CORRELATION ANALYSIS

Example 1:

```
proc contents data=sashelp._all_;
run;
proc print data=sashelp.class;
run;
```

The SAS System

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83.0
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5

Contains 19 observation

```
proc corr data=sashelp.class;
      var height weight;
run;
```

The CORR Procedure

2 Variables:	Height Weight

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Height	19	62.33684	5.12708	1184	51.30000	72.00000
Weight	19	100.02632	22.77393	1901	50.50000	150.00000

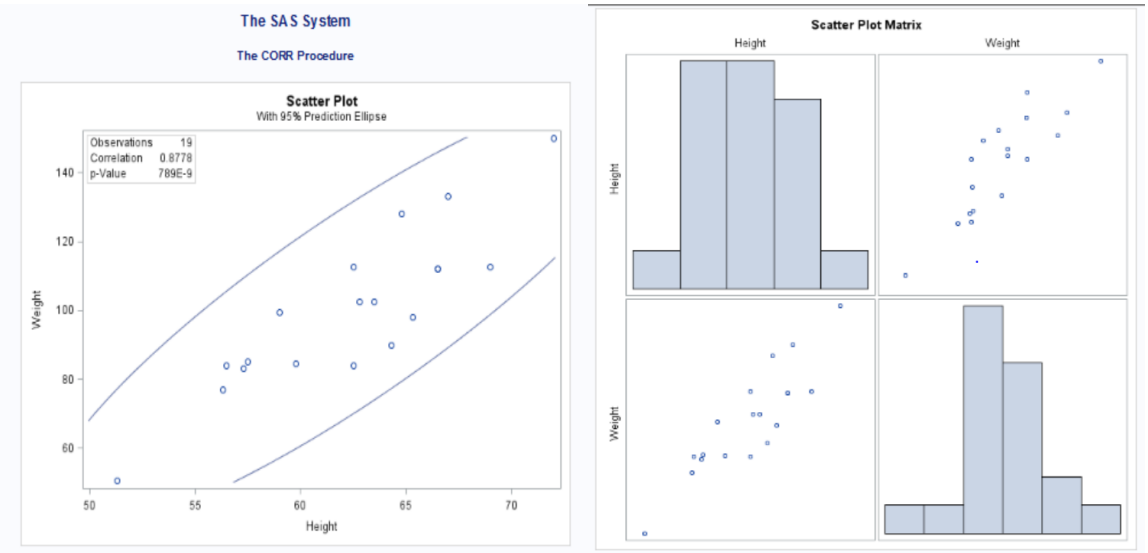
Pearson Correlation Coefficients, N = 19
Prob > |r| under H0: Rho=0

	Height	Weight
Height	1.00000	0.87779 <.0001
Weight	0.87779 <.0001	1.00000

Since the Pearson correlation coefficient is 0.87779, we can conclude that there is strong positive relationship between height and weight.

Since the probability value <.0001 we can conclude that correlation is highly significant.

```
proc corr data=sashelp.class plots=matrix (histogram) plot=scatter();
var height weight; run;
```



Scatter plot shows that the weight and height are linearly related as weight increase, height also increase.

EXAMPLE 2:

```
proc print data=sashelp.iris;
run;
```

Obs	Species	SepalLength	SepalWidth	PetalLength	PetalWidth
1	Setosa	50	33	14	2
2	Setosa	46	34	14	3
3	Setosa	46	36	10	2
4	Setosa	51	33	17	5
5	Setosa	55	35	13	2

```
NOTE: There were 150 observations read from the data set SASHELP.IRIS.
```

```
proc corr data=sashelp.iris;
run;
```

The CORR Procedure

4 Variables:	SepalLength SepalWidth PetalLength PetalWidth

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
SepalLength	150	58.43333	8.28066	8765	43.00000	79.00000	Sepal Length (mm)
SepalWidth	150	30.57333	4.35866	4586	20.00000	44.00000	Sepal Width (mm)
PetalLength	150	37.58000	17.65298	5637	10.00000	69.00000	Petal Length (mm)
PetalWidth	150	11.99333	7.62238	1799	1.00000	25.00000	Petal Width (mm)

