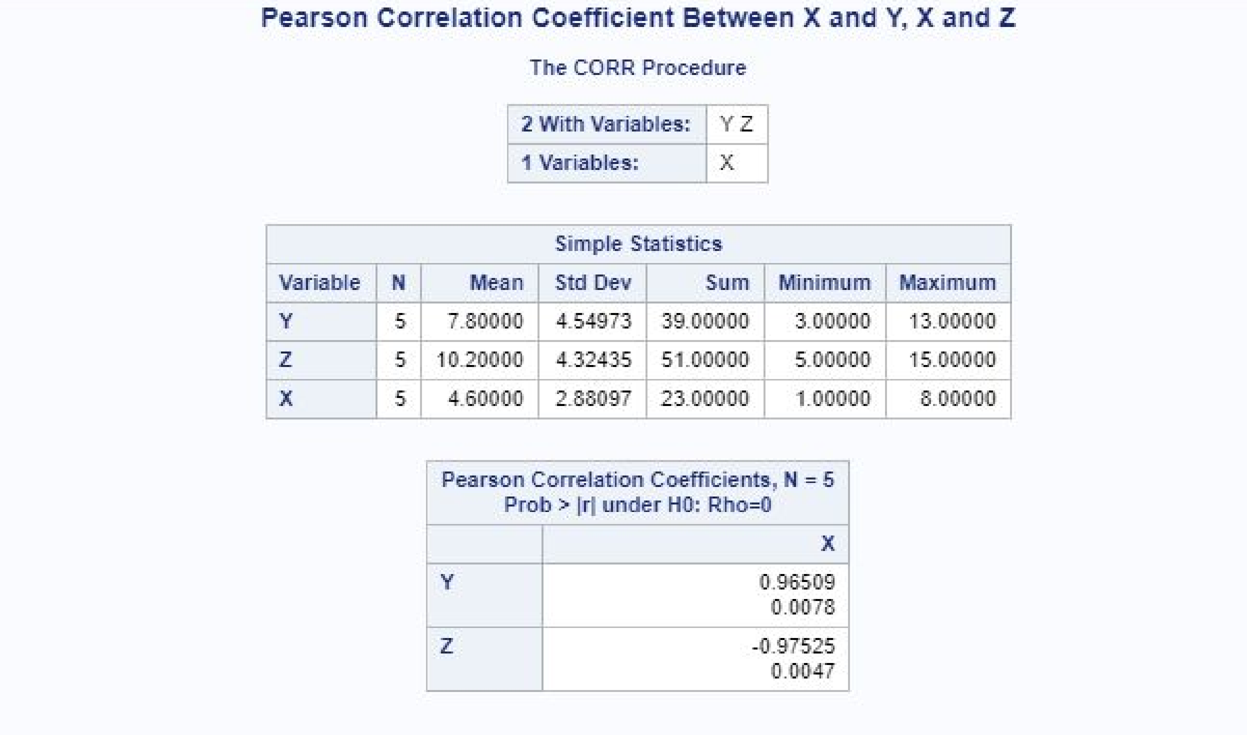
performed by Inna Williams

```
#############################################################
data Problem_5_1_5_5;
input  X Y Z;

cards;
1 3 15
7 13 7
8 12 5
3 4 14
4 7 10
;
/*   a   */
proc corr data=Problem_5_1_5_5;
title "Pearson Correlation Coefficient Between X and Y, X and
Z";
var X;
with Y Z;
Run;
proc sgplot data=Problem_5_1_5_5;
scatter x=Z y=X;
run;
/*   b  */
proc corr data=Problem_5_1_5_5;
title "Pearson Correlation Coefficient Between variables X, Y
Z";
var X Y Z;
run;
```



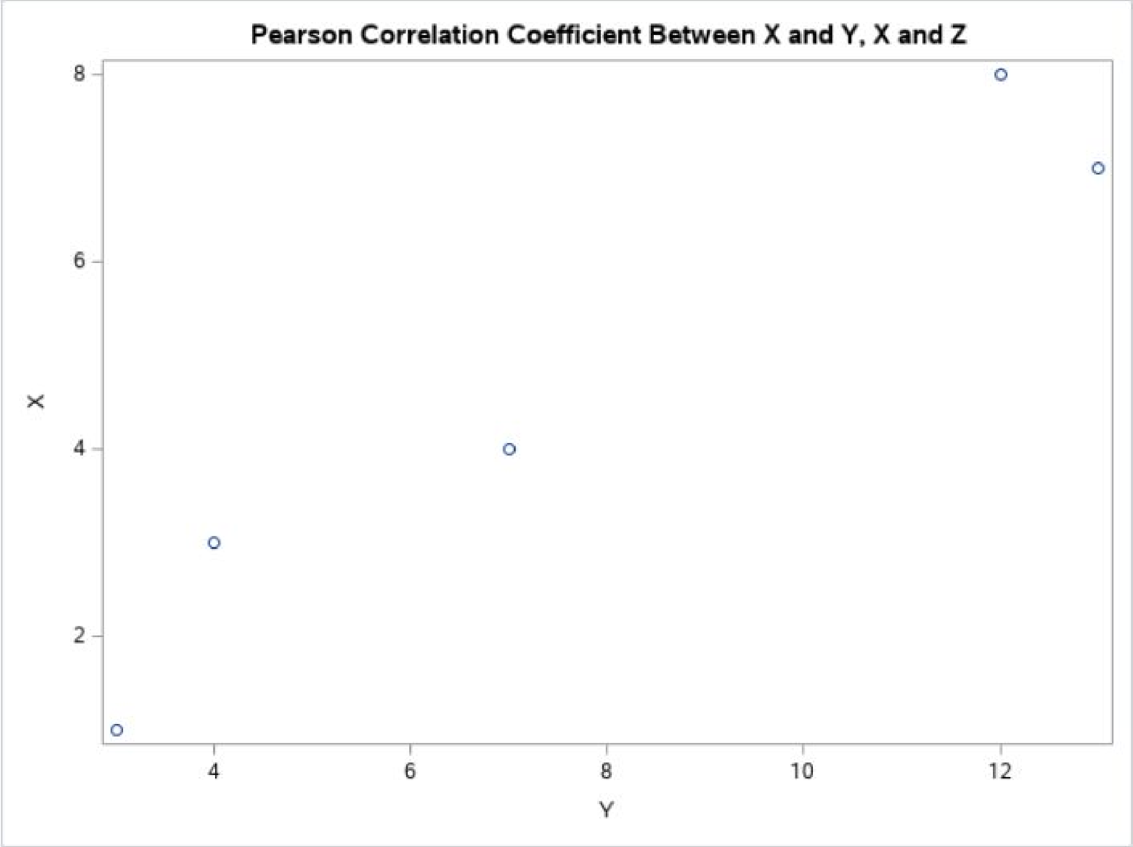
Pearson Correlation Coefficient Between X and Y, X and Z

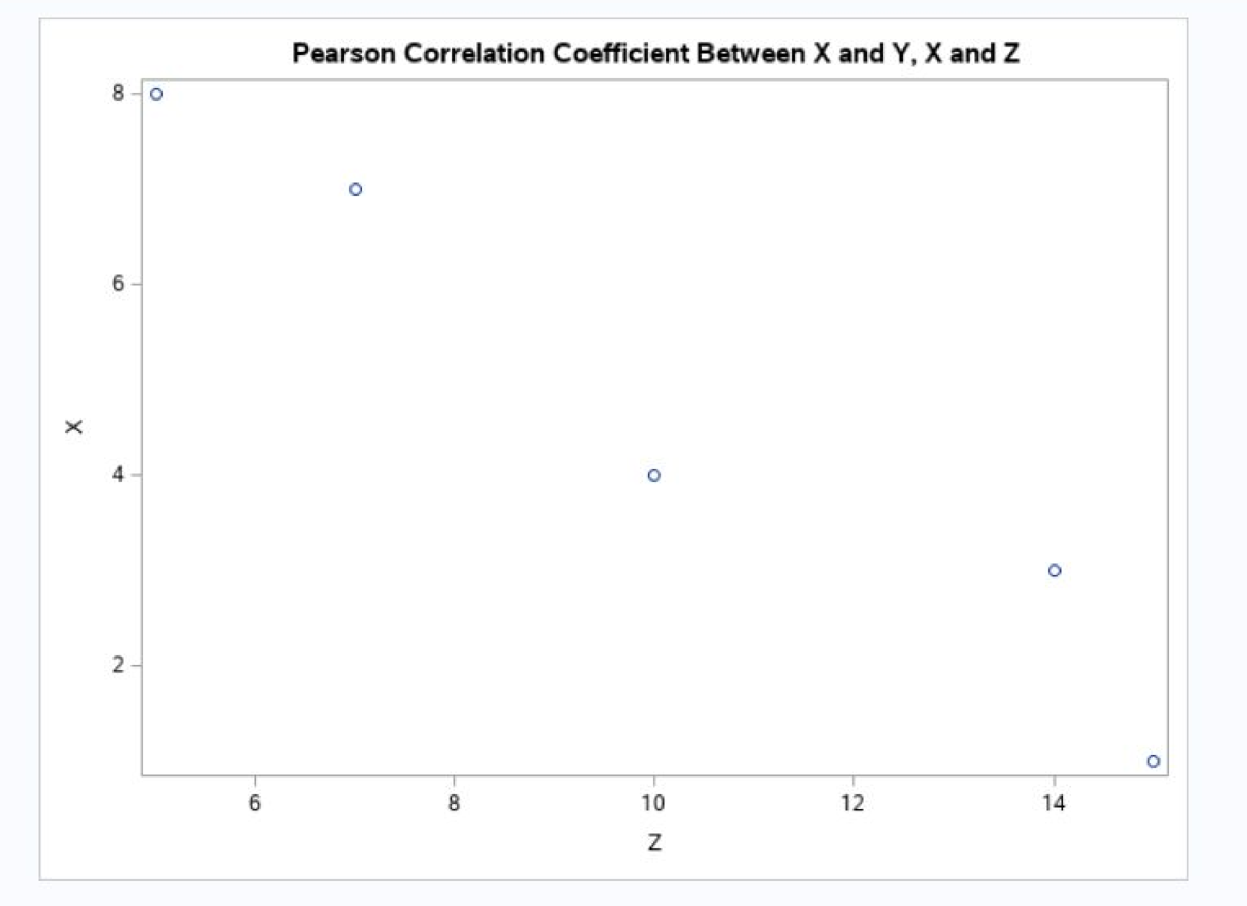
The CORR Procedure

2 With Variables:	Y Z
1 Variables:	X

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y	5	7.80000	4.54973	39.00000	3.00000	13.00000
Z	5	10.20000	4.32435	51.00000	5.00000	15.00000
X	5	4.60000	2.88097	23.00000	1.00000	8.00000

Pearson Correlation Coefficients, N = 5 Prob > |r| under H0: Rho=0	
	X
Y	0.96509 0.0078
Z	-0.97525 0.0047

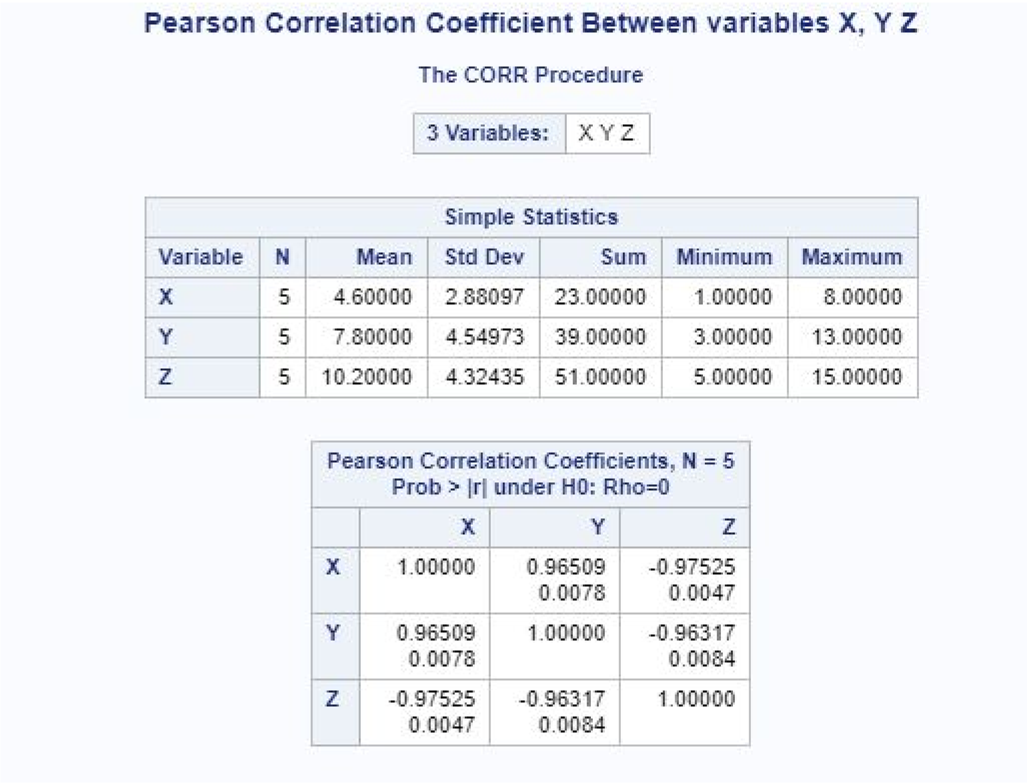




Interpretation Of X Vs Y and X vs Z

X vs Y R=0.96509 p=0.0078
This means that X and Y have Very strong
linear positive correlation. If Y increasing then
X also increasing.
X vs Z R=-0.97525 p=0.0047
This means that X and Y have Very strong
linear negative correlation. If Y increasing then
X decreasing.

