STAT 3A03 Applied Regression Analysis With SAS

Fall 2017

Assignment 3 Solution Set

Q. 1 **a)**

```
PROC REG Data=S3A3.Races plot=none;
    Model Time=Distance Climb;
    Plot Time*Pred.;
    Plot Student.*Pred.;
    Plot Student.*nqq.;
run;
```

The regression output is

The REG Procedure
Model: MODEL1
Dependent Variable: Time

Number of Observations Read	35
Number of Observations Used	35

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	281686567	140843283	181.66	<.0001
Error	32	24810082	775315		
Corrected Total	34	306496649			

Root MSE	880.51977	R-Square	0.9191
Dependent Mean	3472.57143	Adj R-Sq	0.9140
Coeff Var	25.35642		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-539.48291	258.16073	-2.09	0.0447
Distance	1	373.07268	36.06841	10.34	<.0001
Climb	1	0.66289	0.12305	5.39	<.0001

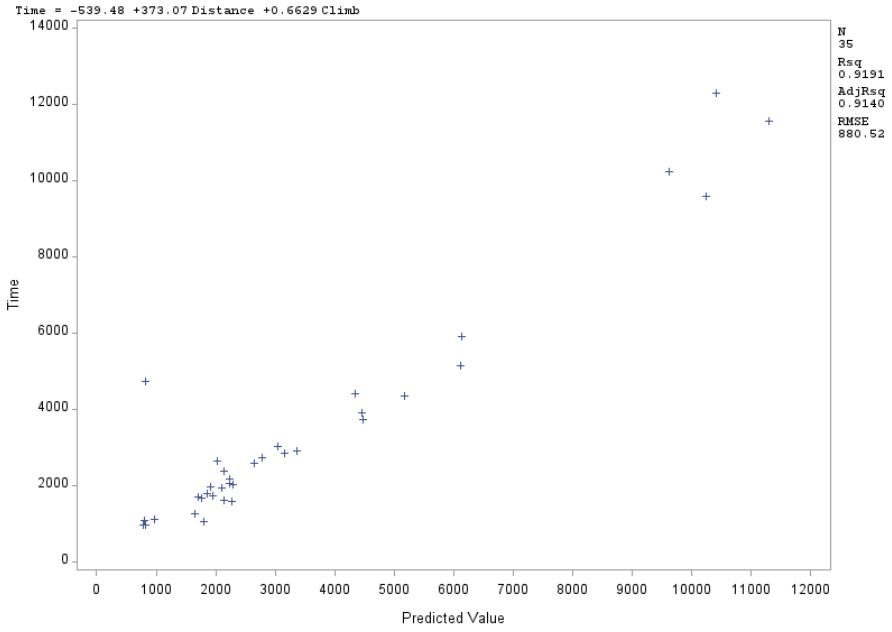
The fitted model is therefore

$$\widehat{\text{Time}} = -539.48 + 373.07 \times \text{Distance} + 0.663 \times \text{Climb}$$

The model is clearly significant since the p-value for the F test is very small and the R^2 value is almost 92%. [3 marks]

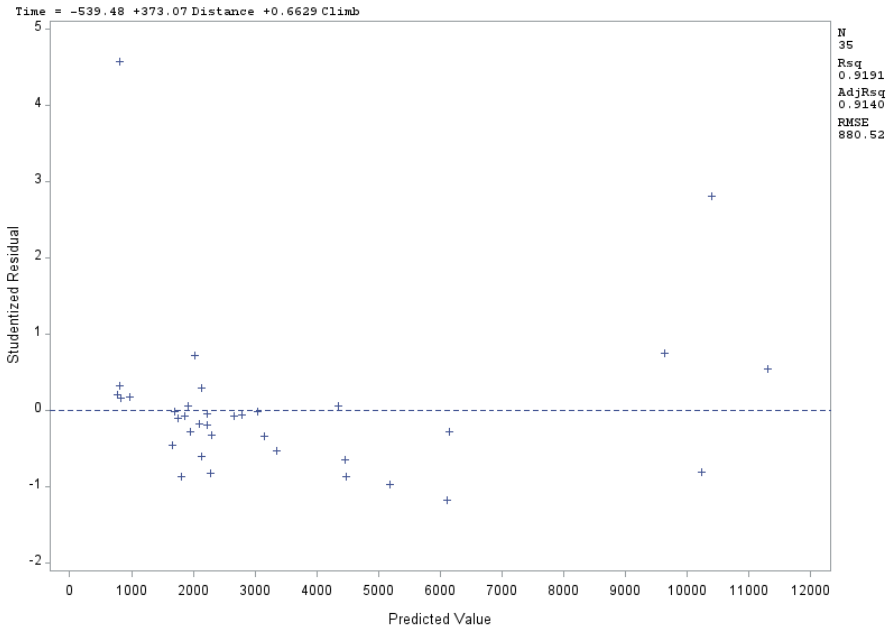
b) We use the usual the three model diagnostic plots from the code above to look for violations of the model assumptions.

The plot of fitted values against the observed record times is



From this plot there does not seem to be any great deviation from linearity although there does appear to be one clear outlier and there are a group of 4 races with longer observed times somewhat distant from the other races. These may be quite influential. [3 marks]

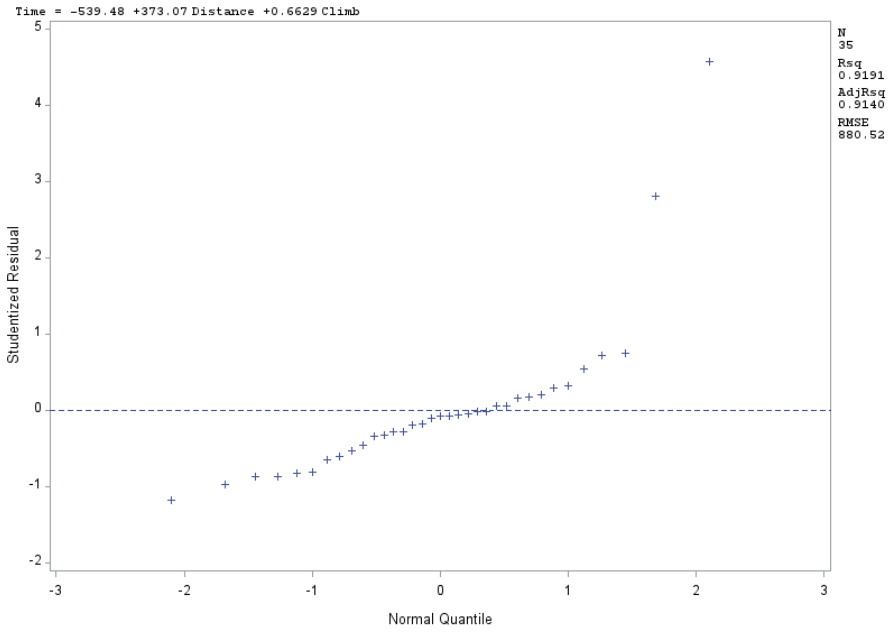
The plot of (internally) studentized residuals is



This plot does not show any marked pattern so the linearity assumption is likely okay. The

outlier spotted on the previous plot is quite evident here too. One possible cause for concern might be that the four races with longer record times also appear to show more variability in their residuals than the rest of the races. [3 marks]

Finally we look at the normal quantile-quantile plot of the studentized residuals.