Pearson Correlation Coefficients, N = 150
Prob > |r| under H0: Rho=0

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength Sepal Length (mm)	1.00000	-0.11757 0.1519	0.87175 <.0001	0.81794 <.0001
SepalWidth Sepal Width (mm)	-0.11757 0.1519	1.00000	-0.42844 <.0001	-0.36613 <.0001
PetalLength Petal Length (mm)	0.87175 <.0001	-0.42844 <.0001	1.00000	0.96287 <.0001
PetalWidth Petal Width (mm)	0.81794 <.0001	-0.36613 <.0001	0.96287 <.0001	1.00000

```
proc corr data=sashelp.iris;
var sepallength petallength;
run;
```

The CORR Procedure

2 Variables:	SepalLength PetalLength

Simple Statistics

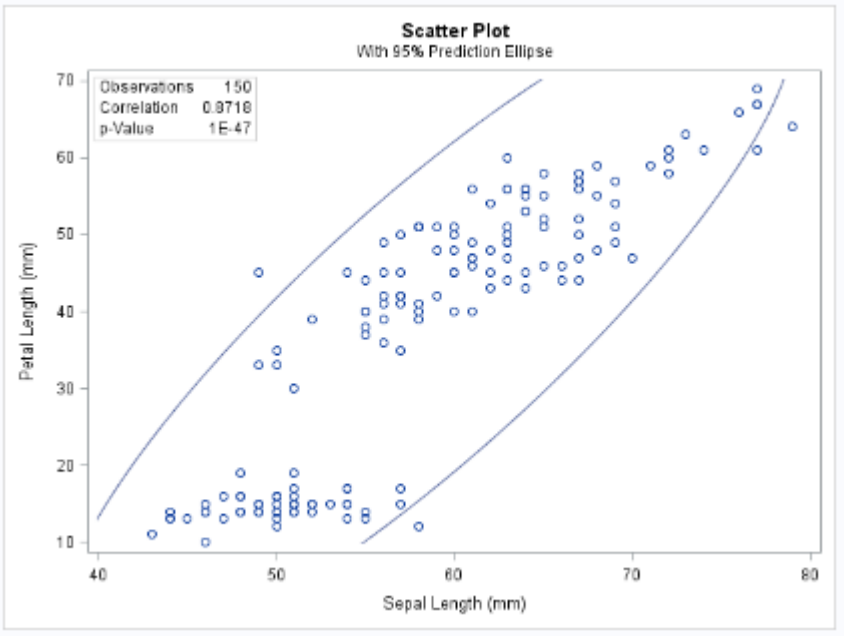
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
SepalLength	150	58.43333	8.28066	8765	43.00000	79.00000	Sepal Length (mm)
PetalLength	150	37.58000	17.65298	5637	10.00000	69.00000	Petal Length (mm)

