Name: Connor Raymond Stewart
Student Number: 101041125

STAT 2509 Assignment Two:

①

a) The Scatter diagram on Page two of the SAS output Shows an approximatly linear relationship.
the line in the graph Shows that as x increases, So does y, indicating a Positive relationship.

b) The SLR Model used:

$$y = B_0 + B_1 X_1 + \varepsilon$$

Where ε is the error term

the Assumptions are:

1) x's observed w/ no error
2) y's are independently distributed with mean: $E(y) = B_0 + B_1 X$
3) Veriance of y's Constant
4) y's or errors are $N \sim (E(y), \sigma^2)$

c)
$$\hat{B}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{6251 - \frac{(738)(98)}{13}}{45580 - (738)^2/13}$$

$$= \frac{8939}{47896} \doteq 0.186633539 \doteq \boxed{0.1866}$$

the least Squares estimates of $\hat{B}_0$ & $\hat{B}_1$ are as follows:

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x} = \frac{\sum y_i}{n} - \hat{B}_1\left(\frac{\sum x_i}{n}\right) = \frac{98}{13} - \left(\frac{8939}{47896}\right)\left(\frac{738}{13}\right)$$

$$\doteq \frac{98}{13} - 10.59504246 \doteq -3.056580925 \doteq \boxed{-3.0566}$$

thus, for $y = \hat{B}_0 + \hat{B}_1 x_1$; we get the following least-Squares fitted Regression line:

$$y = -3.0566 + 0.1866 x_1$$

} y is the Percent of market Shares
x_1 is the evaluation Procedure

d) to verify the fitted regression line goes through $(\bar{x}, \bar{y})$:

$$(\bar{x}, \bar{y}) = \left(\frac{\sum x_i}{n}, \frac{\sum y_i}{n}\right) = \left(\frac{738}{13}, \frac{98}{13}\right)$$

thus:

$$y = \hat{B}_0 + \hat{B}_1 x_1 = \frac{98}{13} \Rightarrow y = -3.0566 + 0.1866\left(\frac{738}{13}\right)$$

$$= 7\tfrac{279}{520} \doteq 7.536538462 \doteq 7.53$$

$$\frac{98}{13} \doteq 7.538461538 \doteq 7.53$$

thus: $y = \hat{B}_0 + \hat{B}_1 x_1 \doteq \frac{98}{13}$ So Since $y \doteq \bar{y}$, we can say $(\bar{x}, \bar{y})$ is verified.

e) Obtain the Residuals for this data set:

$$e_i = y_i - \hat{y}_i$$

$$e_1 = y_1 - \hat{y}_1 \rightarrow \hat{y}_1 = \left[\frac{98}{13} - \left(\frac{8939}{47896}\right)\left(\frac{738}{13}\right)\right] + \left(\frac{8939}{47896}\right)(27)$$

$$\doteq 1.982524637$$

$$e_1 = 2 - 1.982524637 = 0.017475363 \doteq 0.017475$$

for the remaining, See excel Sheet on Next page

Evaluation Procedure (X)	% of Market Shares (Y)	Residual (e_i)	Sum of Residuals
27	2	0.017475363	-6.21725E-15
39	3	-1.222127109	
73	10	-0.567667446	
66	9	-0.261232671	
33	4	0.897674127	
43	6	1.031338734	
47	5	-0.715195423	
55	8	0.791736262	
60	7	-1.141431435	
68	9	-0.634499749	
70	10	-0.007766828	
75	13	2.059065475	
82	12	-0.2473693	

$\Rightarrow \sum e_i = -6.21725E{-15} \doteq 0$

$\therefore$ the residuals sum to zero

Since this is a fitted linear regression, it makes sense that the residuals sum to zero.

$$\begin{cases} e_i = y_i - \hat{y}_i \\ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \end{cases}$$

f) Find s^2:

