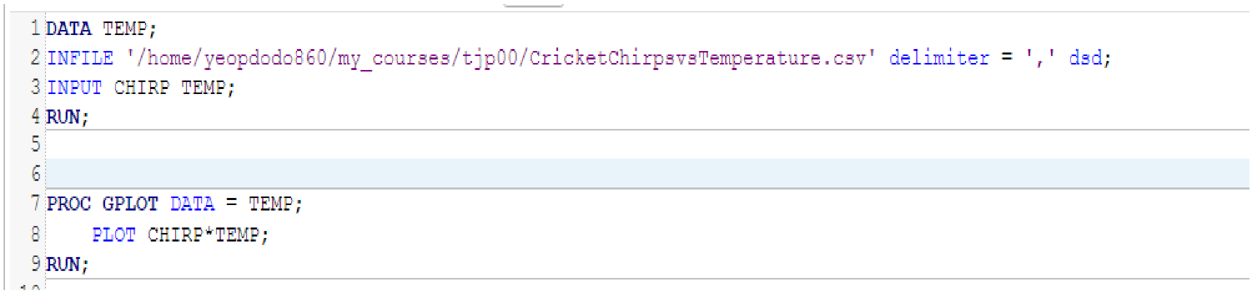
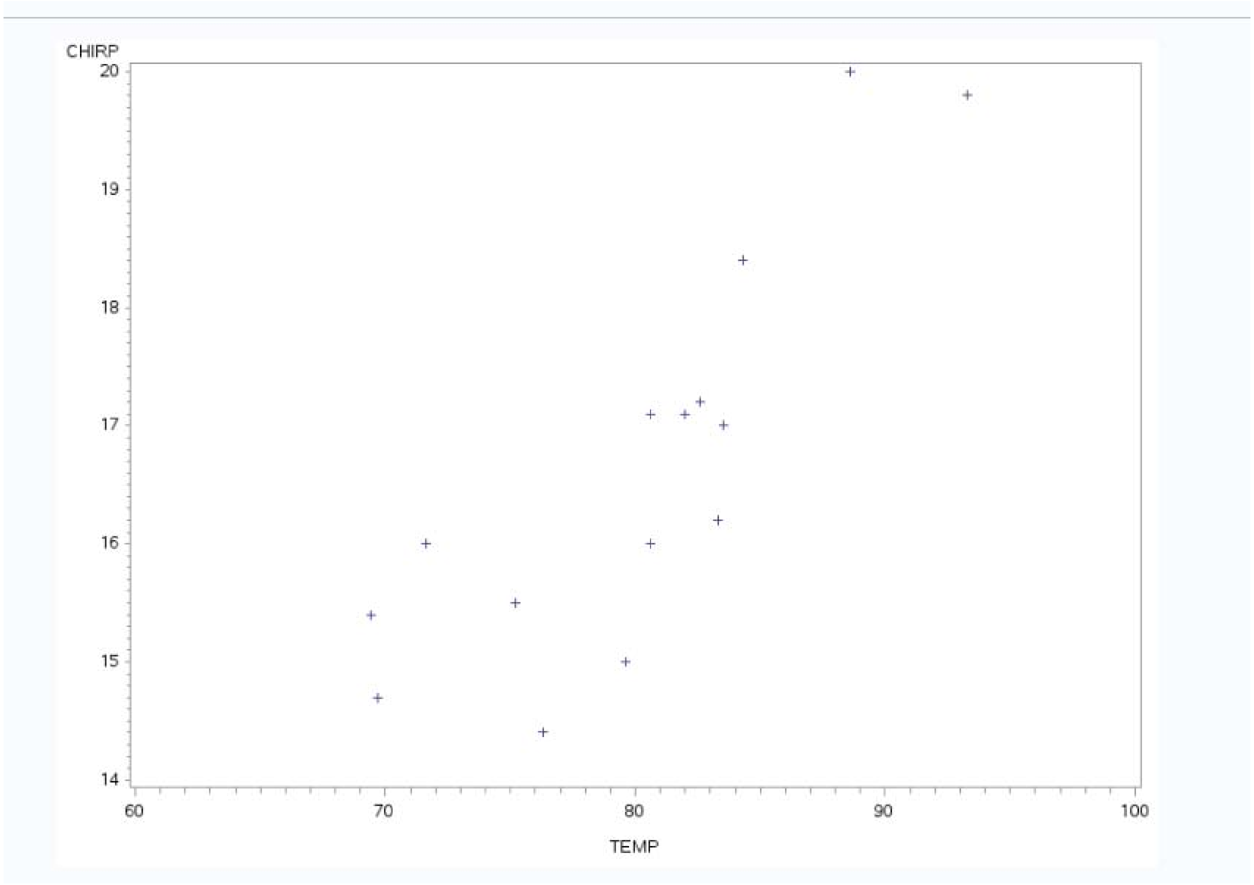
1a. Use SAS to create a scatterplot. What initial impressions do you have about correlation?

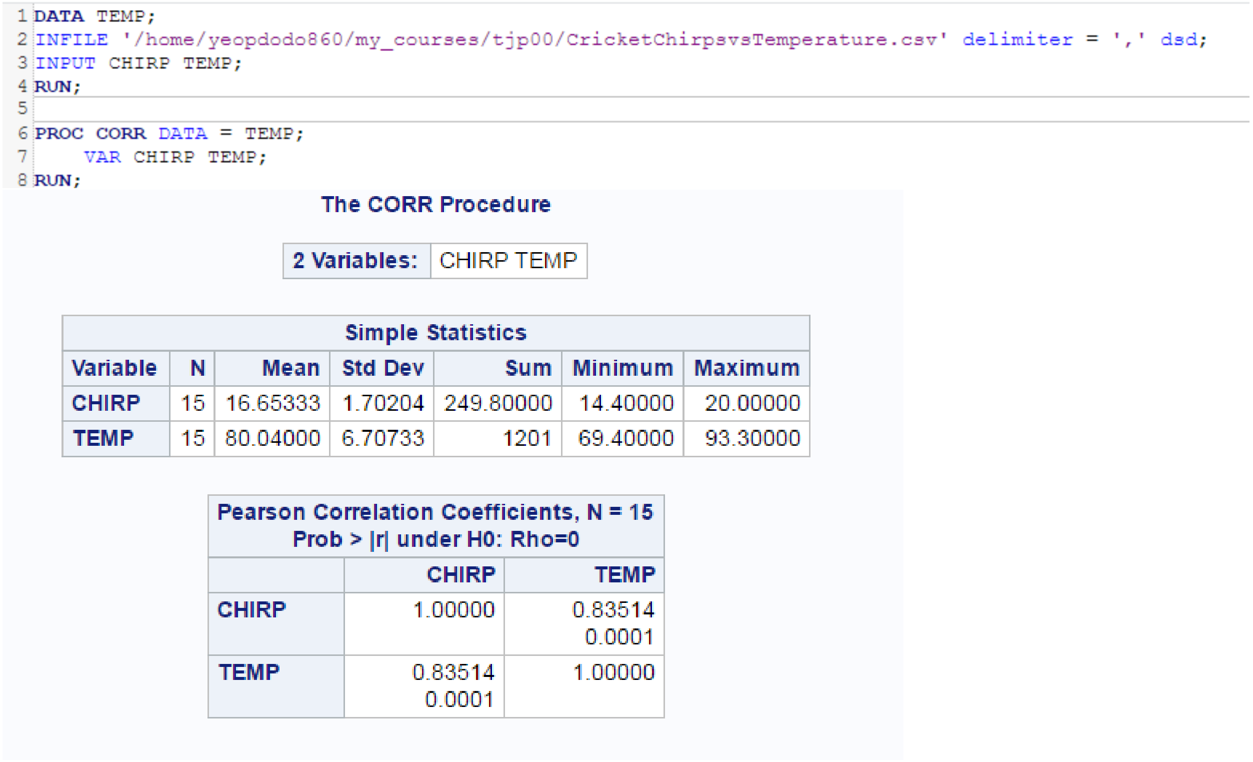


```
DATA TEMP;
INFILE '/home/yeopdodo860/my_courses/tjp00/CricketChirpsvsTemperature.csv' delimiter = ',' dsd;
INPUT CHIRP TEMP;
RUN;


PROC GPLOT DATA = TEMP;
    PLOT CHIRP*TEMP;
RUN;
```



