

# **Statistics I: Introduction to ANOVA, Regression, and Logistic Regression**

Course Notes

*Statistics I: Introduction to ANOVA, Regression, and Logistic Regression Course Notes* was developed by Melinda Thielbar, Mike Patetta, and Paul Marovich. Additional contributions were made by John Amrhein, Marc Huber, Dan Kelly, Bob Lucas, Jill Tao, and Catherine Truxillo. Editing and production support was provided by the Curriculum Development and Support Department.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

**Statistics I: Introduction to ANOVA, Regression, and Logistic Regression Course Notes**

Copyright © 2007 by SAS Institute Inc., Cary, NC 27513, USA. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

---

Book code E70060, course code LWSTAT1, prepared date 03Apr06.

LWSTAT1\_005

## Table of Contents

Course Description .....	vi
Prerequisites .....	vii
General Conventions .....	viii
<b>Chapter 1     Introduction to Statistics .....</b>	<b>1-1</b>
1.1    Fundamental Statistical Concepts .....	1-2
1.2    Examining Distributions .....	1-8
1.3    Confidence Intervals for the Mean .....	1-32
1.4    Hypothesis Testing .....	1-42
1.5    Chapter Summary .....	1-52
<b>Chapter 2     Analysis of Variance (ANOVA) .....</b>	<b>2-1</b>
2.1    One-Way ANOVA: Two Populations.....	2-2
2.2    ANOVA with More than Two Populations .....	2-22
2.3    Two-Way ANOVA with Interactions.....	2-48
2.4    Chapter Summary .....	2-61
<b>Chapter 3     Regression .....</b>	<b>3-1</b>
3.1    Exploratory Data Analysis .....	3-2
3.2    Simple Linear Regression .....	3-27
3.3    Concepts of Multiple Regression .....	3-47
3.4    Model Building and Interpretation.....	3-61
3.5    Chapter Summary .....	3-83

<b>Chapter 4    Regression Diagnostics.....</b>	<b>4-1</b>
4.1 Examining Residuals .....	4-2
4.2 Influential Observations.....	4-14
4.3 Collinearity .....	4-24
4.4 Chapter Summary .....	4-43
<b>Chapter 5    Categorical Data Analysis.....</b>	<b>5-1</b>
5.1 Describing Categorical Data .....	5-2
5.2 Tests of Association .....	5-17
5.3 Introduction to Logistic Regression.....	5-36
5.4 Multiple Logistic Regression.....	5-58
5.5 Logit Plots (Self-Study) .....	5-78
5.6 Chapter Summary .....	5-84
<b>Appendix A    Exercises and Solutions.....</b>	<b>A-1</b>
Exercises.....	A-2
Chapter 1.....	A-2
Chapter 2.....	A-4
Chapter 3.....	A-5
Chapter 4.....	A-8
Chapter 5.....	A-9
Solutions to Exercises .....	A-11
Chapter 1.....	A-11
Chapter 2.....	A-16
Chapter 3.....	A-36
Chapter 4.....	A-66
Chapter 5.....	A-77

**Appendix B Sampling from SAS Data Sets .....** **B-1**

B.1 Random Samples ..... B-2

**Appendix C Additional Topics.....** **C-1**

C.1 Paired *t*-Tests..... C-3

C.2 Two-Sample *t*-Tests..... C-7

C.3 Output Delivery System..... C-17

C.4 Nonparametric ANOVA ..... C-27

C.5 Partial Leverage Plots ..... C-40

**Appendix D Percentile Definitions.....** **D-1**

D.1 Calculating Percentiles..... D-2

**Appendix E Advanced Programs.....** **E-1**

E.1 Interaction Plot..... E-2

**Appendix F Randomization Technique .....** **F-1**

F.1 Randomize Paints..... F-2

**Appendix G Basic Statistics Guidelines for Analysis .....** **G-1**

G.1 Guidelines for Analysis..... G-2

**Appendix H Additional Resources.....** **H-1**

H.1 References..... H-2

## Course Description

This course focuses on the following key areas: statistical inference, analysis of variance, multiple regression, categorical data analysis, and logistic regression. You learn to construct graphs to explore and summarize data, construct confidence intervals for means, test hypotheses, apply multiple comparison techniques in ANOVA, assess and correct collinearity in multiple regression, use diagnostic statistics to identify potential outliers in multiple regression, use chi-square statistics to detect associations among categorical variables, and fit a multiple logistic regression model.

### To learn more...



---

#### SAS Education

A full curriculum of general and statistical instructor-based training is available at any of the Institute's training facilities. Institute instructors can also provide on-site training.

For information on other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to [training@sas.com](mailto:training@sas.com). You can also find this information on the Web at [support.sas.com/training/](http://support.sas.com/training/) as well as in the Training Course Catalog.



---

#### SAS Publishing

For a list of other SAS books that relate to the topics covered in this Course Notes, USA customers can contact our SAS Publishing Department at 1-800-727-3228 or send e-mail to [sasbook@sas.com](mailto:sasbook@sas.com). Customers outside the USA, please contact your local SAS office.

Also, see the Publications Catalog on the Web at [support.sas.com/pubs](http://support.sas.com/pubs) for a complete list of books and a convenient order form.

## Prerequisites

Before selecting this course, you should

- have completed an undergraduate course in statistics covering *p*-values, hypothesis testing, analysis of variance, and regression
- be able to execute SAS programs and create SAS data sets. You can gain this experience by completing the SAS® Programming I: Essentials course.

## General Conventions

This section explains the various conventions used in presenting text, SAS language syntax, and examples in this book.

### Typographical Conventions

You will see several type styles in this book. This list explains the meaning of each style:

UPPERCASE ROMAN	is used for SAS statements and other SAS language elements when they appear in the text.
<i>italic</i>	identifies terms or concepts that are defined in text. Italic is also used for book titles when they are referenced in text, as well as for various syntax and mathematical elements.
<b>bold</b>	is used for emphasis within text.
monospace	is used for examples of SAS programming statements and for SAS character strings. Monospace is also used to refer to variable and data set names, field names in windows, information in fields, and user-supplied information.
<u>select</u>	indicates selectable items in windows and menus. This book also uses icons to represent selectable items.

### Syntax Conventions

The general forms of SAS statements and commands shown in this book include only that part of the syntax actually taught in the course. For complete syntax, see the appropriate SAS reference guide.

```
PROC CHART DATA = SAS-data-set;  
    HBAR | VBAR chart-variables </ options>;  
RUN;
```

This is an example of how SAS syntax is shown in text:

- **PROC** and **CHART** are in uppercase bold because they are SAS keywords.
- **DATA=** is in uppercase to indicate that it must be spelled as shown.
- *SAS-data-set* is in italic because it represents a value that you supply. In this case, the value must be the name of a SAS data set.
- **HBAR** and **VBAR** are in uppercase bold because they are SAS keywords. They are separated by a vertical bar to indicate they are mutually exclusive; you can choose one or the other.
- *chart-variables* is in italic because it represents a value or values that you supply.
- </ *options*> represents optional syntax specific to the HBAR and VBAR statements. The angle brackets enclose the slash as well as *options* because if no options are specified you do not include the slash.
- **RUN** is in uppercase bold because it is a SAS keyword.

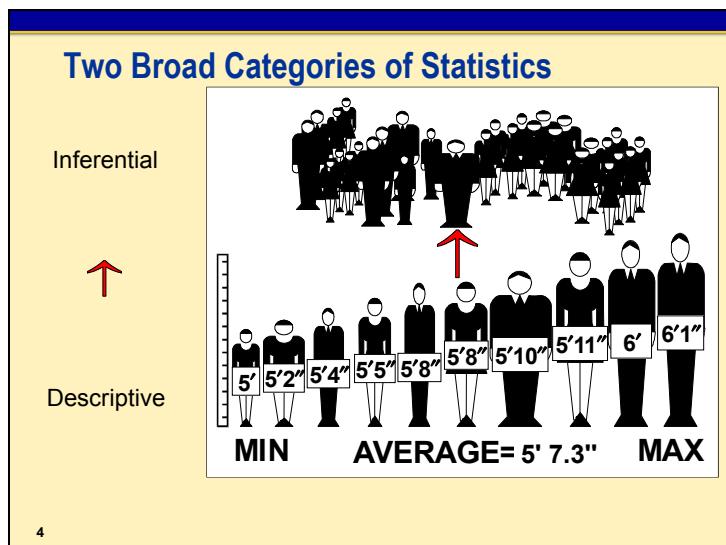
# Chapter 1 Introduction to Statistics

<b>1.1 Fundamental Statistical Concepts .....</b>	<b>1-2</b>
<b>1.2 Examining Distributions .....</b>	<b>1-8</b>
<b>1.3 Confidence Intervals for the Mean.....</b>	<b>1-32</b>
<b>1.4 Hypothesis Testing.....</b>	<b>1-42</b>
<b>1.5 Chapter Summary.....</b>	<b>1-52</b>

## 1.1 Fundamental Statistical Concepts

### Objectives

- Decide what tasks to complete before you analyze your data.
- Distinguish between populations and samples.



*Descriptive statistics* are used to organize, summarize, and focus on the main characteristics of your data. Summarizing your data in such a manner also makes it more usable.

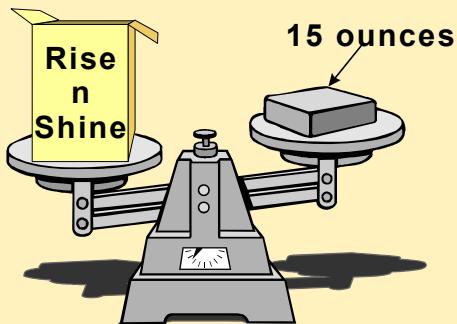
*Inferential statistics* make generalizations or inferences from your data to a larger set of data, based on probability theory.

## Defining the Problem

Before you begin any analysis, you should complete certain tasks.

1. Outline the purpose of the study.
2. Document the study questions.
3. Define the population of interest.
4. Determine the need for sampling.
5. Define the data collection protocol.

### Cereal Example



7

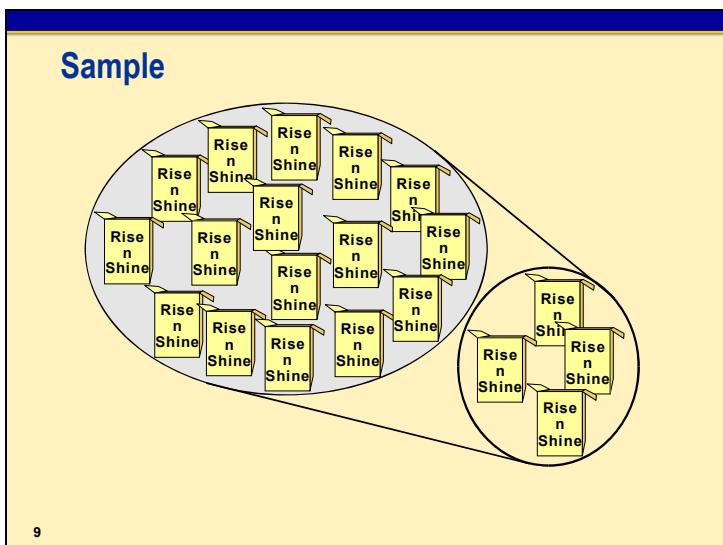
Example: The manufacturer of Rise n Shine cereal wants to test whether the company is producing the specified amount of cereal. Each box is supposed to contain 15 ounces. There are approximately one million boxes of Rise n Shine cereal available to customers.

### Defining the Problem

The purpose of the study is to determine whether Rise n Shine cereal boxes contain 15 ounces of cereal.

The study question is whether the average amount of cereal in Rise n Shine boxes is equal to 15 ounces.

8



A *population* is a collection of all objects about which information is desired.

In our example, the population is all Rise n Shine cereal boxes in the country.

Populations can be categorized as either concrete or theoretical:

- A population is referred to as *concrete* if you can identify every subject in the population. For example, at any one point in time (that is, as of June 30, 2004), you can identify each person on the company payroll. These people constitute a concrete population.
- A population is referred to as *theoretical* if the population is constantly changing. For example, because Rise n Shine cereal continues to be produced and packaged, the population changes almost continuously.

Because there are approximately one million cereal boxes in the grocery stores, you would need to record approximately one million measurements to examine the entire population.

Is it feasible to examine the entire population?

No, the population consists of approximately one million measurements. This would require too much time and too many resources to conduct the study and analyze the results.

A *sample* is a subset of the population. The sample should be random to help ensure that it is representative of the population. A representative sample has characteristics that are similar to the population's characteristics.

For the cereal example, that means the average weight of cereal in a representative sample of Rise n Shine boxes should be close to the average weight of all Rise n Shine boxes.

## Parameters and Statistics

Statistics are used to approximate population parameters.

	Population Parameters	Sample Statistics
Mean	$\mu$	$\bar{x}$
Variance	$\sigma^2$	$s^2$
Standard Deviation	$\sigma$	$s$

10

*Parameters* are characteristics of populations. Because populations usually cannot be measured in their entirety, parameter values are generally unknown. *Statistics* are quantities calculated from the values in the sample.

Suppose you have  $x_1, x_2, \dots, x_n$ , a sample from some population.

- $\bar{x} = \frac{1}{n} \sum x_i$  the mean is an average, a typical value in the distribution.
- $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$  the variance measures the sample variability.
- $s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$  the standard deviation measures variability. It is reported in the same units as the mean.

## Describing Your Data

The goals when you are describing data are to

- screen for unusual data values
- inspect the spread and shape of continuous variables
- characterize the central tendency
- draw preliminary conclusions about your data.

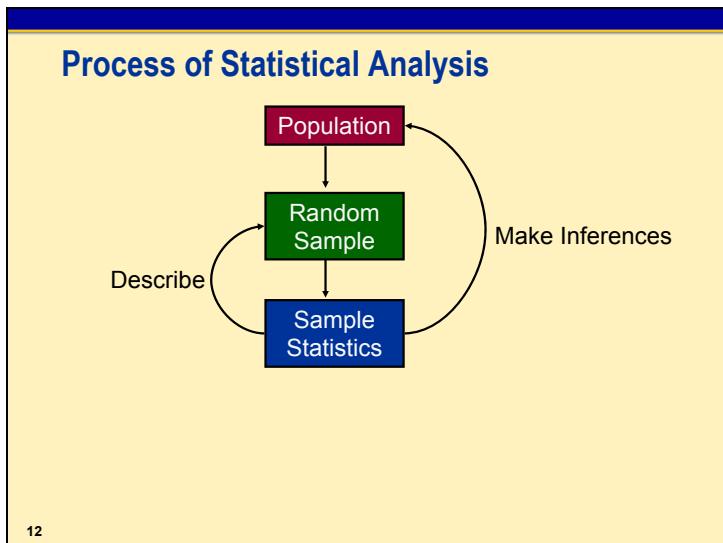
11

After you select a random sample of the data, you can start describing the data. Although you want to draw conclusions about your population, you first want to explore and describe your data before you use inferential statistics.

### Why?

- Data must be as error-free as possible.
- Unique aspects, such as data values that cluster or show some unusual shape, could be missed.
- An extreme value of a variable could be missed and cause gross errors in the interpretation of the statistics.

 Some scientists have suggested that all great scientific discoveries have been due to outliers. The outlying observation indicates an event that is unexpected and does not follow existing theories. In resolving the anomaly, new theories are born.



12

These processes are involved in a statistical analysis:

1. Identify the population of interest.
2. Draw a random sample.
3. Compute sample statistics to describe the sample.
4. Use sample information to make inferences about the population.

## 1.2 Examining Distributions

### Objectives

- Examine distributions of data.
- Explain and interpret measures of location, dispersion, and shape.
- Use the MEANS and UNIVARIATE procedures to produce descriptive statistics.
- Use the UNIVARIATE procedure to generate histograms and normal probability plots.

16

### Cereal Data Set

brand	weight	idnumber
-------	--------	----------



.	.
.	.
.	.
.	.
.	.
.	.

17

Example: A consumer advocacy group wants to determine whether Rise n Shine cereal boxes contain 15 ounces of cereal. A random sample of 40 boxes is selected. The identification number of each box (**idnumber**) and the amount of cereal in ounces (**weight**) are recorded. The data is stored in the **sasuser.b\_rise** data set.



You might feel that the consumer advocacy group would be more interested in whether the cereal manufacturer is “cheating” the consumer by packaging less than 15 ounces of cereal. This possibility is discussed later in the chapter.

### Assumption for This Course

- The sample drawn is *representative* of the population.
  - In other words, the sample characteristics should reflect the characteristics of the population as a whole.

18

One sampling method that helps ensure a representative sample is *simple random sampling*.

Simple random sampling helps ensure that the samples obtained will provide accurate population parameter estimates.

In a simple random sample, every member of the population has an equal chance of being included.

 If you select the sample using valid statistical sampling methods, then you do not need to assume that the sample is representative. Statistical theory can show that sample statistics are accurate estimates of population characteristics.

In the cereal example, each box has an equal chance of being selected from the population.

 See Appendix B, “Sampling from SAS Data Sets,” for information on how to generate random samples without replacement and with replacement.

Why not select cereal boxes from one grocery store near your home?

When you only select boxes that are easily available to you, you are using *convenience sampling*.

A *biased* sample is one that is not representative of the population from which it is drawn. Convenience sampling can lead to biased samples.

In the cereal example, the average weight of a biased sample might not be close to the true average of the population. This can cause the consumer advocacy group to draw erroneous conclusions about the cereal Rise n Shine.

## Distributions

When you examine the distribution of values for the variable **weight**, you can find out

- the range of possible data values
- the frequency of data values
- whether the data values accumulate in the middle of the distribution or at one end.

19

A *distribution* is a collection of data values that are arranged in order, along with the relative frequency. For any kind of problem, it is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.

For the cereal example, these questions can be addressed using graphical techniques.

- Are the values of **weight** symmetrically distributed?
- Are any values of **weight** unusual?

You can answer these questions using descriptive statistics.

- What is the best estimate of the average of the values of **weight** for the population?
- What is the best estimate of the average spread or dispersion of the values of **weight** for the population?

### “Typical Values” in a Distribution

- Mean: the sum of all the values in the data set divided by the number of values

$$\frac{\sum_{i=1}^n x_i}{n}$$

- Median: the middle value (also known as the 50<sup>th</sup> percentile)
- Mode: the most common or frequent data value

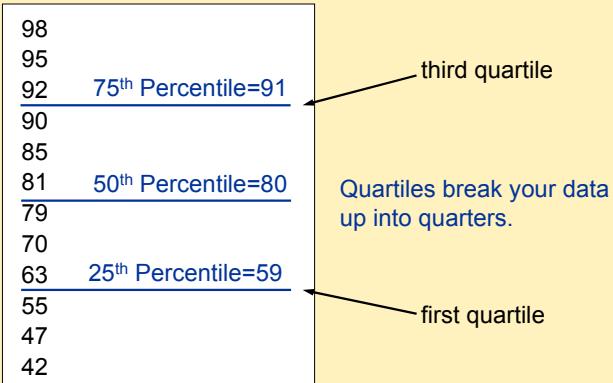
20

Descriptive statistics that locate the center of your data are called *measures of central tendency*. The most common measure of central tendency is the sample mean.

A property of the sample mean is that the sum of the differences of each data value from the mean is always 0. That is,  $\sum(x_i - \bar{x}) = 0$ .

