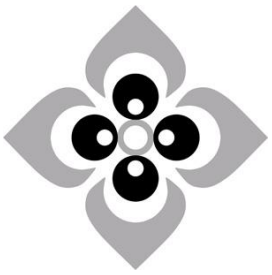


[Academic Script]

Linear Regression and Multiple Linear Regression in SAS

Subject:	Business Economics
Course:	B. A. (Hons.), 5th Semester, Undergraduate
Paper No. & Title:	Paper – 502 Computational Techniques for Management
Unit No. & Title:	Unit – 3 Econometric Problem Solving
Lecture No. & Title:	Lecture – 3 Linear Regression and Multiple Linear Regression in SAS

Academic Script

1. Introduction

Hello friends! In this final session on SAS I'll be talking about few statistical techniques namely Regression analysis, Anova test, simple linear regression and multiple linear regression.

Linear regression is the starting point of econometric analysis. The linear regression model has a dependent variable that is a continuous variable, while the independent variables can take any form (continuous, discrete, or indicator variables). A simple linear regression model has only one independent variable, while a multiple linear regression model has two or more independent variables. The linear regression is typically estimated using OLS (ordinary least squares).

2. Regression Analysis

We can also perform Regression analysis between dependent variable and independent variables. The dependent variable here is mass per gallon which is dependent upon the independent variables weight, length and foreign. So the command used to perform this regression analysis is:

```
* Regression model - mpg is dependent variable and weight, length and foreign are
independent variables;
proc reg data=auto;
model mpg = weight length foreign;
run;
```

looking at the p values of the output statistical significance could be gauged. The screen shows the output of regression command executed.

The REG Procedure

Model: MODEL1

Dependent Variable: mpg

Number of Observations Read 26

Number of Observations Used 26

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3378.69701	126.23234	14.84	<.0001
Error	22	187.14915	8.50678		
Corrected Total	25	565.84615			

Root MSE	2.91664	R-Square	0.6693
Dependent Mean	20.92308	Adj R-Sq	0.6242
Coeff Var	13.93982		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	44.96858	9.32268	4.82	<.0001
Weight	1	-0.00501	0.00219	-2.29	0.0320
Length	1	-0.04306	0.07693	-0.56	0.5813
Foreign	1	-1.26921	1.63213	-0.78	0.4451

Here by looking on the output generated we can see that price and weight have negative relation. Whereas the next two variables those are length and foreign, are statistically insignificant. This can be inferred by looking at their p values which are above 0.5

3. Anova Test

Lastly we will perform ANOVA test where we will check the mean of mass per gallon for the category of data i.e. domestic and foreign is same or not. We can perform this test using proc glm and the command for the same would be:

```
* ANOVA test - if the mean mpg is the same for foreign and domestic cars;
proc glm data=auto;
class foreign;
model mpg = foreign;
run;
```

The GLM Procedure

Class Level Information

Class	Levels	Values
foreign	2	0 1

Number of Observations Read	26
Number of Observations Used	26

The SAS System

The GLM Procedure

Dependent Variable: mpg

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	90.6882591	90.6882591	4.58	0.0427
Error	24	475.1578947	19.7982456		
Corrected Total	25	565.8461538			

R-Square	Coeff Var	Root MSE	mpg Mean
0.160270	21.26610	4.449522	20.92308

Source	DF	Type I SS	Mean Square	F Value	Pr > F
foreign	1	90.68825911	90.68825911	4.58	0.0427

Source	DF	Type III SS	Mean Square	F Value	Pr > F
foreign	1	90.68825911	90.68825911	4.58	0.0427

Here we can see that the p value is less than 0.5 hence it is statistically significant.

Importing the csv file

So now let us start by importing the csv file needed

For that the command needed to be given is:

proc import out= work.data

```
datafile= "C:\Econometrics\Data\regression_auto.csv"
dbms=csv replace; getnames=yes; datarow=2;
run;
```

Now the data that we are using ofr this example contains the information shown on the screen.

