Statistics 512: Homework 3

Solutions

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 $V = \frac{\theta_1 C}{\theta_2 + C}$, 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 $\frac{1}{C}$:

$$\frac{1}{V} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{C}$$

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

   Solution: The relationship between V and C does not appear to be linear. V increases rapidly for small values of C but tends to level off as C approaches 1. See Figure 1.

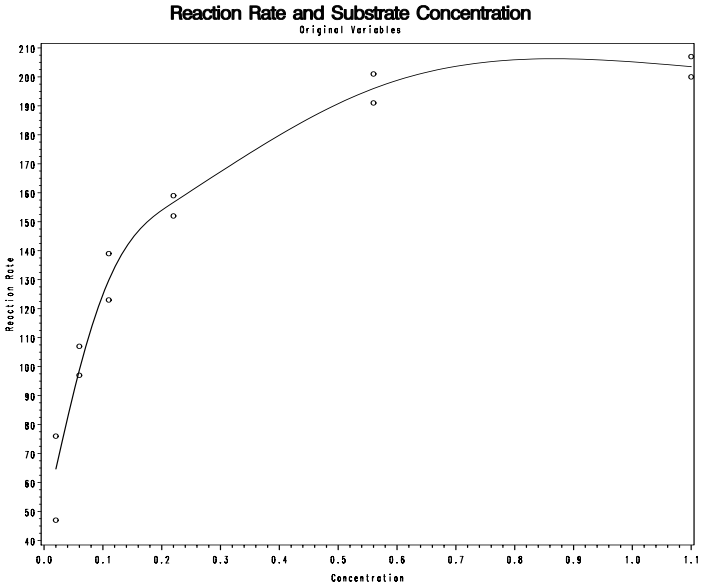

Figure 1: Scatterplot for Problem 1a

   (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?

   Solution: The relationship between $1/V$ and $1/C$ appears to be linear. However, the assumption of constant variance seems to be violated, since the observations with $1/C = 50$ have much larger residuals than do the other observations. See Figure 2.

   (c) How is the distribution of $\frac{1}{C}$ different from the distribution of C? Are there any points that may be more influential in determining the fit?

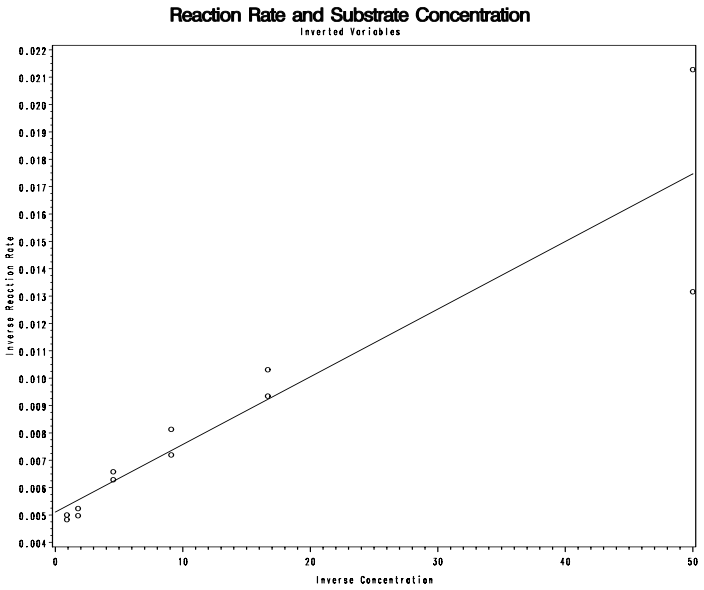

Figure 2: Scatterplot for Problem 1b

Solution: The variable C has mean 0.345, median 0.0165, and standard deviation 0.4, with values ranging from 0.02 and 1.1 . The variable $1/C$ has mean 13.83, median 6.82, and standard deviation 17.8, with values ranging from 0.91 to 50. Neither variable seems to come from a normal distribution. The values for $1/C$ show a large gap in their distribution, with almost all the values between approximately 1 and 17, and then two values of 50. The values for C show a similar and less extreme gap, with all but two of the values between 0.02 and 0.06, and then two values of 1.1 . The points with $1/C = 50$ will likely be very influential in the fit of the straight line, since they are so far away from the others.