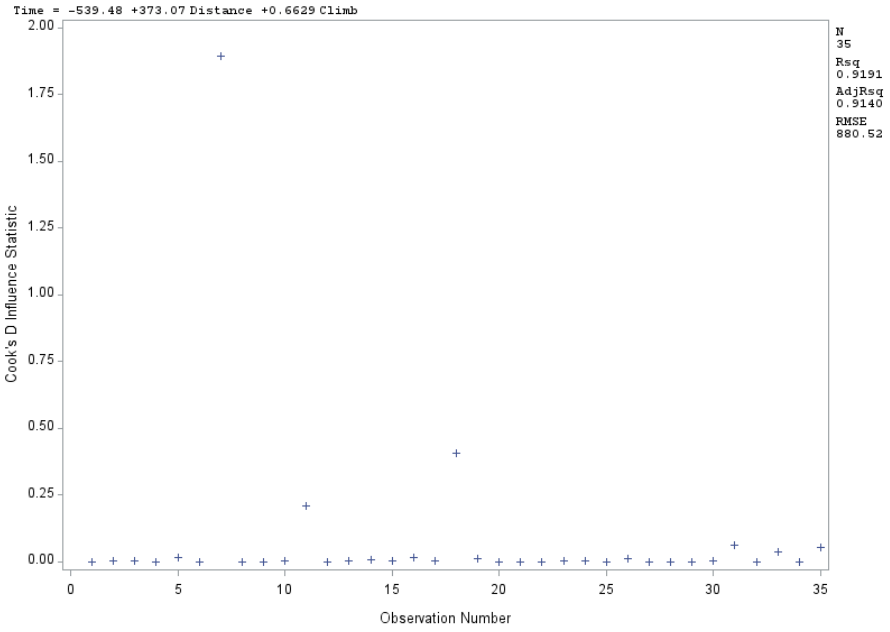
This plot looks somewhat non-linear but that is almost entirely because of 2 outliers in the upper tail. This suggests that the problem may be with individual points rather than a failure of the overall model assumption of normality. [3 marks]

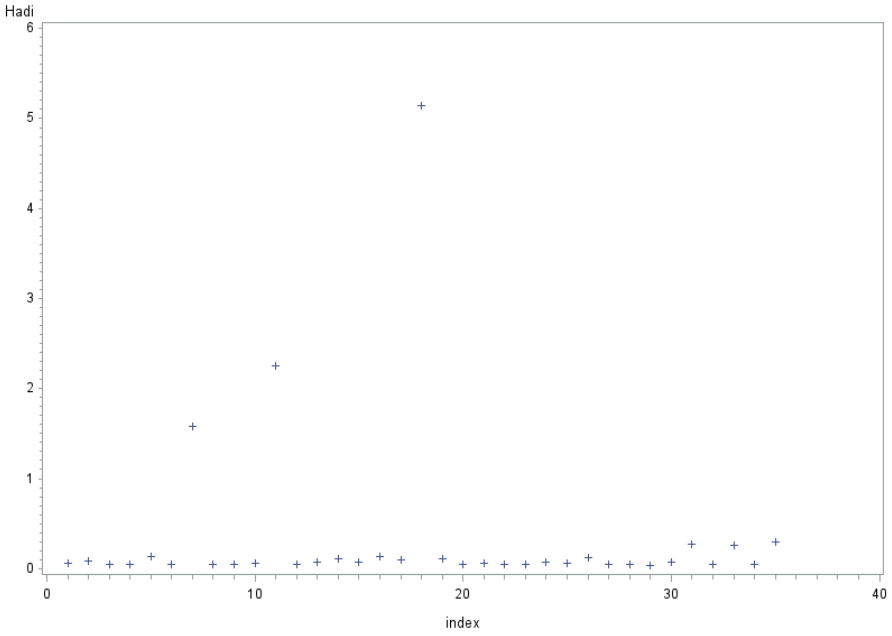
c) The code for the three plots required in this question is

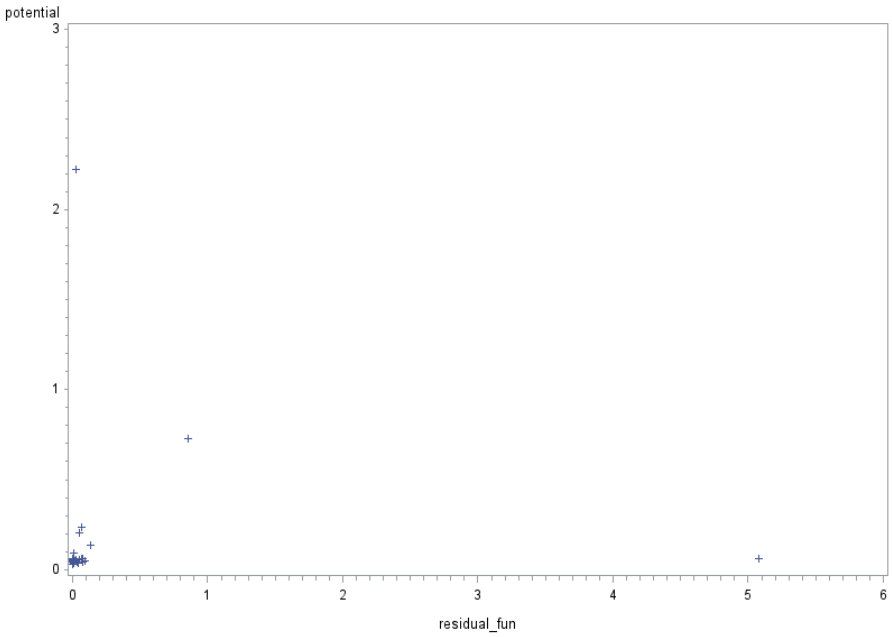
```
PROC REG Data=S3A3.Races plot=none noprint;
    Model Time=Distance Climb;
    Plot CookD.*Obs.;
    Output Out=Races_out
        Predicted=Fitted
        Residual=Res_raw
        H=Leverage;
run;

PROC REG Data=S3A3.Races plot=none;
    model Time=Distance Climb;
    ods output anova=races_anova; /* save the ANOVA table */
run;

Data races_anova;
    set races_anova;
    If source='Error' then call symput ('races.sse', ss);
    If source='Model' then call symput ('races.p', df);
run;

Data Races_out;
    set Races_out;
```

```
    index=_N_;
    d=Res_raw/sqrt(&races_sse);
    potential=Leverage/(1-Leverage);
    residual_fun=(&races_p+1)*d**2/((1-Leverage)*(1-d**2));
    Hadi=potential+residual_fun;
run;

PROC GPLOT Data=Races_out;
    Plot Hadi*index;
    Plot Potential*Residual_fun;
run;
```