Ho -> correlation =zero
Ha -> correlation not = zero
p-value < 0.0078, p=0.0047 (less then 0.05 level).This means that
we reject Ho We have an evidence that correlations
between X and Y, and X and Z is not zero.
Graph shows also that X and Y and X and Z have linear relationship.



Pearson Correlation Coefficient Between variables X, Y Z

The CORR Procedure

3 Variables:	X Y Z

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X	5	4.60000	2.88097	23.00000	1.00000	8.00000
Y	5	7.80000	4.54973	39.00000	3.00000	13.00000
Z	5	10.20000	4.32435	51.00000	5.00000	15.00000

Pearson Correlation Coefficients, N = 5
Prob > |r| under H0: Rho=0

	X	Y	Z
X	1.00000	0.96509 0.0078	-0.97525 0.0047
Y	0.96509 0.0078	1.00000	-0.96317 0.0084
Z	-0.97525 0.0047	-0.96317 0.0084	1.00000

Interpretation for X, Y, Z COrrelations

**X vs Y R=0.96509 p=0.0078
This means that X and Y have Very strong
linear positive correlation. If Y increasing then
X is also increasing.**

**X Vs Z R=-0.97525 p=0.0047
This means that X and Y have Very strong
linear negative correlation. If Z increasing then
X is decreasing.**

**Y vs X R=0.96509 p=0.0078
This means that Y and X have Very strong
linear positive correlation. If X increasing then
Y is also increasing.**

**Y Vs Z R-0.96317 p=0.0084
This means that Y and Z have Very strong
linear negative correlation. If Z increasing then
Y is decreasing.**

**Z vs X R=0.97525 p=0.0047
This means that Z and X have Very strong
linear positive correlation. If X increasing then**

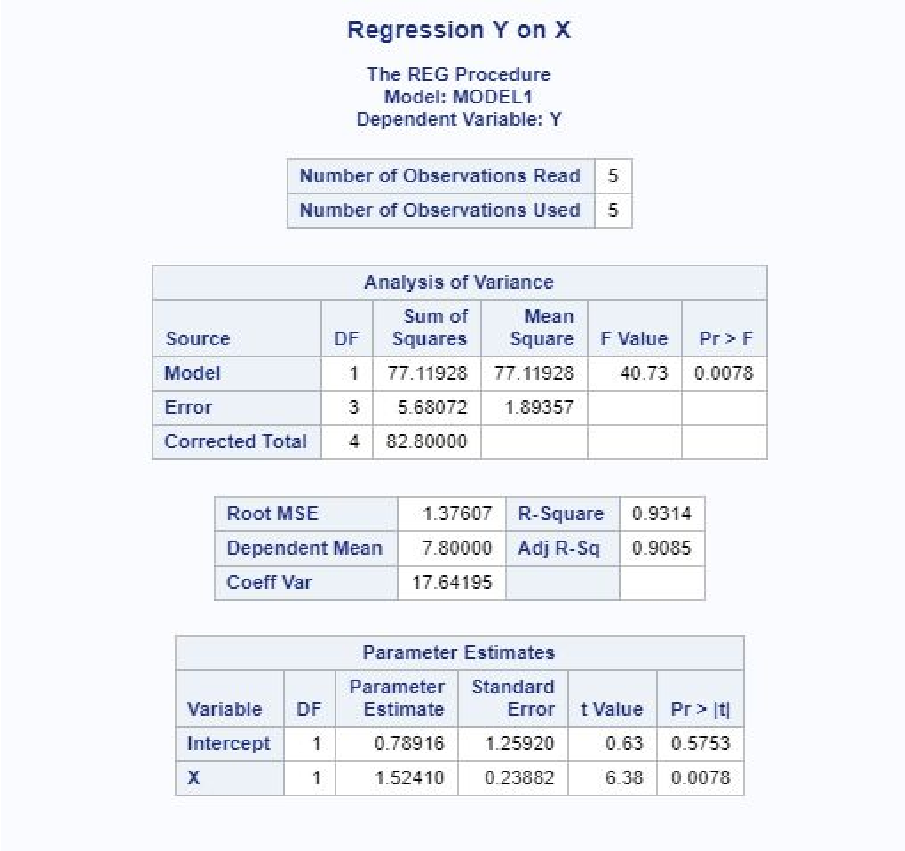
Z is also increasing.

Z Vs Y R=0.96317 p=0.0084
This means that Z and Y have Very strong
linear negative correlation. If Y increasing then
Z is decreasing.

for each case above
Ho -> correlation =zero
Ha -> correlation not = zero
p-value < ... (less than 0.05 level).This means that
we reject Ho We have an evidence that correlations between X,Y Z
are not zero.

###
5_5 a
###

```
proc reg data=Problem_5_1_5_5;
title "Regression Y on X";
model  Y=X;
Run;
```



Regression Y on X

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	5
Number of Observations Used	5

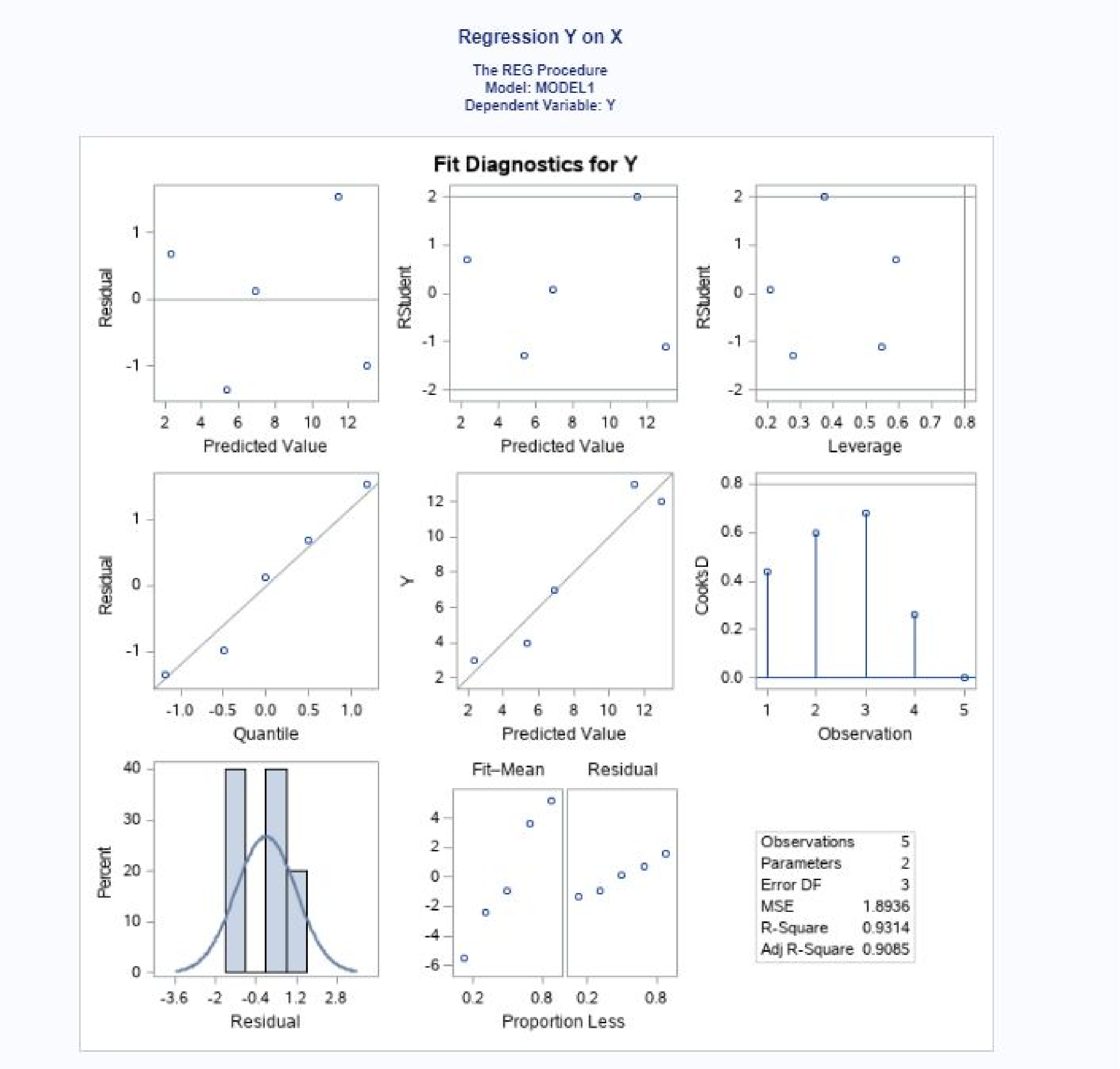
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	77.11928	77.11928	40.73	0.0078
Error	3	5.68072	1.89357		
Corrected Total	4	82.80000			

Root MSE	1.37607	R-Square	0.9314
Dependent Mean	7.80000	Adj R-Sq	0.9085
Coeff Var	17.64195		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.78916	1.25920	0.63	0.5753
X	1	1.52410	0.23882	6.38	0.0078

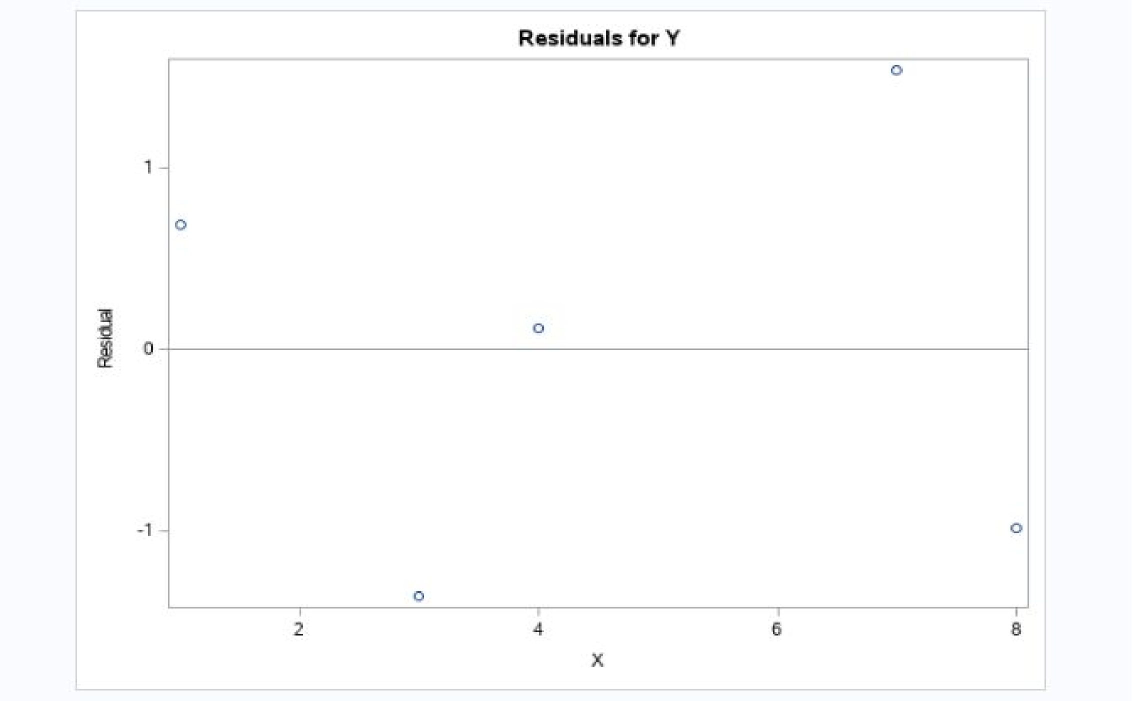


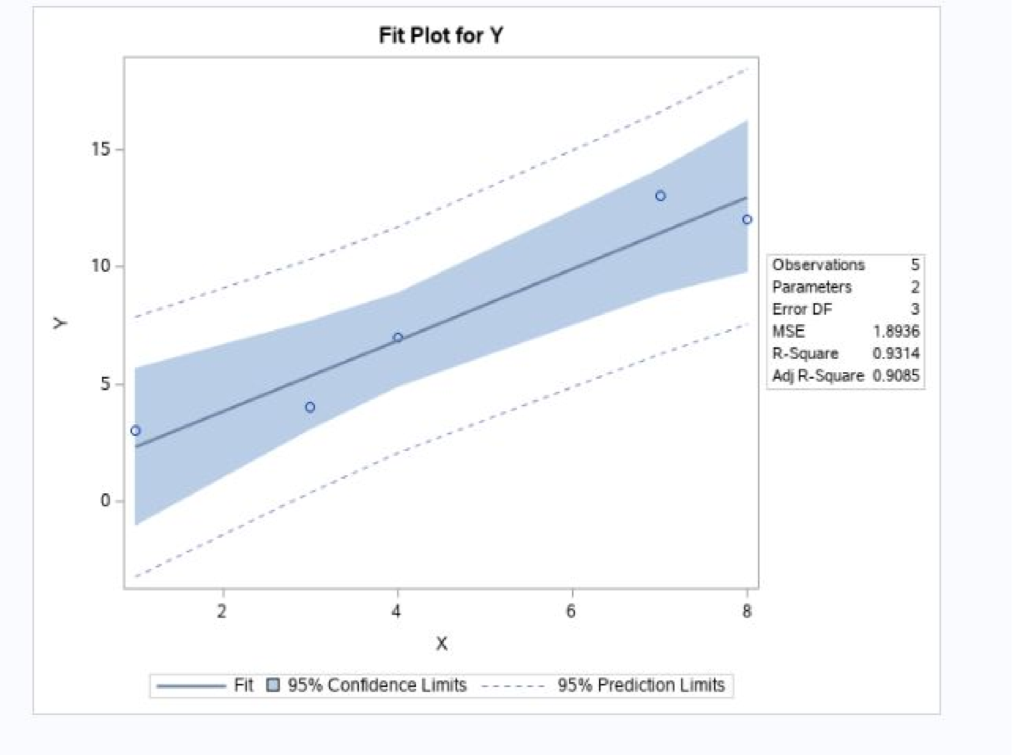
Regression Y on X

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Fit Diagnostics for Y

Observations	5
Parameters	2
Error DF	3
MSE	1.8936
R-Square	0.9314
Adj R-Square	0.9085





