Chapter 6 Analysis of Variance (ANOVA) in SAS

In this chapter, we mainly introduce PROC GLM for conducting ANOVA. Besides, PROC TTEST will be briefly mentioned.

6.1 Analysis of Variance (ANOVA) with PROC GLM

Analysis of variance (ANOVA) is a widely used statistical method, especially in experimental design to examine differences between various treatment groups.

Conceptually, it partitions the variation in the dependent variable to different sources (model and residual) and examines the ratio between the variation explained by model to the unexplained variation (residual). If the ratio is large enough according to a specific F-distribution or the p-value of the F-test is smaller than pre-specified alpha value, we reject the null hypothesis which indicates significant group differences.

- Omnibus ANOVA test examines if all group means are equal. This test is rejected if at least one group has a different mean value from others.
- If the omnibus ANOVA test is rejected, we conduct post-hoc comparisons to find where the difference lies, that is to compare each pair of the group means. To do so, we need to specify a correction method for adjusting inflated Type-I error rate.
- More on stats? Come to EPS702.

Example data: *sashelp.cars*. We examine if there is a difference between car types on the highway MPG. Code:

```
*One-way ANOVA with PROC GLM*;
proc glm data=sashelp.cars;
  class type;
  model MPG_Highway = type;
  means type / tukey;
run;
```

- PROC GLM line: specify dataset
- CLASS statement: indicate the categorical independent variable / grouping variable
- MODEL statement: specify the model
- MEANS statement: request for the group means and conduct post-hoc comparisons
    - Methods for post-hoc comparisons include Tukey (TUKEY), Bonferroni (BON), Scheffe (SCHEFFE), and least significant differences (LSD; no adjustment).

Results:

The GLM Procedure

Class Level Information		
Class	Levels	Values
Type	6	Hybrid SUV Sedan Sports Truck Wagon

Number of Observations Read	428
Number of Observations Used	428

Grouping/categorical independent variable

Number of valid responses

The GLM Procedure

Dependent Variable: MPG_Highway MPG (Highway)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6743.47900	1348.69580	77.64	<.0001
Error	422	7331.03268	17.37212		
Corrected Total	427	14074.51168			

ANOVA table

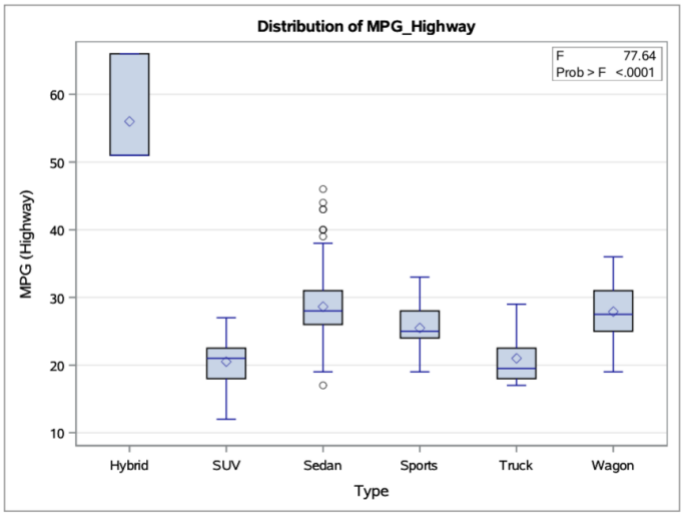
R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.479127	15.52701	4.167987	26.84346

R-square: amount of variation explained by model (SSM/SST)
Root MSE: square root of mean square residual

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Type	5	6743.478998	1348.695800	77.64	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Type	5	6743.478998	1348.695800	77.64	<.0001

Model effects: model variation due to car types



Side-by-side boxplot for highway MPG with different car types

A template of ANOVA table for one-way ANOVA

Source	**DF**	**Sum of Squares**	**Mean Square**	**F value**	**Pr > F**
Model	k-1	SS_M	$MS_M = SS_M/(k-1)$	$F = MS_M/MS_E$	p value
Error	n-k	SS_E	$MS_E = SS_E/(n-k)$		
Corrected Total	n-1	SS_T			

Note: k = number of groups (car types); n = sample size. $SS_T = SS_M + SS_E$.

