Modelling Procedures in SAS

Module Content

Analysis

- PROC CORR
- PROC GENMOD
- PROC TTEST
- PROC GPLOT
- PROC BOXPLOT
- PROC LATTICE

Sampling

- PROC SURVEYSELECT
- PROC HPSPLIT

Modelling

Supervised Learning

Regression

- PROC REG
- PROC DISCRIM
- PROC GLM

Classification

- PROC LOGISTIC

Unsupervised Learning

Cluster Analysis

- PROC FACTOR
- PROC PRINCOM

Factor Analysis

- PROC VARCLUS
- PROC FASTCLUS
- PROC CLUSTER

Scoring

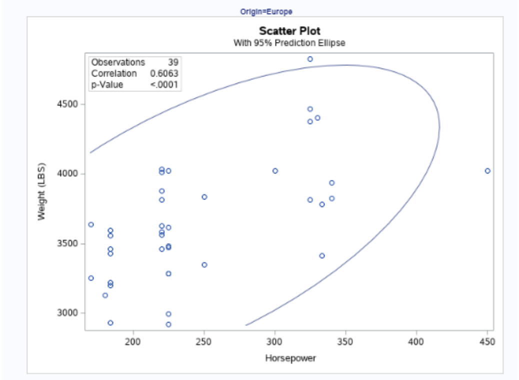
- PROC SCORE

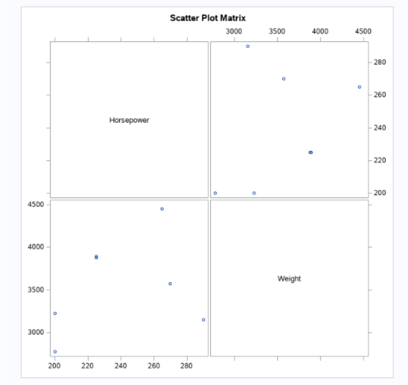
Proc Corr

Pearson correlation coefficient is a measure of linear relationship between two variables. Value of the coefficient is between -1 and +1.

```
Proc corr data = cars2 ;
        VAR horsepower weight ;
        BY origin;
run;
```

It can be run without By variable to get combined correlations.
plots = matrix/scatter option will add the visual correlation matrix/plot with observations as below





2 Variables:	Horsepower Weight

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Horsepower	7	239.28571	35.75712	1675	200.00000	290.00000	
Weight	7	3566	561.72939	24960	2778	4451	Weight (LBS)

Pearson Correlation Coefficients, N = 7
Prob > |r| under H0: Rho=0

	Horsepower	Weight
Horsepower	1.00000	0.30775 0.5019
Weight Weight (LBS)	0.30775 0.5019	1.00000

The CORR Procedure

Origin=Europe

2 Variables:	Horsepower Weight

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Horsepower	39	246.00000	65.12054	9594	170.00000	450.00000	
Weight	39	3655	429.35248	142539	2921	4824	Weight (LBS)

Pearson Correlation Coefficients, N = 39
Prob > |r| under H0: Rho=0

	Horsepower	Weight
Horsepower	1.00000	0.60631 <.0001
Weight Weight (LBS)	0.60631 <.0001	1.00000

PROC GENMOD

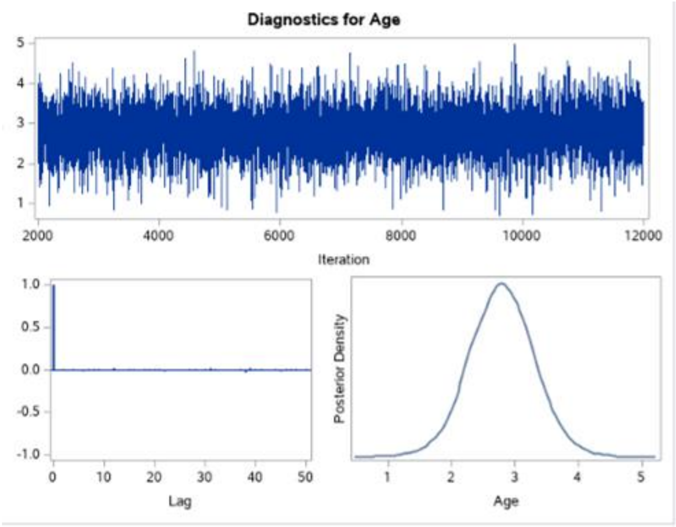
- PROC GENMOD helps us to do Bayesian analysis for distributions like normal,gamma, Gaussian, Poisson and binomial. It also provides Bayesian analysis for links like logit, probit, identity, log, etc. Model parameters are treated as random variables in Bayesian analysis, and inference is based on the posterior distribution of the parameters.

```
proc genmod data=class;

model height=age / dist=uniform;

bayes outpost=class2; run;
```

the dist= option specifies the kind of distribution
outpost = option saves samples (posterior) to the POST dataset.



