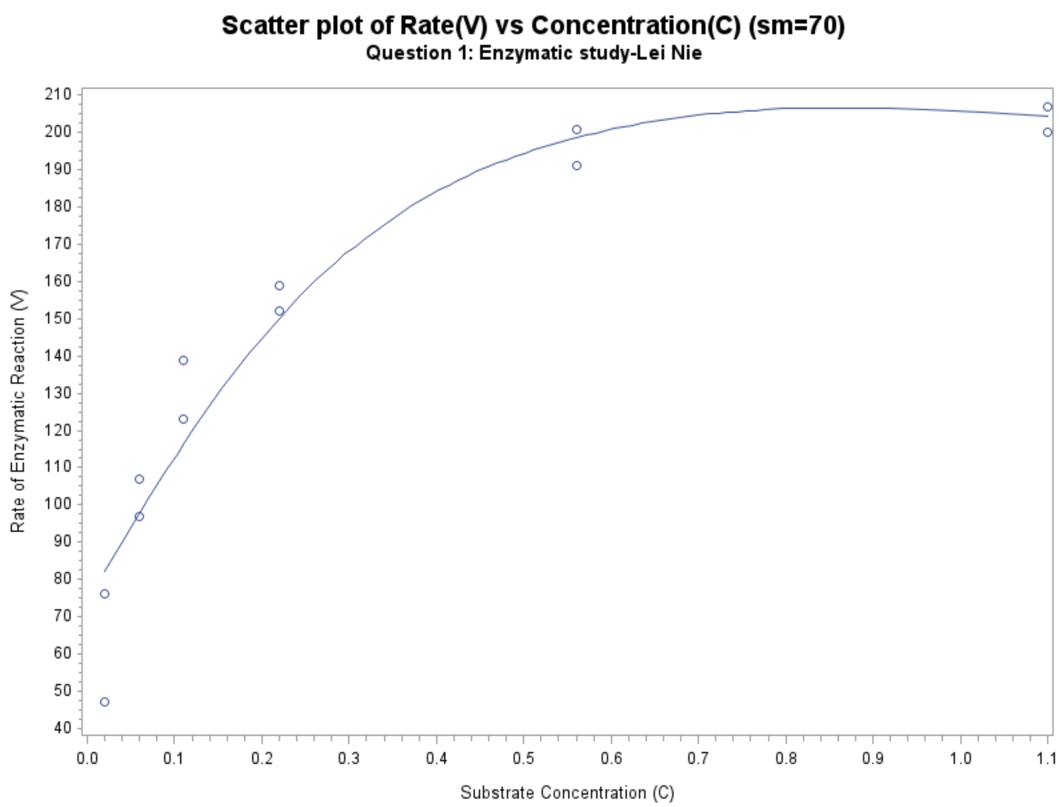
STAT 512: Homework 3

Name: Lei Nie

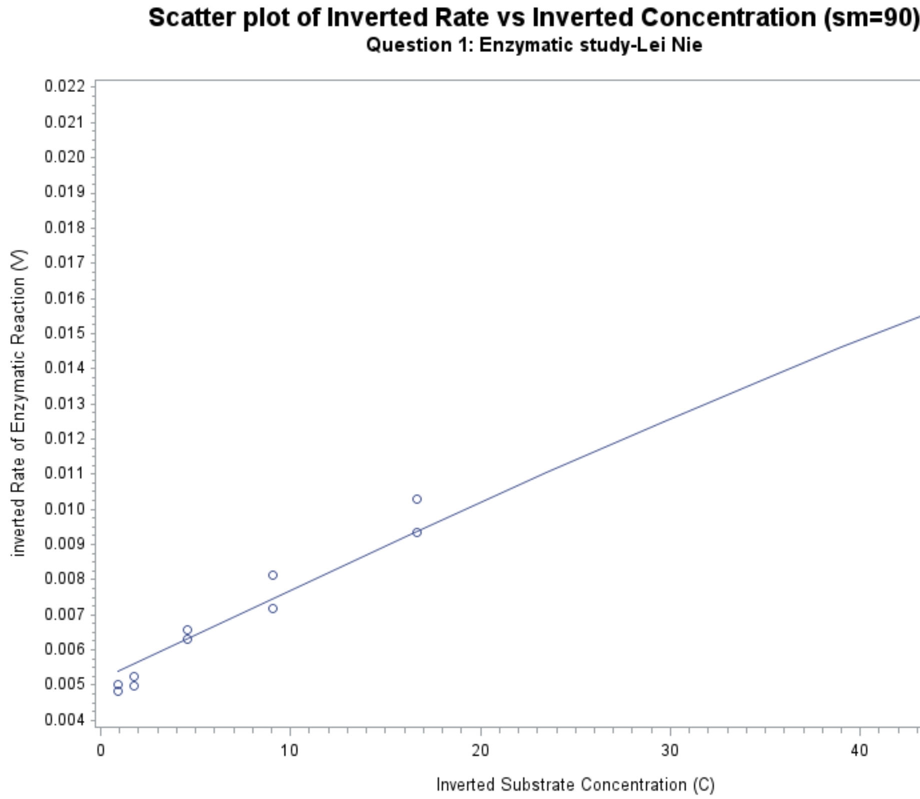
1. Consider the following data set that describes the relationship between the rate of an enzymatic reaction (V) and the substrate concentration (C). A common model used to describe the relationship between rate and concentration is the Michaelis-Menten model where, θ_1 is the maximum rate of the reaction and θ_2 describes how quickly the reaction will reach its maximum rate. With this mode, $\frac{1}{V}$ can be written as a linear model with explanatory variable.