The UNIVARIATE Procedure
Variable: c
Basic Statistical Measures

Location		Variability	
Mean	0.345000	Std Deviation	0.39862
Median	0.165000	Variance	0.15890
Mode	0.020000	Range	1.08000
		Interquartile Range	0.50000

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
0.02	2	0.22	8
0.02	1	0.56	9
0.06	4	0.56	10
0.06	3	1.10	11
0.11	6	1.10	12

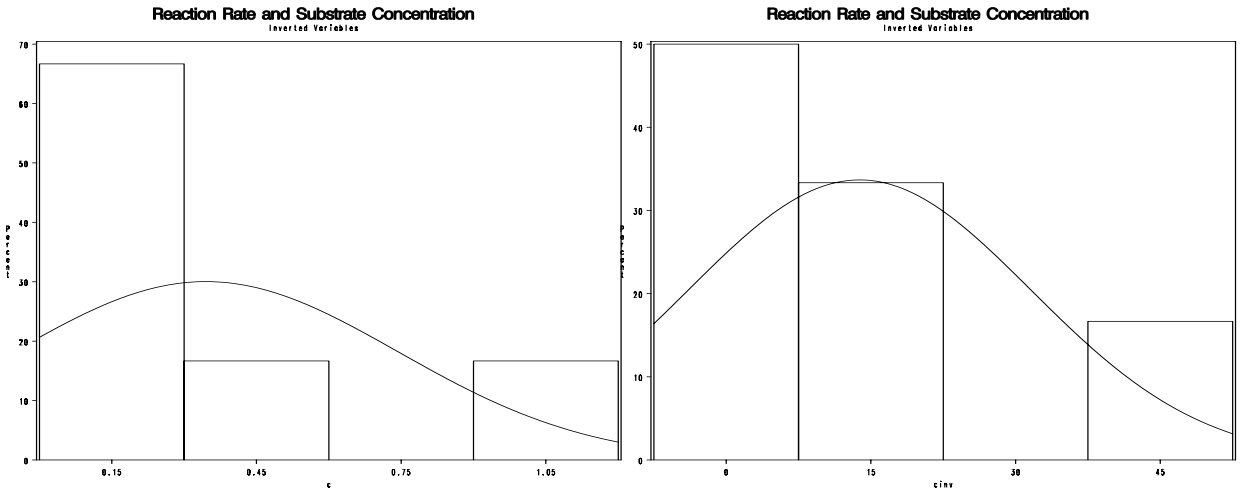

Figure 3: Histograms for Problem 1c

The UNIVARIATE Procedure
Variable: cinv
Basic Statistical Measures

Location		Variability	
Mean	13.83297	Std Deviation	17.77123
Median	6.81818	Variance	315.81673
Mode	0.90909	Range	49.09091
		Interquartile Range	14.88095

Extreme Observations

------Lowest------		------Highest-----	
Value	Obs	Value	Obs
0.909091	12	9.09091	6
0.909091	11	16.66667	3
1.785714	10	16.66667	4
1.785714	9	50.00000	1
4.545455	8	50.00000	2

(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?

Solution: The least squares line is $1\hat{/}V = 0.00511 + 0.000247(1/C)$. The residual plot (Figure 3) suggests that the variance is not constant (heterscedasticity).

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

(e) Convert this regression line back into the original nonlinear model and plot the predicted

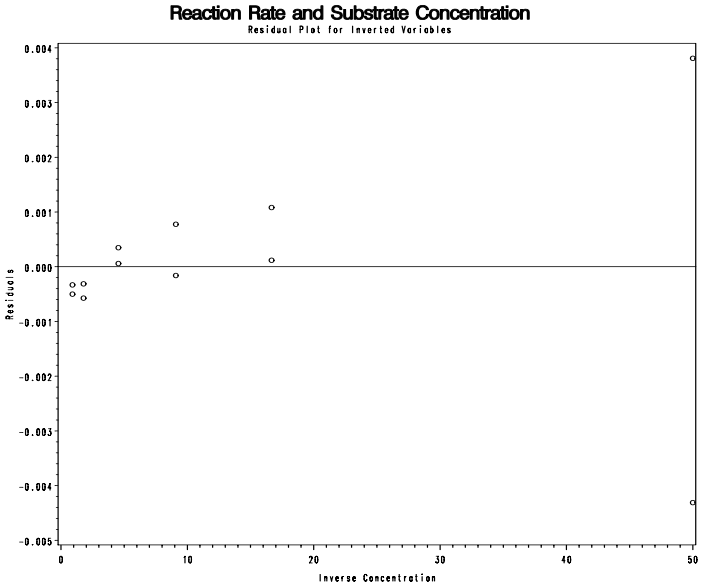

Figure 4: Residual Plot for Problem 1d

curve on a scatterplot of V vs C. Comment on the fit.