Partial output for post-hoc comparisons (mean difference followed by its 95% CI, significance)

The GLM Procedure

Tukey's Studentized Range (HSD) Test for MPG_Highway

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	422
Error Mean Square	17.37212
Critical Value of Studentized Range	4.04870

Comparisons significant at the 0.05 level are indicated by *.**

Type Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
Hybrid - Sedan	27.3702	20.4417	34.2987	***
Hybrid - Wagon	28.1000	20.8746	35.3254	***
Hybrid - Sports	30.5102	23.4133	37.6071	***
Hybrid - Truck	35.0000	27.6929	42.3071	***
Hybrid - SUV	35.5000	28.4407	42.5593	***
Sedan - Hybrid	-27.3702	-34.2987	-20.4417	***
Sedan - Wagon	0.7298	-1.5701	3.0297	
Sedan - Sports	3.1400	1.2828	4.9972	***
Sedan - Truck	7.6298	5.0850	10.1746	***
Sedan - SUV	8.1298	6.4220	9.8375	***
Wagon - Hybrid	28.1000	35.3254	20.8746	***

6.2 T-Tests with PROC TTEST

a. Independent samples

Let's create a subset of data responses that only contain the cars with front and rear drivetrains.

```
*Create a subset of data cars*;
data cars_sub;
  set sashelp.cars;
   where drivetrain in ('Front' 'Rear');
run;
```

Next, we conduct independent samples t-test using PROC TTEST.

```
*Perform t-test with PROC TTEST*;
proc ttest data=cars_sub;
   class drivetrain;
   var MPG_Highway;
run;
```

- PROC TTEST line: specify dataset
- CLASS statement: indicate grouping variable
- VAR statement: specify dependent variable. If we include multiple variables here, a separate t-test will be conducted for each variable with the same grouping variable.

Results:

The TTEST Procedure

Variable: MPG_Highway (MPG (Highway))

DriveTrain	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Front		226	29.5044	5.8858	0.3915	18.0000	66.0000
Rear		110	25.0364	3.0044	0.2865	18.0000	36.0000
Diff (1-2)	Pooled		4.4681	5.1266	0.5960		
Diff (1-2)	Satterthwaite		4.4681		0.4851		

Descriptive statistics in general

DriveTrain	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Front		29.5044	28.7329	30.2759	5.8858	5.3886	6.4848
Rear		25.0364	24.4686	25.6041	3.0044	2.6530	3.4639
Diff (1-2)	Pooled	4.4681	3.2957	5.6405	5.1266	4.7656	5.5473
Diff (1-2)	Satterthwaite	4.4681	3.5138	5.4223			

Descriptive statistics that are relevant for computing a t-statistic, particularly mean and standard deviation

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	334	7.50	<.0001
Satterthwaite	Unequal	333.24	9.21	<.0001

Main t-test table

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	225	109	3.84	<.0001

Equal/homogenous variance test

- Meet: Pooled
- Does not meet: Satterthwaite

The PROC TTEST also provides a few graphical displays, such as histograms, boxplots, and normal QQ plots for the dependent variable by groups. Check out the graphs in SAS.

b. Paired samples

Example data: a company wants to evaluate which type of pizza is more popular. Ten participants were given both types of pizza that weighed exactly 16 oz in a sequence. After fifteen minutes, the remainders of the pizza were weighed. The design of the experiment was counterbalanced.

```
*Paired samples t-test with PROC TTEST*;

data sasdata.pizza;
  input subject A B;
  datalines;
  1  12.9 16.0
  2   5.7  7.5
  3  16.0 16.0
  4  14.3 15.7
  5   2.4 13.2
  6   1.6  5.4
  7  14.6 15.5
  8  10.2 11.3
  9   4.3 15.4
  10  6.6 10.6
  ;
run;

proc ttest data=sasdata.pizza;
  paired A*B;
run;
```

- PAIRED statement: use * to denote different treatment groups.