PROC TTEST

- T-test is an analytical test used to test if there is a significant difference between two sets of data, or if mean of data differs significantly from a predicted value.

```
Proc sql;
create table carset as
select make, invoice, length , type, weight, horsepower from cars
Where make in ('Audi', 'BMW');
Quit;

proc ttest data = carset alpha = 0.01 h0 = 0;
var horsepower;
run;

proc ttest data = carset; paired weight*length;
run;
```

One Sample T test

The TTEST Procedure

Variable: Horsepower

N	Mean	Std Dev	Std Err	Minimum	Maximum
23	234.9	58.5657	12.2118	150.0	345.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
234.9	209.5	260.2	58.5657	45.2944	82.8910

<table><tr><th>DF</th><th>t Value</th><th>Pr > |t|</th></tr><tr><td>22</td><td>19.23</td><td><.0001</td></tr></table>

Paired T test

The TTEST Procedure

Difference: Weight - Length

N	Mean	Std Dev	Std Err	Minimum	Maximum
23	3629.4	711.7	148.4	2900.0	5658.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3629.4	3321.6	3937.2	711.7	550.4	1007.3

<table><tr><th>DF</th><th>t Value</th><th>Pr > |t|</th></tr><tr><td>22</td><td>24.46</td><td><.0001</td></tr></table>

PROC GPLOT

? Not working in SAS university Edition

Proc Gplot helps us to graphically present the information. Its alternative to proc plot. Notice the quit in the end which is needed to run this procedure.

```
proc plot data=sashelp.cars;
plot enginesize*msrp=make;
run;
quit;
proc gplot data=sashelp.cars;
plot enginesize*msrp=make;
run;
quit;
```

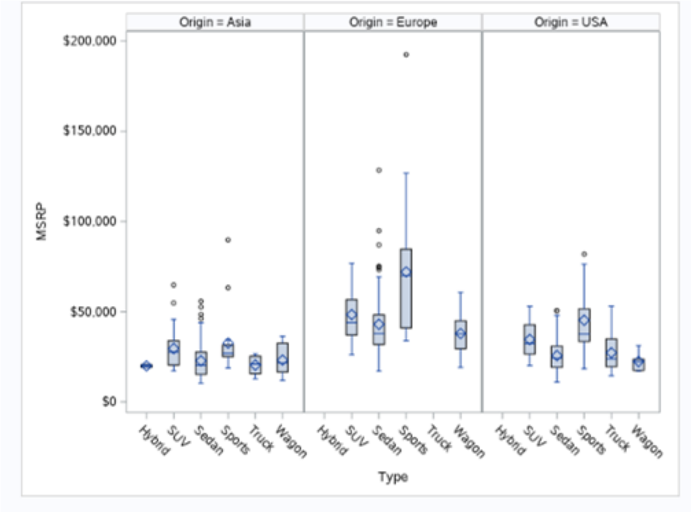
PROCs for BOXPLOT

A box-and-whiskers plot displays the *mean, quartiles, and minimum* and *maximum observations* for a group, organized in panels side by side.

```
proc sgplot data=sashelp.cars;
   vbox msrp/ category=type;
   run;

proc sgpanel data=sashelp.cars;
  panelby origin /rows=1 columns=3;
   vbox msrp/ category=type;
  run;
```

Use hbox for horizontal panel



PROC ANOVA

- The analysis of variance for balanced data can be performed by this procedure.

```
data heart;
set sashelp.heart;
run;
proc anova data=heart;
class smoking_status;
model cholesterol=smoking_status;
run;
```

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Smoking_Status	5	Heavy (16-25) Light (1-5) Moderate (6-15) Non-smoker Very Heavy (> 25)

Number of Observations Read	5209
Number of Observations Used	5049

The ANOVA Procedure

Dependent Variable: Cholesterol

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	22345.12	5586.28	2.77	0.0257
Error	5044	10168843.77	2016.03		
Corrected Total	5048	10191188.89			

R-Square	Coeff Var	Root MSE	Cholesterol Mean
0.002193	19.74114	44.90020	227.4448

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Smoking_Status	4	22345.11719	5586.27930	2.77	0.0257

PROC Lattice

This proc is used to perform the analysis of variance and simple covariance for data from an experiment with a lattice design – rectangular or square lattices.

```
proc lattice data=test;

run;
```

The Lattice Procedure

Analysis of Variance for Yield			
Source	DF	Sum of Squares	Mean Square
Replications	1	212.18	212.18
Blocks within Replications (Adj.)	8	501.84	62.7300
Component B	8	501.84	62.7300
Treatments (Unadj.)	24	559.28	23.3033
Intra Block Error	16	218.48	13.6550
Randomized Complete Block Error	24	720.32	30.0133
Total	49	1491.78	30.4445

Additional Statistics for Yield	
Variance of Means in Same Block	15.7915
Variance of Means in Different Blocks	17.9280
Average of Variance	17.2159
LSD at .01 Level	12.1189
LSD at .05 Level	8.7959
Efficiency Relative to RCBD	174.34

