

## **SAS II: Inferential Statistics**



Updated: August 2012

## Table of Contents

Section 1: Introduction.....	3
1.1 About this Document .....	3
1.2 Prerequisites .....	3
Section 2: Setting Up the Data.....	3
2.1 Introduction to the data .....	3
2.2 Syntax Conventions .....	5
Section 3: Summarizing and Describing Data .....	6
3.1 Descriptive Statistics.....	6
3.3 Frequencies .....	8
3.4 Creating Attractive Output.....	9
3.5 Crosstabulations .....	10
Section 4: Inferential Statistics .....	11
4.1 Chi-Square .....	11
4.2 <i>T</i> -test .....	13
4.3 Correlation .....	16
4.4 Regression.....	17
4.5 General Linear Model .....	20
4.6 Univariate GLM.....	21
4.7 Multivariate GLM.....	24
Conclusion .....	26
References.....	27

## Section 1: Introduction

### 1.1 About this Document

This document is the second module of a two-part tutorial series. It is intended to provide users who have some experience with SAS (*e.g.*, understanding a temporary versus permanent SAS dataset) with the programming tools needed to perform descriptive and inferential statistics in the SAS system. A single data set, *cars\_1993*, is used for all of the examples. If you are not familiar with SAS or need more information on how to get SAS to read your data, consult the first module of this two part tutorial, [SAS I: Getting Started](#).

This tutorial is best thought of as a sequential progression of common tasks involved in analyzing a dataset. In Section 2: Setting Up the Data, you will import the data, and apply formats and labels. In Section 3: Summarizing and describing, you will assess the distributions of the variables and subscales, and compute descriptive statistics for categorical and quantitative variables. Section 4: Inferential Statistics provides examples of inferential statistics such as regression and ANOVA, as well as interpretation of output.

### 1.2 Prerequisites

Knowledge of basic SAS programming such as the data step and procedure step are necessary. Introductions to these topics can be found in the first module, [SAS I: Getting Started](#).

---

## Section 2: Setting Up the Data

### 2.1 Introduction to the data

Throughout this document, a single data set, *cars\_1993*, is used for all of the examples. We will now download four versions of this dataset. First create the following folder on your computer: **C:\SAS-examples**. This (and the previous) tutorial will use the following files:

```
cars_1993.sas7bdat
cars_1993_excel.xls
cars_1993_text.txt
SAS Syntax March 2007.sas
```

The files used in this tutorial are located in a single ZIP file located [HERE](#). Download the file to your desktop and extract. Then, move the four individual files to the **C:\SAS-examples** directory.

SAS provides information about the *cars\_1993* file, which is reproduced below (units of measurement have been added for illustrative purposes):

**Name:** cars\_1993

**Analysis:** descriptive statistics, t-tests, ANOVA, Regression, ANCOVA, data transformation

**Reference:** This represents a subset of the information reported in the 1993 Cars Annual Auto Issue published by Consumer Reports and from Pace New Car and Truck 1993 Buying Guide

**Description:** A random sample of 92 1993 model cars is contained in this data set. The information for each car includes: manufacturer, model, type (small, compact, sporty, midsize, large, or van), price (in thousands of dollars), city mpg, highway mpg, engine size (liters), horsepower, fuel tank size (gallons), weight (pounds), and origin (US or non-US). The data are excellent for doing descriptive statistics by groups or an ANOVA or regression with price as the response variable. Note that violations of the assumptions are probably present and transformation of the response variable is most likely necessary.

Below, a portion of the data set is shown in a SAS viewtable.

	Manufacturer	Model	type	Price	CityMPG	HighwayMPG	EngineSize	Horsepower	FuelTank	Passengers	Weight	Origin
1	Mazda	RX-7	3	32.5	17	25	1.3	255	20	2	2895	non-US
2	Chevrolet	Corvette	3	38	17	25	5.7	300	20	2	3380	US
3	Hyundai	Scoupe	3	10	26	34	1.5	92	11.9	4	2285	non-US
4	Honda	Prelude	3	19.8	24	31	2.3	160	15.9	4	2865	non-US
5	Honda	Accord	2	17.5	24	31	2.2	140	17	4	3040	non-US
6	Honda	Civic	1	12.1	42	46	1.5	102	11.9	4	2350	non-US
7	Geo	Strom	3	12.5	30	36	1.6	90	12.4	4	2475	non-US
8	Ford	Festiva	1	7.4	31	33	1.3	63	10	4	1845	US
9	Dodge	Stealth	3	25.8	18	24	3	300	19.8	4	3805	US
10	Ford	Mustang	3	15.9	22	29	2.3	105	15.4	4	2850	US
11	Geo	Metro	1	8.4	46	50	1	55	10.6	4	1695	non-US
12	Ford	Probe	3	14	24	30	2	115	15.5	4	2710	US
13	Suzuki	Swift	1	8.6	39	43	1.3	70	10.6	4	1965	non-US
14	Subaru	Justy	1	8.4	33	37	1.2	73	9.2	4	2045	non-US
15	Toyota	Celica	3	18.4	25	32	2.2	135	15.9	4	2950	non-US
16	Volkswagen	Corrado	3	23.3	18	25	2.8	178	18.5	4	2810	non-US
17	Volkswagen	Fox	1	9.1	25	33	1.8	81	12.4	4	2240	non-US
18	Pontiac	Firebird	3	17.7	19	28	3.4	160	15.5	4	3240	US
19	Mazda	323	1	8.3	29	37	1.6	82	13.2	4	2325	non-US
20	Lexus	SC300	4	35.2	18	23	3	225	20.6	4	3515	non-US
21	Mercury	Capri	3	14.1	23	26	1.6	100	11.1	4	2450	US
22	Pontiac	LeMans	1	9	31	41	1.6	74	13.2	4	2350	US
23	Plymouth	Laser	3	14.4	23	30	1.8	92	15.9	4	2640	US
24	BMW	535i	4	30	22	30	3.5	208	21.1	4	3640	non-US
25	Chevrolet	Camaro	3	15.1	19	28	3.4	160	15.5	4	3240	US
26	Nissan	Sentra	1	11.8	29	33	1.6	110	13.2	5	2545	non-US


## 2.2 Syntax Conventions

In this tutorial, uppercase letters will be used to indicate SAS keywords that should be entered as shown. Lowercase letters will be used to indicate items supplied by the user, such as data set names and variable names. These conventions are for illustrative purposes only: you can enter them in your program in any mixture of uppercase and lowercase. This tutorial uses the same color coding conventions as the SAS enhanced editor. As a good programming practice, a space will be placed before each terminating semicolon because this facilitates scanning for missing semicolons when troubleshooting program errors.

### SAS Library

If you're going to use an already existing SAS data set or want to create a permanent SAS data set, the first step in most SAS programs involves creating a SAS library to specify the location of the data set (or where you would like to save the raw data as a SAS system file). A SAS library is best thought of as a pointer to a directory or folder on a computer that contains the SAS data set(s). In the example below, the *cars* data set is stored on the C drive of a computer in the directory *SAS-examples*. In order to read these data, we need to create a SAS library and assign it a library name. Here, a library named *project* is created that points to the directory called *SAS-examples* on the C drive.

```
LIBNAME project 'C:\SAS-examples' ;
```

This syntax is typed at the beginning of the Enhanced Program Editor window and is run by highlighting it and clicking on the running man icon located on the toolbar .

