School of Mathematics and Statistics
Te Kura Mātai Tatauranga

STAT 292 **Assignment 4** **Due by 12pm (midday), Fri 12 June 2020**

There are four questions, worth a total of 100 marks. Question 1 starts on page 2.

Assignment Guidelines (once more)

You are encouraged to discuss assignments with other students, but your submitted work must be your own.

The following Assignment Guidelines are helpful for all the assignments in Parts 2 and 3 of the course.

When you carry out a statistical test of hypothesis, you should state the following, **when relevant**:

- Model equation.
- Assumptions about the data, and comments about whether diagnostic graphs support those assumptions.
- Null and alternative hypotheses.
- ANOVA Table (if relevant), p-values.
- Statistical conclusions. For example, "We reject H_0 and conclude H_A, that μ_1 and μ_2 differ at the 5% significance level".
- Interpretation of the statistical conclusions back to the original problem, using the original meaning of the response variable and any factors or covariates. For example, if comparing heights of two groups, "Female and male adults have different mean heights, with males being taller on average".

1. **Comprehension Test**

Children in a school class are given a test of comprehension of English, marked out of 100. The children are from three different ethnic groups, which is thought to be an important factor. The question of interest is whether there are sex differences after allowing for ethnicity. The data follow:

	Females	Males
Ethnic group E1	67 66 75 76 71 70 72	63 72 62 61 69 64 71 68 56
E2	69 57 55 63 65 55	59 47 49
E3	30 47	39 33

(a) A two-way ANOVA was run on the data, with SAS output given on pages 3 to 6. Present the results from the ANOVA following the usual Assignment Guidelines, as given on page 1.

(b) If a one-way ANOVA is done with factor Sex, the resulting ANOVA table is:

Source	DF	Sum of Squares	Mean Square	F value	p-value
Sex	1	144.166	144.166	0.99	0.3292
Error	27	3942.662	146.025		
Total	28	4086.828			

Briefly discuss the outcomes of the separate tests for Sex presented in parts (a) and (b). Are the conclusions different? Give reasons to explain your answer.

SAS Output for Comprehension Test

Linear Models

The GLM Procedure

Class Level Information		
Class	Levels	Values
Ethnicity	3	E1 E2 E3
Sex	2	F M

Number of Observations Read	29
Number of Observations Used	29

Dependent Variable: Comprehension

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3365.438697	673.087739	21.46	<.0001
Error	23	721.388889	31.364734		
Corrected Total	28	4086.827586			

R-Square	Coeff Var	Root MSE	Comprehension Mean
0.823484	9.275400	5.600423	60.37931

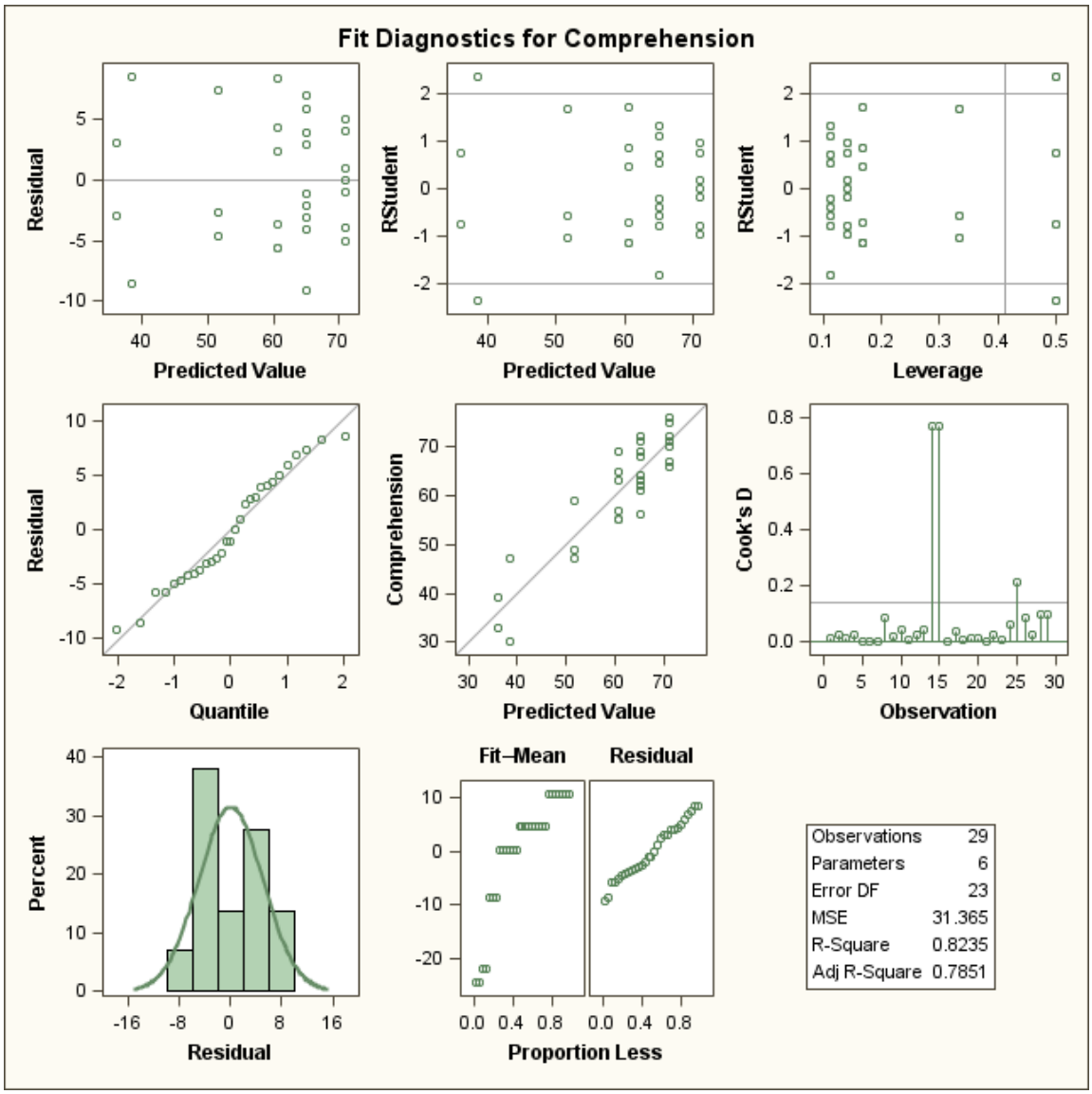
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Ethnicity	2	3060.640086	1530.320043	48.79	<.0001
Sex	1	275.113176	275.113176	8.77	0.0070
Ethnicity*Sex	2	29.685435	14.842718	0.47	0.6289