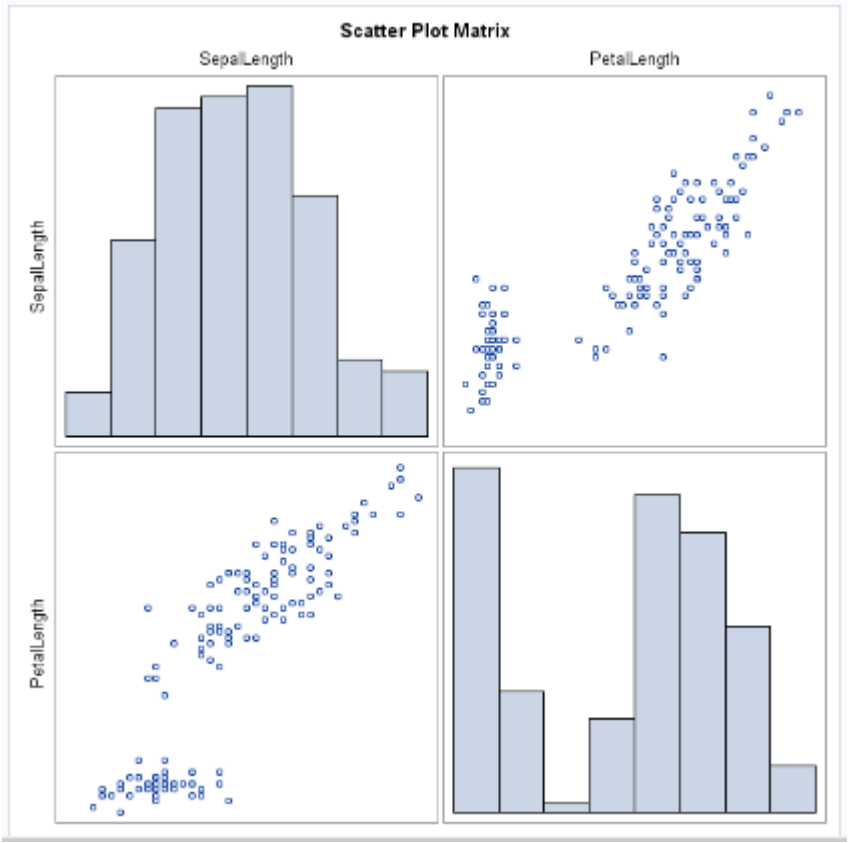
Pearson Correlation Coefficients, N = 150
Prob > |r| under H0: Rho=0

	SepalLength	PetalLength
SepalLength Sepal Length (mm)	1.00000	0.87175 <.0001
PetalLength Petal Length (mm)	0.87175 <.0001	1.00000

Based on the Pearson correlation coefficient, the value is 0.87175. Thus, we can conclude that there is strong positive relationship between sepal length and petal length.

```
proc corr data=sashelp.iris plots=matrix(histogram) plots=scatter();
var sepallength petallength;
run;
```





From the scatterplot of the variable sepal length and petal length, we can see that the points going from bottom left to the upper right, which is saying the correlation is positive.

EXAMPLE 3:
Use fish data from SAS help

```
title 'Fish Measurement Data';
proc corr data=sashelp.fish nomiss plots=matrix(histogram);
var Height Width Length3 Weight3;
run;
```

