STAT 3A03 Applied Regression Analysis With SAS

Fall 2017

Assignment 5 Solution Set

Q. 1 **a)** The code that I used and the output is as follows

```
PROC GLM Data=S3A3.Wool plots=none;
  Class Amp Len Load;
  Model Cycles=Amp Len Load;
  Output Out=woolout
    predicted=fitted
student=student;
run;
```

The GLM Procedure

Dependent Variable: Cycles

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	15559598.44	2593266.41	11.11	<.0001
Error	20	4669019.85	233450.99		
Corrected Total	26	20228618.30			

R-Square	Coeff Var	Root MSE	Cycles Mean
0.769187	56.09291	483.1677	861.3704

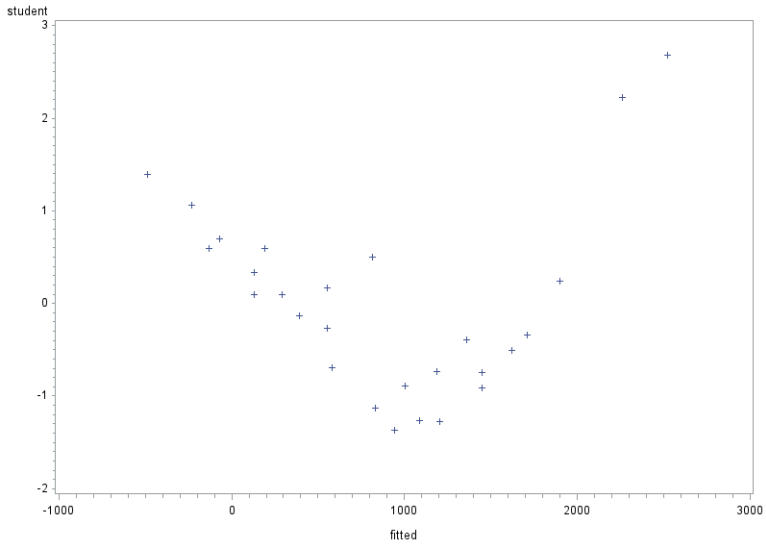
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Amp	2	5624248.963	2812124.481	12.05	0.0004
Len	2	8182252.519	4091126.259	17.52	<.0001
Load	2	1753096.963	876548.481	3.75	0.0413

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Amp	2	5624248.963	2812124.481	12.05	0.0004
Len	2	8182252.519	4091126.259	17.52	<.0001
Load	2	1753096.963	876548.481	3.75	0.0413

[2 marks]

A plot of the studentized residuals against the fitted values is as follows

```
PROC GPLOT Data=woolout;
    Plot student*fitted;
run;
```



There is a clear pattern in these residuals suggesting non-linearity in the model. [2 marks]

b) Here is the code and output adding in all three pairwise interactions.

```
PROC GLM Data=S3A3.Wool plots=none;
    Class Amp Len Load;
    Model Cycles=Amp Len Load Amp*Len Amp*Load Len*Load;
    Output Out=woolout1
        predicted=fitted
        student=student;
run;
```

The GLM Procedure

Dependent Variable: Cycles

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	20131626.44	1118423.69	92.25	<.0001
Error	8	96991.85	12123.98		
Corrected Total	26	20228618.30			

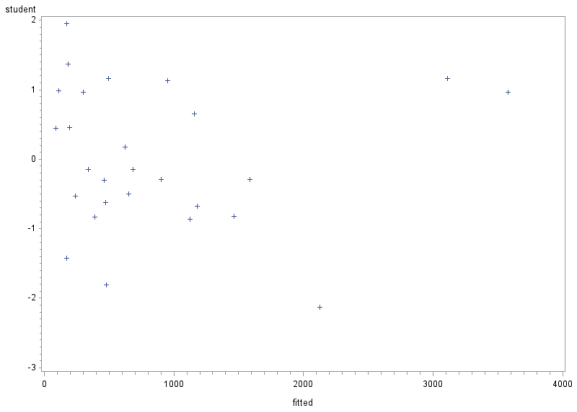
R-Square	Coeff Var	Root MSE	Cycles Mean
0.995205	12.78300	110.1090	861.3704

We notice that the R^2 value has now increased from 76.9% to 99.5% in this model. [3 marks]
A test of the null hypothesis that we can omit the interaction terms has test statistic

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(4669019.85 - 96991.85)/(20 - 8)}{12123.98} \\ &= 31.43 \end{aligned}$$

If the true model is the reduced (no interaction) model then this will be an observation from an $F(12, 8)$ distribution. From Table A.5 in the text book we see that the 1% critical value for this distribution is 5.67. Since the observed value of F is greater than 5.67 we can conclude that the p-value is less than 0.01 and so we reject H_0 and conclude that the model including interactions is a significantly better fit. [3 marks]
Furthermore a plot of the studentized residuals against the fitted value shows that the non-linearity problem has now been solved also.

