responses being dependent on a linear prediction of the predictors but also the errors, thus assumption 2 is violated.

3. **The variance of the errors ϵ_i and hence the variance of response y_i is constant, $i = 1, \dots, n$**

Interpreting the residuals vs fitted values plot we see that there is no apparent fan shape. The variation within the residuals along the red line seems to be constant, so this assumption of the general linear model is not violated.

4. **The errors ϵ_i and thus the responses y_i are normally distributed, $i = 1, \dots, n$.**

We can interpret the normal Q-Q plot and observe that there is a deviation in the upper right corner of the plot from the straight line. We are unable to verify this assumption, thus there is an immediate impact on inference when using this model. Further analysis of the standardised residuals vs fitted values plot shows that observations 2 and 18 are outliers which deviate the rest of the standardised residuals. To interpret these influential points, we look at the leverage vs standardised residuals plot and observe that observation 2 lies outside the 0.5 Cook's distance and rather close to 1. This indicates that there is high influence of this point. Observations 2 and 18 are highly influential as they contain high standardised residuals compared to the rest of the data. This indicates that there is high influence of this point. However, when we interpret the residuals vs fitted values plot, the observations' residuals seem to be small. This is because the large influential points pull the line of best fit towards themselves, thus the residuals of the observations are decreased but the leverage is increased.

b)

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.56123 -0.49604  0.09069  0.45717  0.98057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3351     0.6009 -10.543 2.23e-09 ***
x4            4.1542     1.7104   2.429   0.0252 *
x5           15.0160     2.6057   5.763 1.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.6875 on 19 degrees of freedom
Multiple R-squared:  0.9584,    Adjusted R-squared:  0.954
F-statistic: 218.9 on 2 and 19 DF,  p-value: 7.584e-14
```

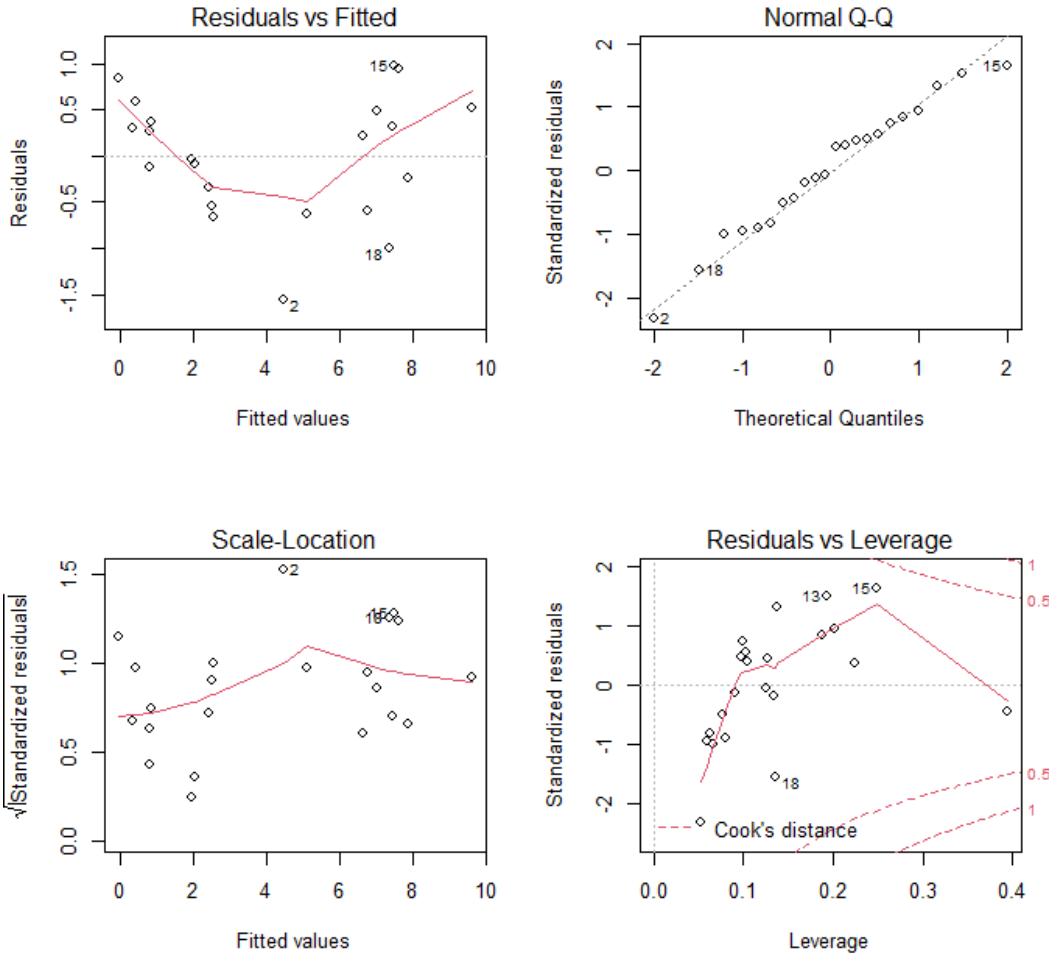
Our final optimal model is:

$$\hat{y} = -6.335 + 4.154x_4 + 15.016x_5$$

based on the findings from part (I) and (II). From previous calculations, specifically in (II) using forward and backward selection, the best model we obtain is the model with predictors x4 and x5. Although this is the only model with the lowest PRESS statistic and