They seem to be correlated since the spots from the plot are close to each other

b. Do a PROC Corr. State the value of the correlation coefficient, and state whether it indicates a weak, moderate or strong correlation.



```
DATA TEMP;
INFILE '/home/yeopdodo860/my_courses/tjp00/CricketChirpsvsTemperature.csv' delimiter = ',' dsd;
INPUT CHIRP TEMP;
RUN;

PROC CORR DATA = TEMP;
    VAR CHIRP TEMP;
RUN;
```

The CORR Procedure

2 Variables:	CHIRP TEMP

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CHIRP	15	16.65333	1.70204	249.80000	14.40000	20.00000
TEMP	15	80.04000	6.70733	1201	69.40000	93.30000

Pearson Correlation Coefficients, N = 15
Prob > |r| under H0: Rho=0

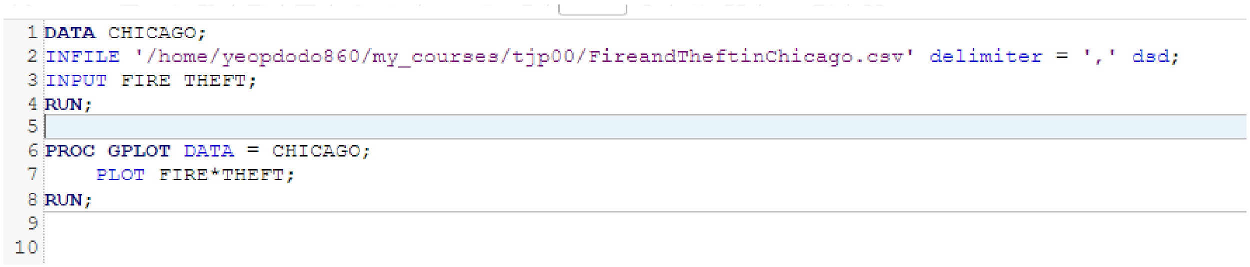
	CHIRP	TEMP
CHIRP	1.00000	0.83514 0.0001
TEMP	0.83514 0.0001	1.00000

They have a strong correlation because they have about 0.84 correlation value

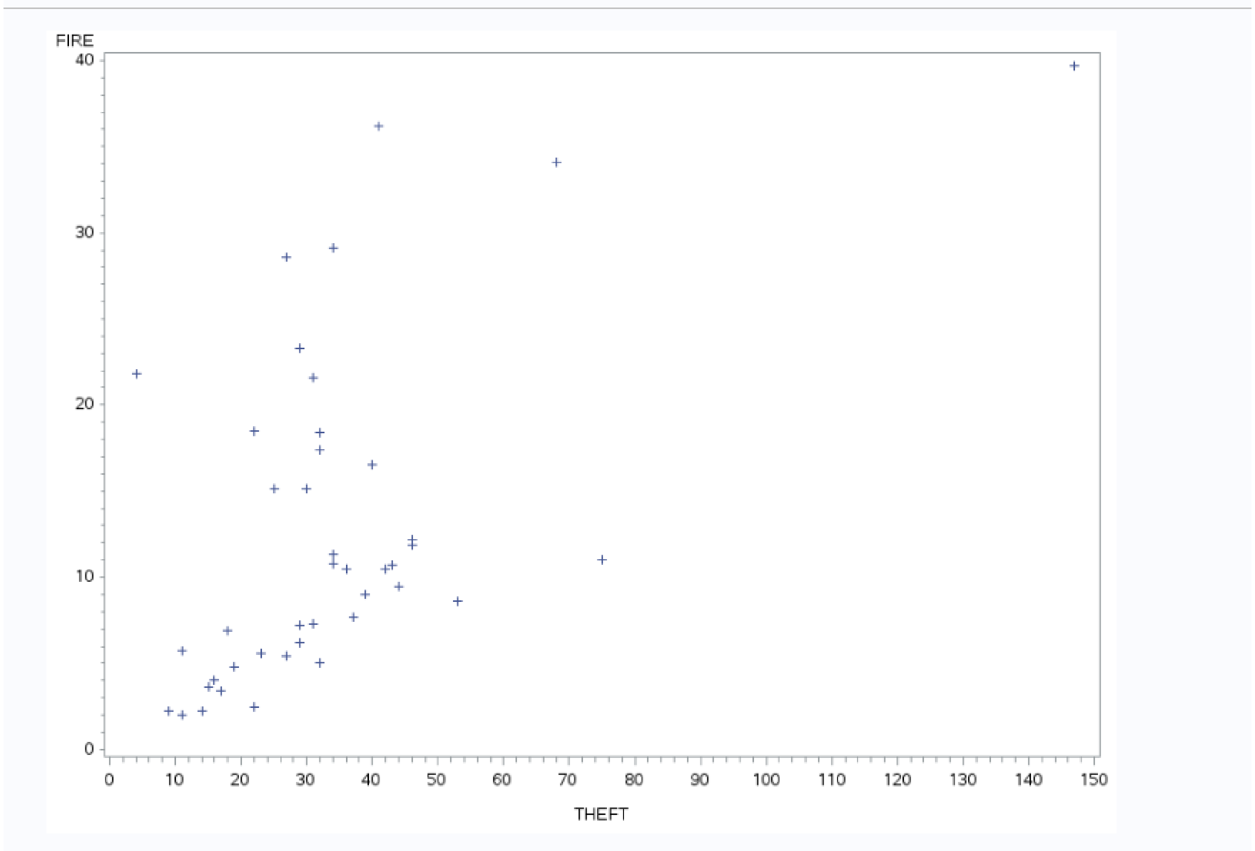
c. State the p-value. What does it tell you about the statistical significance of the correlation?

The null hypothesis should not be rejected since p value is 0.83514 > significance level

2. a. Use SAS to create a scatterplot. What initial impressions do you have about correlation?