Here are the three plots produced





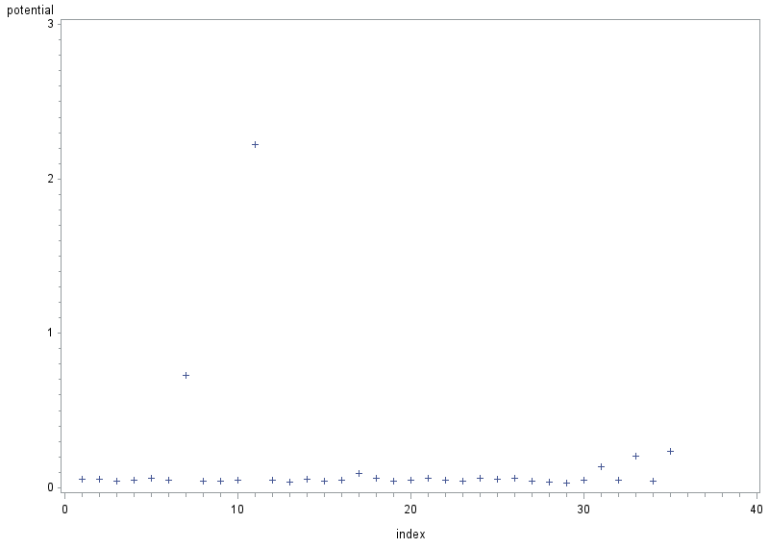


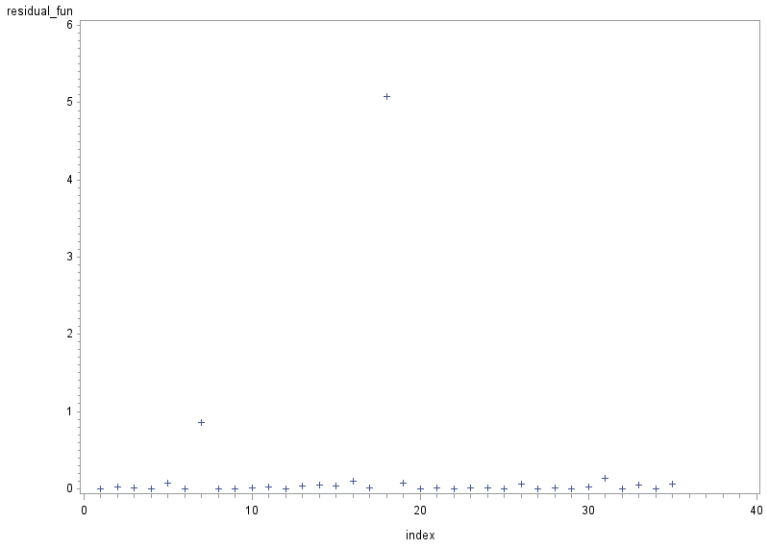
[6 marks]

From these plots we see that there are three points which are suspect. They are point 7 (Bens of Jura), point 11 (Lairig Ghru) and point 18 (Knock Hill). From the potential-residual plot we see that one of these is a point of very high leverage, one is an outlier (high residual) and one seems to be a combination of the two. To identify which is which, however we need plots of the residual function and potential function against the id.

```
PROC GPLOT Data=Races_out;
   Plot Potential*index;
   Plot Residual_fun*index;
run;
```

```
quit;
```





From these plots it is clear that point 11 (Lairig Ghru) is the point with highest influence and point 18 (Knock Hill) is the major outlier. Point 7 (Bens of Jura) has moderately high potential and also moderately high residual. [2 marks]

d) To remove a point we can simply edit the SAS dataset or we can use a data step as follows

```
Data Races_del7;
    Set S3A3.Races;
    If _N_=7 Then Delete;
run;

Data Races_del11;
    Set S3A3.Races;
    If _N_=11 Then Delete;
run;
```

```
Data Races_del18;
    Set S3A3.Races;
    If _N_=18 Then Delete;
run;
```

We can then run the usual regression on each of these three new datasets. In the table below I give some summary measures from these three fits as well as the original fit

Deleted	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	MSE	R^2
None	−539.48	373.07	0.663	775315	0.9191
7	−336.46	395.77	0.405	604511	0.9173
11	−460.82	345.02	0.723	793220	0.8972
18	−811.78	381.87	0.711	278995	0.9716

The major change in the model when we remove point 7 is that the intercept is reduced and the coefficient of climb is also reduced quite a lot. Looking at the original dataset we notice that this race had the largest amount of vertical climb (7500 feet) so the point is most influential in terms of the climb parameter. Removing the point reduces the error sum of squares but does not really change the R^2 suggesting that most of the reduction in the error sum of squares is due to a reduction in the total variability of the record times (total sum of squares).

When we remove point 11, all of the parameters change but none do very much. Looking at the original dataset we see that this race is by far the longest race in the dataset and so it will have high leverage. In fact, however, it does not look like this point is really changing the model and its removal actually decreases the R^2. Despite it being a point of high leverage, it is not an influential point and can safely remain in the model.

Finally removing point 18 does not have a very big change on any of the parameters except maybe the intercept. The big change caused by removing this point is that the mean-squared error goes down by a huge amount and the R^2 increases to 97%. We see that this race is actually a relatively short (3 miles) and flat (350 feet climb) race but yet had a record time of 4719 seconds (78 minutes, 39 seconds)! This point seems quite suspect and should be removed from the analysis. In fact it is now known that this point was entered incorrectly and the correct record time should have been 1119 (18 minutes, 39 seconds).