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{13-2} = \frac{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right) - \frac{(S_{xy})^2}{S_{xx}}}{11}$$

$$= \frac{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right) - \frac{\left(\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right)^2}{\sum x_i^2 - (\sum x_i)^2/n}}{11}$$

$$= \frac{\left(878 - \frac{(78)^2}{13}\right) - \frac{\left(6251 - \frac{(738)(78)}{13}\right)^2}{45580 - (738)^2/13}}{11}$$

$$\doteq \frac{\overbrace{10.8986763}^{=SSE}}{11} \doteq \underline{0.990788754} \doteq 0.9908$$

So, $\boxed{s^2 = 0.990788754 \doteq 0.9908}$

this is our estimate of σ^2, so $s^2 \approx \sigma^2$

g) Using the t-test w/ $\alpha = 0.95$:

Step II) $$t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \frac{(8939/47896)}{\left(\frac{\sqrt{0.990788754}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}}\right)} = \frac{8939/47896}{\left(\frac{\sqrt{0.990788754}}{\sqrt{45580 - \frac{738^2}{13}}}\right)}$$

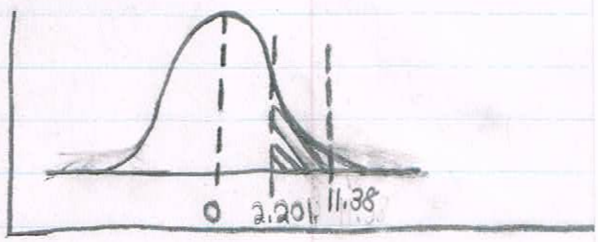
$$\doteq \underline{11.38091302} \doteq 11.3809 \doteq 11.38$$

Step I) Claim: the given claim is that there's a significant linear relationship between the evaluation procedure & percentage of market shares:

$H_0: B_1 = 0$

$H_a: B_1 \neq 0$

Step III) Find the rejection region:



$$t_{\alpha/2;n-2} = t_{0.025;11} = \underline{2.201}$$

Step IV) Since $t >$ (or $<$) 2.201 (R.R), we reject H_0 & conclude at 95% confidence, $\therefore$ there is a linear relationship between evaluation score & percentage of market shares.

h) Find a 95% Confidence interval for $\hat{B}_1$:

$$B_1 \in \hat{B}_1 \pm t_{n-2;\alpha/2} \cdot s/\sqrt{S_{xx}} \quad [SV, bV]$$

$$t_{11;0.025} = 2.201$$

$$\downarrow$$

$$\hat{B}_1 \pm (2.201)\left(\sqrt{0.990788754} / [45580 - (738)^2/13]^{1/2}\right)$$

$$(0.18663353?) \pm 0.036093801$$

$$\text{LD } (0.150539737, 0.22272734) \doteq (0.1505, 0.2227)$$

$\boxed{\therefore \text{ the true value of } B_1 \text{ lies between } (0.1505, 0.2227)}$

i) Setup an ANOVA table:

$$TSS = S_{yy} = \sum y_i^2 - (\sum y_i)^2/n = 878 - 98^2/13 \doteq 139.2307692 \doteq \boxed{139.23077}$$

$$SSR = S_{xy}^2/S_{xx} = \frac{\left[\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right]^2}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{\left[6251 - \frac{(738)(98)}{13}\right]^2}{45580 - 738^2/13}$$

$$= \frac{\left[687\tfrac{8}{13}\right]^2}{3684\,4/13} \doteq 128.3320927 \doteq \underline{128.33209}$$