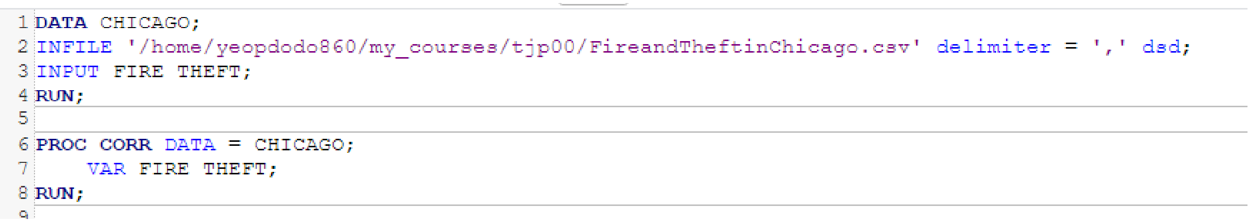


```
DATA CHICAGO;
INFILE '/home/yeopdodo860/my_courses/tjp00/FireandTheftinChicago.csv' delimiter = ',' dsd;
INPUT FIRE THEFT;
RUN;

PROC GPLOT DATA = CHICAGO;
    PLOT FIRE*THEFT;
RUN;


```

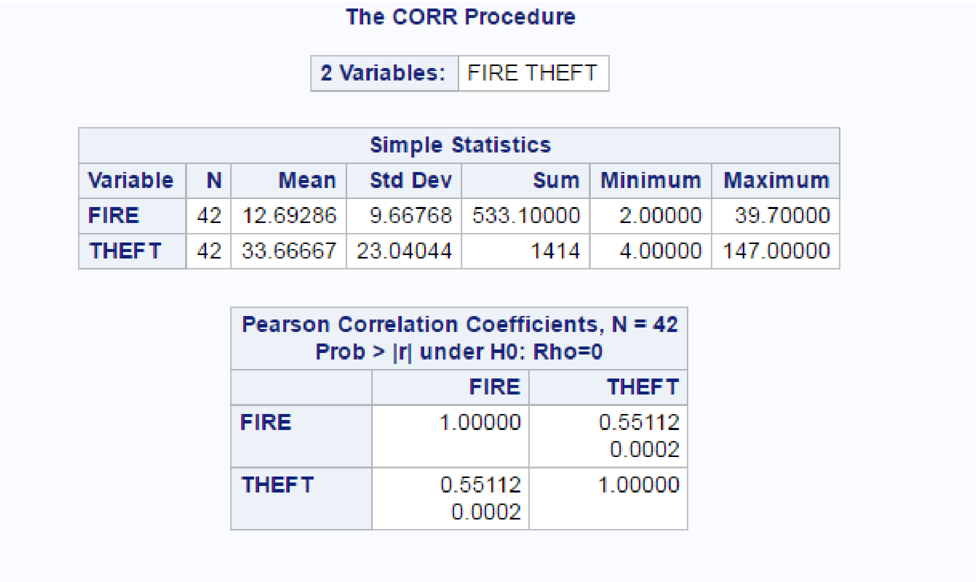


They seem to be correlated in some parts but not in some parts.

b. Do a PROC Corr. State the value of the correlation coefficient, and state whether it indicates a weak, moderate or strong correlation.



```
DATA CHICAGO;
INFILE '/home/yeopdodo860/my_courses/tjp00/FireandTheftinChicago.csv' delimiter = ',' dsd;
INPUT FIRE THEFT;
RUN;

PROC CORR DATA = CHICAGO;
    VAR FIRE THEFT;
RUN;
```



The CORR Procedure

2 Variables:	FIRE THEFT

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
FIRE	42	12.69286	9.66768	533.10000	2.00000	39.70000
THEFT	42	33.66667	23.04044	1414	4.00000	147.00000

Pearson Correlation Coefficients, N = 42
Prob > |r| under H0: Rho=0

	FIRE	THEFT
FIRE	1.00000	0.55112 0.0002
THEFT	0.55112 0.0002	1.00000

They seem to have a moderate correlation since it has 0.55.

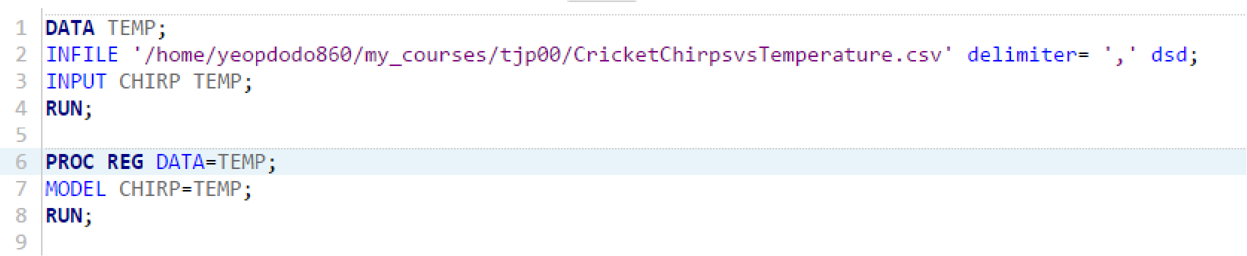
c. State the p-value. What does it tell you about the statistical significance of the correlation?

The null hypothesis should still not be rejected since p value is 0.55 > significance level.

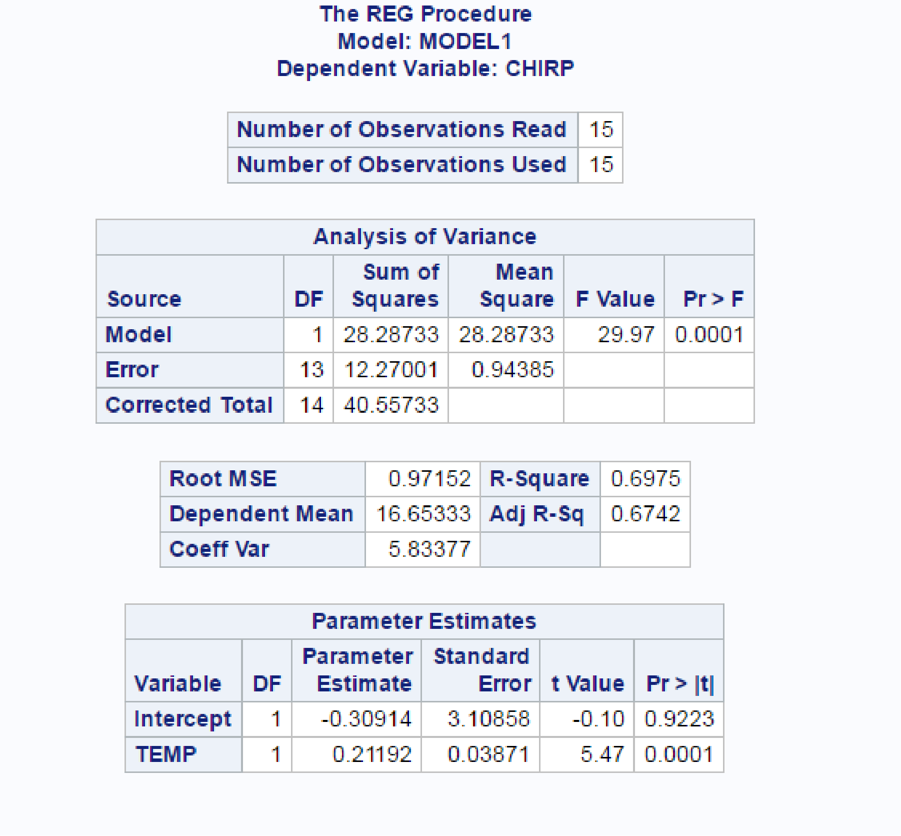
Part 2

1.

