

Stat 346 Homework 3

Nathan Jarus

Feb. 20, 2014

1 Problem 1

1.1 Part a

The residuals are not normally distributed, indicating a nonlinear relationship. Possible fixes include non-linear regression or a transformation on X.

1.2 Part b

These residuals seem to be normally distributed with constant variance. This indicates that the model is a good fit.

1.3 Part c

The errors are not normally distributed. This might be fixed with a transformation on Y.

1.4 Part d

The variance on the residuals is not constant. This might be fixed with a transformation on Y.

2 Problem 2

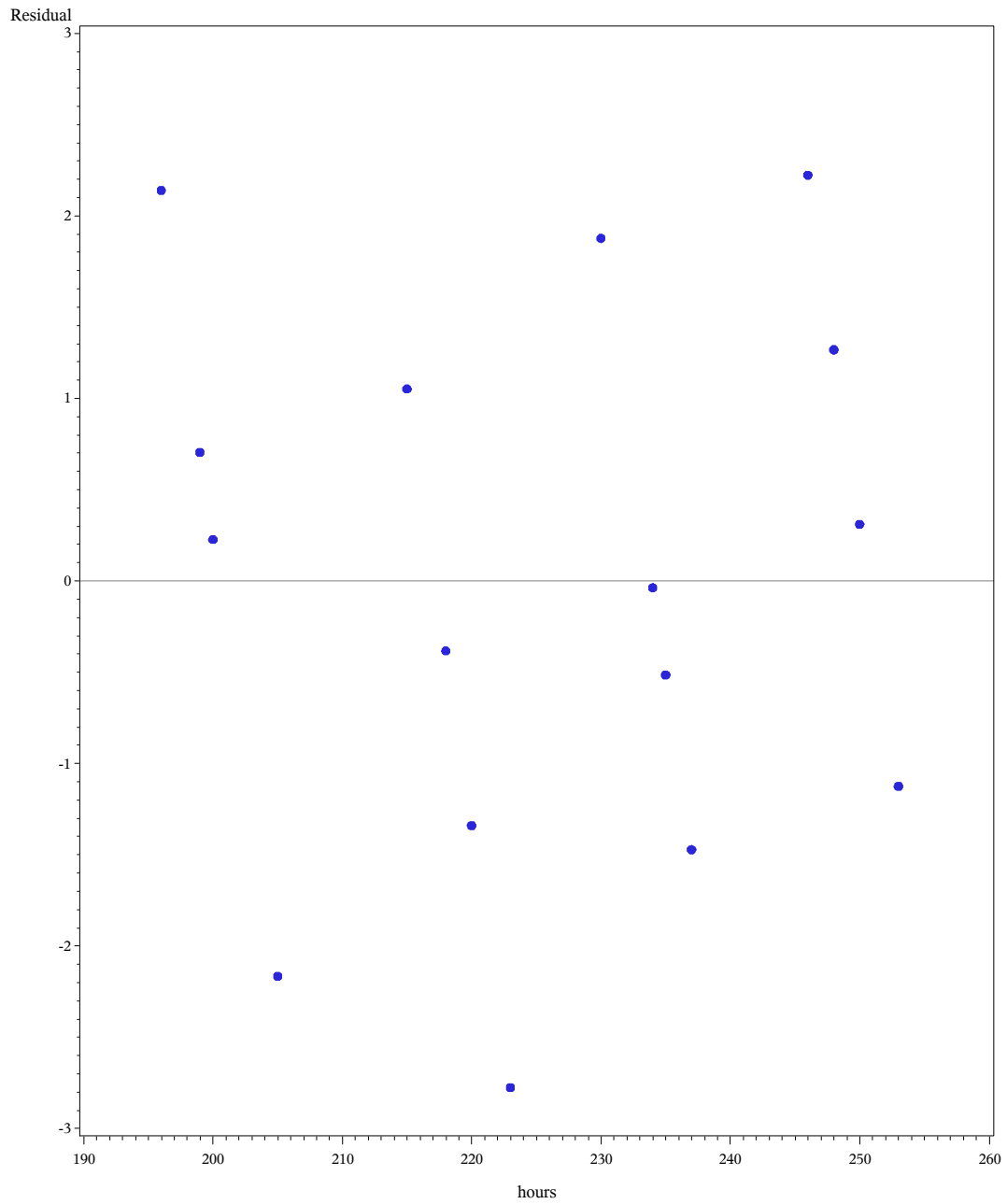
2.1 Part a

Plots 2.1 and 2.1. The residuals seem normally distributed and support a linear trend.

2.2 Part b

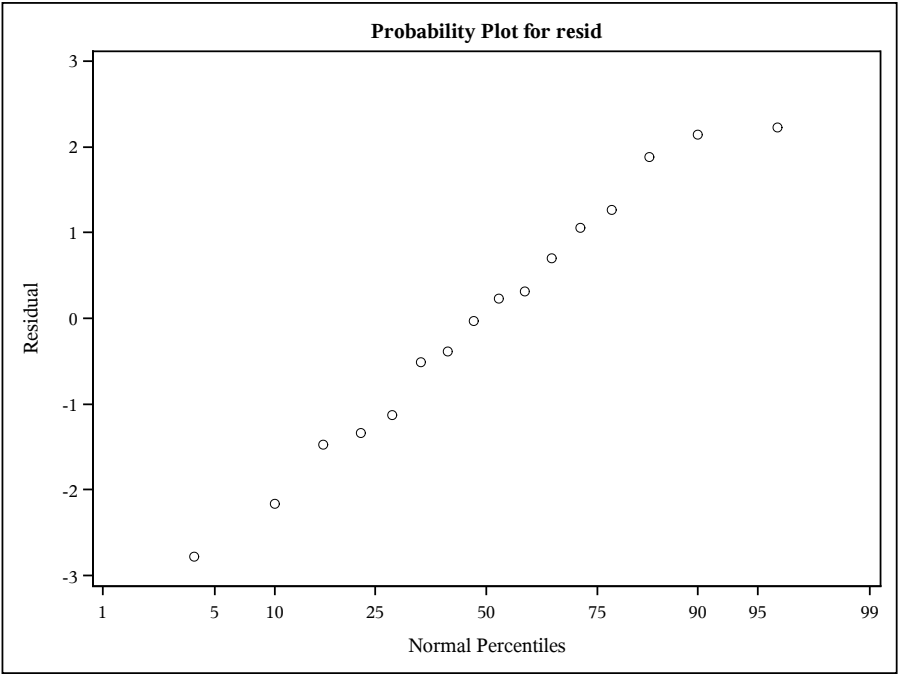
$$\begin{aligned}s\{b_1\} &= 2.125 \times 10^{-2} \\ t(1 - 0.05/2, 16 - 2) &= t(0.975, 14) = 2.145 \\ b_1 &= 0.47833 \pm 2.145 * 2.125 \times 10^{-2} \\ &= (0.43275, 0.52392)\end{aligned}$$

00:27 Thursday, February 20, 2014 1

Plastic Hardness
Residuals vs. time

Plastic Hardness
Normal Probability

00:27 Thursday, February 20, 2014 1



2.3 Part c

$$\hat{Y} = 37.7759$$

$$s\{\hat{Y}\} = 0.5851$$

$$\hat{Y} \pm t(1 - 0.05/2, 16 - 2)s\{\hat{Y}\} = 2.145 * 0.5851$$

$$(36.5209, 39.0309)$$

2.4 Part d

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{34.342802}{1280} = 0.9731$$

2.5 Part e

1. General Linear Test
2. ANOVA
3. T-test

3 Problem 3

3.1 Part a

	ORIGINAL REGRESSION	REGRESSION WITH OUTLIER
Fitted Regression Equation	$\hat{Y} = 2.11405 + 0.03883X$	$\hat{Y} = 3.04977 + 0.00090502X$
R-Square	0.0726	0.0011
MSE	0.38829	0.41822
$SE\{b_1\}$	0.01277	0.00250
P-Value	0.0029	0.7178

3.2 Part b

Plots 3.2 and 3.2 for the original dataset. Plots 3.2 and 3.2 for the edited dataset. The residuals plot is much more extreme.