Solution: The predicted line fits the data pretty well for small C but seriously underestimates V for large values of C. The highly influential points with $C = 0.02$ $(1/C = 50)$ are bringing the line down substantially. See Figure 5.

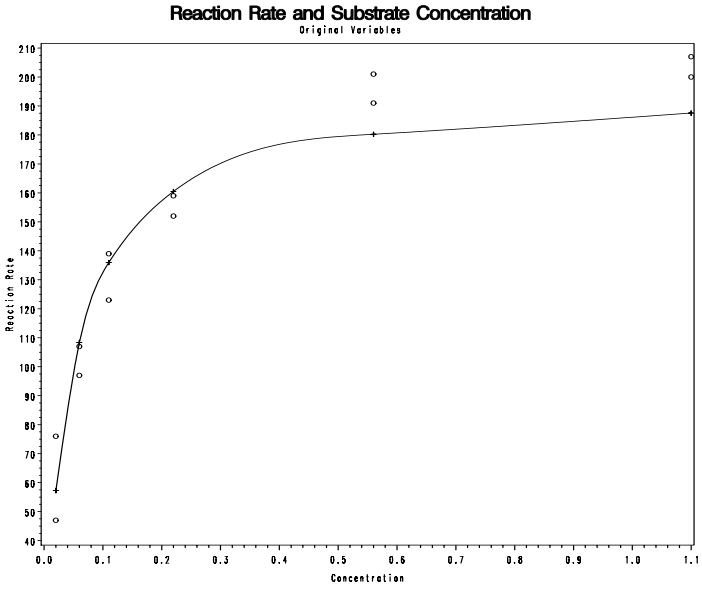

Figure 5: Scatterplot for Problem 1e

For the next 3 questions, use the grade point average data described in the text with Problem 1.19 .

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

Solution: Using `proc univariate`, we see there are 120 observations ranging between 14 and 35 with a mean of 24.725 and median of 24; their standard deviation is 4.472 . There are no extreme observations (i.e., ones far away from the others) as shown in the histogram (Figure 6). The distribution appears to be reasonably symmetric but not completely so; it has a slight skew to the right.

```
                    The UNIVARIATE Procedure
                      Variable:  testscore
                             Moments

N                          120    Sum Weights               120
Mean                    24.725    Sum Observations         2967
Std Deviation      4.47206549    Variance            19.9993697
Skewness           -0.1363553    Kurtosis            -0.5596968
Uncorrected SS          75739    Corrected SS          2379.925

Coeff Variation    18.0872214    Std Error Mean      0.40824186


                  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


      Stem Leaf                     #             Boxplot
        35 0                        1                |
        34 0                        1                |
        33 0                        1                |
        32 0000                     4                |
        31 0000                     4                |
        30 0000000                  7                |
        29 0000000                  7                |
        28 0000000000              10             +-----+
        27 0000000000              10             |     |
        26 0000000000              10             |     |
        25 0000000000              10             *-----*
        24 000000000000            12             |  +  |
```

```
23 00000                    5          |      |
22 0000                     4          |      |
21 000000000                9          +------+
20 0000000000              10             |
19 000                      3             |
18 0000000                  7             |
17                                        |
16 000                      3             |
15 0                        1             |
14 0                        1             |
   ----+----+----+----+
```