SAS Output for Comprehension Test

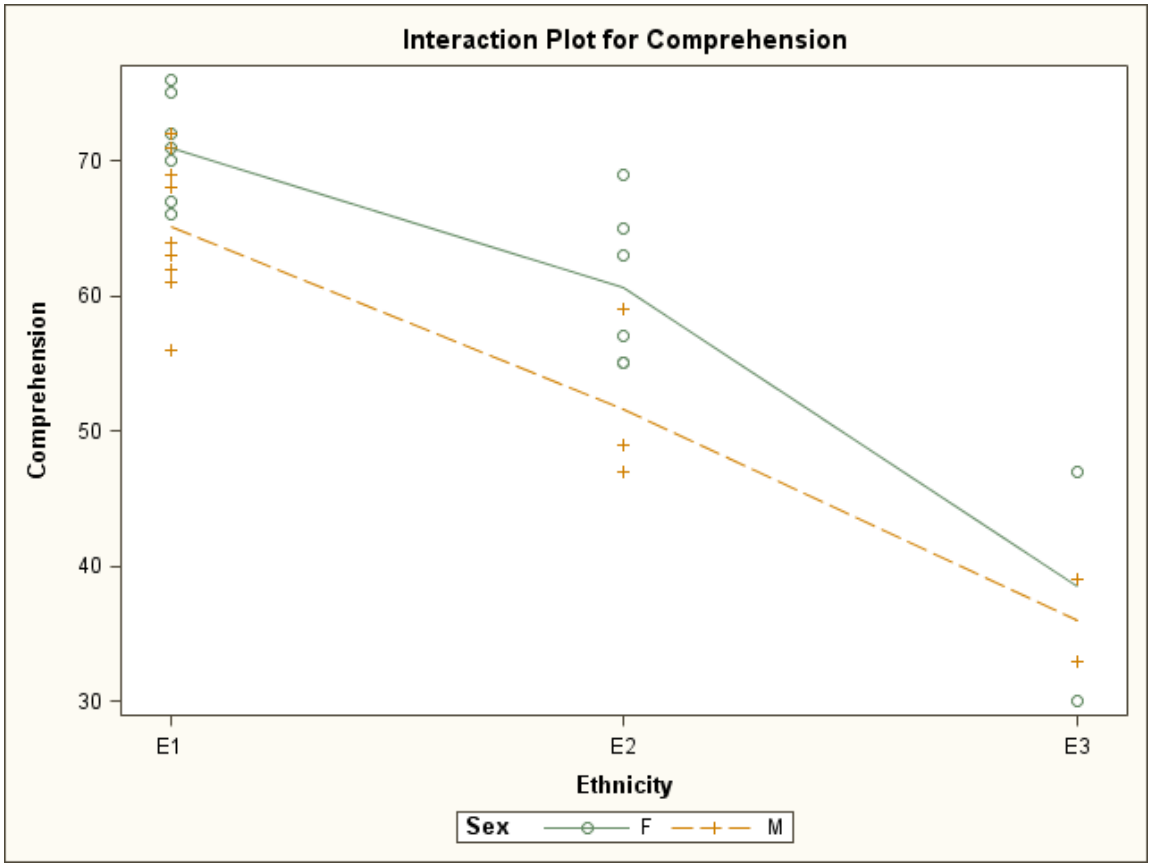
Linear Models

The GLM Procedure

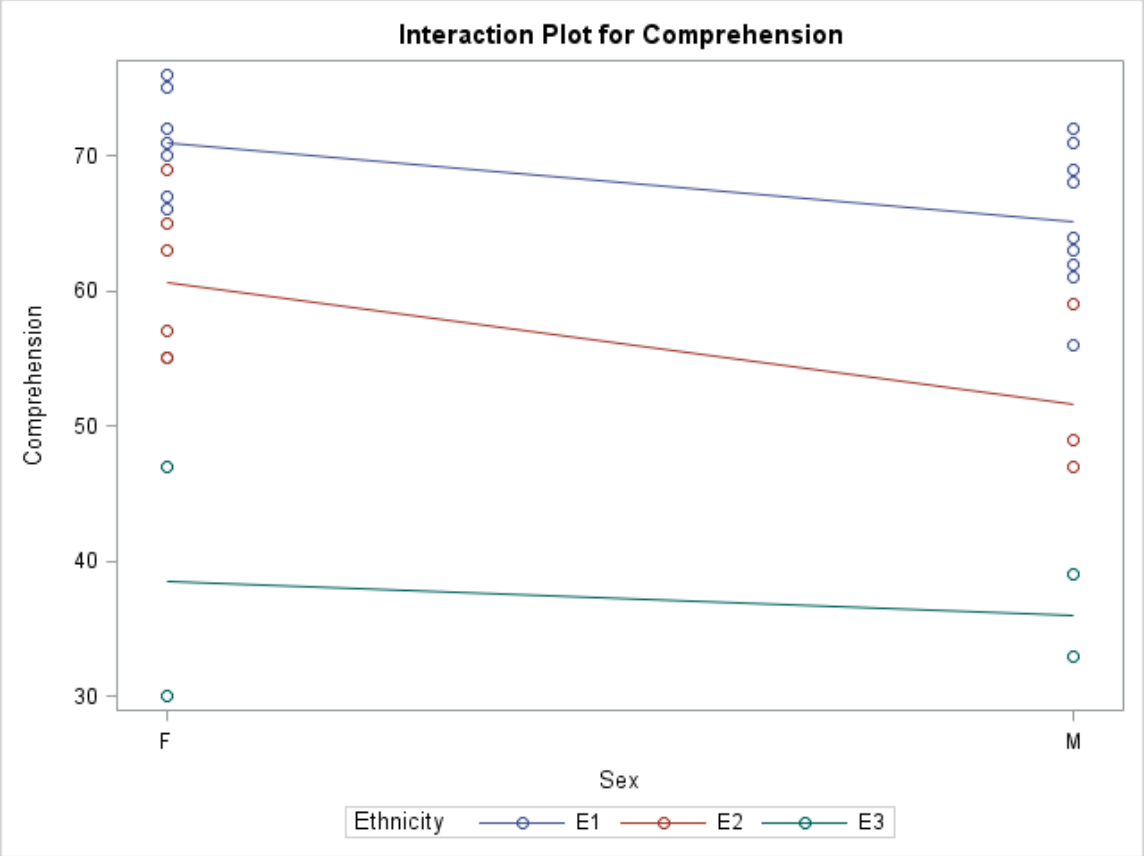
1



SAS Output for Comprehension Test



SAS Output for Comprehension Test



Note: E1 is the top line, E2 the middle line and E3 the lowest. (The lines are different colours, but that doesn't show up if viewed or printed in black and white.)

2. **Invertebrates in Mussel Clumps**

The following data are from Peake and Quinn (1993), Temporal variation in species-area curves for invertebrates in clumps of an intertidal mussel, *Ecography* **16**, 269-277. The two variables used in this question are:

$x = \log_{10}(\text{Area})$ of each of 25 mussel clumps (in dm^2), and

$Y =$ number of different species of macroinvertebrates in each clump.