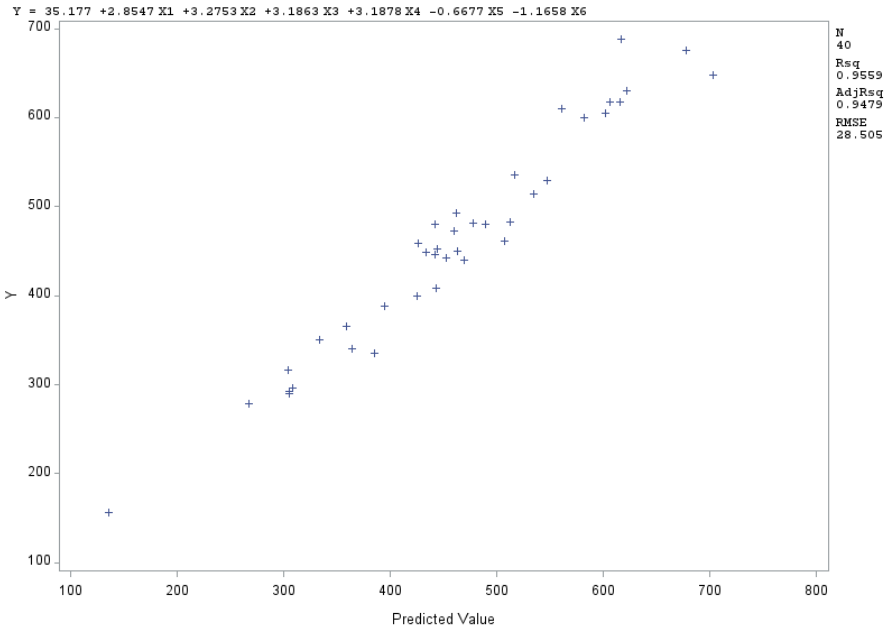
[5 marks]

Q. 2 Textbook 4.12

a) We will examine three plots: the response against the fitted values plot to assess violations of linearity, the residuals against fitted values plot to assess homogeneity of the error variance and the normal qq plot of the residuals to assess non-normality.

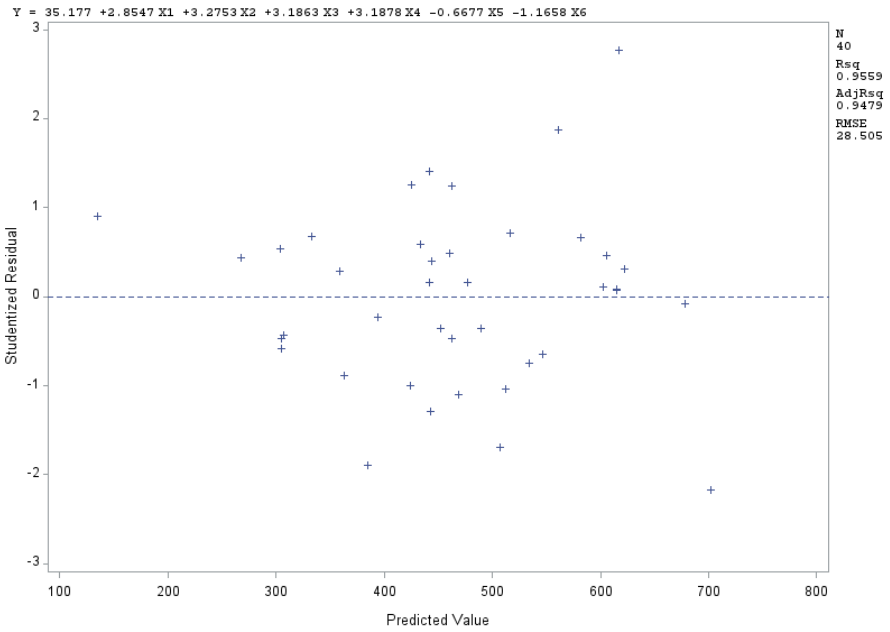
```
PROC REG Data=S3A3.Tab4_8 plots=none;
    Model Y=X1 X2 X3 X4 X5 X6;
    PLOT Y*Predicted.;
    PLOT Student.*Predicted.;
    PLOT Student.*nqq.;
run;
quit;
```

First we examine the response against fitted value plot.



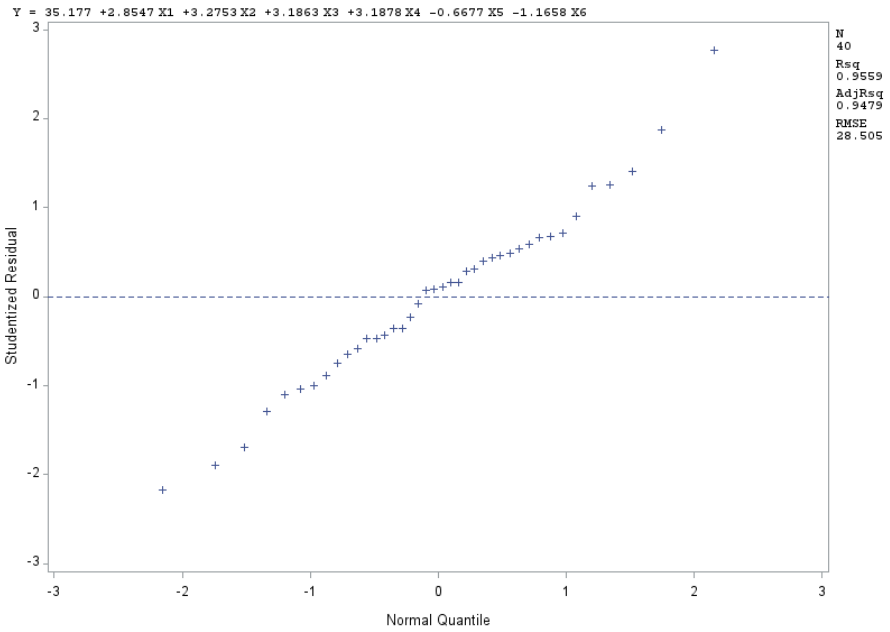
This plot is clearly linear and so we can accept that the linear model is a reasonable model to use for this dataset. **[3 marks]**

The plot of the studentized residuals against the fitted values is



From this plot we see that there is some evidence of increasing variance as the fitted value increases but the evidence is not overwhelming and may be more related to one or two outliers near the larger fitted values. **[3 marks]**