Adjusted Treatment Means for Yield	
Treatment	Mean
1	19.0681
2	16.9728
3	14.6463
4	14.7687
5	12.8470
6	13.1701
7	9.0748
8	6.7483
9	8.3707
10	8.4489
11	23.5511
12	12.4558
13	12.6293
14	20.7517
15	19.3299
16	12.6224
17	10.5272
18	10.7007
19	7.3231
20	11.4013
21	11.6259
22	18.5306
23	12.2041
24	17.3265
25	15.4048

```
data test;
input Group Block Treatment Yield @@;
datalines;
1 1 1 6
1 1 2 7
1 1 3 5
1 1 4 8
1 1 5 6
1 2 6 16
1 2 7 12
1 2 8 12
1 2 9 13
1 2 10 8
1 3 11 17
1 3 12 7
1 3 13 7
1 3 14 9
1 3 15 14
1 4 16 18
1 4 17 16
1 4 18 13
1 4 19 13
1 4 20 14
1 5 21 14
1 5 22 15
1 5 23 11
1 5 24 14
1 5 25 14
2 1 1 24
2 1 6 13
2 1 11 24
2 1 16 11
2 1 21 8
2 2 2 21
2 2 7 11
2 2 12 14
2 2 17 11
2 2 22 23
2 3 3 16
2 3 8 4
2 3 13 12
2 3 18 12
2 3 23 12
2 4 4 17
2 4 9 10
2 4 14 30
2 4 19 9
2 4 24 23
2 5 5 15
2 5 10 15
2 5 15 22
2 5 20 16
2 5 25 19

;
run;
```

Stratified Sampling: PROC Survey select

Proc SURVEYSELECT is used to create samples from datasets. This can be useful in creating a validation dataset for a model, selecting a random set of people to survey or forming a control group of customers to assess marketing campaign effectiveness.

```
proc surveyselect data = cars
          out = outdata
          method = srs
          samprate = (0.333 0.333 0.333)
          seed = 123;
 strata origin;
run;

Proc freq data = cars;
tables origin ;
run;

Proc freq data = outdata;
tables origin ;
run;
```

Use method = urs for unrestricted random sampling (allowing replacements) and method = srs for simple random sampling which restricts replacements.

Sampsize can be used to define size of sample and samprate can be used to express it in %age.

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Strata Variable	Origin

Input Data Set	CARS
Random Number Seed	123
Number of Strata	3
Total Sample Size	143
Output Data Set	CONTROL

The FREQ Procedure

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	158	36.92	158	36.92
Europe	123	28.74	281	65.65
USA	147	34.35	428	100.00

The FREQ Procedure

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	53	37.06	53	37.06
Europe	41	28.67	94	65.73
USA	49	34.27	143	100.00

?

PROC HPSPLIT

- The HPSPLIT procedure is a high-performance proc that is used to build tree-based models for classification and regression problems. The procedure produces classification trees (for categorical dependent variable), and regression trees(for continuous dependent variable). The output of model is displayed as a set of if-then statements and hence they are called trees.

```
proc hpsplit data=shmeq maxdepth=7 maxbranch=2;
target BAD; input DELIN DERO JOB NINQ RSON / level=nom;
input CLAGE CLN DEBTINC LOAN MORTDU VALUE YOJ / level=int; criterion entropy;
prune misc/ N <= 6; partition fraction(validate=0.2); rules file='rules1.txt';
score out=scored2;
run;
```

Regression: PROC REG

Linear regression models the relationship between a dependent variable and set of independent variable (s).

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

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
9	Jeffrey	M	13	62.5	84.0
10	John	M	12	59.0	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90.0
13	Louise	F	12	56.3	77.0
14	Mary	F	15	66.5	112.0
15	Philip	M	16	72.0	150.0
16	Robert	M	12	64.8	128.0
17	Ronald	M	15	67.0	133.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

PROC Reg: Output

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	19
Number of Observations Used	19