For more information on SAS libraries and SAS data sets please see the [SAS I: Getting Started](#) tutorial.

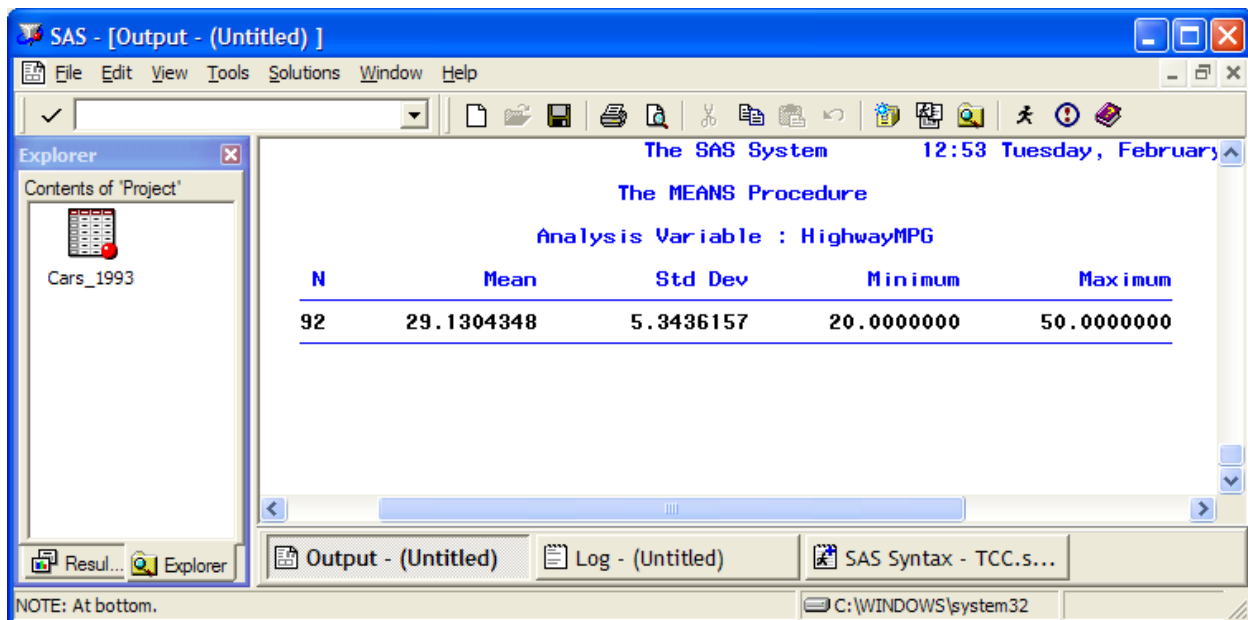
## Section 3: Summarizing and Describing Data

### 3.1 Descriptive Statistics

A common first step in data analysis is to summarize information about variables in your data set, such as the averages and variances of variables. For variables that are measured on a continuous or interval scale, the **MEANS** procedure is often used:

```
* PROC MEANS is used for continuous data;
PROC MEANS DATA = project.cars_1993 ;
  VAR highwaympg ;
RUN ;
```

Running this syntax produces the output below.



The screenshot shows the SAS Output window titled "SAS - [Output - (Untitled)]". The main content area displays the following output:

```
The SAS System      12:53 Tuesday, February 14, 2006
The MEANS Procedure
Analysis Variable : HighwayMPG
```

N	Mean	Std Dev	Minimum	Maximum
92	29.1304348	5.3436157	20.0000000	50.0000000

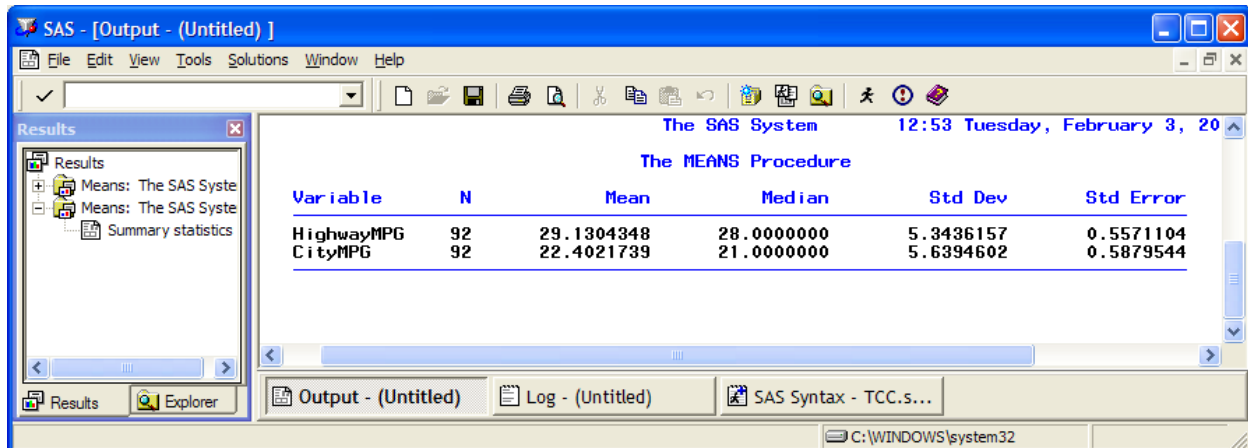
The output window also shows a file explorer on the left with "Cars\_1993" selected, and a taskbar at the bottom with "Output - (Untitled)", "Log - (Untitled)", and "SAS Syntax - TCC.s..." open. A status bar at the bottom indicates "NOTE: At bottom." and "C:\WINDOWS\system32".

This output contains several pieces of information that can be useful to you in understanding the descriptive qualities of your data. The number of cases in the data set is recorded under the column labeled *N*. Information about the range of variables is contained in the *Minimum* and *Maximum* columns. For example, *HighwayMPG* (highway miles per gallon) ranged from 20 MPG to 50 MPG. The average highway MPG is shown in the *Mean* column. Variability can be assessed by examining the values in the *Std. Dev* column. The standard deviation measures the amount of variability in the distribution of a variable. Thus, if individual data points are spread

widely apart, the standard deviation will be large. Conversely, if data points are very similar to each other, the standard deviation will be quite small. The standard deviation describes the standard amount variables differ from the mean. For example, a car capable of 34.47 MPG on the highway is one standard deviation above the mean value of 29.13 MPG on the highway.

You can use the **MEANS** procedure to calculate descriptive statistics for more than one variable at a time by listing additional variables in the **VAR** statement. In addition, you can specify particular descriptive statistics to include in the output by listing these as options in the **PROC** statement. In the example below, the sample size (N), mean, median, standard deviation, and standard error are requested for the *highwaympg* and *citympg* variables.