Fish Measurement Data

The CORR Procedure

4 Variables:	Height Width Length3 Weight3

Simple Statistics

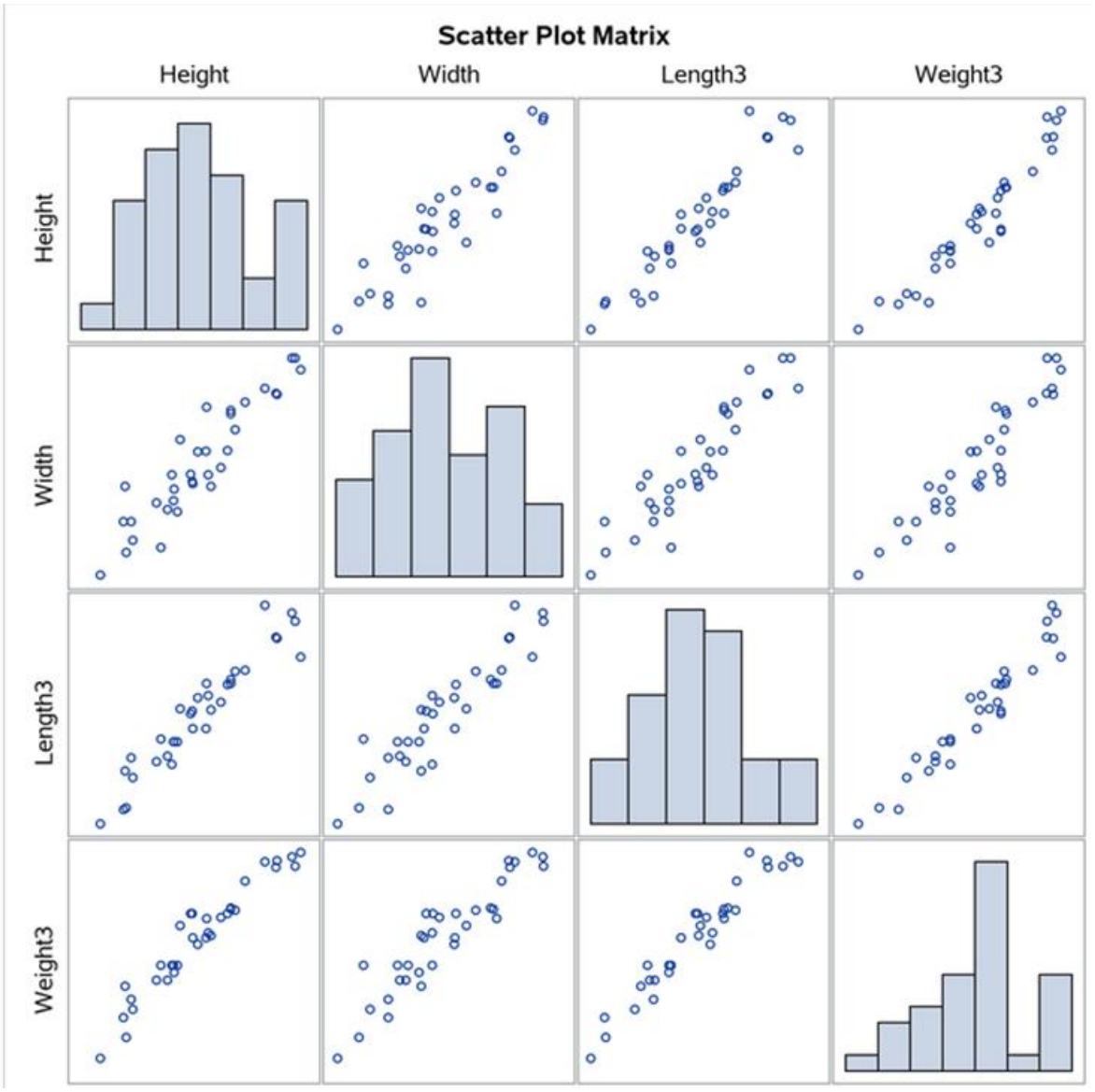
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Height	34	15.22057	1.98159	517.49950	11.52000	18.95700
Width	34	5.43805	0.72967	184.89370	4.02000	6.74970
Length3	34	38.38529	4.21628	1305	30.00000	46.50000
Weight3	34	8.44751	0.97574	287.21524	6.23168	10.00000

Output 2.8.2: Pearson Correlation Coefficients

Pearson Correlation Coefficients, N = 34 Prob > |r| under H0: Rho=0				
	Height	**Width**	**Length3**	**Weight3**
Height	1.00000	0.92632 <.0001	0.95492 <.0001	0.96261 <.0001
Width	0.92632 <.0001	1.00000	0.92171 <.0001	0.92789 <.0001
Length3	0.95492 <.0001	0.92171 <.0001	1.00000	0.96523 <.0001
Weight3	0.96261 <.0001	0.92789 <.0001	0.96523 <.0001	1.00000

Since the Pearson correlation coefficient is more than 0.7, we can conclude that there is strong positive relationship between height, width and length.