   a. *Generate a scatterplot of V vs C. Comment on the shape.*



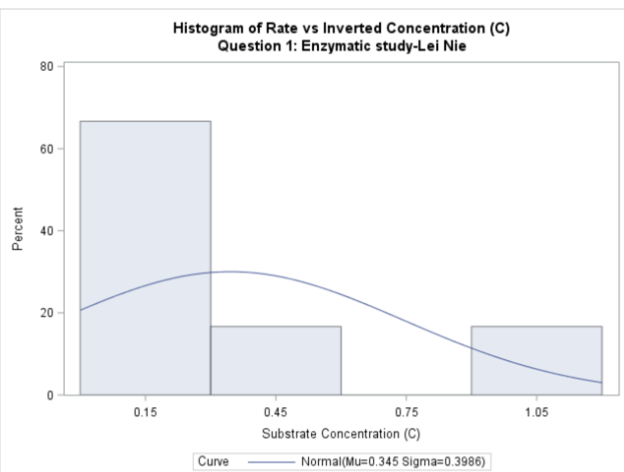
The shape of the plot does not look linear at all. V increases slower when C is large.

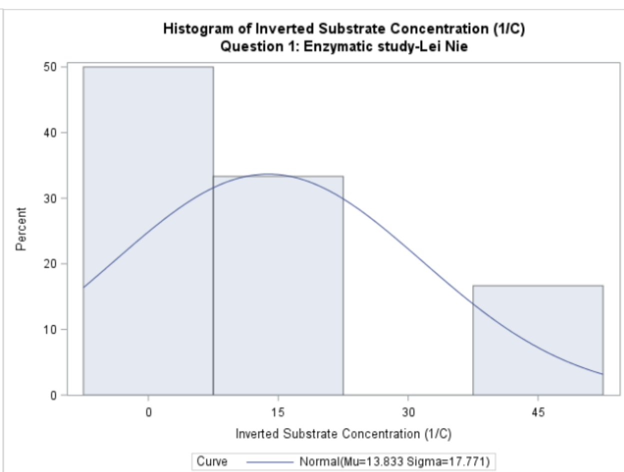
b. *Define new variables for $\frac{1}{V}$ and $\frac{1}{C}$ in SAS and generate a scatterplot of the new variables. Does the fit appear linear? Do any assumptions appear to be violated?*



The line seems to be linear. However, the constant variance assumption is violated since the difference of residuals are wider when $\frac{1}{C}$ is larger. When $\frac{1}{C} = 1$, the residuals are closer than those when $\frac{1}{C} = 50$.

c. *How is the distribution of C^{-1} different from the distribution of C? Are there any points that may be more influential in determining the fit?*