The mean is the physical balancing point of your data.

### Percentiles



21

*Percentiles* locate a position in your data larger than a given proportion of data values.

Commonly reported percentile values are

- the 25<sup>th</sup> percentile, also called the first quartile
- the 50<sup>th</sup> percentile, also called the median
- the 75<sup>th</sup> percentile, also called the third quartile.

## The Spread of a Distribution: Dispersion

Measure	Definition
<i>range</i>	the difference between the maximum and minimum data values
<i>interquartile range</i>	the difference between the 25 <sup>th</sup> and 75 <sup>th</sup> percentiles
<i>variance</i>	a measure of dispersion of the data around the mean
<i>standard deviation</i>	a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

22

Measures of dispersion enable you to characterize the dispersion, or spread, of the data.

$$\text{Formula for sample variance: } s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$



Another measure of variation is the coefficient of variation (C.V.), which is the standard deviation as a percentage of the mean. It is defined as  $\frac{s}{\bar{x}} \times 100$ .

## The MEANS Procedure

General form of the MEANS procedure:

```
PROC MEANS DATA=SAS-data-set <options>;
  VAR variables;
  RUN;
```

23

The MEANS procedure is a Base SAS procedure for generating descriptive statistics for your data.

Selected MEANS procedure statement:

VAR specifies numeric variables for which you want to calculate descriptive statistics. If no VAR statement appears, all numeric variables in the data set are analyzed.

-  For assistance with the correct syntax and options for a SAS procedure you can type **help** followed by the name of the procedure in the command box. This opens the Help window for that procedure. After you are in the appropriate Help window, select **syntax** to see all options available for that procedure.



## Descriptive Statistics

Example: Use the PRINT procedure to list the first 10 observations in the data set **sasuser.b\_rise**. Then use PROC MEANS to generate descriptive statistics for **weight**.

```
/* c1demo01 */
options nodate nonumber;
proc print data=sasuser.b_rise (obs=10);
  title 'Listing of the Cereal Data Set';
run;
```

Listing of the Cereal Data Set

Obs	brand	weight	idnumber
1	Rise n Shine	15.0136	33081197
2	Rise n Shine	14.9982	37070397
3	Rise n Shine	14.9930	60714297
4	Rise n Shine	15.0812	9589297
5	Rise n Shine	15.0418	85859397
6	Rise n Shine	15.0639	99108497
7	Rise n Shine	15.0613	70847197
8	Rise n Shine	15.0255	53750297
9	Rise n Shine	15.0176	3873197
10	Rise n Shine	15.0122	43493297

```
/* c1demo02 */
proc means data=sasuser.b_rise maxdec=4;
  var weight;
  title 'Descriptive Statistics Using PROC MEANS';
run;
```

Selected PROC MEANS statement option:

MAXDEC= specifies the maximum number of decimal places to use when printing numeric values.

Descriptive Statistics Using PROC MEANS

The MEANS Procedure

Analysis Variable : weight

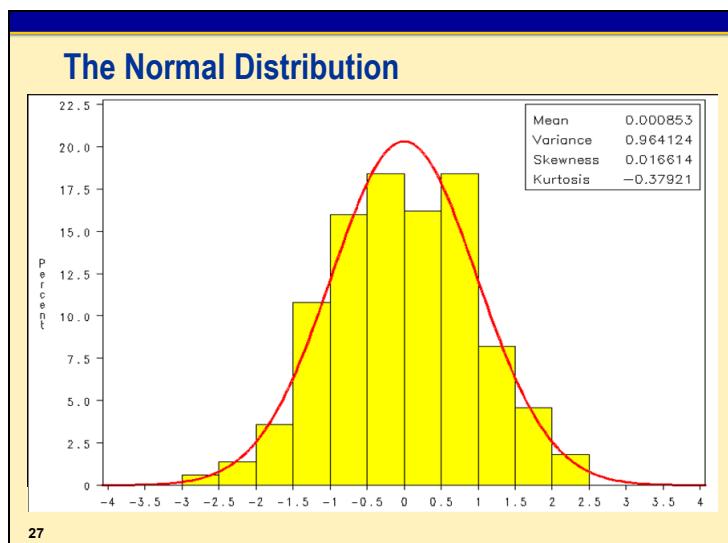
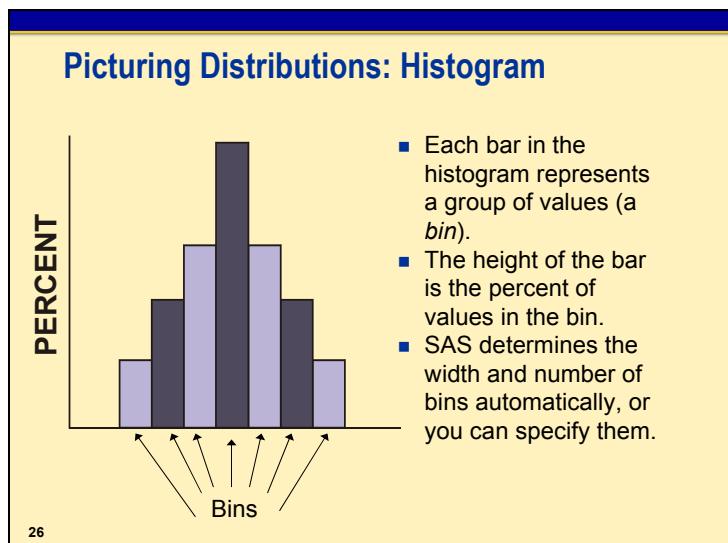
N	Mean	Std Dev	Minimum	Maximum
40	15.0360	0.0265	14.9831	15.0980

By default, PROC MEANS prints the number of nonmissing observations, the mean, the standard deviation, the minimum value, and the maximum value. You can add options to the MEANS statement to request additional statistics.

```
proc means data=sasuser.b_rise  
    maxdec=4  
    n mean median std var q1 q3;  
var weight;  
title 'Selected Descriptive Statistics for weight';  
run;
```

When you add options to request specific statistics, only the statistics requested appear in the output.

Selected Descriptive Statistics for weight						
The MEANS Procedure						
Analysis Variable : weight						
N	Mean	Median	Std Dev	Variance	Lower Quartile	Upper Quartile
40	15.0360	15.0348	0.0265	0.0007	15.0160	15.0525



The normal distribution is a commonly used distribution in statistics. It is characterized by its bell shape and its two parameters: the mean and the standard deviation.

In evaluating distributions, it is useful to look at measures of the shape of the distribution compared to the normal. Two such measures are skewness and kurtosis, which are defined over the next few pages.

Theoretically, the normal distribution has skewness=0 and kurtosis=0, although skewness and kurtosis measures from samples of a normal distribution typically vary somewhat from zero.

## The Normal Distribution

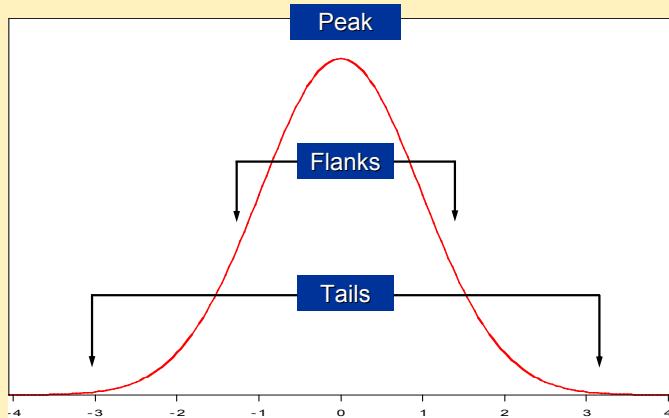
The normal distribution

- is *symmetric*. If you draw a line down the center, you get the same shape on either side.
- is *fully characterized by the mean and standard deviation*. Given those two parameters, you know all there is to know about the distribution.
- is bell shaped.
- has mean  $\approx$  median  $\approx$  mode.

The red line on each of the following graphs represents the shape of the normal distribution with the mean and variance estimated from the sample data.

28

## Characteristics of the Bell Curve

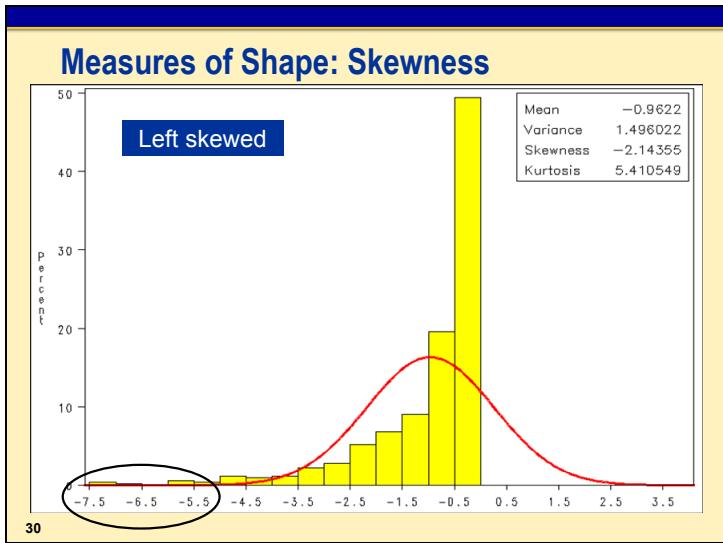


29

To understand skewness and kurtosis, it can be useful to think of the bell-shaped curve of the normal distribution as consisting of three parts:

- The peak, or center, of the curve is where most of the observations occur. Generally, the peak is within 1 standard deviation of the mean.
- The flanks are the areas beyond the peak between roughly 1 and 2 standard deviations from the mean. These observations, although not typical, are still not unusual. Slightly less than 30% of the observations are expected to fall between 1 and 2 standard deviations from the mean.
- The tails are the areas far from the center of the distribution, usually considered to be beyond 2 standard deviations in the normal distribution. Observations in the tails account for only about 5% of the normal distribution.

Although these components of the normal curve (the peak, the flanks, and the tails) are not formally defined here, they can be useful as a tool for describing your data. A distribution can be compared to the normal distribution in terms of how observations tend to be distributed across the peak, flanks, and tails.

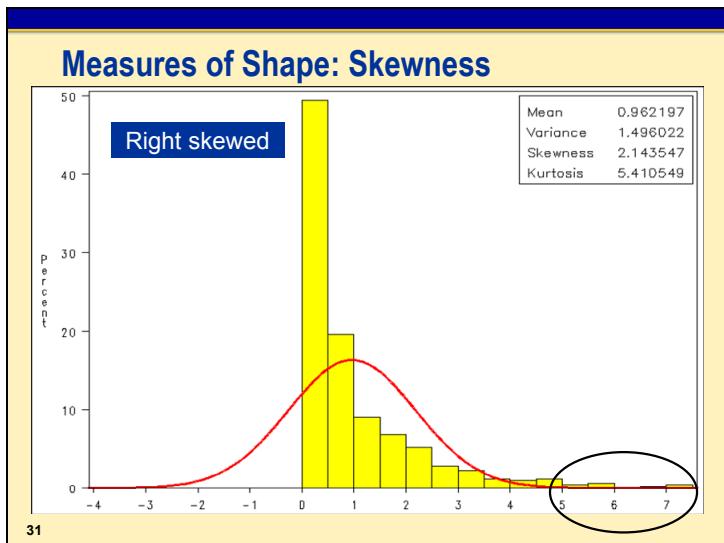


One measure of the shape of a distribution is skewness. The *skewness* statistic measures the tendency of your distribution to be more spread out on one side than the other. A distribution that is approximately symmetric has a skewness statistic close to 0.

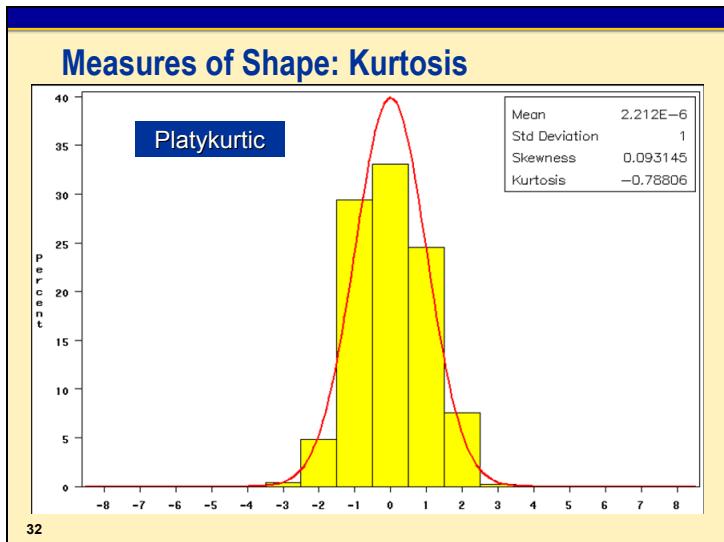
If your distribution is more spread out on the

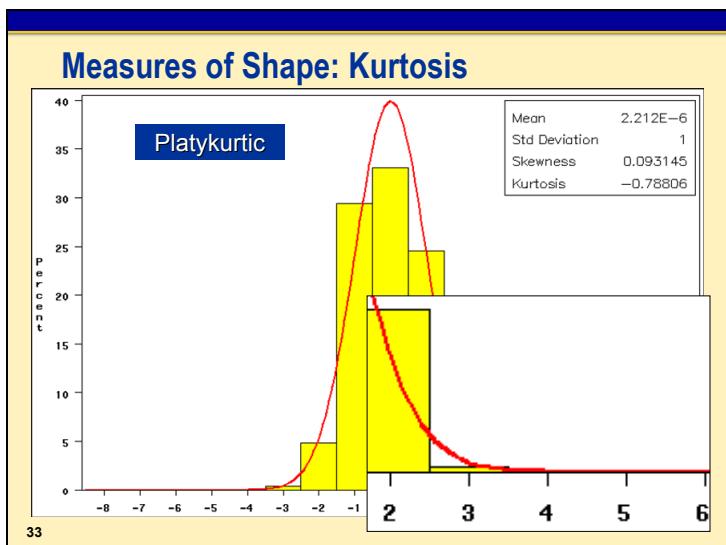
- left side, then the statistic is negative, and the mean is less than the median. This is sometimes referred to as a *left-skewed* or *negatively skewed* distribution.
- right side, then the statistic is positive, and the mean is greater than the median. This is sometimes referred to as a *right-skewed* or *positively skewed* distribution.

Notice in the picture above that the greatest concentration of observations is just above the peak of the normal reference curve and that a large portion of observations are spread out in the tail on the left side. Few (if any) observations are found in the ride-side flank or tail. This histogram represents a distribution with a large negative skewness statistic and is therefore a **left-skewed** distribution.



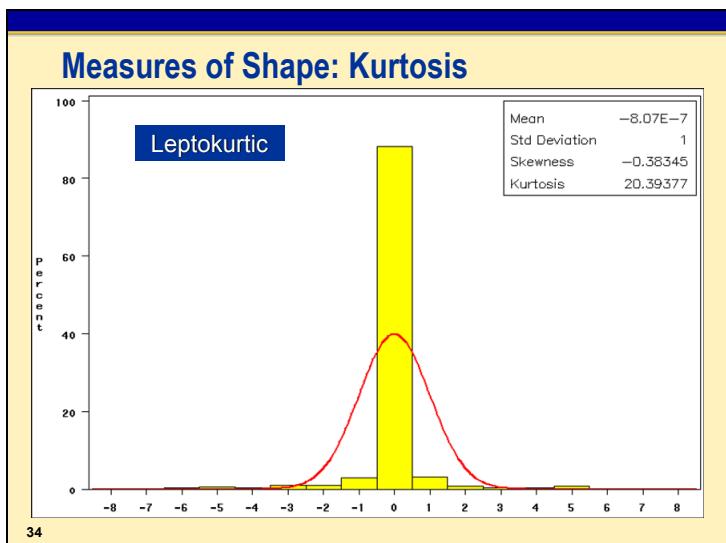
The picture above shows that the greatest concentration of observations is just below the peak of the normal reference curve and that a large portion of observations is spread out in the tail on the right side. Few (if any) observations are found in the left-side flank or tail. This histogram represents a distribution with a large positive skewness statistic and is therefore a **right-skewed** distribution.

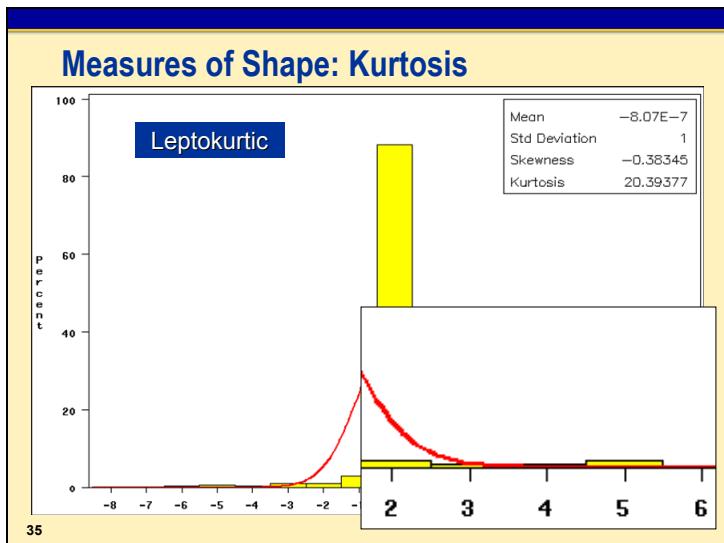




The *kurtosis* statistic measures the tendency of your data to be distributed toward the center or toward the tails of the distribution. A distribution that is approximately normal has a kurtosis statistic close to 0.

If your kurtosis statistic is negative, the distribution is said to be *platykurtic* compared to the normal. If the distribution is symmetric, a platykurtic distribution tends to have a larger-than-normal proportion of observations in the flanks, a smaller-than-normal proportion of observations in the tails, and/or a somewhat flat peak. A platykurtic distribution is often referred to as *light-tailed*.





If your kurtosis statistic is positive, the distribution is said to be *leptokurtic* compared to the normal. If the distribution is symmetric, a leptokurtic distribution tends to have a larger-than-normal proportion of observations in the extreme tails, a smaller-than-normal proportion of observations in the flanks, and/or a taller peak than the normal. A leptokurtic distribution is often referred to as *heavy-tailed*. Leptokurtic distributions are also sometimes referred to as *outlier-prone distributions*.

Distributions that are asymmetric also tend to have nonzero kurtosis. In these cases, understanding kurtosis is considerably more complex than in situations where the distribution is approximately symmetric.

## The UNIVARIATE Procedure

General form of the UNIVARIATE procedure:

```
PROC UNIVARIATE DATA=SAS-data-set <options>;
  VAR variables;
  ID variable;
  HISTOGRAM variables </ options>;
  PROBPLOT variables </ options>;
RUN;
```

36

The UNIVARIATE procedure not only computes descriptive statistics, it also provides greater detail on the distributions of the variables.

Selected UNIVARIATE procedure statements:

VAR	specifies numeric variables to analyze. If no VAR statement appears, then all numeric variables in the data set are analyzed.
ID	specifies a variable used to label the five lowest and five highest values in the output.
HISTOGRAM	creates high-resolution histograms.
PROBPLOT	creates a high-resolution probability plot, which compares ordered variable values with the percentiles of a specified theoretical distribution.