3.3 Part c

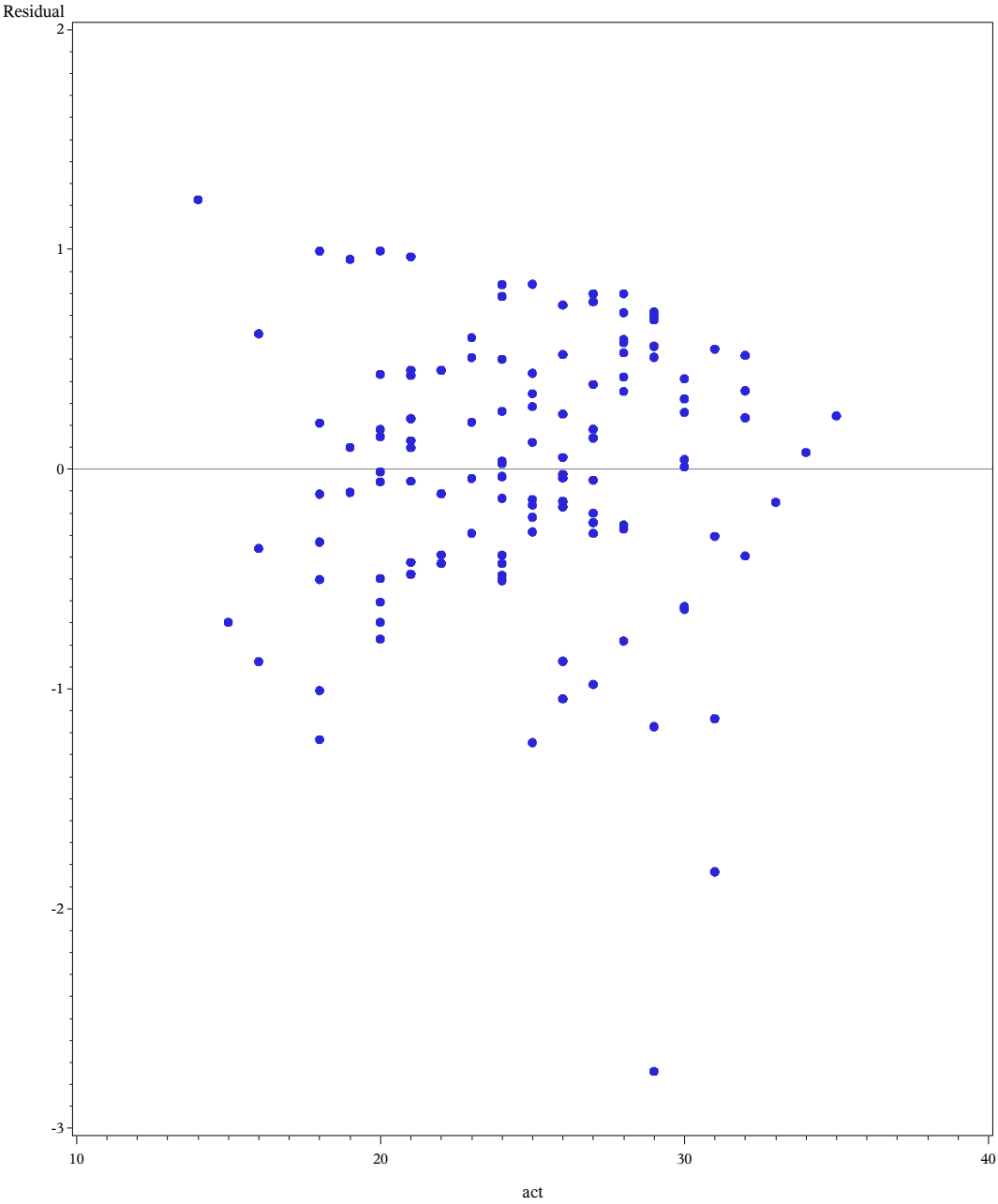
Yes, you could construct a sequence diagram over the dates, or perhaps school years, at which the GPA and ACT scores were collected.