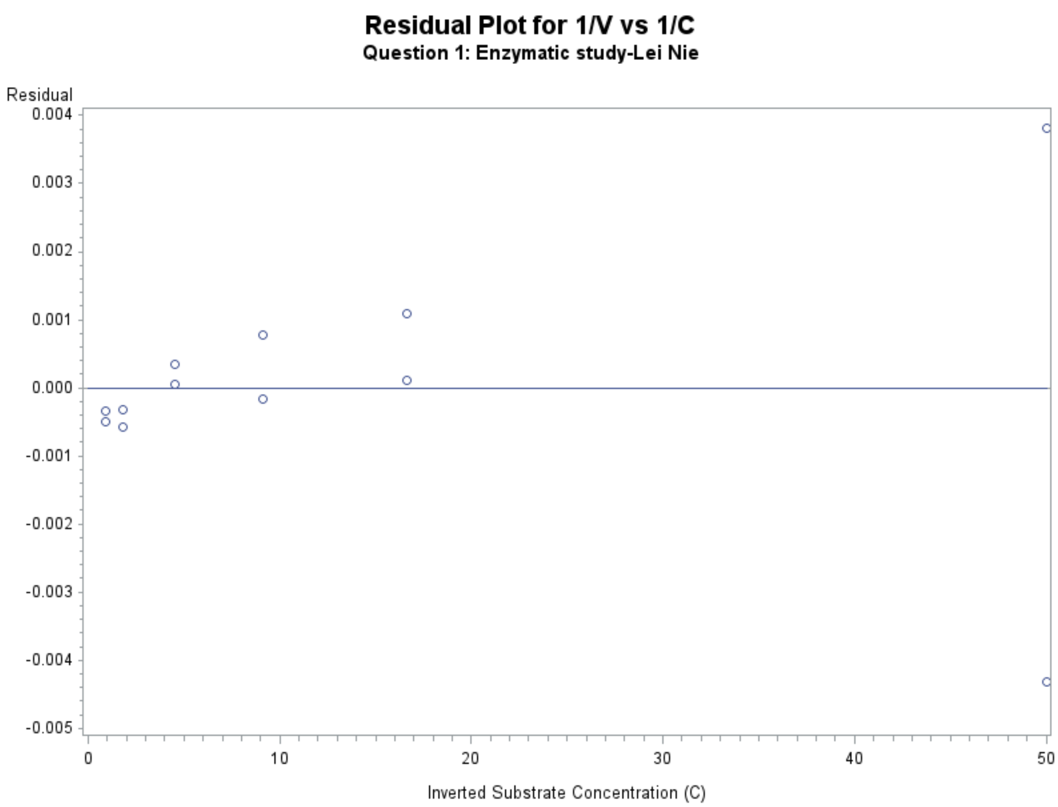




The distribution of $\boldsymbol{C}$ has a mean of 0.345, and standard deviation of 0.3986. The distribution of $\boldsymbol{C}^{-1}$ has a mean of 13.833, and standard deviation of 17.771. Both distributions are right-skewed, but the distribution of $\boldsymbol{C}$ is more skewed than that of $\boldsymbol{C}^{-1}$. There are two influential observations for both histogram. They are corresponding to the two extreme values with $\boldsymbol{C} = \mathbf{1.1}$ on the left plot, and $\boldsymbol{C}^{-1} = \mathbf{50}$ on the right plot. However, since $\boldsymbol{C}^{-1}$ has a smoother distribution than $\boldsymbol{C}$, the influence of two extreme observations is smaller on the distribution of $\boldsymbol{C}^{-1}$ than on that of $\boldsymbol{C}$.

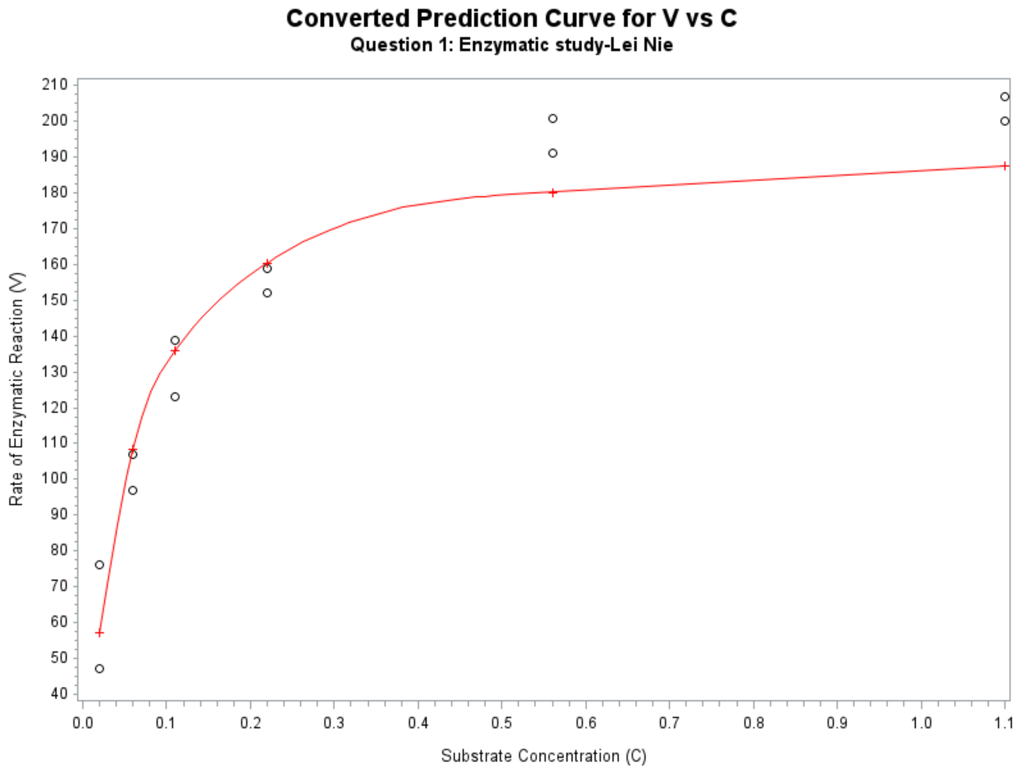
d. *Determine the least squares regression line for $\frac{1}{V}$ vs $\frac{1}{C}$. Save the residuals and predicted values. Does the residual plot suggest any problems?*

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.00511	0.00070400	7.25	<.0001
cinv	**1**	0.00024722	0.00003210	7.70	<.0001