Observations	5
Parameters	2
Error DF	3
MSE	1.8936
R-Square	0.9314
Adj R-Square	0.9085

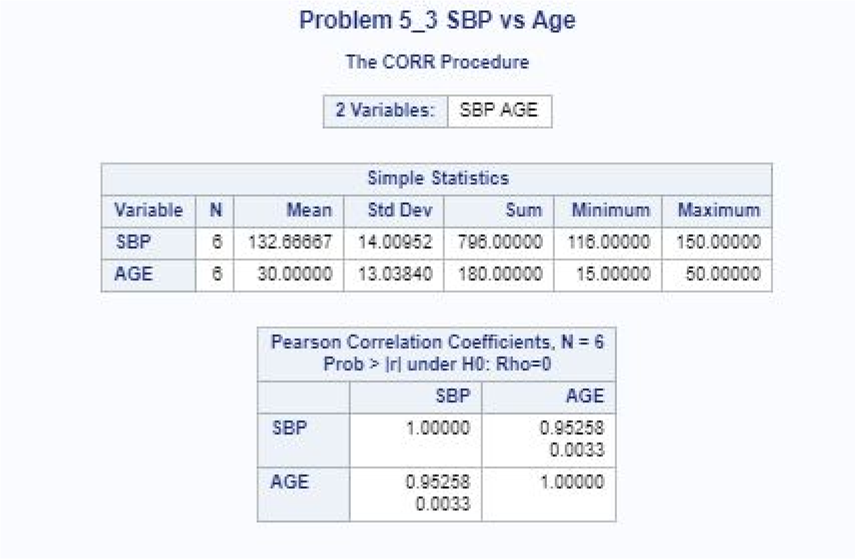
```
##########################################################
   5_5  b)
##########################################################
Intercept = 0.78916 prob|t| = 0.5753
Slope     = 1.52410 prob|t| = 0.0078

##########################################################
   5_5 c)
##########################################################
Ho -> slope is zero
Ha -> slope is not zero
p-value < 0.0078 (less than 0.05 level).This means that
we reject Ho We have an evidence that slope not equal zero

Ho -> intercept is zero
Ha -> Intercept is not zero
p-value = 0.5753 (> than 0.05 level).This means that
we fail to reject Ho We do not have enough  evidence
to conclude that the slope is not  equal zero.


#############################################################
    5_3
#############################################################
data Problem_5_3;
input AGE SBP;
cards;
15 116
20 120
25 130
30 132
40 150
50 148
;

proc corr data=Problem_5_3;
title "Problem 5_3 SBP vs Age";
var SBP AGE;
Run;
```



Problem 5_3 SBP vs Age

The CORR Procedure

2 Variables:	SBP AGE

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SBP	6	132.66667	14.00952	796.00000	116.00000	150.00000
AGE	6	30.00000	13.03840	180.00000	15.00000	50.00000

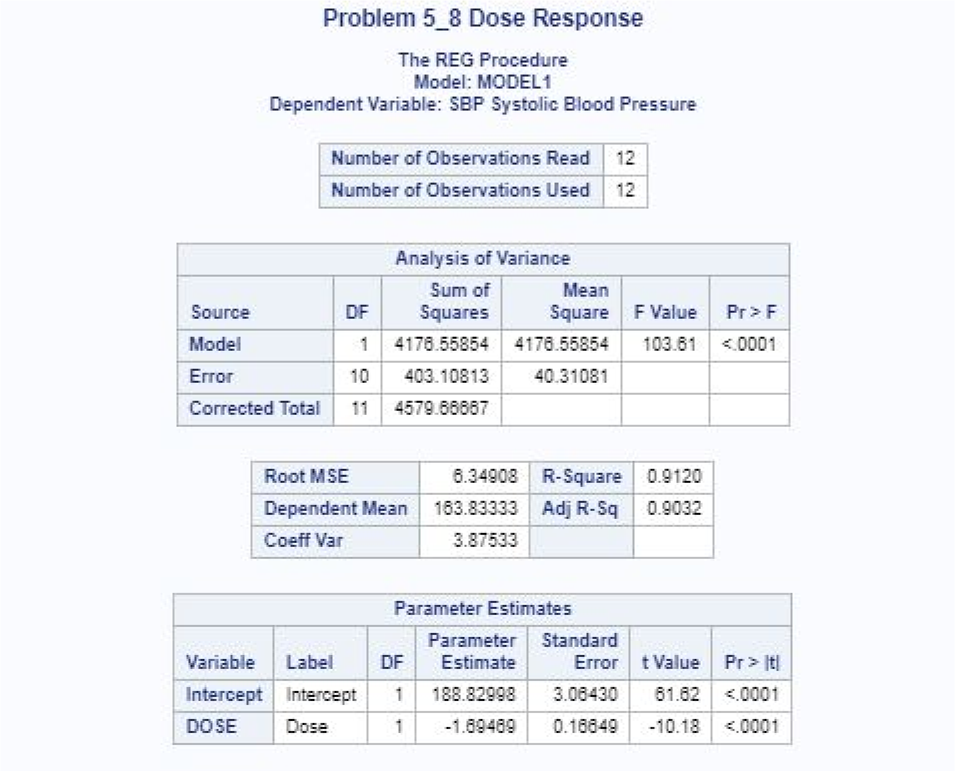
Pearson Correlation Coefficients, N = 6 Prob > |r| under H0: Rho=0		
	SBP	AGE
SBP	1.00000	0.95258 0.0033
AGE	0.95258 0.0033	1.00000

SBP vs Age R=0.95258 p=0.0033
This means that SBP and AGE have Very strong linear positive correlation. If AGE increasing then SBP is also increasing.

Age vs SBP R=0.95258 p=0.0033
This means that AGE and SBP have Very strong linear negative correlation. If SBP increasing then AGE is also increasing.

```
##############################################################
     5_8
##############################################################
data DOSE_RESPONCE;
input DOSE SBP DBP;
Label
    DOSE='Dose'
    SBP='Systolic Blood Pressure'
    DBP='Diastolic Blood Pressure'
;
cards;
4 180 110
1 190 108
4 178 100
8 170 100
8 180 98
```

```
8 168 88
16 160 80
16 172 86
16 170 86
32 140 80
32 130 72
32 128 70
;
proc reg data=DOSE_RESPONCE;
title 'Problem 5_8 Dose Response';
model SBP DBP =DOSE;
Run;
```



Problem 5_8 Dose Response

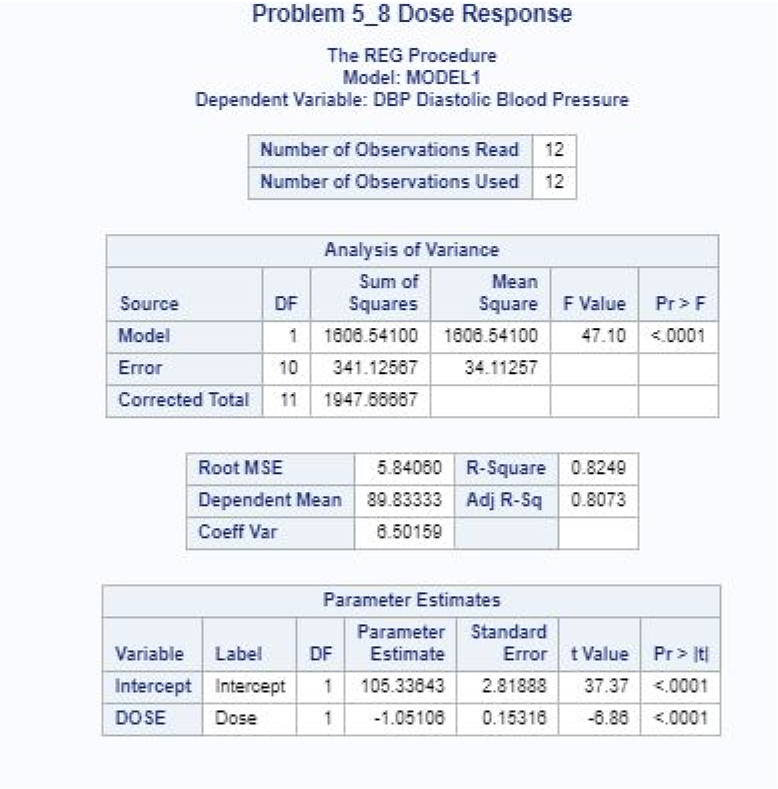
The REG Procedure
Model: MODEL1
Dependent Variable: SBP Systolic Blood Pressure

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4176.55854	4176.55854	103.61	<.0001
Error	10	403.10813	40.31081		
Corrected Total	11	4579.66667			

Root MSE	6.34908	R-Square	0.9120
Dependent Mean	163.83333	Adj R-Sq	0.9032
Coeff Var	3.87533		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	188.82998	3.06430	61.62	<.0001
DOSE	Dose	1	-1.69469	0.16649	-10.18	<.0001



Problem 5_8 Dose Response

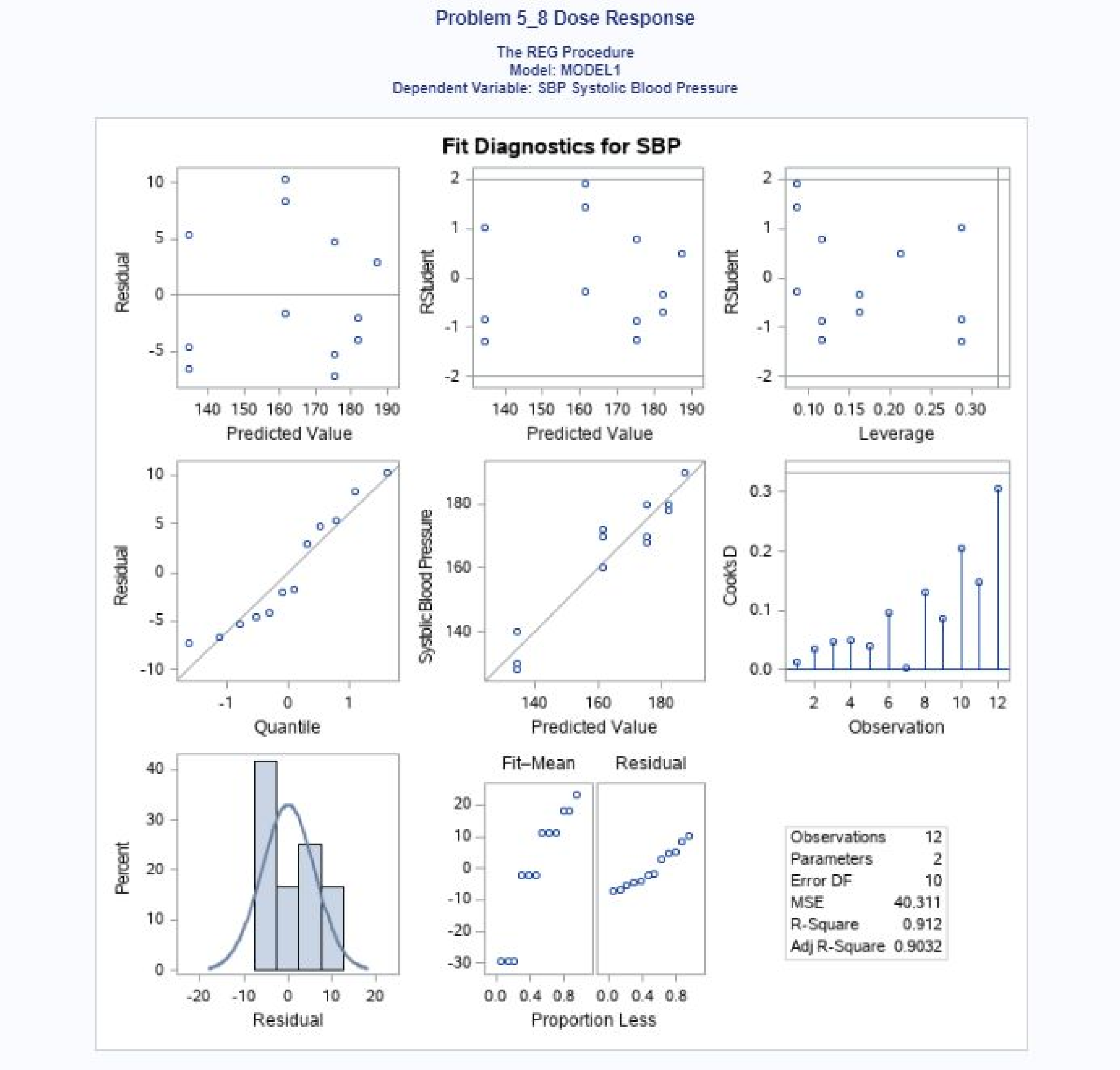
The REG Procedure
Model: MODEL1
Dependent Variable: DBP Diastolic Blood Pressure

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1606.54100	1606.54100	47.10	<.0001
Error	10	341.12567	34.11257		
Corrected Total	11	1947.66667			

Root MSE	5.84060	R-Square	0.8249
Dependent Mean	89.83333	Adj R-Sq	0.8073
Coeff Var	6.50159		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	105.33643	2.81888	37.37	<.0001
DOSE	Dose	1	-1.05106	0.15316	-6.86	<.0001