The normal quantile-quantile plot of the studentized residuals is



This plot is very linear and so we are quite happy to accept the assumption of normality of the errors. **[3 marks]**

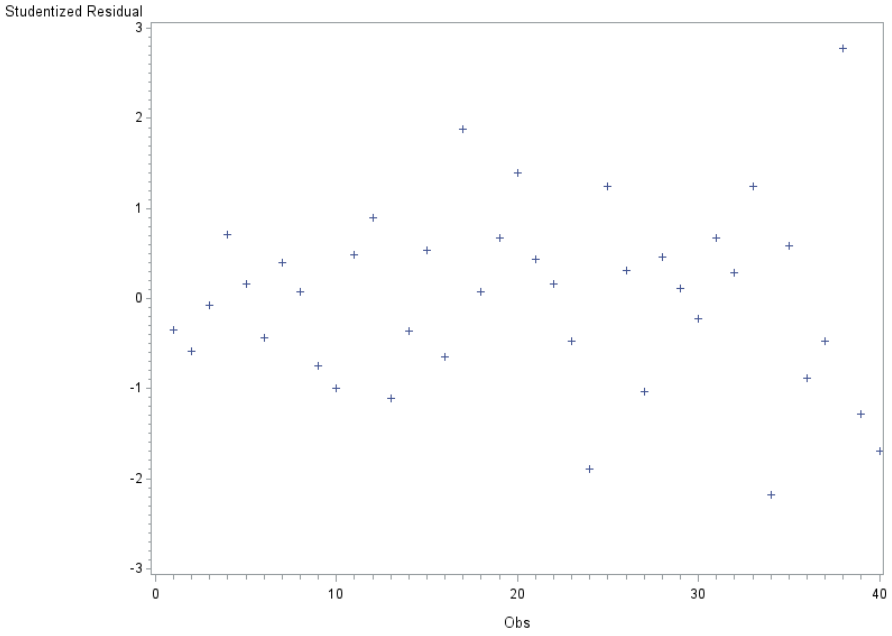
b) Here is the code that is used to compute these quantities and save them in a dataset.

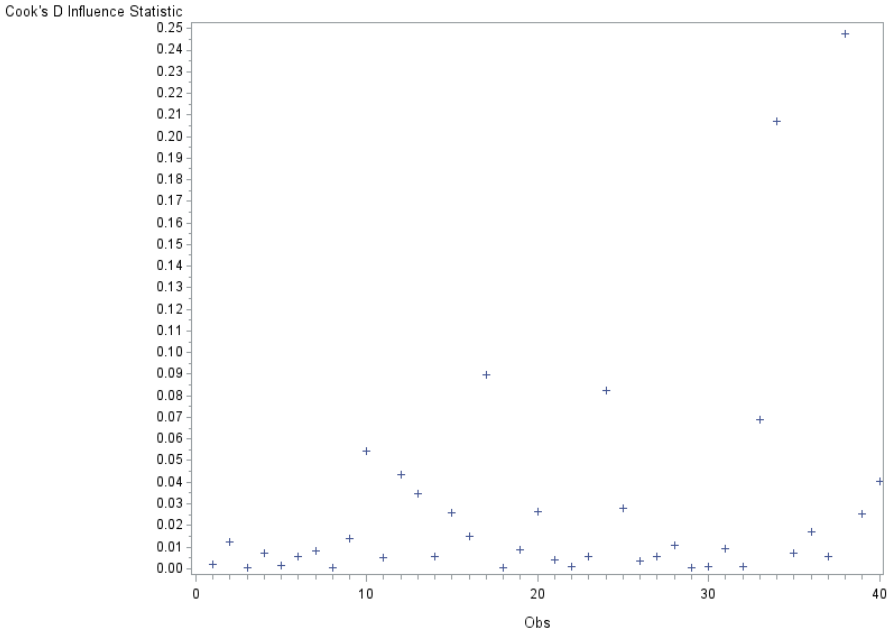
```
PROC REG Data=S3A3.Tab4_8 plots=none;
    Model Y=X1-X6;
    ods output anova=Tab4_8_anova; /* save the ANOVA table */
    output out=Tab4_8out
        residual=res_raw
        student=res_stud
        CookD=C
        DFFITS=DFITS
        H=Leverage;
run;

Data Tab4_8_anova;
    set Tab4_8_anova;
    If source='Error' then call symput ('Q2_sse', ss);
    If source='Model' then call symput ('Q2_p', df);
run;

data Tab4_8out;
    set S3A3.Tab4_8out;
    Obs=_N_;
    d=res_raw/sqrt(&Q2_sse);
    potential=Leverage/(1-Leverage);
    residual_fun=(&Q2_p+1)*d**2/((1-Leverage)*(1-d**2));
    Hadi=potential+residual_fun;
run;
```

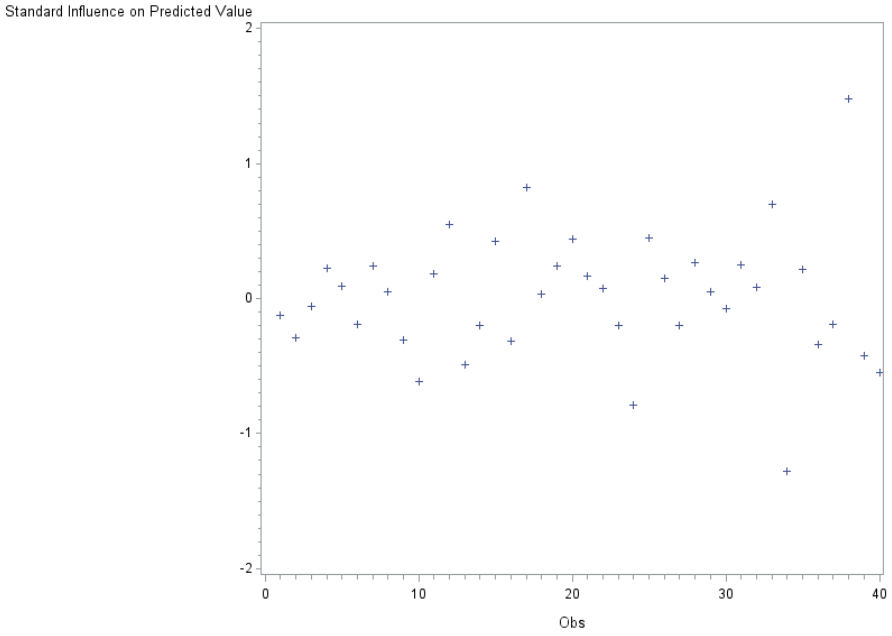
[6 marks]