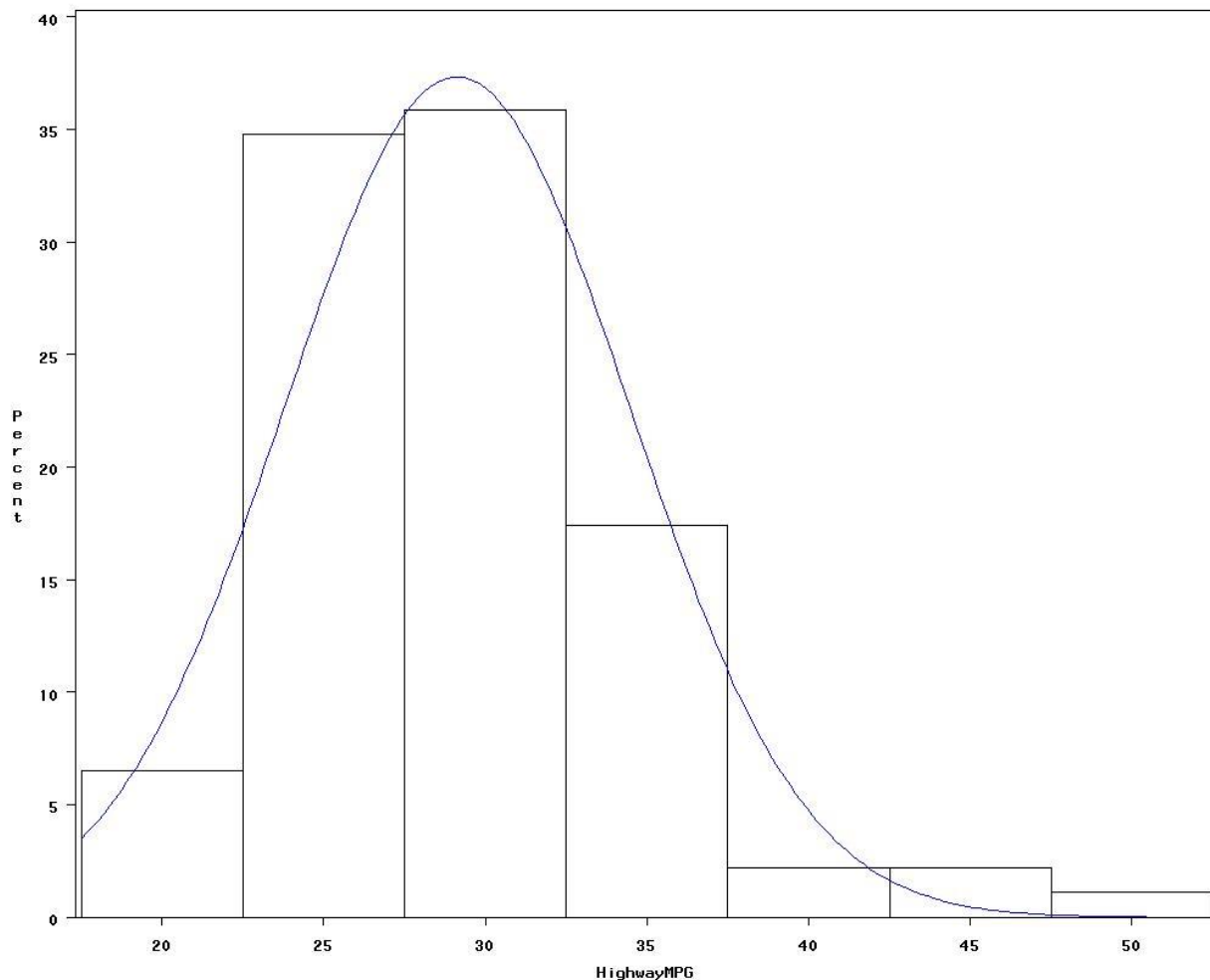
```
* Requesting specific statistics in PROC MEANS ;
PROC MEANS DATA = project.cars_1993 n mean median stddev stderr ;
  VAR highwaympg citympg ;
RUN ;
```



For continuous variables, it is also useful to examine the normality assumption using the **UNIVARIATE** procedure.

```
* Examining normality assumption for a continuous variable ;
PROC UNIVARIATE DATA = project.cars_1993 ;
VAR highwaympg ;
HISTOGRAM highwaympg /normal ;
run;
```

In this example, we request descriptive statistics for *highwaympg*, as well as a histogram and normality tests for the variable. The resulting histogram is shown below:



The variable looks a little skewed, and the normality tests also printed in the output suggest that the variable is significantly skewed. In this tutorial, we will ignore the violation of the normality assumption. In your own research, however, you may wish to consider transforming the data to make it more normal. For example, a square-root or log transformation of the variable may prove helpful.

### 3.3 Frequencies

While the descriptive statistics procedure described above is useful for summarizing data measured on a continuous or interval scale, the **MEANS** and **UNIVARIATE** procedures will not prove helpful for interpreting categorical data. Instead, it is more useful to investigate the numbers of cases that fall into various categories. The **FREQ** or “frequency” procedure allows you to obtain the number of cars in our data set that are of US and non-US origin. As can be seen, 52% of the cars were of US origin.



```
* PROC FREQ is used for categorical data ;
PROC FREQ DATA=project.cars_1993 ;
  TABLES origin ;
RUN ;
```

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
US	48	52.17	48	52.17
non-US	44	47.83	92	100.00

### 3.4 Creating Attractive Output

A useful feature that is introduced in the syntax below is the *Output Delivery System* (ODS). The *Output Delivery System* is a method of delivering output in a variety of formats and of making the formatted output easy to access. For example, to output the frequency table above as an .rtf formatted document, you must first specify the format, e.g., *ODS RTF FILE =*. Second, it is necessary to designate the file directory where you plan to save the document, e.g., “C:\temp\” as well as the name of the specific file you are about to create, e.g., “C:\temp\freq.rtf”. Finally, you must include a **CLOSE** statement at the end of the syntax, e.g., *ODS RTF CLOSE*, as shown below.

```
* Using the Output Delivery System with the FREQ procedure ;
ODS RTF FILE = "c:\sas-examples\freq.rtf" ;
PROC FREQ DATA=project.cars_1993 ;
  TABLES origin ;
RUN ;
ODS RTF CLOSE ;
```

This syntax will create a frequency table that appears as follows:

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
US	48	52.17	48	52.17
non-US	44	47.83	92	100.00

You can create similarly attractive tables in html by selecting:

**Tools**

**Options**

**Preferences**

Click on the Results tab and select *Create HTML*. SAS will now create attractive html tables for all analyses, until such time as you turn off the option. For the rest of the tutorial, ODS-style output tables are shown in the tutorial; you can duplicate these by adding the *ODS RTF FILE* command before any block of syntax and the *ODS RTF CLOSE* command at the end of the block; alternatively, you can turn on the *Create HTML* option.

### 3.5 Crosstabulations

Although one-way frequency tables show you the numbers of cases in each level of a categorical variable, they do not provide information about the relationships between variables. For example, you could create one-way frequency tables to show the number of cars that fall into each category type (compact, sporty, etc.), OR the number of cars of either US or non-US origin; however, a crosstabulation is required to demonstrate the relation between these two variables. For example, crosstabulations can be used to determine the number of cars that are sporty and of US origin or the number of compact cars that are of non-US origin.