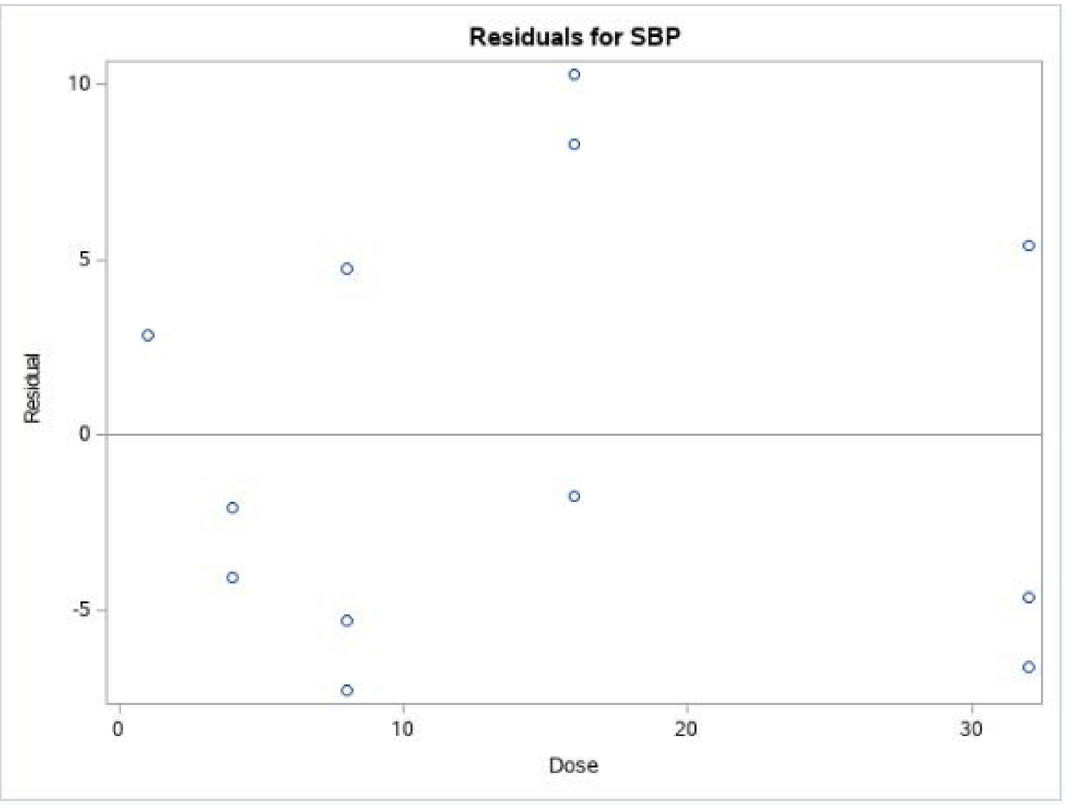


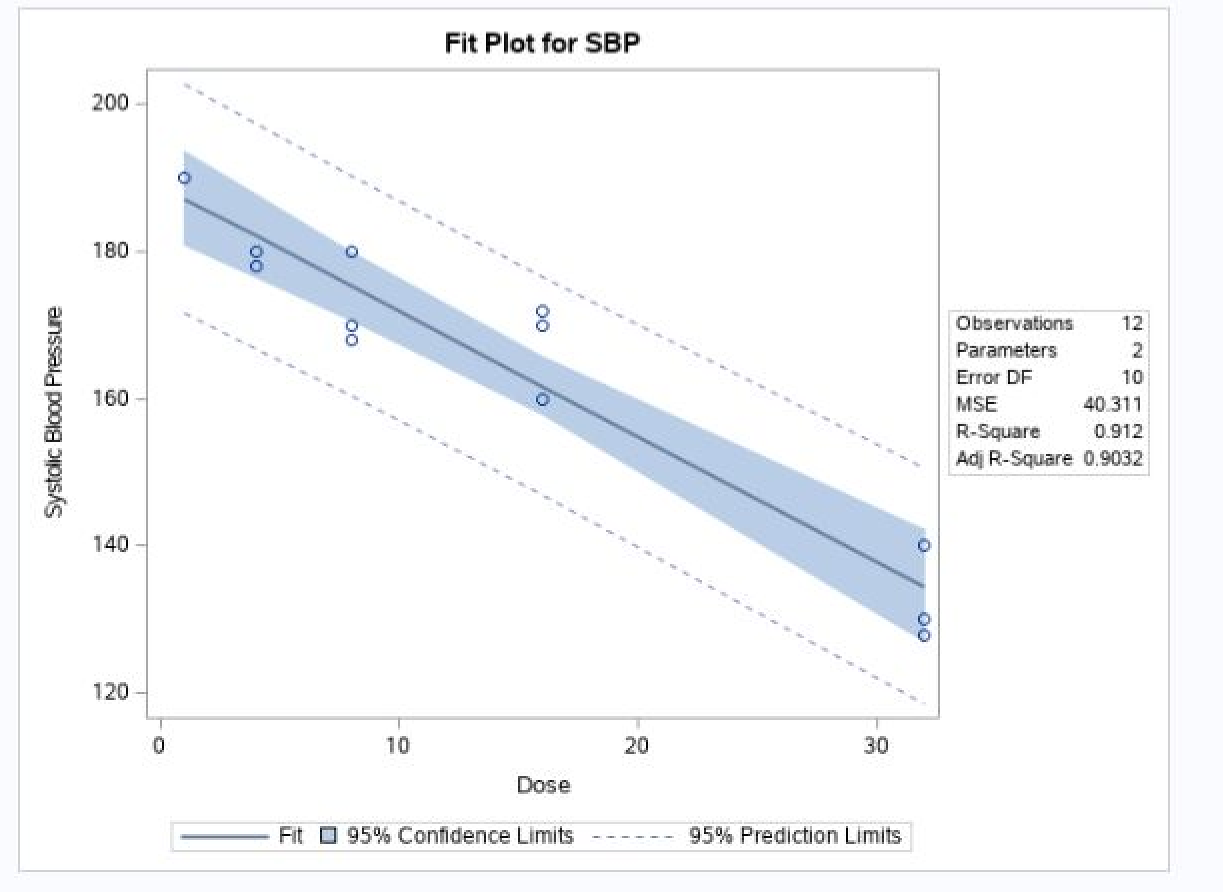
Problem 5_8 Dose Response

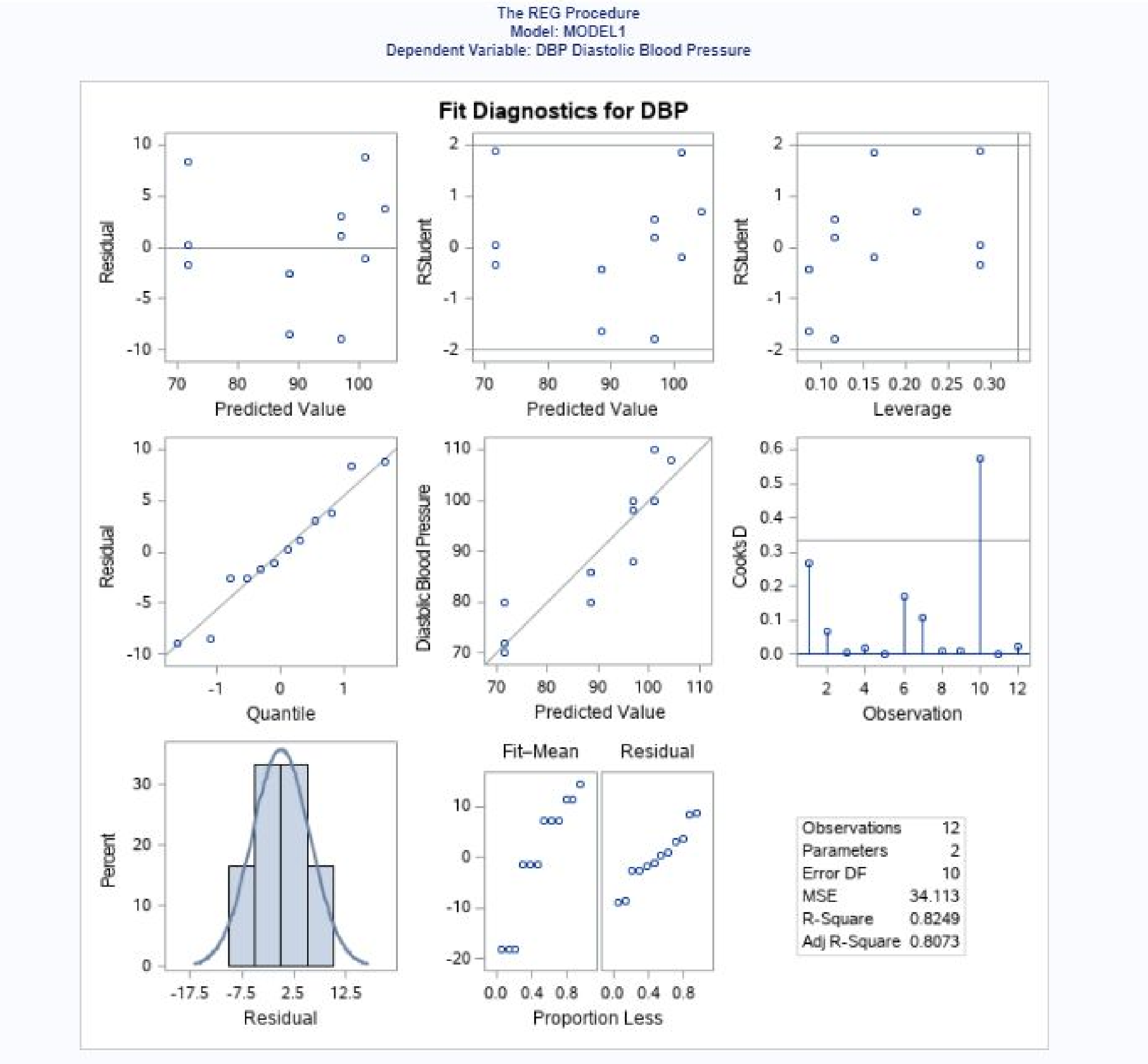
The REG Procedure
Model: MODEL1
Dependent Variable: SBP Systolic Blood Pressure

Fit Diagnostics for SBP

Observations	12
Parameters	2
Error DF	10
MSE	40.311
R-Square	0.912
Adj R-Square	0.9032


Residuals for SBP
Residual
Dose
10
5
0
-5
0
10
20
30

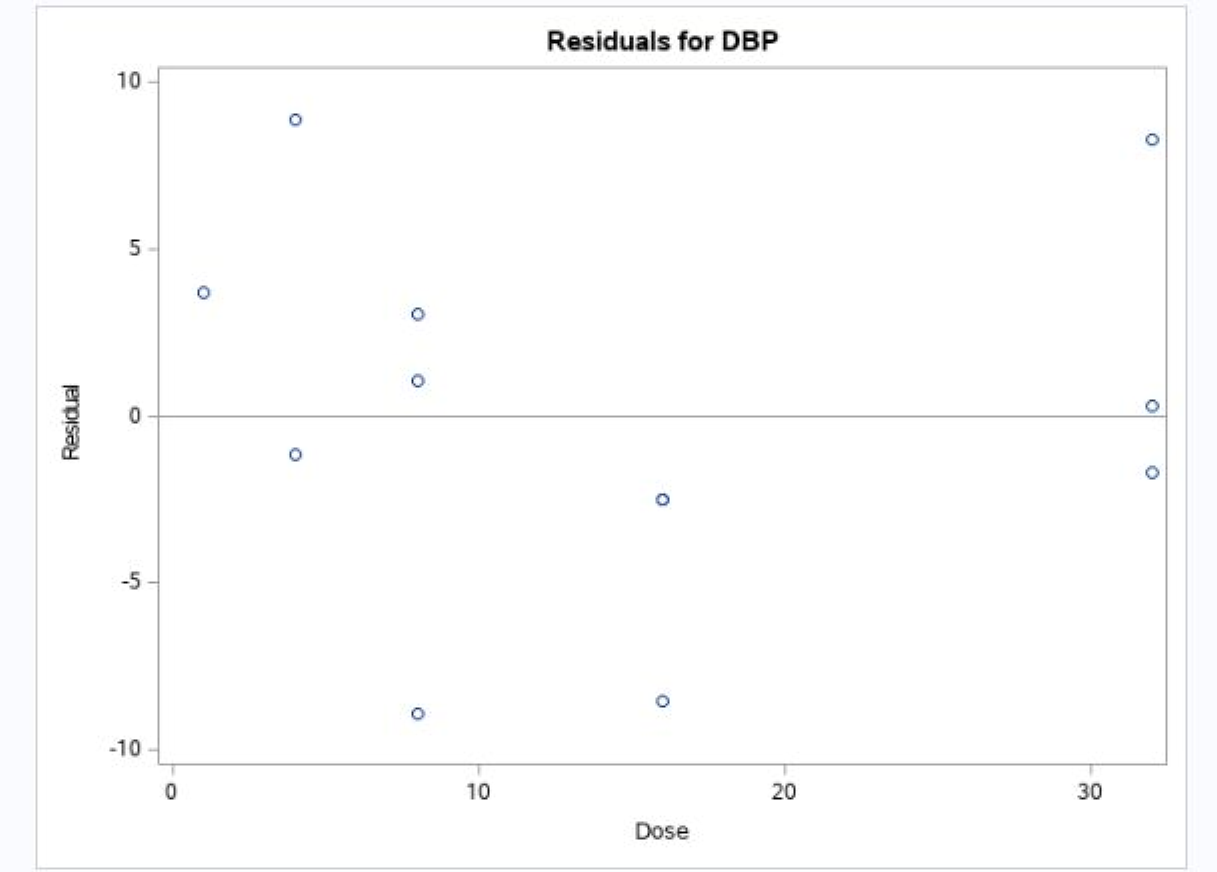

Fit Plot for SBP
200
180
160
140
120
Systolic Blood Pressure
0
10
20
30
Dose
Observations 12
Parameters 2
Error DF 10
MSE 40.311
R-Square 0.912
Adj R-Square 0.9032
Fit
95% Confidence Limits
95% Prediction Limits