Note: Using log(Area) gives a straighter regression line than Area, which is why it is used. This is a transformation of x, not Y; it has been done to improve linearity, not to stabilise variances.

The data follow. Decide if there is a useful linear relationship between x and Y, i.e. if x is a useful linear predictor of Y.

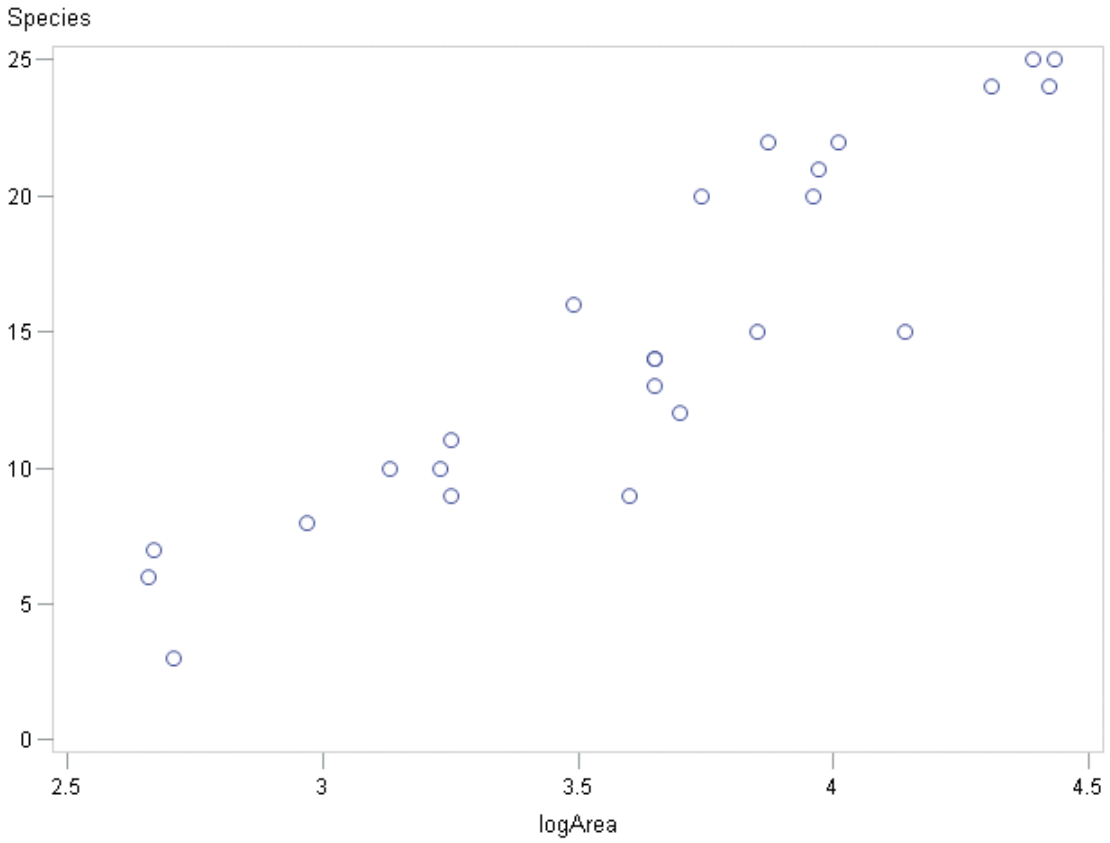
Clump	logArea	Species
1	2.71	3
2	2.67	7
3	2.66	6
4	2.97	8
5	3.13	10
6	3.25	9
7	3.23	10
8	3.25	11
9	3.49	16
10	3.60	9
11	3.65	13
12	3.65	14
13	3.70	12
14	3.65	14
15	3.74	20
16	3.87	22
17	3.85	15
18	3.96	20
19	4.01	22
20	3.97	21
21	4.14	15
22	4.31	24
23	4.39	25
24	4.43	25
25	4.42	24

(a) A scatterplot of the data is given on page 8. Give comments on whether you think the plot shows (i) linearity, (ii) constant variance.