make	mpg	weight	weight1	price	foreign	repairs	length
AMC	22	2930	2.93	4099	0	3	186
AMC	17	3350	3.35	4749	0	3	173
AMC	22	2640	2.64	3799	0	3	168
Audi	17	2830	2.83	9690	1	5	189
Audi	23	2070	2.07	6295	1	3	174
BMW	25	2650	2.65	9735	1	4	177
Buick	20	3250	3.25	4816	0	3	196
Buick	15	4080	4.08	7827	0	4	222
Buick	18	3670	3.67	5788	0	3	218
Buick	26	2230	2.23	4453	0	3	170
Buick	20	3280	3.28	5189	0	3	200
Buick	16	3880	3.88	10372	0	3	207
Buick	19	3400	3.4	4082	0	3	200
Cadillac	14	4330	4.33	11385	0	3	221
Cadillac	14	3900	3.9	14500	0	2	204
Cadillac	21	4290	4.29	15906	0	3	204
Chevrolet	29	2110	2.11	3299	0	3	163
Chevrolet	16	3690	3.69	5705	0	4	212
Chevrolet	22	3180	3.18	4504	0	3	193
Chevrolet	22	3220	3.22	5104	0	2	200
Chevrolet	24	2750	2.75	3667	0	2	179
Chevrolet	19	3430	3.43	3955	0	3	197
Datsun	23	2370	2.37	6229	1	4	170
Datsun	35	2020	2.02	4589	1	5	165
Datsun	24	2280	2.28	5079	1	4	170
Datsun	21	2750	2.75	8129	1	4	184

We can see that we have 26 observations of the cars, than mass per gallon (mpg) which will be a dependent variable we want to explain, weight of the car is shown in the next column besides

which I have generated weight1 variable by diving the weight by 1000. Hence the weight now derived is in thousand pounds. Next is the price of the car in a particular year in hundreds. Next column shows whether it is foreign or not. And these are the variables that we will use as independent analysis.

4. Descriptive Statistics

Now the first thing we will do is to look at the descriptive statistics using proc means with mpg as dependent variable and weight1 price and foreign as independent ones.

```
* Descriptive statistics;
proc means data=data;
var mpg weight1 price foreign;
run;
```

The output would be as shown on the screen:

Variable	N	Mean	Std Dev	Minimum	Maximum
mpg	26	20.9230769	4.7575042	14.0000000	35.0000000
weight1	26	3.0992308	0.6950794	2.0200000	4.3300000
price	26	6651.73	3371.12	3299.00	15906.00
foreign	26	0.2692308	0.4523443	0	1.0000000

Here as we can see there 26 observation with mpg as 20.923 mean, weight is about 3000 pounds, 6651 for price and 0.29% car are foreign.

Detailed descriptive statistics

We can also look into more detailed descriptive statistics by using proc univariate and variable mpg.

```
* Detailed descriptive statistics;
```

```
proc univariate data=data;
var mpg;
run;
```

The output is shown on the Screen

The UNIVARIATE Procedure
Variable: mpg

Moments

N	26	Sum Weights	26
Mean	20.9230769	Sum Observations	544
Std Deviation	4.75750419	Variance	22.6338462
Skewness	0.93547297	Kurtosis	1.79270004
Uncorrected SS	11948	Corrected SS	565.846154
Coeff Variation	22.7380715	Std Error Mean	0.93302334

Basic Statistical Measures

Location		Variability	
Mean	20.92308	Std Deviation	4.75750
Median	21.00000	Variance	22.63385
Mode	22.00000	Range	21.00000
		Interquartile Range	6.00000

Tests for Location: Mu0=0

Test	Statistic		p Value	
Student's t	t	22.42503	Pr > |t|	<.0001
Sign	M	13	Pr >= |M|	<.0001
Signed Rank	S	175.5	Pr >= |S|	<.0001

Quantiles	
Quantile	Estimate
90%	26
75% Q3	23
50% Median	21
25% Q1	17
10%	15
5%	14
1%	14
0% Min	14

Extreme Observations

Lowest		Highest	
Value	**Obs**	**Value**	**Obs**
14	15	24	25
14	14	25	6
15	8	26	10
16	18	29	17
16	12	35	24

Here we can see so many statistical details like mean median and mode, the percentages, highest and lowest observations to outliers and can thus be determined.

5. Correlations

Next thing we can do is to find out correlation using proc corr, putting in equation of all the variables we have by executing the command:

```
*Correlations;
proc corr data=data;
var mpg weight1 price foreign;
run;
```

The output of executed command would look like:

The CORR Procedure

4 Variables: mpg weight1 price foreign

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
mpg	26	20.92308	4.75750	544.00000	14.00000	35.00000
weight1	26	3.09923	0.69508	80.58000	2.02000	4.33000
price	26	6652	3371	172945	3299	15906
foreign	26	0.26923	0.45234	7.00000	0	1.00000

Pearson Correlation Coefficients, N = 26

Prob > |r| under H0: Rho=0

	mpg	weight1	price	foreign
mpg	1.00000	-0.80816	-0.43846	0.40034
		<.0001	0.0251	0.0427
weight1	-0.80816	1.00000	0.55607	-0.60107

Pearson Correlation Coefficients, N = 26 Prob > |r| under H0: Rho=0

	mpg	weight1	price	foreign
	<.0001		0.0032	0.0012
price	-0.43846	0.55607	1.00000	0.08352
	0.0251	0.0032		0.6850
foreign	0.40034	-0.60107	0.08352	1.00000
	0.0427	0.0012	0.6850	