a. Do a linear regression using PROC REG. State the estimated linear regression equation.



```
DATA TEMP;
INFILE '/home/yeopdodo860/my_courses/tjp00/CricketChirpsvsTemperature.csv' delimiter= ',' dsd;
INPUT CHIRP TEMP;
RUN;

PROC REG DATA=TEMP;
MODEL CHIRP=TEMP;
RUN;

```



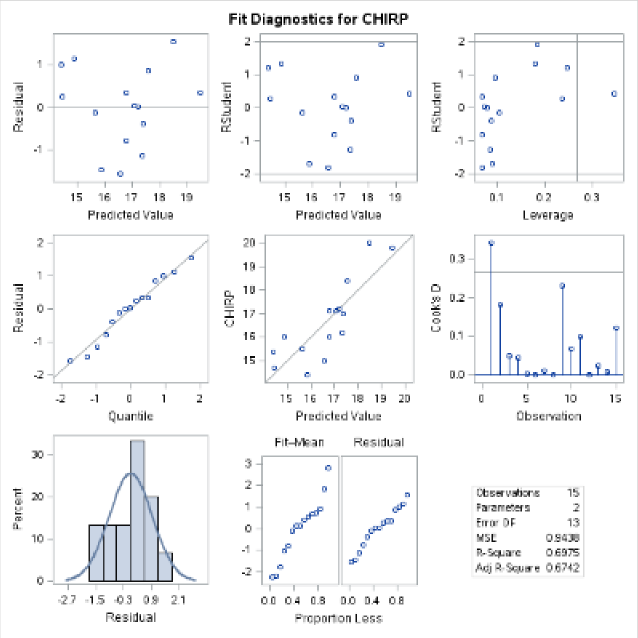
The REG Procedure
Model: MODEL1
Dependent Variable: CHIRP

Number of Observations Read	15
Number of Observations Used	15

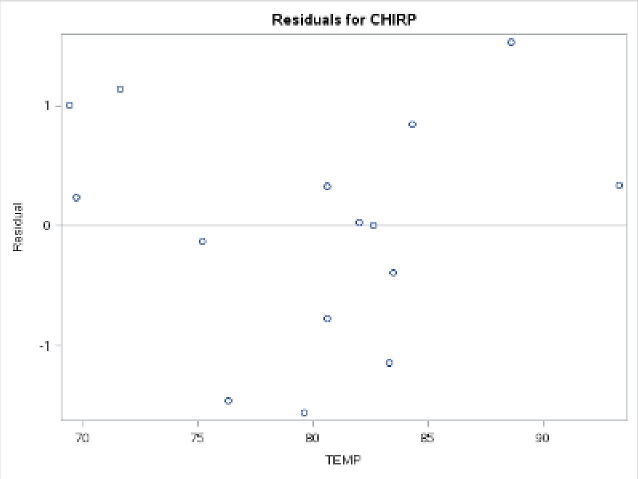
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	28.28733	28.28733	29.97	0.0001
Error	13	12.27001	0.94385		
Corrected Total	14	40.55733			

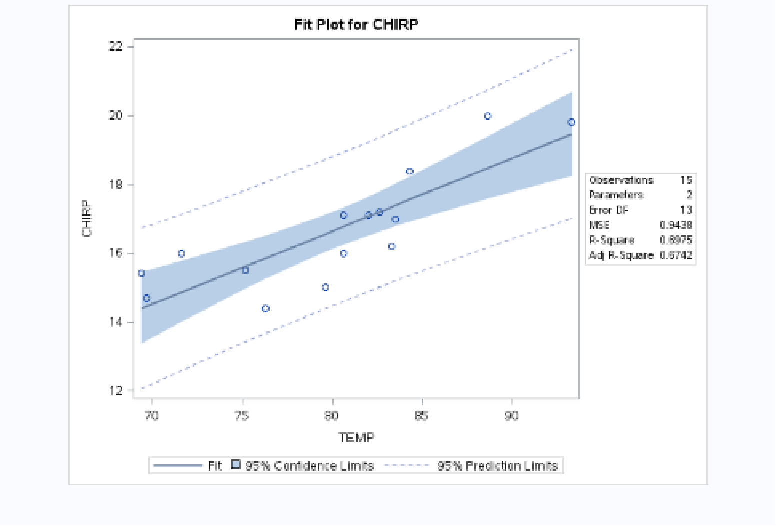
Root MSE	0.97152	R-Square	0.6975
Dependent Mean	16.65333	Adj R-Sq	0.6742
Coeff Var	5.83377		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.30914	3.10858	-0.10	0.9223
TEMP	1	0.21192	0.03871	5.47	0.0001



```
Observations       15
Parameters          2
Error DF           13
MSE            0.9438
R-Square       0.6975
Adj R-Square   0.6742
```





b. Interpret the regression output and state whether the following indicate that the regression equation is reliable and should be used.

i. The p-value for the ANOVA table.

p value is 0.83514

ii. R –Square value. Include an interpretation of what this number tells us.

R-Square	0.6975

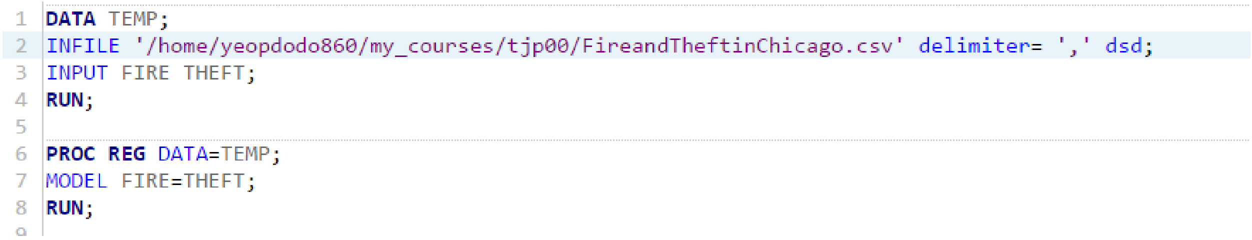
Approximately 67.75% of the variability in chirp can be explained by or attributed to variability in temp.

iii. The p-value for parameter (β_0 and β_1) estimates. Include a conclusion about the statistical significance of the linear regression equation.