4 Problem 4

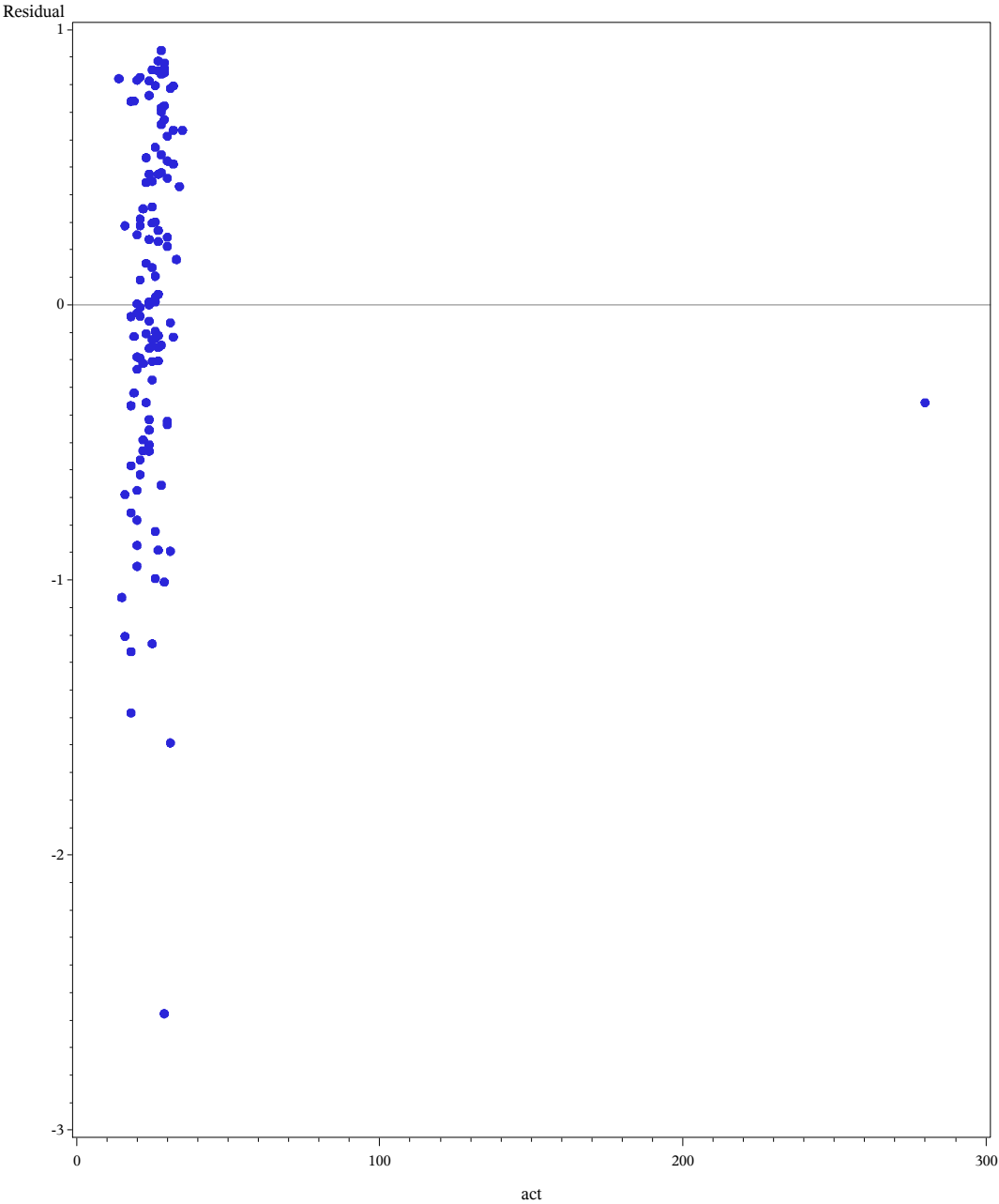
4.1 Part a

As shown in Plot 4.1, the relationship is not linear.

GPA and ACT
Residuals vs. time

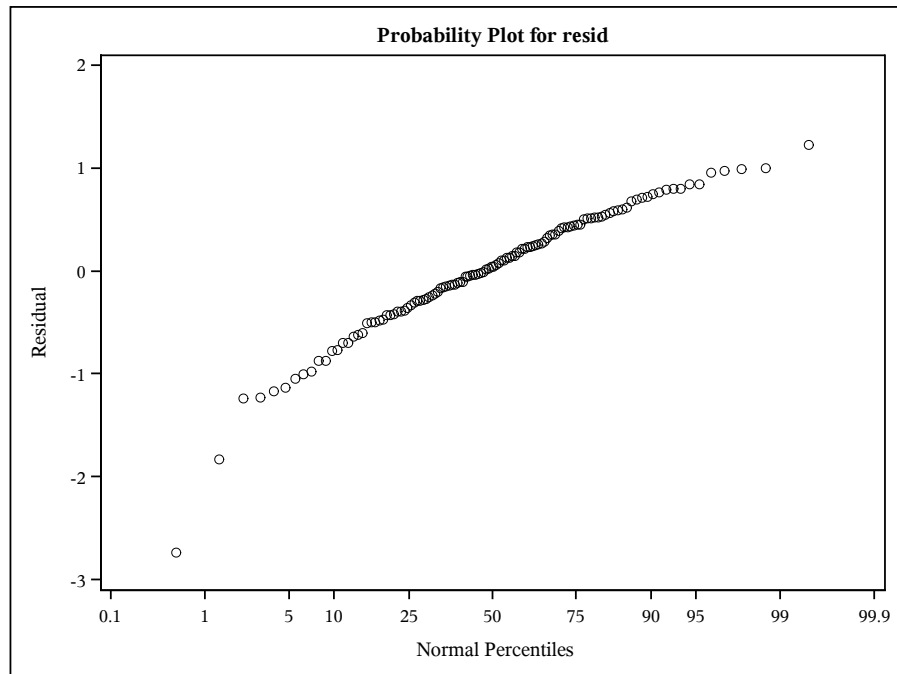


GPA and ACT with typo
Residuals vs. time



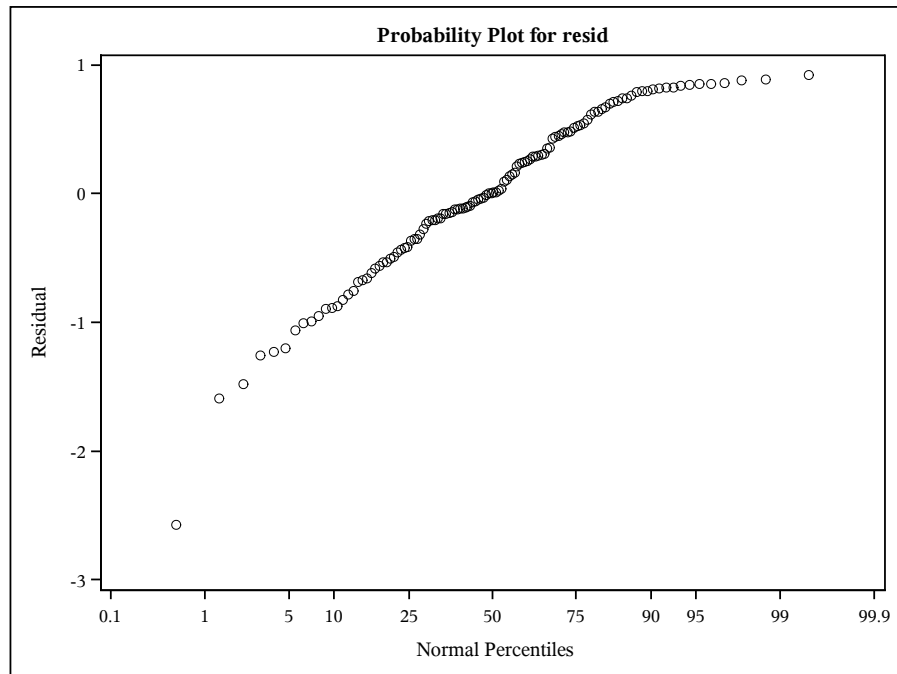
GPA and ACT
Normal Probability

01:31 Thursday, February 20, 2014 1



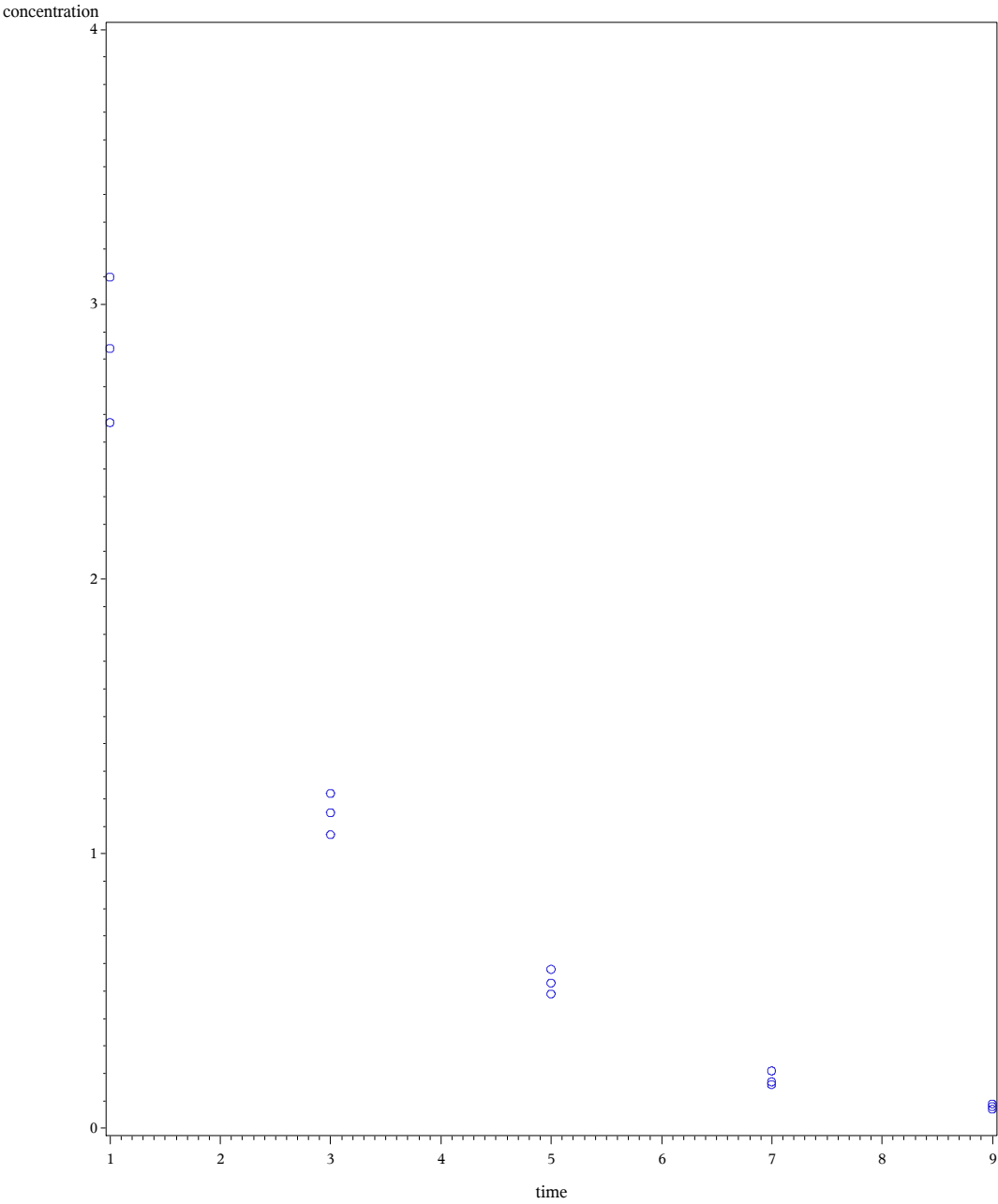
***GPA and ACT with typo
Normal Probability***