(b) Output from a simple linear regression using logArea to predict the number of species is given on pages 9 and 10. Present a report on this analysis that includes (as usual) the model equation, hypotheses, assumptions, comments on whether the analysis is valid, plus statistical conclusions and interpretation.

SAS Output for Mussel Clumps

Scatter Plot



SAS Output for Mussel Clumps

Linear Regression Results

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: Species

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	868.50179	868.50179	117.85	<.0001
Error	23	169.49821	7.36949		
Corrected Total	24	1038.00000			

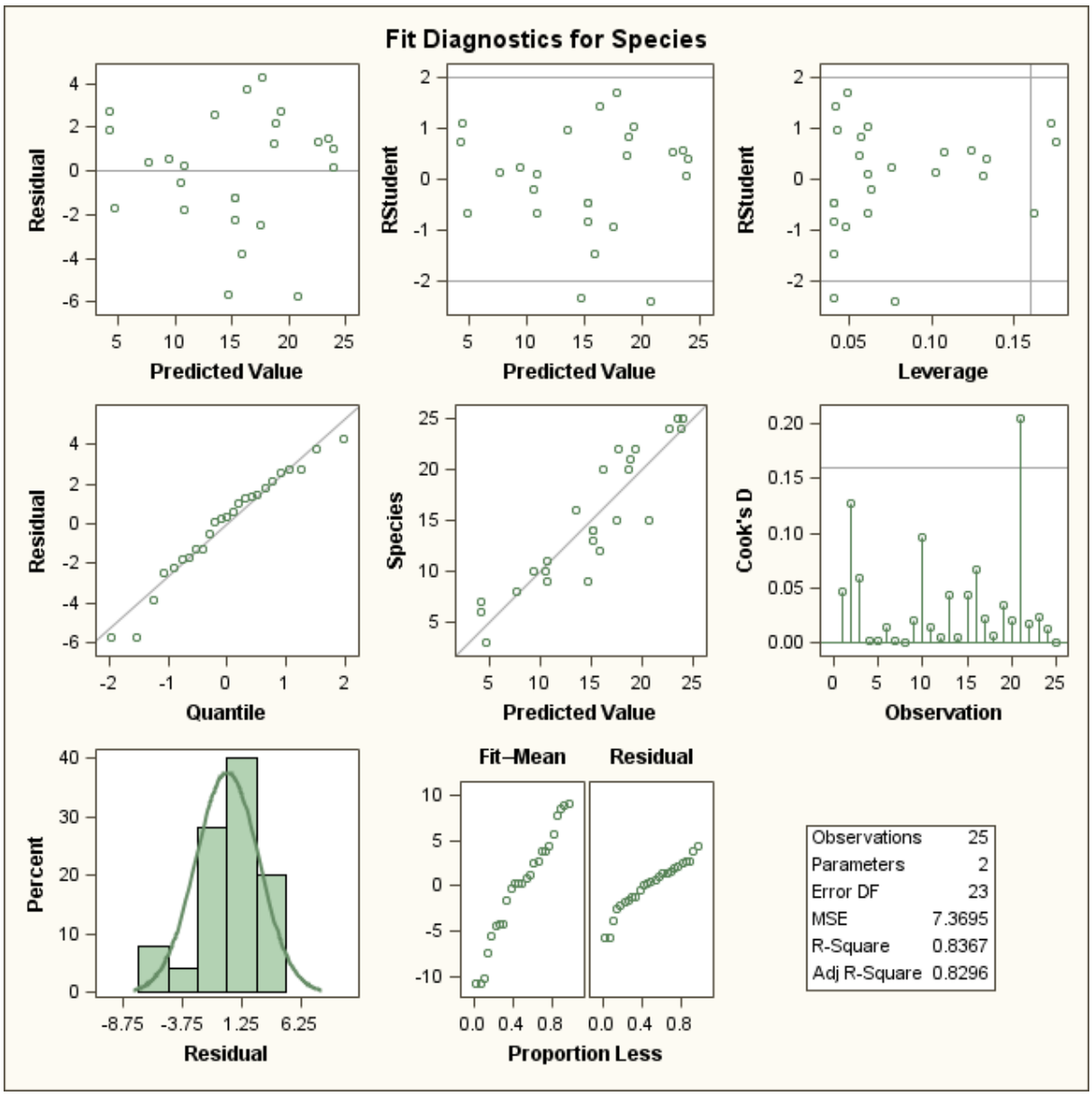
Root MSE	2.71468	R-Square	0.8367
Dependent Mean	15.00000	Adj R-Sq	0.8296
Coeff Var	18.09787		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-25.64136	3.78287	-6.78	<.0001
logArea	1	11.20214	1.03189	10.86	<.0001

SAS Output for Mussel Clumps

Linear Regression Results



3. **Coarse Woody Debris in Lakes**

Christensen *et al.* (1996, *Ecological Applications* **6(4)**, 1143-1149) studied the relationships between coarse woody debris (CWD), shoreline vegetation and lake development in a sample of 16 lakes in North America. Coarse woody debris is useful in providing a habitat for various fish species. It is known to be related to the riparian (river-bank, lake-edge) tree density, irrespective of whether or not humans are present. The objective is to find out whether, after allowing for riparian tree density, human habitation is having an effect on the CWD.

The variables below were taken around the shoreline and near-shore water:

L10CABIN = $\log_{10}$ of 1 + density of cabins (number km^{-1}),
RIP.DENS = density of riparian trees (trees km^{-1}), and
CWD.BASA = basal area of coarse woody debris (m^2 km^{-1}).