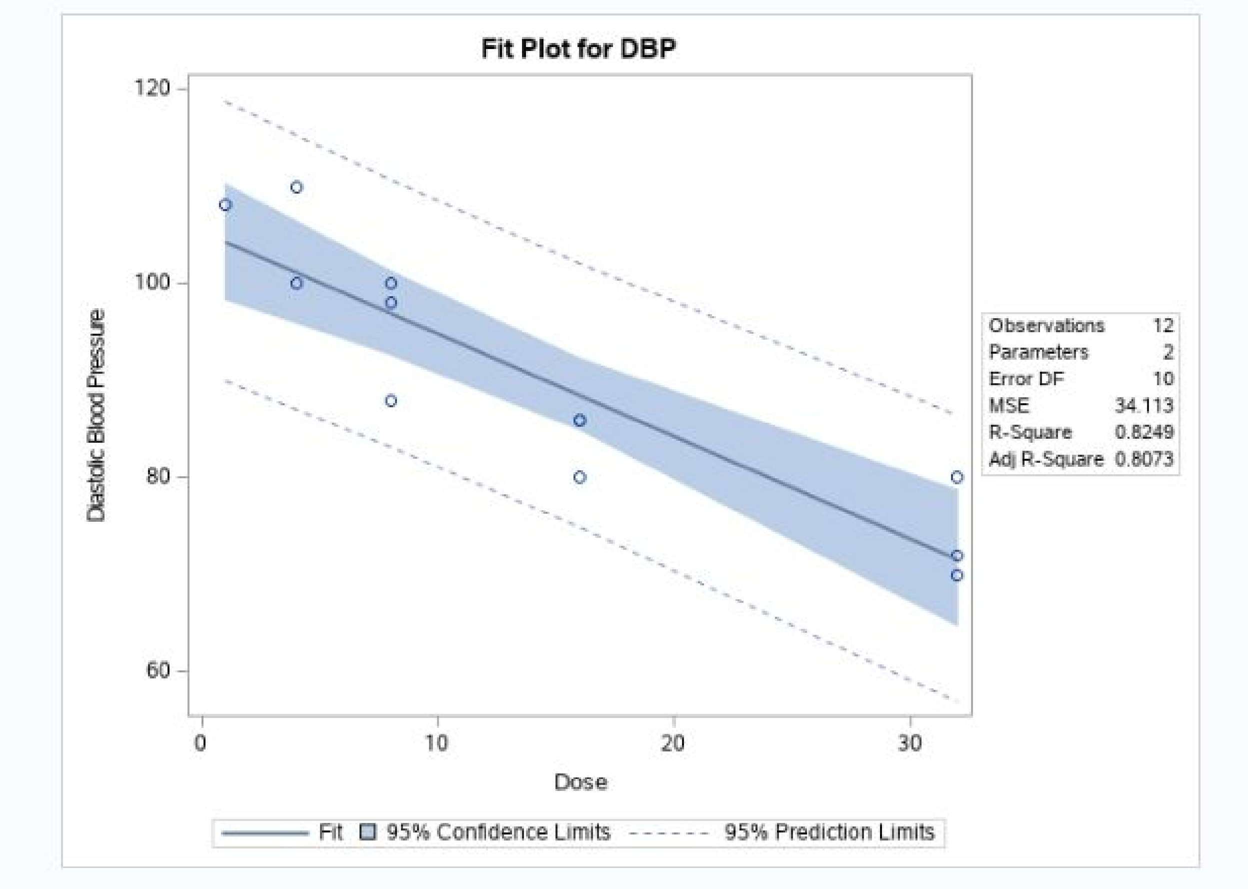


The REG Procedure
Model: MODEL1
Dependent Variable: DBP Diastolic Blood Pressure

Fit Diagnostics for DBP

```
Observations      12
Parameters         2
Error DF          10
MSE           34.113
R-Square      0.8249
Adj R-Square  0.8073
```


Residuals for DBP
10
5
0
-5
-10
Residual
0
10
20
30
Dose



SBP VS DOSE
Intercept = 188.82998 Slope = -1.69469

DBP VS DOSE
Intercept = 105.33643 Slope = -1.05106

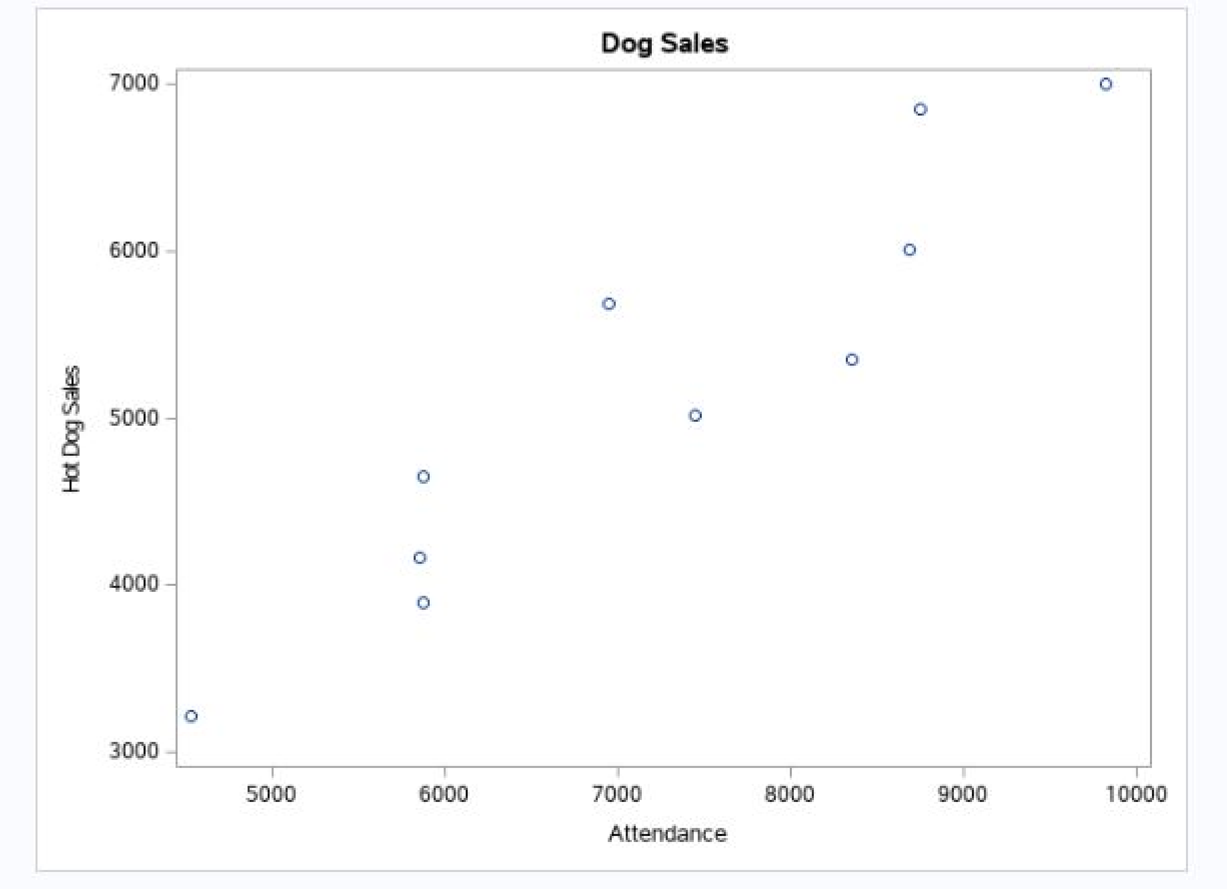
##

The daily attendance and the number of hot dog sales at a local ballpark are studied over a period of games. Given the following data, answer the following questions.

##

(a) Plot the data using proc sgplot.
##

```
data DOG_SALES;
input ATT SALES;
Label
   ATT="Attendance"
   SALES="Hot Dog Sales";
   ;
cards;
8747 6845
5857 4168
8360 5348
6945 5687
8688 6007
4534 3216
7450 5018
5874 4652
9821 7001
5873 3896
;
proc sgplot data=DOG_SALES;
title 'Dog Sales';
scatter x=ATT y=SALES;
Run;
```

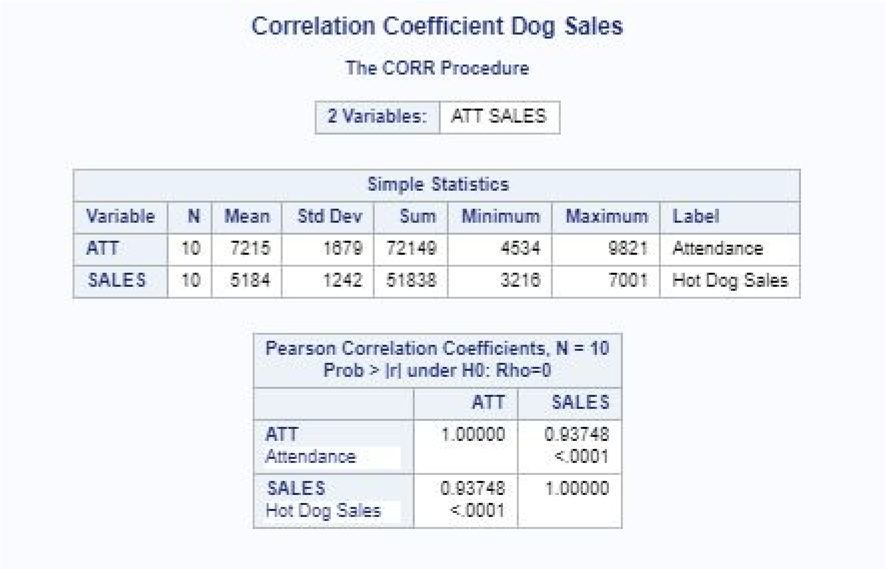


##

(b) Find the correlation coefficient and test its significance.

##

```
proc corr data=DOG_SALES;
title 'Correlation Coefficient Dog Sales';
Var ATT SALES;
Run;
```



Correlation Coefficient Dog Sales

The CORR Procedure

2 Variables:	ATT SALES

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ATT	10	7215	1679	72149	4534	9821	Attendance
SALES	10	5184	1242	51838	3216	7001	Hot Dog Sales

Pearson Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	ATT	SALES
ATT Attendance	1.00000	0.93748 <.0001
SALES Hot Dog Sales	0.93748 <.0001	1.00000

Attendance vs Sale Coefficient of Correlation R=0.93748 with significance p<.0001
This means that Attendance vs Sale have Very strong linear positive correlation. If Sales is increasing then Attendance is also increasing.

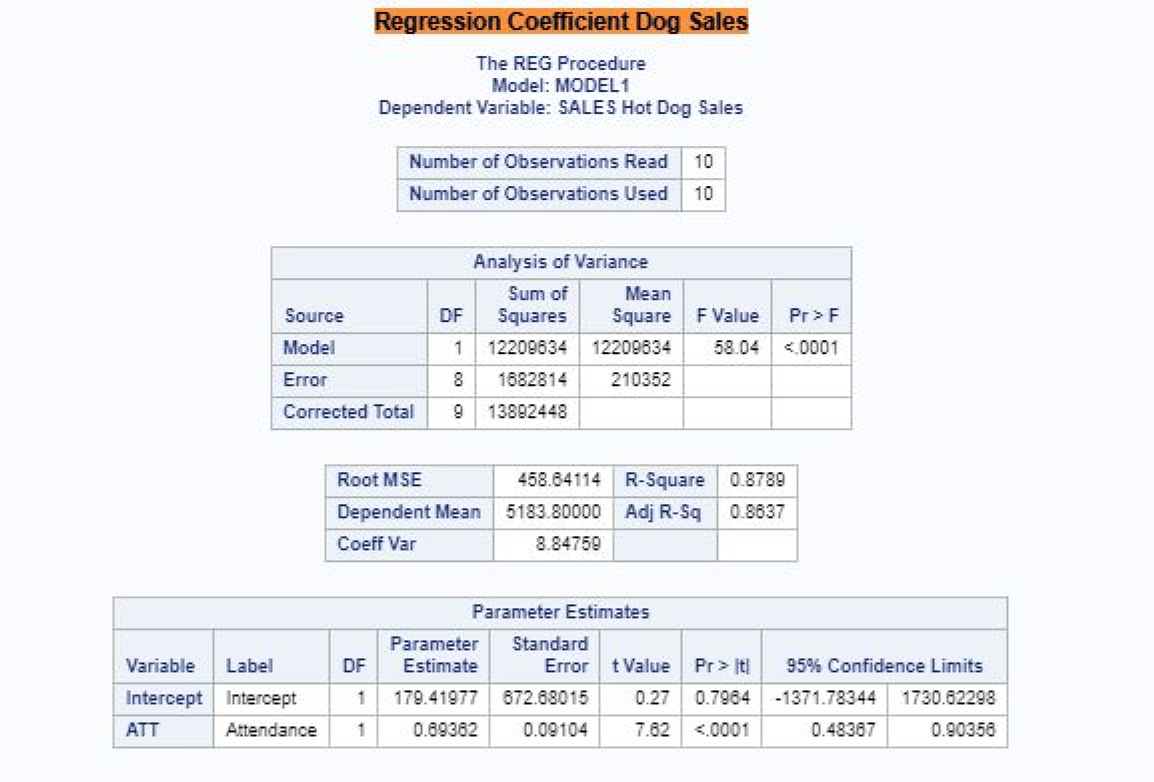
Sale vs Attendance Coefficient of Correlation R=0.93748 with significance p<.0001
This means that Sale vs Attendance have Very strong linear positive correlation. If Attendance is increasing then Sales is also increasing.

testing for significance

Ho -> correlation = zero
Ha -> correlation not = zero
p-value < 0.0001 (less than 0.05 level).This means that we reject Ho We have an evidence that correlation is not Zero.