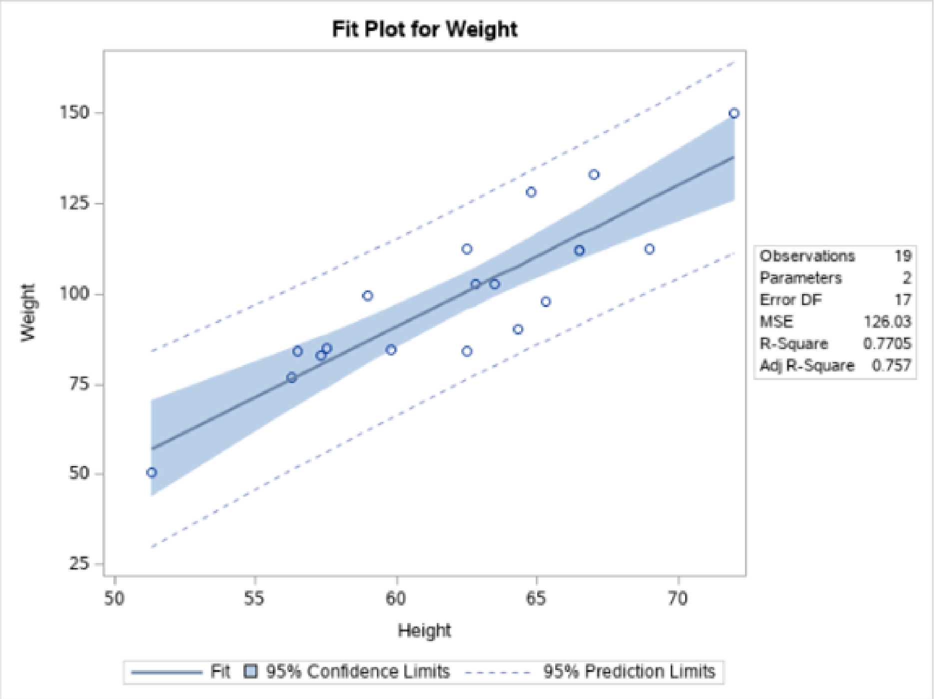
Analysis of Variance

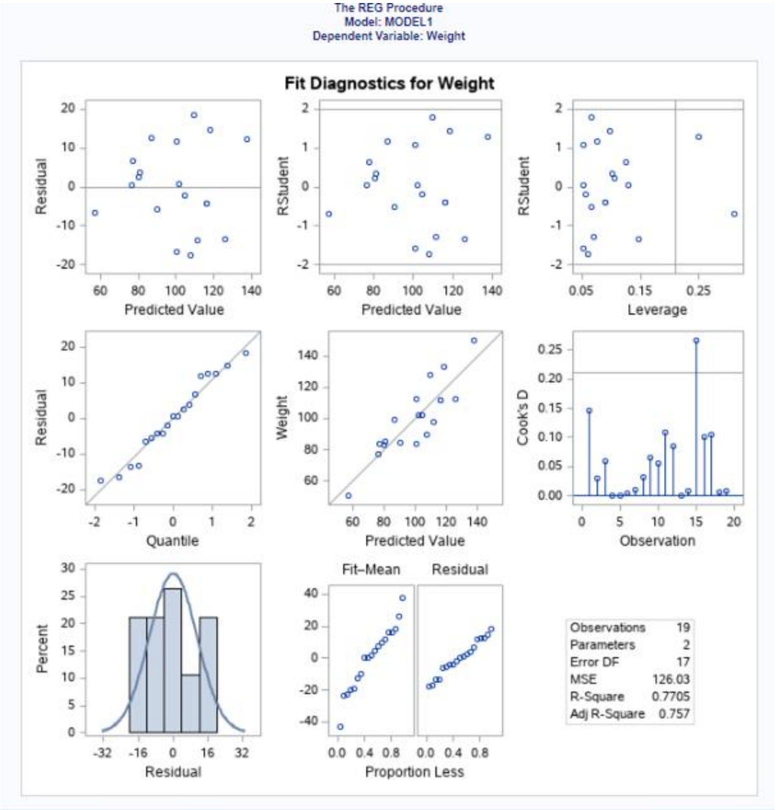
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

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





Parameters to evaluate the model

The R-square of 0.77 indicates that Height accounts for 77% of the variation in Weight

The p-values indicate that the intercept and Height parameter estimates are highly significant.

From the parameter estimates, the fitted model is Weight = 143.03 + 3.9 x Height

PROC REG – Full Syntax

- PROC REG;
  MODEL dependents=<regressors>
  BY variables;
  FREQ variable;
  ID variables;
  VAR variables;
  WEIGHT variable;
  ADD variables;
  CODE <options>;
  DELETE variables;
  MTEST<equation,. . .,equation>;
- OUTPUT<OUT=dataset> <keyword = names> <. . . keyword=names>;
  PLOT<yvariablexvariable> <=symbol> <. . . yvariablexvariable> <=symbol> </ options>;
  PRINT<options> <ANOVA> <MODEL DATA>;
  REFIT ;
  RESTRICT equation,. . .,equation;
  REWEIGHT<condition|ALLOBS> </ options> | <STATUS|UNDO>;
  STORE<options>;
  <label: > TESTequation,<,. . .,equation> </ option>;

Discriminant Analysis : PROC DISCRIM

- PROC DISCRIM is used to do discriminant analysis by which it classifies observations into multiple groups. It is like logistic regression, except that it allows multiple categories to be used and doesn't use maximum likelihood function.

```
data iris1;
set sashelp.iris;
run;


Proc DISCRIM data=iris1
distance anova MANOVA CROSSLISTERR;
class species;
var sepalwidth sepallength petalwidth petallength;
run;
```

The DISCRIM Procedure

Univariate Test Statistics

F Statistics, Num DF=2, Den DF=147

Variable	Label	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
SepalLength	Sepal Length (mm)	8.2807	5.1479	7.9506	0.6187	1.6226	119.26	<.0001
SepalWidth	Sepal Width (mm)	4.3587	3.3969	3.3682	0.4008	0.6688	49.16	<.0001
PetalLength	Petal Length (mm)	17.6530	4.3033	20.9070	0.9414	16.0566	1180.16	<.0001
PetalWidth	Petal Width (mm)	7.6224	2.0465	8.9673	0.9289	13.0613	960.01	<.0001

Average R-Square	
Unweighted	0.7224358
Weighted by Variance	0.8689444

Multivariate Statistics and F Approximations

S=2 M=0.5 N=71

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02343863	199.15	8	288	<.0001
Pillai's Trace	1.19189883	53.47	8	290	<.0001
Hotelling-Lawley Trace	32.47732024	582.20	8	203.4	<.0001
Roy's Greatest Root	32.19192920	1166.96	4	145	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Linear Discriminant Function for Species

Variable	Label	Setosa	Versicolor	Virginica
Constant		-85.20986	-71.75400	-103.26971
SepalLength	Sepal Length (mm)	2.35442	1.56982	1.24458
SepalWidth	Sepal Width (mm)	2.35879	0.70725	0.36853
PetalLength	Petal Length (mm)	-1.64306	0.52115	1.27665
PetalWidth	Petal Width (mm)	-1.73984	0.64342	2.10791

PROC GLM

- The linear regression model is a special case of a general linear model. In this case dependent variable is a continuous normally distributed
  and no class variables exist among the independent variables.