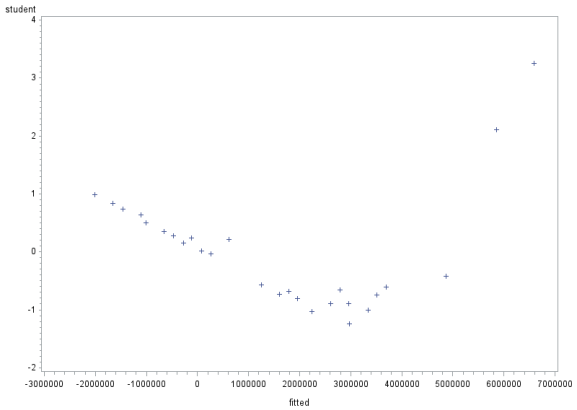
```
PROC GPLOT Data=woolout1;
    Plot student*fitted;
run;
```



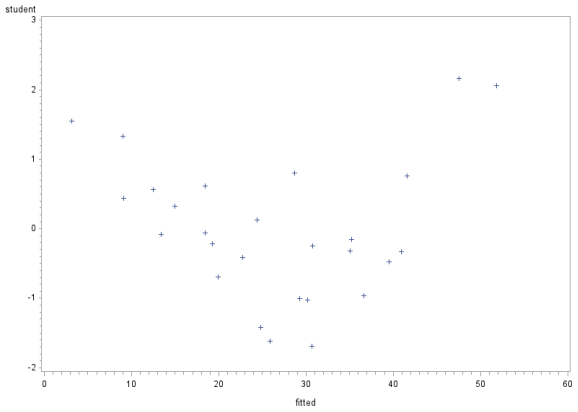
[2 marks]

c) The following plots come from trying the square, square root and log transformations for the response variable and the no interaction model.

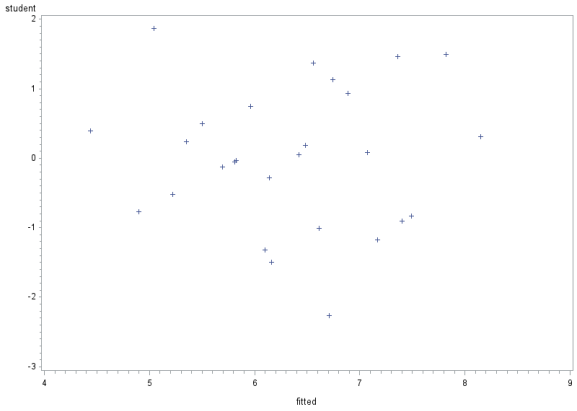
Cycles Squared



Square Root of Cycles



Log of Cycles



From these plots, we see that using a log transformation seems to be the best at removing the non-linearity in the model. [4 marks]

The ANOVA table and other summary statistics for this transformation are

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	22.48517565	3.74752927	104.47	<.0001
Error	20	0.71742240	0.03587112		
Corrected Total	26	23.20259805			

R-Square	Coeff Var	Root MSE	LogCycles Mean
0.969080	2.989807	0.189397	6.334748

We note that on the log scale the R^2 value for the model without interactions is now 96.9%. [3 marks]

d) For the model on the log scale with all pairwise interactions we get the ANOVA table

Dependent Variable: LogCycles

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	23.03668445	1.27981580	61.71	<.0001
Error	8	0.16591360	0.02073920		
Corrected Total	26	23.20259805			

R-Square	Coeff Var	Root MSE	LogCycles Mean
0.992849	2.273352	0.144011	6.334748

The R^2 value for this model on the log scale is now over 99%. [3 marks]

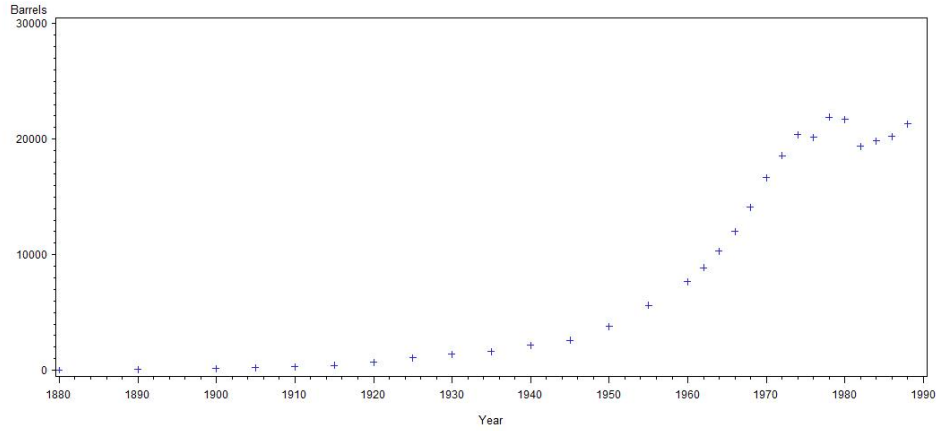
The F statistic to compare this model with the model without interactions is

$$\begin{aligned} F &= \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(df_{\text{red}} - df_{\text{full}})}{\text{MSE}_{\text{full}}} \\ &= \frac{(0.7174224 - 0.1659136)/(20 - 8)}{0.0207392} \\ &= 2.22 \end{aligned}$$

Once again we compare this to the critical values of the $F(12, 8)$ distribution. From Table A.4 we see that the 5% critical value is 3.28 and so (since $2.22 < 3.28$) the p-value of the test is greater than 0.05. Hence we do not reject H_0 and can conclude that there is no evidence that we need the interaction terms in the model on the log scale. [3 marks]

Q. 2 **a)** First we plot the millions of barrels produced against the year as follows

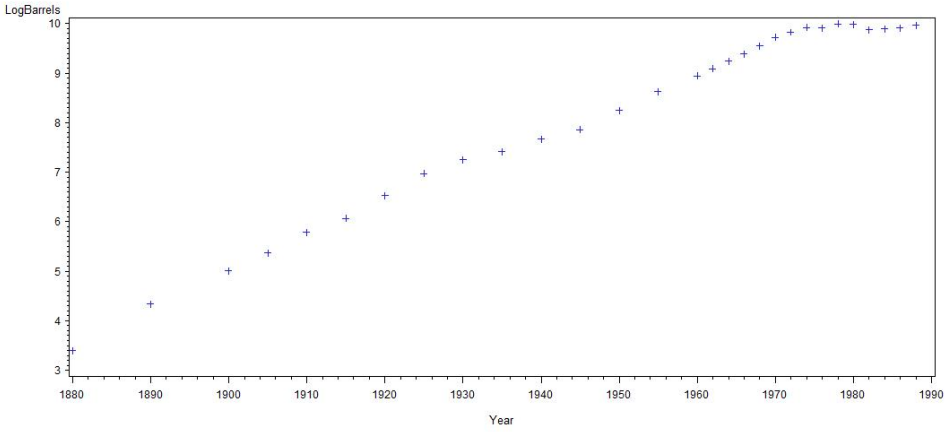
```
PROC GPLOT Data=S3A3.Oil;
  Plot Barrels*Year;
run;
quit;
```