01:31 Thursday, February 20, 2014 1



02:35 Thursday, February 20, 2014 1

Solution Concentration
Scatterplot Concentration vs. Time



4.2 Part b

In the plots in 4.2 and 4.2, we see a large skew towards smaller times. The boxplot especially shows a very nonlinear relationship.

4.3 Part c

The regression equation is

$$\hat{Y} = 2.57533 + 0.324X \quad (1)$$

The p value is < 0.0001 , indicating a significant linear relationship if the SLR assumptions hold. The correlation is 0.8116, which is pretty good.

Solution Concentration

Box Plot

The REG Procedure

Model: MODEL1

Dependent Variable: concentration

Number of Observations Read	15
Number of Observations Used	15

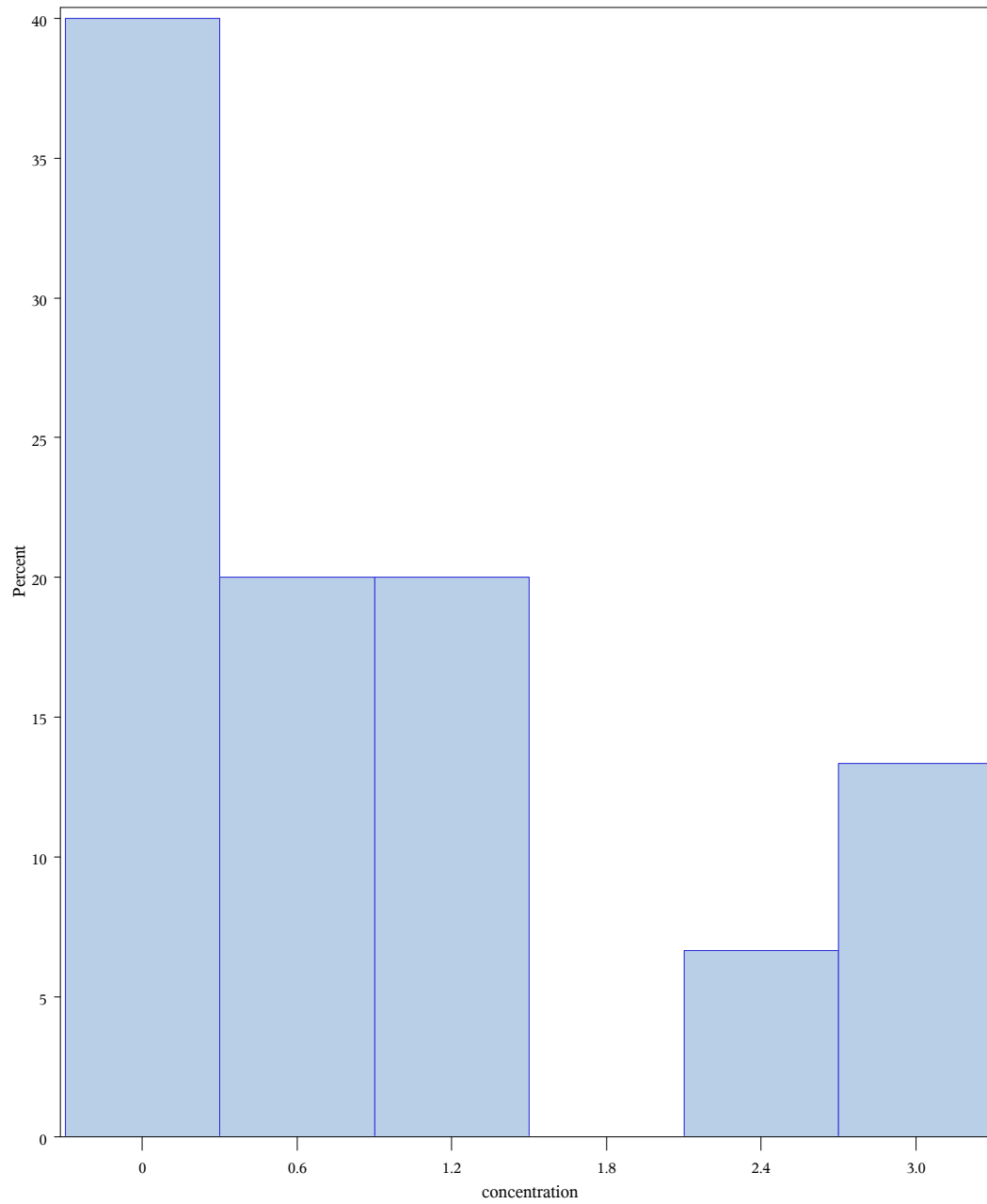
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.59712	12.59712	55.99	<.0001
Error	13	2.92465	0.22497		
Corrected Total	14	15.52177			

Root MSE	0.47431	R-Square	0.8116
Dependent Mean	0.95533	Adj R-Sq	0.7971
Coeff Var	49.64901		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t 	95% Confidence Limits	
Intercept	1	2.57533	0.24873	10.35	<.0001	2.03798	3.11269
time	1	-0.32400	0.04330	-7.48	<.0001	-0.41754	-0.23046