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

The GLM Procedure

Dependent Variable: Weight

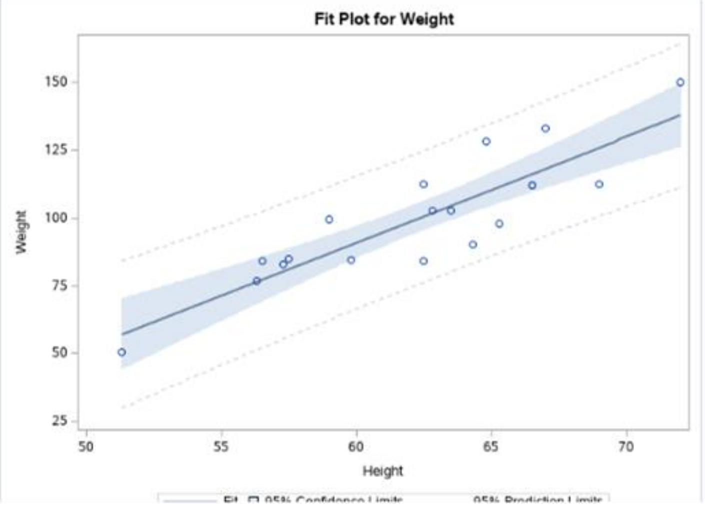
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.249119	7193.249119	57.08	<.0001
Error	17	2142.487723	126.028690		
Corrected Total	18	9335.736842			

R-Square	Coeff Var	Root MSE	Weight Mean
0.770507	11.22330	11.22625	100.0263

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Height	1	7193.249119	7193.249119	57.08	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Height	1	7193.249119	7193.249119	57.08	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-143.0269184	32.27459130	-4.43	0.0004
Height	3.8990303	0.51609396	7.55	<.0001



PROC LOGISTIC

- Proc Logistics is used to perform logistic regression on the categorical data in a classification problem. Its based on Maximum Likelihood (ML) Estimation and **Fisher Scoring** is generally used for iterative estimation of the regression parameters.

```
data cars1;
set sashelp.cars;
if mpg_highway <20 then efficiency=0;
else efficiency =1;
run;

ods graphics on;
proc logistic data=cars1;
model efficiency = enginesize weight;
oddsratio efficiency;
run;
```

- Logistic regression is most used proc for modelling decision variables which are categorical and used across industries.

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CARS
Response Variable	efficiency
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	428
Number of Observations Used	428

Response Profile

Ordered Value	efficiency	Total Frequency
1	0	41
2	1	387

Probability modeled is efficiency=0.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	272.276	103.429
SC	276.335	115.607
-2 Log L	270.276	97.429

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	172.8465	2	<.0001
Score	164.2460	2	<.0001
Wald	47.2099	2	<.0001

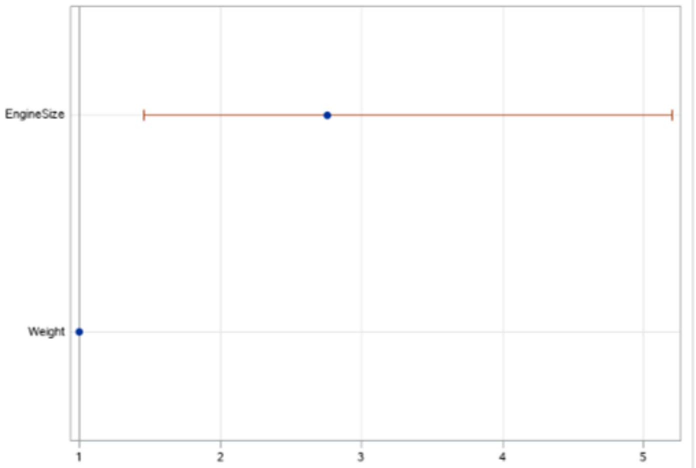
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-21.0881	2.9260	51.9446	<.0001
EngineSize	1	1.0141	0.3237	9.8114	0.0017
Weight	1	0.00350	0.000582	36.0590	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
EngineSize	2.757	1.462	5.200
Weight	1.004	1.002	1.005

Odds Ratios with 95% Wald Confidence Limits



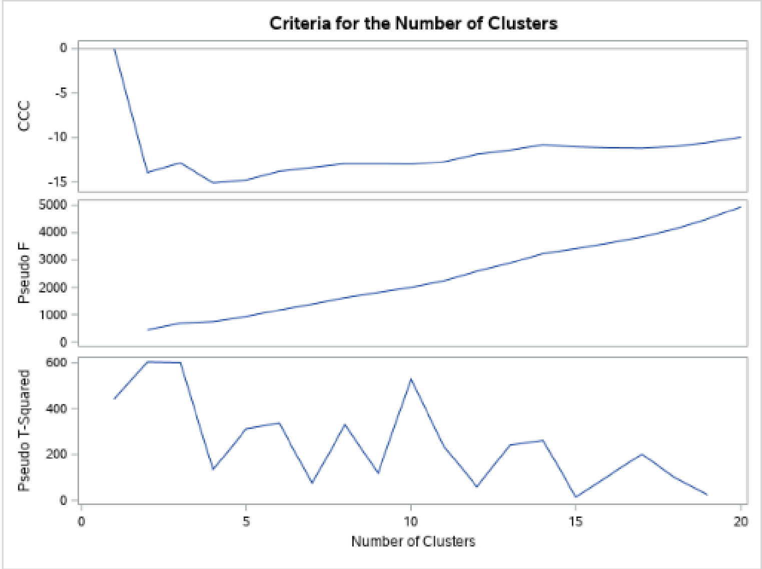
Association of Predicted Probabilities and Observed Responses