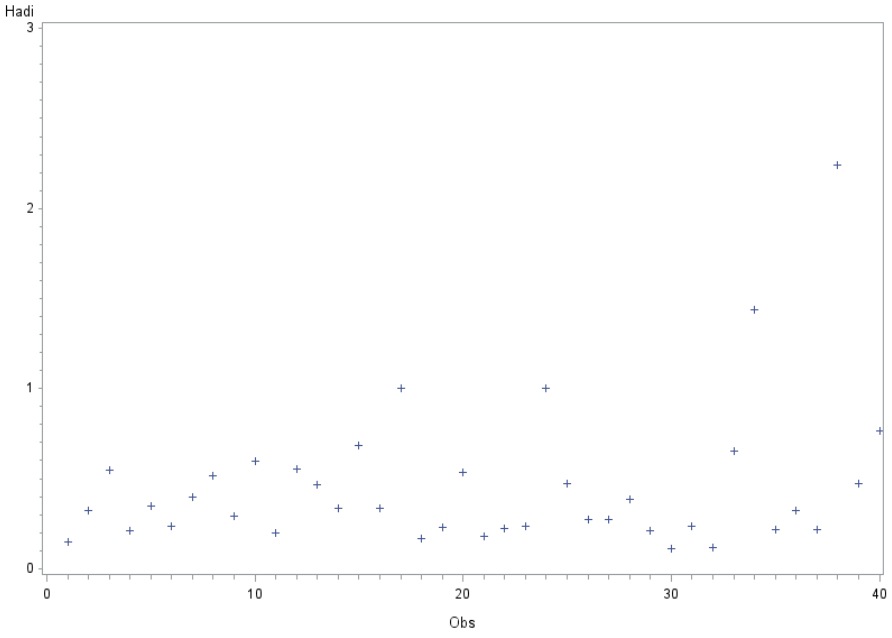
c) Here is the code for the five plots followed by the plots themselves.

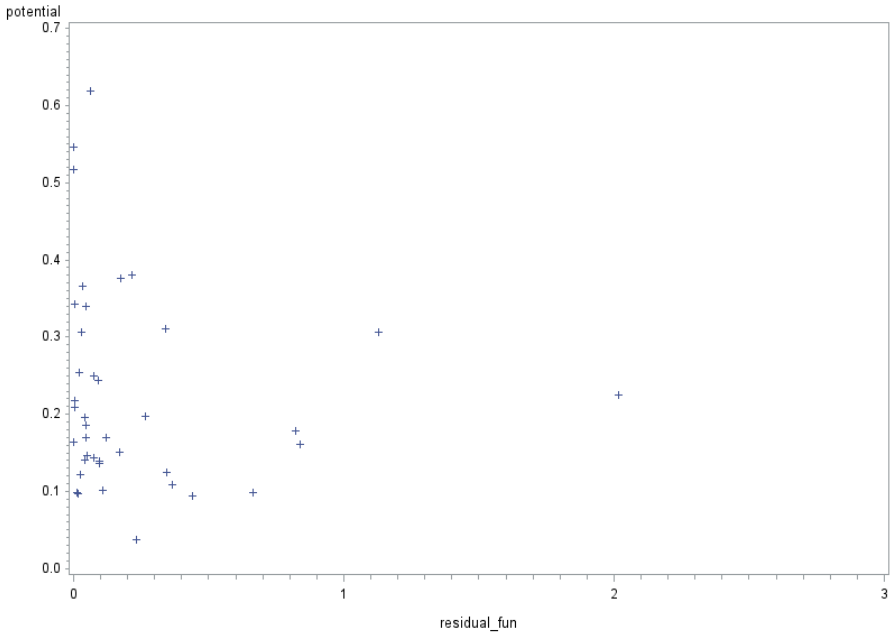
```
PROC GPLOT Data=S3A3.Tab4_8out;
   PLOT res_stud*Obs;
   PLOT C*Obs;
   PLOT DFITS*Obs;
   PLOT Hadi*Obs;
   PLOT potential*residual_fun;
run;
```


Studentized Residual
3
2
1
0
-1
-2
-3
0
10
20
30
40
Obs


Cook's D Influence Statistic
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
Obs


Standard Influence on Predicted Value
2
1
0
-1
-2
0
10
20
30
40
Obs


Hadi
3
2
1
0
0
10
20
30
40
Obs



[5 marks]

d) It appears that observation 38 has a rather high studentized residual. This causes it to also have a very high Cook's distance and Hadi statistic. The potential function for this point, however, is not very large so this point would be classed as an outlier with a larger than expected response value.

Point 34 also has a large Cook's Distance and a residual which is quite large in absolute value (in this case negative). It is not as large an outlier as observation 38 but would still be classed as an outlier with a smaller than expected response.

Looking at the potential-residual plot there do not appear to be any very odd values of the potential function so we would conclude that there are no points of very high influence in this dataset.

One rather odd feature in some of these plots (particularly the studentized residual and DFITS plots) is that there appears to be a "wave-like" pattern in the values across id. This suggests that there may be a violation of the assumption of independence of the errors. **[5 marks]**

Q. 3 Textbook 4.13

a) The simplest way to do this, in SAS, is to fit a model with the first 4 X variables as the predictors and get the added-variable plots for each and then select only that for X_4.

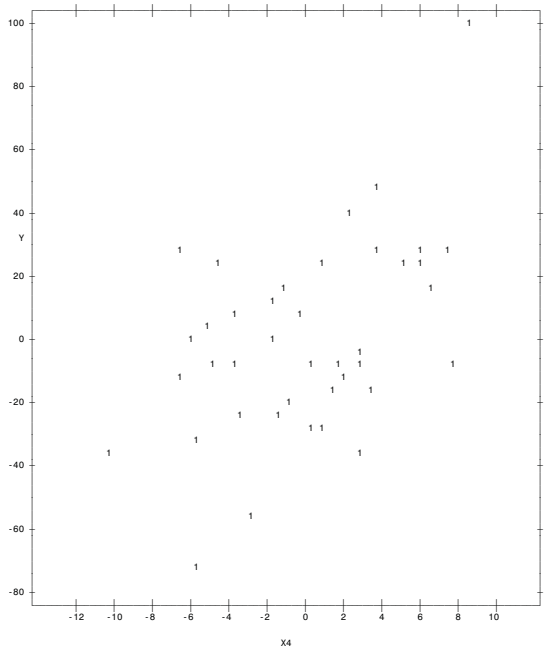
```
PROC REG Data=S3A3.Tab4_8;
    MODEL Y=X1-X4 / partial;
run;
```