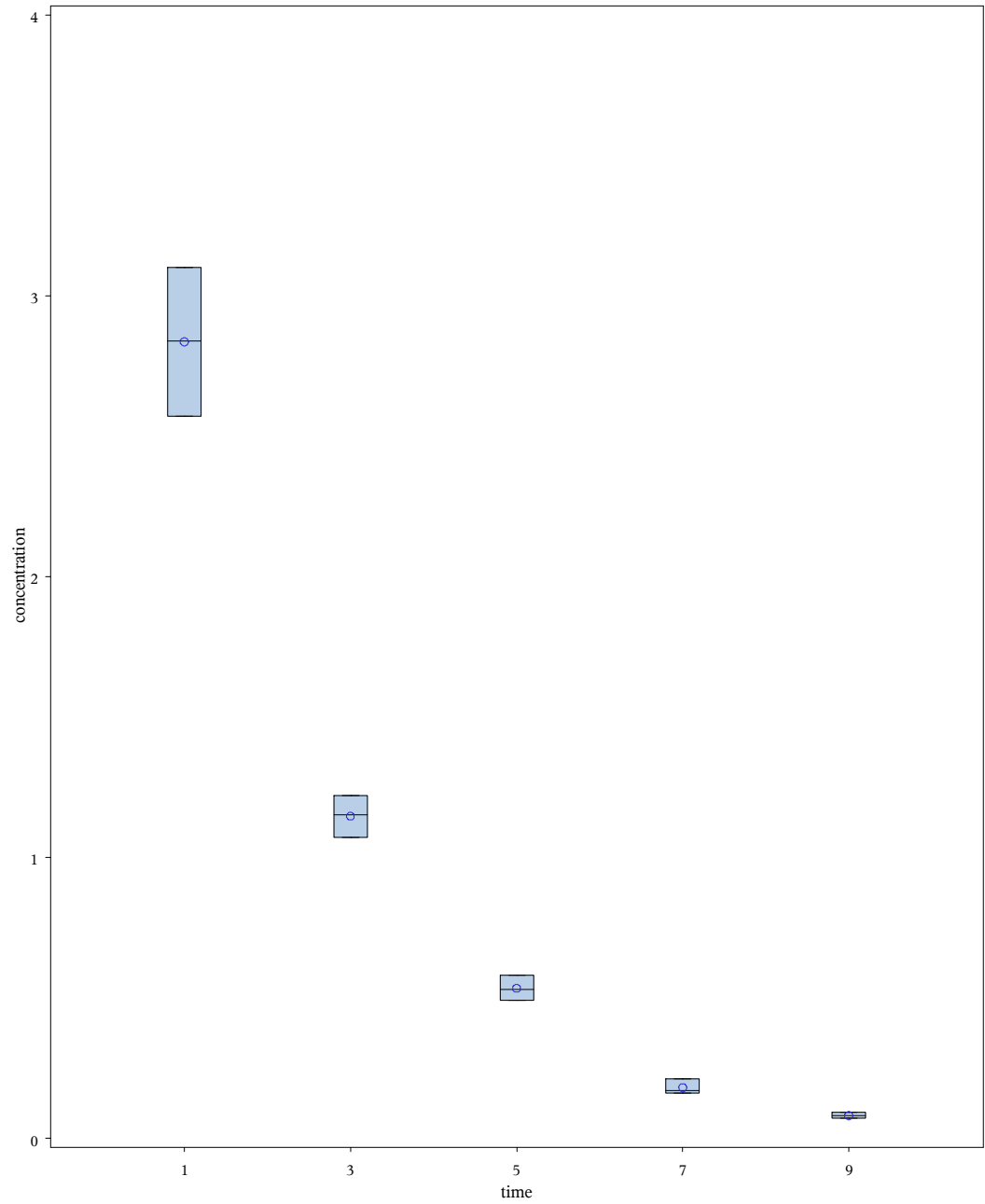
02:35 Thursday, February 20, 2014 1

**Solution Concentration
Histogram**



02:35 Thursday, February 20, 2014 1

Solution Concentration
Box Plot



Solution Concentration***Box Plot******The REG Procedure******Model: MODEL1******Dependent Variable: concentration***

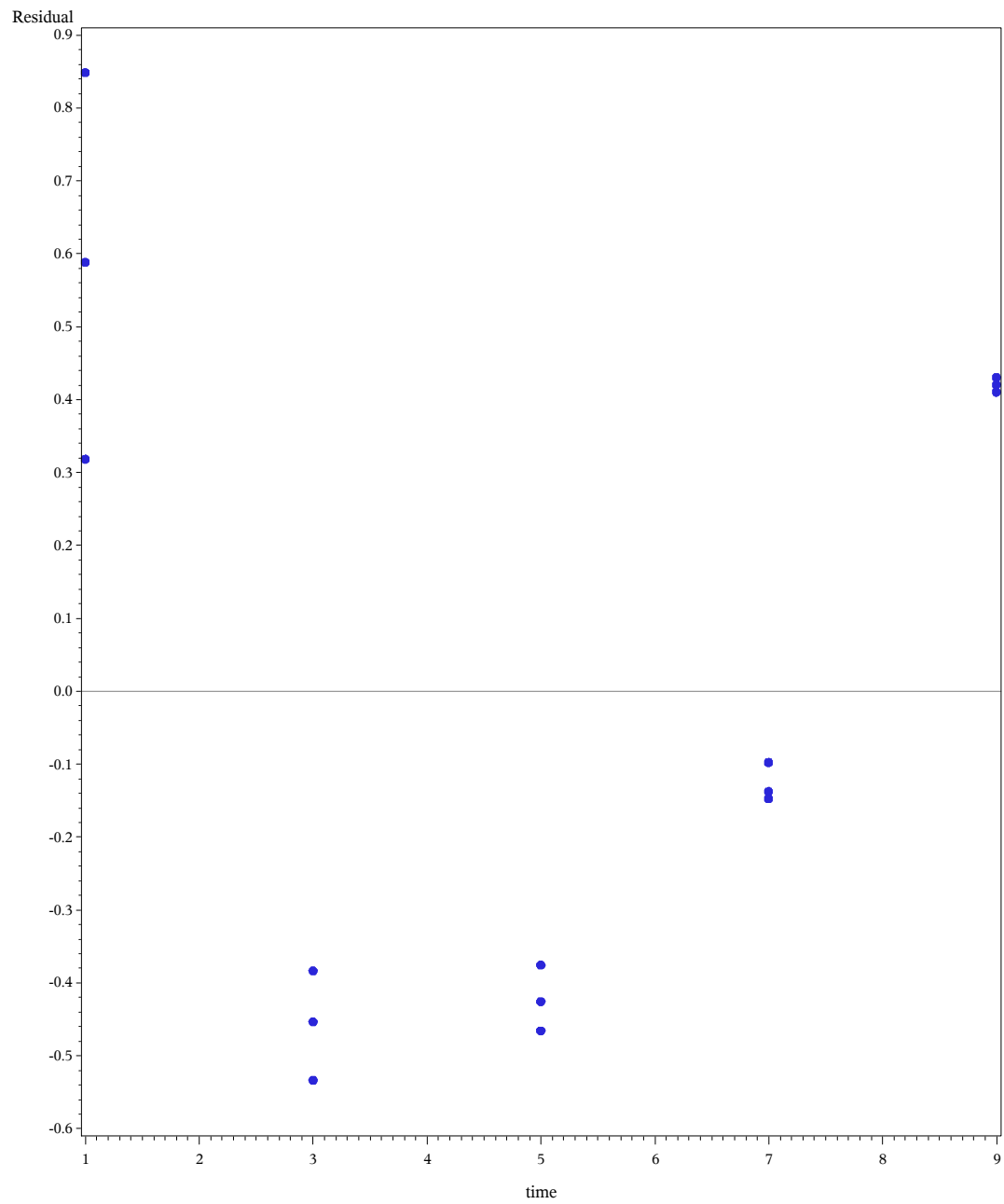
Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	2.8400	2.2513	0.2121	1.1288	3.3738	0.5887
2	2.5700	2.2513	0.2121	1.1288	3.3738	0.3187
3	3.1000	2.2513	0.2121	1.1288	3.3738	0.8487
4	1.2200	1.6033	0.1500	0.5286	2.6780	-0.3833
5	1.1500	1.6033	0.1500	0.5286	2.6780	-0.4533
6	1.0700	1.6033	0.1500	0.5286	2.6780	-0.5333
7	0.4900	0.9553	0.1225	-0.1030	2.0136	-0.4653
8	0.5800	0.9553	0.1225	-0.1030	2.0136	-0.3753
9	0.5300	0.9553	0.1225	-0.1030	2.0136	-0.4253
10	0.1600	0.3073	0.1500	-0.7674	1.3820	-0.1473
11	0.1700	0.3073	0.1500	-0.7674	1.3820	-0.1373
12	0.2100	0.3073	0.1500	-0.7674	1.3820	-0.0973
13	0.0700	-0.3407	0.2121	-1.4632	0.7818	0.4107
14	0.0900	-0.3407	0.2121	-1.4632	0.7818	0.4307
15	0.0800	-0.3407	0.2121	-1.4632	0.7818	0.4207