## Descriptive Statistics

Example: Use the UNIVARIATE procedure to calculate measures of shape and display a high-resolution histogram of the variable **weight**.

```
/* c1demo03 */
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=sasuser.b_rise;
  var weight;
  id idnumber;
  histogram weight;
  title 'Descriptive Statistics Using PROC UNIVARIATE';
run;
```

### PROC UNIVARIATE Output

#### Descriptive Statistics Using PROC UNIVARIATE

The UNIVARIATE Procedure  
Variable: weight

##### Moments

N	40	Sum Weights	40
Mean	15.03596	Sum Observations	601.4384
Std Deviation	0.02654963	Variance	0.00070488
Skewness	0.39889232	Kurtosis	-0.1975717
Uncorrected SS	9043.23122	Corrected SS	0.02749044
Coeff Variation	0.17657424	Std Error Mean	0.00419787

##### Basic Statistical Measures

###### Location Variability

Mean	15.03596	Std Deviation	0.02655
Median	15.03480	Variance	0.0007049
Mode	15.01220	Range	0.11490
		Interquartile Range	0.03650

NOTE: The mode displayed is the smallest of 2 modes with a count of 2.

##### Tests for Location: Mu0=0

Test	-Statistic-	-----	p Value-----
Student's t	t 3581.811	Pr >  t	<.0001
Sign	M 20	Pr >=  M	<.0001
Signed Rank	S 410	Pr >=  S	<.0001

## PROC UNIVARIATE Output (continued)

Quantiles (Definition 5)		
Quantile	Estimate	
100% Max	15.0980	
99%	15.0980	
95%	15.0863	
90%	15.0726	
75% Q3	15.0525	
50% Median	15.0348	
25% Q1	15.0160	
10%	15.0095	
5%	14.9956	
1%	14.9831	
0% Min	14.9831	

Extreme Observations					
-----Lowest-----			-----Highest-----		
Value	idnumber	Obs	Value	idnumber	Obs
14.9831	30834797	37	15.0639	99108497	6
14.9930	60714297	3	15.0812	9589297	4
14.9982	37070397	2	15.0858	73461797	21
15.0093	46028397	14	15.0868	40177297	27
15.0096	59149297	40	15.0980	23573597	35

The output indicates that

- the mean or center point of the data is 15.03596 ounces. This is approximately equal to the median (15.0348), which indicates the distribution is fairly symmetric.
- the standard deviation is 0.02655, which means that the average variability around the mean is approximately 0.027 ounces.
- the distribution is slightly skewed to the right.
- the distribution has lighter tails than the normal distribution.
- the range of the data is 0.1149, the difference between 14.9831 and 15.098.
- the interquartile range focuses on the variation of the middle 50% of the data and is 0.0365.
- the cereal box with the largest amount of cereal has an identification number of 23573597, which is also observation number 35 in the data set.

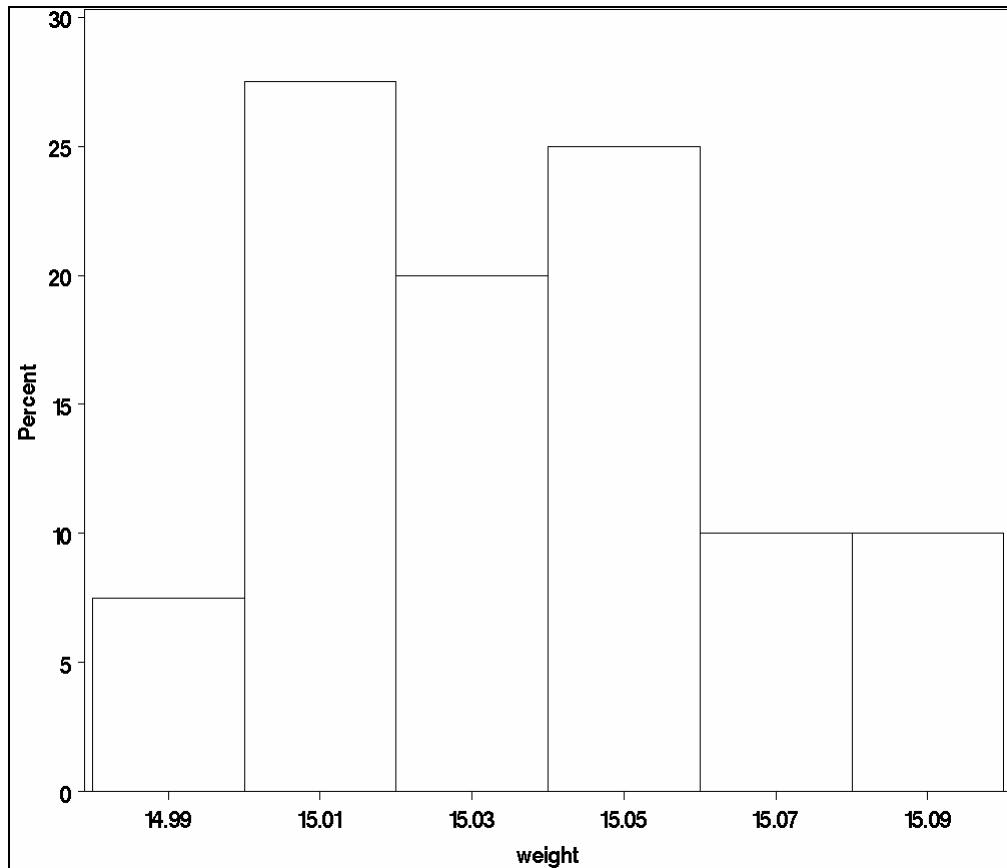
The *mode* is the most frequent data value. The note in the output listing indicates that the mode displayed is the smaller of two modes with a count of 2. If there are no replicated values in your data, the mode does not exist and, therefore, is reported as missing.

 If you would like a table of the modes and their respective frequencies, add the MODES option in the PROC UNIVARIATE statement

In the Quantiles table, Definition 5 indicates that PROC UNIVARIATE is using the default definition for calculating percentile values. You can use the PCTLDEF= option in the PROC UNIVARIATE statement to specify one of five methods. These methods are listed in Appendix C, “Percentile Definitions.”

The bin identified with the midpoint of 15.01 has approximately 27% of the values; in addition, you can state that 27% of the values fall between the bin end points of 15.00 and 15.02. In a similar way, you can state that approximately 7% of the values fall between 14.98 and 15.00.

Partial PROC UNIVARIATE Graph Output



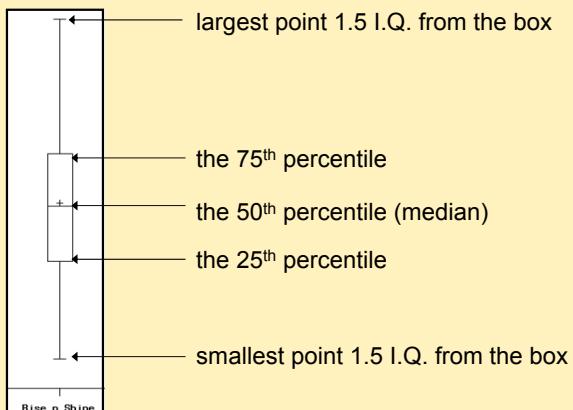
## Graphical Displays of Distributions

You can produce three kinds of plots for examining the distribution of your data values:

- histograms
- box plots
- normal probability plots.

40

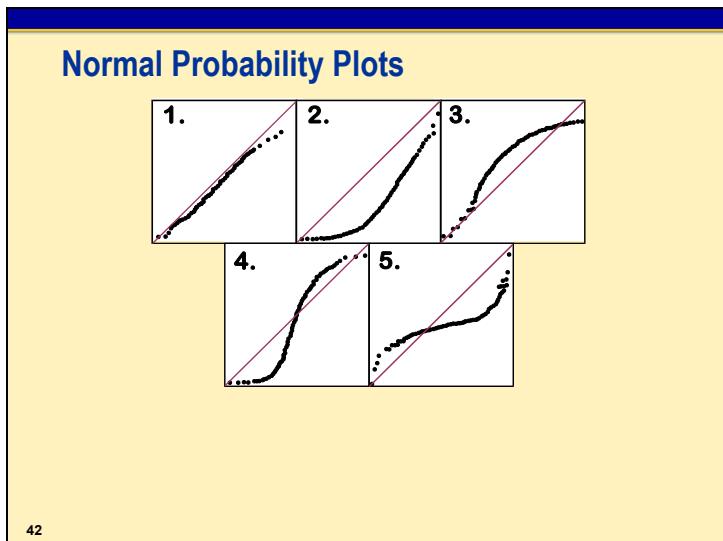
## Box-and-Whisker Plots



41

*Box plots* provide information about the variability of data and the extreme data values. The box represents the middle of your data, and you get a rough impression of the symmetry of your distribution by comparing the mean and median. The whiskers extend from the box as far as the data extends, to a distance of, at most, 1.5 interquartile units. Square symbols denote points that are more than 1.5 interquartile units from the box.

The above plot is of the values of **weight** from the **sasuser.b\_rise** data set. The plot shows that the data is symmetric.



A *normal probability plot* is a visual method for determining whether or not your data comes from a distribution that is approximately normal. The vertical axis represents the actual data values, and the horizontal axis displays the expected percentiles from a standard normal distribution.

The above diagrams illustrate some possible normal probability plots for data from a

1. normal distribution (the observed data follow the reference line)
2. skewed-to-the-right distribution
3. skewed-to-the-left distribution
4. light-tailed distribution
5. heavy-tailed distribution.

## The BOXPLOT Procedure

General form of the BOXPLOT procedure:

```
PROC BOXPLOT DATA=SAS-data-set;
   PLOT analysis-variable*group-variable
      </options>;
RUN;
```

43

The BOXPLOT procedure is from SAS/GRAFH software.

Selected BOXPLOT procedure statement:

PLOT       the *analysis-variable* identifies one or more variables to be analyzed. If you specify more than one analysis variable, enclose the list in parentheses. The *group-variable* specifies the variable that identifies groups in the data. The group variable and at least one analysis variable is required.

If you need to create a box plot without a group variable, you can create a dummy group variable with only one level. For example, with the **sasuser.b\_boston** data set, you use the following code:

```
data race;
   set sasuser.b_boston;
   Dummy='1';
run;

proc boxplot data=race;
   plot tottime*Dummy / boxstyle=schematic;
run;
```



## Examining Distributions

Example: Use the PROBPLOT statement in PROC UNIVARIATE to generate the normal probability plot, and use PROC BOXPLOT to create a box-and-whisker plot for the variable **weight** in the **sasuser.b\_rise** data set.

```
/* c1demo04 */
proc univariate data=sasuser.b_rise;
  var weight;
  id idnumber;
  probplot weight / normal (mu=est sigma=est
                             color=blue w=1);
  title;
run;
```

You cancel all previously defined titles by submitting a TITLE statement.

Selected UNIVARIATE procedure statements:

PROBPLOT creates a high-resolution probability plot, which compares ordered variable values with the percentiles of a specified theoretical distribution.

Selected PROBPLOT statement option:

NORMAL superimposes a reference line on the normal probability plot using the estimates of mu and sigma from the data. In this example, the reference line will be blue with a width of 1.

```
proc boxplot data=sasuser.b_rise;
  plot weight*brand / cboxes=black
                     boxstyle=schematic;
run;
```

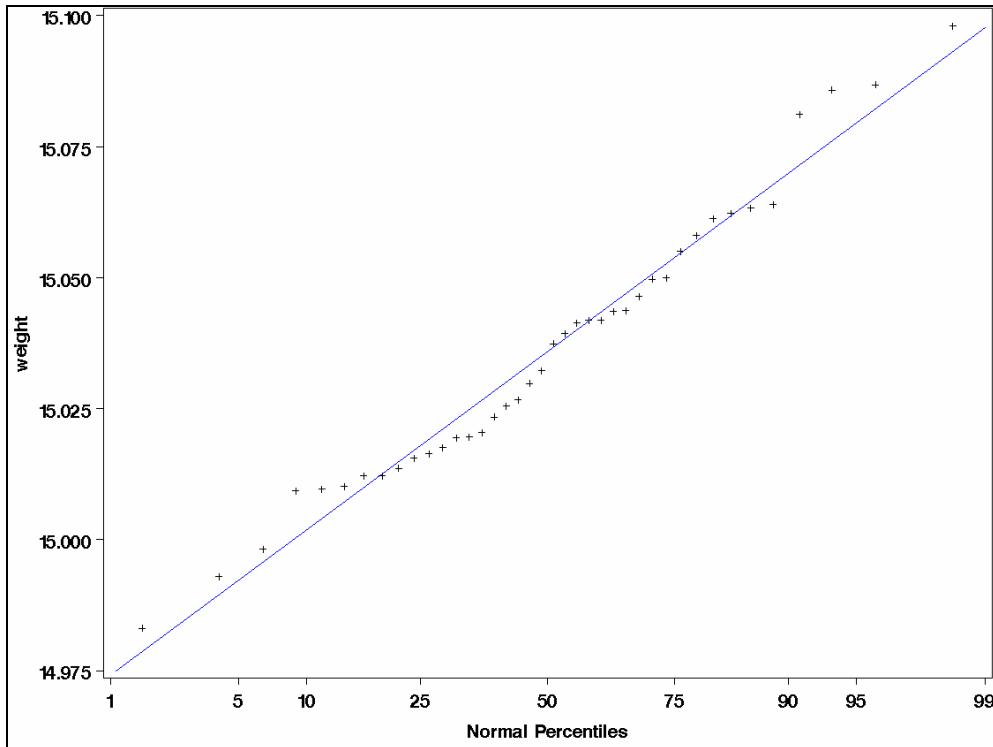
Selected PLOT statement options for the BOXPLOT procedure:

CBOXES= specifies the color of the box plots.

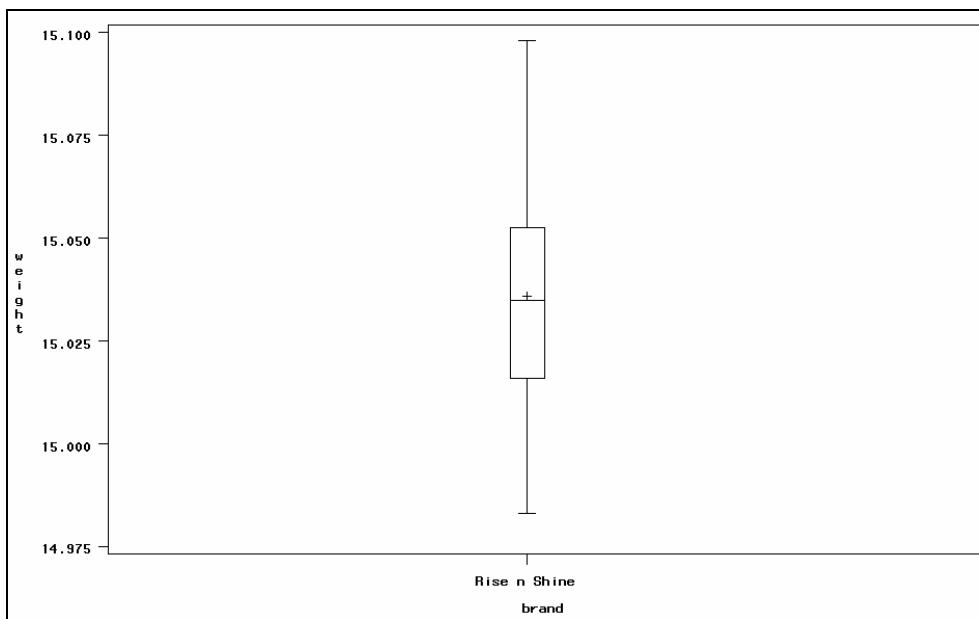
BOXSTYLE= specifies the type of box plot to be drawn. The default type is SKELETAL, which draws lines from the ends of the box to the maximum and the minimum points. The SCHEMATIC option creates a plot with lines drawn to the highest point that is 1.5 interquartile units from the box and the lowest point that is 1.5 interquartile units from the box. Points that are more than 1.5 interquartile units from the box are represented with boxes. You can change the plotting symbol with the IDSYMBOL= option.

The graphically enhanced normal probability plot is shown below using the PROBPLOT statement. The 45-degree line represents where the data values would fall if they came from a normal distribution. The plus signs represent the observed data values. Because the plus signs follow the 45-degree line in the graph below, you can conclude that there does not appear to be any severe departure from the normal distribution.

#### Partial PROC UNIVARIATE Graph Output



The box plot most closely resembles the box plot for a normal distribution with no outliers.



Refer to Exercise 1 for Chapter 1 in Appendix A.

## 1.3 Confidence Intervals for the Mean

### Objectives

- Explain and interpret the confidence intervals for the mean.
- Explain the central limit theorem.
- Calculate confidence intervals using PROC MEANS.

50

### Point Estimates

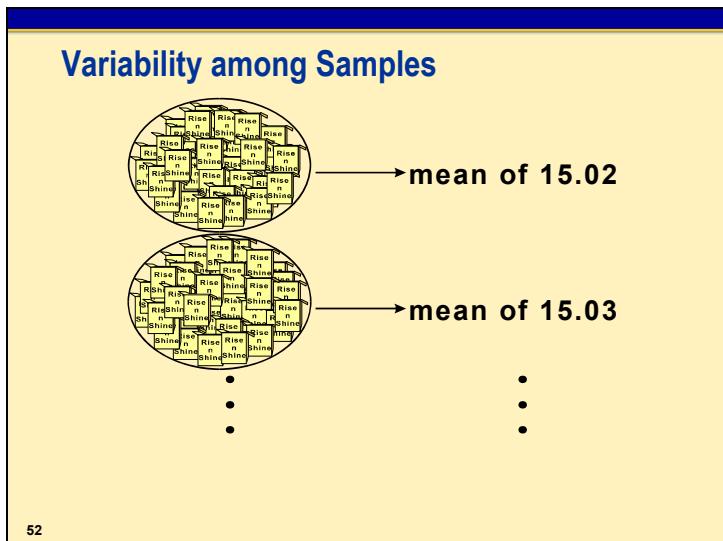
$$\begin{array}{ccc} \overline{X} & \text{estimates} & \mu \\ S & \text{estimates} & \sigma \end{array}$$

51

A *point estimate* is a sample statistic used to estimate a population parameter.

- An estimate of the average **weight** is 15.036, and an estimate of the standard deviation is 0.027.
- Because you only have an estimate of the unknown population mean, you need to know the variability of your estimate.
- A point estimate does not take into account the accuracy of the calculated statistic.

 Variance is the traditional measure of precision. Mean Square Error (MSE) is the traditional measure of accuracy used by statisticians. MSE is equal to variance plus bias-squared. Because the expected value of the sample mean ( $\bar{x}$ ) equals the population mean ( $\mu$ ), MSE equals the variance.



Why are you not absolutely certain that the mean weight for Rise n Shine cereals is 15.036?

The answer is because the sample mean is only an estimate of the population mean. If you collected another sample of cereal boxes, you would have another estimate of the mean.

Therefore, different samples yield different estimates of the mean for the same population. How close these sample means are to one another determines the variability of the estimate of the population mean.

## Standard Error of the Mean

A statistic that measures the variability of your estimate is the *standard error of the mean*.