Mallow's Cp, the deciding factor between the models was that this model contained the lowest AIC score. In terms of adjusted R-squared, this model has the second highest value of 0.9540375 using the regsubset function in (I). To conclude, by using methods to obtain these statistics we can say that this model is the best in terms of predicting new values which are mostly unbiased and contains lower variance of prediction compared to a model which contains more predictors (a more complicated model). From the summary output above, the predictor x4 contains a p-value of 0.0252 of the partial t-test. This suggests that the predictor x4 is not statistically significant.

Repeating a) for the model y~x4+x5:



<u>Residuals vs Fitted</u>

In the residuals vs fitted values plot, there is an immediate violation of the linearity assumptions, because there is a prominent U shape which highlights an incorrectly specified mean. If this model were to have the linearity assumption, the mean response y_i must be a linear combination of the predictors and hence the values should be distributed along the horizontal line in the plot. The homoskedasticity assumption in the general linear model suggests that the variance of the errors $e_i$and responses y_i is constant. There is a violation of the constancy of error variance because we can see a definite fan shape in the residuals.

Normal Q-Q

For the normality assumption to hold, the errors $e_i$and thus the responses $y_i$must be normally distributed. On a normal Q-Q plot, observations must lie approximately on a straight line if the data is to be normally distributed and thus for the assumption to hold. From the plot, we observe that there is an overall linear pattern with a slight deviation in observation 15. Therefore, the normality assumption holds and there is no violation despite observation 15 being slightly influential across the data set.

Scale-Location

Interpreting the scale location plot, we can check if the assumption for homoskedasticity holds or not. There is no noticeable trend or pattern in the residuals and the data points seem to be equally distributed. This means that there is a constancy of error variance and thus the homoskedasticity assumption holds true.

Residuals vs Leverage

From the residuals vs leverage plot, we see that from a), the highly influential observation 2 is no longer present and observation 18 is roughly within the bulk of the data. Although, this time observation 13 and 15 seem to be outliers but they do not contain high leverages compared to the observations in part a). Thus, there are no observations that have a substantial impact on the inferences of interest in this dataset.