Sum of Residuals	0
Sum of Squared Residuals	2.92465
Predicted Residual SS (PRESS)	4.12135

4.4 Part d

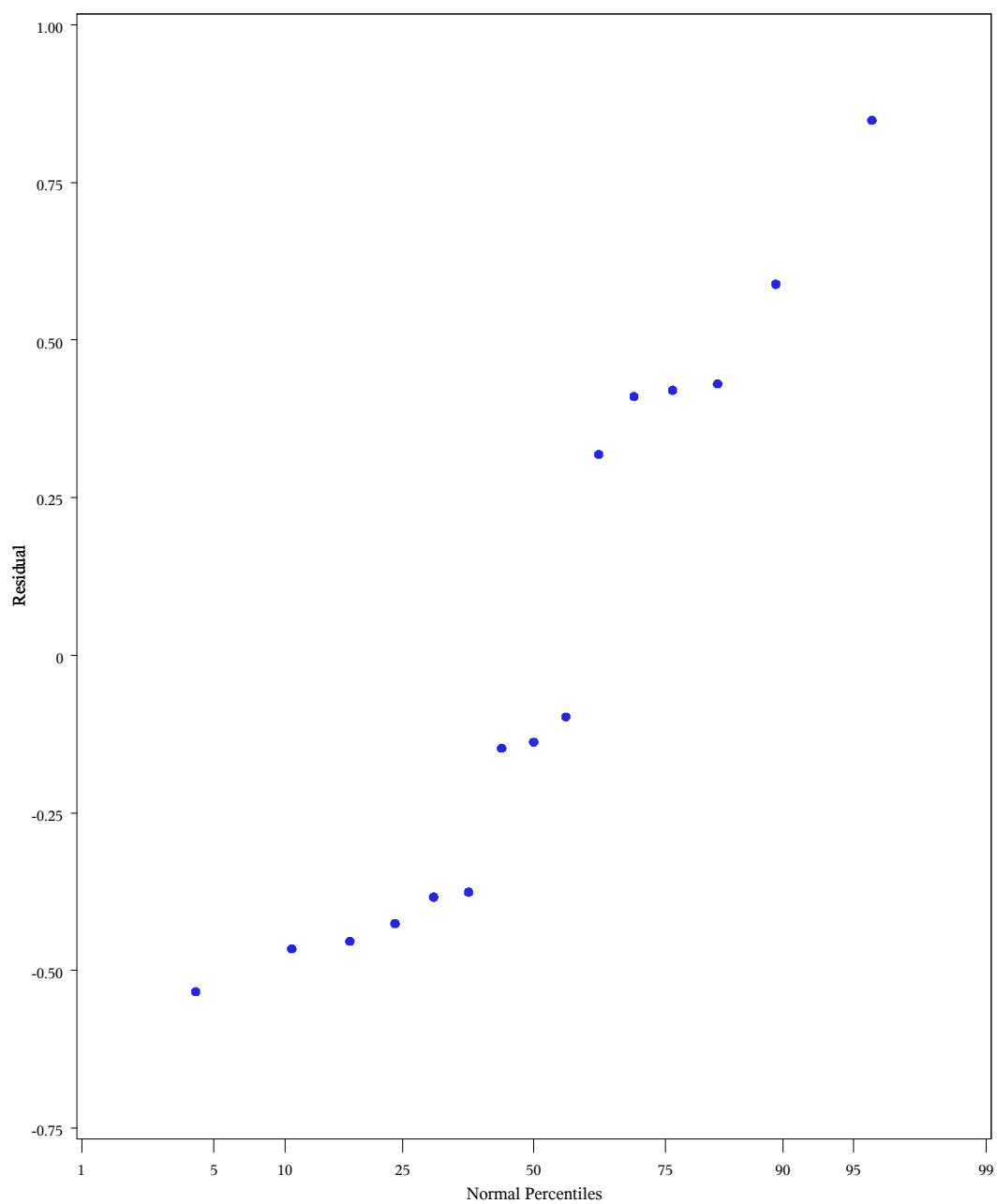
Both the normal probability plot and the residuals plot shown in plot 4.4 and 4.4 indicate that our errors are not normally distributed. This could be rectified by a transformation on Y.

02:35 Thursday, February 20, 2014 1

Solution Concentration
Residuals vs. time

02:35 Thursday, February 20, 2014 1

Solution Concentration
Normal Probability



4.5 Part e

Solution Concentration***Normal Probability******The RSREG Procedure***

Response Surface for Variable concentration	
Response Mean	0.955333
Root MSE	0.474314
R-Square	0.8116
Coefficient of Variation	49.6490

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Covariates	1	12.597120	0.8116	55.99	<.0001
Linear	0	0	0.0000	.	.
Quadratic	0	0	0.0000	.	.
Crossproduct	0	0	0.0000	.	.
Total Model	1	12.597120	0.8116	55.99	<.0001

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	2.767253	0.922418	58.60	<.0001
Pure Error	10	0.157400	0.015740		
Total Error	13	2.924653	0.224973		

Parameter	DF	Estimate	Standard Error	t Value	Pr > t	Parameter Estimate from Coded Data
Intercept	1	2.575333	0.248732	10.35	<.0001	2.575333
time	1	−0.324000	0.043299	−7.48	<.0001	−0.324000

With a lack of fit test, our hypothesis are

$$H_0 : E(Y) = 2.57533 + 0.324X \quad (2)$$

$$H_a : E(Y) \neq 2.57533 + 0.324X \quad (3)$$

To prove a linear relationship, we want to fail to reject H_0 . Unfortunately, our p-value for this test are < 0.0001 , indicating that we do reject H_0 . Thus, we can conclude that the data is not accurately modeled by a linear relationship.