##
(c)
Find the regression line to predict hot dogs sales based on attendance.
##

```
proc reg data=DOG_SALES;
title 'Regression  Coefficient Dog Sales';
model SALES=ATT / clb;
run;
```



Regression Coefficient Dog Sales

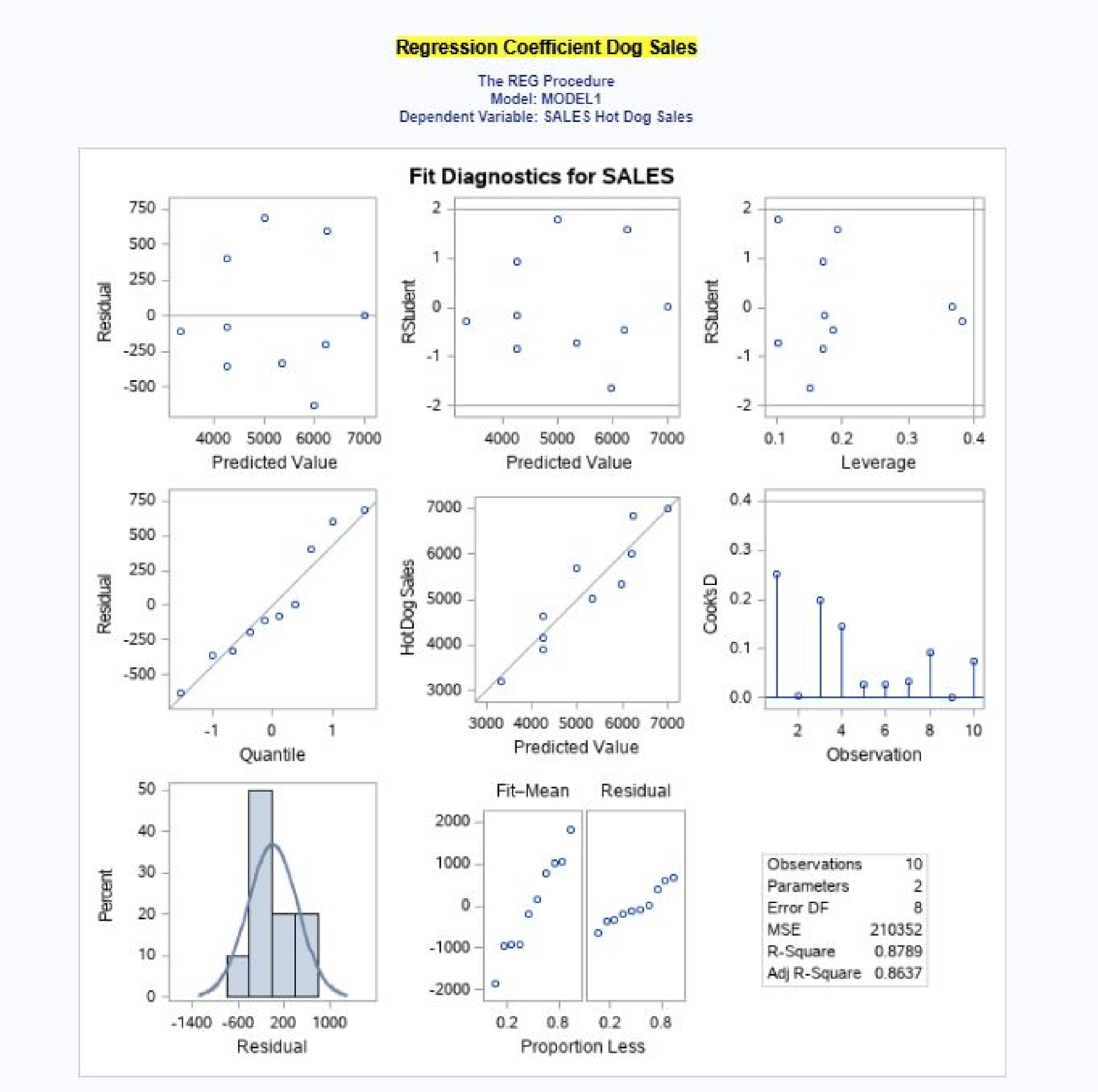
The REG Procedure
Model: MODEL1
Dependent Variable: SALES Hot Dog Sales

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12209634	12209634	58.04	<.0001
Error	8	1682814	210352		
Corrected Total	9	13892448			

Root MSE	458.64114	R-Square	0.8789
Dependent Mean	5183.80000	Adj R-Sq	0.8637
Coeff Var	8.84759		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	179.41977	672.68015	0.27	0.7964	-1371.78344	1730.62298
ATT	Attendance	1	0.69362	0.09104	7.62	<.0001	0.48367	0.90356

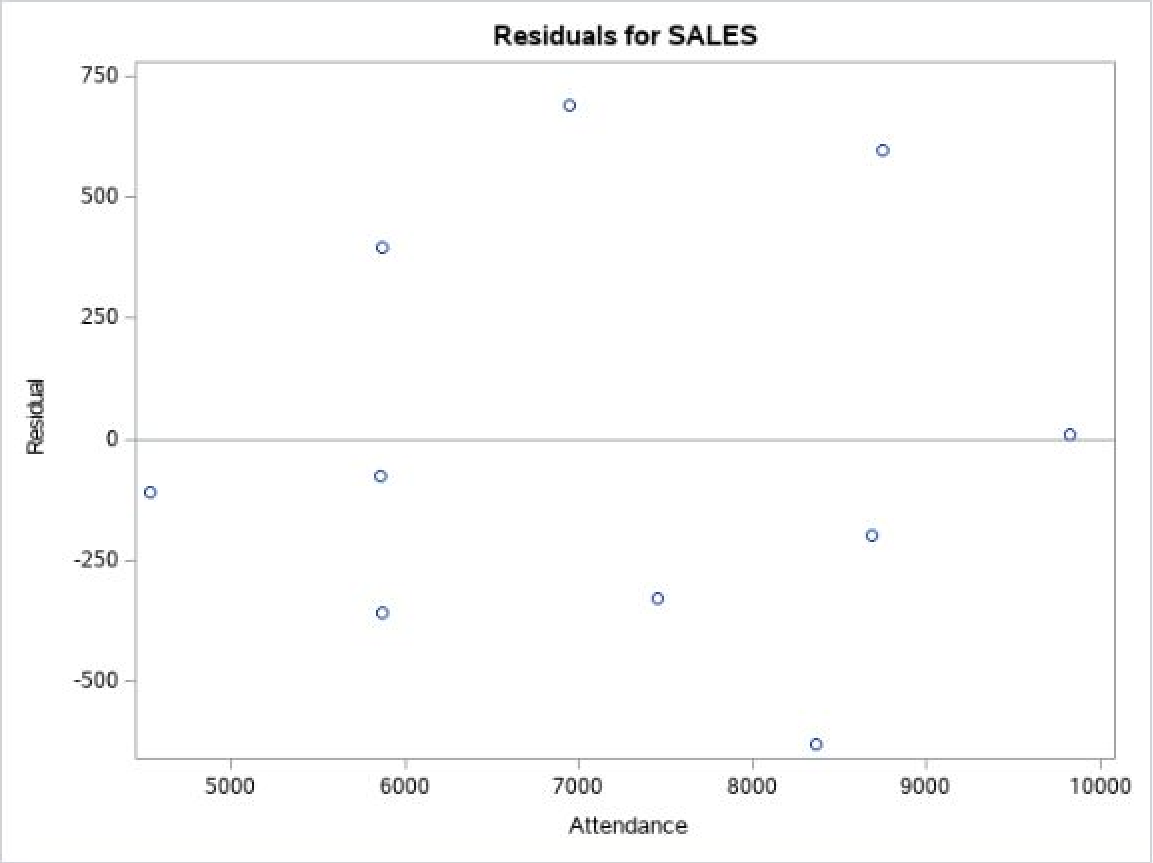


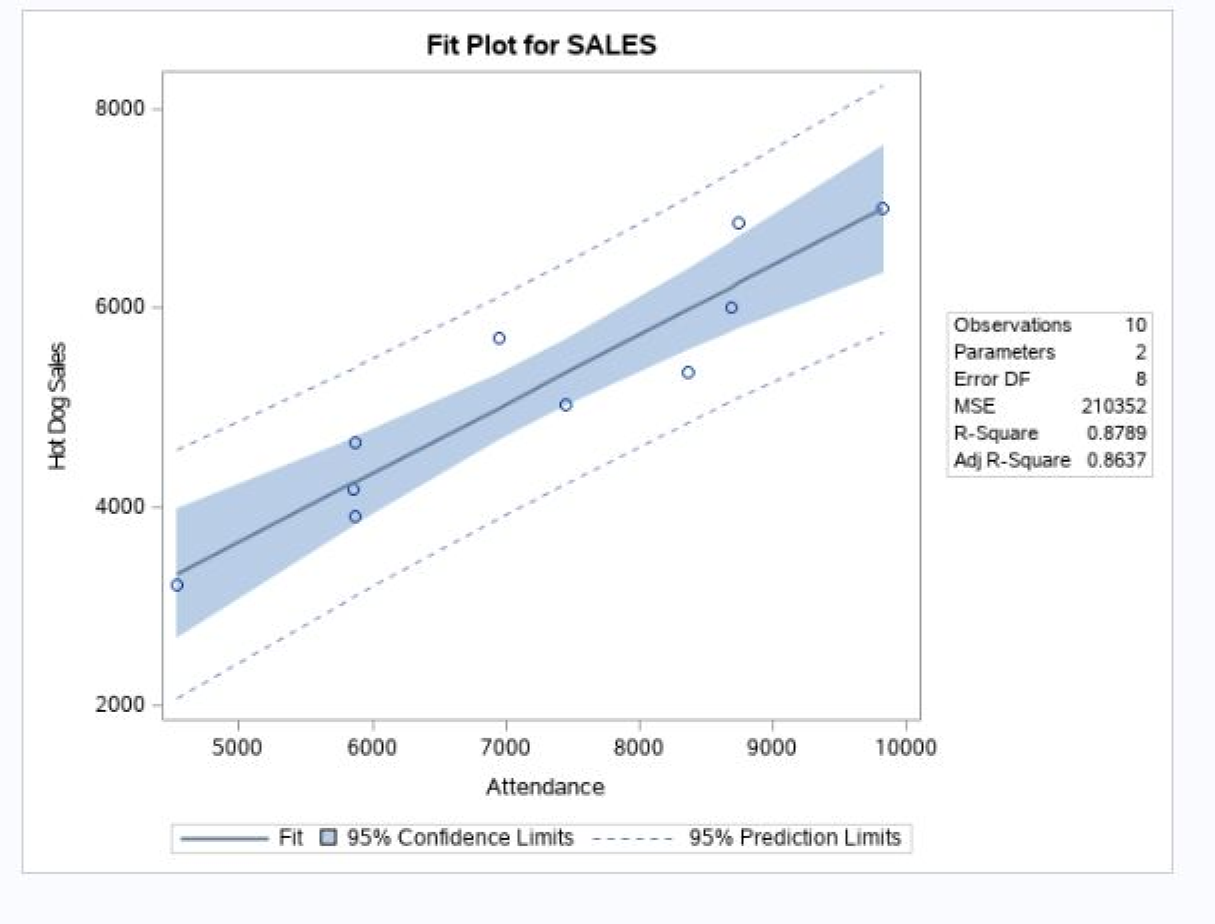
Regression Coefficient Dog Sales

The REG Procedure
Model: MODEL1
Dependent Variable: SALES Hot Dog Sales

Fit Diagnostics for SALES

Observations	10
Parameters	2
Error DF	8
MSE	210352
R-Square	0.8789
Adj R-Square	0.8637


Residuals for SALES
Residual
750
500
250
0
-250
-500
5000
6000
7000
8000
9000
10000
Attendance



Observations	10
Parameters	2
Error DF	8
MSE	210352
R-Square	0.8789
Adj R-Square	0.8637

Regression Line:

hot_dog_sales = 179.41977 + 0.69362   * Attendance

###

(d) What is the estimate of standard deviation (root MSE)?

###

Root MSE = 458.64114
It is the square root of the Residual(or error)

###

(e) Report and interpret the 95% confidence interval for the slope coefficient.

##

```
95% CI for Slope = [0.48367, 0.90356]
0.48367   0.90356
```

with 95% confidence for each attander there is an associated increase in dog sales between [0.48367, 0.90356]

##

2. Although the income tax system is structured so that people with higher incomes should pay a higher percentage of their incomes in taxes, there are many loopholes and tax shelters available for individuals with higher incomes. A sample of 2017 tax returns gave the data listed in the table.