$SSE = 10.89868$ (See Part 1.f)

$$MSR = \frac{SSR}{df_R} = \frac{SSR}{1} = SSR = (S_{xy})^2/S_{xx} \doteq 128.33209 \quad | \quad df_R = \#\text{Slopes} = 1$$

$$MSE = \frac{SSE}{df_E} = s^2 \doteq \underline{0.990788754 \doteq 0.99079} \quad | \quad df_E = n-2$$

$$F = \frac{MSR}{MSE} = \frac{128.3321}{0.990788754} \doteq 129.53$$

this Gives us the ANOVA Table as Follows:

	df	SS	MS	F
Model	1	128.33209	128.33209	129.53
Error	11	10.89868	0.99079	
Total	12	139.23077		

we are testing if there's a linear relationship, we will use the F-test:

I) $H_0: B_1 = 0$, $H_a: B_1 \neq 0$

II) F-test: $F \doteq 129.53$

III) Rejection Region: $F_{\alpha(MSR, MSE)} = F_{0.05(1,11)} = \boxed{4.84}$

$4.84 < F = 129.53$ $\therefore$ we must reject the null hypothesis

IV) $\therefore$ Since $F > 4.84$ we reject H_0, & Conclude at 95% Confidence that there is a linear relationship between the evaluation Procedure & the precentage of Market Shares.

j) Find the values of the Coefficient of Correlation, r, & the coefficient of determination, r^2, & interpret their meanings in this problem:

$$r^2 = \frac{SSR}{TSS} = \frac{\left(\frac{S_{xy}^2}{S_{xx}}\right)}{S_{yy}} = \frac{128.3320929}{139.2307692} \doteq \underline{0.921722214} \doteq 0.9217 \text{ or } \boxed{92.17\%}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

$$= \frac{6251 - \frac{(738)(98)}{13}}{\sqrt{\left(45580 - \frac{738^2}{13}\right)\left(878 - \frac{98^2}{13}\right)}}$$

$$\doteq 0.960063651 \doteq \underline{0.96} \text{ or } 96\%$$

Alternatively:

$$r = \sqrt{r^2} = \sqrt{0.921722214} \doteq 0.960063651 \doteq 0.96 \text{ or } \boxed{96\%}$$

$\therefore$ 92.17% of the total variation in the experiment is explained by the model

↳ thus, Approximatly 7.83% is explained by error $(1-0.9217 = 7.83\%)$

high r^2 $\therefore$ the SLR model's pretty good

Since r is 96%, we can state there's a Strong positive Correlation

Hilroy

(2)

a) Find a 95% Confidence interval for the mean value of the response Variable:

Confidence Interval:

$$\hat{y} \pm t_{n-2;\alpha/2} \cdot S\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$$

$t_{11;0.025} = 2.201$

$n = 13$

$x_p = 79$

$\bar{x} = \frac{\sum x_i}{n} = \frac{738}{13}$

$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 45580 - \frac{738^2}{13}$

[see 1.c] $\hat{y} = [\frac{98}{13} - (\frac{8939}{47896})(\frac{738}{13})] + (\frac{8939}{47896})(79) = 11.68746868$

from 1.f, we see $S^2 = 0.990788754$ so $S = \sqrt{S^2} = \sqrt{0.990788754}$

$\doteq 0.995383722$

thus:

$11.68746868 \pm (2.201)(0.995383722)(1/13 + (79 - \frac{738}{13})^2/(45580 - \frac{738^2}{13}))^{1/2}$

$11.68746868 \pm 1.006502951$

$(10.68096573, 12.69397163)$

$\approx (10.6810, 12.6940)$

So;

the 95% Confidence interval is $(10.6810, 12.6940)$

Prediction Interval:

$$\hat{y} \pm t_{n-2;\alpha/2} \cdot S\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}+1}$$

using the values defined above:

$11.68746868 \pm (2.201)(0.990788754)(1+\frac{1}{13}+(79-\frac{738}{13})^2/(45580-\frac{738^2}{13}))^{1/2}$

$11.68746868 \pm 2.410980344$

$(9.276488338, 14.09844903)$

$\approx (9.2765, 14.0984)$, So this is the 95% value for the P.I.

the width of the Confidence interval & the Prediction interval:

CI: $12.6940 - 10.6810 = 2.013$

PI: $14.0984 - 9.2765 = 4.8219$

Since the Prediction interval has a +1 in the Square root, it's wider then the Confidence interval, Since the addition of 1 increases the Std. error.

this is because the CI is more Conservative, whereas the PI is more estimative. Adding one in the PI increases the Std. error & ∴ increases the width of the interval Since more people are included.

b) On page 4 of the SAS output we see 95% CL mean & 95% CL Predict. These columns show the intervals. On obs #14, we see the total for the whole data set. The left element's the lower-bound of the interval, whereas the right's the upper-bound:

CI: (10.6810, 12.6940)

PI: (9.2765, 14.0984)

The Calculated values for the intervals are;

CI: (10.6810, 12.6940)

PI: (9.2765, 14.0984)

∴ the Calculated Intervals is the Same as the SAS outputted intervals to the fourth decimal Place.

③ Preform a residual analysis to check the SLR model assumptions using SAS:

Page 6 of the SAS output Shows a residual graph:

- No Pattern amaung the y-values is noted, thus Assumption 2 is not violated
  ∴ the graph tests the assumption of independence, since no pattern means no violation
- The residual vs. our x's tests the assumption that the y's have constant Varience: (Pg. 7 of SAS output)
  The y's have no pattern
  ∴ There's no violation of Assumption 3
- The residual histogram tests for normallity, Since the histogram is Skew right, this is a Violation of the assumption of normality: (Page 8)
  ∴ Assumption 4's violated Since the histogram's not normal

Only one assumtion (Number 4) is violated, we Could fix this situation by changing the data points to a logorithmic, exponential, or Square-root Curve. This would be Capable of Changing the data to allow Assumption 4 to hold.

thus, Considering the assumptions are fine, the R^2 is high, the SSE is high Compared to the SSE, & the relationship's Statistically Significant, this leads us to believe the model is pretty good.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: percent

Number of Observations Read	13
Number of Observations Used	13

ANOVA matches Q.1, P.i

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	128.33209	128.33209	129.53	<.0001
Error	11	10.89868	0.99079		
Corrected Total	12	139.23077			

SSE SSR MSR F MSE, S^2 TSS

Root MSE	0.99538	R-Square	0.9217
Dependent Mean	7.53846	Adj R-Sq	0.9146
Coeff Var	13.20407		

r^2

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.05658	0.97102	-3.15	0.0093
eval	1	0.18663	0.01640	11.38	<.0001