The SAS output and the appropriate partial regression plot are

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	581026	145256	186.27	<.0001
Error	35	27293	779.80543		
Corrected Total	39	608319			

Root MSE	27.92500	**R-Square**	0.9551
Dependent Mean	464.92500	**Adj R-Sq**	0.9500
Coeff Var	6.00634		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	28.34689	18.91410	1.50	0.1429
X1	1	1.70065	0.19583	8.68	<.0001
X2	1	2.09068	0.18089	11.56	<.0001
X3	1	2.02086	0.21174	9.54	<.0001
X4	1	3.22952	0.96538	3.35	0.0020



There does look to be an increasing linear trend in the added-variable plot suggesting that including X_4 in the regression is justified. Also the p-value for testing $\beta_4 = 0$ is $p = 0.0020 < 0.05$ confirming that X_4 is a significant predictor even when adjusting for X_1, X_2 and X_3. We therefore keep X_4 in the model for the subsequent analyses. **[6 marks]**

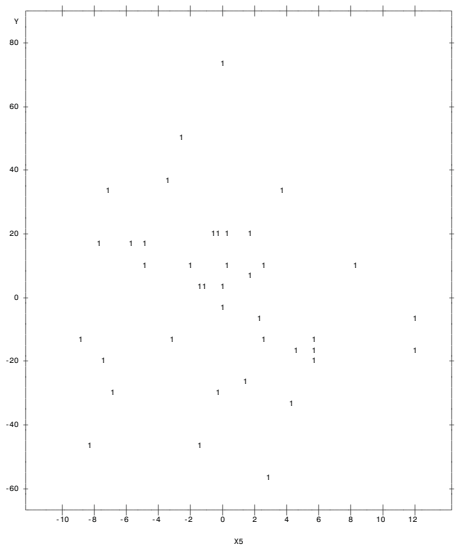
b) Now we consider adding X_5.

```
PROC REG Data=S3A3.Tab4_8;
   MODEL Y=X1-X5 / partial;
run;
quit;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	581467	116293	147.25	<.0001
Error	34	26852	789.76291		
Corrected Total	39	608319			

Root MSE	28.10272	R-Square	0.9559
Dependent Mean	464.92500	Adj R-Sq	0.9494
Coeff Var	6.04457		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.36879	20.18539	1.65	0.1075
X1	1	1.68631	0.19801	8.52	<.0001
X2	1	2.10770	0.18346	11.49	<.0001
X3	1	2.02864	0.21334	9.51	<.0001
X4	1	3.21182	0.97182	3.30	0.0022
X5	1	-0.65746	0.87958	-0.75	0.4599



There does not seem to be any pattern in this plot, the points seem to be a random scatter about 0 suggesting that X_5 will not add anything further to the model which already includes X_1, X_2, X_3 and X_4. This is confirmed by the p-value to test $\beta_5 = 0$ which is $p = 0.4599 > 0.05$ so there is no evidence that $\beta_5 \neq 0$. Therefore we can safely omit X_5 from the model which we do for the rest of the question. **[6 marks]**

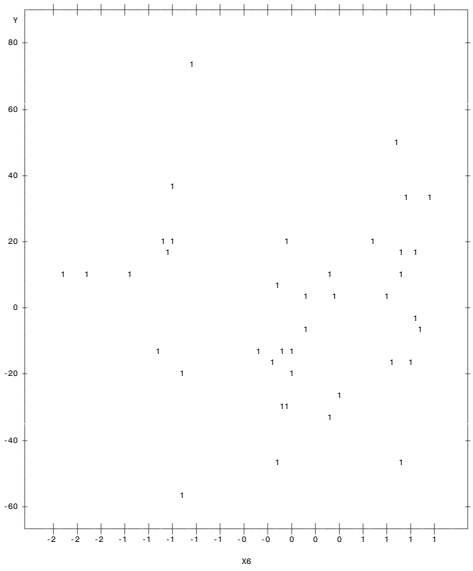
c) We now add X_6 to the model with X_1, X_2, X_3 and X_4.

```
PROC REG Data=S3A3.Tab4_8;
   MODEL Y=X1-X4 X6/ partial;
run;
quit;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	581052	116210	144.91	<.0001
Error	34	27267	801.96389		
Corrected Total	39	608319			

Root MSE	28.31897	**R-Square**	0.9552
Dependent Mean	464.92500	**Adj R-Sq**	0.9486
Coeff Var	6.09108		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	29.76920	20.72003	1.44	0.1599
X1	1	2.66122	5.29618	0.50	0.6186
X2	1	3.05020	5.28985	0.58	0.5680
X3	1	2.97230	5.24657	0.57	0.5748
X4	1	3.20999	0.98490	3.26	0.0025
X6	1	-0.95828	5.27984	-0.18	0.8571



Once again the added variable plot seems to be a random scatter about 0 and the p-value from the t-test for $\beta_6 = 0$ is $p = 0.8571$ suggesting that X_6 is also not needed in this model.

[6 marks]

d) The output for just fitting the model with X_1, X_2 and X_3 is

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	572299	190766	190.66	<.0001
Error	36	36020	1000.55952		
Corrected Total	39	608319			

Root MSE	31.63162	R-Square	0.9408
Dependent Mean	464.92500	Adj R-Sq	0.9359
Coeff Var	6.80360		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	61.92526	18.15890	3.41	0.0016
X1	1	1.63651	0.22076	7.41	<.0001
X2	1	2.17693	0.20280	10.73	<.0001
X3	1	2.01729	0.23984	8.41	<.0001

Comparing the R^2 and adjusted R^2 values as well as the estimates of σ^2 for the four models considered we have the following information

Model	R^2	R^2 (adj)	$\hat{\sigma}^2$
$X_1 + X_2 + X_3$	0.9408	0.9359	31.6316
$X_1 + X_2 + X_3 + X_4$	0.9551	0.9500	27.9250
$X_1 + X_2 + X_3 + X_4 + X_5$	0.9559	0.9494	28.1027
$X_1 + X_2 + X_3 + X_4 + X_6$	0.9552	0.9486	28.3190

The Model with X_1, X_2, X_3 and X_4 has the largest value of the adjusted R^2 and also the smallest residual variance. Even though the value of R^2 does increase when a new variable is added (which it **always** does), the increase is very slight and does not overcome the effect of having an extra parameter in the model. The model without X_4, however, has a substantially lower R^2 and adjusted R^2 as well as a higher residual variance. All of this says that the best of these four models to use is the model with the first 4 X variables as the covariates.