Clearly the relationship is not linear. There appears to be exponential growth for most of the times although that seems to have changed in the later years. [2 marks]
Next we try to log transform the production covariate.

```
Data S3A3.Oil;
  set S3A3.Oil;
  LogBarrels=log(Barrels);
run;

PROC GPLOT Data=S3A3.Oil;
  Plot LogBarrels*Year;
run;
quit;
```



For most of the range of years this seems to have improved the linearity although there is still an issue in the later years. [2 marks]

b) The regression for log of oil production on year results in the following output and diagnostic plots

```
PROC REG Data=S3A3.Oil plots=none;
   Model LogBarrels=Year;
   Plot LogBarrels*Year;
   Plot Student.*Predicted.;
   Plot Student.*nqq.;
   Plot CookD.*Year;
run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: LogBarrels

Number of Observations Read	30
Number of Observations Used	29
Number of Observations with Missing Values	1

Analysis of Variance

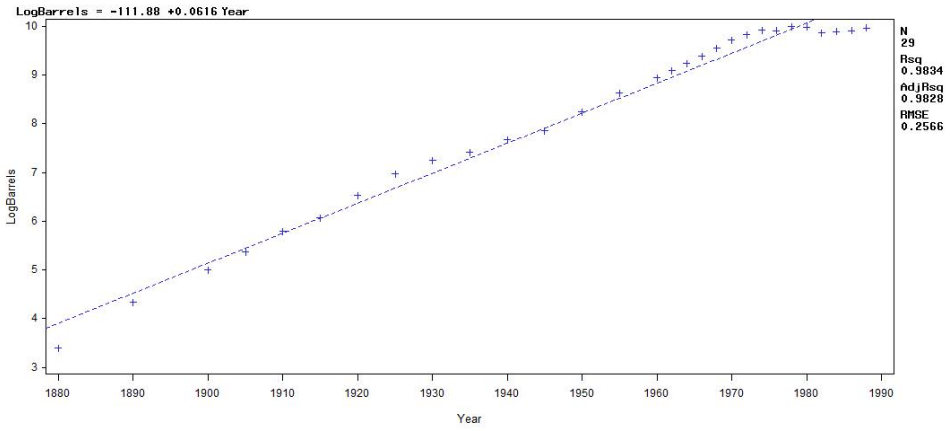
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	105.64736	105.64736	1604.18	<.0001
Error	27	1.77815	0.06586		
Corrected Total	28	107.42551			

Root MSE	0.25663	R-Square	0.9834
Dependent Mean	8.13163	Adj R-Sq	0.9828
Coeff Var	3.15591		

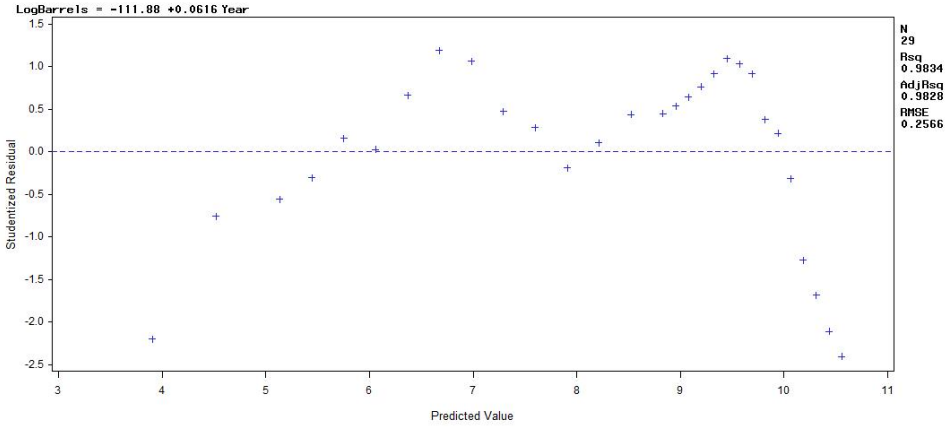
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-111.87532	2.99664	-37.33	<.0001
Year	1	0.06159	0.00154	40.05	<.0001

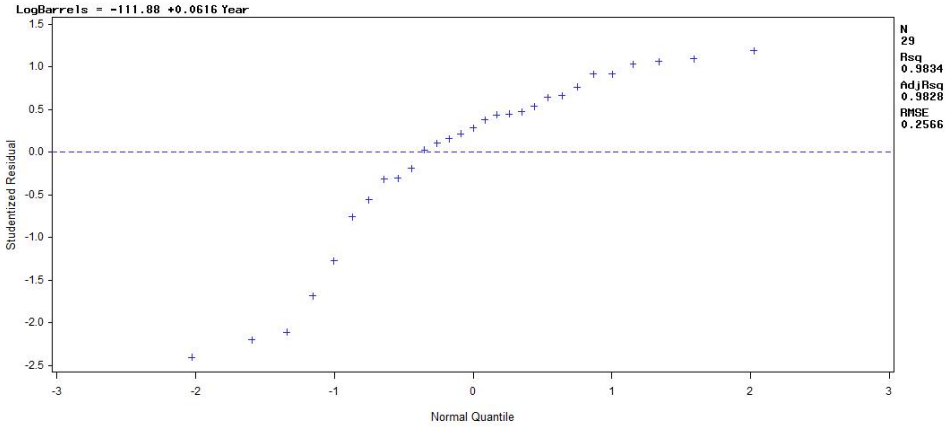
[2 marks]



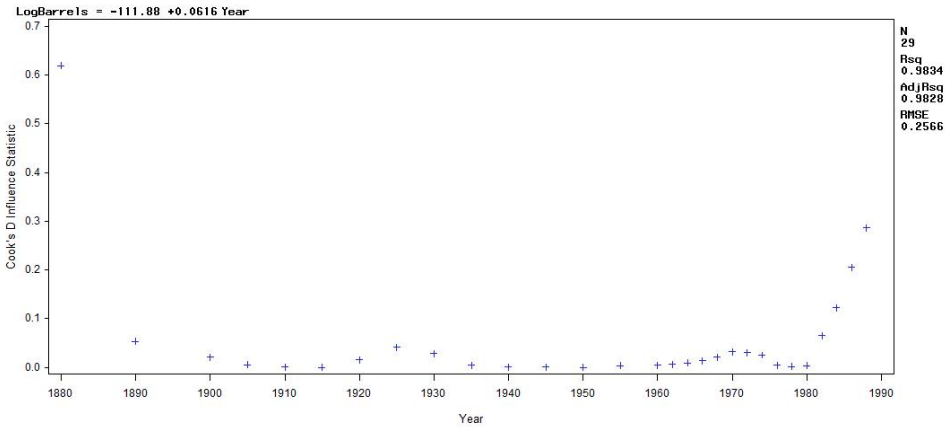
[1 mark]



[1 mark]



[1 mark]



[1 mark]

The fitted model has a very significant F statistic and an R^2 of 98.34% so the model appears to fit quite well. The plots, however, tell another story. although the linear relationship fits most of the years, the production in the last few observations is consistently overestimated resulting in negative residuals. We also see consistent overestimation of production in the early years of the dataset. There is evidence of correlation between successive observations and the normality assumption is quite suspect due to marked curvature in the normal quantile-quantile plot. The very first observation in 1880 seems quite influential based on the Cook's Distance plot and the final few observations also have somewhat elevated Cook's Distance. Overall we would not be at all satisfied that this is an appropriate linear model. [3 marks]

c) Next we shall set up an indicator variable for whether the year was prior to or after the 1973 Oil Crisis. We will also need an interaction variable between this indicator and Year to enable us to fit separate slopes to the two time periods.