2) a)

Q2) a) Show that $\sum_{i=1}^{n} \text{var}(\hat{y}_i) = \sigma^2 tr(H_1) = p\sigma^2$

$$\text{Var}(\hat{y}) = H_1 \text{Var}(y) H_1^T$$
$$= \sigma^2 H_1 H_1 = \sigma^2 H_1$$

From the calculation above, the diagonal entries of $\sigma^2 H_1$ is equal to the variance of the fitted value. Thus, showing that:

$$\sum_{i=1}^{n} \text{Var}(\hat{y}_i) = \sigma^2 Tr(H_1)$$
$$= \sigma^2 tr(X_1^T X_1 (X_1^T X_1)^{-1})$$
$$= \sigma^2 tr(I_{p \times p})$$
$$= p\sigma^2$$

Thus, $$\frac{\sum_{i=1}^{n} \text{Var}(\hat{y}_i)}{\sigma^2} = p$$

b)

Q2) b) we consider:

$$\hat{\sigma}^2 = \frac{y^T(I - X_1(X_1^TX_1)^{-1}X_1^T)y}{n-p} \qquad E(y) = X_1\beta_1 + X_2\beta_2$$

show that $E(\hat{\sigma}^2) = \sigma^2 + \frac{1}{n-p}\beta_2^TX_2^T(I-H_1)X_2\beta_2$

$$E(\hat{\sigma}^2) = \frac{1}{n-p}E(y^T(I - X_1(X_1^TX_1)^{-1}X_1^T)y) \qquad (\text{var}(y) = \sigma^2)$$

$$= \frac{1}{n-p}(\sigma^2 \text{tr}(I - X_1(X_1^TX_1)^{-1}X_1^T) + (X_1\beta_1 + X_2\beta_2)^T(I - X_1(X_1^TX_1)^{-1}X_1^T)(X_1\beta_1 + X_2\beta_2)$$

Note: ~~$(A+B)^T = A^T + B^T$~~

$$= \frac{1}{n-p}(\sigma^2\text{tr}(I) - \sigma^2\text{tr}(X_1(X_1^TX_1)^{-1}X_1^T) + ((X_1\beta_1)^T + (X_2\beta_2)^T)(I - X_1(X_1^TX_1)^{-1}X_1^T)(X_1\beta_1 + X_2\beta_2)$$

(we use $\text{tr}(XY) = \text{tr}(YX)$)

$$= \frac{1}{n-p}(\sigma^2(n - \text{tr}(X_1^TX_1(X_1^TX_1)^{-1}) + ((X_1\beta_1)^T + (X_2\beta_2)^T)(I-H_1)(X_1\beta_1 + X_2\beta_2))$$

$$= \frac{1}{n-p}(\sigma^2(n-p) + ((X_1\beta_1)^T + (X_2\beta_2)^T)(I-H_1)(X_1\beta_1 + X_2\beta_2))$$

note: $(XY)^T = Y^TX^T$

$$= \sigma^2 + \frac{1}{n-p}(\beta_1^TX_1^T + \beta_2^TX_2^T)(I-H_1)(X_1\beta_1 + X_2\beta_2)) \quad \text{note: } (I-H)X = 0$$

$$= \sigma^2 + \frac{1}{n-p}((\beta_1^TX_1^T + \beta_2^TX_2^T)((I-H_1)X_1\beta_1 + (I-H_1)X_2\beta_2)$$

$$= \sigma^2 + \frac{1}{n-p}((\beta_1^TX_1^T + \beta_2^TX_2^T)((I-H_1)X_2\beta_2)$$

$$= \sigma^2 + \frac{1}{n-p}(\beta_1^TX_1^T(I-H_1)X_2\beta_2 + \beta_2^TX_2^T(I-H_1)X_2\beta_2)$$

$$= \sigma^2 + \frac{1}{n-p}(\beta_1^TX_1^T(I - X_1(X_1^TX_1)^{-1}X_1^T)X_2\beta_2 + \beta_2^TX_2^T(I-H)X_2\beta_2)$$

$$= \sigma^2 + \frac{1}{n-p}(\beta_1^TX_1^TX_2\beta_2 - \beta_1^T\underbrace{X_1^TX_1(X_1^TX_1)^{-1}}_{I}X_1^TX_2\beta_2 + \beta_2^TX_2^T(I-H_1)X_2\beta_2)$$

$$= \sigma^2 + \frac{1}{n-p}(\beta_1^TX_1^TX_2\beta_2 - \beta_1^TX_1^TX_2\beta_2 + \beta_2^TX_2^T(I-H_1)X_2\beta_2)$$

$$= \sigma^2 + \frac{1}{n-p}(\beta_2^TX_2^T(I-H_1)X_2\beta_2)$$

$$\therefore \quad E(\hat{\sigma}^2) = \sigma^2 + \frac{1}{n-p}(\beta_2^TX_2^T(I-H_1)X_2\beta_2) \text{ as required}$$

c)