The fitted model is $\widehat{\frac{1}{V}} = 0.00511 + 0.00024722 * \frac{1}{C}$. The residual plot shows violation to constant variance assumption.

e. *Convert this regression line back into the original nonlinear model and plot the predicted curve on a scatterplot of V vs C. Comment on the fit.*



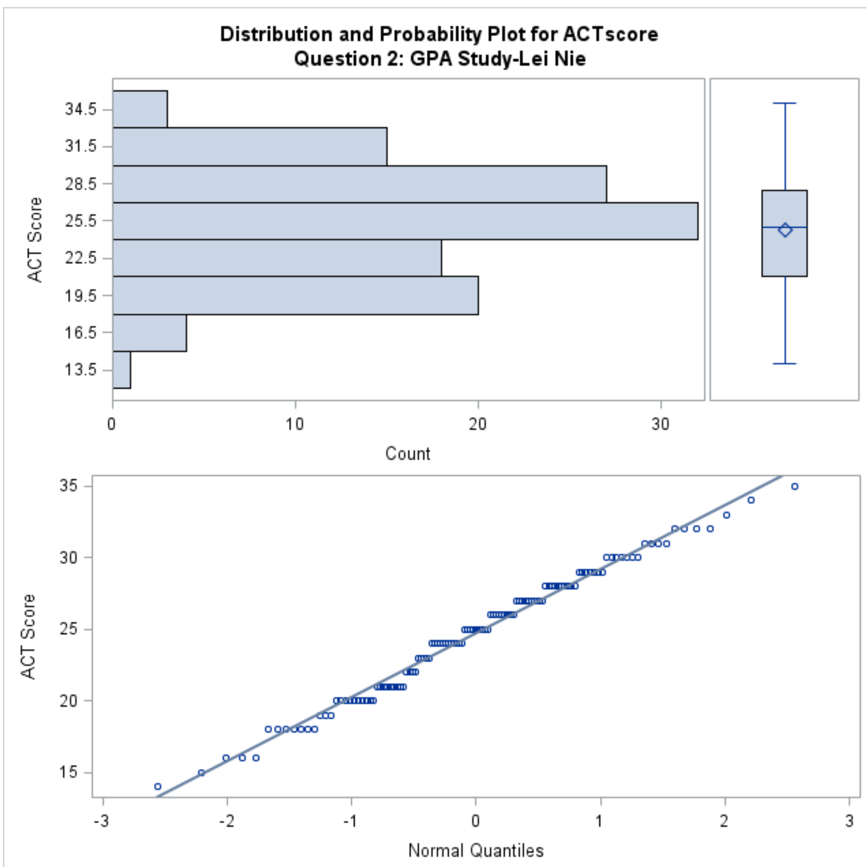
The prediction curve fits the original data pretty well when C is small. When C is large, the prediction value tends to smaller than the observation value.

2. *Describe the distribution of the explanatory variable. Show the plots and output that were helpful in learning about this variable.*

The UNIVARIATE Procedure
Variable: ACTscore (ACT Score)

Basic Statistical Measures			
Location		**Variability**	
Mean	24.72500	**Std Deviation**	4.47207
Median	25.00000	**Variance**	19.99937
Mode	24.00000	**Range**	21.00000
		Interquartile Range	7.00000

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
14	2	32	84
15	48	32	104
16	119	33	15
16	52	34	80
16	32	35	106



There are 120 observed ACT scores in total with range = 21, mean = 24.725, median = 25, standard deviation = 4.47207. There are no extreme influential observations based on the SAS output. The histogram and the box plot both show that the distribution of ACT scores is approximately symmetric and normal. The QQ plot shows the same trend.

3. Run the linear regression to predict GPA from the entrance test score and obtain the residuals (do not include a list of the residuals in your solution).