LAKE	AREA	RIP.DENS	CWD.BASA	L10CABIN
Bay	69	1270	121	0
Bergner	9	1210	41	0
Crampton	24	1800	183	0
Long	8	1875	130	0
Roach	20	1300	127	0
Tenderfoot	175	2150	134	0.20412
Palmer	254	1330	65	0.462398
Street	22	964	52	0.6627578
Laura	240	961	12	0.7075702
Annabelle	85	1400	46	0.763428
Joyce	12	1280	54	0.845098
Lake hills	25	976	97	0.8864907
Towanda	58	771	1	1.10721
Black oak	234	833	4	1.1238516
Johnson	31	883	1	1.2552725
Arrowhead	40	956	4	1.40824

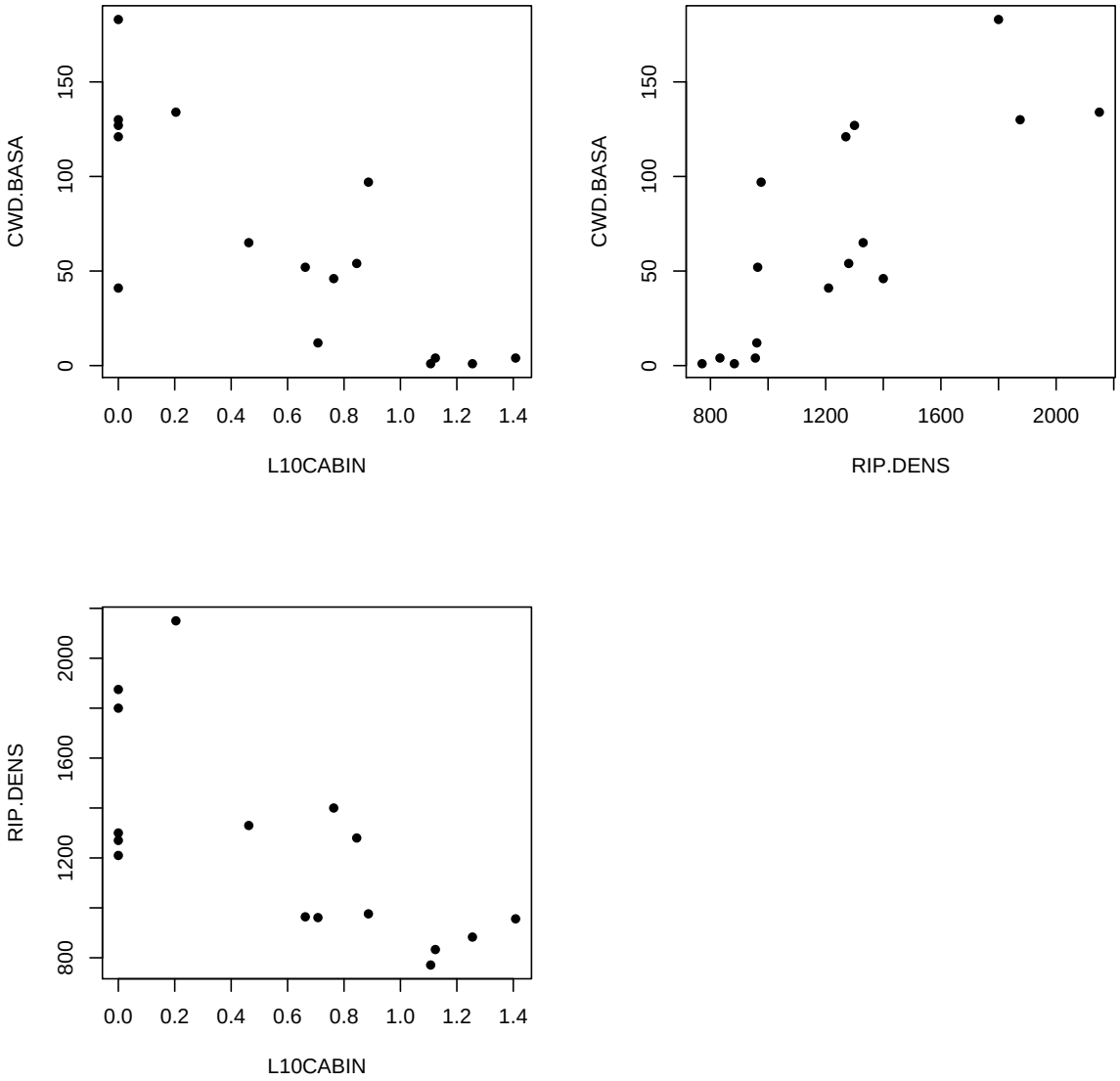
(a) Let Y = CWD.BASA, X_1 = RIP.DENS and X_2 = L10CABIN. Plots of Y vs. X_1, Y vs. X_2 and X_1 vs. X_2 are given on page 12. Comment on any relationships you see.

(b) SAS output for the following models is presented on pages 13 to 16. Diagnostic graphs are shown for the last model.

i. Regression of Y on the predictor X_1
ii. Regression of Y on the predictor X_2
iii. Regression of Y on the two predictors X_1 and X_2

For each analysis above, present the model equation, hypotheses and conclusions. For the third analysis, comment on whether or not the model assumptions are satisfied.

(c) Which of the hypothesis tests from the three presented models gives the answer to the question of interest in this situation? Explain the answer.



Scatterplots: CWD by L10CABIN, CWD by RIP.DENS, RIP.DENS by L10CABIN

Coarse Woody Debris SAS Output

Linear Regression Results

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: CWD.BASA

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32054	32054	24.30	0.0002
Error	14	18466	1318.96866		
Corrected Total	15	50520			

Root MSE	36.31761	**R-Square**	0.6345
Dependent Mean	67.00000	**Adj R-Sq**	0.6084
Coeff Var	54.20539		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-77.09908	30.60801	-2.52	0.0246
RIP.DENS	1	0.11552	0.02343	4.93	0.0002

Linear Regression Results

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: CWD.BASA

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32840	32840	26.00	0.0002
Error	14	17680	1262.86950		
Corrected Total	15	50520			

Root MSE	35.53688	**R-Square**	0.6500
Dependent Mean	67.00000	**Adj R-Sq**	0.6250
Coeff Var	53.04011		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	121.96875	13.96871	8.73	<.0001
L10CABIN	1	-93.30142	18.29646	-5.10	0.0002