```
Data S3A3.Oil;
  Set S3A3.Oil;
  If Year < 1973 then Ind1973=0;
  else Ind1973=1;
  Interaction=Ind1973*Year;
run;
```

[1 mark]

Now we fit a model with both the indicator variable and the interaction variable.

```
PROC REG Data=S3A3.Oil plots=none;
   Model LogBarrels=Year Ind1973 Interaction;
   Plot LogBarrels*Predicted.;
   Plot Student.*Predicted.;
   Plot Student.*nqq.;
   Plot CookD.*Year;
run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: LogBarrels

Number of Observations Read	30
Number of Observations Used	29
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	107.09025	35.69675	2661.87	<.0001
Error	25	0.33526	0.01341		
Corrected Total	28	107.42551			

Root MSE	0.11580	R-Square	0.9969
Dependent Mean	8.13163	Adj R-Sq	0.9965
Coeff Var	1.42411		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-121.31370	1.76711	-68.65	<.0001
Year	1	0.06650	0.00091254	72.87	<.0001
Ind1973	1	132.18496	17.78712	7.43	<.0001
Interaction	1	-0.06697	0.00898	-7.46	<.0001

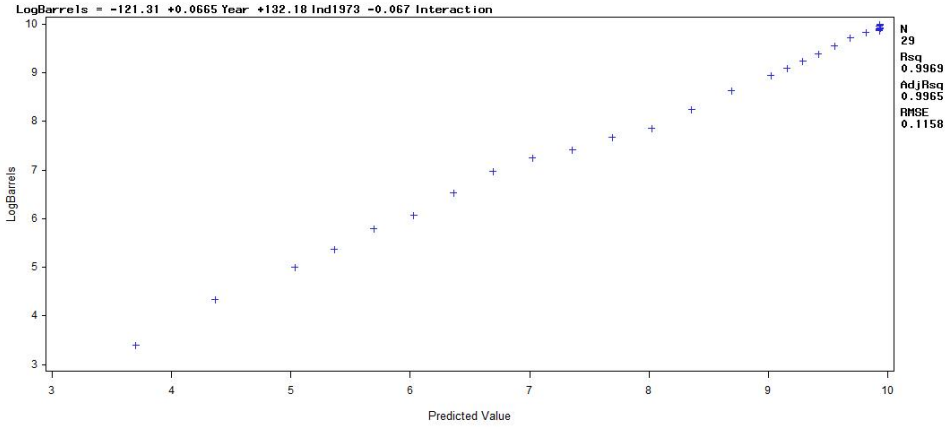
This results in the following fitted lines

$$\widehat{\log(\text{Barrels})} = \begin{cases} -121.314 + 0.06650\text{Year} & \text{if Year } < 1973 \\ 10.871 - 0.00047\text{Year} & \text{if Year } \geqslant 1973 \end{cases}$$