4.6 Part f

Running Box-Cox on the data, we get the following results:

Solution Concentration

Normal Probability

The TRANSREG Procedure

Box-Cox Transformation Information for time				
Lambda		R-Square	Log Like	
−3.00		0.88	−37.2367	
−2.75		0.88	−32.9674	
−2.50		0.89	−28.6758	
−2.25		0.90	−24.3345	
−2.00		0.91	−19.9035	
−1.75		0.92	−15.3228	
−1.50		0.94	−10.5023	
−1.25		0.95	−5.3101	
−1.00		0.97	0.4221	
−0.75		0.98	6.7164	
−0.50	+	0.99	12.2568	*
−0.25		0.99	12.8678	<
0.00		0.97	9.1486	
0.25		0.95	5.0555	
0.50		0.91	1.5428	
0.75		0.86	−1.4700	
1.00		0.81	−4.1511	
1.25		0.76	−6.6264	
1.50		0.71	−8.9803	
1.75		0.66	−11.2682	
2.00		0.61	−13.5262	
2.25		0.57	−15.7777	
2.50		0.54	−18.0375	
2.75		0.51	−20.3149	
3.00		0.48	−22.6154	
< - Best Lambda * - 95% Confidence Interval + - Convenient Lambda				

From this, we chose to transform the data with the following equation:

$$Y' = Y^{-0.25} \quad (4)$$

The resulting scatterplot of the transformed data, plot4.6, indicates a much stronger linear relationship.

4.7 Part g

The transformed regression equation is

$$\hat{Y}' = 0.56720 + 0.1398X \quad (5)$$

The p-value for the slope is < 0.0001 , also indicating a strong linear relationship assuming the SLR assumptions hold. The correlation value is 0.9725, much better than the non-transformed correlation.

4.8 Part h

The residual and normal probability plots in 4.8 and 4.8 indicate a much more normal distribution of error terms. The residual plot is not quite perfectly independently distributed, but it is much less erratic than the non-transformed plot.

4.9 Part i

Solution Concentration ** -0.25

Normal Probability

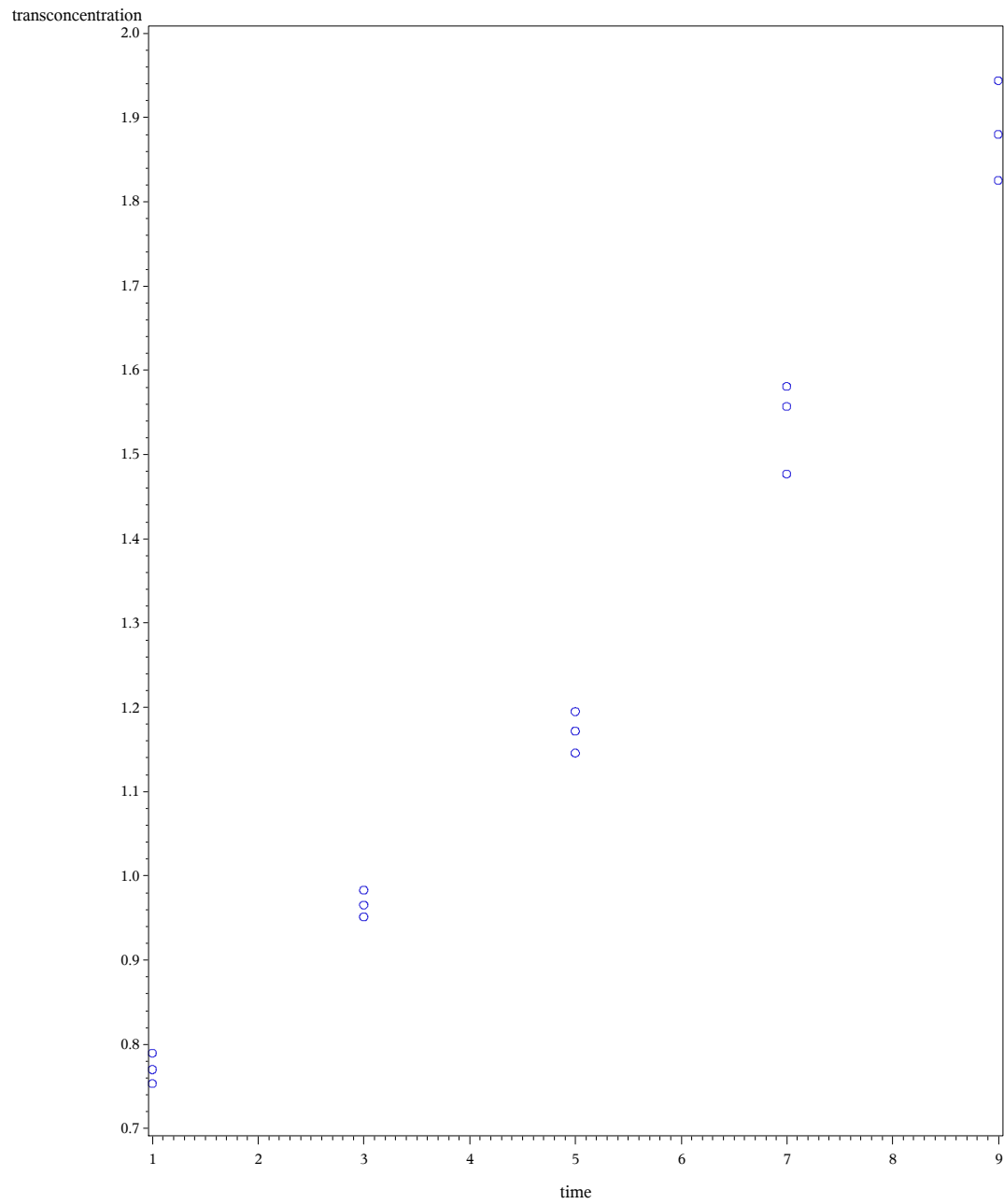
The RSREG Procedure

Response Surface for Variable transconcentration	
Response Mean	1.266211
Root MSE	0.071415
R-Square	0.9725
Coefficient of Variation	5.6401

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Covariates	1	2.345379	0.9725	459.87	<.0001
Linear	0	0	0.0000	.	.
Quadratic	0	0	0.0000	.	.
Crossproduct	0	0	0.0000	.	.
Total Model	1	2.345379	0.9725	459.87	<.0001