Linear Regression Results

The REG Procedure

Model: Linear_Regression_Model

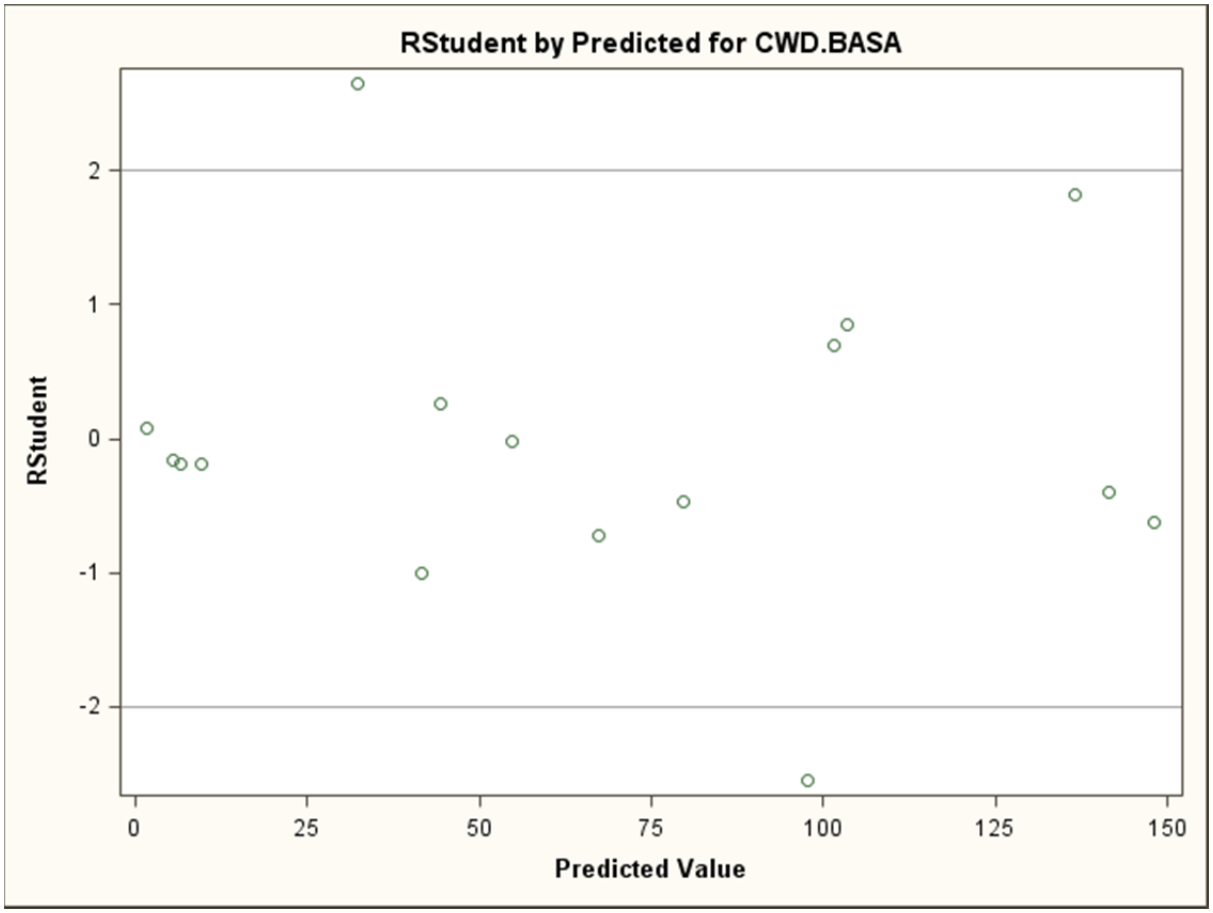
Dependent Variable: CWD.BASA

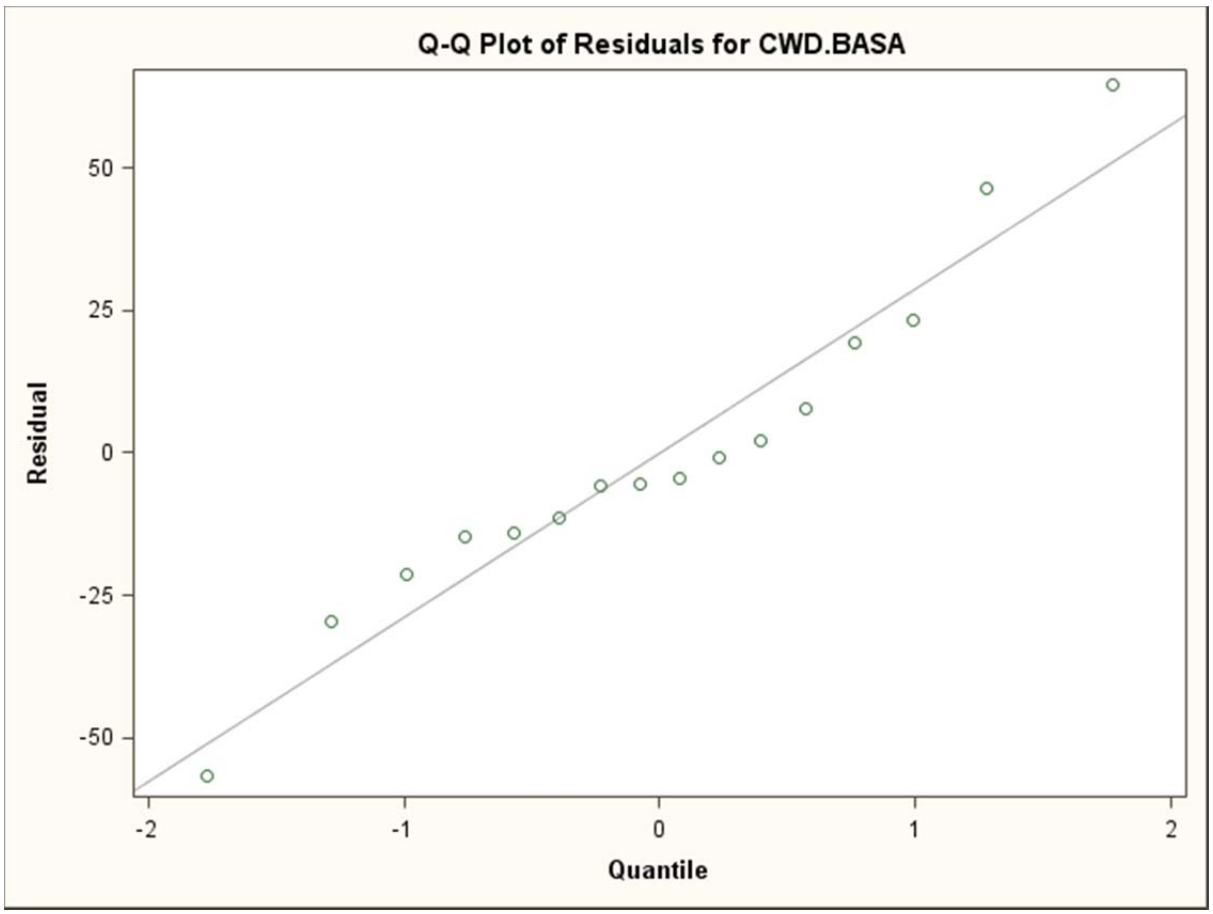
Number of Observations Read	16
Number of Observations Used	16