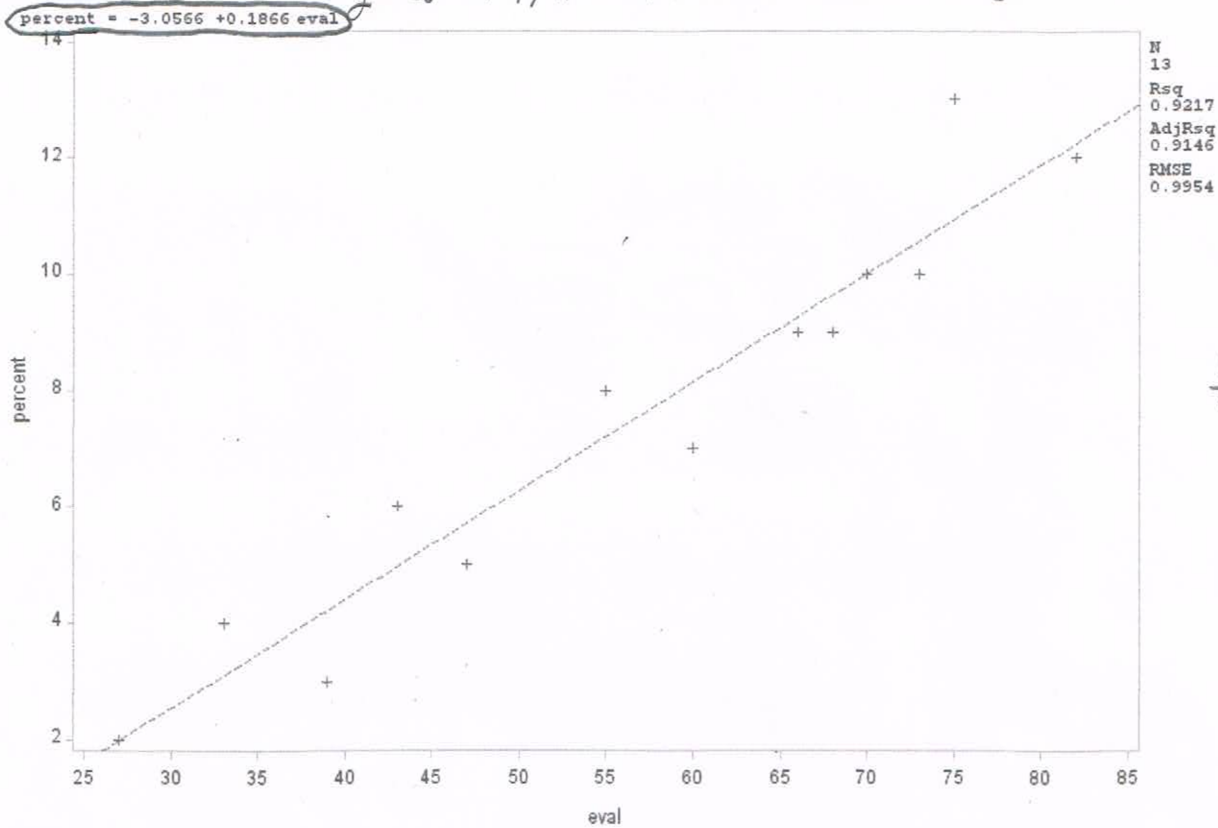
t-value in q.1.g.

Connor, 101041125

a) Scatter diagram

The REG Procedure

$y=\hat{B}_0+\hat{B}_1 x_1$, this matches the results in c)



$y=mx+b$

Connor, 101041125

$y = B_0 + B_1 X_1$

$X_1 \rightarrow$ Score

$y \rightarrow$ % market shares

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: percent

Number of Observations Read	14
Number of Observations Used	13
Number of Observations with Missing Values	1

ANOVA matches Q.1, P. i

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	128.33209 SSR	128.33209 MSR	129.53 F-value	<.0001
Error	11	10.89868 SSE	0.99079 MSE, s^2		
Corrected Total	12	139.23077 TSS			

Root MSE	0.99538	R-Square	0.9217 r^2
Dependent Mean	7.53846	Adj R-Sq	0.9146
Coeff Var	13.20407		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.05658	0.97102	-3.15	0.0093
eval	1	0.18663	0.01640	11.38	<.0001

question 1.g's t-value

Connor, 101041125

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: percent

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	2	1.9825	0.5608	0.7481	3.2169	-0.5321	4.4972	0.0175
2	3	4.2221	0.4014	3.3386	5.1056	1.8599	6.5844	-1.2221
3	10	10.5677	0.3835	9.7236	11.4117	8.2199	12.9155	-0.5677
4	9	9.2612	0.3148	8.5683	9.9542	6.9634	11.5590	-0.2612
5	4	3.1023	0.4776	2.0510	4.1536	0.6723	5.5323	0.8977
6	6	4.9687	0.3567	4.1837	5.7536	2.6415	7.2959	1.0313
7	5	5.7152	0.3192	5.0127	6.4177	3.4145	8.0159	-0.7152
8	8	7.2083	0.2776	6.5973	7.8192	4.9338	9.4827	0.7917
9	7	8.1414	0.2811	7.5227	8.7601	5.8649	10.4179	-1.1414
10	9	9.6345	0.3319	8.9041	10.3649	7.3251	11.9439	-0.6345
11	10	10.0078	0.3511	9.2349	10.7806	7.6846	12.3309	-0.007767
12	13	10.9409	0.4069	10.0453	11.8366	8.5741	13.3078	2.0591
13	12	12.2474	0.4974	11.1526	13.3421	9.7982	14.6965	-0.2474
14	.	11.6875	0.4573	10.6810	12.6940	9.2765	14.0984	.