###

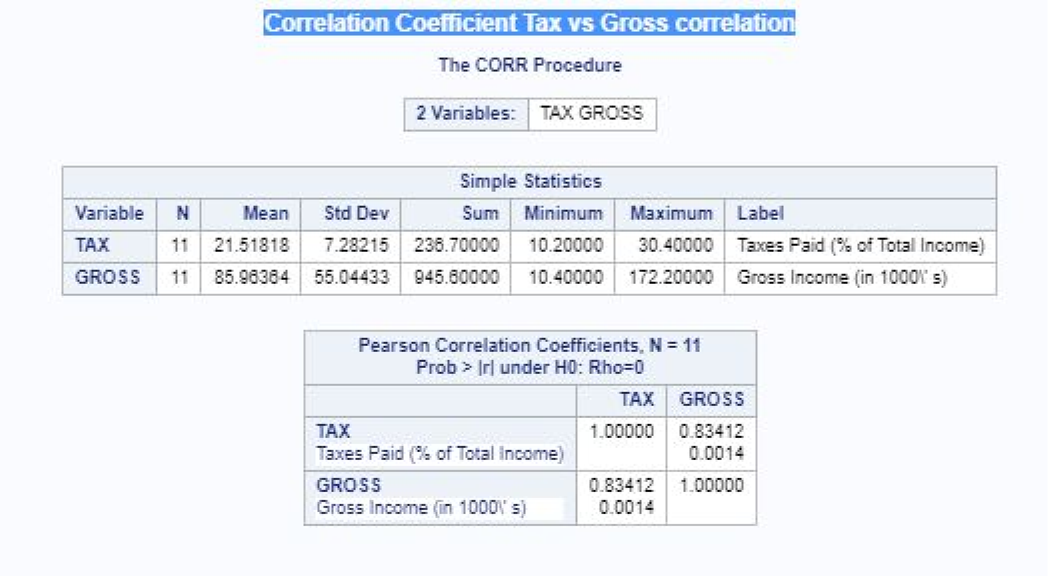
```
data TAX_DATA;
input INDIVIDUAL GROSS TAX;
Label
    GROSS='Gross Income (in 1000\'' s)'
    TAX='Taxes Paid (% of Total Income)'
    ;
cards;
1 44.2 16.0
2 92.0 20.1
3 17.0 11.1
4 54.0 24.3
5 10.4 10.2
6 172.2 30.4
7 63.9 27.3
8 125.9 27.9
9 83.6 16.2
10 167.7 29.8
11 114.7 23.4
;
```

###

(a) Compute the sample correlation coefficient and interpret the results.

##

```
proc corr data=TAX_DATA;
title 'Correlation Coefficient Tax vs Gross correlation';
var TAX GROSS ;
Run;
```



Correlation Coefficient Tax vs Gross correlation

The CORR Procedure

2 Variables:	TAX GROSS

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
TAX	11	21.51818	7.28215	236.70000	10.20000	30.40000	Taxes Paid (% of Total Income)
GROSS	11	85.96364	55.04433	945.60000	10.40000	172.20000	Gross Income (in 1000\' s)

Pearson Correlation Coefficients, N = 11 Prob > |r| under H0: Rho=0		
	TAX	GROSS
TAX Taxes Paid (% of Total Income)	1.00000	0.83412 0.0014
GROSS Gross Income (in 1000\' s)	0.83412 0.0014	1.00000

TAX vs Gross : Correlation Coefficient = 0.83412 significance = 0.0014
This means that Correlation Coefficient > 0.8 and therefore
it is strong positive correlation. If Gross is increasing then Tax is increasing.
Ho -> correlation between Tax and Gross Income equal zero
Ha -> correlation between Tax and Gross Income !=zero
because significance =0.0014 < 0.05 we reject Ho. We have a evidence that correlation between Tax and Gross Income is not zero.

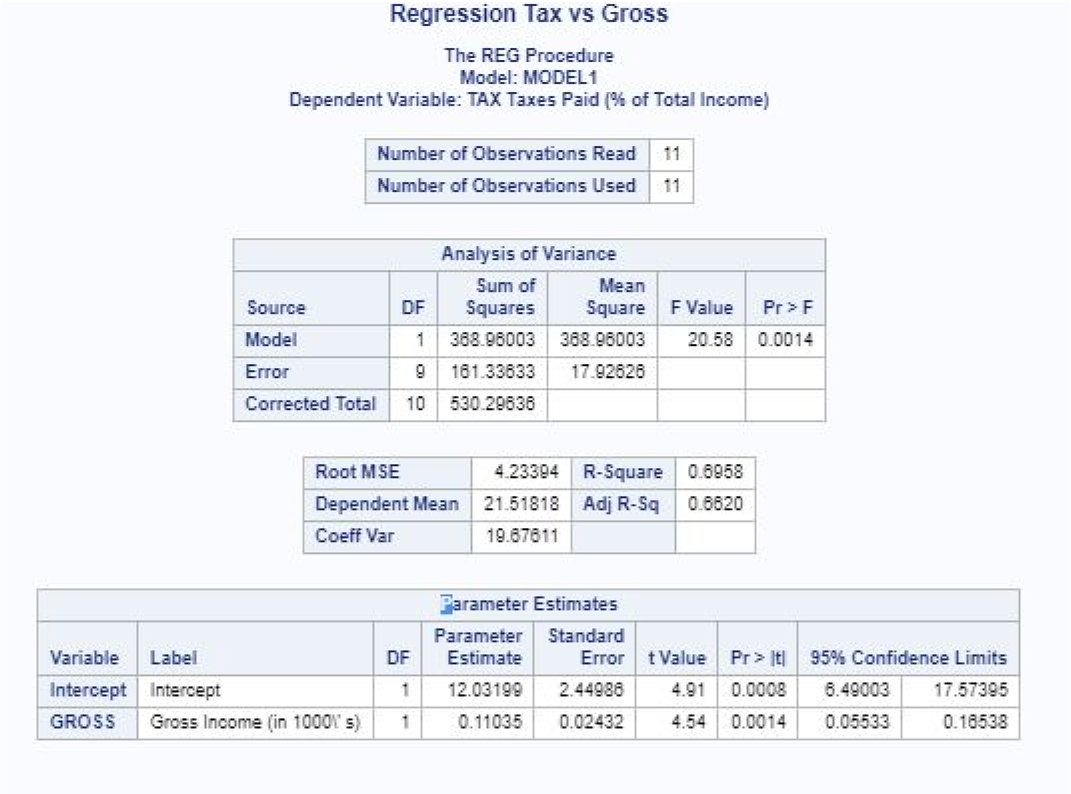
###

(b) Compute r-squared and interpret results.

###

```
proc reg data=TAX_DATA;
```

```
title 'Regression Tax vs Gross';
model TAX=GROSS /clb;
run;
```



Regression Tax vs Gross

The REG Procedure
Model: MODEL1
Dependent Variable: TAX Taxes Paid (% of Total Income)

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	368.96003	368.96003	20.58	0.0014
Error	9	161.33633	17.92626		
Corrected Total	10	530.29636			

Root MSE	4.23394	R-Square	0.6958
Dependent Mean	21.51818	Adj R-Sq	0.6620
Coeff Var	19.67611		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	12.03199	2.44986	4.91	0.0008	6.49003	17.57395
GROSS	Gross Income (in 1000\' s)	1	0.11035	0.02432	4.54	0.0014	0.05533	0.16538

R-squared = Sum Of Squares Of Model/Sum Of Squares Of Total
    = 368.960/530.296 = 0.6958

R-squared = 0.6958

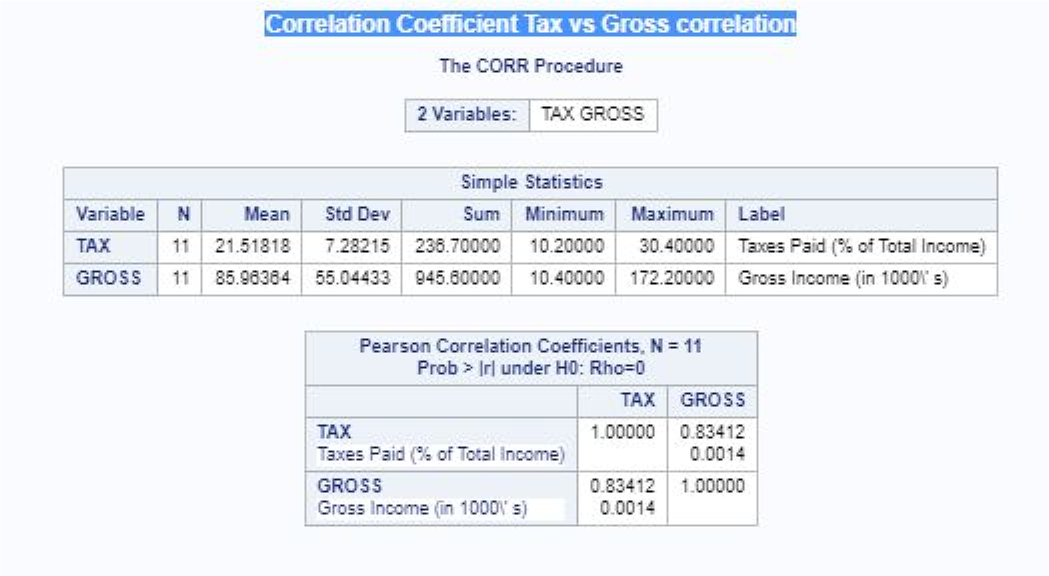
R-square is the proportion of the variance explained by independent variables.

R-square means that out of total variance 69.58% explained by Variance of the model. (in our case this is a Gross) or 69.58% of total tax variance can be predicted by the variable Gross income.

##

(c) Is the correlation coefficient significant at = 0:05?

###



Correlation Coefficient Tax vs Gross correlation

The CORR Procedure

2 Variables:	TAX GROSS

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
TAX	11	21.51818	7.28215	236.70000	10.20000	30.40000	Taxes Paid (% of Total Income)
GROSS	11	85.96364	55.04433	945.60000	10.40000	172.20000	Gross Income (in 1000\' s)

Pearson Correlation Coefficients, N = 11
Prob > |r| under H0: Rho=0

	TAX	GROSS
TAX Taxes Paid (% of Total Income)	1.00000	0.83412 0.0014
GROSS Gross Income (in 1000\' s)	0.83412 0.0014	1.00000

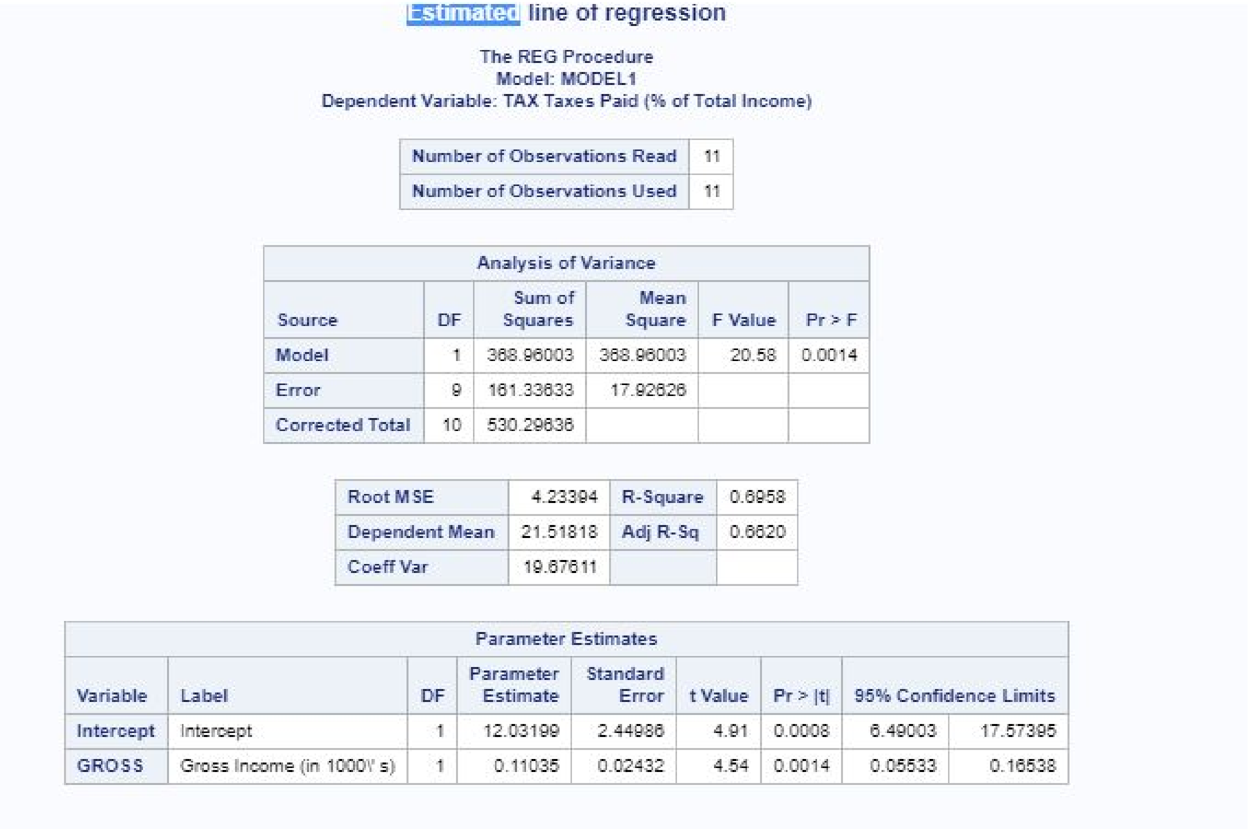
Coefficient Of Correlation = 0.83412 p=0.0014
Ho - >Correlation between Tax paid and Gross Income is zero
Ha -> correlation between Tax paid and Gross income is not zero
p-value =0.0014 < alpha=0.05 -> We reject Ho. We conclude that we have evidence that correlation between tax paid and the gross income is not zero for alpha=0.05.

###

(d) Compute the estimated line of regression.
(e)What is the estimate of standard deviation (root MSE)?