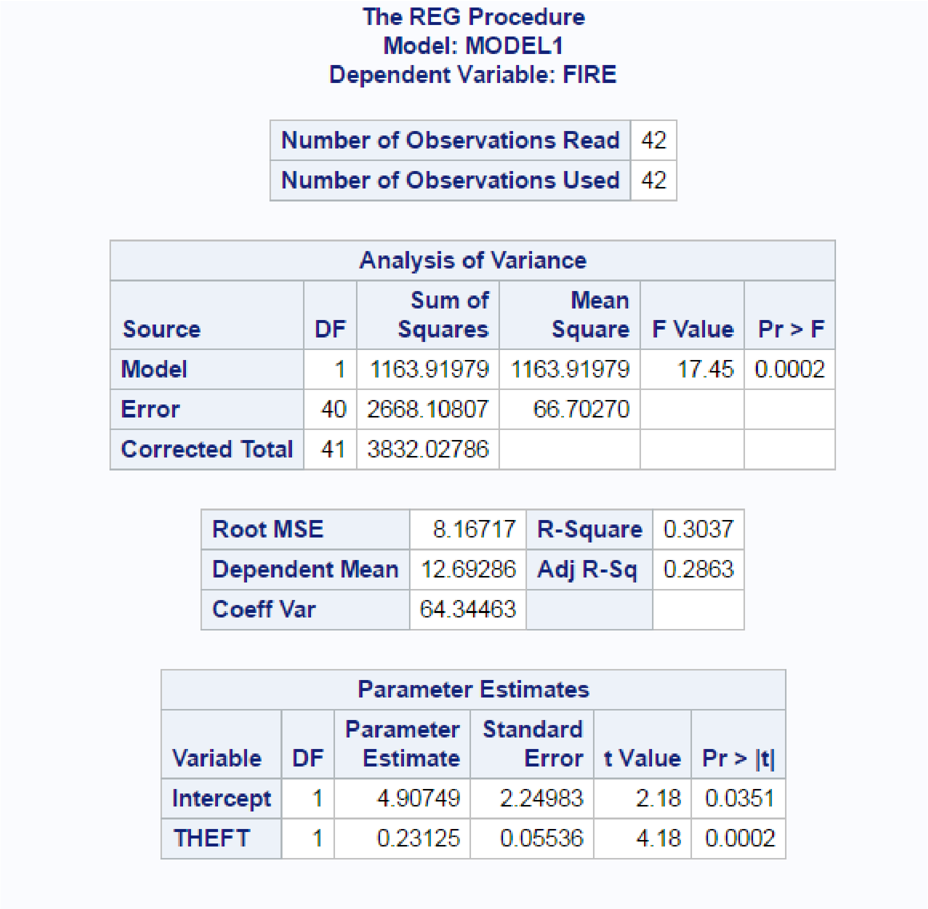
$\beta_0 =$ 0.9223 ,   changes in the predictor is not related to changes in the response since larger than alpha 0.05. less meaningful to the model

$\beta_1 =$ 0.0001   changes in the predictor is related to changes in the response since less than alpha 0.05. more meaningful to the model

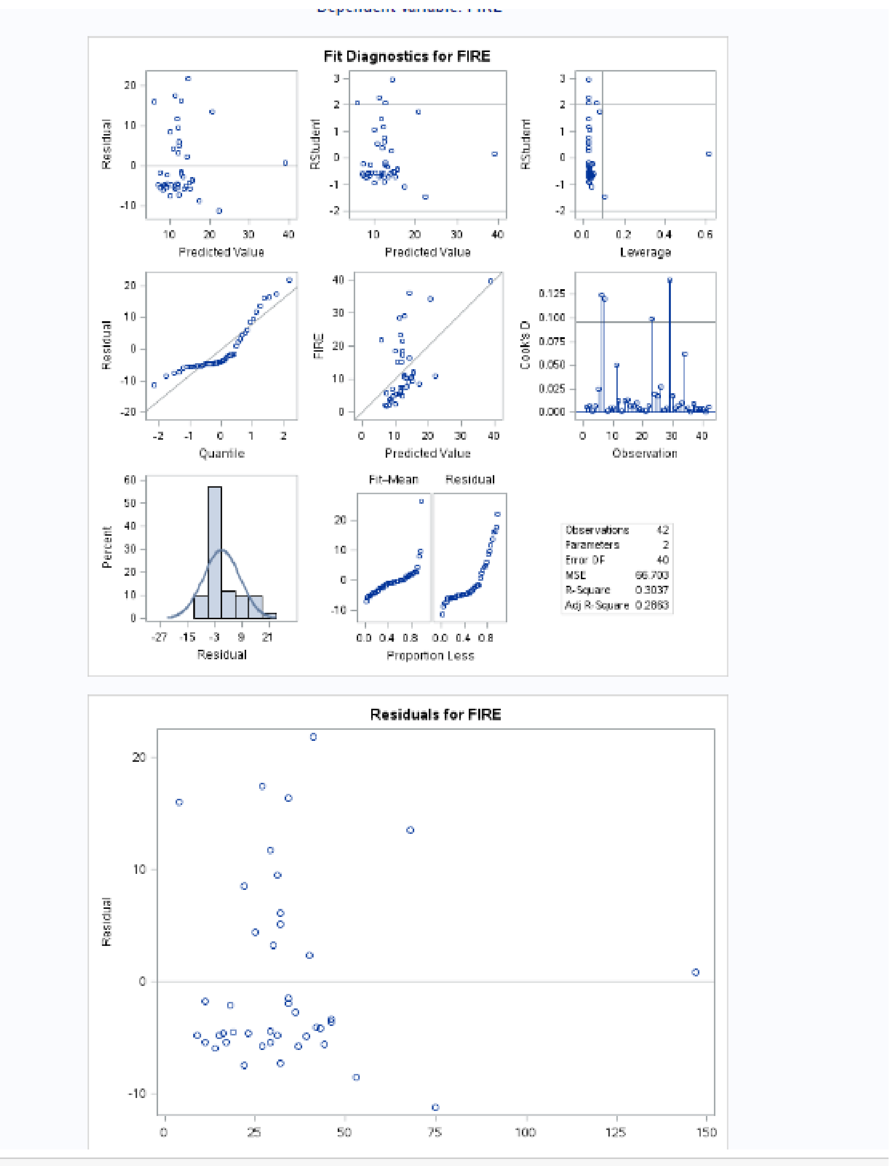
1.a. Do a linear regression using PROC REG. State the estimated linear regression equation.



```
DATA TEMP;
INFILE '/home/yeopdodo860/my_courses/tjp00/FireandTheftinChicago.csv' delimiter= ',' dsd;
INPUT FIRE THEFT;
RUN;

PROC REG DATA=TEMP;
MODEL FIRE=THEFT;
RUN;
```



```
                        The REG Procedure
                          Model: MODEL1
                    Dependent Variable: FIRE

              Number of Observations Read   42
              Number of Observations Used   42

                      Analysis of Variance
                             Sum of         Mean
Source             DF       Squares       Square   F Value   Pr > F
Model               1    1163.91979   1163.91979     17.45   0.0002
Error              40    2668.10807     66.70270
Corrected Total    41    3832.02786

          Root MSE            8.16717   R-Square   0.3037
          Dependent Mean     12.69286   Adj R-Sq   0.2863
          Coeff Var          64.34463

                       Parameter Estimates
                    Parameter   Standard
Variable     DF      Estimate      Error   t Value   Pr > |t|
Intercept     1       4.90749    2.24983      2.18     0.0351
THEFT         1       0.23125    0.05536      4.18     0.0002
```