Q2) c) $\sum_{i=1}^{n} Bias^2(\hat{y}_i) = (E(y) - E(\hat{y}))^T (E(y) - E(\hat{y}))$

$= \beta_2^T X_2^T (1 - H_1) X_2 \beta_2$

Proof: $(E(y) - E(\hat{y}))^T (E(y) - E(\hat{y}))$

$= ((X_1\beta_1 + X_2\beta_2) - H_1(X_1\beta_1 + X_2\beta_2))^T ((X_1\beta_1 + X_2\beta_2) - H_1(X_1\beta_1 + X_2\beta_2))$

$= (X_1\beta_1 + X_2\beta_2 - X_1\underbrace{(X_1^TX_1)^{-1}X^TX_1}_{I}\beta_1 - H_1X_2\beta_2)^T (X_1\beta_1 + X_2\beta_2 - X_1\underbrace{(X_1^TX_1)^{-1}X^TX_1}_{I}\beta_1 - H_1X_2\beta_2)$

$= (X_2\beta_2 - H_1X_2\beta_2)^T (X_2\beta_2 - H_1X_2\beta_2)$ H_1 is symmetric

$= (\beta_2^TX_2^T - \beta_2^TX_2^TH_1)(X_2\beta_2 - H_1X_2\beta_2)$

$= \beta_2^TX_2^TX_2\beta_2 - \beta_2^TX_2^TH_1X_2\beta_2 - \beta_2^TX_2^TH_1X_2\beta_2 + \beta_2^TX_2^TH_1H_1X_2\beta_2$ since $H_1^2 = H_1$ (idempotent)

$= \beta_2^TX_2^TX_2\beta_2 - 2\beta_2^TX_2^TH_1X_2\beta_2 + \beta_2^TX_2^TH_1X_2\beta_2$

$= \beta_2^TX_2^TX_2\beta_2 - \beta_2^TX_2^TH_1X_2\beta_2$

$= \beta_2^TX_2^T(X_2\beta_2 - H_1X_2\beta_2)$

$= \beta_2^TX_2^T(1 - H_1)X_2\beta_2$ as required.

d)

Q2) d) If $\frac{(n-p)(\hat{\sigma}^2-\sigma^2)}{\sigma^2}$ is an unbiased estimator of $\sum_{i=1}^{n}\frac{Bias^2(\hat{y}_i)}{\sigma^2}$. Then, $E\left[\frac{(n-p)(\hat{\sigma}^2-\sigma^2)}{\sigma^2}\right] = \sum_{i=1}^{n}\frac{Bias^2(\hat{y}_i)}{\sigma^2}$

Proof: $L.H.S = E\left[\frac{(n-p)(\hat{\sigma}^2-\sigma^2)}{\sigma^2}\right]$

$$= \frac{(n-p)}{\sigma^2}E(\hat{\sigma}^2-\sigma^2)$$

$$E(\hat{\sigma}^2-\sigma^2) = E(\hat{\sigma}^2)-\sigma^2$$

$$= \sigma^2 + \frac{1}{n-p}\beta_2^T X_2^T (I-H_1)X_2\beta_2 - \sigma^2$$

Therefore, $L.H.S = \left(\frac{n-p}{\sigma^2}\right)\left[\frac{1}{n-p}\beta_2^T X_2^T (I-H_1)X_2\beta_2\right]$

$$= \frac{1}{\sigma^2}\beta_2^T X_2^T (I-H_1)X_2\beta_2$$

$R.H.S = \frac{\sum_{i=1}^{n} Bias^2(\hat{y}_i)}{\sigma^2} = \frac{1}{\sigma^2}\beta_2^T X_2^T (I-H_1)X_2\beta_2$

Therefore, $\frac{(n-p)(\hat{\sigma}^2-\sigma^2)}{\sigma^2}$ is an unbiased estimator of $\frac{\sum_{i=1}^{n} Bias^2(\hat{y}_i)}{\sigma^2}$.

$$\sum_{i=1}^{n}\frac{MSE(\hat{y}_i)}{\sigma^2} = \sum_{i=1}^{n}\frac{Var(\hat{y}_i)}{\sigma^2} + \sum_{i=1}^{n}\frac{Bias^2(\hat{y}_i)}{\sigma^2}$$

$$\frac{Var(\hat{y}_i)}{\sigma^2} = p$$

$$E\left(p + \frac{(n-p)(\hat{\sigma}^2-\sigma^2)}{\sigma^2}\right)$$

$$= p + E\left[\frac{(n-p)(\hat{\sigma}^2-\sigma^2)}{\sigma^2}\right]$$

$$= p + \frac{\sum_{i=1}^{n} Bias^2(\hat{y}_i)}{\sigma^2} = \sum_{i=1}^{n}\frac{MSE(\hat{y}_i)}{\sigma^2}$$

Therefore, $p + \frac{(n-p)(\hat{\sigma}^2 - \sigma^2)}{\sigma^2}$ is a sensible estimator of $\frac{\sum_{i=1}^{n} MSE(\hat{y}_i)}{\sigma^2}$