It differs from the sample standard deviation because

- the sample standard deviation deals with the variability of your data
- the standard error of the mean deals with the variability of your sample mean.

$$\text{Standard error of the mean} = \frac{s}{\sqrt{n}}$$

53

The standard error of the mean is computed as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

$s$  is the sample standard deviation

$n$  is the sample size.

The standard error of the mean for the variable **weight** is  $0.02654963 / 6.324555$ , or approximately 0.004. This is a measure of how much error you can expect when you use the sample mean to predict the population mean. Therefore, the smaller the standard error is, the more precise your sample estimate is.



Precision and accuracy are equivalent for an unbiased estimator. You can improve the precision of an estimate by increasing the sample size.

## Confidence Intervals

### 95% Confidence



- A 95% confidence interval states that you are 95% certain that the true population mean lies between two calculated values.
  - In other words, if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

54

A *confidence interval*

- is a range of values that you believe to contain the population parameter of interest
- places an upper and lower bound around a sample statistic.

To construct a confidence interval, a significance level must be chosen.

A 95% confidence interval is commonly used to assess the variability of the sample mean. In the cereal example, you interpret a 95% confidence interval by stating that you are 95% confident that the interval contains the mean number of ounces of cereal for your population.

Do you want to be as confident as possible?

- Yes, but if you increase the confidence level, the width of your interval increases.
- As the width of the interval increases, it becomes less useful.

## Assumption about Confidence Intervals

The types of confidence intervals in this course assume that the sample means are normally distributed.

55

### Confidence Interval for the Mean

$$\bar{x} \pm t \cdot s_{\bar{x}} \quad \text{or} \quad (\bar{x} - t \cdot s_{\bar{x}}, \bar{x} + t \cdot s_{\bar{x}})$$

where

$\bar{x}$  is the sample mean.

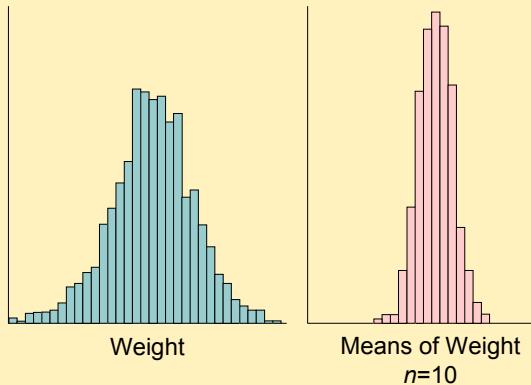
$t$  is the  $t$  value corresponding to the confidence level and  $n-1$  degrees of freedom, where  $n$  is the sample size.

$s_{\bar{x}}$  is the standard error of the mean.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

56

### Distribution of Sample Means



57

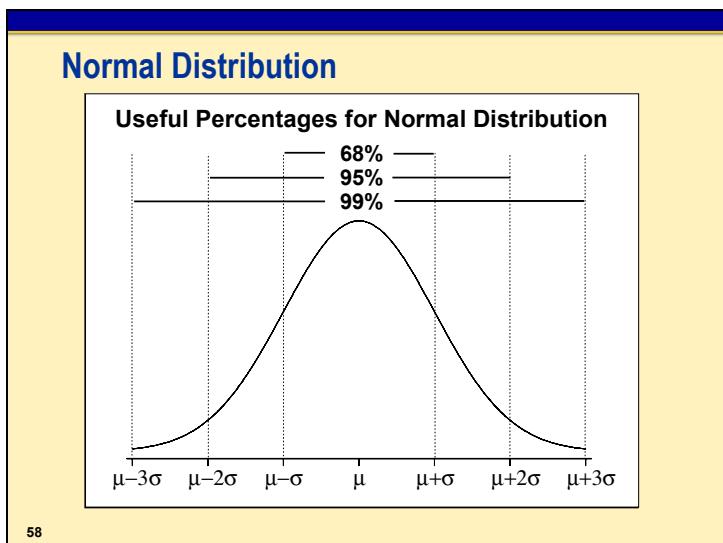
What is a distribution of sample means?

In the cereal example, it is the distribution of all possible sample means of ounces of cereal.

In general terms, suppose 500 random samples, all with the same sample size of 10, are taken from an identified population. Note the following:

- The histogram on the left is the distribution of all 5000 observations.
- The histogram on the right, however, represents the distribution of the 500 sample means.

The variability of the distribution of sample means is smaller,  $s_{\bar{x}} = \frac{s}{\sqrt{10}}$ , than the variability of the distribution of the 5000 observations, which has a standard deviation of  $s$ .



Why does the distribution of sample means have to be normally distributed?

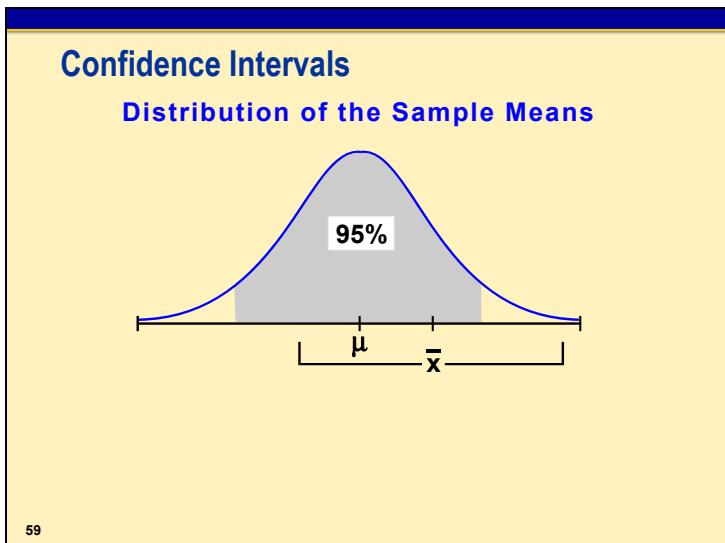
The normal distribution defines probabilities. For example, approximately

- 68% of the data falls within 1 standard deviation of the mean
- 95% of the data falls within 2 standard deviations of the mean
- 99% of the data falls within 3 standard deviations of the mean.

If the distribution of sample means is normal, you can use the probabilities associated with the normal distribution when constructing a confidence interval. The probability corresponds to the confidence level.

Therefore, if you construct a 95% confidence interval, you have a 95% probability of constructing a confidence interval that contains the population mean.

If the distribution of sample means is not normal, you have no idea what range of values corresponds to a 95% confidence interval (unless the distribution of sample means follows another known distribution).



The graph above is the distribution of sample means. The shaded region represents 95% of the area in the distribution.

When constructing a 95% confidence interval, the width of the interval

- covers 95% of the area under the distribution of sample means when it is centered over  $\mu$ , the population mean
- corresponds to a 95% probability of capturing the population mean when the interval is constructed.

Therefore, if the sample mean falls in the shaded region in the distribution of sample means, the interval constructed will contain the population mean.

Notice that  $\mu$  is captured in this interval.

## Verifying the Normality Assumption

To satisfy the assumption of normality, you can either

- verify that the population distribution is approximately normal, or
- apply the central limit theorem.

60

## Central Limit Theorem

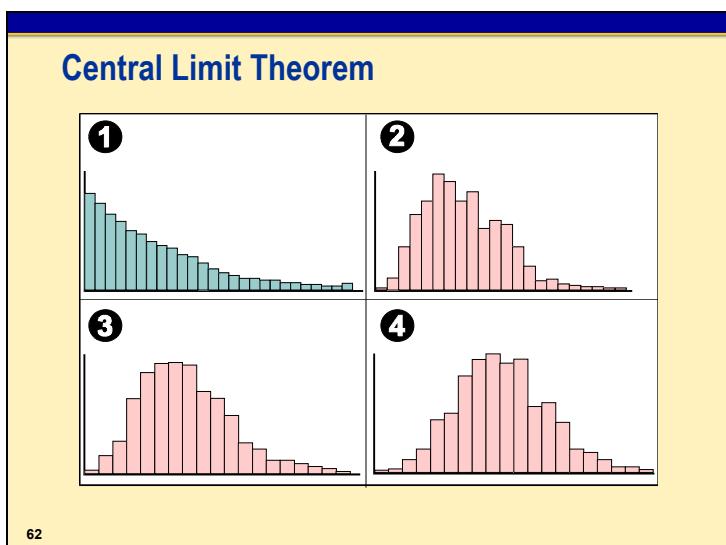
The central limit theorem states that the distribution of sample means is approximately normal, regardless of the distribution's shape, if the sample size is large enough.

"Large enough" is usually about 30 observations: more if the data is heavily skewed, fewer if the data is symmetric.

61

To apply the central limit theorem, your sample size should be at least 30. The central limit theorem holds even if you have no reason to believe the population distribution is not normal.

Because the sample size for the cereal example is 40, you can apply the central limit theorem and satisfy the assumption of normality for the confidence intervals.



62

The graphs illustrate the tendency of a distribution of sample means to approach normality as the sample size increases.

The first chart is a histogram of data values drawn from an exponential distribution. The remaining charts are histograms of the sample means for samples of differing sizes drawn from the same exponential distribution.

1. Data from an exponential distribution
2. 1000 samples of size 5
3. 1000 samples of size 10
4. 1000 samples of size 30



For the sample size of 30, the distribution is approximately bell-shaped and symmetric, even though the sample data is highly skewed.



## Confidence Intervals

Example: Use the MEANS procedure to generate a 95% confidence interval for the mean of **weight** in the **sasuser.b\_rise** data set.

```
/* c1demo05 */
proc means data=sasuser.b_rise n mean stderr clm;
  var weight;
  title '95% Confidence Interval for WEIGHT';
run;
```

Selected PROC MEANS statement options:

- N prints the number of nonmissing values.
- MEAN prints the mean.
- CLM calculates confidence limits for the mean.
- STDERR calculates the standard error of the mean.

The output is shown below.

95% Confidence Interval for WEIGHT				
The MEANS Procedure				
Analysis Variable : weight				
N	Mean	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
40	15.0359600	0.0041979	15.0274690	15.0444510

In the cereal example, you are 95% confident that the population mean ounces for the Rise n Shine cereal boxes is contained in the interval 15.0275 and 15.0445. Because the interval between the upper and lower limits is small from a practical point of view, you can conclude that the sample mean is a fairly accurate estimate of the population mean.

How do you increase the accuracy of your estimate using the same confidence level? If you increase your sample size, you reduce the standard error of the sample mean and therefore reduce the width of your confidence interval. Thus, your estimate will be more accurate.

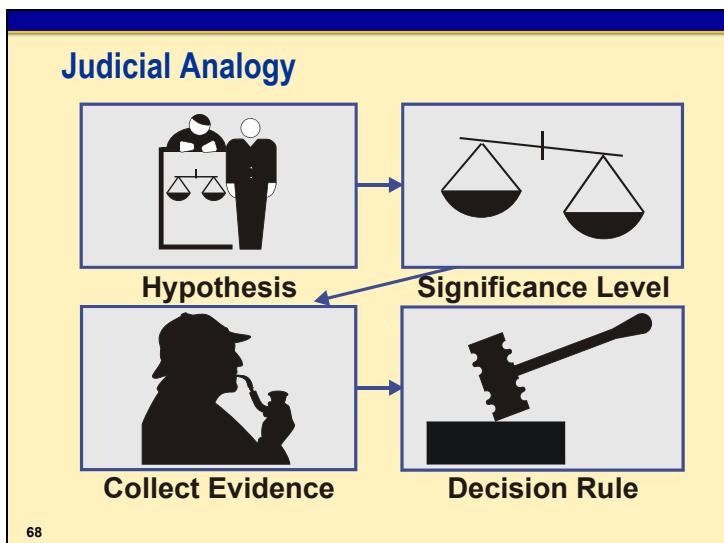
Does 95% of all cereal weights for all Rise n Shine boxes fall between 15.0275 and 15.0445? No, confidence intervals deal with the variability of your sample mean.

- ✍ You can use the ALPHA= option in the PROC MEANS statement to construct confidence intervals with a different confidence level.

## 1.4 Hypothesis Testing

### Objectives

- Define some common terminology related to hypothesis testing.
- Perform hypothesis testing using the UNIVARIATE procedure.



In a criminal court, you put defendants on trial because you suspect they are guilty of a crime. But how does the trial proceed?

Determine the null and alternative hypotheses. The *alternative* hypothesis is your initial research hypothesis (the defendant is guilty). The *null* is the logical opposite of the alternative hypothesis (the defendant is not guilty).

Select a *significance level* as the amount of evidence needed to convict. In a court of law, the evidence must prove guilt “beyond a reasonable doubt.”

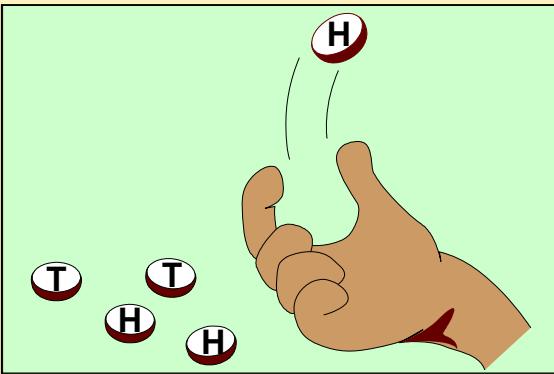
Collect evidence.

Use a *decision rule* to make a judgment. If the evidence is

- sufficiently strong, reject the null hypothesis.
- not strong enough, fail to reject the null hypothesis. Note that failing to prove guilt does not prove that the defendant is innocent.

Statistical hypothesis testing follows this same basic path.

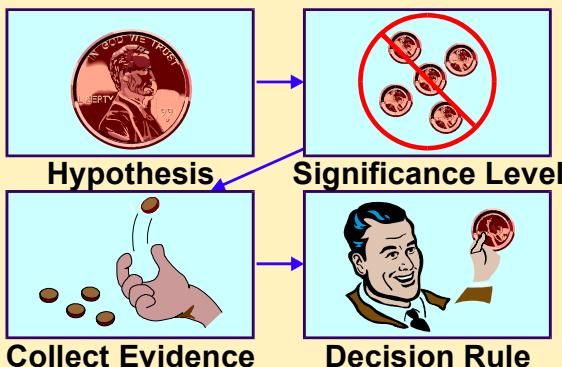
### Coin Example



69

Suppose you want to know whether a coin is fair. You cannot flip it forever, so you decide to take a sample. Flip it five times and count the number of heads and tails.

### Coin Analogy



71

Test whether a coin is fair.

1. You suspect that the coin is **not** fair but recall the legal example and begin by assuming the coin is fair.
2. You select a significance level. If you observe five heads in a row or five tails in a row, you conclude the coin is not fair; otherwise, you decide there is not enough evidence to show the coin is not fair.
3. You flip the coin five times and count the number of heads and tails.
4. You evaluate the data using your decision rule and make a decision that there is
  - enough evidence to reject the assumption that the coin is fair
  - not enough evidence to reject the assumption that the coin is fair.

## Types of Errors

You used a decision rule to make a decision, but was the decision correct?

DECISION	ACTUAL	
	$H_0$ Is True	$H_0$ Is False
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

72

Recall that you start by assuming that the coin is fair.

The probability of a Type I error, often denoted  $\alpha$ , is the probability that you reject the null hypothesis when it is true. It is also called the significance level of a test. In the

- legal example, it is the probability that you conclude the person is guilty when he or she is innocent
- coin example, it is the probability that you conclude the coin is not fair when it is fair.

The probability of a Type II error, often denoted  $\beta$ , is the probability that you fail to reject the null hypothesis when it is false. In the

- legal example, it is the probability that you fail to find the person guilty when he or she is guilty
- coin example, it is the probability that you fail to find the coin is not fair when it is not fair.

The power of a statistical test is equal to  $1-\beta$ , where  $\beta$  is the Type II error rate. This is the probability that you correctly reject the null hypothesis.

## Modified Coin Experiment

Flip a fair coin 100 times and decide whether it is fair.

<b>55 Heads 45 Tails</b>  $p\text{-value}=.37$	<b>40 Heads 60 Tails</b>  $p\text{-value}=.06$
<b>37 Heads 63 Tails</b>  $p\text{-value}=.01$	<b>15 Heads 85 Tails</b>  $p\text{-value}<.001$

73

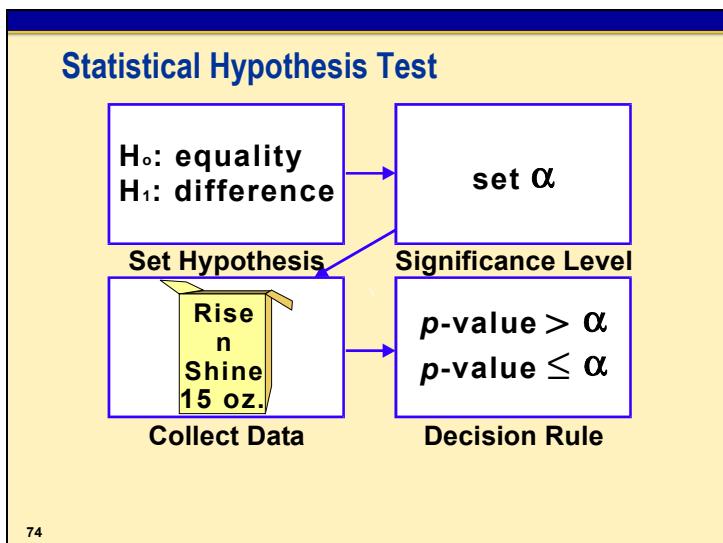
If you flip a coin 100 times and count the number of heads, you do not doubt that the coin is fair if you observe exactly 50 heads. However, you might be

- somewhat skeptical that the coin is fair if you observe 40 or 60 heads
- even more skeptical that the coin is fair if you observe 37 or 63 heads
- highly skeptical that the coin is fair if you observe 15 or 85 heads.

In this situation, the greater the difference between the number of heads and tails, the more evidence you have that the coin is not fair.

A *p*-value measures the probability of observing a value as extreme or more extreme than the one observed. For example, if your null hypothesis is that the coin is fair and you observe 40 heads (60 tails), the *p*-value is the probability of observing a difference in the number of heads and tails of 20 or more from a fair coin tossed 100 times.

If the *p*-value is large, you would often see a difference this large in experiments with a fair coin. If the *p*-value is small, however, you would rarely see differences this large from a fair coin. In the latter situation, you have evidence that the coin is not fair.



74

In statistics,

1. the null hypothesis, denoted  $H_0$ , is your initial assumption and is usually one of equality or no relationship. For the cereal example,  $H_0$  is that the mean population weight for Rise n Shine cereal is 15 ounces.
2. the significance level is usually denoted by  $\alpha$ , the Type I error rate.
3. the strength of the evidence is measured by a  $p$ -value.
4. the decision rule is
  - fail to reject the null hypothesis if the  $p$ -value is greater than or equal to  $\alpha$
  - reject the null hypothesis if the  $p$ -value is less than  $\alpha$ .



You never conclude that two things are the same or have no relationship; you can only fail to show a difference or a relationship.

## Comparing $\alpha$ and the $p$ -Value

In general, you

- reject the null hypothesis if  $p$ -value  $< \alpha$
- fail to reject the null hypothesis if  $p$ -value  $\geq \alpha$ .

75

It is important to clarify that

- $\alpha$ , the probability of Type I error, is specified by the experimenter before collecting data
- the  $p$ -value is calculated from the collected data.