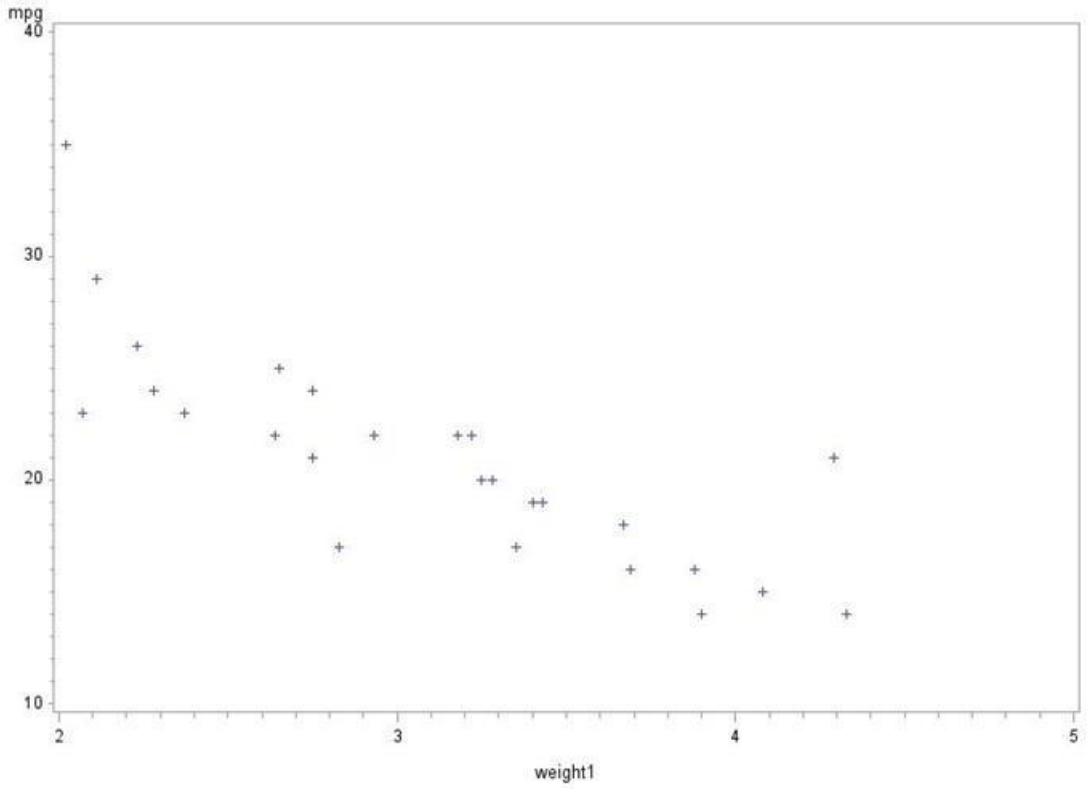
Here we can see that the highest correlation is between weight and mass per gallon (mpg), which is the independent and dependent variable, is -0.8

6. Plotting the Data

Now we can also plot the data using proc gplot, and we are plotting mpg*weight1.

```
* Plotting the data;
proc gplot data=data;
plot mpg*weight1;
run;
```

So this is how it looks like when the data plotting command is executed as shown on the screen. On the X axis we have independent variable weight1 and on y axis we have dependent variable mass per gallon (mpg). And the observations are scattered in the XY plane. When we observe the data we can make out that it has a negative relation. So when we are trying to fit a regression line we will generate a line as close as possible to all the points.



7. Simple Linear Regression

So now let us talk about simple linear regression on this example. Here we will take one dependent variable and one independent variable as the definition of simple linear regression states. The command for the same would be:

```
* Simple linear regression;
proc reg data=data;
model mpg = weight1;
run;
```

The results are as shown on the screen:

The REG Procedure

Model: MODEL1

Dependent Variable: mpg

Number of Observations Read	26
Number of Observations Used	26

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	369.56777	369.5677	45.19	<.0001
Error	24	196.27838	8.17827		
Corrected Total	25	565.84615			

Root MSE	2.85977	R-Square	0.6531
Dependent Mean	20.92308	Adj R-Sq	0.6387
Coeff Var	13.66800		

Parameter Estimates

Variable	DF	Parameter Estimates	Standard Error	t Value	Pr > |t|
Intercept	1	38.06646	2.61118	14.58	<.0001
weight1	1	-5.53150	0.82286	-6.72	<.0001