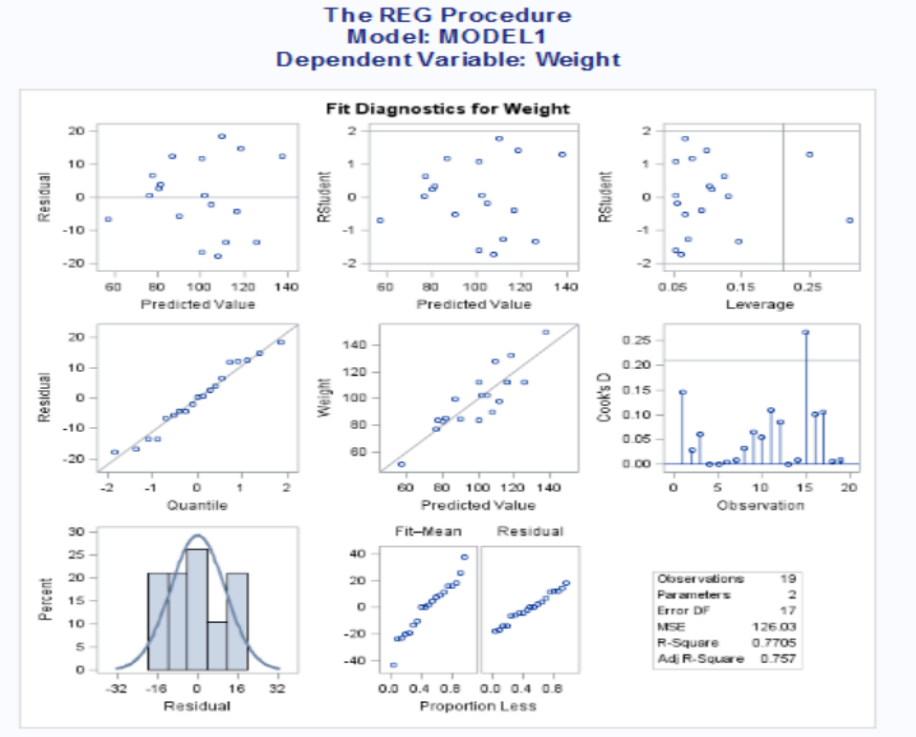
Since the Pearson correlation coefficient is 0.87779, we can conclude that there is strong positive relationship between height and weight.


Scatter Plot Matrix
Height
Width
Length3
Weight3
Height
Width
Length3
Weight3

REGRESSION ANALYSIS

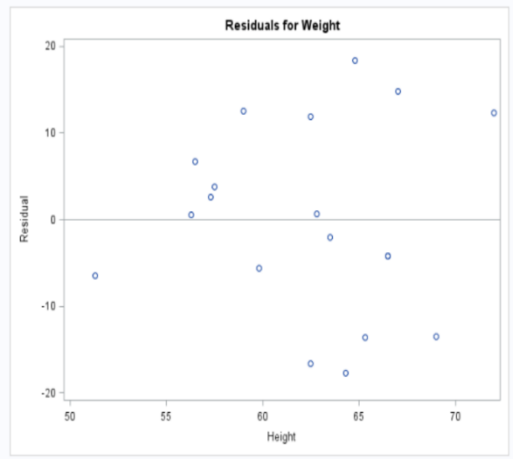
Example 1:

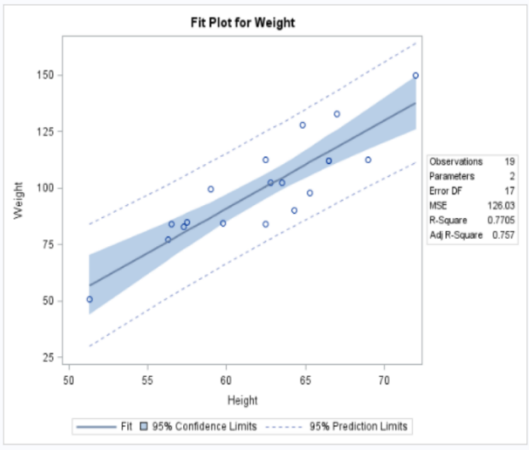
```
proc reg data=sashelp.class; model weight = height; run; quit;
```



From residual versus predicted value plot, we can see that the data are randomly scattered which indicate that the regression is independent and has constant variance.

From histogram of percent versus residual plot, we can see that the data are the data is normally distributed.





Plot of residual versus height show that data is linear.

```
proc reg data=sashelp.class rsquare;
   eq1: model  weight=height;
   eq2: model  weight=height age;
run;
```

The SAS System

The REG Procedure
Model: eq1
Dependent Variable: Weight

Number of Observations Read	19
Number of Observations Used	19

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

Root MSE	11.22625	R-Square	0.7705
Dependent Mean	100.02632	Adj R-Sq	0.7570
Coeff Var	11.22330		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-143.02692	32.27459	-4.43	0.0004
Height	1	3.89903	0.51609	7.55	<.0001