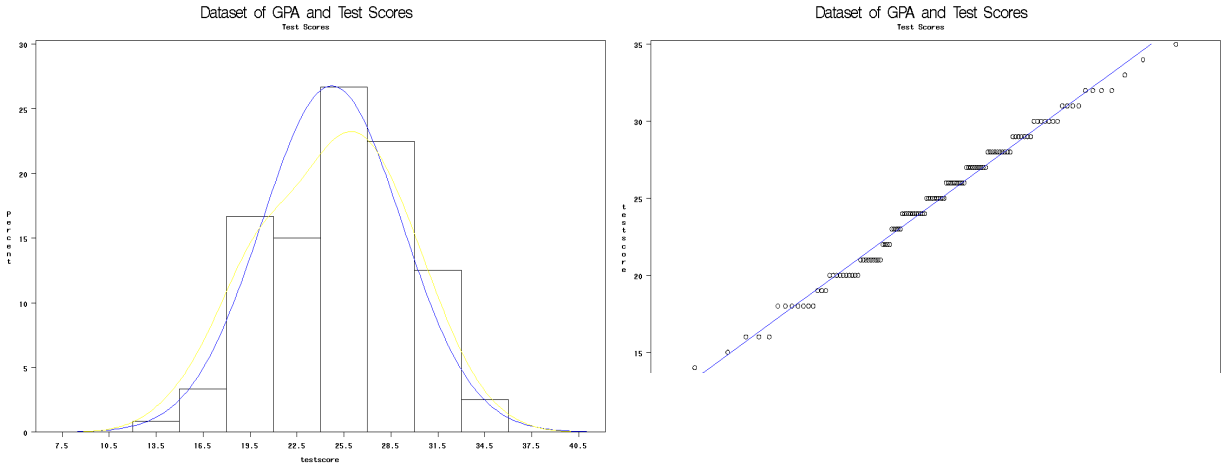

Figure 6: Graphs for Problem 2

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.

   Solution: The given `proc univariate` output shows that the residuals sum to zero.

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

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

Solution: There does not appear to be any obvious pattern or outlier in this residual plot (Figure 7). It looks like a random scatter of points, and the variance is reasonably constant.

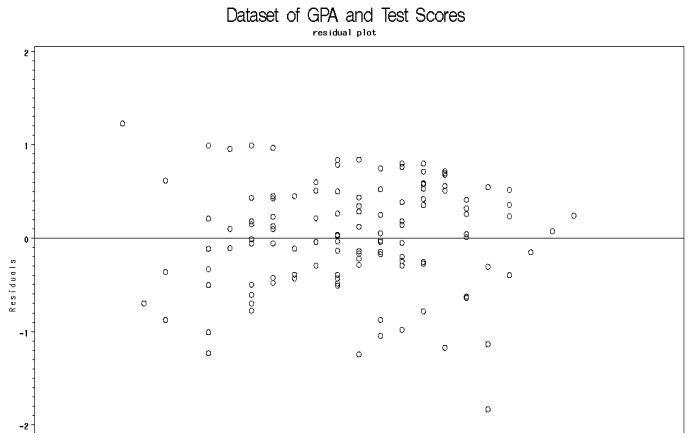

Figure 7: Scatterplot for Problem 3b

(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.

Solution: There is no obvious pattern over time. See Figure 8.

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

Solution: The residuals appear reasonably normal if somewhat asymmetric, since the histogram appears fairly normal, and the qqplot is fairly linear (Figure 9). There is some suggestion of a concave down shape to the qq-plot, but it is not too bad.

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). You can do this in a data step. For example, `data a2; set a1; if _n_ eq 120 then gpa = 29.48`; an alternative is simply to edit the data file.

(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.

Solution: The outlier has a huge impact on these results. The slope becomes nearly double its original value and is no longer significantly different from zero. The changes in R^2 and s^2 are also very extreme. The outlier greatly inflates the estimated variance and makes the R^2 almost zero.