To get crosstabulations you need to use the **TABLES** statement to request a two-way table. Create a crosstabulation of *origin* BY *type* by placing an asterisk between the variables of interest (without any spaces around the asterisk).

```
* Crosstabs are used to examine two categorical variables ;
PROC FREQ DATA = project.cars_1993 ;
    TABLES origin*type ;
RUN ;
```

Table of Origin by Type							
Origin	Type						Total
Frequency	Compact	Large	Midsize	Small	Sporty	Van	
Percent							
Row Pct							
Col Pct							
US	7	11	10	7	8	5	48
	7.61	11.96	10.87	7.61	8.70	5.43	52.17
	14.58	22.92	20.83	14.58	16.67	10.42	
	43.75	100.00	47.62	33.33	57.14	55.56	
non-US	9	0	11	14	6	4	44
	9.78	0.00	11.96	15.22	6.52	4.35	47.83
	20.45	0.00	25.00	31.82	13.64	9.09	
	56.25	0.00	52.38	66.67	42.86	44.44	
Total	16	11	21	21	14	9	92
	17.39	11.96	22.83	22.83	15.22	9.78	100.00

The cross tabulation provides several interesting observations about the data. As can be seen in the upper left hand corner of the output, the first row of numbers in each cell are the *observed frequencies* of the cases, or the actual counts. The second row in each cell contains the *percentage* of all possible values of *origin\*type* that are in the particular cell. For example, the cell at the intersection of the column *Large* and the row *US* contains counts of cars that are large and of US origin. These cars represent 11.96% of the cars in the 1993 sample. The next row in each cell presents the *row percentage*, or the percentage of those cars of either US or non-US origin. For example, for non-US origin, 9.09% of cars were vans, while 20.45% were compact cars. The last row in each cell contains the *column percentages*, or the percentage of cars of a particular *type*. For example, of all cars that were classified as “Sporty,” 57.14% were of US origin and the remaining 42.86% were of non-US origin. Likewise, 33.33% of the “Small” cars were of US origin while the remaining 66.67% were of non-US origin. These findings indicate an association between certain types of cars and origin. For example, all “Large” cars are of US origin and the majority of “Small” cars (66.67%) are of non-US origin. The following section will discuss how to further examine this relationship with inferential statistics.

---

## Section 4: Inferential Statistics

### 4.1 Chi-Square

Categorical variables are qualitative variables in which cases are classified in one and only one of the possible levels or groupings. The chi-square test for independence is used in situations where you have two *categorical variables*. To conduct a chi-square analysis, simply add the

**/CHISQ** option to the **TABLES** command in the **FREQUENCY** procedure. Options should be placed after the **TABLES** command following a forward slash (/).

In continuation of the crosstabulation example above, the **NOCOL**, **NOPERCENT**, **EXPECTED**, and **CELLCHI2**, options are added to the syntax as well as the **CHISQ** option. The **NOCOL** and **NOPERCENT** options simply request that the column percentages and total percentages be omitted from the output table. By reducing the amount of output in each cell, the table becomes less cluttered and facilitates interpretation. The **EXPECTED** option requests that the expected cell frequency for each cell be provided on the output table (the number of cars expected in each cell if there were no association between car type and origin). A rule of thumb is that you should have no expected counts of zero, and less than 20% of expected counts less than 5. If you violate these guidelines, it may be necessary to collapse some of your categories in order to get higher expected counts. The **CELLCHI2** option reports each cell's contribution to the total Pearson chi-square statistic. Recall that the Chi-square statistic is calculated as

$$\sum ((\text{observed} - \text{expected})^2 / \text{expected})$$

Therefore, the larger the **CELLCHI2** value, the more it contributes to the overall Chi-square value. If the Chi-square statistic indicates a significant association between two variables, then the **CELLCHI2** values can be used to infer which cells are contributing most to that association.

```
* Chi-square for origin by rectype ;
PROC FREQ DATA = project.cars_1993 ;
  TABLES origin*type / CHISQ EXPECTED CELLCHI2 NOPERCENT NOCOL ;
RUN ;
```

As can be seen in the output below, the chi-square is 13.88, with a p-value of 0.0164. Because the p-value is less than 0.05, the chi-square is statistically significant. Scanning the crosstabulation table reveals that the two large car cells contribute most to the overall Chi-square since their individual cell Chi-square values are 4.82 and 5.26 for cars of US and non-US origin respectively. Therefore, these two cells contribute 10.08 ( $4.82 + 5.26 = 10.08$ ) to the overall Chi-square value of 13.88. In addition, the two cells representing small cars contribute 1.43 and 1.56 for cars of US and non-US origin respectively, thus contributing 2.99 to the overall Chi-square value of 13.88. Together, these 4 cells contribute 13.07 to the overall Chi-square value of 13.88 and can be used to infer the association between type of car and origin.

To understand this association, compare observed and expected cell frequencies. For example, the observed number of large cars of US origin was 11 but the expected count was only 5.7. In other words, more large cars are of US origin than was expected by chance. Conversely, there are fewer large cars (0.0) of non-US origin than expected by chance (5.3). The opposite pattern is seen for small cars. There are fewer small cars of US origin (7) than expected by chance (11.0) and more small cars of non-US origin (14) than expected by chance (10.0). Therefore, this pattern indicates that US car manufacturers tended to produce larger cars, while foreign manufacturers tended to produce smaller cars in 1993.

Table of Origin by Type							
Origin	Type						Total
Frequency Expected Cell Chi-Square Row Pct	Compact	Large	Midsize	Small	Sporty	Van	
US	7	11	10	7	8	5	48
	8.3478	5.7391	10.957	10.957	7.3043	4.6957	
	0.2176	4.8225	0.0835	1.4287	0.0663	0.0197	
	14.58	22.92	20.83	14.58	16.67	10.42	
non-US	9	0	11	14	6	4	44
	7.6522	5.2609	10.043	10.043	6.6957	4.3043	
	0.2374	5.2609	0.0911	1.5586	0.0723	0.0215	
	20.45	0.00	25.00	31.82	13.64	9.09	
<b>Total</b>	16	11	21	21	14	9	92

Statistic	DF	Value	Prob
Chi-Square	5	13.8801	0.0164
Likelihood Ratio Chi-Square	5	18.1502	0.0028
Mantel-Haenszel Chi-Square	1	6.3187	0.0119
Phi Coefficient		0.3884	
Contingency Coefficient		0.3621	
Cramer's V		0.3884	

## 4.2 *T*-test

The *t*-test is a useful technique for comparing the mean values of two sets of numbers. The comparison will provide you with a statistic for evaluating whether the difference between two means is statistically significant. *T*-tests can be used either to compare two independent groups (independent-samples *t*-test) or to compare observations from two measurement occasions for the same group (paired comparison *t*-test). To conduct a *t*-test, two sets of numbers should both be drawn from continuous, normally distributed populations. If you have more than two groups, or more than two variables in a single group that you want to compare, you should use General Linear Model (proc GLM) in SAS. The GLM procedure will be covered in detail later in this document.

If you are using the *t*-test to compare two groups, the groups should be randomly drawn from normally distributed and independent populations. For example, if you were comparing the average *highwaympg* between cars of US versus non-US origin, the two independent populations

are: (1) US cars, and (2) non-US cars. This demarcation of cars results in two non-overlapping groups. For the independent samples  $t$ -test, the TTEST procedure is used, as illustrated below. The CLASS statement contains the variable that distinguishes the groups being compared, while the VAR statement identifies the dependent variable.

```
* Independent-samples t-test comparing origin in terms of highwaympg ;
PROC TTEST DATA = project.cars_1993 ;
  CLASS origin ;
  VAR highwaympg ;
RUN ;
```

In the output, the summary statistics and confidence intervals are displayed first. For the variable *highwaympg*, the mean of the US group (28.1) minus the mean of the non-US group (30.2) equals an average difference of -2.1. Also included is a *confidence interval*, a range of values that has a 95% probability of containing the parameter being estimated (in this case, the mean). The confidence interval is often abbreviated as “CI,” and the upper and lower limits of the confidence interval are often abbreviated as “lower CL” and “upper CL.”