C.I. P.I.

Sum of Residuals	0
Sum of Squared Residuals	10.89868
Predicted Residual SS (PRESS)	15.06091

Connor, 101041125

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: percent

Number of Observations Read	14
Number of Observations Used	13
Number of Observations with Missing Values	1

this ANOVA table matches the Calculations made in Part 1.i.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	128.33209 SSR	128.33209	129.53 MSR	<.0001 F-value
Error	11	10.89868 SSE	0.99079 MSE s^2		
Corrected Total	12	139.23077 TSS			

Same as f)

Root MSE	0.99538	R-Square	0.9217 r^2
Dependent Mean	7.53846	Adj R-Sq	0.9146
Coeff Var	13.20407		

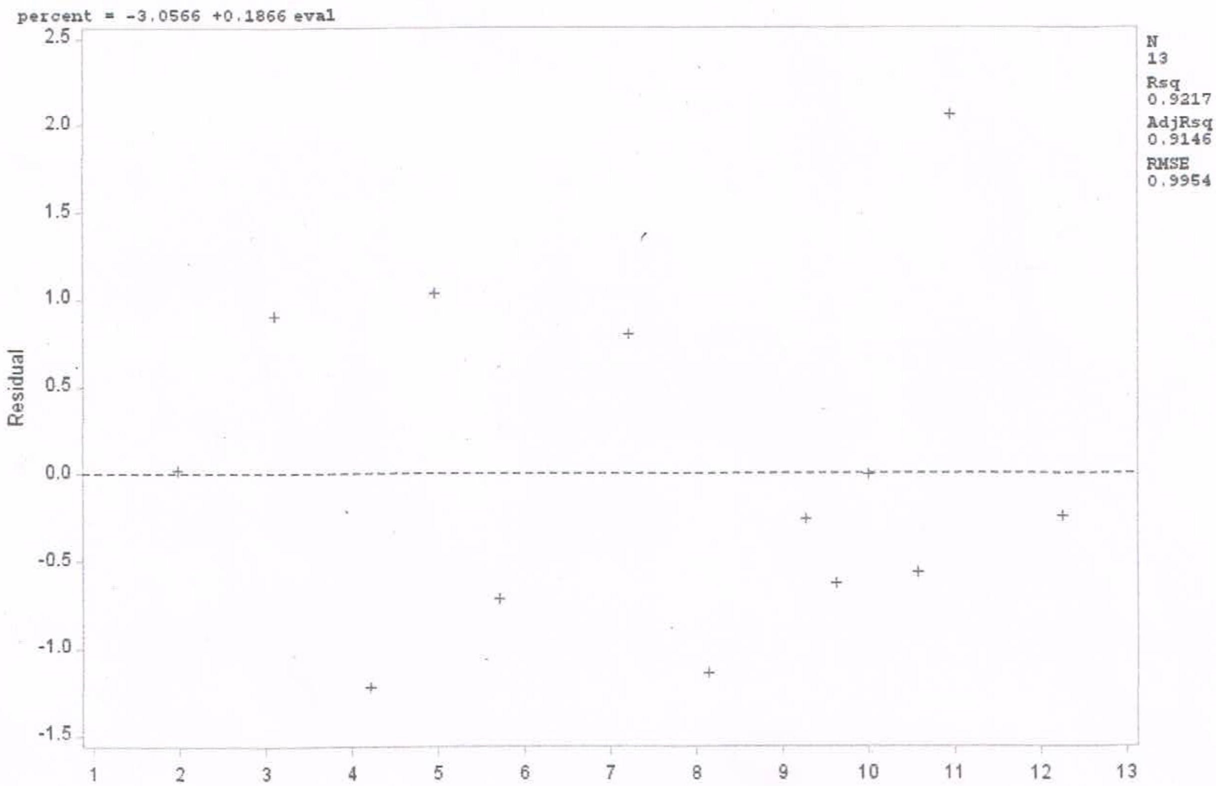
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.05658	0.97102	-3.15	0.0093
eval	1	0.18663	0.01640	11.38	<.0001