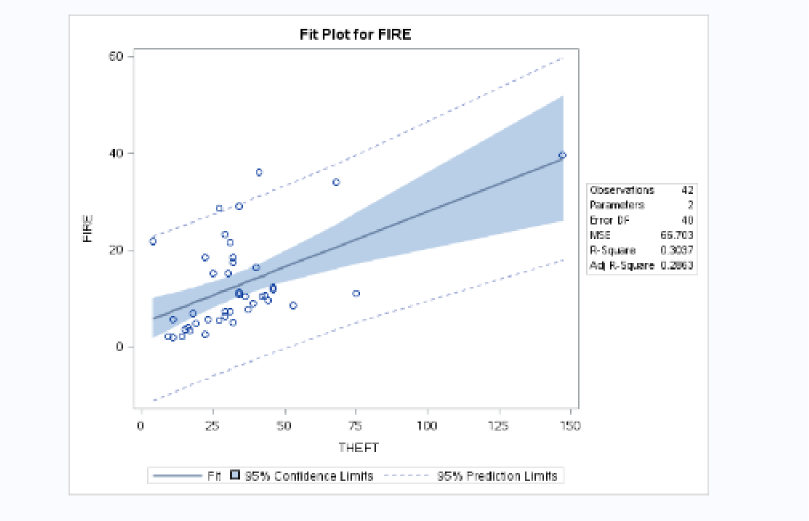


Dependent Variable: FIRE

Fit Diagnostics for FIRE

Observations	42
Parameters	2
Error DF	40
MSE	66.703
R-Square	0.3037
Adj R-Square	0.2863

Residuals for FIRE



b. Interpret the regression output and state whether the following indicate that the regression equation is reliable and should be used.

i. The p-value for the ANOVA table.

p value is 0.55

ii. R –Square value. Include an interpretation of what this number tells us.

R-Square	0.3037

Approximately 30.37% of the variability in fire can be explained by or attributed to variability in theft.

iii. The p-value for parameter (β_0 and β_1) estimates. Include a conclusion about the statistical significance of the linear regression equation.

$\beta_{0} =$ 0.0351, changes in the predictor is related to changes in the response since less than alpha 0.05. more meaningful to the model

$\beta_{1} =$ 0.0002 changes in the predictor is related to changes in the response since less than alpha 0.05. more meaningful to the model

Part 3

1.

a. The correlation coefficient and slope of the estimated linear regression equation resulted in low p-values, indicating that the linear regression model is reliable. Why do you think the p-value for the intercept so high?

Changes in the predictor are not related to changes in the response since larger than alpha 0.05. It is less meaningful to the model.

2.

a. Looking at the scatterplot, what does it tell you about the reliability and usefulness of the linear regression model?

They seem to have many outliers so the linear regression model may not be reliable and useful.

b. What does the value of r^2 tell you about the reliability and usefulness of the linear regression model?

The R- square value of this plot is 0.3037 which give a moderate level of usefulness of the linear regression model.

c. If you did not produce residual plots for Online Assignment #9, do so now.

i. What does the “predicted value vs residual values” plot tell you about the reliability and usefulness of the linear regression model?

The predicted value vs residual values plot shows that the linear regression model is neither really useful nor reliable because most of the plots show that they are not closely related to each other.

ii. What does the normal probability plot tell you about the reliability and usefulness of the linear regression model?

According to the normal distribution, the linear regression model may have outliers so it is not much useful.

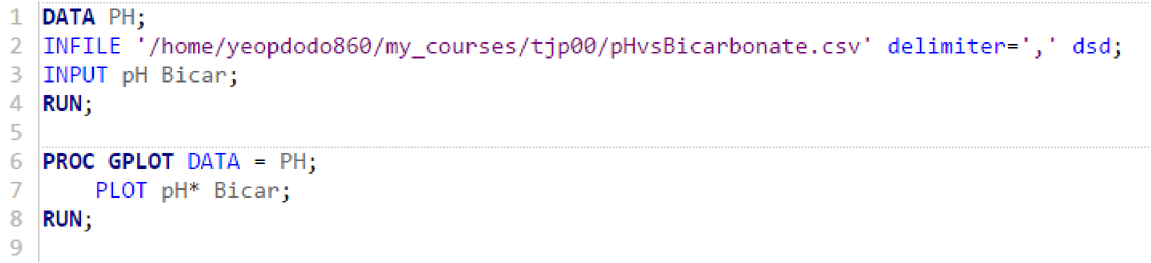
iii. What does the boxplot tell you about the reliability and usefulness of the linear regression model?

The box plot shows that the model definitely has outliers.

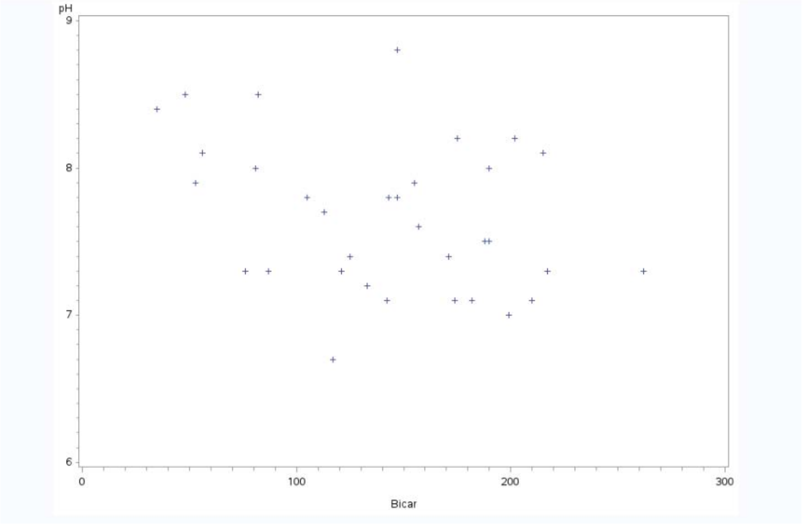
3.

a. Use SAS to create a scatterplot. What initial impressions do you have about correlation?

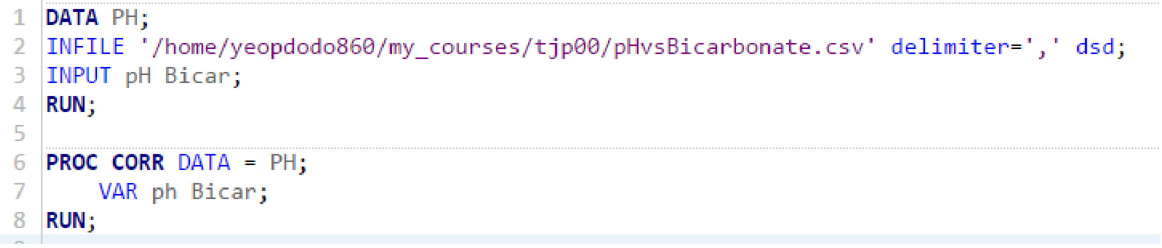
They seem to negatively correlate each other if I imagine a line that average the spots.