Variable	Origin	N	Lower CL Mean	Mean	Upper CL Mean	Std Dev	Minimum	Maximum
HighwayMPG	US	48	26.94	28.146	29.351	4.1513	20	41
HighwayMPG	non-US	44	28.298	30.205	32.111	6.2713	21	50
HighwayMPG	Diff (1-2)		-4.245	-2.059	0.1271	5.2717		

For the difference score, the confidence interval range is between -4.2 and 0.1 (commonly expressed as:  $-4.2 \leq \mu \leq 0.1$ ). That is, the interval between -4.2 and 0.1 has a 95% probability of containing the true difference score between the two groups. Only confidence intervals that do not contain 0 will also have significant probability tests. If a confidence interval for a mean difference includes 0, the data are consistent with a population mean difference of 0. If the difference is 0, the population means are equal. If the confidence interval for a difference excludes 0, the data are not consistent with equal population means. Therefore, one of the first things to look at is whether a confidence interval for a difference contains 0. If 0 is not in the interval, a difference has been established. If a CI contains 0, then a difference has not been established.

Next are the tests for equal group means and equal variances. A group test statistic for the equality of means is reported for equal and unequal variances. To assess whether to refer to the statistics for equal or unequal variances, look at the test for equality of variances; this test ( $F(43,47) = 2.28, p = 0.006$ ) does indicate a significant difference in the two variances; that is, the variances are not equal, so the Satterthwaite  $t$  statistic for unequal variance should be used to assess mean differences. Based on the Satterthwaite statistic, cars of US and non-US origin do not differ in highway mpg [ $t(73.6) = -1.84, p = 0.07$ ].

T-Tests					
Variable	Method	Variances	DF	t Value	Pr >  t
HighwayMPG	Pooled	Equal	90	-1.87	0.0646
HighwayMPG	Satterthwaite	Unequal	73.6	-1.84	0.0699

Equality of Variances					
Variable	Method	Num DF	Den DF	F Value	Pr > F
HighwayMPG	Folded F	43	47	2.28	0.0063

When it is not feasible to assume that two groups of data are independent, or you would like to see if two continuous variables differ, we instead use the paired  $t$ -test, an analysis that takes into account the correlation of the variables. The primary advantage of this method is that the utilization of the positive or negative linear correlation will result in higher power to detect existing differences between the means. In the paired  $t$ -test, the differences between paired observations are assumed to be normally distributed. The most common usage of the paired  $t$ -test is for pre- and post-test assessment for efficacy of intervention. For the *cars\_1993* data set, we will explore whether there is a significant difference between highway mpg and city mpg. The **PAIRED** command in the **TTEST** procedure with an asterisk (*i.e.*, **\***) between variables will produce a paired  $t$ -test, such as in the example below.

```
* Paired-samples t-test comparing highway and city mpg ;
PROC MEANS DATA = project.cars_1993 ;
  VAR highwaympg*citympg ;
RUN ;

PROC TTEST DATA = project.cars_1993 ;
  PAIRED highwaympg*citympg ;
RUN ;
```

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
HighwayMPG	HighwayMPG	92	29.1304348	5.3436157	20.0000000	50.0000000
CityMPG	CityMPG	92	22.4021739	5.6394602	15.0000000	46.0000000

Difference	N	Lower CL Mean	Mean	Upper CL Mean	Std Dev	Minimum	Maximum
HighwayMPG - CityMPG	92	6.3422	6.7283	7.1143	1.8641	2	11

T-Tests			
Difference	DF	t Value	Pr >  t
HighwayMPG - CityMPG	91	34.62	<.0001

In this sample of 1993 cars, highway mpg is significantly higher than city mpg ( $t = 34.62$ ,  $df = 91$ ,  $p < 0.001$ ).

### 4.3 Correlation

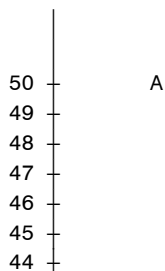
Correlation is one of the most common forms of data analysis, for several reasons: it can provide an analysis that stands on its own, it underlies many other analyses, and it can be a good way to support conclusions after primary analyses have been completed. Correlation measures the linear relationship between two variables. A correlation coefficient has a value ranging from -1 to 1. Values that are closer to the absolute value of 1 indicate that there is a strong relationship between the variables being correlated, whereas values closer to 0 indicate that there is little or no linear relationship. The sign of a correlation coefficient describes the type of relationship between the variables being correlated. A positive correlation coefficient indicates that there is a positive linear relationship between the variables: as one variable increases in value, so does the other. A negative coefficient denotes an inverse relationship. Thus, if it is hypothesized that heavier cars will also have lower higher highway mpg, then one might expect a negative correlation (*i.e.*, as weight increases, mpg decreases).

Prior to conducting a correlation analysis, it is advisable to plot the two variables to visually inspect the relationship between them. The syntax below demonstrates a simple method for plotting two variables with SAS.

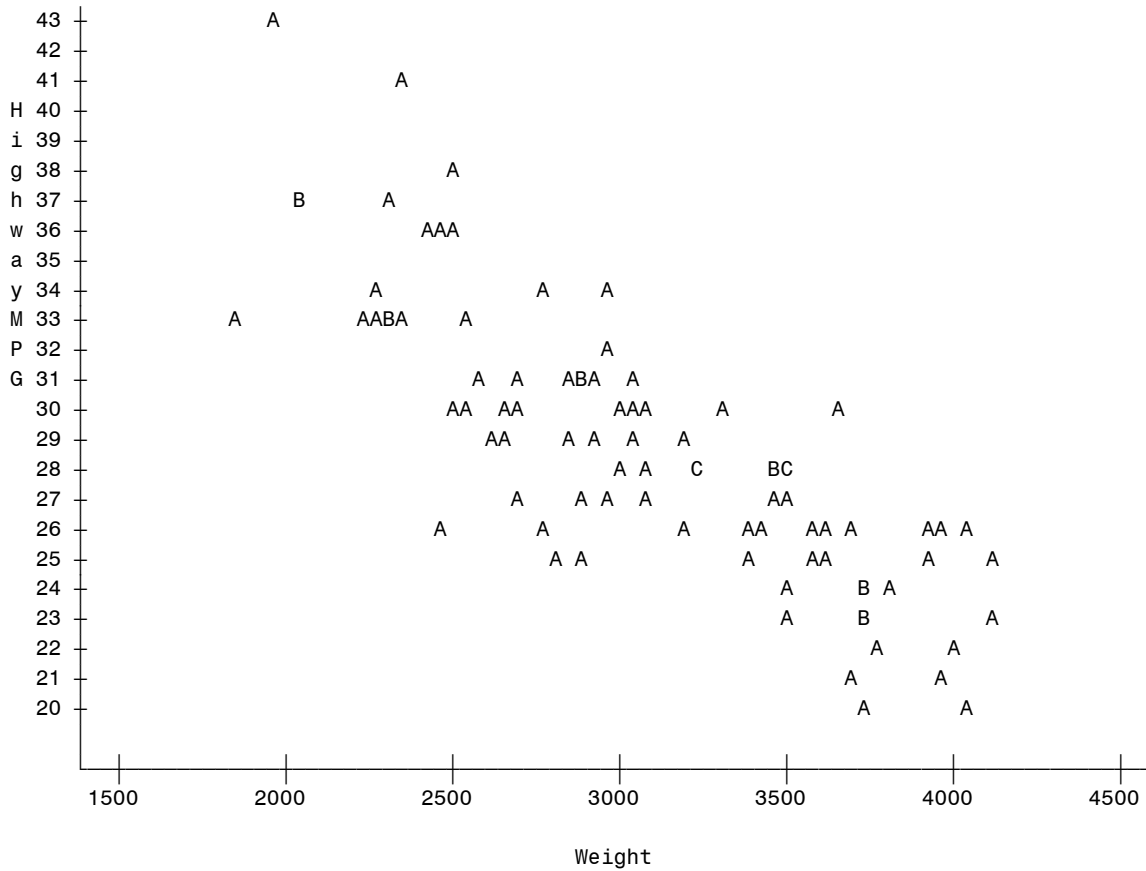
```
* Creating a scatterplot of highwaympg and weight ;
PROC PLOT DATA = project.cars_1993 ;
  PLOT highwaympg*weight ;
RUN ;
```