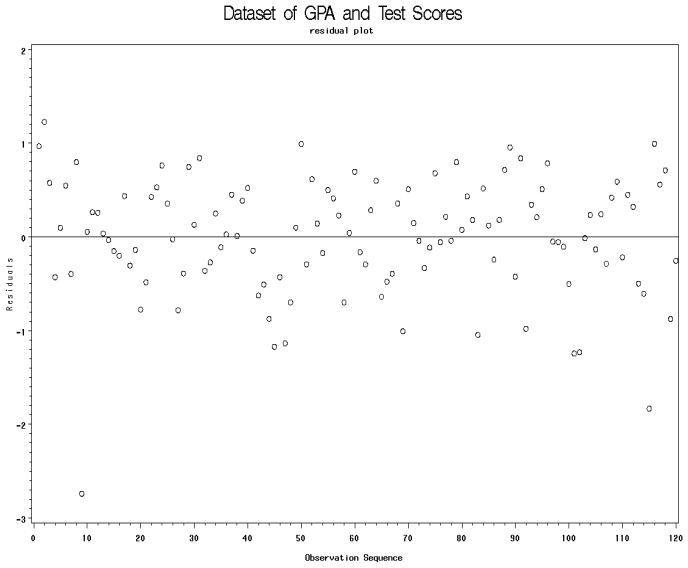

Figure 8: Scatterplot for Problem 3c

Result	*Original*	*With Outlier*
Equation	$Y = 2.114 + 0.0388X$	$Y = 1.432 + 0.0753X$
t-test	$t = 3.04$	$t = 1.48$
SE	0.0133	0.0509
p-value	0.0029	0.1414
Conclusion	Reject H_0	Do not reject H_0
R^2	0.0726	0.0182
s^2	0.388	6.163

GPA and Test Scores Original:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.50789	13.50789	2.19	0.1414
Error	118	727.28965	6.16347		
Corrected Total	119	740.79754			

Root MSE	2.48263	R-Square	0.0182
Dependent Mean	3.29515	Adj R-Sq	0.0099
Coeff Var	75.34206		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.43243	1.27850	1.12	0.2648
testscore	1	0.07534	0.05089	1.48	0.1414

GPA and Test Scores with Outlier:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.50789	13.50789	2.19	0.1414

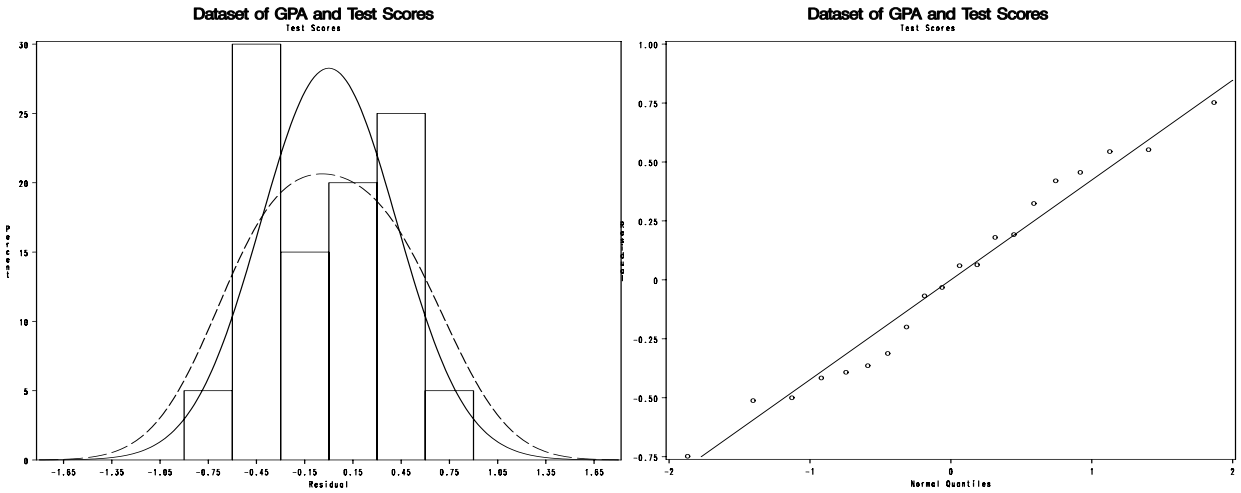

Figure 9: Graphs for Problem 3d

Error	118	727.28965	6.16347	
Corrected Total	119	740.79754		

Root MSE	2.48263	R-Square	0.0182
Dependent Mean	3.29515	Adj R-Sq	0.0099
Coeff Var	75.34206		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.43243	1.27850	1.12	0.2648
testscore	1	0.07534	0.05089	1.48	0.1414

(b) Repeat parts (b), (c), and (d) from the previous problem and explain how these plots help you to detect the unusual observation.