Results:

The TTEST Procedure

Difference: A - B

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-3.8000	3.9822	1.2593	-11.1000	0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-3.8000	-6.6487	-0.9513	3.9822	2.7391	7.2699

DF	t Value	Pr > |t|
9	-3.02	0.0145

Descriptive statistics for mean difference

t-test results

The output also includes a few graphical displays for mean difference. Check them out in SAS.

6.3 Regression analysis with PROC GLM

a. Linear regression with continuous independent variables.

Let's perform regression analysis using sashelp.class data with PROC GLM. Code:

```
*Regression analysis with PROC GLM*;
proc glm data=sashelp.class;
   model Weight = Height;
run;
```

Results (Open Chapter 5; see p.3 for a comparison with PROC REG):

The GLM Procedure

Number of Observations Read	19
Number of Observations Used	19

Number of valid responses: Same

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			

ANOVA table: same

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

Summary statistics: similar

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

Model effects: unique in PROC GLM

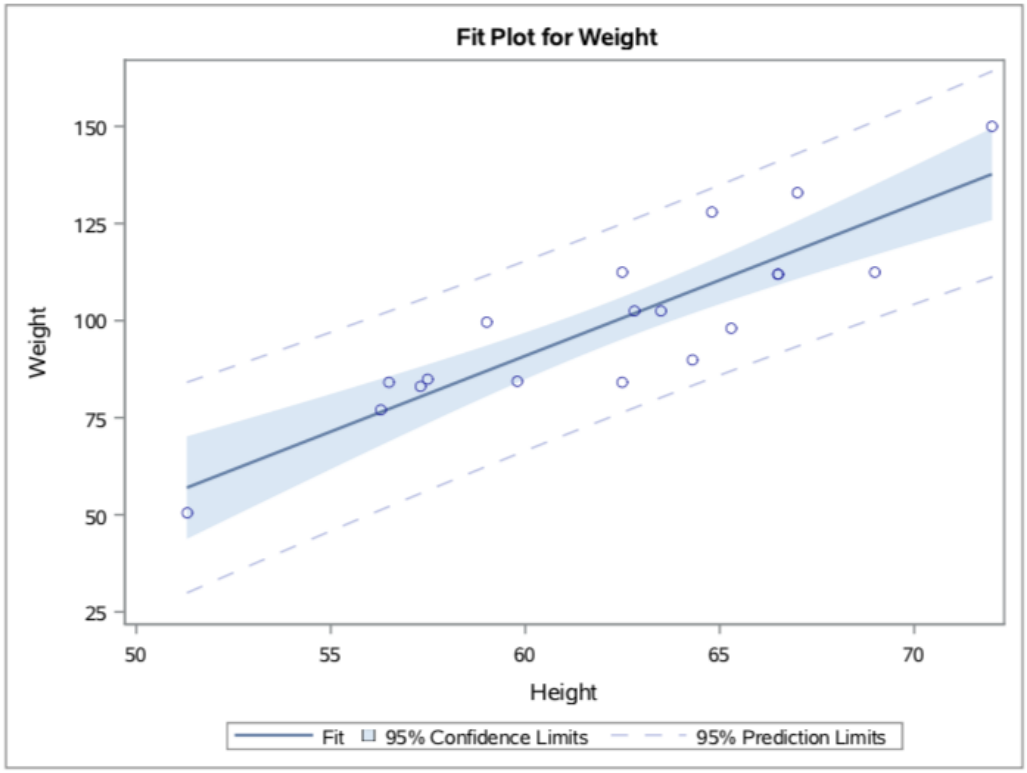
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-143.0269184	32.27459130	-4.43	0.0004
Height	3.8990303	0.51609395	7.55	<.0001

Parameter estimates: same

In terms of graphical displays and model diagnostics, PROC GLM only provides one graph (showing prediction line with prediction limits and confidence limits). PROC REG provides a variety of residual plots for assessing model-data fit.

The GLM Procedure

Dependent Variable: Weight



b. Regression analysis with mixed-types of predictors

The PROC GLM can easily deal with a mixture of categorical and continuous predictors. This is a unique benefit. Let's see an example predicting Weight using Height and Sex (sashelp.class) with equal and varying slopes.