Connor, 101041125

Same as g) Step II
thus this is the t-value

The REG Procedure



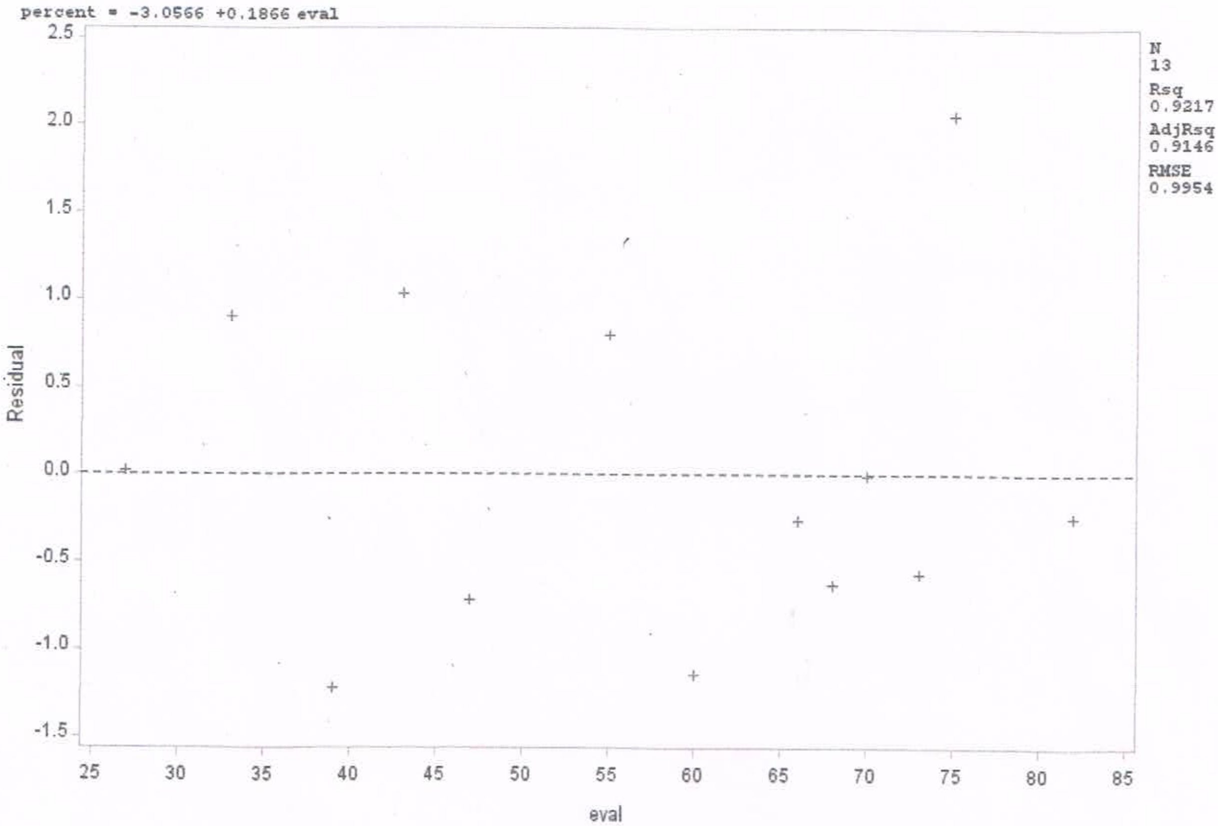
Connor, 101041125

This graph tests the assumption of independence

There's no pattern among the y-values

∴ Assumption two's not violated

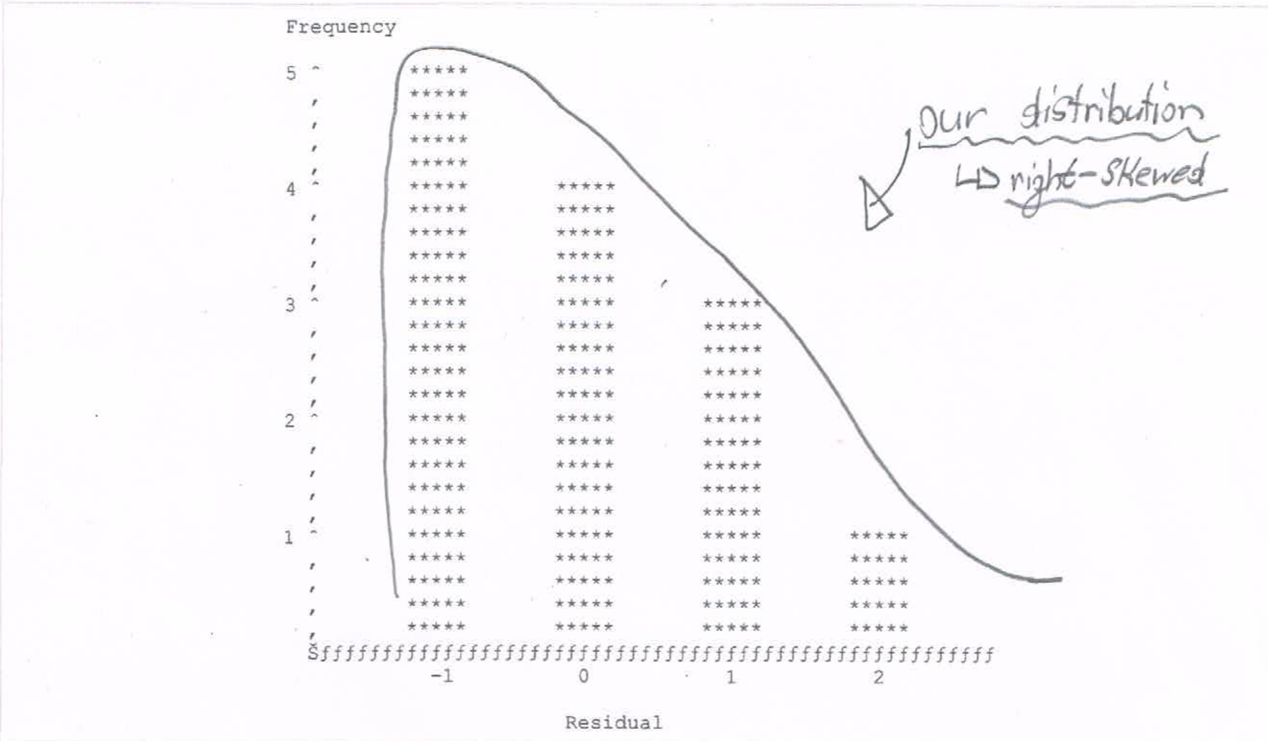
The REG Procedure



Connor, 101041125

This graph looks at the varience of y's by showing our residuals vs. our x's
The y's have no Pattern
∴ Assumption 3's not violated

The SAS System



Connor, 101041125

The residual histogram tests for normallity, Since the histogram is Skew-right this violates the assumption of normallity. However, We could fix this by changing the scale of the data points to logerithmic, exponential, or Square-root.

So Although the histogram violates assumption 4, the experement can be altered easily to So it does not.