In most statistical hypothesis tests, you compare  $\alpha$  and the associated  $p$ -value to make a decision.

Remember,  $\alpha$  is set ahead of time based on the circumstances of the experiment. The level of  $\alpha$  is chosen based on the cost of making a Type I error. It is also a function of your knowledge of the data and theoretical considerations.

For the cereal example,  $\alpha$  was set to 0.05, based on the consequences of making a Type I error (if you conclude that the mean cereal weight is not 15 ounces when it really is 15 ounces). For example, if making a Type I error causes serious problems, you might want to lower your significance level.

## Performing a Hypothesis Test

To test the null hypothesis  $H_0: \mu = \mu_0$ , SAS software calculates the  $t$  statistic

$$t = \frac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$$

77

For the cereal example,  $\mu_0$  is the hypothesized value of 15 ounces,  $\bar{x}$  is the sample mean weight of the cereal, and  $s_{\bar{x}}$  is the standard error of the mean.

- This statistic measures how far  $\bar{x}$  is from the hypothesized mean.
- To reject a test with this statistic, the  $t$  statistic should be much higher or lower than 0 and have a small corresponding  $p$ -value.
- The results of this test are valid if the distribution of sample means is normally distributed.

### Performing a Hypothesis Test

The null hypothesis is rejected when the actual value of interest is either less than or greater than the hypothesized value.

$$H_0: \mu = 15.00$$

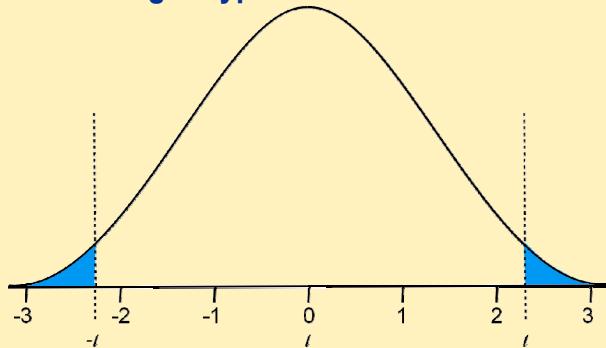
$$H_1: \mu \neq 15.00$$

$$t = \frac{(\bar{x} - 15)}{s_{\bar{x}}}$$

78

For the cereal example, if discrepancies in either direction (above 15 ounces or below 15 ounces) are of interest, then the cereal manufacturer would conduct a two-sided test of the hypothesis.

### Performing a Hypothesis Test



The  $t$  statistic can be positive or negative.

79

For a two-sided test of hypothesis, the rejection region is contained in both tails of the  $t$  distribution. If the  $t$  statistic falls in the rejection region, then you reject the null hypothesis. Otherwise, you fail to reject the null hypothesis.

The area in each of the tails corresponds to  $\frac{\alpha}{2}$  or 2.5%.

It is also possible to have a one-sided test of hypothesis where the question is whether the mean of the population is greater than or less than a certain amount. For example, a consumer advocacy group might suspect that Rise n Shine is not giving consumers enough cereal. Their hypothesis would therefore be as follows:

$$H_0: \mu \geq 15$$

$$H_1: \mu < 15$$

The *p*-value for a one-sided test of hypothesis is half the *p*-value for a two-sided test of hypothesis. Therefore, in order to perform a one-sided test, you must do the following:

1. Check to see if the *t* statistic is the right sign (negative if  $H_1$  is  $<$ , positive if  $H_1$  is  $>$ ).
2. If the sign of the *t* statistic is correct, then divide the reported *p*-value by 2.
3. Compare the new *p*-value to alpha.



## Hypothesis Testing

Example: Use the MU0= option in the UNIVARIATE procedure to test the hypothesis that the mean of the cereal example is equal to 15 ounces.

```
/* c1demo06 */
proc univariate data=sasuser.b_rise mu0=15;
  var weight;
  title 'Testing Whether the Mean of Cereal = 15 Ounces';
run;
```

Selected PROC UNIVARIATE statement option:

MU0 = specifies the value of the mean or location parameter in the null hypothesis for tests of location.

Partial PROC UNIVARIATE Output

Tests for Location: Mu0=15				
Test	-Statistic-		-----p Value-----	
Student's t	t 8.566258	Pr >  t	<.0001	
Sign	M 17	Pr >=  M	<.0001	
Signed Rank	S 396	Pr >=  S	<.0001	

The  $t$  statistic and  $p$ -value are labeled Student's  $t$  and  $\text{Pr} > |t|$ , respectively.

- The  $t$  statistic value is 8.566258 and the  $p$ -value is <.0001.
- Therefore, you can reject the null hypothesis at the 0.05 level. Thus, there is enough evidence to conclude that the mean is not equal to 15 ounces.



Refer to Exercises 2 and 3 for Chapter 1 in Appendix A.

## 1.5 Chapter Summary

Statistics provide information about your data so you can answer questions and make informed decisions. The two major branches are descriptive and inferential statistics. When you analyze data, it is imperative to state the purpose(s) of the analysis, identify specific questions to be answered, identify the population of interest, determine the need for sampling, and finally evaluate the data collection process.

Descriptive statistics describe the characteristics of the data. They include measures of location, dispersion, and shape. Some measures of location are the mean, median, and percentiles. Measures of dispersion describe the variability in a set of values and include the range, interquartile range, variance, standard deviation, and coefficient of variation. Skewness and kurtosis are measures of shape and enable you to compare your data's distribution to symmetric and normal distributions respectively.

The initial stage of data analysis includes an examination of the distribution of the data. A distribution is a collection of data values arranged in order, along with the relative frequency. In a symmetric distribution, the right side of the distribution is a mirror image of the left side and the mean is equal to the median. In a skewed distribution, many data values accumulate at one end of the distribution.

Box-and-whisker plots and normal probability plots, when used in conjunction with the mean, median, skewness and kurtosis, can help determine whether the data is normally distributed.

A population is the set of all measurement values of interest. Most of the time, you cannot collect information for the entire population, so you select a sample. A sample is a subset of the population. If the sample is a random sample, it helps ensure that it is representative of the population as a whole. Descriptive statistics describe the sample's characteristics, and inferential statistics draw conclusions about the population.

A point estimate is a sample statistic used to estimate a population parameter. A point estimate does not take into account the variability of the calculated statistic. Therefore, rather than relying on the absolute accuracy of the point estimates, you use confidence intervals to estimate population parameters. A confidence interval is a range of values that you believe to contain the population parameter of interest.

Confidence intervals for the mean make the assumption that the sample means are normally distributed. This normality can be verified by assessing the normality of the data or by invoking the central limit theorem. The central limit theorem states that as the sample size becomes sufficiently large for independent random samples, the distribution of the sample means becomes approximately normal.

There are four basic steps when conducting a test of hypothesis.

1. Determine the null and alternative hypotheses. The null hypothesis,  $H_0$ , is your initial assumption and is usually one of equality or no relationship.
2. Select a significance level: the amount of evidence needed to reject the null hypothesis. The significance level is usually denoted by  $\alpha$  and is the Type I error rate. This is the probability that you incorrectly reject the null hypothesis.
3. Collect evidence. The strength of the evidence is measured by a  $p$ -value.
4. Use a decision rule to make a judgment. You fail to reject the null hypothesis if the  $p$ -value is greater than or equal to  $\alpha$ . You reject the null hypothesis if the  $p$ -value is less than  $\alpha$ .

The one-sample  $t$ -test for the mean is based on the assumption that the sample means are normally distributed. SAS automatically generates a two-tailed  $t$ -test, so care must be taken when using a one-sided test. Only the researcher/analyst knows whether a one-sided test is appropriate.

When you conduct a test of hypothesis, there are two types of errors you can make. A Type I error is when you incorrectly reject the null hypothesis. The probability of making a Type I error is denoted by  $\alpha$ . A Type II error is when you fail to reject the null hypothesis and the null hypothesis is false. The probability of making a Type II error is denoted by  $\beta$ . The power of a statistical test is equal to  $1-\beta$  and is the probability that you correctly reject the null.

```
PROC MEANS DATA=SAS-data-set <options>;
  VAR variables;
  RUN;
```

```
PROC UNIVARIATE DATA=SAS-data-set <options>;
  VAR variables;
  ID variable;
  BY variable;
  HISTOGRAM variables / <options>;
  PROBPLOT variables / <options>;
  RUN;
```

```
PROC BOXPLOT DATA=SAS-data-set <options>;
  PLOT analysis-variable*group-variable
    < / options>;
  RUN;
```

# Chapter 2 Analysis of Variance (ANOVA)

<b>2.1 One-Way ANOVA: Two Populations .....</b>	<b>2-2</b>
<b>2.2 ANOVA with More than Two Populations .....</b>	<b>2-22</b>
<b>2.3 Two-Way ANOVA with Interactions .....</b>	<b>2-48</b>
<b>2.4 Chapter Summary.....</b>	<b>2-61</b>

## 2.1 One-Way ANOVA: Two Populations

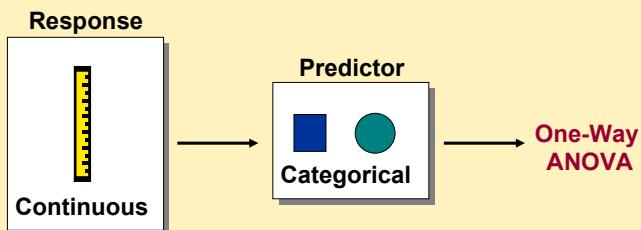
### Objectives

- Analyze differences between two population means using the GLM procedure.
- Verify the assumptions of analysis of variance.

3

### Overview

Are there any differences among the population means?



4

*Analysis of variance* (ANOVA) is a statistical technique used to compare the means of two or more groups of observations or treatments. In this section, you apply analysis of variance to examine problems where there are two treatments. For this type of problem, you have a

- continuous dependent variable, or *response* variable
- discrete independent variable also called a *predictor* or *explanatory* variable.

## Research Questions for One-Way ANOVA

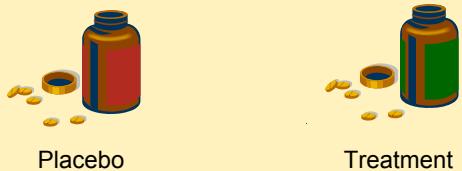
Do men, on average, have higher salaries than women?



5

## Research Questions for One-Way ANOVA

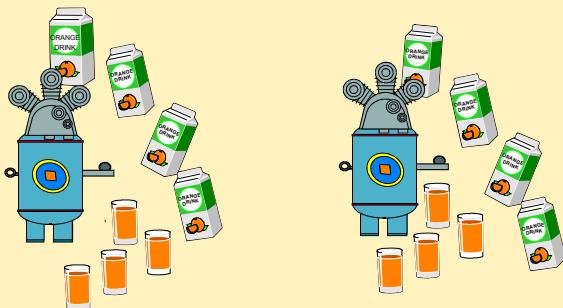
Do people in the treatment group have a higher average T cell count than people in the control group?



6

## Research Questions for One-Way ANOVA

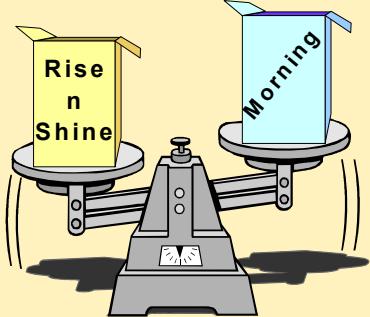
Do two different factory lines produce juice boxes with the same average amount of juice?



7

**Research Questions for One-Way ANOVA**

Do the brands Morning and Rise n Shine have the same average amount of cereal per box?



8

Example: The same manufacturer makes Rise n Shine and Morning cereal. They want to make sure their two different processes are putting the same amount of cereal in each box. Both brands should have 15 ounces of cereal per box. A random sample of both brands is selected and the number of ounces of cereal is recorded. The data is stored in a data set named sasuser.b\_cereal.

The variables in the data set are

- brand** two groups, Rise n Shine and Morning, corresponding to the two brands
- weight** weight of the cereal in ounces
- idnumber** the identification number for each cereal box.



## Descriptive Statistics across Groups

Example: Print the data in the **sasuser.b\_cereal** data set and create descriptive statistics of the two populations.

```
/* c2demo01 */
proc print data=sasuser.b_cereal (obs=15);
  title 'Partial Listing of Cereal Data';
run;
```

Part of the data is shown below.

Partial Listing of Cereal Data			
OBS	BRAND	WEIGHT	ID
1	Morning	14.9982	61469897
2	Rise n Shine	15.0136	33081197
3	Morning	15.0100	68137597
4	Rise n Shine	14.9982	37070397
5	Morning	15.0052	64608797
6	Rise n Shine	14.9930	60714297
7	Morning	14.9733	16907997
8	Rise n Shine	15.0812	9589297
9	Morning	15.0037	93891897
10	Rise n Shine	15.0418	85859397
11	Morning	14.9957	38152597
12	Rise n Shine	15.0639	99108497
13	Morning	15.0099	59666697
14	Rise n Shine	15.0613	70847197
15	Morning	14.9943	47613397

```
/* c2demo02 */
options reset=all;

proc univariate data=sasuser.b_cereal;
  class brand;
  var weight;
  probplot weight / normal
    (mu=est sigma=est color=blue w=1);
  title 'Univariate Analysis of the Cereal Data';
run;

proc sort data=sasuser.b_cereal out=b_cereal;
  by brand;
run;

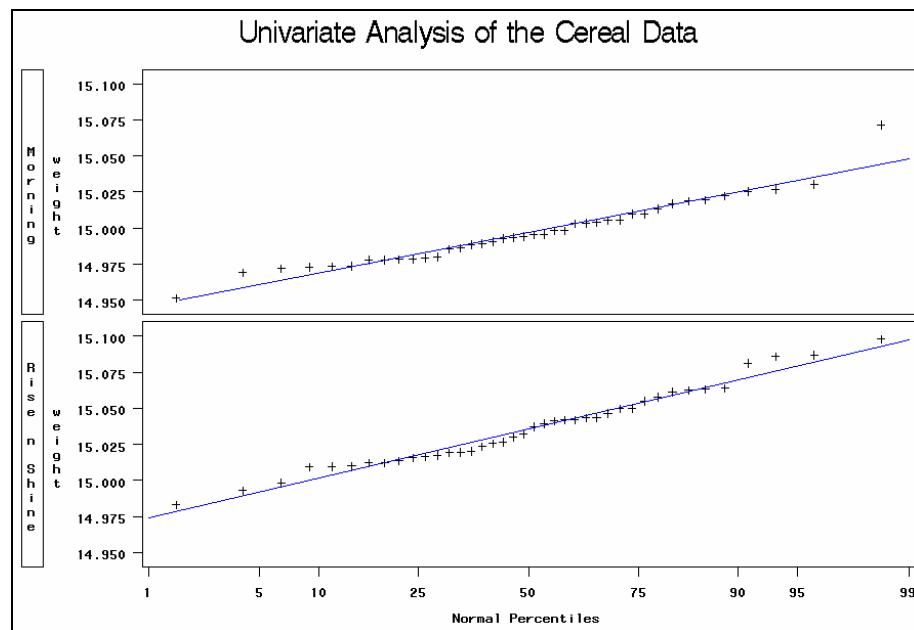
proc boxplot data=b_cereal;
  plot weight*brand / cboxes=black boxstyle=schematic;
run;
```

 Sorting is not required for the CLASS statement in PROC UNIVARIATE. For the BOXPLOT procedure, however, you must sort the data in ascending order by the second variable in the PLOT statement.

Selected PROC SORT statement option:

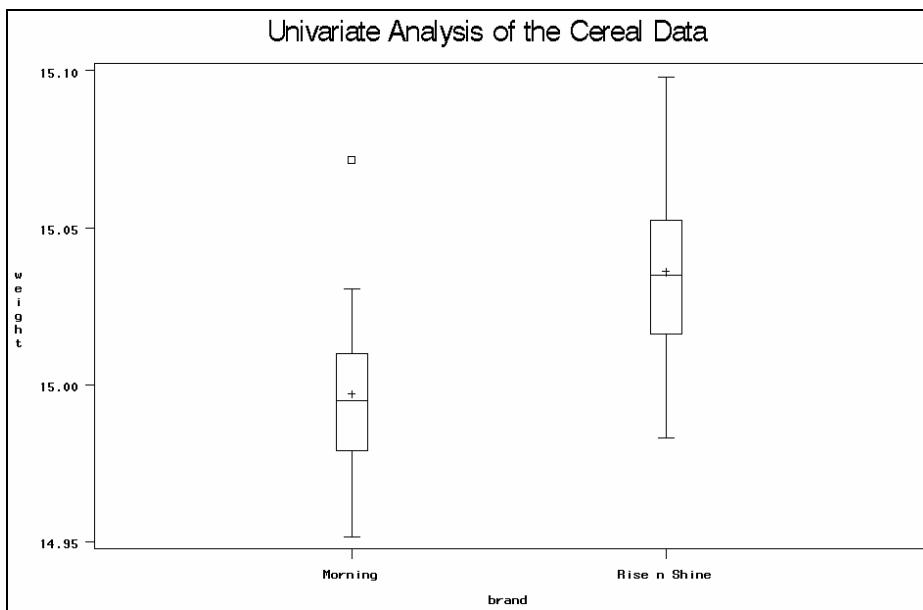
OUT= specifies a name for the output data set. If the OUT= option is omitted, then the DATA= data set is sorted and the sorted version replaces the original data set.

Partial PROC UNIVARIATE Output



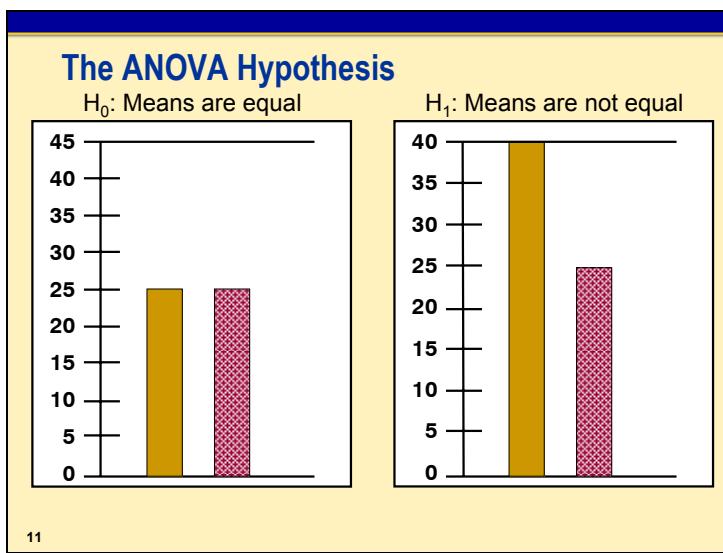
The CLASS statement in PROC UNIVARIATE produces side-by-side high-resolution plots. For full-screen plots for each value of the CLASS variable, use a BY statement instead. (When using a BY statement, the data must be sorted by the BY variable.)

The normal probability plot shows no serious departures from normality, allowing for the one extreme point in the Morning sample. There appears to be no pattern in either sample that reflects skewness or kurtosis.



The box-and-whisker plots provide further evidence that both samples are normally distributed. The outlier is also visible in the schematic box plot.

By comparing the box-and-whisker plots, you can see that the weights of the brand Rise n Shine have a larger mean and slightly more variability than Morning cereal weights.