- Equal slopes (of Weight-Height relationship for both females and males).
    - This is equivalent to Analysis of Covariance (ANCOVA).

Code:

```
*Mixed-types of predictors using PROC GLM (Equal slopes)*;
proc glm data=sashelp.class;
  class Sex;
  model Weight = Height Sex;
run;
```

- CLASS statement: identify the categorical variable.
- MODEL statement: specify the model.

Results:

The GLM Procedure

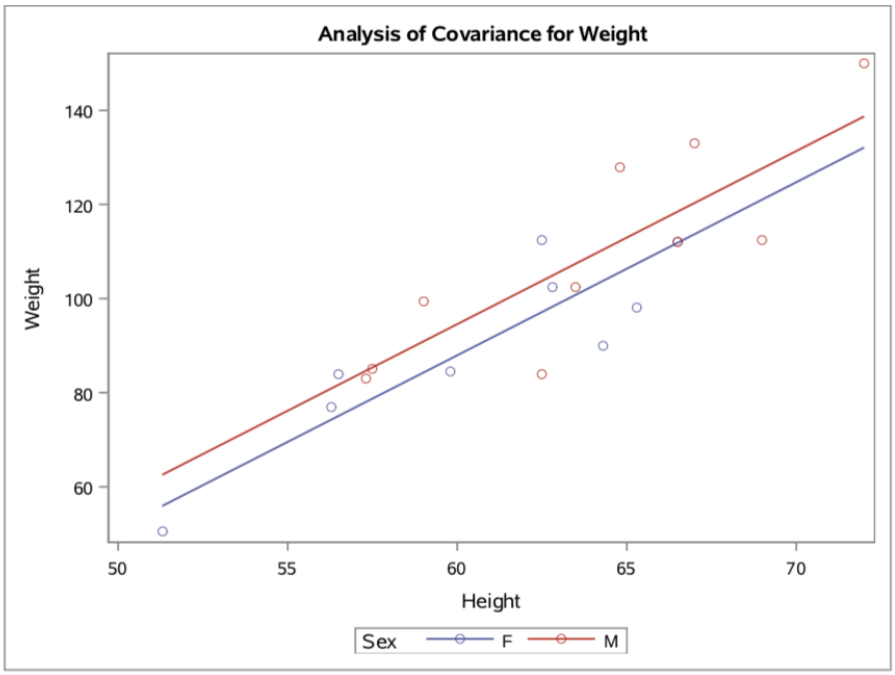
Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7377.963619	3688.981809	30.15	<.0001
Error	16	1957.773223	122.360826		
Corrected Total	18	9335.736842			

R-Square	Coeff Var	Root MSE	Weight Mean
0.790293	11.05877	11.06168	100.0263

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Height	1	7193.249119	7193.249119	58.79	<.0001
Sex	1	184.714500	184.714500	1.51	0.2370

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Height	1	5696.840666	5696.840666	46.56	<.0001
Sex	1	184.714500	184.714500	1.51	0.2370



- Varying slopes (of Weight-Height relationship for females and males)

```
*Mixed-types of predictors using PROC GLM (Varying slopes)*;
proc glm data=sashelp.class;
  class Sex;
  model Weight = Height Sex Height*Sex;
run;
```

- MODEL statement: specify the model including the interaction (using *).

Results:

The GLM Procedure

Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7402.992420	2467.664140	19.15	<.0001
Error	15	1932.744422	128.849628		
Corrected Total	18	9335.736842			

R-Square	Coeff Var	Root MSE	Weight Mean
0.792974	11.34821	11.35120	100.0263

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Height	1	7193.249119	7193.249119	55.83	<.0001
Sex	1	184.714500	184.714500	1.43	0.2498
Height*Sex	1	25.028801	25.028801	0.19	0.6657

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Height	1	5654.260964	5654.260964	43.88	<.0001
Sex	1	15.202417	15.202417	0.12	0.7360
Height*Sex	1	25.028801	25.028801	0.19	0.6657

Dependent Variable: Weight