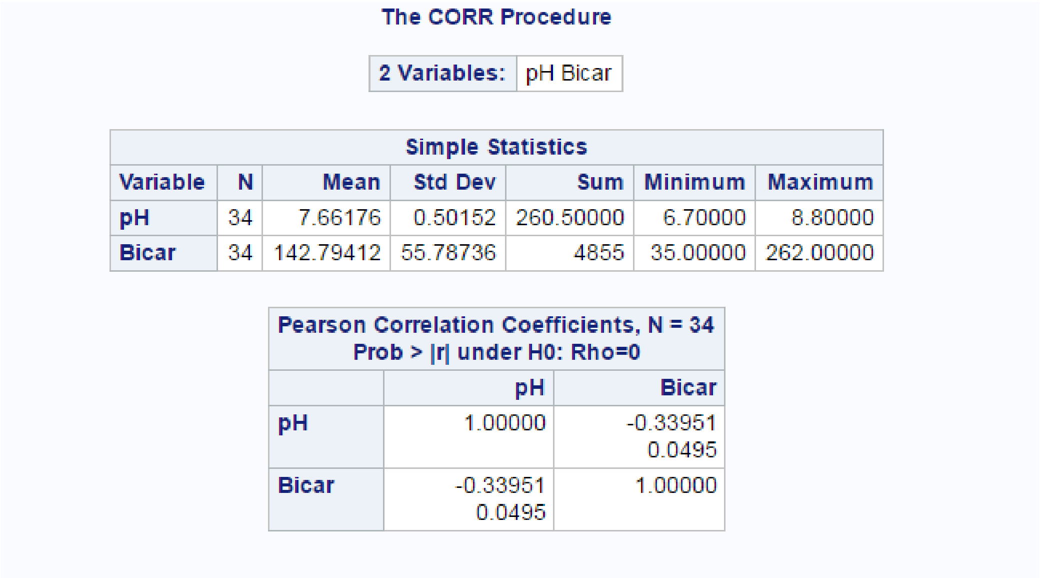
```
DATA PH;
INFILE '/home/yeopdodo860/my_courses/tjp00/pHvsBicarbonate.csv' delimiter=',' dsd;
INPUT pH Bicar;
RUN;

PROC GPLOT DATA = PH;
    PLOT pH* Bicar;
RUN;

```



b. Do a PROC Corr. State the value of the correlation coefficient, and state whether it indicates a weak, moderate or strong correlation.



```
DATA PH;
INFILE '/home/yeopdodo860/my_courses/tjp00/pHvsBicarbonate.csv' delimiter=',' dsd;
INPUT pH Bicar;
RUN;

PROC CORR DATA = PH;
    VAR ph Bicar;
RUN;
```



The CORR Procedure

2 Variables:	pH Bicar

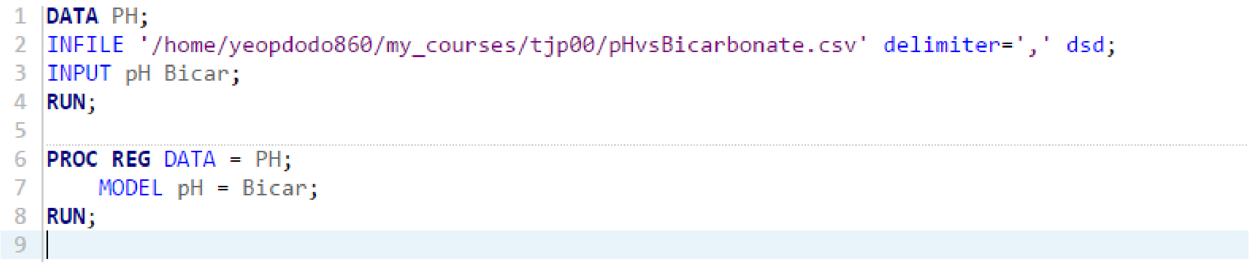
Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
pH	34	7.66176	0.50152	260.50000	6.70000	8.80000
Bicar	34	142.79412	55.78736	4855	35.00000	262.00000

Pearson Correlation Coefficients, N = 34 Prob > |r| under H0: Rho=0		
	pH	**Bicar**
pH	1.00000	-0.33951 0.0495
Bicar	-0.33951 0.0495	1.00000

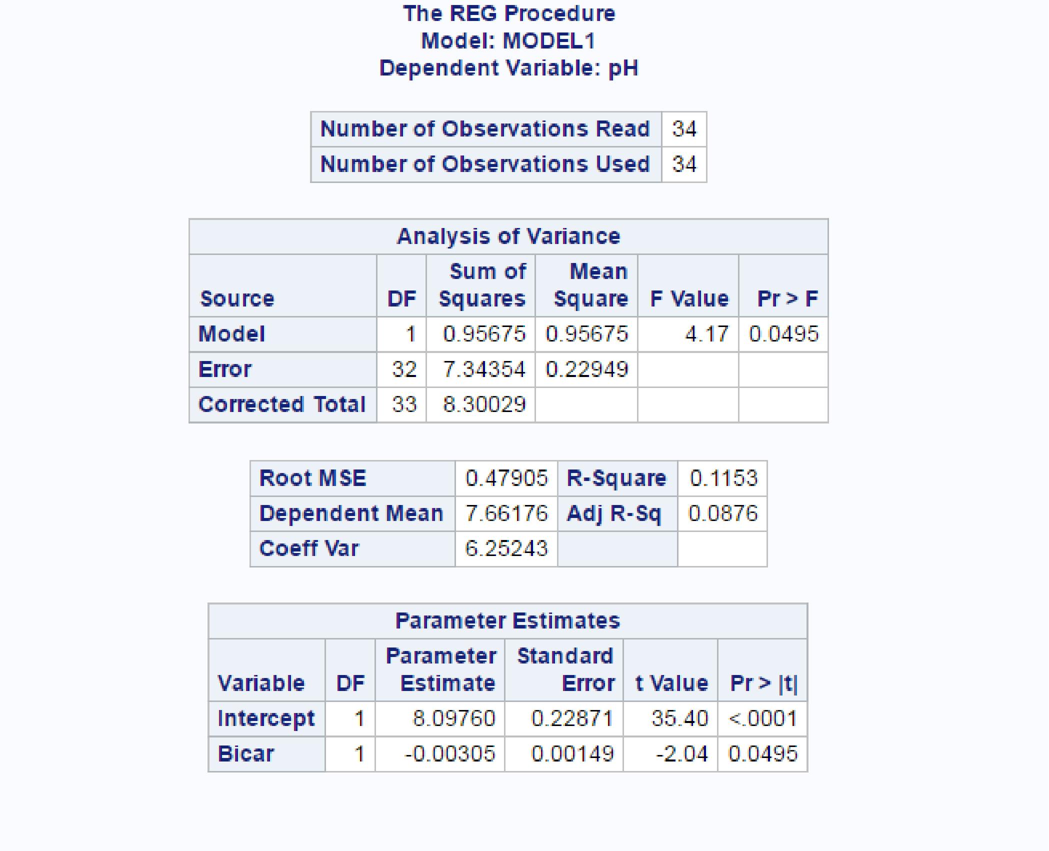
correlation coefficient   -0.33951   0.0495

They have moderate negative correlation.

c. Do a linear regression using PROC REG. State the estimated linear regression equation.



```
DATA PH;
INFILE '/home/yeopdodo860/my_courses/tjp00/pHvsBicarbonate.csv' delimiter=',' dsd;
INPUT pH Bicar;
RUN;

PROC REG DATA = PH;
    MODEL pH = Bicar;
RUN;