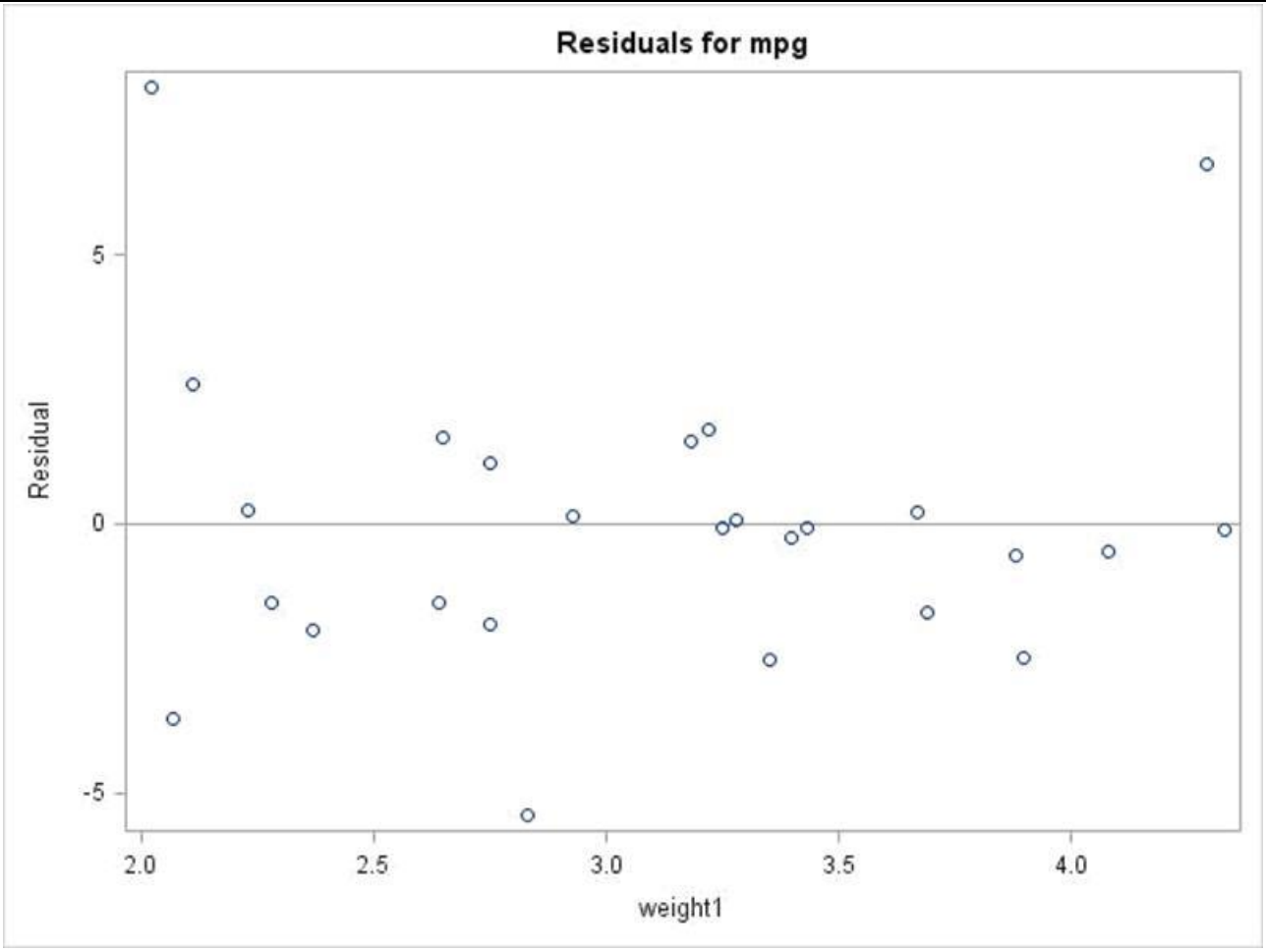
The third table shows parameter estimates on which we are going to concentrate. Notice that we have an intercept and next we have an independent variable weight1. The third column under the heading of parameter estimates are the coefficients of the variables. The interpretation of this figures says that if weight increases by one unit which in our case is one thousand pound than we will have a reduction of 5.53 mass per gallon, which is our dependent variable. Here I used the word reduction because the parameter estimate shows a negative sign. Now we