The SAS System

The REG Procedure
Model: eq2
Dependent Variable: Weight

Number of Observations Read	19
Number of Observations Used	19

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7215.63710	3607.81855	27.23	<.0001
Error	16	2120.09974	132.50623		
Corrected Total	18	9335.73684			

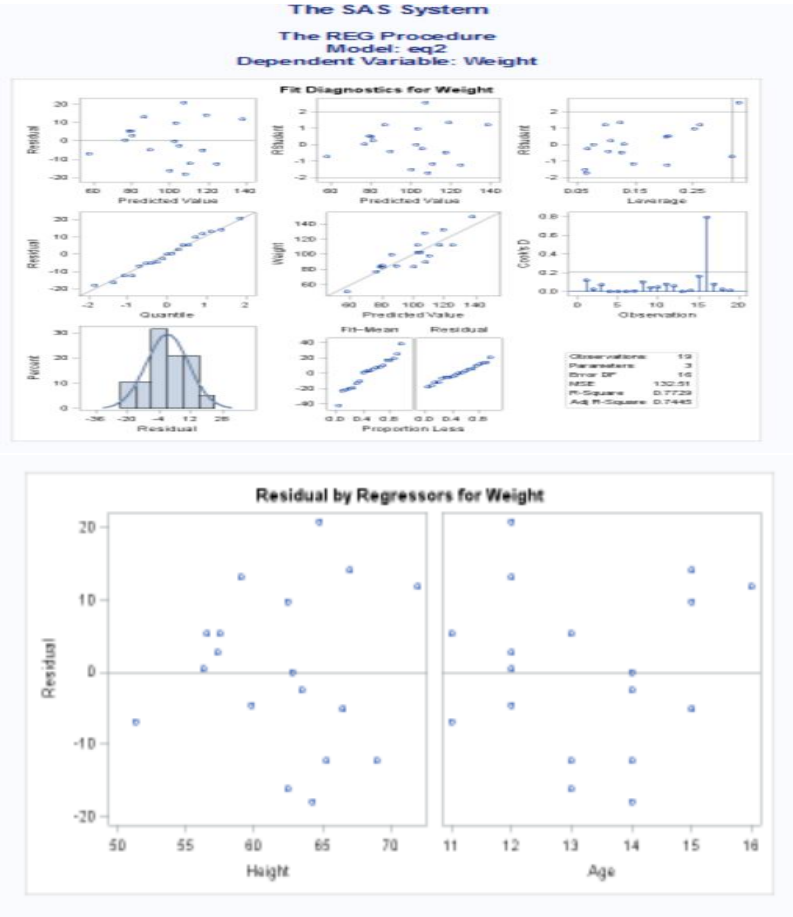
Root MSE	11.51114	R-Square	0.7729
Dependent Mean	100.02632	Adj R-Sq	0.7445
Coeff Var	11.50811		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-141.22376	33.38309	-4.23	0.0006
Height	1	3.59703	0.90546	3.97	0.0011
Age	1	1.27839	3.11010	0.41	0.6865

Model for eq1 is significant as the p-value is less than 0.0001. Height is significant in estimating weight as the p-value is less than 0.0001. The coefficient parameter for intercept and height are -143.02692 and 3.89903 respectively.

Model for eq2 also significant as the p-value is less than 0.0001. Height is significant in estimating weight as the p-value is less than 0.0001While age is not significant in estimating weight as the p-value is 0.6865 greater than 0.05. The coefficient parameter for intercept, height and age are -141.122376, 3.59703and 1.27839 respectively.



Plot of residual versus height and residual versus age show random patterns which indicate that the models are independent and have constant variance.

REFERENCES

Beers B. (2019). *Regression Definition.* Retrieved from https://www.investopedia.com/terms/r/regression.asp

Institute of digital research and education. Retrieved from https://stats.idre.ucla.edu/sas/webbooks/reg/chapter1/regressionwith-saschapter-1-simple-and-multiple-regression/

SAS/STAT/STAT(R) 9.3 User's Guide. Retrieved from https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_reg_sect054.htm

SAS Tutorials: Pearson Correlation with PROC CORR. Retrieved from https://libguides.library.kent.edu/SAS/PearsonCorr