Small differences between sample means are usually present. The objective is to determine whether these differences are significant. In other words, is the difference more than what might be expected to occur by chance?

The assumptions for ANOVA are

- independent observations

- normally distributed error terms for each treatment
- approximately equal error variances for each treatment.

### The ANOVA Model

Weight = Base Level + Brand + Unaccounted for Variation

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

12

$Y_{ik}$  the  $k^{\text{th}}$  value of the response variable for the  $i^{\text{th}}$  treatment.

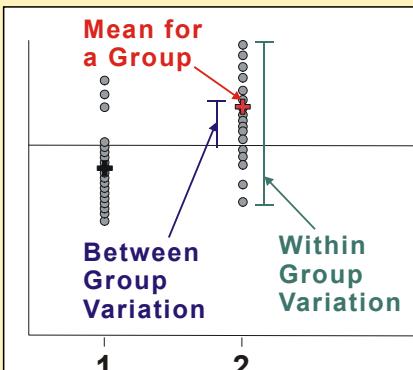
$\mu$  the overall population mean of the response, for instance cereal weight.

$\tau_i$  the difference between the population mean of the  $i^{\text{th}}$  treatment and the overall mean,  $\mu$ . This is referred to as the *effect* of treatment  $i$ .

$\varepsilon_{ik}$  the difference between the observed value of the  $k^{\text{th}}$  observation in the  $i^{\text{th}}$  group and the mean of the  $i^{\text{th}}$  group. This is called the *error term*.

 Because you are interested only in these two specific brands, **cereal** is considered fixed. In some references this would be considered a fixed effect.

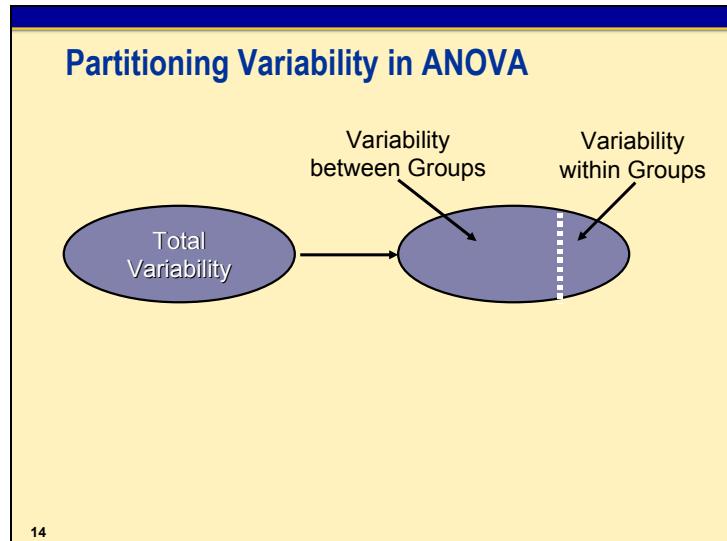
### Sums of Squares



13

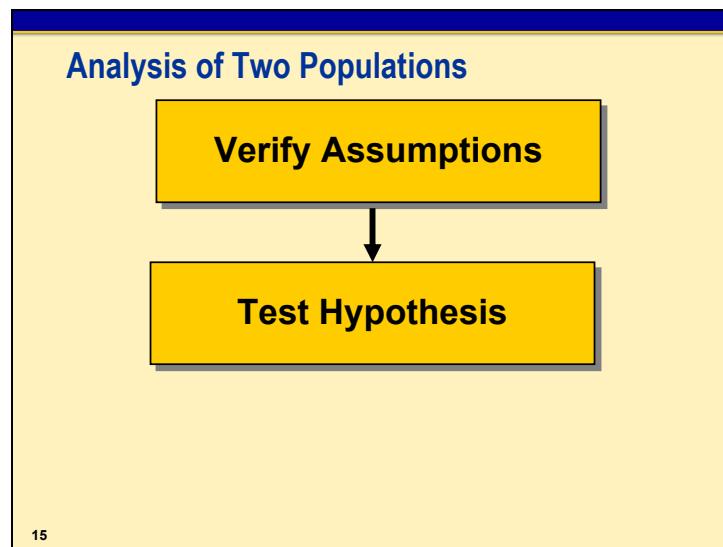
As its name implies, analysis of variance analyzes the variances of the data to determine whether there is a difference between the group means.

Between Group Variation	the sum of the squared differences between the mean for each group and the overall mean, $\sum n_i(\bar{\tau}_i - \bar{\mu})^2$ .
Within Group Variation	the sum of the squared differences between each observed value and the mean for its group, $\sum \sum (Y_{ij} - (\bar{\mu} + \bar{\tau}_i))^2$ .
Total Variation	the sum of the squared differences between each observed value and the overall mean, $\sum \sum (Y_{ij} - \bar{\mu})^2$ .

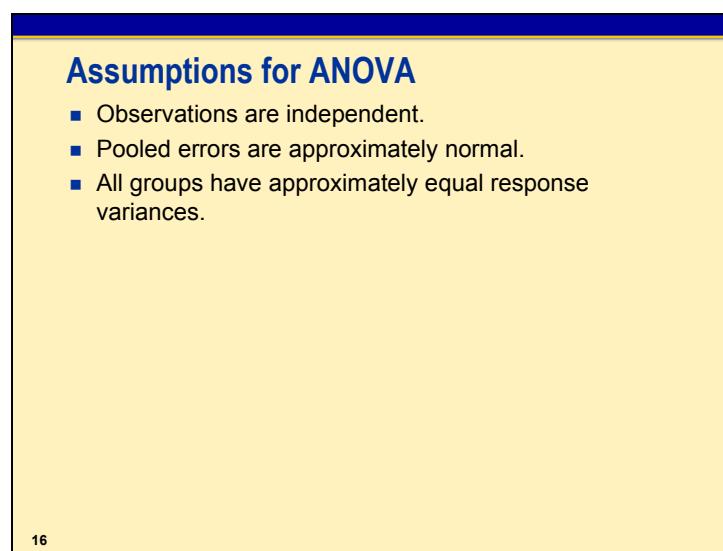


In ANOVA, the corrected total sum of squares is partitioned into two parts, the Model Sum of Squares and the Error Sum of Squares.

Model Sum of Squares (SSM)	the variability explained by the independent variable and therefore represented by the <b>between</b> treatment sums of squares.
Error Sum of Squares (SSE)	the variability not explained by the independent variable. Also referred to as <b>within</b> treatment variability or residual sum of squares.
Total Sum of Squares (SST)	the <b>overall</b> variability in the response variable. $SST = SSM + SSE$ .



15



16

Pooled error terms refer to the error terms for all groups; that is, each individual group does not have to be normally distributed as long as the errors as a whole are normally distributed.

One assumption of ANOVA is approximately equal error variances for each treatment. Although you can get an idea about the equality of variances by looking at the descriptive statistics and plots of the data, you should also consider a formal test for homogeneity of variances. The GLM procedure has a homogeneity of variance test option (HOVTEST).

The observations being independent implies that  $\epsilon_{ij}$ s in the theoretical model are independent. The independence assumption should be verified with good data collection. In some cases, the residuals can be used to verify this assumption.

## Predicted and Residual Values

The predicted value in ANOVA is the group mean.

A residual is the difference between the observed value of the response and the predicted value of the response variable.

	brand	weight	predicted	residual
1	Morning	14.9982	14.9970125	0.0011875
2	Rise n Shine	15.0136	15.03596	-0.02236
3	Morning	15.0100	14.9970125	0.0129875
4	Rise n Shine	14.9982	15.03596	-0.03776
5	Morning	15.0052	14.9970125	0.0081875
6	Rise n Shine	14.9930	15.03596	-0.04296

17

The residuals from the ANOVA are calculated as (the actual value – the predicted value). These residuals can be examined with PROC UNIVARIATE to determine normality. With a reasonably sized sample, only severe departures from normality are considered a problem.

In ANOVA with more than one predictor variable, the HOVTEST option is unavailable. In those circumstances, you can plot the residuals against their predicted values to verify that the variances are equal. The result will be a set of vertical lines equal to the number of groups. If the lines are approximately the same height, the variances are approximately equal. Descriptive statistics can also be used to determine whether the variances are equal.

## The GLM Procedure

General form of the GLM procedure:

```
PROC GLM DATA=SAS-data-set;
  CLASS variables;
  MODEL dependents=independents </ options>;
  MEANS effects </ options>;
  LSMEANS effects </ options>;
  OUTPUT OUT=SAS-data-set keyword=variable...;
RUN;
QUIT;
```

20

Selected GLM procedure statements:

- CLASS        specifies classification variables for the analysis.
- MODEL      specifies dependent and independent variables for the analysis.
- MEANS      computes unadjusted means of the dependent variable for each value of the specified effect.
- LSMEANS     produces adjusted means for the outcome variable, broken out by the variable specified and adjusting for any other explanatory variables included in the MODEL statement.
- OUTPUT      specifies an output data set that contains all variables from the input data set and variables that represent statistics from the analysis.

 PROC GLM supports RUN-group processing, which means the procedure stays active until a PROC, DATA, or QUIT statement is encountered. This enables you to submit additional statements followed by another RUN statement without resubmitting the PROC statement.

## The GPLOT Procedure

General form of the GPLOT procedure:

```
SYMBOL <options>;
AXISn <options>;
PROC GPLOT DATA=SAS-data-set;
  PLOT vertical-variable*horizontal-variable
    </ options>;
RUN;
QUIT;
```

22

The GPLOT procedure is a SAS/GPGRAPH procedure that produces scatter plots.

Selected SAS/GPGRAPH global statements:

**SYMBOL** defines the appearance of the plotting symbol and plot lines and optionally specifies the type and additional characteristics of the plot line.

**AXIS**n specifies detailed definitions of individual axis characteristics including the range of values and scaling for the axis, and the number of major and minor tick marks. The value of n can range from 1 to 99.

Selected GPLOT procedure statements:

**PLOT** specifies the vertical axis variable and the horizontal axis variable.

 PROC GPLOT supports RUN-group processing.

## Assessing ANOVA Assumptions

- Good data collection methods help ensure the independence assumption.
- The UNIVARIATE procedure can be used on data output from PROC GLM to test the assumption that pooled residuals are approximately normal.
- The GLM procedure produces a hypothesis test with the HOVTEST option. Null for this hypothesis test is that the variances are approximately equal for all populations.

23



## The GLM Procedure

Example: Test the equality of means for the **sasuser.b\_cereal** data set using PROC GLM.  
Also test for equality of variances and output the residuals for plotting.

```
/* c2demo03 */
options ls=75 ps=45;
proc glm data=sasuser.b_cereal;
  class brand;
  model weight=brand;
  means brand / hovtest;
  output out=check r=resid p=pred;
  title 'Testing for Equality of Means with PROC GLM';
run;
quit;
```

Selected MEANS statement option:

HOVTEST performs Levene's test for homogeneity (equality) of variances. The null hypothesis for this test is that the variances are equal. Levene's test is the default.

```
options reset=all;

proc gplot data=check;
  plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
  symbol v=star h=3pct;
  axis1 w=2 major=(w=2) minor=none offset=(10pct);
  axis2 w=2 major=(w=2) minor=none;
  title 'Plot of Residuals vs. Predicted Values for Cereal Data Set';
run;
quit;
```

Selected PLOT statement options:

HAXIS= associates an AXIS statement with the horizontal axis.

VAXIS= associates an AXIS statement with the vertical axis.

Selected SYMBOL statement options:

V= specifies the plotting symbol.

H= specifies the height of the plotting symbol in CELLS (the default), CM (centimeters), IN (inches), or PCT (percent).

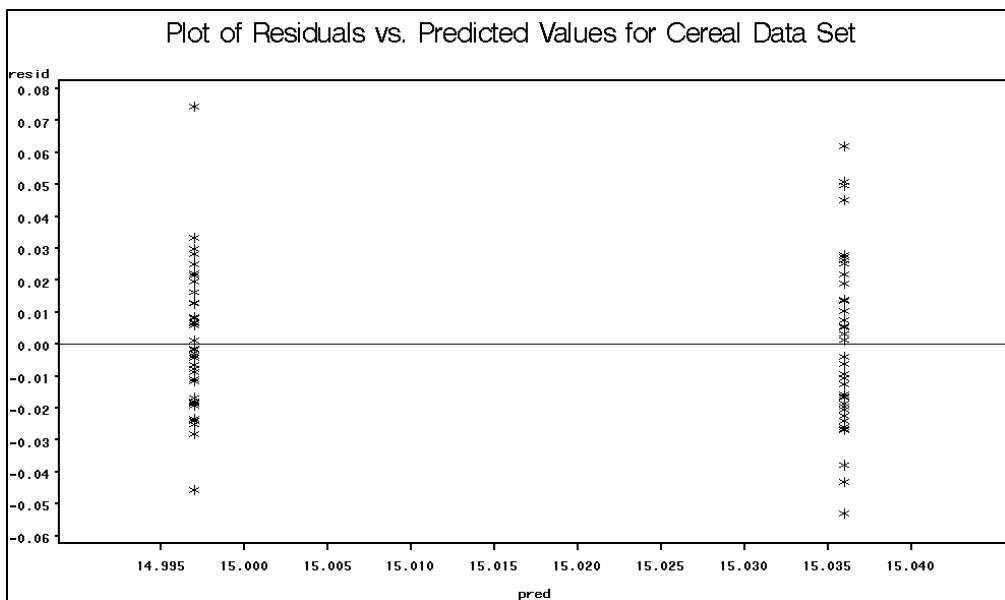
Selected AXIS statement options:

W= specifies the thickness of the axis line.

MAJOR= defines the appearance of major tick marks.

MINOR= defines the appearance of minor tick marks.

## PROC GPLOT Output



The graph above is a plot of the residuals versus the fitted values from the ANOVA model. Essentially, you are looking for a random scatter within each group. Any patterns or trends in this plot can indicate model assumption violations.

-  If you do not have SAS/GPLOT software, a similar graph can be generated using PROC PLOT to provide the same information as PROC GPLOT.

```
proc univariate data=check normal;
  var resid;
  histogram / normal;
  probplot / normal (mu=est sigma=est color=blue w=1);
  title;
run;
```

Selected PROC UNIVARIATE statement option:

**NORMAL** produces goodness-of-fit tests for the normal distribution. Be cautious when interpreting the results, as these tests tend to be sensitive to the sample size.

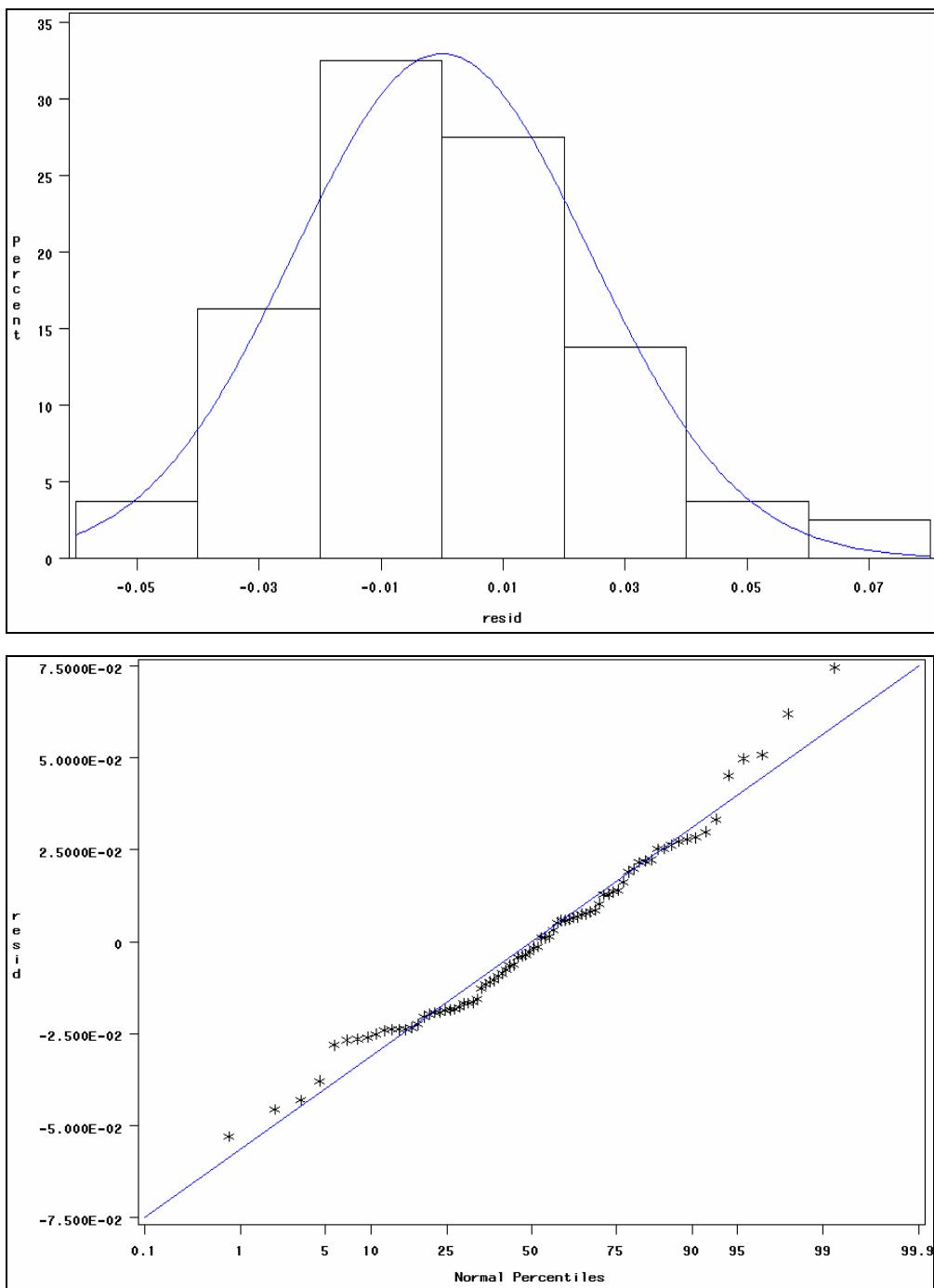
Output from the UNIVARIATE procedure helps to verify the assumption of normality of the residuals. The normal probability plot, the histogram with the normal curve superimposed on it, and the statistics found in the Goodness-of-Fit Tests for Normal Distribution table do not indicate any major departures from normality.

Partial PROC UNIVARIATE Output

The UNIVARIATE Procedure			
Variable: resid			
Moments			
N	80	Sum Weights	80
Mean	0	Sum Observations	0
Std Deviation	0.02423107	Variance	0.00058714
Skewness	0.56777406	Kurtosis	0.54402045
Uncorrected SS	0.04638442	Corrected SS	0.04638442
Coeff Variation	.	Std Error Mean	0.00270912
Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	0.02423
Median	-0.00211	Variance	0.0005871
Mode	-0.02376	Range	0.12745
		Interquartile Range	0.03233

NOTE: The mode displayed is the smallest of 4 modes with a count of 2.

The high-resolution plots provide evidence that the residuals are normally distributed.