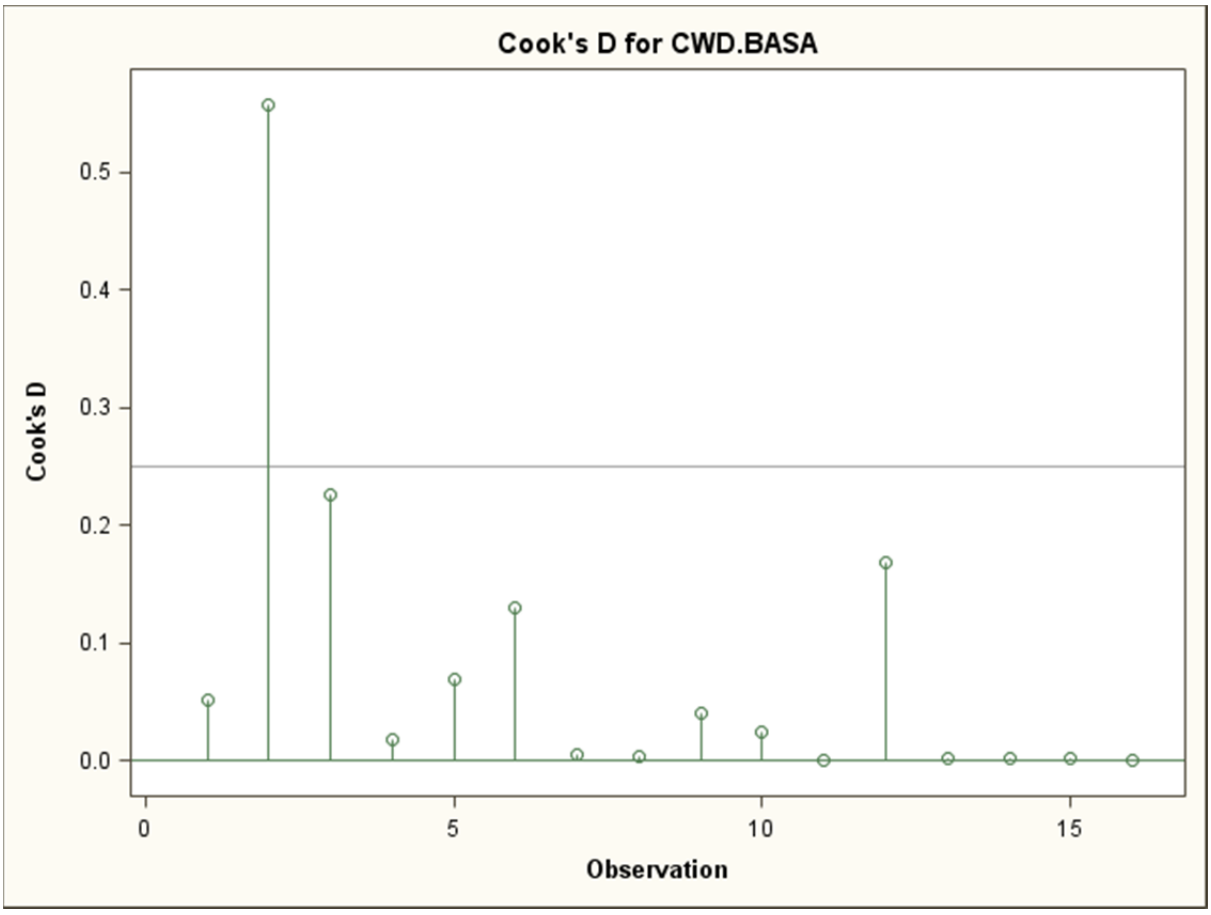
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	38041	19020	19.81	0.0001
Error	13	12479	959.93185		
Corrected Total	15	50520			

Root MSE	30.98277	R-Square	0.7530
Dependent Mean	67.00000	Adj R-Sq	0.7150
Coeff Var	46.24294		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.16485	46.22822	0.39	0.7007
RIP.DENS	1	0.06572	0.02823	2.33	0.0367
L10CABIN	1	-56.26481	22.53059	-2.50	0.0267


RStudent by Predicted for CWD.BASA
RStudent
2
1
0
-1
-2
0
25
50
75
100
125
150
Predicted Value


Q-Q Plot of Residuals for CWD.BASA
Residual
50
25
0
-25
-50
-2
-1
0
1
2
Quantile


Cook's D for CWD.BASA
Cook's D
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
Observation

4. **Age of Teeth**

In forensic work, scientists estimate the age of a skeleton by counting teeth cementum annulation (i.e. growth rings). Two teeth preparation methods, A and B, are compared by estimating the ages (Y) of twenty teeth of known age (X). The teeth are randomly allocated to the two methods, ten to each, as follows.