Percent Concordant	96.6	Somers' D	0.933
Percent Discordant	3.4	Gamma	0.933
Percent Tied	0.0	Tau-a	0.162
Pairs	15867	c	0.966

PROC Cluster

PROC CLUSTER performs hierarchical clustering of observations using distance data methods like complete linkage, average linkage, the centroid method, density linkage etc.

```
proc cluster data= CARS method=ave ccc pseudo PRINT=25
plots=den(height=rsq);
var Wheelbase;
id make;
run;
```



The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	69.0862352		1.0000	1.0000

Root-Mean-Square Total-Sample Standard Deviation	8.311813

Root-Mean-Square Distance Between Observations	11.75468

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
20	CL54	CL49	72	0.0006	.996	.998	-10	4945	.	
19	CL28	CL311	11	0.0007	.995	.998	-11	4499	22.6	
18	CL38	CL31	26	0.0008	.994	.997	-11	4131	98.1	
17	CL29	CL68	48	0.0008	.993	.997	-11	3842	201	
16	CL84	CL217	11	0.0009	.992	.996	-11	3621	.	
15	CL312	CL32	4	0.0010	.991	.996	-11	3419	13.4	
14	CL25	CL40	62	0.0012	.990	.995	-11	3239	261	
13	CL27	CL26	63	0.0021	.988	.995	-11	2888	241	
12	CL24	CL16	30	0.0026	.986	.994	-12	2586	57.9	
11	CL30	CL17	75	0.0039	.982	.992	-13	2238	235	
10	CL20	CL23	117	0.0044	.977	.991	-13	2003	528	
9	CL18	CL21	37	0.0054	.972	.988	-13	1817	117	
8	CL14	CL22	91	0.0076	.964	.985	-13	1626	330	
7	CL19	CL15	15	0.0124	.952	.980	-13	1392	74.6	
6	CL13	CL11	138	0.0191	.933	.973	-14	1174	337	
5	CL8	CL12	121	0.0352	.898	.961	-15	928	312	
4	CL9	CL7	52	0.0575	.840	.939	-15	743	134	
3	CL10	CL6	255	0.0764	.764	.890	-13	687	601	
2	CL3	CL5	376	0.2538	.510	.751	-14	443	604	
1	CL2	CL4	428	0.5099	.000	.000	0.00	.	443	

PROC VARCLUS

- PROC VARCLUS performs clustering of variables, it separates a set of variables using hierarchical clustering.

```
Proc varclus data=SASHELP.IRIS MAXCLUSTERS=5;
var PetalWidth SepalWidth;
run;
```

Oblique Principal Component Cluster Analysis

Observations	150	Proportion	1
Variables	2	Maxeigen	0

Clustering algorithm converged.

Cluster Summary for 1 Cluster					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	2	2	1.366126	0.6831	0.6339

Total variation explained = 1.366126 Proportion = 0.6831

Cluster 1 will be split because it has the largest second eigenvalue, 0.633874, which is greater than the MAXEIGEN=0 value.

Clustering algorithm converged.

Cluster Summary for 2 Clusters					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	1	1	1	1.0000	
2	1	1	1	1.0000	

Total variation explained = 2 Proportion = 1.0000

2 Clusters		R-squared with			
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	PetalWidth	1.0000	0.1340	0.0000	Petal Width (mm)
Cluster 2	SepalWidth	1.0000	0.1340	0.0000	Sepal Width (mm)

Standardized Scoring Coefficients			
Cluster		1	2
PetalWidth	Petal Width (mm)	1.00000	0.00000
SepalWidth	Sepal Width (mm)	0.00000	1.00000

Cluster Structure			
Cluster		1	2
PetalWidth	Petal Width (mm)	1.00000	-0.36613
SepalWidth	Sepal Width (mm)	-0.36613	1.00000

Inter-Cluster Correlations		
Cluster	1	2
1	1.00000	-0.36613
2	-0.36613	1.00000

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	1.366126	0.6831	0.6831	0.633874	0.6831	
2	2.000000	1.0000	1.0000	0.000000	1.0000	0.0000

PROC FASTCLUS

- PROC FASTCLUS performs k-means clustering based on distances computed from one or more variables. This is especially used for large data sets. This procedure uses Euclidean distances by default.

```
Proc fastclus data=cars maxclusters=15;
var EngineSize Cylinders;
run;
```

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=5 Maxiter=1

Initial Seeds		
Cluster	EngineSize	Cylinders
1	5.70000000	8.00000000
2	8.30000000	10.00000000
3	5.50000000	12.00000000
4	3.20000000	6.00000000
5	2.00000000	3.00000000

Criterion Based on Final Seeds =	0.3190

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	87	0.3835	1.2793		4	2.4630
2	2	0.7500	0.7500		3	2.7472
3	3	0.2041	0.3333		2	2.7472
4	190	0.3103	1.0478		5	2.3059
5	146	0.2829	1.1043		4	2.3059

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
EngineSize	1.10859	0.43038	0.850699	5.697896
Cylinders	1.55844	0.13568	0.992492	132.189156
OVER-ALL	1.35183	0.31940	0.944701	17.083671

Pseudo F Statistic =	1806.60

Approximate Expected Over-All R-Squared =	0.81471

Cubic Clustering Criterion =	31.942

WARNING: The two values above are invalid for correlated variables.