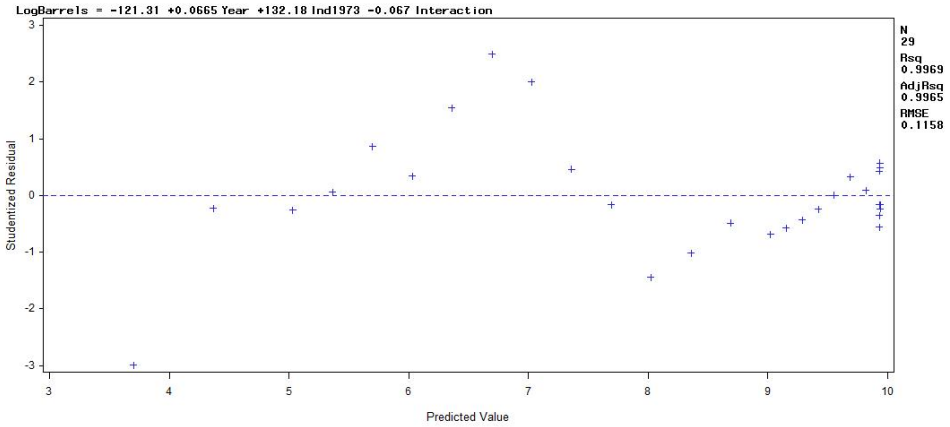
[4 marks]

This analysis clearly shows that the pattern of production subsequent to 1973 was very different from that prior to that date. Prior to 1973 oil production was steadily increasing at an exponential rate whereas after 1973 it became essentially static over time. [1 mark]

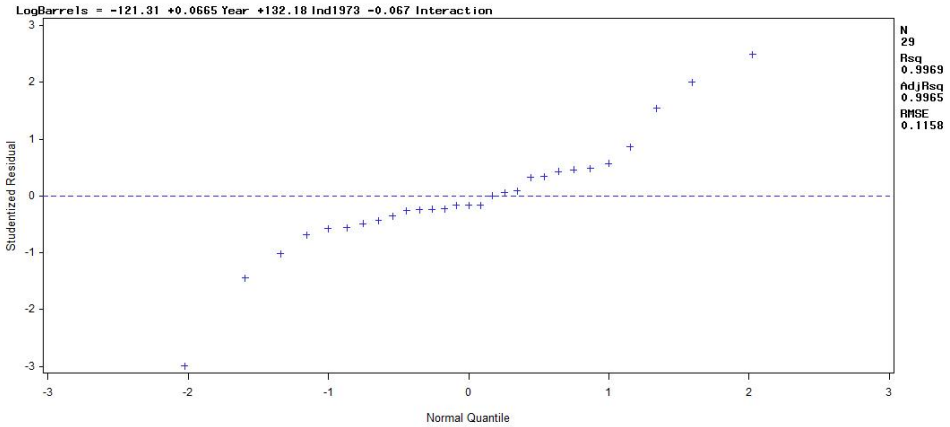
The diagnostic plots are:



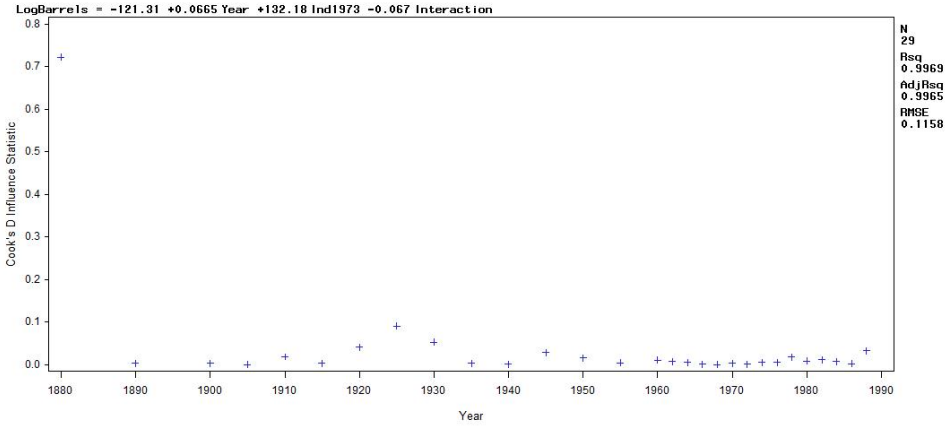
[1 mark]



[1 mark]



[1 mark]



[1 mark]

The observed value against fitted value plot now seems much more linear and the influence of the last few years has been greatly reduced. Some major issues that still remain, however, are that there appears to be correlation between the errors (not unexpected for data gathered over time like this) and the normality assumption still seems suspect. The first observation still has rather elevated Cook's Distance and very large (negative) residual which is a sign that the model does not fit well that long ago. [2 marks]

Q. 3 a)

$$\begin{aligned}\sum_{i=1}^{n} w_i(x_i - \overline{x}_w) &= \sum_{i=1}^{n} w_i x_i - \overline{x}_w \sum_{i=1}^{n} w_i \\ &= \sum_{i=1}^{n} w_i x_i - \left(\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}\right) \sum_{i=1}^{n} w_i \\ &= \sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i x_i \\ &= 0\end{aligned}$$

[2 marks]

Hence we can write

$$\begin{aligned}\sum_{i=1}^{n} w_i x_i(x_i - \overline{x}_w) &= \sum_{i=1}^{n} w_i x_i(x_i - \overline{x}_w) - \overline{x}_w \sum_{i=1}^{n} w_i(x_i - \overline{x}_w) \\ &= \sum_{i=1}^{n} w_i x_i(x_i - \overline{x}_w) - \sum_{i=1}^{n} w_i \overline{x}_w(x_i - \overline{x}_w) \\ &= w_i(x_i - \overline{x}_w)^2\end{aligned}$$

[2 marks]