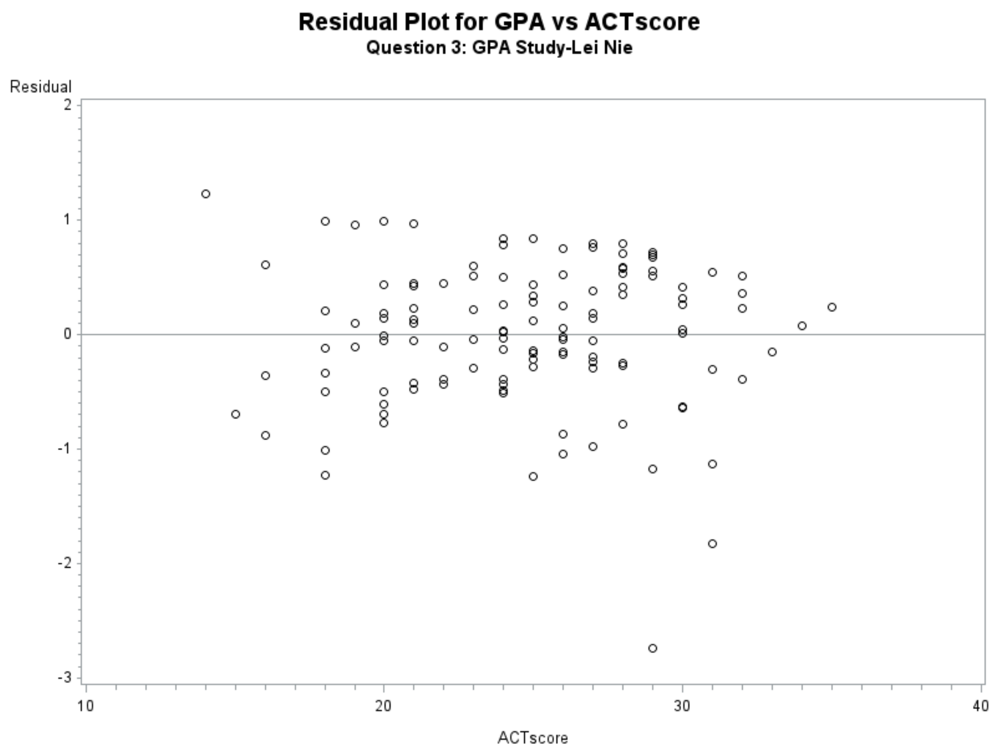
   a. *Verify that the sum of the residuals is zero by running proc univariate with the output from the regression.*

The UNIVARIATE Procedure
Variable: resid (Residual)

Moments			
N	120	**Sum Weights**	120
Mean	0	**Sum Observations**	0
Std Deviation	0.62050134	**Variance**	0.38502191
Skewness	-1.0067279	**Kurtosis**	2.50187662
Uncorrected SS	45.8176078	**Corrected SS**	45.8176078
Coeff Variation	.	**Std Error Mean**	0.05664376

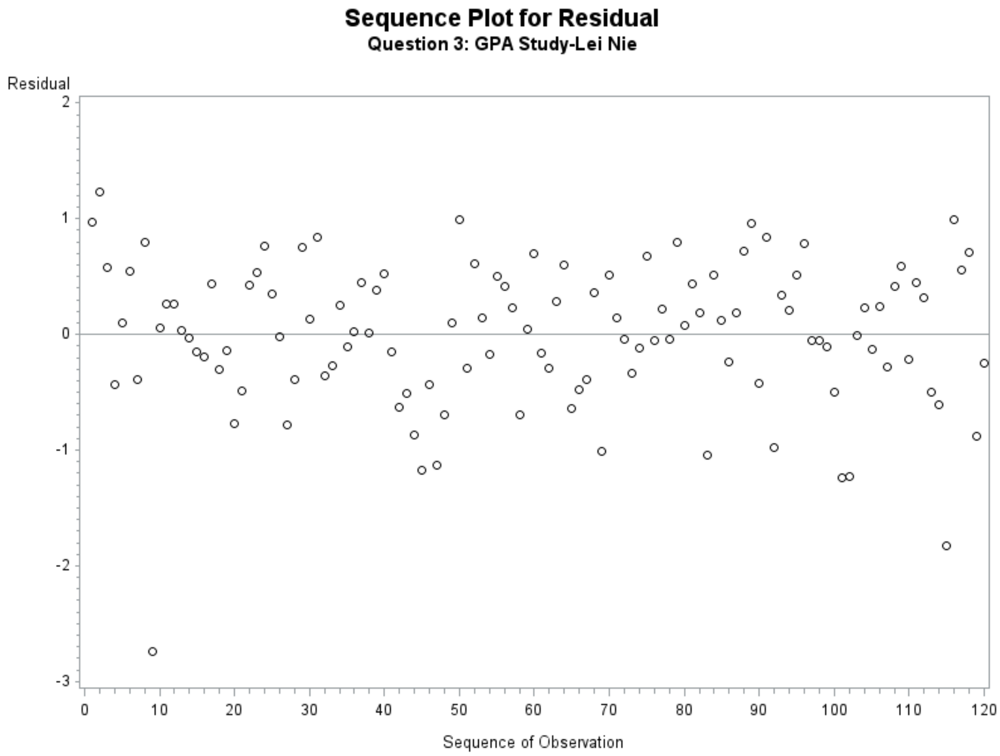
The residual sums to zero.

   b. *Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.*



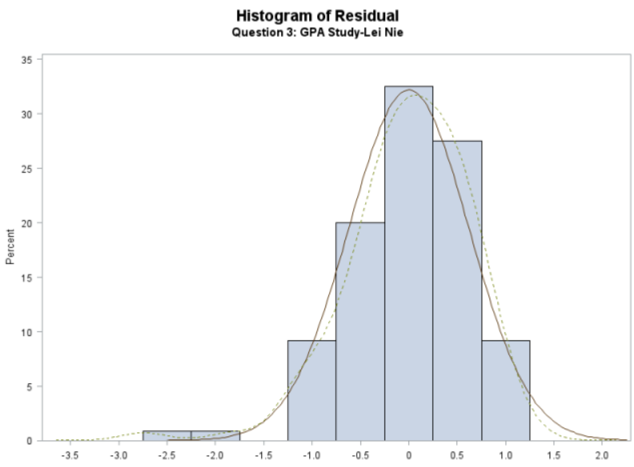
The residual plot doesn't show any obvious pattern or influential points.