Cluster Means		
Cluster	EngineSize	Cylinders
1	4.72068966	8.00000000
2	7.55000000	10.00000000
3	5.66666667	12.00000000
4	3.28315789	6.00000000
5	2.06575342	4.04166667

Cluster Standard Deviations		
Cluster	EngineSize	Cylinders
1	0.542418771	0.000000000
2	1.060660172	0.000000000
3	0.288675135	0.000000000
4	0.438890528	0.000000000
5	0.325351451	0.232799923

PROC Factor

- This procedure computes a variety of common factors and rotations and used for variable reduction.

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	12
Number of Records Used	12
N for Significance Tests	12

Means and Standard Deviations from 12 Observations		
Variable	Mean	Std Dev
Population	6241.667	3439.9943
School	11.442	1.7865
Employment	2333.333	1241.2115
Services	120.833	114.9275
HouseValue	17000.000	6367.5313

Correlations					
	Population	School	Employment	Services	HouseValue
Population	1.00000	0.00975	0.97245	0.43887	0.02241
School	0.00975	1.00000	0.15428	0.69141	0.86307
Employment	0.97245	0.15428	1.00000	0.51472	0.12193
Services	0.43887	0.69141	0.51472	1.00000	0.77765
HouseValue	0.02241	0.86307	0.12193	0.77765	1.00000

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

2 factors will be retained by the MINEIGEN criterion.

Factor Pattern		
	Factor1	Factor2
Population	0.58096	0.80642
School	0.76704	-0.54476
Employment	0.67243	0.72605
Services	0.93239	-0.10431
HouseValue	0.79116	-0.55818

Variance Explained by Each Factor	
Factor1	Factor2
2.8733136	1.7966601

Final Communality Estimates: Total = 4.669974				
Population	School	Employment	Services	HouseValue
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041

```
data test;
input Population School Employment Services
House_Value;
cards;
  5500   12.8   2500   270   24000
  1000   10.9   640    10    10030
  3440   8.8    1000   10    9300
  3800   13.6   1700   140   25000
  4000   12.8   1600   140   25000
  8200   8.3    2600   60    12000
  1200   11.4   400    10    16000
  9100   11.5   3300   60    14000
  9900   12.5   3400   180   18000
  9600   13.7   3600   390   25000
  9600   9.6    3300   80    12000
  9400   11.4   4000   100   13000
  ;
proc factor data=test simple corr;
run;
```

ROTATE= (PROMAX/VARIMAX for example) option can be used produce rotations

PROC PRINCOM

- Like Proc FACTOR, proc PRINCOMP also does PCA or dimensionality reduction. It’s a linear combination of variables where weights are chosen using explanation of highest variation using eigen values

- procPRINCOMP is slightly faster than PROC FACTOR

```
proc princomp data=SEco out = test1 outstat=stat;
run;
```

Observations	12
Variables	5

Simple Statistics

	Population	School	Employment	Services	HouseValue
Mean	6241.666667	11.44166667	2333.333333	120.8333333	17000.00000
StD	3439.994274	1.78654483	1241.211529	114.9275134	6367.53128

Correlation Matrix

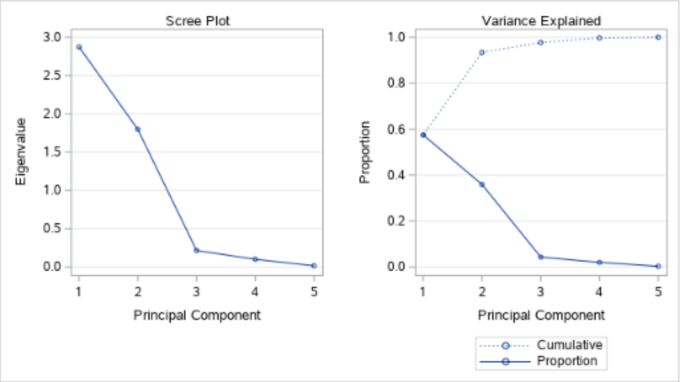
	Population	School	Employment	Services	HouseValue
Population	1.0000	0.0098	0.9724	0.4389	0.0224
School	0.0098	1.0000	0.1543	0.6914	0.8631
Employment	0.9724	0.1543	1.0000	0.5147	0.1219
Services	0.4389	0.6914	0.5147	1.0000	0.7777
HouseValue	0.0224	0.8631	0.1219	0.7777	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Population	0.342730	0.601629	0.059517	0.204033	0.689497
School	0.452507	-.406414	0.688822	-.353571	0.174861
Employment	0.396695	0.541665	0.247958	0.022937	-.698014
Services	0.550057	-.077817	-.664076	-.500386	-.000124
HouseValue	0.466738	-.416429	-.139649	0.763182	-.082425