Method A	X = true age	49	13	38	55	44	56	7	66	18	39
	Y = estimated age	50	14	38	57	44	55	7	63	20	38
Method B	X = true age	51	59	32	37	12	38	4	28	58	24
	Y = estimated age	51	59	29	34	10	35	5	25	57	22

A confirmatory analysis using a model with terms True Age (i.e. X), Method and True Age×Method is required.

(a) Give the model equation for the required confirmatory analysis.

(b) SAS output from a fitted model is given on pages 18 to 20. Present a report on this analysis that includes any necessary assumptions, comments on their validity, hypotheses, statistical conclusions at a 5% significance level, and interpretation plus discussion.

Linear Models

The GLM Procedure

Class Level Information		
Class	Levels	Values
Method	2	A B

Number of Observations Read	20
Number of Observations Used	20

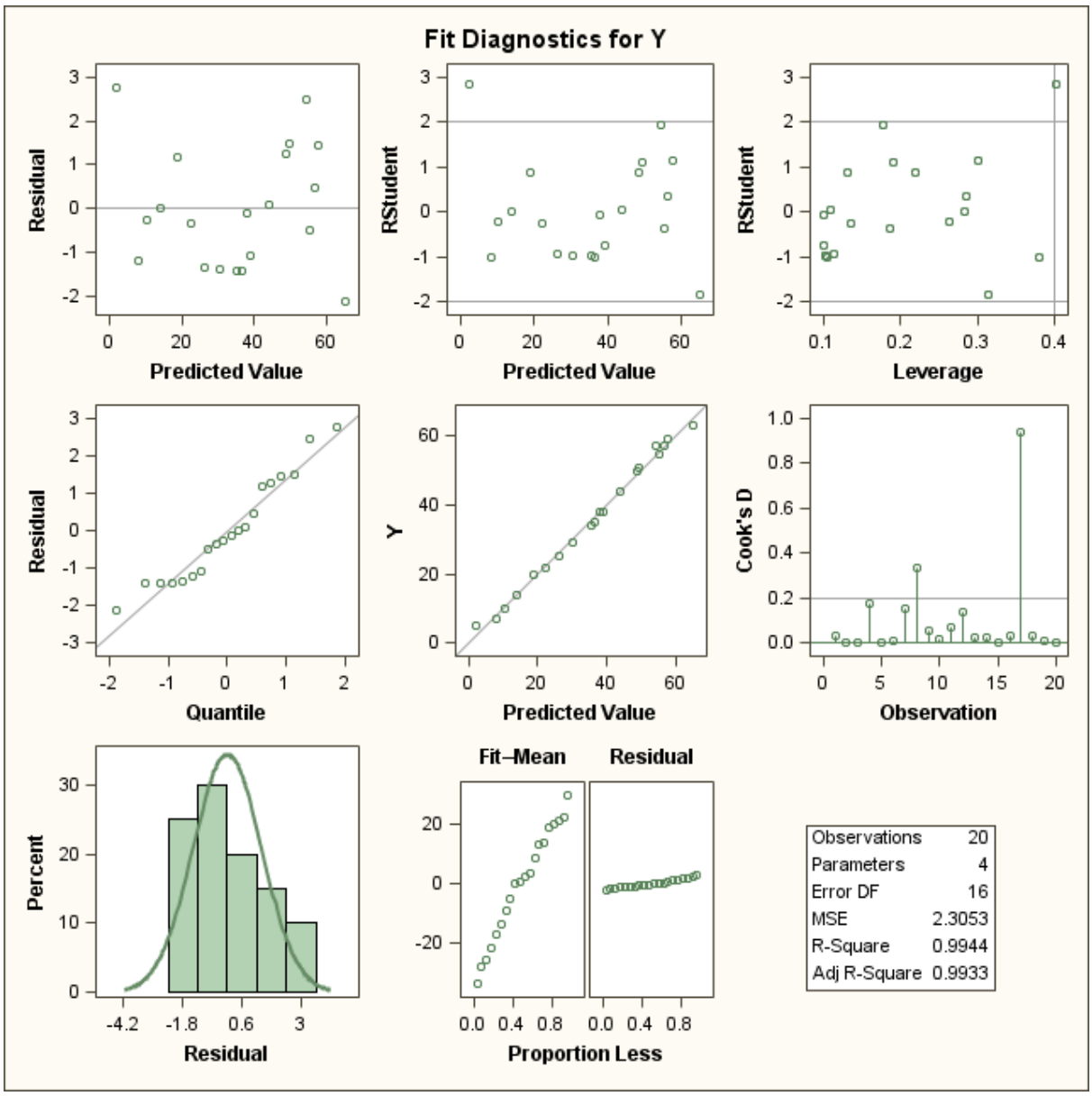
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6543.664660	2181.221553	946.16	<.0001
Error	16	36.885340	2.305334		
Corrected Total	19	6580.550000			

R-Square	Coeff Var	Root MSE	Y Mean
0.994395	4.258997	1.518333	35.65000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	6525.535206	6525.535206	2830.62	<.0001
Method	1	15.413619	15.413619	6.69	0.0199
X*Method	1	2.715836	2.715836	1.18	0.2938

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X	1	6350.006837	6350.006837	2754.48	<.0001
Method	1	10.463729	10.463729	4.54	0.0490
X*Method	1	2.715836	2.715836	1.18	0.2938


Fit Diagnostics for Y
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Y
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 20
Parameters 4
Error DF 16
MSE 2.3053
R-Square 0.9944
Adj R-Square 0.9933

Data and fitted lines: Method A line (dashed) is above Method B line (solid)