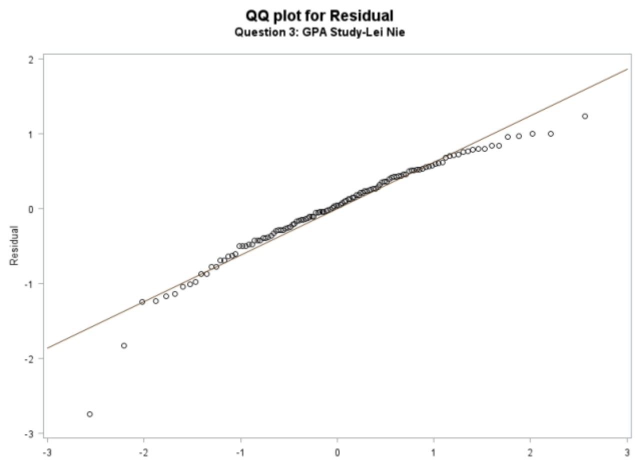
c. Plot the residuals versus the order in which the data appear in the data file and briefly describe the plot noting any unusual patterns or points.



The plot doesn't show any obvious pattern or influential points.

d. Examine the distribution of the residuals by getting a histogram and a normal probability plot of the residuals by using the histogram and qq-plot statements in proc univariate. What do you conclude?





The residual plot seems approximately normal and reasonably symmetric. The QQ plot seems reasonably linear.

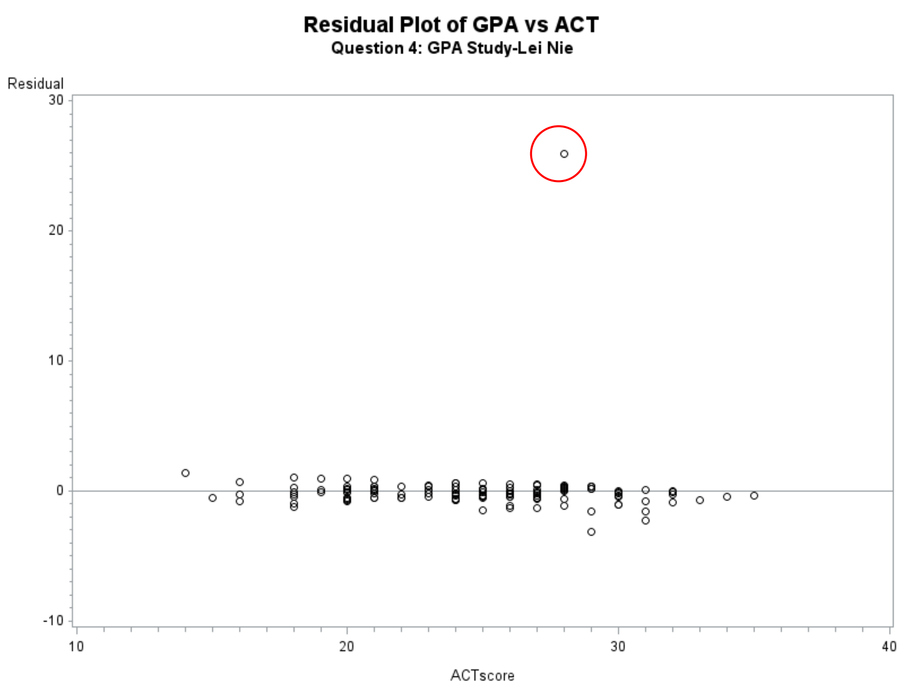
4. Change the data set by changing the value of the GPA for the last observation from 2.948 to 29.48 (e.g., a typo).

   a. *Make a table comparing the results of this analysis with the results of the analysis of the original data. Include in the table the following: fitted equation, t-test for the slope, with standard error and p-value, R^2, and the estimate of σ^2. Summarize the differences.*

	Original	**Typo**
fitted equation	$\hat{Y} = 2.114 + 0.0388X$	$\hat{Y} = 1.432 + 0.0753X$
t-test for the slope	3.04	1.48
Standard Error for the slope	0.0128	0.0509
P-value for the slope	0.0029 (Reject)	0.1414 (Fail to reject)
R^2	0.0726	0.0182
Estimate of σ^2	0.388	6.163