By symmetry we must also have $\sum_{i=1}^{n} w_i(y_i - \overline{y}_w) = 0$ and so we can write

$$\begin{aligned}\sum_{i=1}^{n} w_i x_i(y_i - \overline{y}_w) &= \sum_{i=1}^{n} w_i x_i(y_i - \overline{y}_w) - \overline{x}_w \sum_{i=1}^{n} w_i(y_i - \overline{y}_w) \\ &= \sum_{i=1}^{n} w_i x_i(y_i - \overline{y}_w) - \sum_{i=1}^{n} w_i \overline{x}_w(y_i - \overline{y}_w) \\ &= w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w)\end{aligned}$$

[2 marks]

b) From our notes the weighted least squares criterion is

$$S_w(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_i)^2$$

[1 mark]

The partial derivatives of this criterion are

$$\begin{aligned}\frac{\partial S_w}{\partial \beta_0} &= -2\sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial S_w}{\partial \beta_0} &= -2\sum_{i=1}^{n} w_i x_i(y_i - \beta_0 - \beta_1 x_i)\end{aligned}$$

and so we need to solve the two equations

$$\sum_{i=1}^{n} w_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (1)$$

$$\sum_{i=1}^{n} w_i x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (2)$$

[2 marks]

From Equation (1) we see that

$$\begin{aligned}
& \textstyle\sum_{i=1}^{n} w_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &=& 0 \\
\Rightarrow\ & \textstyle\sum_{i=1}^{n} w_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} w_i - \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i &=& 0 \\
\Rightarrow\ & \hat{\beta}_0 \textstyle\sum_{i=1}^{n} w_i &=& \textstyle\sum_{i=1}^{n} w_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i \\
\Rightarrow\ & \hat{\beta}_0 &=& \dfrac{\sum w_i y_i}{\sum w_i} - \hat{\beta}_1 \dfrac{\sum w_i x_i}{\sum w_i} \\
\Rightarrow\ & \hat{\beta}_0 &=& \overline{y}_w - \hat{\beta}_i \overline{x}_w
\end{aligned}$$

[2 marks]

Inserting this into Equation (2) we get

$$\begin{aligned}
& \textstyle\sum_{i=1}^{n} w_i x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &=& 0 \\
\Rightarrow\ & \textstyle\sum_{i=1}^{n} w_i x_i(y_i - \overline{y}_w + \hat{\beta}_1 \overline{x}_w - \hat{\beta}_1 x_i) &=& 0 \\
\Rightarrow\ & \textstyle\sum_{i=1}^{n} w_i x_i(y_i - \overline{y}_w) - \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i(x_i - \overline{x}_w) &=& 0 \\
\Rightarrow\ & \textstyle\sum_{i=1}^{n} w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w) - \hat{\beta}_1 \sum_{i=1}^{n} w_i(x_i - \overline{x}_w)^2 &=& 0 \\
\Rightarrow\ & \hat{\beta}_1 &=& \dfrac{\sum w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w)}{\sum w_i(x_i - \overline{x}_w)^2}
\end{aligned}$$

[2 marks]

c) I will show the unbiasedness of $\hat{\beta}_1$ first. From above we can write

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w)}{\sum w_i(x_i - \overline{x}_w)^2} \\
&= \frac{\sum w_i(x_i - \overline{x}_w) y_i - \overline{y}_w \sum w_i(x_i - \overline{x}_w)}{\sum w_i(x_i - \overline{x}_w)^2} \\
&= \frac{\sum w_i(x_i - \overline{x}_w) y_i}{\sum w_i(x_i - \overline{x}_w)^2} \\
&= \sum_{i=1}^{n} \frac{w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} y_i
\end{aligned}$$

Replacing the observed y_i with the corresponding random variable Y_i we get the random variable $\hat{\beta}_1$ to be

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} Y_i$$

[3 marks]

Now taking expectations we get

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_1 \mid x_1, \ldots, x_n\right) &= \sum_{i=1}^{n} \frac{w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \mathrm{E}(Y_i \mid x_i) \\
&= \sum_{i=1}^{n} \frac{w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \frac{\sum w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} + \beta_1 \frac{\sum w_i x_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \\
&= \beta_0 \frac{0}{\sum w_j(x_j - \overline{x}_w)^2} + \beta_1 \frac{\sum w_i(x_i - \overline{x}_w)^2}{\sum w_j(x_j - \overline{x}_w)^2} \\
&= \beta_1
\end{aligned}$$

[3 marks]

Now for $\hat{\beta}_0$ we have

$$\begin{aligned}
\hat{\beta}_0 &= \overline{y}_w - \hat{\beta}_1 \overline{x}_w \\
&= \frac{\sum w_i y_i}{\sum w_i} - \overline{x}_w \sum_{i=1}^{n} \frac{w_i(x_i - \overline{x}_w)}{\sum w_i(x_i - \overline{x}_w)^2} y_i \\
&= \sum_{i=1}^{n} \left(\frac{w_i}{\sum w_j} y_i - \sum_{i=1}^{n} \frac{\overline{x}_w w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} y_i \right) \\
&= \sum_{i=1}^{n} \left(\frac{w_i}{\sum w_j} - \sum_{i=1}^{n} \frac{\overline{x}_w w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \right) y_i
\end{aligned}$$

and so the random variable $\hat{\beta}_0$ is given by

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{w_i}{\sum w_j} - \sum_{i=1}^{n} \frac{\overline{x}_w w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \right) Y_i$$

[3 marks]

Taking expectations again we get

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_0 \mid x_1, \ldots, x_n\right) &= \sum_{i=1}^{n} \left(\frac{w_i}{\sum w_j} - \sum_{i=1}^{n} \frac{\overline{x}_w w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \right) (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \left(\frac{\sum w_i}{\sum w_j} - \frac{\overline{x}_w \sum w_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \right) + \beta_1 \left(\frac{\sum w_i x_i}{\sum w_j} - \frac{\overline{x}_w \sum w_i x_i(x_i - \overline{x}_w)}{\sum w_j(x_j - \overline{x}_w)^2} \right) \\
&= \beta_0 \left(1 + \frac{0}{\sum w_j(x_j - \overline{x}_w)^2} \right) + \beta_1 \left(\overline{x}_w - \frac{\overline{x}_w \sum w_i(x_i - \overline{x}_w)^2}{\sum w_j(x_j - \overline{x}_w)^2} \right) \\
&= \beta_0
\end{aligned}$$