##

```
proc reg data=TAX_DATA;
title 'Regression Tax vs Gross';
model TAX=GROSS /clb;
run;
```



Estimated line of regression

The REG Procedure
Model: MODEL1
Dependent Variable: TAX Taxes Paid (% of Total Income)

Number of Observations Read	11
Number of Observations Used	11

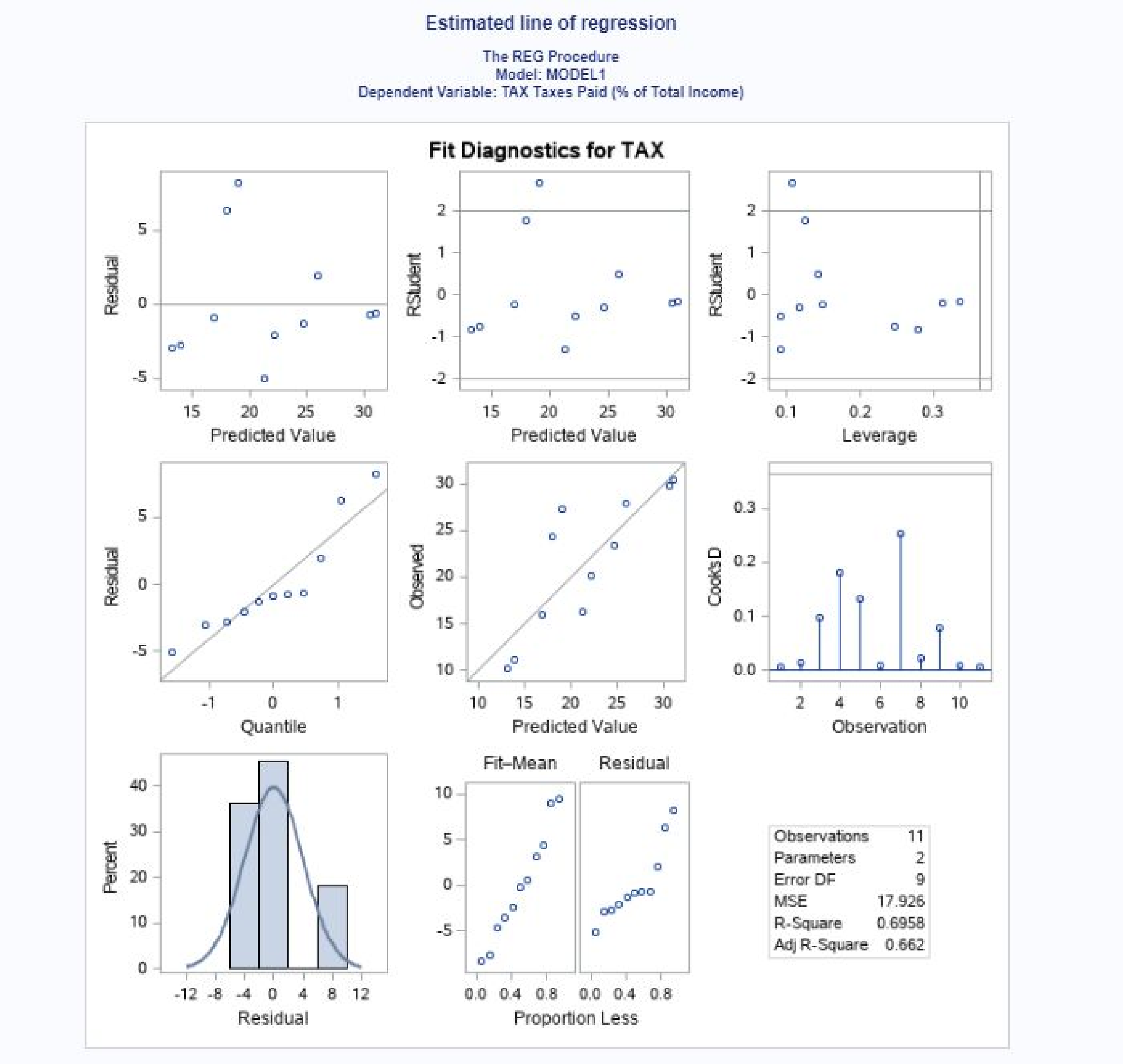
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	368.96003	368.96003	20.58	0.0014
Error	9	161.33633	17.92626		
Corrected Total	10	530.29636			

Root MSE	4.23394	R-Square	0.6958
Dependent Mean	21.51818	Adj R-Sq	0.6620
Coeff Var	19.67611		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	12.03199	2.44986	4.91	0.0008	6.49003	17.57395
GROSS	Gross Income (in 1000\' s)	1	0.11035	0.02432	4.54	0.0014	0.05533	0.16538

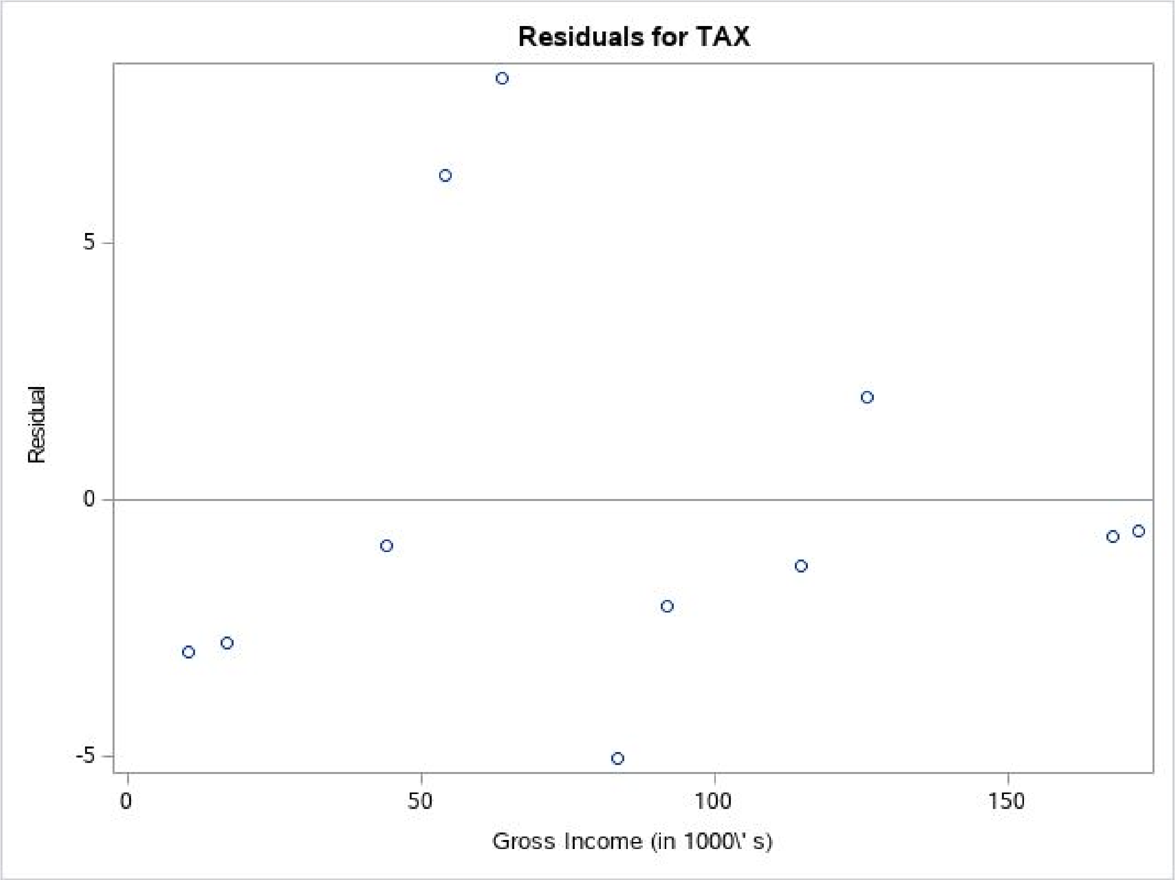


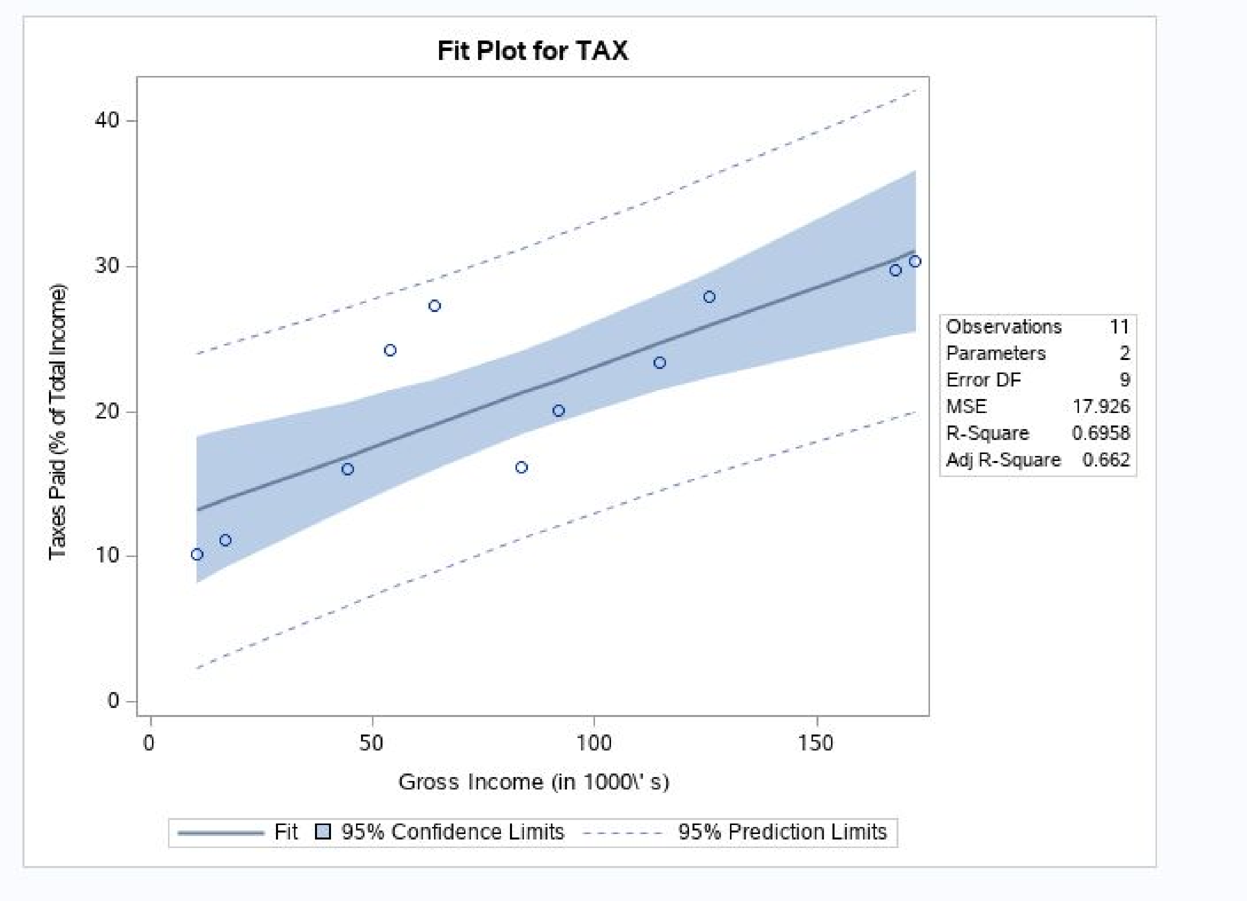
Estimated line of regression

The REG Procedure
Model: MODEL1
Dependent Variable: TAX Taxes Paid (% of Total Income)

Fit Diagnostics for TAX

Observations	11
Parameters	2
Error DF	9
MSE	17.926
R-Square	0.6958
Adj R-Square	0.662


Residuals for TAX
Residual
5
0
-5
0
50
100
150
Gross Income (in 1000\' s)



We calculated Intercept and slope, therefore estimated line of regression is:

TaxPaid = 12.03199 + 0.11035 * GrossIncome

The estimate of standard deviation (square root MSE):
root MSE = 4.23394
Root MSE is a square root of the mean square residual.

##

(f) Predict the mean percentage of income paid in taxes by individuals with a gross income of $80,000. Report and interpret the confidence interval for this estimate.

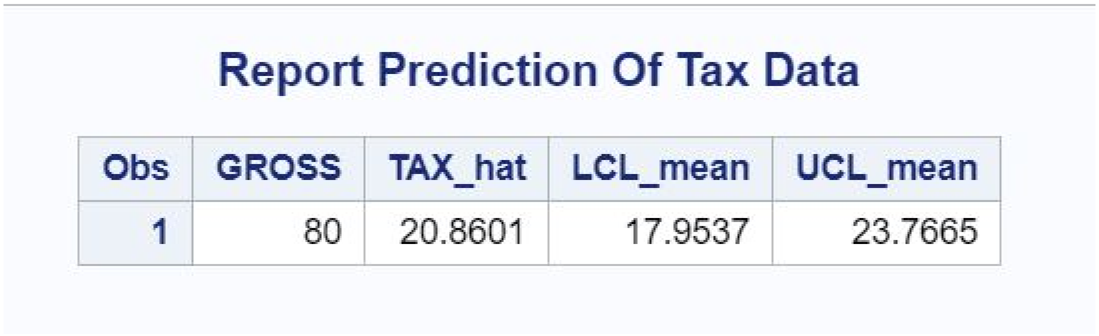
###

TAX($80000) = 12.03199 + 0.11035 * 80 = 20.85999

```
data TAX_DATA;
set TAX_DATA end=last;
output;
if last then do;
TAX=.;
GROSS=80;
output;
end;
run;

proc reg data=TAX_DATA;
model TAX=GROSS / clb;
output out=TAX_DATA_out (where=(TAX=.)) predicted=TAX_hat
LCLM=LCL_mean UCLM=UCL_mean;
run;

proc print data=TAX_DATA_OUT ;
title 'Report Prediction Of Tax Data';
var GROSS TAX_hat LCL_mean UCL_mean;
Run;
```



Report Prediction Of Tax Data

Obs	GROSS	TAX_hat	LCL_mean	UCL_mean
1	80	20.8601	17.9537	23.7665

```
for 80000 of gross income
estimated tax paid = 20.8601% with
95% CI= [17.9537 ,  23.7665]
the mean tax of 80000 income will be in the interval
[17.9537 , 23.7665].
```