[7 marks]

Q. 4 **a)** Jackknifing involves deleting each observation in turn and refitting the model. **[2 marks]**

This is useful in regression since we can examine how the parameter estimates or fitted values change when we delete a point. If there is a large change then the deleted point has a large influence on the model and should be flagged for further investigation. **[3 marks]**

b) **(i)**

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{j=1}^{n} x_j \\
&= \frac{1}{n}\left(\sum_{j\neq} x_j + x_i\right) \\
&= \frac{n-1}{n}\overline{x}_{(i)} + \frac{x_i}{n} \\
&= \overline{x}_{(i)} - \frac{\overline{x}_{(i)}}{n} + \frac{x_i}{n} \\
&= \overline{x}_{(i)} + \frac{x_i - \overline{x}_{(i)}}{n}
\end{aligned}$$

[4 marks]

(ii) First note that $S_{xx(i)} = \sum_{j\neq i}\left(x_j - \overline{x}_{(i)}\right)^2$.

Hence we have

$$\begin{aligned}
S_{xx} &= \sum_{j=1}^{n}(x_j - \overline{x})^2 \\
&= \sum_{j=1}^{n}\left(x_j - \overline{x}_{(i)} - \frac{x_i - \overline{x}_{(i)}}{n}\right)^2 \\
&= \sum_{j\neq i}\left(x_j - \overline{x}_{(i)} - \frac{x_i - \overline{x}_{(i)}}{n}\right)^2 + \left(x_i - \overline{x}_{(i)} - \frac{x_i - \overline{x}_{(i)}}{n}\right)^2 \\
&= \sum_{j\neq i}\left(x_j - \overline{x}_{(i)}\right)^2 - \frac{2(x_i - \overline{x}_{(i)})}{n}\sum_{j\neq i}\left(x_j - \overline{x}_{(i)}\right) \\
&\quad + \frac{(n-1)(x_i - \overline{x}_{(i)})^2}{n^2} + \left(\frac{n-1}{n}\right)^2\left(x_i - \overline{x}_{(i)}\right)^2 \\
&= S_{xx(i)} + \frac{(n-1) + (n-1)^2}{n^2}\left(x_i - \overline{x}_{(i)}\right)^2 \\
&= S_{xx(i)} + \frac{n-1}{n}\left(x_i - \overline{x}_{(i)}\right)^2
\end{aligned}$$

[5 marks]

(iii) Similarly we have

$$\begin{aligned}
S_{xy} &= \sum_{j=1}^{n} (x_j - \overline{x})\,(y_j - \overline{y}) \\
&= \sum_{j=1}^{n} \left(x_j - \overline{x}_{(i)} - \frac{x_i - \overline{x}_{(i)}}{n}\right)\left(y_j - \overline{y}_{(i)} - \frac{y_i - \overline{y}_{(i)}}{n}\right) \\
&= \sum_{j\neq i} \left(x_j - \overline{x}_{(i)} - \frac{x_i - \overline{x}_{(i)}}{n}\right)\left(y_j - \overline{y}_{(i)} - \frac{y_i - \overline{y}_{(i)}}{n}\right) \\
&\quad + \left(x_i - \overline{x}_{(i)} - \frac{x_i - \overline{x}_{(i)}}{n}\right)\left(y_i - \overline{y}_{(i)} - \frac{y_i - \overline{y}_{(i)}}{n}\right) \\
&= \sum_{j\neq i} \left(x_j - \overline{x}_{(i)}\right)\left(y_j - \overline{y}_{(i)}\right) - \frac{\left(x_i - \overline{x}_{(i)}\right)}{n}\sum_{j\neq i}\left(y_j - \overline{y}_{(i)}\right) - \frac{\left(y_i - \overline{y}_{(i)}\right)}{n}\sum_{j\neq i}\left(x_j - \overline{x}_{(i)}\right) \\
&\quad + \frac{(n-1)\left(x_i - \overline{x}_{(i)}\right)\left(y_i - \overline{y}_{(i)}\right)}{n^2} + \left(\frac{n-1}{n}\right)^2 \left(x_i - \overline{x}_{(i)}\right)\left(y_i - \overline{y}_{(i)}\right) \\
&= S_{xy(i)} + \frac{(n-1) + (n-1)^2}{n^2}\left(x_i - \overline{x}_{(i)}\right)\left(y_i - \overline{y}_{(i)}\right) \\
&= S_{xx(i)} + \frac{n-1}{n}\left(x_i - \overline{x}_{(i)}\right)\left(y_i - \overline{y}_{(i)}\right)
\end{aligned}$$

[5 marks]

c) The case-deletion least squares estimates are

$$\hat{\beta}_{1(i)} = \frac{S_{xy(i)}}{S_{xx(i)}} \qquad \hat{\beta}_{0(i)} = \overline{y}_{(i)} - \hat{\beta}_{1(i)}\overline{x}_{(i)}.$$

From part b(i) we can see that

$$\overline{x}_{(i)} = \frac{n\overline{x} - x_i}{n-1}$$

and from the other two parts of part (b) we have

$$S_{xx(i)} = S_{xx} - \frac{n-1}{n}\left(x_i - \frac{n\overline{x} - x_i}{n-1}\right)^2 = S_{xx} - \frac{n}{n-1}(x_i - \overline{x})^2$$

and similarly that

$$S_{xy(i)} = S_{xy} - \frac{n-1}{n}\left(x_i - \frac{n\overline{x} - x_i}{n-1}\right)\left(y_i - \frac{n\overline{y} - y_i}{n-1}\right) = S_{xx} - \frac{n}{n-1}(x_i - \overline{x})(y_i - \overline{y})$$

[3 marks]

Hence we

$$\begin{aligned}
\hat{\beta}_{1(i)} &= \frac{(n-1)S_{xy} - n(x_i - \overline{x})(y_i - \overline{y})}{(n-1)S_{xx} - n(x_i - \overline{x})^2} \\
\hat{\beta}_{0(i)} &= \frac{n\overline{y} - y_i}{n-1} - \hat{\beta}_{1(i)} \frac{n\overline{x} - x_i}{n-1} \\
&= \frac{n}{n-1}(\overline{y} - \hat{\beta}_{1(i)}\overline{x}) - \frac{1}{n-1}(y_i - \hat{\beta}_{1(i)} x_i)
\end{aligned}$$

[3 marks]