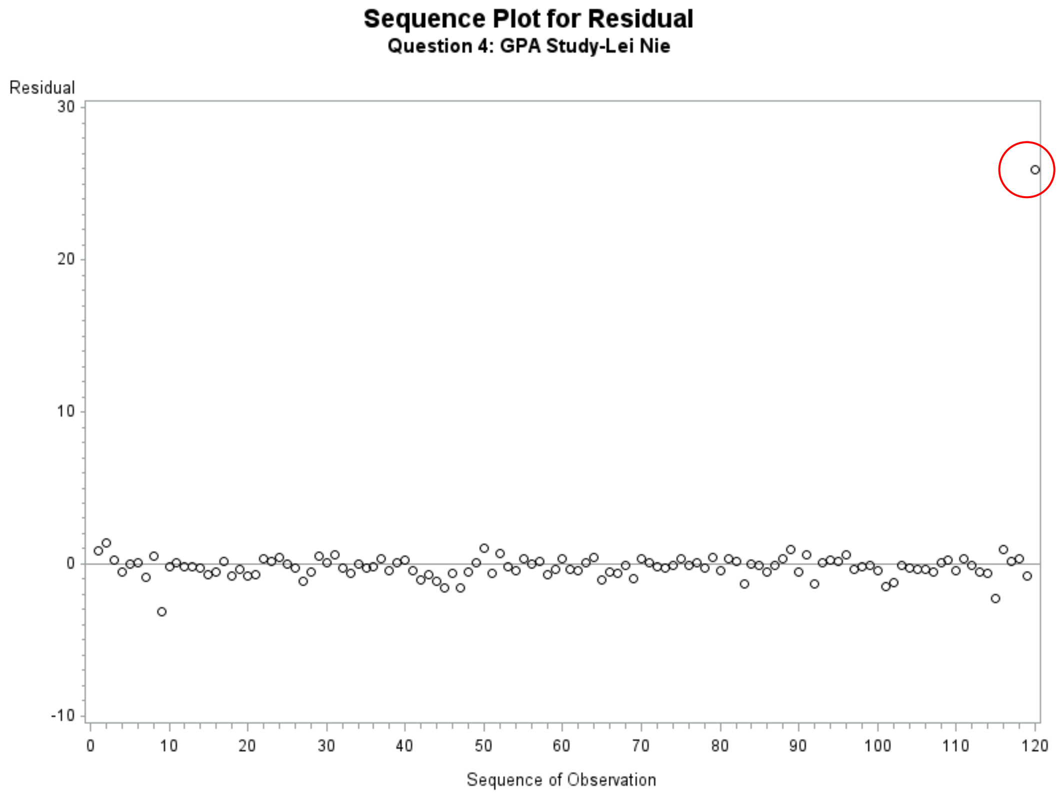
The outlier doubles the slope value and turns the slope from being significant to being insignificant. The R^2 and $\widehat{\sigma^2}$ change a lot after adding the outlier into the dataset. With the increase of the error variance, R^2 becomes very small.

   b. *Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.*



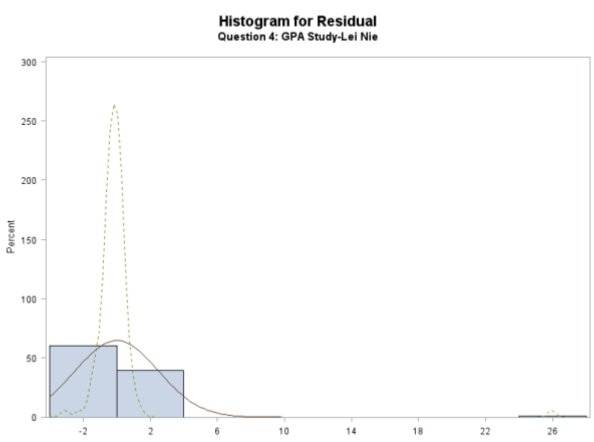
The existence of the outlier is pretty obvious.

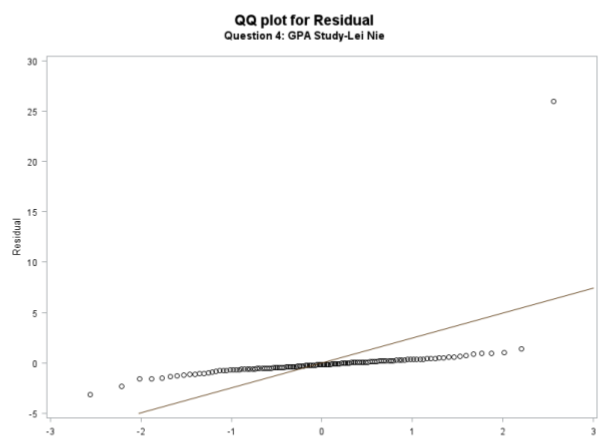
c. *Plot the residuals versus the order in which the data appear in the data file and briefly describe the plot noting any unusual patterns or points.*



The sequence plot shows that the last observation is the outlier identified in the residual plot above.

d. *Examine the distribution of the residuals by getting a histogram and a normal probability plot of the residuals by using the histogram and qq-plot statements in proc univariate. What do you conclude?*





The histogram and the QQ plot clearly shows a violation to the normality assumption.