An asterisk should be placed between the two variables that will be plotted. The name of the variable to be plotted on the Y-axis should be placed first (before the asterisk). In this example, *highwaympg* is plotted on the Y-axis, whereas *weight* is plotted along the X.

Plot of HighwayMPG\*Weight. Legend: A = 1 obs, B = 2 obs, etc.







The plot indicates a strong negative relationship between weight and highway mpg.

To obtain a correlation in SAS, adapt the following syntax:

```
* Correlate highwaympg and weight ;
PROC CORR DATA = project.cars_1993 ;
  VAR highwaympg weight ;
RUN ;
```

Pearson Correlation Coefficients, N = 92 Prob >  r  under H0: Rho=0		
	HighwayMPG	Weight
HighwayMPG	1.00000	-0.80943 <.0001
Weight	-0.80943 <.0001	1.00000

A strong negative correlation is found between highway MPG and automobile weight ( $r = -0.81$ ,  $N = 92$ ,  $p < .0001$ ). The output suggests that highway MPG and automobile weight are not independent phenomena.

### 4.4 Regression

Regression is a technique that can be used to investigate the effect of one or more predictor variables on an outcome variable. Regression allows you to make statements about how well one or more *predictors* (or independent variables) will predict the value of a *criterion variable* (or dependent variable). For example, if you were interested in investigating which variables in the cars database were good predictors of highway MPG, you could create a regression equation to explore whether weight, horsepower, and engine size help to predict highway MPG. Here, the hypothesis is that lighter cars with smaller engines and less horsepower will have better (higher) highway MPG.

To conduct a regression analysis in SAS, use the following syntax. After the keyword **MODEL**, the dependent or outcome variable is specified, followed by an equal sign and then the predictors, or independent variables. In SAS, variables specified in the **MODEL** statement must be numeric variables in the data set being analyzed.

```

/*REGRESSION OF HIGHWAYMPG ON WEIGHT HORSEPOWER AND ENGINESIZE */
PROC REG DATA = project.cars_1993 ;
    MODEL highwaympg = weight horsepower enginesize / STB ;
RUN ;

```

Note: specifying the option **STB** in the **MODEL** statement tells SAS to print the *Standardized Betas*. This statistic will be explained below.

The first table in the output is an ANOVA table that indicates whether the overall model is a good fit. More specifically, it describes whether or not the overall variance predicted by this multiple regression model is greater than would be expected by chance. You could conceptualize the meaning of the *F* test in two different ways. First, the *F* statistic represents a test of the null hypothesis that the expected values of the regression coefficients are all equal to zero. Second, the *F* statistic also indicates whether the *R-square*, or proportion of variance in the dependent variable accounted for by the predictors, is significantly different from zero. If the null hypothesis were true, then there would be no relationship between the dependent variable and the predictor variables. In this case, however, the null hypothesis is rejected, as indicated by a large *F* value ( $F = 59.89$ ) and a small significance level ( $p < .0001$ ). Thus, the three predictor variables in the present example could be used to predict the dependent variable, highway MPG.

The *R-Square* indicates the amount of variance explained by a given set of predictor variables. In this example, the value is 0.67, which indicates that 67% of the variance in the dependent variable is explained by the independent variables in the model. Since the size of R-square is related to the number of predictors in the equation, it has been suggested that a modified statistic that adjusts for the number of predictors in the model is more accurate than the R-square. The *adjusted R square* adjusts R-square by dividing each sum of square by its degrees of freedom (Neter, Kutner, Nachtsheim, & Wasserman, 1996). In this example, the value of R-square was 0.67, while the value of Adjusted R-square was 0.66.

The REG Procedure  
Model: MODEL1  
Dependent Variable: HighwayMPG

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1744.12044	581.37348	59.89	<.0001
Error	88	854.31434	9.70812		
Corrected Total	91	2598.43478			

Root MSE 3.11579 R-Square 0.6712  
Dependent Mean 29.13043 Adj R-Sq 0.6600  
Coeff Var 10.69598

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Standardized Estimate
Intercept	1	53.36223	2.06370	25.86	<.0001	0
Weight	1	-0.00850	0.00109	-7.79	<.0001	-0.94007
Horsepower	1	-0.01030	0.00975	-1.06	0.2937	-0.10037
EngineSize	1	1.24301	0.61527	2.02	0.0464	0.24228

The final section of the output, the *Parameter Estimates* table, provides information about the effects of the individual predictor variables. Generally, two types of information are presented: coefficients and significance tests. The coefficients indicate the increase in the value of the dependent variable for each unit increase in the predictor variable. For example, the unstandardized coefficient for *weight*, -0.0085, indicates that as a car increases in one unit of weight, its predicted highway MPG will decrease by 0.0085 units. A well-known problem with the interpretation of unstandardized coefficients is that their values are dependent on the scale of the variable for which they were calculated, which makes it difficult to assess the relative influence of independent variables through a comparison of unstandardized coefficients. For example, the unstandardized coefficient for engine size (1.2430) is much greater than that for weight (-0.0085). However, the scales differ greatly since car weight is measured in pounds and engine size is measured in liters. In order to compare the relative influence of each variable, we turn to the standardized coefficients, or *Beta coefficients*. Beta coefficients are based on data expressed in a standardized, or *z*-score form. Thus, all variables have a mean of zero and a standard deviation of one and are expressed in the same scale. Examining the Beta coefficients for *weight* and *enginesize*, we can see that *weight* is about 4 times as influential as *enginesize* in terms of relative predictive power.

In addition to the coefficients, the table also provides a significance test for each of the independent variables in the model. The significance test evaluates the null hypothesis that the unstandardized regression coefficient for the predictor is zero when all other predictors' coefficients are fixed to zero. This test is presented as a *t* statistic. For example, examining the *t* statistic for the variable *weight*, you can see that it is associated with a significance value of <0.0001, indicating that the null hypothesis, that states that this variable's regression coefficient

is zero when all other predictor coefficients are fixed to zero, can be rejected. That is, *weight* is a significant predictor of *highwaympg*. Likewise, *enginesize* is also a significant predictor of *highwaympg* ( $t = 2.02$ ,  $df = 1, 88$ ,  $p = 0.046$ ) but horsepower is not ( $t = -1.06$ ,  $df = 1, 88$ ,  $p = 0.294$ ).