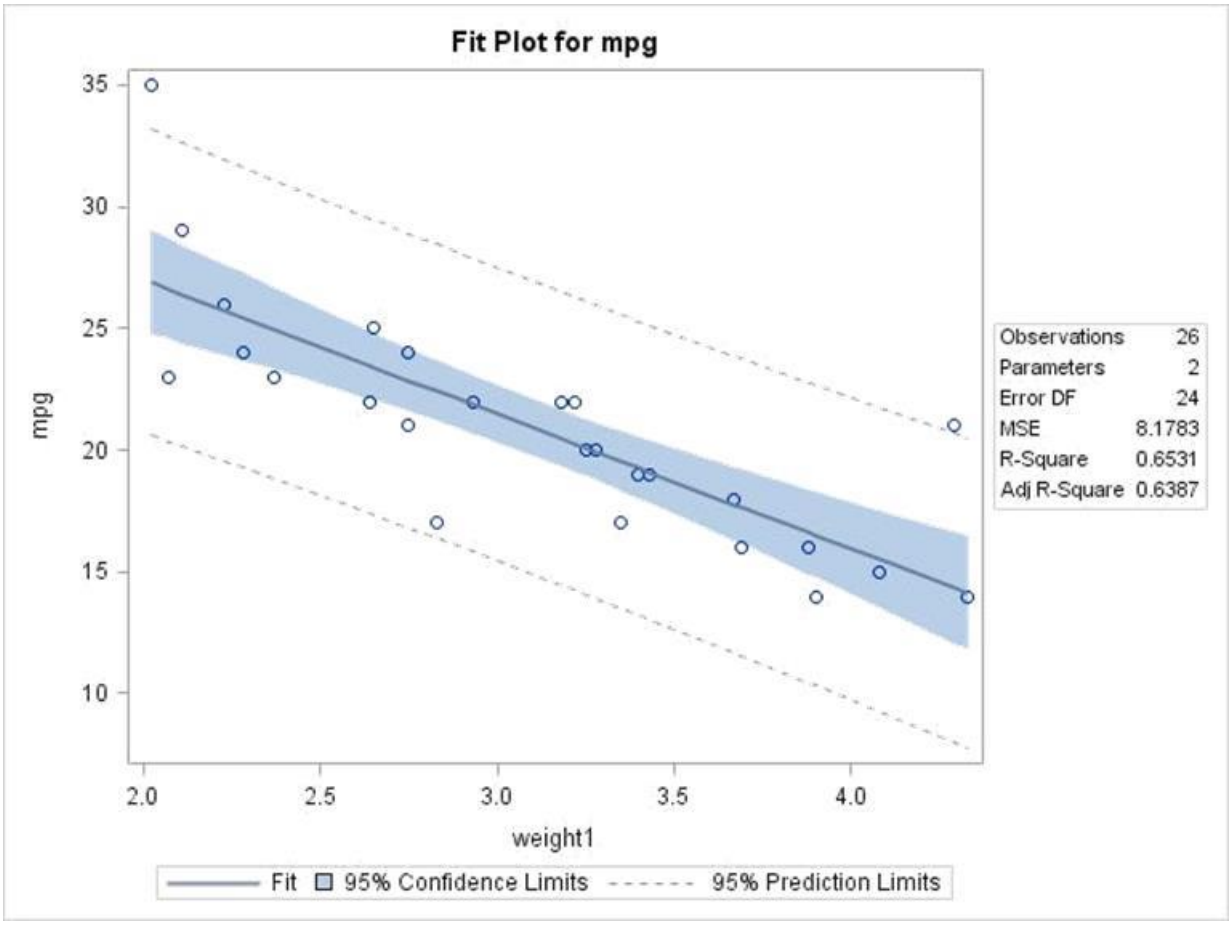
want to know whether this coefficient is significantly different from zero and to do that we will look at the standard error, also the t value which is nothing but parameter divided by standard errorand lastly the P value which should be less than 0.05. As we can see that P value is less than 0.05 we can say that is significantly negative relationship. We can also look into the ANOVA table for Model, Error and total variation. Note that we have one independent variable so the degree of freedom DF is 1. Also the total variation comes by total observation-1 = 25. In this table also the P value is less than 0.05 which means it is significant in nature and hence also called significantly different than 0.

Now as we can see on the screen we have some interesting graphs.