[3 marks]

Q. 4 Here is the code used to construct the transformed variables and to fit the transformed model and the output.

a)
```
Data S3A3.Chroma;
     Set S3A3.Chroma;
     LogAmt=log(Amount);
     LogOut=log(Output);
run;

PROC REG Data=S3A3.Chroma plots=none;
    Model LogOut=LogAmt;
    Plot Student.*LogAmt;
    Plot Student.*nqq.;
    Output Out=ChromaOut
         Predicted=Fitted
         Student=Res_stud;
run;
```

Dependent Variable: LogOut

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	68.82878	68.82878	10816.7	<.0001
Error	18	0.11454	0.00636		
Corrected Total	19	68.94332			

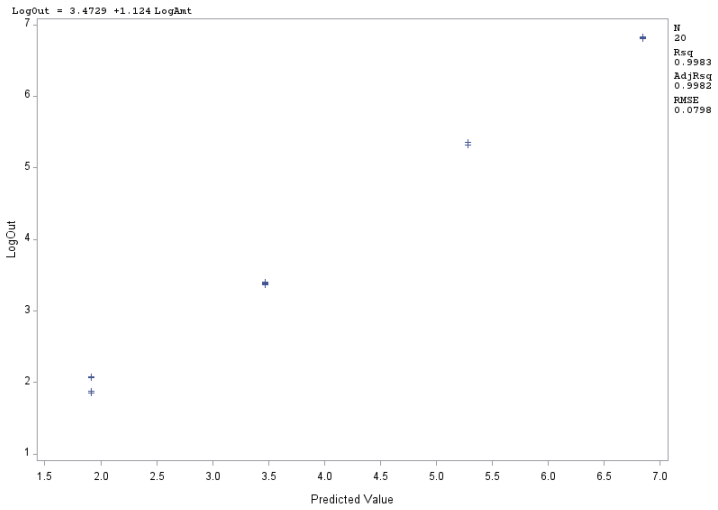
Root MSE	0.07977	**R-Square**	0.9983
Dependent Mean	4.37741	**Adj R-Sq**	0.9982
Coeff Var	1.82231		

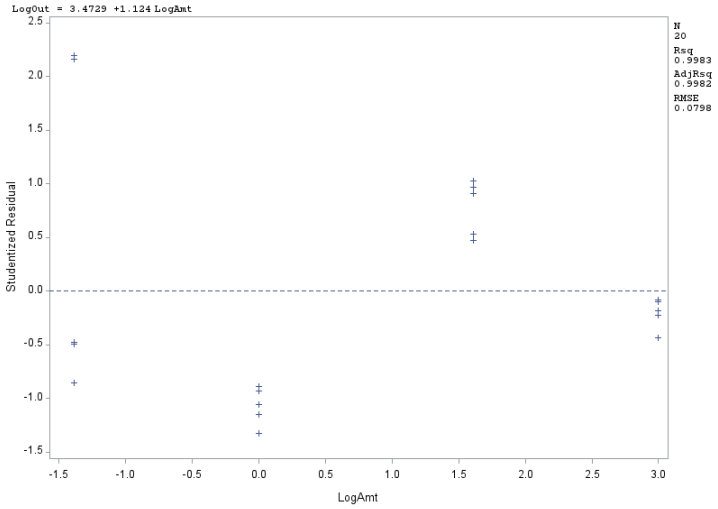
Parameter Estimates

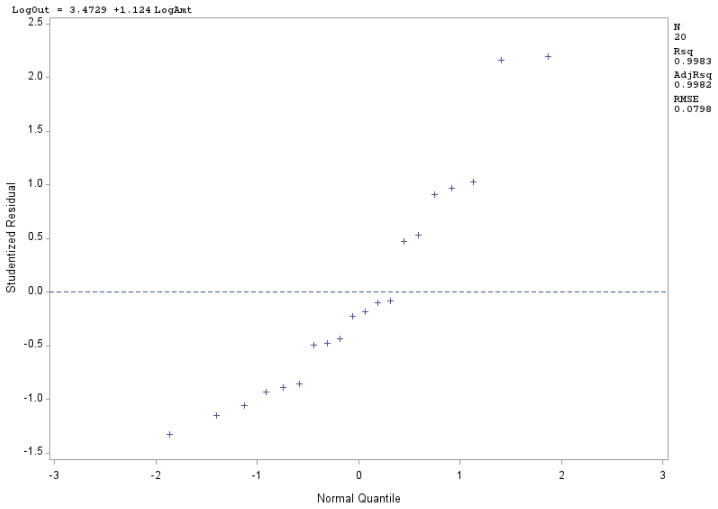
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.47291	0.01984	175.01	<.0001
LogAmt	1	1.12399	0.01081	104.00	<.0001

[2 marks]

The usual diagnostic plots on the residuals for the fitted model are







We can't see much from the first of these in this case. The residual against fitted value plot shows that the variability for the smallest amount is much greater than for the other amounts. There is also evidence of non-linearity since the residuals for when the amount is 1 are all negative whereas when the amount is 5 they are all positive. The normality assumption also seems somewhat suspect here. [5 marks]

b) We can adapt the code given in SAS Lab 11 for this problem to get the weights

```
PROC MEANS Data=ChromaOut NWay Noprint;
    CLASS Amount;
    VAR Res_stud;
    OUTPUT Out=temp
        Mean=Mean
        Var=Var;
run;

PROC PRINT Data=temp;
run;
```