## Partial PROC UNIVARIATE Output

The UNIVARIATE Procedure Fitted Distribution for resid					
Parameters for Normal Distribution					
Parameter	Symbol	Estimate			
Mean	Mu	0			
Std Dev	Sigma	0.024231			

Goodness-of-Fit Tests for Normal Distribution					
Test	---Statistic---	-----p Value-----			
Kolmogorov-Smirnov	D 0.07711238	Pr > D >0.150			
Cramer-von Mises	W-Sq 0.08755262	Pr > W-Sq 0.167			
Anderson-Darling	A-Sq 0.61754096	Pr > A-Sq 0.105			

The output below is the result of the HOVTEST option in the MEANS statement. Levene's test for homogeneity of variances is the default. The null hypothesis is that the variances for the treatments are equal. The *p*-value indicates that you do not reject the null hypothesis. Therefore, the assumption of equal variances appears to be satisfied.

Testing for Equality of Means with PROC GLM					
The GLM Procedure					
Levene's Test for Homogeneity of weight Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
brand	1	9.237E-7	9.237E-7	1.12	0.2942
Error	78	0.000065	8.283E-7		

 If at this point you determined that the variances were not equal, you would add the WELCH option to the MEANS statement. This requests Welch's (1951) variance-weighted one-way ANOVA. This alternative to the usual ANOVA is robust to the assumption of equal variances. This is similar to the unequal variance *t*-test for two populations. See Appendix B, "Additional Topics," for more information.

After you are satisfied that the assumptions are met, turn your attention to the first page of the PROC GLM output, which specifies the number of levels, the values of the class variable, and the number of observations read versus the number of observations used. If any row has missing data for a predictor or response variable, that row is dropped from the analysis.

Testing for Equality of Means with PROC GLM

The GLM Procedure

Class Level Information

Class	Levels	Values
brand	2	Morning Rise n Shine
Number of observations read		80
Number of observations used		80

The second page of the output contains all of the information that is needed to test the equality of the treatment means.

Testing for Equality of Means with PROC GLM

The GLM Procedure

Dependent Variable: weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.03033816	0.03033816	51.02	<.0001
Error	78	0.04638442	0.00059467		
Corrected Total	79	0.07672257			

R-Square	Coeff Var	Root MSE	weight Mean
0.395427	0.162394	0.024386	15.01649

Source	DF	Type I SS	Mean Square	F Value	Pr > F
brand	1	0.03033816	0.03033816	51.02	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
brand	1	0.03033816	0.03033816	51.02	<.0001

This output is divided into three parts:

- the analysis of variance table
- descriptive information
- information about the class variable in the model

Look at each of these parts separately.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.03033816	0.03033816	51.02	<.0001
Error	78	0.04638442	0.00059467		
Corrected Total	79	0.07672257			

In general, *degrees of freedom* (DF) can be thought of as the number of independent pieces of information.

- Model DF is the number of treatments minus 1.
- Corrected total DF is the sample size minus 1.

*Mean squares* are calculated by taking sums of squares and dividing by the corresponding degrees of freedom.

- Mean square for error (MSE) is an estimate of  $\sigma^2$ , the constant variance assumed for all treatments.
- If  $\mu_1 = \mu_2$ , the mean square for the model (MSM) is also an estimate of  $\sigma^2$ .
- If  $\mu_1 \neq \mu_2$ , MSM estimates  $\sigma^2$  plus a positive constant.
- $$F = \frac{MSM}{MSE}$$
.

Based on the above, if the *F* statistic is significantly larger than 1, it supports rejecting the null hypothesis, concluding that the treatment means are not equal.

The *F* statistic and corresponding *p*-value are reported in the analysis of variance table. Because the reported *p*-value is less than 0.0001, you conclude that there is a statistically significant difference between the means.

R-Square	Coeff Var	Root MSE	weight Mean
0.395427	0.162394	0.024386	15.01649

The *coefficient of determination*,  $R^2$ , denoted in this table as R-Square, is a measure of the proportion of variability explained by the independent variables in the analysis. This statistic is calculated as

$$R^2 = \frac{SSM}{SST}$$

The value of  $R^2$  is between 0 and 1. The value is

- close to 0 if the independent variables do not explain much variability in the data
- close to 1 if the independent variables explain a relatively large proportion of variability in the data.

Although values of  $R^2$  closer to 1 are preferred, judging the magnitude of  $R^2$  depends on the context of the problem.

The coefficient of variation (denoted Coeff Var) expresses the root MSE (the estimate of the standard deviation for all treatments) as a percent of the mean. It is a unitless measure that is useful in comparing the variability of two sets of data with different units of measure.

The weight Mean is the mean of all of the data values in the variable **weight** without regard to **brand**.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
brand	1	0.03033816	0.03033816	51.02	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
brand	1	0.03033816	0.03033816	51.02	<.0001

For a one-way analysis of variance (only one classification variable), the information about the class variable in the model is an exact duplicate of the model line of the analysis of variance table.



Refer to Exercise 1 for Chapter 2 in Appendix A.

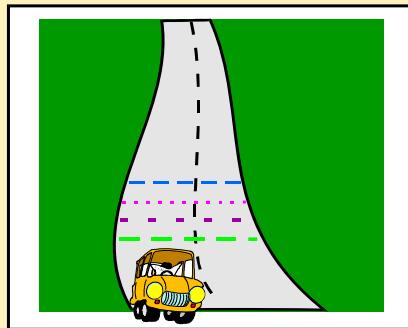
## 2.2 ANOVA with More than Two Populations

### Objectives

- Recognize the difference between a completely randomized design and a randomized block design.
- Differentiate between observed data and designed experiments.
- Analyze data from the different types of designs using the GLM procedure.

29

### Defining the Objectives



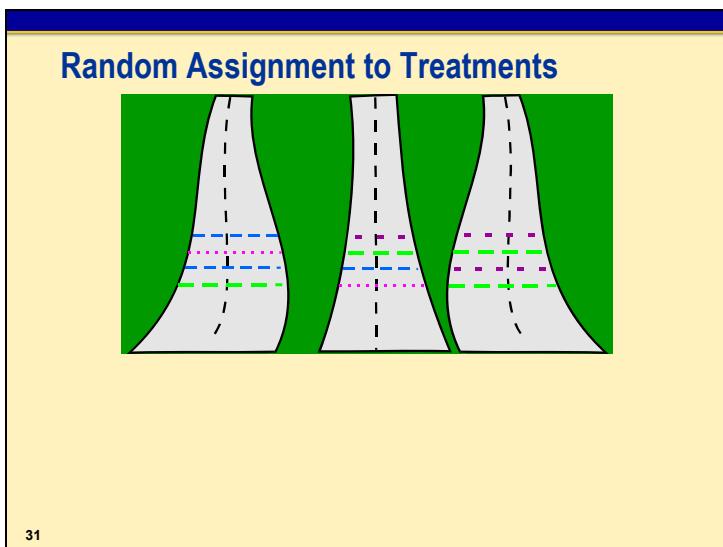
30

Question: Which paint formula is the brightest on the town roads?

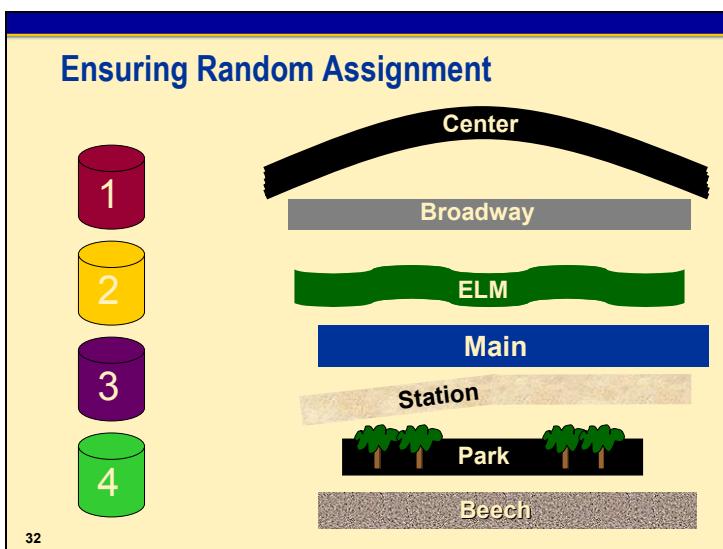
Population: The seven busiest roads in town

In this case, you have a designed experiment. Treatments are assigned, and observed values of the response are recorded. It is also possible to have observed data, where treatments are not assigned, but instead observed from the individuals in the sample.

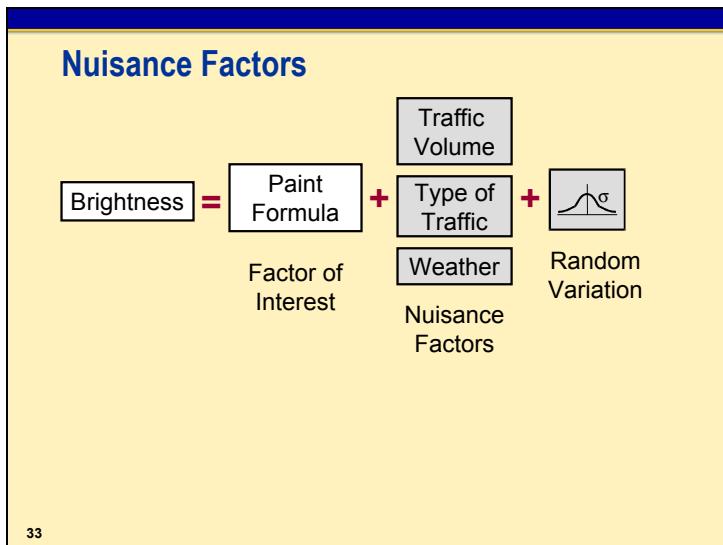
The question is specific, indicating that you are interested only in the effect paint formula has on brightness. The target population is also specific, indicating that inferences are only to be drawn on the seven busiest roads in the town.



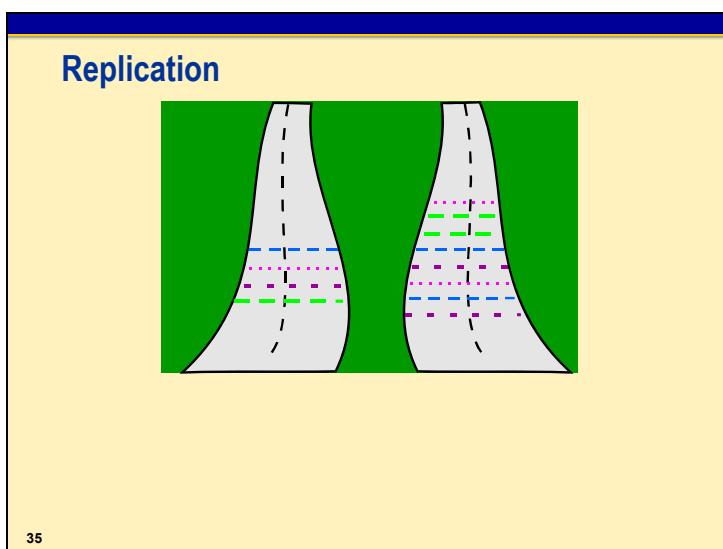
Example: You have identified the seven roads to paint and the four paint formulas to test. You plan to paint 4 stripes of paint on each road, a total of 28 stripes. One paint formula is randomly assigned to each of the 28 stripes.



Careful planning is required to ensure that the paints are randomly assigned to each of the 28 stripes. Appendix E, "Randomization Technique," contains a possible program to accomplish this task.



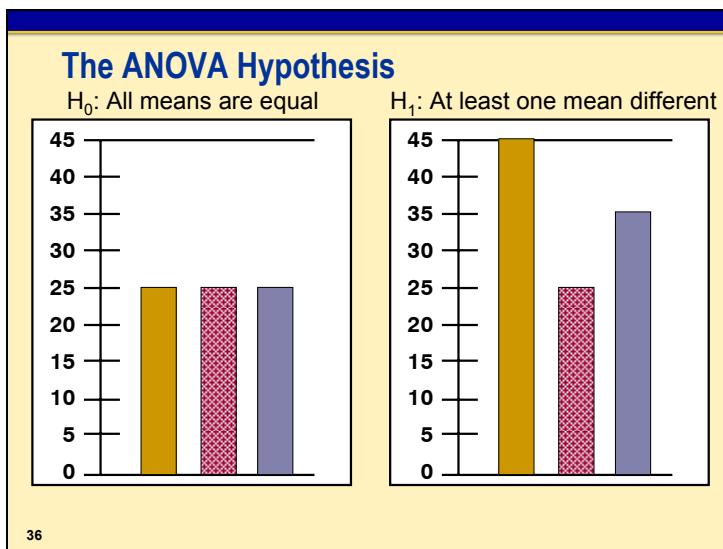
Factors that can affect the outcome but are not of interest in the experiment are called *nuisance factors*. The variation due to nuisance factors becomes part of the random variation.



A *replication* occurs when you assign each treatment to more than one experimental unit.

In the picture, there is one stripe of each paint formula applied to the road on the left. If you are concerned that a sample size of one for each treatment is insufficient, then you might consider dividing each stripe into two pieces and measuring the brightness of each piece. You reason that this gives you two observations, or replicates, for each treatment. What is wrong with this approach?

You cannot apply different treatments (paint formulas) to part of a stripe of paint, but only to each stripe. By dividing the stripes and using each piece as an experimental unit, you have done pseudo-replication but not true replication. To have true replication you would have to paint more stripes as shown in the picture on the road on the right.

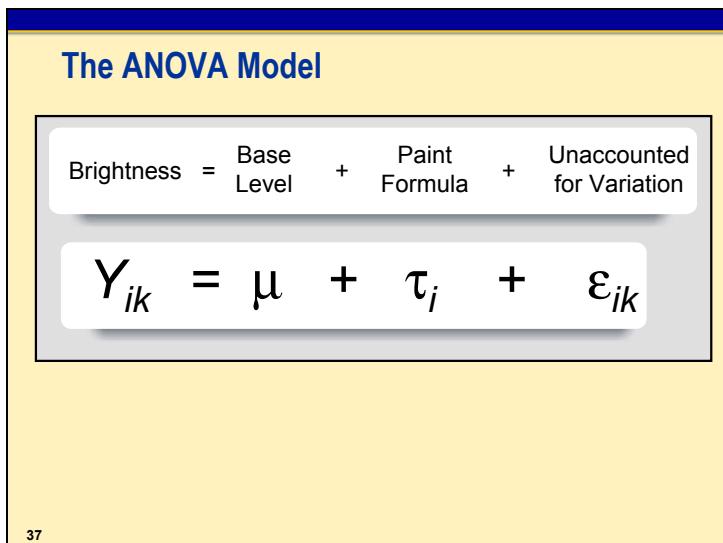


The basic concepts are the same when analyzing more than two populations as they are when analyzing two populations. The model and its assumptions are identical.

Consider the experiment to determine the best paint formula for roads with a completely randomized design. You want to determine whether the brightness of the paint is significantly different for the various paint formulas. There are seven roads, and four paint formulas are randomly assigned to each road.

Recall that the objective is to determine whether there are differences between population means. Now, with more than two populations, you are testing the hypothesis

- $H_0$ : all means are equal
- $H_1$ : at least one mean is different from one of the other means.



The model is the same as ANOVA for two treatments.



## Analysis of More than Two Populations

Example: Analyze the road paint data stored in the **sasuser.b\_roads** data set.

The variables in the data set are

**road** the name of the road

**paint** the paint formula used

**bright** the brightness of the paint after one month on the road (candellas/m<sup>2</sup>).

Print the data set.

```
/* c2demo04 */
proc print data=sasuser.b_roads;
  title 'Paint Data';
run;
```

Paint Data			
Obs	road	paint	bright
1	Center St.	1	43
2	Broadway	1	46
3	Main St.	1	47
4	Main St.	3	54
5	Elm St.	1	55
6	Station Rd.	1	56
7	Center St.	1	59
8	Center St.	4	61
9	Main St.	3	62
10	Center St.	4	62
11	Park Dr.	3	63
12	Main St.	2	64
13	Park Dr.	1	64
14	Broadway	4	64
15	Broadway	2	64
16	Broadway	3	65
17	Station Rd.	3	67
18	Station Rd.	3	67
19	Elm St.	3	68
20	Beech St.	4	71
21	Elm St.	4	72
22	Beech St.	2	75
23	Beech St.	4	75
24	Beech St.	2	76
25	Park Dr.	4	77
26	Elm St.	2	79
27	Station Rd.	2	79
28	Park Dr.	2	84

Initially, you want to examine the data to identify any unusual values and get a general idea about the distribution of the data. The UNIVARIATE procedure provides much of the information needed. (The PROC UNIVARIATE output is not shown in the text.)

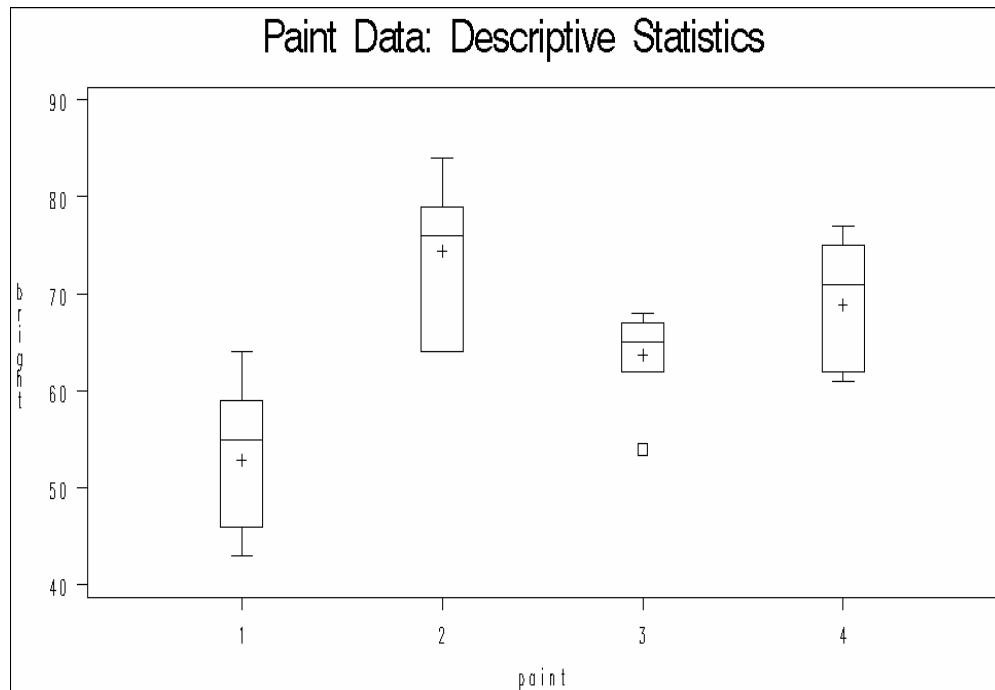
```
/* c2demo05 */
goptions reset=all;

proc univariate data=sasuser.b_roads;
  class paint;
  var bright;
  title 'Paint Data: Descriptive Statistics';
run;

proc sort data=sasuser.b_roads out=b_roads;
  by paint;
run;

proc boxplot data=b_roads;
  plot bright*paint / cboxes=black boxstyle=schematic;
run;
```

PROC BOXPLOT Output



There do not appear to be any unusual data values. There do appear to be differences between the mean brightness for the different types of paint. Specifically, paint formula 1 seems to have lower brightness than the other paint formulas. But are the differences more than could reasonably occur by chance alone? In other words, are the differences statistically significant?