```



The REG Procedure
Model: MODEL1
Dependent Variable: pH

Number of Observations Read	34
Number of Observations Used	34

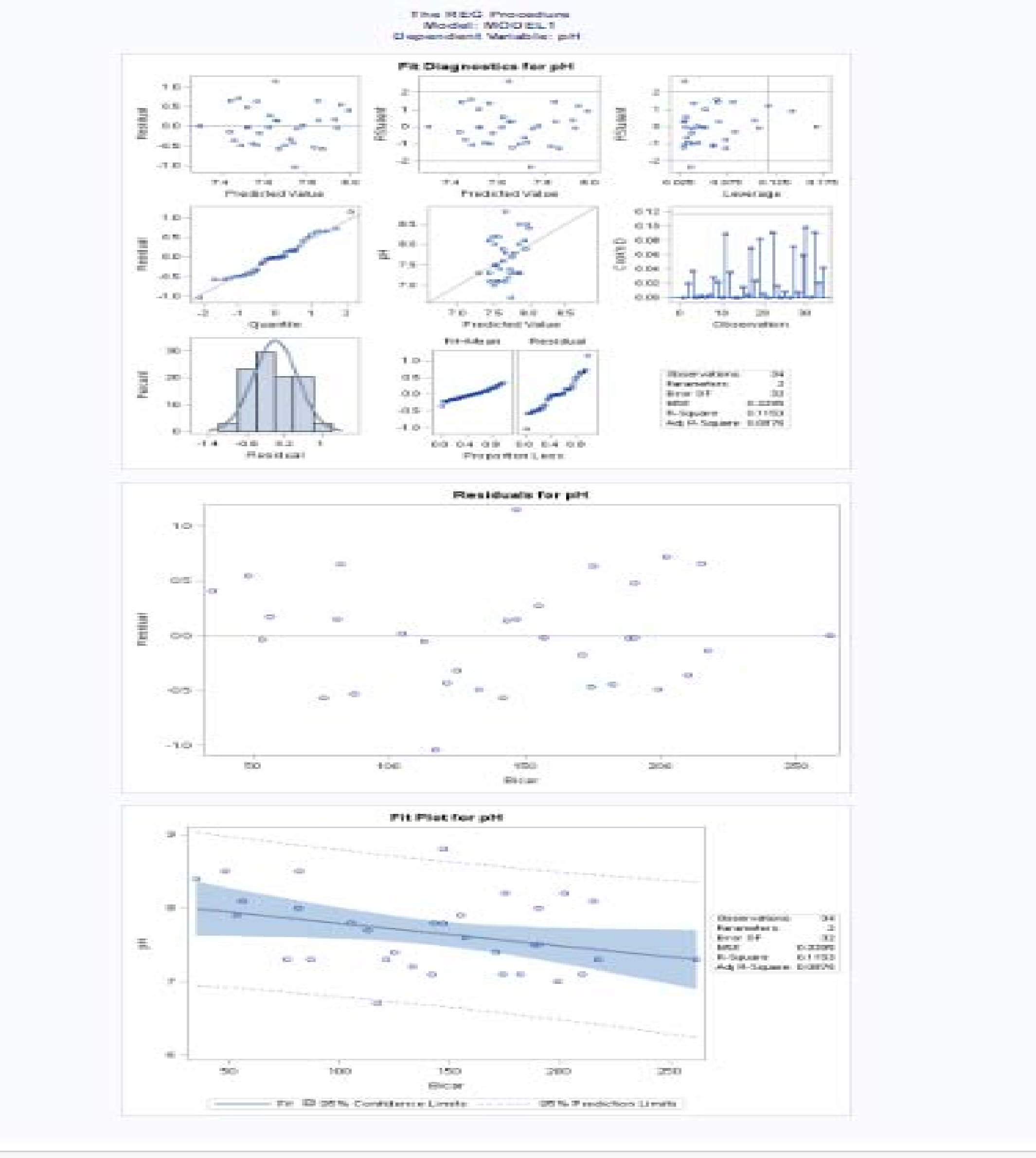
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.95675	0.95675	4.17	0.0495
Error	32	7.34354	0.22949		
Corrected Total	33	8.30029			

Root MSE	0.47905	R-Square	0.1153
Dependent Mean	7.66176	Adj R-Sq	0.0876
Coeff Var	6.25243		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.09760	0.22871	35.40	<.0001
Bicar	1	-0.00305	0.00149	-2.04	0.0495


The REG Procedure
Model: MODEL1
Dependent Variable: pH

Observations	34
Parameters	2
Error DF	32
MSE	0.2395
R-Square	0.1153
Adj R-Square	0.0876

d. Interpret the regression output and state whether the following indicate that the regression equation is reliable and should be used.
i. The p-value for the ANOVA table.

p-value = 0.0495

ii. R –Square value.

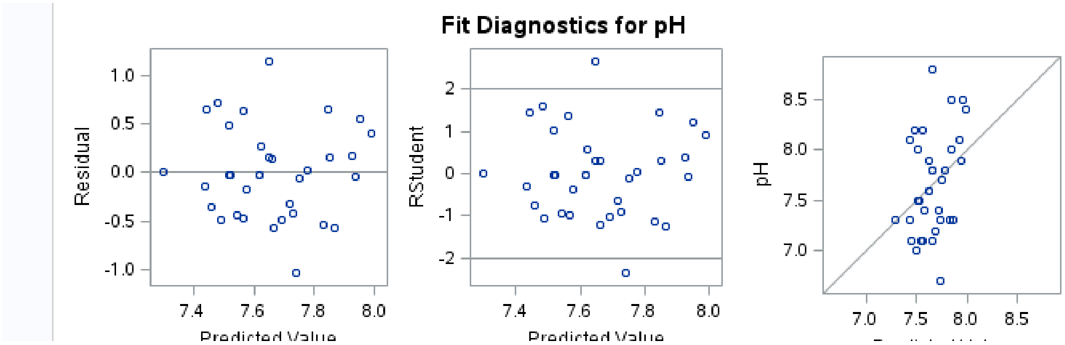
R-Square	0.1153

Approximately 11.53% of the variability in pH can be explained by or attributed to variability in Bicar.

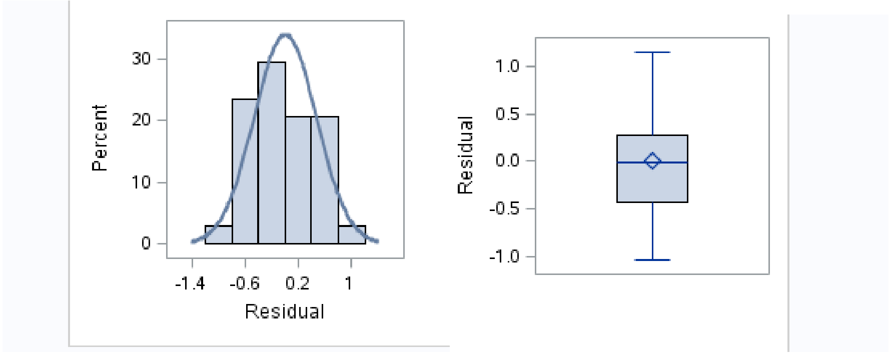
iii. The p-value for parameter (β_0 and β_1) estimates.

β_0 <.0001
$\beta_1 =$ 0.0495

e. Do an analysis of the residuals. What do the “predicted value vs residual values” plot, the normal probability plot, and the boxplot tell you about the reliability and usefulness of the linear regression model?



The predicted value vs residual values plots show that the linear regression model is quite reliable and useful.



The normal probability plot shows that the linear regression is quite useful because the distribution looks well distributed.

The box plot shows that the linear regression is pretty useful because there is no significant outliers and the mean, median, and a box look well distributed.