Solution: The outlier is quite noticeable on the residual plots as a point far away from all the others (Figure 10 left). The sequence plot (Figure 10 right) additionally shows that the outlier is that last observation in the list. The qqplot (Figure 11 right) shows that the data do not fit on the straight line with the estimated mean and standard deviation, and the outlier again appears quite separate from all the others. The histogram (Figure 11 left) shows a distribution that is far from normal, since there is an observation very far out in the tail. All together, these plots clearly show there is one outlying point which should be investigated. (Since GPA cannot be 29.48, this would appear to be a data entry error.)

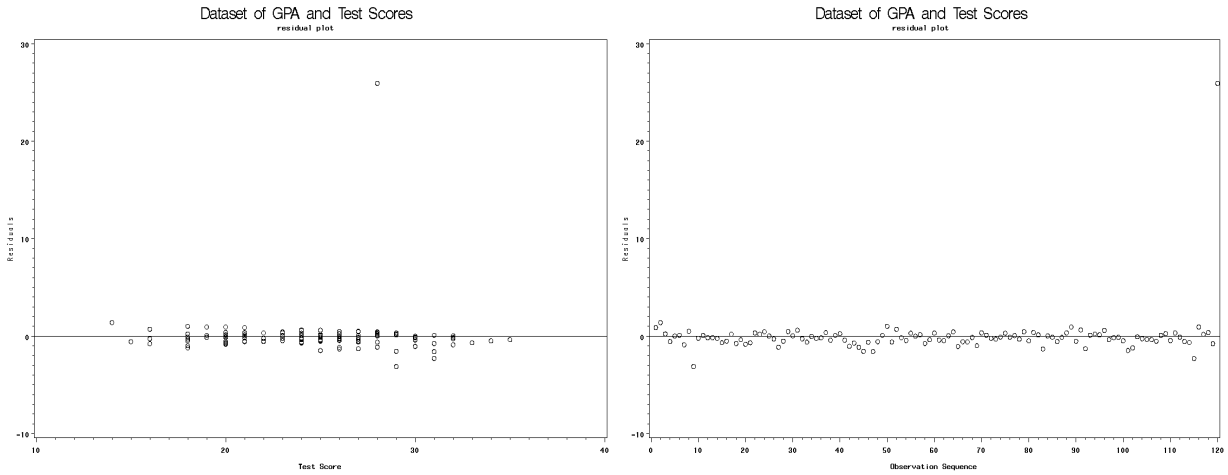

Figure 10: Graphs for Problem 4b

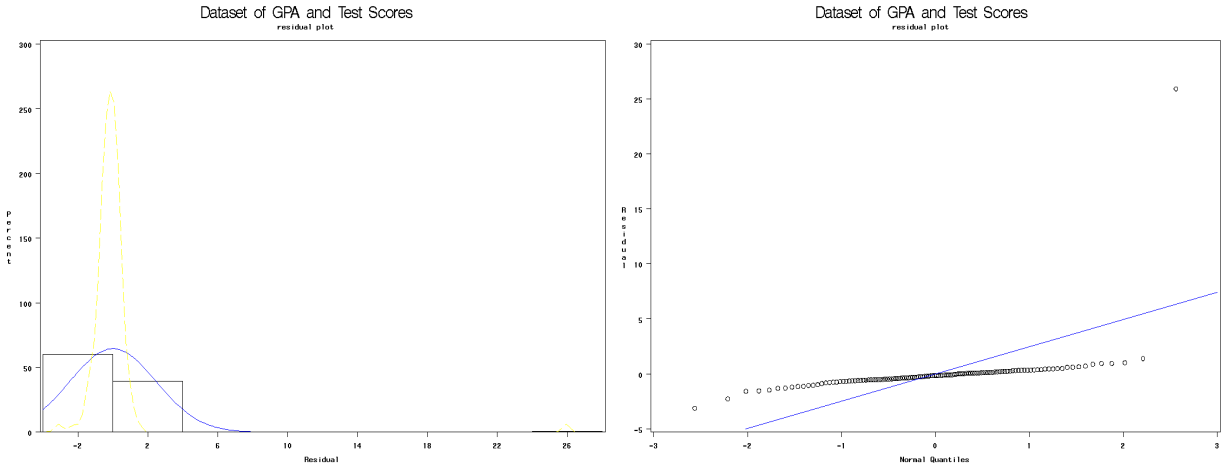

Figure 11: Graphs for Problem 4b