## 4.5 General Linear Model

Analysis of variance (ANOVA) is available in the **GLM** procedure. Analysis of variance can be used in many situations to determine whether there are differences between groups on the basis of one or more outcome variables, or if a continuous variable is a good predictor of one or more dependent variables. There are three general varieties of the GLM: univariate models, multivariate models, and repeated measures models. Use the *univariate general linear model* when you have only a single dependent variable, but have one or more independent variables (which may be fixed between-subjects factors, random between-subjects factors, or covariates). Use the *multivariate general linear model* when you have more than one dependent variable, and independent variables are either between-subjects factors or covariates. Use the *repeated measures general linear model* when you have more than one measurement occasion for a dependent variable and have fixed between-subjects factors or covariates as independent variables. Because it is beyond the scope of this document to cover all three varieties of the general linear model in detail, we will focus on the univariate version of the general linear model, with some attention given to topics that are unique to the multivariate general linear model. The features of the univariate general linear model given here are also useful for understanding multivariate models.

In our example ANOVA models, we would like to use the between-subjects factors *origin* and *type*. However, as we saw earlier in our crosstabulation of origin and type, some of the cells had very low numbers of subjects; in particular, there were zero observations in the cell for large non-US cars. This poses a problem for ANOVA; a rule of thumb is that you should have a minimum of 10 observations per cell for this type of analysis. Accordingly, we need to collapse *type* into a smaller set of categories.

```
* Recode type into two categories based on size ;
DATA project.cars_1993 ;
  SET project.cars_1993 ;
  IF type = '.' THEN size = . ;
  ELSE IF type = 'Small' OR type = 'Compact'
    OR type = 'Sporty' THEN size = 1 ;
  ELSE IF type = 'Midsize' OR type = 'Large'
    OR type = 'Van' THEN size = 2 ;
RUN ;
```

It's also useful to create a format for newly recoded variables, so that you can remember what they mean later (for more information about formats, see SAS I: Getting Started Tutorial).

```
* Define value labels for the new variable ;
PROC FORMAT ;
  VALUE fsize          1 = 'Small/Compact/Sporty'
                     2 = 'Midsize/Large/Van';
```

**RUN ;**

Now we are ready to run some ANOVAs.

## 4.6 Univariate GLM

The univariate general linear model is used to compare differences between group means (ANOVA) and/or to estimate the effect of covariates on a single dependent variable (ANCOVA). We begin with a simple 2x2 ANOVA, including *origin* and *size* as between-subjects factors.

```
* A two-way ANOVA with origin and size as factors ;
PROC GLM DATA = project.cars_1993 ;
  CLASS origin size ;
  MODEL highwaympg = origin size origin*size ;
  LSMEANS origin size origin*size ;
  FORMAT size fsize. ;
RUN ;
```

The **CLASS** statement tells SAS that *origin* and *size* are categorical variables. Make sure to include *only* categorical variables in the **CLASS** statement. If you were to include a quantitative variable (such as IQ) in the **CLASS** statement, SAS would create a category for each unique value found in that column. As you might imagine, this analysis would be unwieldy and difficult to interpret, as your factor could have over a hundred levels. On the next line of syntax, we find the **MODEL** statement, which specifies the independent and dependent variables in the model. On the next line, the **LSMEANS** statement requests the *least squared means* for the class (*i.e.*, categorical) variables. Variables that are not in the **CLASS** statement are not categorical and cannot be included in the **LSMEANS** statement. In our analysis, there are only two levels for each factor, so there is no need for any follow-up pairwise comparisons. However, if you had three levels or more in a factor, you would also find it useful to specify the **PDIF** option in the **LSMEANS** statement.

*Omnibus F-test table*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	954.754765	318.251588	17.04	<.0001
Error	88	1643.680018	18.678182		
Corrected Total	91	2598.434783			

*Type I Sum of Squares F-test table*

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Origin	1	97.2965250	97.2965250	5.21	0.0249
size	1	785.4091705	785.4091705	42.05	<.0001
Origin*size	1	72.0490694	72.0490694	3.86	0.0527

*Type III Sum of Squares F-test table*

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origin	1	7.9180756	7.9180756	0.42	0.5167
size	1	823.3341632	823.3341632	44.08	<.0001
Origin*size	1	72.0490694	72.0490694	3.86	0.0527

Output from ANOVAs or ANCOVAs that contain interaction terms are commonly interpreted in the following manner:

- Omnibus  $F$  Test
- Interaction
- Main Effects

Similar to the multiple regression example, the Omnibus  $F$  Test tests the null hypothesis,  $H_0$ :  $Full\ Model = 0$ . In other words, the null hypothesis should be rejected before an analyst can interpret the individual parameters, as failure to reject the omnibus null means that all of the independent variables are no better than *zero* independent variables.

The main effects are shown in the *Type III Sums of Squares F table*. By default, SAS prints the  $F$  tables for both the Type I and Type III sums of squares. Type III sums of squares are the *partial sums of squares*, or the variance that is unique to each independent variable above and beyond the other independent variables in the equation. Type I sums of squares, on the other hand, assess the value of adding each independent variable to a model in a stepwise or sequential manner. These sums of squares are computed by sequentially building up the model. For example, in ANOVA, the sums of squares for each variable depend on the particular order the variables are sequentially placed into the model. (In SAS, the order will depend upon the left-to-right specification of the **MODEL** statement).

Researchers use Type I sums of squares when they are interested in the sequential contribution of each variable. The order a variable is placed in a Type I sums of squares model is important in that the first variable has the first crack at all the potential variance in the dependent variables. Subsequent variables, then, receive a dependent variable that is the residual of the first analysis. In other words, the second IV predicts a dependent variable that contains no variance attributable to the first dependent variable. In Type III sums of squares, all IVs are entered simultaneously. Any variance in the DV or IVs that is common between two or more IVs is excluded. Type III is more commonly-used analysis than Type I sums of squares, and so here we interpret the Type III results. The output shows that *size* ( $F_{1,88} = 44.08, p < 0.001$ ) is statistically significant but *origin* ( $F_{1,88} = 0.42, p = 0.52$ ) is not. The interaction between the two is not quite significant ( $F_{1,88} = 3.86, p = 0.05$ ). Some researchers would report this as marginally significant ( $p < 0.10$ ), while others would report it as non-significant; this practice differs depending on the field and the particular audience for the research.

Perhaps you wanted to expand this analysis to include car weight as a *covariate*. A covariate is a quantitative independent variable, often entered in models to reduce error variance. By removing the effects of the relationship between the covariate and the dependent variable, you can often

get a better estimate of the amount of variance that is being accounted for by the factors in the model. Covariates can also be used to measure the linear association between the covariate and a dependent variable, as is done in regression models. In this situation, a linear relationship indicates that the dependent variable increases or decreases in value as the covariate increases or decreases in value. For example, you may want to determine whether highway MPG varies with car origin and size while controlling for car weight. Car weight is included as a covariate since you know that it is negatively correlated with highway MPG from your previous analyses.