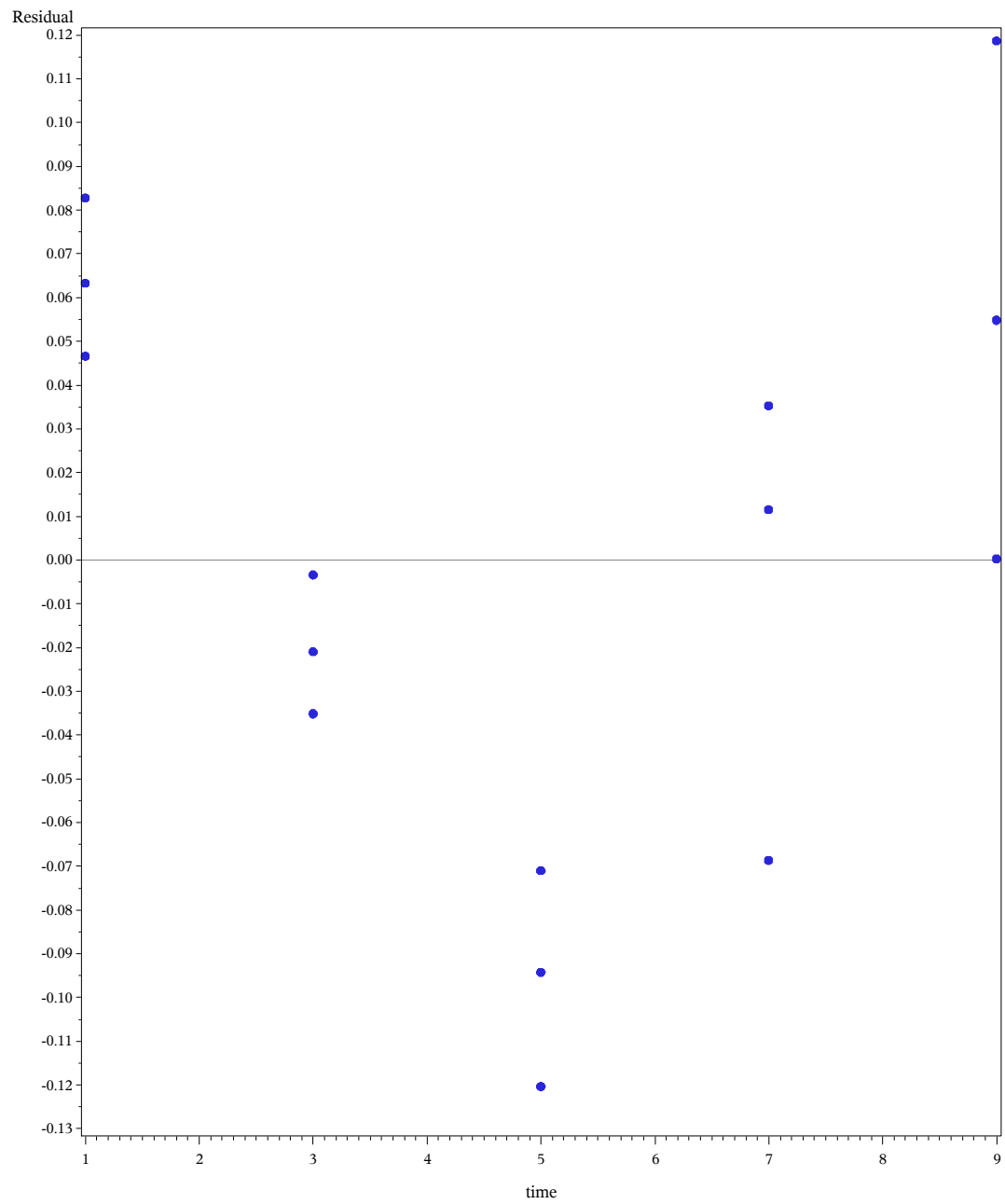
02:35 Thursday, February 20, 2014 1

Solution Concentration ** -0.25
Scatterplot Concentration vs. Time



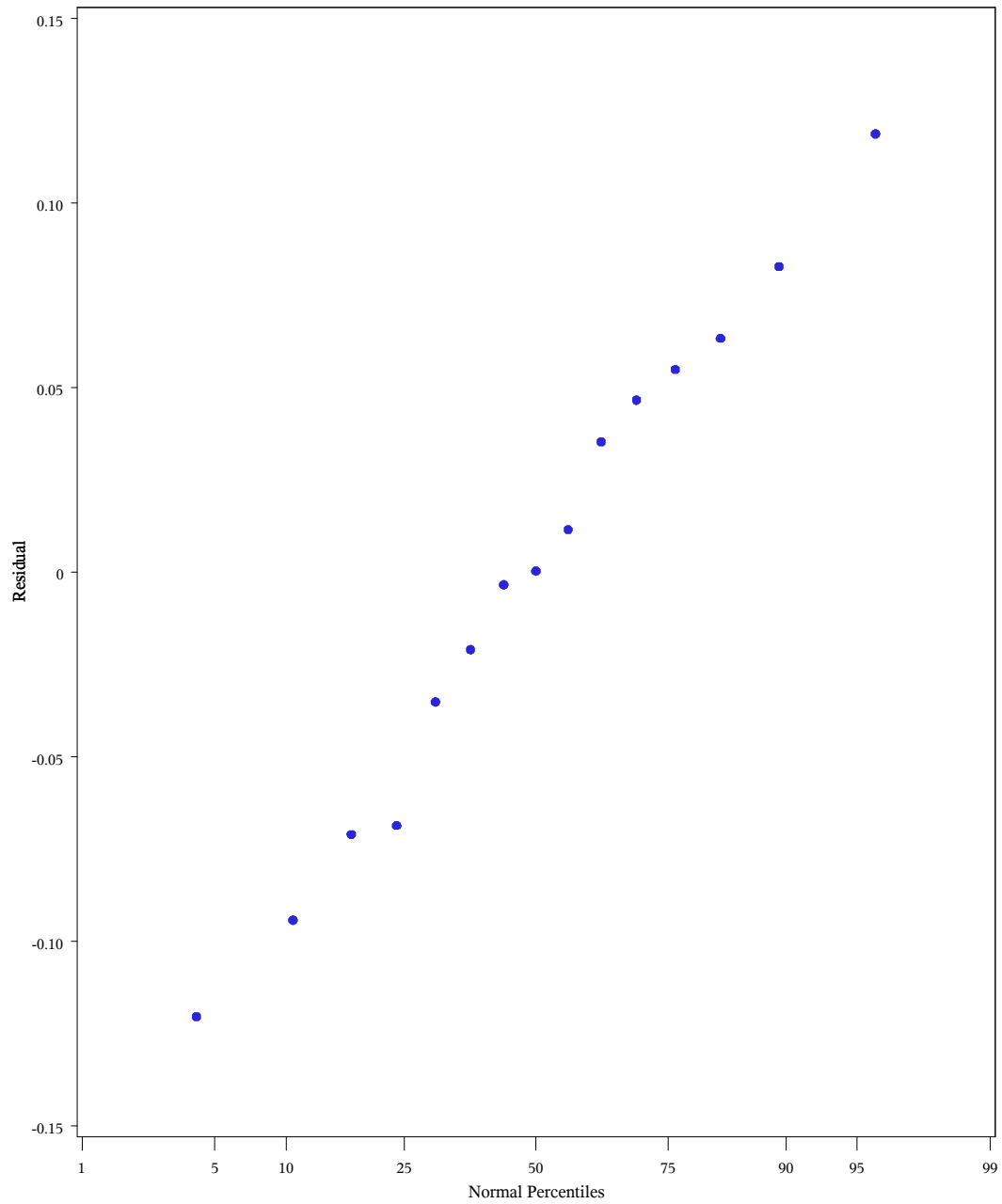
02:35 Thursday, February 20, 2014 1

Solution Concentration ** -0.25
Residuals vs. time



02:35 Thursday, February 20, 2014 1

Solution Concentration ** -0.25
Normal Probability



Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	0.050971	0.016990	11.08	0.0016
Pure Error	10	0.015330	0.001533		
Total Error	13	0.066301	0.005100		

Parameter	DF	Estimate	Standard Error	t Value	Pr > t	Parameter Estimate from Coded Data
Intercept	1	0.567197	0.037450	15.15	<.0001	0.567197
time	1	0.139803	0.006519	21.44	<.0001	0.139803

To perform a lack-of-fit test on the transformed data, we start with the following hypotheses:

$$H_0 : E(Y) = 0.56720 + 0.1398X \quad (6)$$

$$H_a : E(Y) \neq 0.56720 + 0.1398X \quad (7)$$

The resulting p-value is 0.0016, indicating that we fail to reject H_0 . Thus, we can conclude that our transformed data can be modeled accurately by linear regression.

5 Problem 5, KNN #3.23

Full Model:

$$Y_{ij} = \mu_j + \epsilon_{ij} \quad (8)$$

Degrees of freedom: $n - c$

Reduced Model:

$$Y_{ij} = \beta_1 X_j + \epsilon_{ij} \quad (9)$$

Degrees of freedom: $n - 2$