PROC SCORE

- PROC SCORE is a SAS Post Processing procedure, where we use values obtained already from the processed dataset and do our execution. This allows to reduce the processing time. Proc score combined data from two data set – raw data and the one which has the coefficient processed from another procedure.

```
proc reg data=sashelp.cars outest=test1;
model mpg_city=weight horsepower length;
run;

proc score data=sashelp.cars score=test1 type=parms predict out=test2;
var weight horsepower length;
run;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	39.43701	2.25174	17.51	<.0001
Weight	Weight (LBS)	1	-0.00346	0.00034211	-10.10	<.0001
Horsepower		1	-0.02572	0.00283	-9.09	<.0001
Length	Length (IN)	1	-0.00784	0.01518	-0.52	0.6059

7/13/2020

Results: WORK.TEST2

Obs	Make	Model	Type	Origin	DriveTrain	MSRP	Invoice	EngineSize	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase	Length	MODEL1
1	Acura	MDX	SUV	Asia	All	$36,945	$33,337	3.5	6	265	17	23	4451	106	189	15.7602
2	Acura	RSX Type S 2dr	Sedan	Asia	Front	$23,820	$21,761	2.0	4	200	24	31	2778	101	172	23.3461
3	Acura	TSX 4dr	Sedan	Asia	Front	$26,990	$24,647	2.4	4	200	22	29	3230	105	183	21.6980
4	Acura	TL 4dr	Sedan	Asia	Front	$33,195	$30,299	3.2	6	270	20	28	3575	108	186	18.6821
5	Acura	3.5 RL 4dr	Sedan	Asia	Front	$43,755	$39,014	3.5	6	225	18	24	3880	115	197	18.6993
6	Acura	3.5 RL w/Navigation 4dr	Sedan	Asia	Front	$46,100	$41,100	3.5	6	225	18	24	3893	115	197	18.6543
7	Acura	NSX coupe 2dr manual S	Sports	Asia	Rear	$89,765	$79,978	3.2	6	290	17	24	3153	100	174	19.7200
8	Audi	A4 1.8T 4dr	Sedan	Europe	Front	$25,940	$23,508	1.8	4	170	22	31	3252	104	179	22.4248
9	Audi	A41.8T convertible 2dr	Sedan	Europe	Front	$35,940	$32,506	1.8	4	170	23	30	3638	105	180	21.0832
10	Audi	A4 3.0 4dr	Sedan	Europe	Front	$31,840	$28,846	3.0	6	220	20	28	3462	104	179	20.4133
11	Audi	A4 3.0 Quattro 4dr manual	Sedan	Europe	All	$33,430	$30,366	3.0	6	220	17	26	3583	104	179	19.9852
12	Audi	A4 3.0 Quattro 4dr auto	Sedan	Europe	All	$34,480	$31,388	3.0	6	220	18	25	3627	104	179	19.8432
13	Audi	A6 3.0 4dr	Sedan	Europe	Front	$36,640	$33,129	3.0	6	220	20	27	3561	109	192	19.9693
14	Audi	A6 3.0 Quattro 4dr	Sedan	Europe	All	$39,640	$35,992	3.0	6	220	18	25	3880	109	192	18.8670
15	Audi	A4 3.0 convertible 2dr	Sedan	Europe	Front	$42,490	$38,325	3.0	6	220	20	27	3814	105	180	19.1892
16	Audi	A4 3.0 Quattro convertible 2dr	Sedan	Europe	All	$44,240	$40,075	3.0	6	220	18	25	4013	105	180	18.5015
17	Audi	A6 2.7 Turbo Quattro 4dr	Sedan	Europe	All	$42,840	$38,840	2.7	6	250	18	25	3836	109	192	18.2476
18	Audi	A6 4.2 Quattro 4dr	Sedan	Europe	All	$49,690	$44,936	4.2	8	300	17	24	4024	109	193	16.3042
19	Audi	A8 L Quattro 4dr	Sedan	Europe	All	$69,190	$64,740	4.2	8	330	17	24	4399	121	204	14.1507
20	Audi	S4 Quattro 4dr	Sedan	Europe	All	$48,040	$43,556	4.2	8	340	14	20	3825	104	179	16.0729
21	Audi	RS 6 4dr	Sports	Europe	Front	$84,600	$76,417	4.2	8	450	15	22	4024	109	191	12.4624
22	Audi	TT 1.8 convertible 2dr (coupe)	Sports	Europe	Front	$35,940	$32,512	1.8	4	180	20	28	3131	95	159	22.7425
23	Audi	TT 1.8 Quattro 2dr (convertible)	Sports	Europe	All	$37,390	$33,891	1.8	4	225	20	28	2921	96	159	22.3109
24	Audi	TT 3.2 coupe 2dr (convertible)	Sports	Europe	All	$40,590	$36,739	3.2	6	250	21	29	3351	96	159	20.1821

Missing Values and Outlier Treatment

In the modelling process, it is a common step to impute missing data.

- Replace missing values or outliers with mean of the continuous variable.
- Replace missing values with median value of the ordinal categorical variables.
- Replace missing values with mode value of the nominal categorical variables.

Example:

```
proc stdize data=old reponly
        method=median
        out=new;
        var Var1 Var2 Var3;
run;
```

Thank You