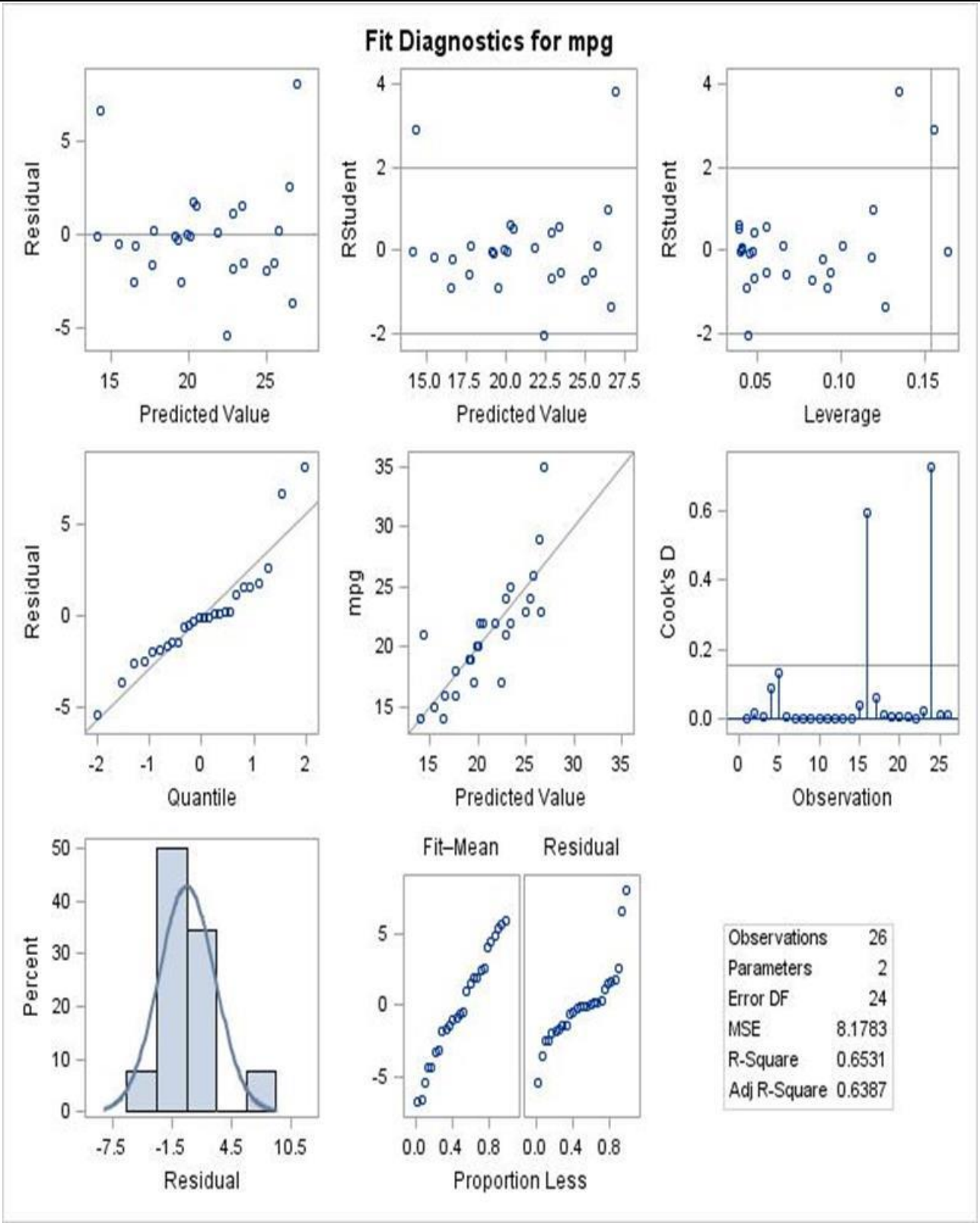
This graph is somewhat similar to scattered plot diagram that we saw earlier. On the X axis we have independent variable weight1 and on y axis we have dependent variable mass per gallon (mpg). And the observations are scattered in the XY plane. The circles are the actual observations and the line is the regression line. The circles are the actual value for dependent value y and the regression line would contain the predicted values.

Now the next graph of residual is nothing but flattening of this regression line as a straight horizontal line. And we can see that the difference between actual observation and predicted value is the error. This is the best thing about simple linear regression that is has only one dependent and one independent variable and hence it can be easily plotted.





And these are some more graphs:



8. Multiple Linear Regression

So now let us talk about multiple linear regression on this example. Here we will take one dependent variable and more than one independent variable as the definition of multiple linear regression states. Here we will take mpg as dependent variable

and weight1 price and foreign as independent variables. The command for the same would be:

```
* Multiple linear regression;
proc reg data=data;
model mpg = weight1 price foreign;
run;
```

The results are as shown on the screen:

The REG Procedure

Model: MODEL1

Dependent Variable: mpg

Number of Observations Read	26
Number of Observations Used	26

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	382.07964	127.35988	15.25	<.0001
Error	22	183.76652	8.35302		
Corrected Total	25	565.84615			

Root MSE	2.89016	R-Square	0.6752
Dependent Mean	20.92308	Adj R-Sq	0.6309
Coeff Var	13.81326		