Appendix: SAS code

```
*Question 1;
data data1;
input C V;
cards;
0.02 76
;
proc print data=data1;run;

*Question 1a;
title1 'Scatter plot of Rate(V) vs
Concentration(C) (sm=70)';
title2 'Question 1: Enzymatic
study-Lei Nie';
axis1 label=('Substrate
Concentration (C)');
axis2 label=(angle=90 'Rate of
Enzymatic Reaction (V)');
symbol1 v=circle i=sm70;
proc gplot data=data1;
plot V*C/haxis=axis1
vaxis=axis2;run;

*Question 1b;
data data1b;
set data1;
vinv=1/V;
cinv=1/C;run;
title1 'Scatter plot of Inverted
Rate vs Inverted Concentration
(sm=90)';
title2 'Question 1: Enzymatic
study-Lei Nie';
axis1 label=('Inverted Substrate
Concentration (C)');
axis2 label=(angle=90 'inverted
Rate of Enzymatic Reaction (V)');
symbol1 v=circle i=sm90;
proc gplot data=data1b;
plot vinv*cinv/haxis=axis1
vaxis=axis2;run;

*Question 1c;
ods graphics on;
title1 'Histogram of Substrate
Concentration (C)';
title2 'Question 1: Enzymatic
study-Lei Nie';
proc univariate data=data1;
var C;
histogram C/normal odstitle=title1
odstitle2=title2;
label C='Substrate Concentration
(C)';run;
title1 'Histogram of Inverted
Substrate Concentration (1/C)';
title2 'Question 1: Enzymatic
study-Lei Nie';
proc univariate data=data1b;
var cinv;
histogram cinv/normal
odstitle=title1 odstitle2=title2;
label cinv='Inverted Substrate
Concentration (1/C)';run;
ods graphics off;

*Question 1d;
proc reg data=data1b;
model vinv=cinv;
output out=out1 r=resid p=pred;run;
symbol1 v=circle i=rl;
title1 'Residual Plot for 1/V vs
1/C';
title2 'Question 1: Enzymatic
study-Lei Nie';
axis1 label=('Inverted Substrate
Concentration (C)');
proc gplot data=out1;
plot resid*cinv/vref=0
haxis=axis1;run;

*Question 1e;
data invert;
set out1;
predv = 1/pred;
symbol1 v = circle i = none c =
black;
symbol2 v = plus i = sm5 c = red;
title1 'Converted Prediction Curve
for V vs C';
title2 'Question 1: Enzymatic
study-Lei Nie';
axis1 label=('Substrate
Concentration (C)');
axis2 label=(angle=90 'Rate of
Enzymatic Reaction (V)');
proc gplot data = invert;
plot V*C predv*C /haxis=axis1
vaxis=axis2 overlay;run;


*Question 2/3/4;
data data2;
input GPA ACTscore;
seq=_n_;
datalines;
 3.897    21
```

```
 2.948    28
 ;
 proc print data=data2;run;

*Question 2;
ods graphics on;
proc univariate data=data2
plot(odstitle2='Question 2: GPA
Study-Lei Nie');
var ACTscore;
label ACTscore='ACT Score';run;
ods graphics off;

*Question 3a;
proc reg data=data2;
model GPA=ACTscore;
output out=out2 r=resid p=pred;run;
proc univariate data=out2;
var resid;run;

*Question 3b;
title1 'Residual Plot for GPA vs
ACTscore';
title2 'Question 3: GPA Study-Lei
Nie';
proc gplot data=out2;
plot resid*ACTscore / vref=0;run;


*Question 3c;
title1 'Sequence Plot for Residual';
title2 'Question 3: GPA Study-Lei
Nie';
axis1 label=('Sequence of
Observation');
proc gplot data=out2;
plot resid*seq/ haxis=axis1
vref=0;run;

*Question 3d;
title1 'QQ plot for Residual';
title2 'Question 3: GPA Study-Lei
Nie';
proc univariate data=out2;
var resid;
histogram resid / normal
kernel(L=2);
qqplot resid /normal (L=1 mu=est
sigma=est);run;


*Question 4;
data data4;
set data2;
if seq eq 120 then GPA = 29.48;
proc print data=data4;run;

*Question 4a;
proc reg data=data4;
model GPA=ACTscore;
output out=out4 r=resid p=pred;run;

*Question 4b;
title1 'Residual Plot of GPA vs
ACT';
title2 'Question 4: GPA Study-Lei
Nie';
proc gplot data=out4;
plot resid*ACTscore / vref=0;run;

*Question 4c;
title1 'Sequence Plot for Residual';
title2 'Question 4: GPA Study-Lei
Nie';
axis1 label=('Sequence of
Observation');
proc gplot data=out4;
plot resid*seq/ haxis=axis1
vref=0;run;

*Question 4d;
title1 'QQ plot for Residual';
title2 'Question 4: GPA Study-Lei
Nie';
proc univariate data=out4;
var resid;
histogram resid / normal
kernel(L=2);
qqplot resid /normal (L=1 mu=est
sigma=est);run;
```