Obs	amount	_TYPE_	_FREQ_	Mean	Var
1	0.25	1	5	0.50423	2.34418
2	1	1	5	-1.06795	0.03096
3	5	1	5	0.78009	0.06755
4	20	1	5	-0.20390	0.02024

```
Data temp;
    set temp;
    If (_N_=1) then call symput('var1', Var);
    If (_N_=2) then call symput('var2', Var);
    If (_N_=3) then call symput('var3', Var);
    If (_N_=4) then call symput('var4', Var);
run;

PROC MEANS Data=ChromaOut noprint;
     Var Res_stud;
     Output Out=temp1
         Mean=Mean
   Var=Var;
run;

Data temp;
    set temp;
    Call symput('vare', Var);
run;

Data ChromaOut;
    Set ChromaOut;
    If (Amount=0.25) then var=&var1;
    If (Amount=1) then var=&var2;
    If (Amount=5) then var=&var3;
    If (Amount=20) then var=&var4;
    w=&vare/var;
run;
```

Looking at the resulting dataset we see that the weights are

$$w_i = \begin{cases} 0.4507 & \text{if } x_i = 0.25 \\ 34.1236 & \text{if } x_i = 1 \\ 15.6430 & \text{if } x_i = 5 \\ 52.2046 & \text{if } x_i = 20 \end{cases}$$

[6 marks]

c) We will now use these weights in a weighted regression

```
PROC REG Data=ChromaOut;
    Model LogOut=LogAmt;
    Weight w;
    Plot Student.*LogAmt;
    Plot Student.*nqq.;
run;
```

Dependent Variable: LogOut

Number of Observations Read	20
Number of Observations Used	20

Weight: w

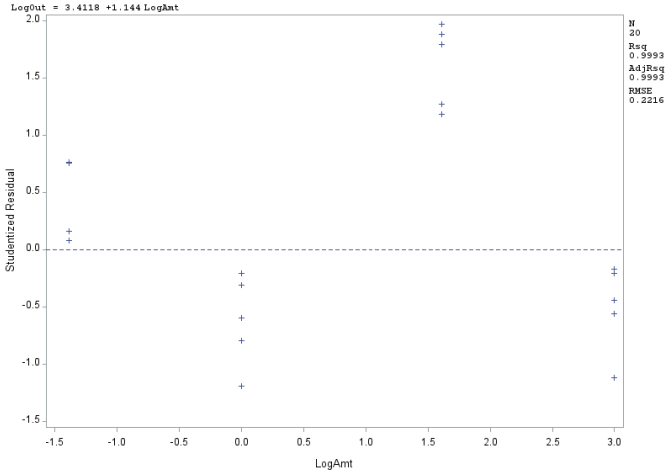
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1244.78315	1244.78315	25359.4	<.0001
Error	18	0.88354	0.04909		
Corrected Total	19	1245.66669			

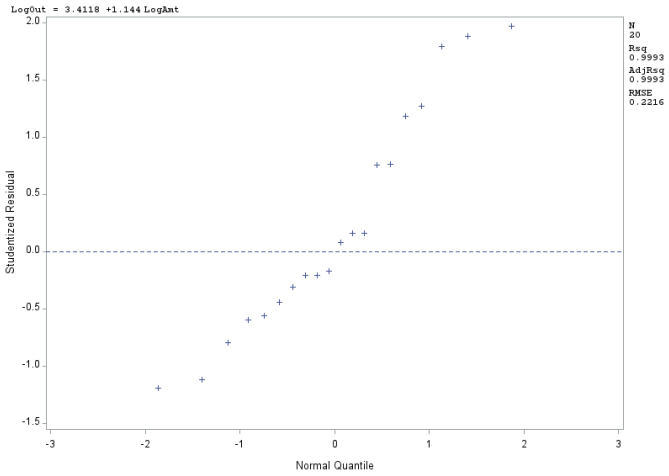
Root MSE	0.22155	**R-Square**	0.9993
Dependent Mean	5.43278	**Adj R-Sq**	0.9993
Coeff Var	4.07807		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.41177	0.01603	212.86	<.0001
LogAmt	1	1.14399	0.00718	159.25	<.0001

There hasn't been much change in the values of the estimates in this model but the root mean square error is much larger than before although the R^2 is also larger. [2 marks]

The useful diagnostic plots are





The spread of the studentized residuals at each predicted value is now much closer to constant. The normality of the residuals based on the weighted regression also seems more plausible than before. There is still a major issue with non-linearity however. [5 marks]

Report on the Analysis

In this analysis I fitted a linear model to predict the output of the gas chromatograph from the log known amounts of substance used. Both variables were transformed to the log scale. A standard analysis revealed that the variability of the output was much larger for the lowest level of amount and so the model assumptions were violated. To try to fix this problem, I estimated the variance of the residuals for each of the amounts and used a weighted least squares regression with weights equal to the total variance of the residuals divided by the variance for the amount used in each observation. The fitted weighted regression line was

$$\log(\text{output}) \;=\; 3.412 + 1.144 \times \log(\text{amount})$$

The intercept and slope in this model are both significantly different from 0 with both p-values less than 0.0001. The model explained 99.93% of the variability in the output of the gas chromatograph.

From the plot of the residuals of this model against the fitted values we see that the weighted regression did indeed alleviate the problem of non-constant variance. There is, however, a strong indication that a linear model is not appropriate here since the residuals for any given amount are generally all positive (under-estimation) or all negative (over-estimation). This suggests that the relationship is not linear and so either a different transformation is required or there is a problem with the calibration of the gas chromatograph. The normal quantile-quantile plot does not suggest any major departures from normality or outliers in the data.

[5 marks]