If you intend to conduct an analysis of covariance, you should test for interactions between covariates and factors. Doing so will determine whether you have met the *homogeneity of regression slopes* assumption, which states that the regression slopes for all groups in your analysis are equal. This assumption is important because the means for each group are adjusted by averaging the slopes for each group so that group differences in the covariate are removed from the dependent variable. Thus, it is assumed that the relationship between the covariate and the dependent variable is the same at all levels of the independent variables. In our example, we would meet this assumption if the “US” and “non-US” cars demonstrated the same slope for the relationship between highway MPG and car weight, and the two different size groups also had the same slope between highway MPG and car weight. Therefore, to run an ANCOVA you need to run two analyses. One analysis is the actual ANCOVA model and does not include interaction terms between the covariate and the independent variables. The other analysis tests the *homogeneity of slopes* assumption by including interaction terms. If the interactions are statistically significant, then the homogeneity of slopes assumption has been violated.

```
* First, we have to test the homogeneity of regression slopes ;
PROC GLM DATA = project.cars_1993 ;
  CLASS origin size ;
  MODEL highwaympg = origin size weight origin*weight size*weight ;

RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origin	1	63.4584704	63.4584704	6.93	0.0100
size	1	14.4984311	14.4984311	1.58	0.2117
Weight	1	656.9018999	656.9018999	71.75	<.0001
Weight*Origin	1	64.4483802	64.4483802	7.04	0.0095
Weight*size	1	19.1706694	19.1706694	2.09	0.1515

The Type III table for this analysis shows that the interaction between *weight* and *size* is not statistically significant, but the interaction between *weight* and *origin* is. Thus, the homogeneity of slopes assumption has been violated and a standard ANCOVA analysis is not appropriate. For the sake of illustration, however, we will proceed with the standard ANCOVA, which contains all possible interactions between the categorical variables, but no interactions with the covariate.

```
* The standard ANCOVA ;
PROC GLM DATA = project.cars_1993 ;
```

```

CLASS origin size ;
MODEL highwaympg = origin size origin*size weight ;
  LSMEANS origin size origin*size ;
  FORMAT size fsize. ;

```

**RUN;**

Source	DF	Type III SS	Mean Square	F Value	Pr > F
<b>Origin</b>	1	0.0193008	0.0193008	0.00	0.9645
<b>size</b>	1	7.2137155	7.2137155	0.75	0.3901
<b>Origin*size</b>	1	42.5862620	42.5862620	4.40	0.0388
<b>Weight</b>	1	802.4086265	802.4086265	82.98	<.0001

With the inclusion of *weight* as a covariate, *size* is no longer significant, but the *origin\*size* interaction (only marginally significant before) is now significant at  $p < 0.05$ . To interpret the meaning of this interaction, look at the LSMEANS table for *origin\*size*:

Origin	size	HighwayMPG LSMEAN
US	Midsize/Large/Van	30.2006607
US	Small/Compact/Sporty	27.9097456
non-US	Midsize/Large/Van	28.7636262
non-US	Small/Compact/Sporty	29.2866906

If you were to plot these means using a program like Excel, you could see that the effect of size seems a little stronger among U.S. cars, with a difference of more than 2 mpg (27.9 vs. 30.2) between the two size categories. In addition, once the impact of weight is controlled, larger U.S. cars seem more efficient on the highway than smaller cars. In contrast, among non-U.S. cars, the effect of size is smaller, with a difference of less than 1 mpg between the two size categories, and the effect of size is in a different direction: smaller cars are more efficient on the highway than larger cars.

Of course, the results of this analysis may not be trustworthy, since we know that the homogeneity of regression slopes assumption has been violated. How else might you analyze the data? Several options are available to you; perhaps the easiest option to implement and interpret would be to divide weight into classes and use that new categorical variable as an additional between-subjects factor in the analysis.

## 4.7 Multivariate GLM

The multivariate version of the general linear model (aka multivariate analysis of variance – MANOVA) has many similarities to the univariate model described above. However, the key



difference between models is that MANOVA allows multiple dependent variables, whereas the univariate model only permits a single dependent variable. As an example, you could expand your original 2x2 ANOVA to include two dependent variables, city mpg and highway mpg.

To run a MANOVA you must include a **MANOVA** statement in your PROC GLM syntax. The **H=** in the **MANOVA** statement designates the variables to include in the MANOVA. In this example, we use the keyword **\_ALL\_**, which tells SAS to include all of the dependent variables in the **MODEL** statement.

```
* A MANOVA with highway and city mileage-per-gallon ;
PROC GLM DATA = project.cars_1993 ;
  CLASS origin size ;
  MODEL highwaympg citympg = origin size origin*size ;
  MANOVA H = _ALL_ ;
  LSMEANS origin size origin*size ;
  FORMAT size fsize. ;
```

```
RUN ;
```

The null hypothesis tested in a univariate model is  $H_0: \mu_1 = \mu_2 = \mu_3$  (population means are equal) while the multivariate null hypothesis is:  $H_0: \mu_1 = \mu_2 = \mu_3$  (population mean vectors are equal). That is, failure to reject the multivariate null hypothesis results in the conclusion that all mean vectors of the dependent variables are equal (Stevens, 1996). In contrast, rejection of the null hypothesis indicates that the linear relationships between the dependent variables are different for each group. Like the ANOVA, the order of interpretation for MANOVA is:

- Multivariate Omnibus Null
- Interactions
- Main Effects

Scroll down to the bottom of the output to find the multivariate tables, titled *MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall [name] Effect*. These tables contain multiple statistics such as Wilks' Lambda, Pillai's Trace, the Hotelling-Lawley Trace, and Roy's largest root. These statistics assess whether or not group effects can be discerned across the outcome variables by the given IV. For example, the table below shows the MANOVA output for *size*. Although four test statistics are provided, the most widely known and used is Wilks' Lambda. Each of these test statistics can be converted to an approximate *F* value and these are shown on the output also. Thus, the multivariate *F* statistics are not actually *F* ratios, but are approximate values that have corresponding critical values. As can be seen in the output below, the *F* values for the overall *size* effect are significant for all four multivariate test statistics.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall size Effect					
H = Type III SSCP Matrix for size					
E = Error SSCP Matrix					
S=1 M=0 N=42.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.65404222	23.01	2	87	<.0001
Pillai's Trace	0.34595778	23.01	2	87	<.0001
Hotelling-Lawley Trace	0.52895328	23.01	2	87	<.0001
Roy's Greatest Root	0.52895328	23.01	2	87	<.0001

The output also includes univariate ANOVAs for each dependent variable. The following Type III tables are pulled from the Highway MPG and City MPG sections of output:

**Dependent Variable: HighwayMPG**

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origin	1	7.9180756	7.9180756	0.42	0.5167
size	1	823.3341632	823.3341632	44.08	<.0001
Origin*size	1	72.0490694	72.0490694	3.86	0.0527

**Dependent Variable: CityMPG**

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origin	1	53.1984529	53.1984529	2.65	0.1073
size	1	906.1549331	906.1549331	45.08	<.0001
Origin*size	1	35.3630637	35.3630637	1.76	0.1881

The factor *size* was significant for each of the dependent variables.

## Conclusion

In this course you learned how to:

- Manage data
- Perform elementary descriptive statistics
- Perform basic inferential statistics
- Interpret basic output

For more information on the analytic procedures introduced in this document, or to learn about the other procedures available in SAS/STAT, consult the SAS Online Documentation, available from the SAS support website at [support.sas.com](http://support.sas.com).

## References

Stevens, J. (1996). *Applied Multivariate Statistics for the Social Sciences (3<sup>rd</sup> ed.)*. New Jersey: LEA.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics (3<sup>rd</sup> ed.)*. New York: Harper Collins.

Winer, B. J. (1971), *Statistical Principles in Experimental Design, (2<sup>nd</sup> ed.)*. New York: McGraw-Hill Book Co.