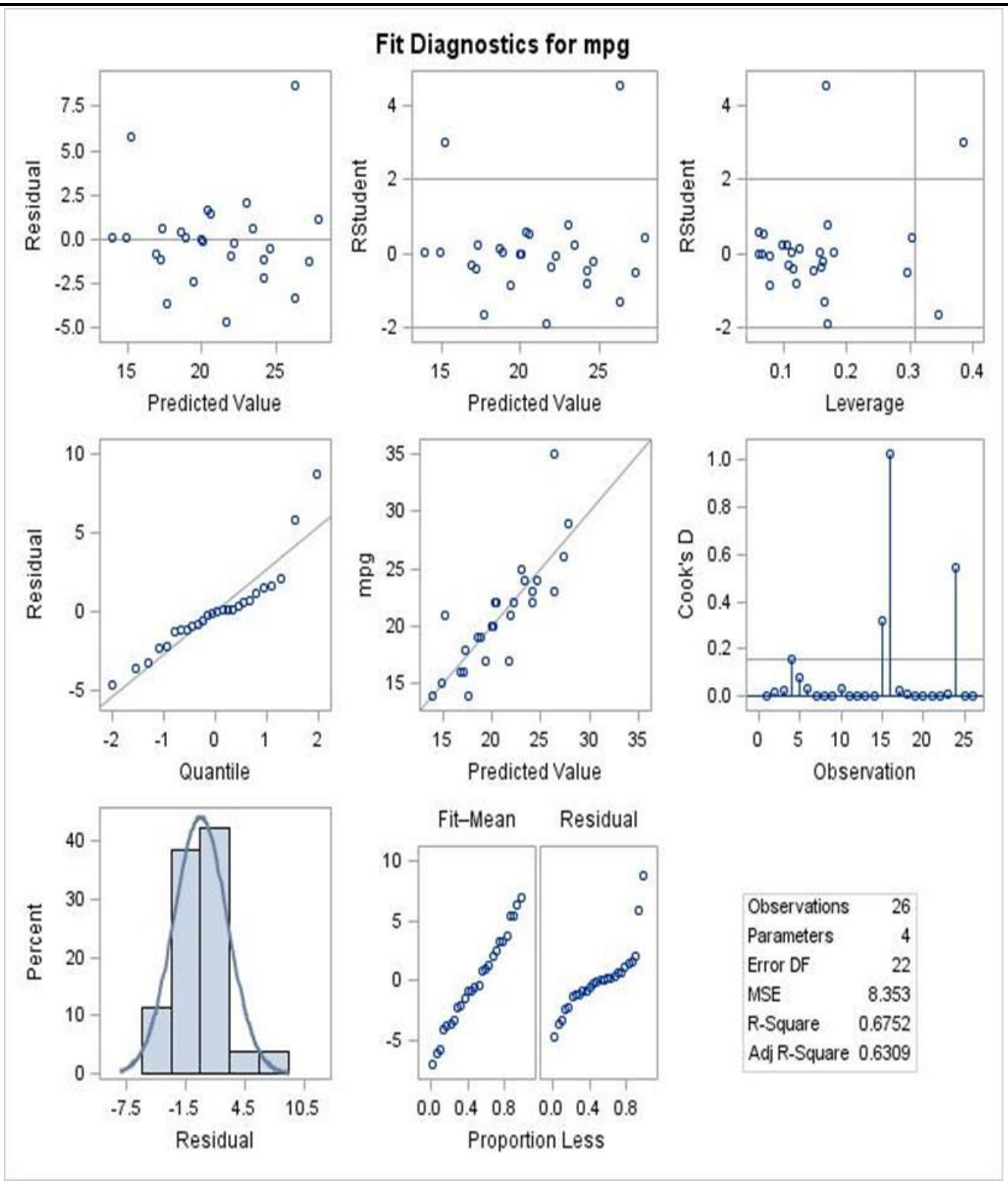
Parameter Estimates

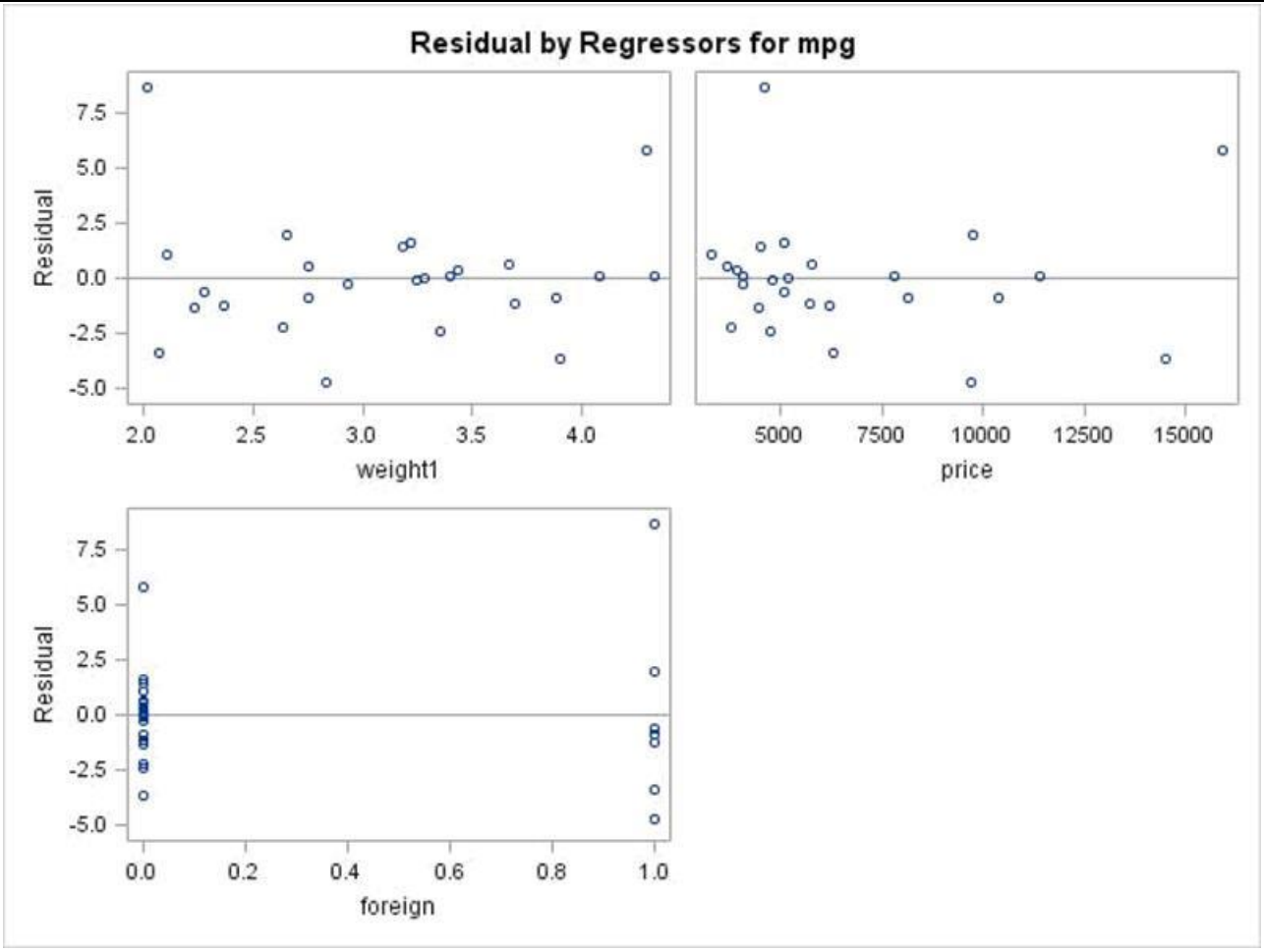
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	42.16620	4.26475	9.89	<.0001
weight1	1	-7.12111	1.60467	-4.44	0.0002
price	1	0.00022578	0.00026535	0.85	0.4040
foreign	1	-2.50713	2.05657	-1.22	0.2357

The third table shows parameter estimates on which we are going to concentrate. Notice that we have an intercept and next

we have now more independent variables; weight1, price and foreign. The third column under the heading of parameter estimates are the coefficients of the variables. The interpretation of this figures will again be same as mentioned earlier with only slightest difference in the value of weight1 as -7.12 which says that if weight increases by one unit which in our case is one thousand pound than we will have a reduction of 7.12 mass per gallon, which is our dependent variable. Now we want to know whether this coefficient is significantly different from zero and to do that we will look at the standard error, also the t value which is nothing but parameter divided by standard errorand lastly the P value which should be less than 0.05. As we can see that P value is less than 0.05 we can say that is significantly negative relationship. We can also look into the ANOVA table for Model, Error and total variation. Note that we have three independent variable so the degree of freedom DF is 3. Also the total variation comes by total observation-1 = 25. In this table also the P value is less than 0.05 which means it is significant in nature and hence also called significantly different than 0.

Now as we can see on the screen we have some interesting graphs. But as there are many more variables added now it becomes really difficult to understand from graphs hence table are better option.


Fit Diagnostics for mpg
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
mpg
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 26
Parameters 4
Error DF 22
MSE 8.353
R-Square 0.6752
Adj R-Square 0.6309



9. Summary

So friends let us summarize today's session. Today we discussed the Analysis part of SAS in which we looked upon how to perform regression analysis on dependency of dependent variable over independent variables and ANOVA test.

Also, we learnt two important statistical techniques in SAS namely simple linear regression and multiple linear regression. We also saw the interpretation of different variables and values shown in the output tables as well as the graph plotted for the same.

I hope the session would have given you brief insights on the analysis part of SAS program. Thank you for joining us.