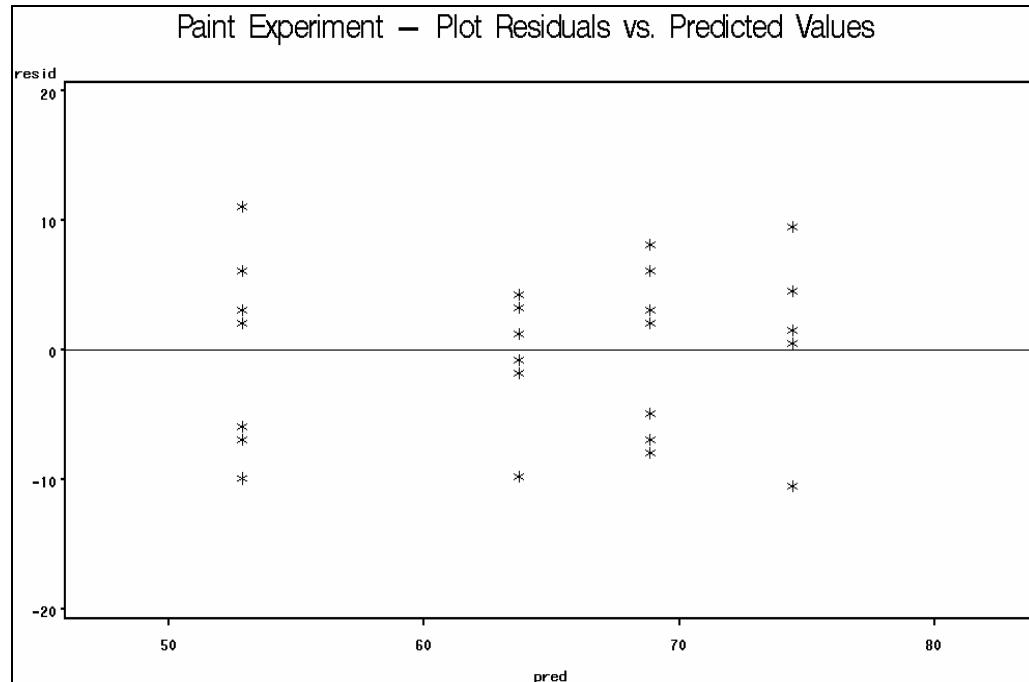
You can use the GLM procedure to test the null hypothesis that the means are equal. This program runs PROC GLM and also uses the UNIVARIATE and GPLOT procedures to check the assumptions of the ANOVA model.

```
/* c2demo06 */
proc glm data=sasuser.b_roads;
  class paint;
  model bright=paint;
  means paint / hovtest;
  output out=check r=resid p=pred;
  title 'Paint Data: Test Differences Between Means';
run;

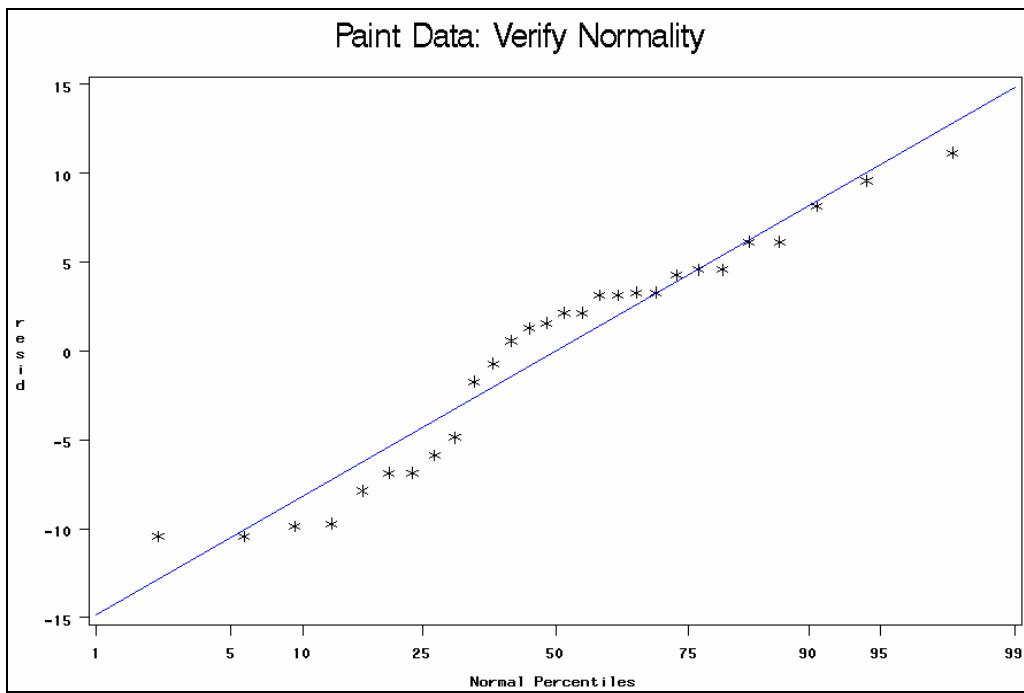
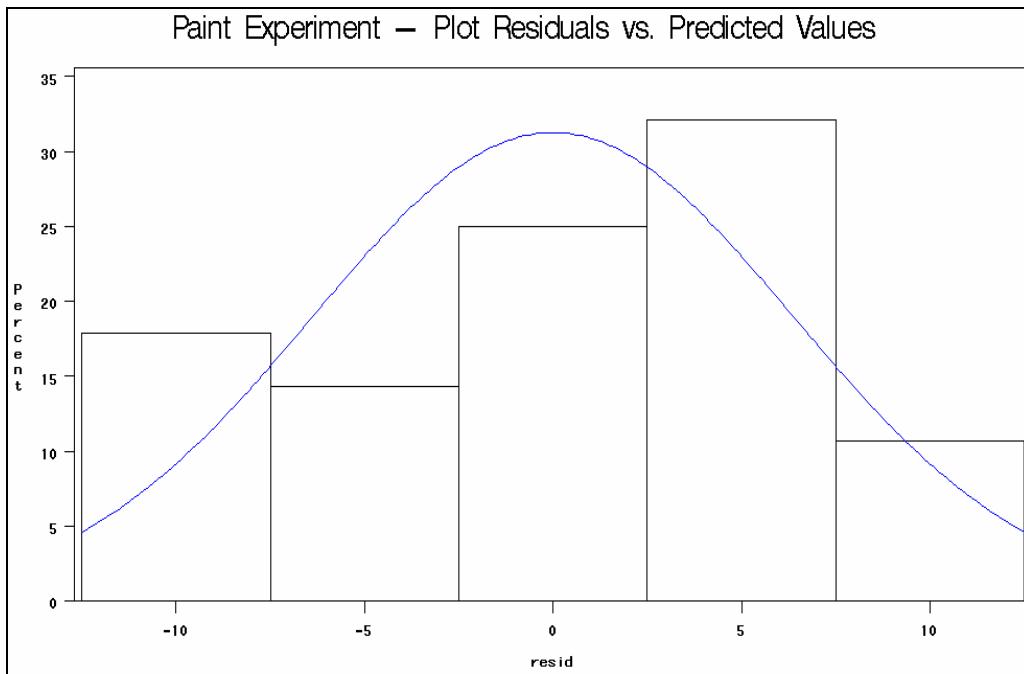
proc univariate data=check;
  var resid;
  histogram / normal;
  probplot resid / normal (mu=est sigma=est w=1);
  title 'Paint Data: Verify Normality';
run;

proc gplot data=check;
  plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
  symbol v=star h=3pct;
  axis1 w=2 major=(w=2) minor=none offset=(10pct);
  axis2 w=2 major=(w=2) minor=none;
  title 'Paint Data: Verify Assumptions';
run;
quit;
```

Based on the plot of the residuals, there do not appear to be any extreme violations of the assumptions.



The histogram and the normal probability plot shown below do not indicate any severe departure from the normality assumption.



The three statistics found in the Goodness-of-Fit Tests for Normal Distribution table provide mixed signals. Two of the normality tests are significant (0.041 and 0.049) at the 5 percent level of significance. Remember that these tests for normality should not be used exclusively to validate the normality assumption.

#### Partial PROC UNIVARIATE Output

Paint Data: Verify Normality			
The UNIVARIATE Procedure			
Variable: resid			
Moments			
N	28	Sum Weights	28
Mean	0	Sum Observations	0
Std Deviation	6.37953076	Variance	40.6984127
Skewness	-0.2920394	Kurtosis	-0.9980851
Uncorrected SS	1098.85714	Corrected SS	1098.85714
Coeff Variation	.	Std Error Mean	1.20561799
Basic Statistical Measures			
Location		Variability	
Mean	0.0000	Std Deviation	6.37953
Median	1.8571	Variance	40.69841
Mode	-10.4286	Range	21.57143
		Interquartile Range	10.78571
NOTE: The mode displayed is the smallest of 3 modes with a count of 2.			
Goodness-of-Fit Tests for Normal Distribution			
Test	---	Statistic-----	p Value-----
Kolmogorov-Smirnov	D	0.15128939	Pr > D 0.097
Cramer-von Mises	W-Sq	0.13197489	Pr > W-Sq 0.041
Anderson-Darling	A-Sq	0.73596819	Pr > A-Sq 0.049

After reviewing this information regarding the residuals, look at the part of the PROC GLM output that shows Levene's test for equality of variances.

Paint Data: Test Differences Between Means					
The GLM Procedure					
Levene's Test for Homogeneity of bright Variance					
ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
paint	3	4505.0	1501.7	0.97	0.4224
Error	24	37090.8	1545.5		

The *p*-value of 0.4224 indicates that you do not reject the null hypothesis that the variances for the treatments are equal. Therefore, the analysis of variance procedure appears to be appropriate.

Now that you are reasonably sure that the assumptions of the ANOVA model have been met, turn your attention to the Class Level Information table and the ANOVA table.

The first page of PROC GLM output, shown below, specifies the number of levels and the values of the class variable, as well as the number of observations.

Paint Data: Test Differences Between Means			
The GLM Procedure			
Class Level Information			
Class	Levels	Values	
paint	4	1 2 3 4	
Number of Observations Read 28			
Number of Observations Used 28			

Part of the second page of the PROC GLM output is shown below.

Paint Data: Test Differences Between Means						
The GLM Procedure						
Dependent Variable: bright						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	3	1770.107143	590.035714	12.89	<.0001	
Error	24	1098.857143	45.785714			
Corrected Total	27	2868.964286				

With a *p*-value less than or equal to 0.0001, you reject the null hypothesis that all treatment means are equal.

At this point, you know there is **at least** one treatment mean that is different from one other treatment mean, but you cannot be sure which one(s) are different. Some insight can be gained by looking at the side-by-side box plots from PROC UNIVARIATE and the page of the PROC GLM output produced by the MEANS statement.

Paint Data: Test Differences Between Means			
The GLM Procedure			
Level of paint	N	-----bright-----	
1	7	52.8571429	7.69043933
2	7	74.4285714	7.67804539
3	7	63.7142857	4.82059076
4	7	68.8571429	6.46602844

It appears that paint formula 1 has lower brightness than the other formulas, and paint formula 2 results in the highest brightness. Multiple comparison techniques can be used to determine whether these are statistically significant differences.

## Multiple Comparison Methods

Comparisonwise Error Rate	Number of Comparisons	Experimentwise Error Rate
.05	1	.05
.05	3	.14
.05	6	.26
.05	10	.40

EER  $\leq 1 - (1 - \alpha)^{nc}$  where nc=number of comparisons

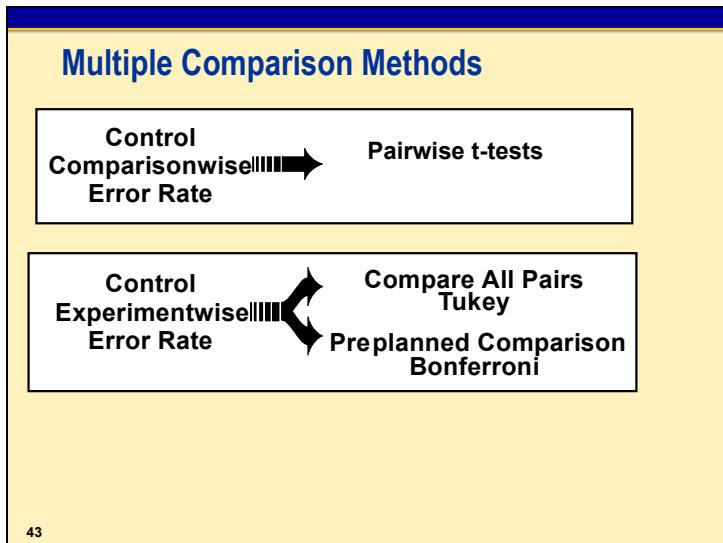
42

When you control the comparisonwise error rate (CER), you fix the level of alpha for a single comparison, without taking into consideration all the pairwise comparisons you are making.

The experimentwise error rate (EER) uses an alpha that takes into consideration all the pairwise comparisons you are making. Presuming no differences exist, the chance you falsely conclude at least one difference exists is much higher when you consider all possible comparisons.

If you want to make sure that the error rate is 0.05 for the entire set of comparisons, use a method that controls the experimentwise error rate at 0.05.

- ✍ There is some disagreement among statisticians about the need to control the experimentwise error rate.



All of these multiple comparison methods are requested with options in the LSMEANS statement of PROC GLM.

This course addresses these options:

Comparisonwise Control      ADJUST=T

Experimentwise Control      ADJUST=TUKEY or ADJUST=BONFERRONI

- There are many other options available that control the experimentwise error rate. For information about these options, see the SAS online documentation.

## Bonferroni's Method

Bonferroni's multiple comparison method

- is used only for preplanned comparisons
- adjusts for multiple comparisons by dividing the alpha level by the number of comparisons made
- ensures an experimentwise error rate less than or equal to alpha
- is the most conservative method.

44

Bonferroni's method is not generally considered appropriate for comparisons made after looking at the data, because the adjustment is made based on the number of comparisons you intend to do. If you look at the data to determine how many and what comparisons to make, you are using the data to determine the adjustment.

A conservative method tends to find fewer significant differences than might otherwise be found.

While Bonferroni's method can be used for all pairwise comparisons, Tukey's method is generally less conservative and more appropriate.

## Tukey's Multiple Comparison Method

This method is appropriate when considering pairwise comparisons only.

The experimentwise error rate is

- equal to alpha when **all** pairwise comparisons are considered
- less than alpha when **fewer** than all pairwise comparisons are considered.

45

A pairwise comparison examines the difference between two treatment means. All pairwise comparisons are all possible combinations of two treatment means.

Tukey's multiple comparison adjustment is based on conducting all pairwise comparisons and guarantees the Type I experimentwise error rate is equal to alpha for this situation. If you choose to do fewer than all pairwise comparisons, then this method is more conservative.



## Multiple Comparison Methods

Example: Use the LSMEANS statement in PROC GLM to produce comparison information on the means of the treatments.

```
/* c2demo07 */
proc glm data=sasuser.b_roads;
  class paint;
  model bright=paint;
  lsmeans paint / pdiff=all adjust=t;
  title 'Paint Data: Multiple Comparisons';
run;
quit;
```

Selected LSMEANS statement options.

PDIFF= requests  $p$ -values for the differences, the probability of seeing a difference between two means that is as large as the observed means or larger if the two population means are actually the same. You can request to compare all means using PDIFF=ALL. You can also specify which means to compare. See the documentation for LSMEANS under the GLM procedure for details.

ADJUST= specifies the adjustment method for multiple comparisons. If no adjustment method is specified, the Tukey method is used by default. The T option asks that no adjustment be made for multiple comparisons. The TUKEY option uses Tukey's adjustment method. The BON option uses the Bonferroni method. For a list of available methods, check the documentation for LSMEANS under the GLM procedure.

 The MEANS statement can be used for multiple comparisons. However, the results can be misleading if the groups that are specified have different numbers of observations.

## Partial PROC GLM Output

The GLM Procedure Least Squares Means			
paint	bright LSMEAN	LSMEAN Number	
1	52.8571429	1	
2	74.4285714	2	
3	63.7142857	3	
4	68.8571429	4	

Least Squares Means for effect paint Pr >  t  for H0: LSMean(i)=LSMean(j)				
Dependent Variable: bright				
i/j	1	2	3	4
1		<.0001	0.0062	0.0002
2	<.0001		0.0068	0.1365
3	0.0062	0.0068		0.1679
4	0.0002	0.1365	0.1679	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

The output above is from the LSMEANS statement. The note is meant as a warning that SAS is making no adjustments for the number of comparisons.

The first part of the output shows the adjusted means for each group. The second part of the output shows *p*-values from pairwise comparisons of all possible combinations of means. Notice that row 2--column 4 has the same *p*-value as row 4--column 2 because the same two means are being compared in each case. Both are displayed as a convenience to the user. Notice also that row 1--column 1, row 2--column 2, and so forth, are left blank, because it does not make any sense to compare a mean to itself.

By default, SAS only controls the comparisonwise error rate. In order to control the experimentwise error rate, you can use Bonferroni's or Tukey's method.

Example: Use the Bonferroni and Tukey methods for multiple comparisons to test differences between the treatment means for the variable **bright** in the **sasuser.b\_roads** data set.

```
proc glm data=sasuser.b_roads;
  class paint;
  model bright=paint;
  lsmeans paint / pdiff=all adjust=bon;
  title 'Paint Data: Multiple Comparisons BON';
run;
quit;
```

```
proc glm data=sasuser.b_roads;
  class paint;
  model bright=paint;
  lsmeans paint / pdiff=all adjust=tukey;
  title 'Paint Data: Multiple Comparisons TUKEY';
run;
quit;
```

## Partial PROC GLM Output

Paint Data: Multiple Comparisons BON			
The GLM Procedure			
Least Squares Means			
Adjustment for Multiple Comparisons: Bonferroni			
paint	bright	LSMEAN	Number
1	52.8571429	1	
2	74.4285714	2	
3	63.7142857	3	
4	68.8571429	4	

Least Squares Means for effect paint				
Pr >  t  for H0: LSMean(i)=LSMean(j)				
Dependent Variable: bright				
i/j	1	2	3	4
1		<.0001	0.0371	0.0011
2	<.0001		0.0407	0.8193
3	0.0371	0.0407		1.0000
4	0.0011	0.8193	1.0000	

The output indicates that paint formula 1 is different from all other formulas and that paint formulas 2 and 3 are different from each other at the .05 level. Notice that the *p*-values are larger when the experimentwise error rate is controlled.

## Partial PROC GLM Output (continued)

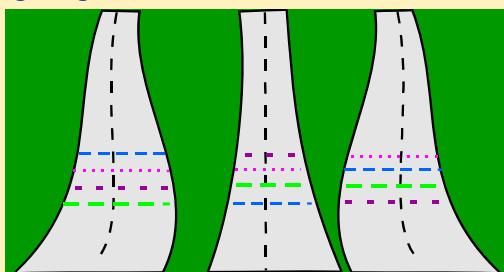
The GLM Procedure			
Least Squares Means			
Adjustment for Multiple Comparisons: Tukey			
paint	bright LSMEAN	LSMEAN Number	
1	52.8571429	1	
2	74.4285714	2	
3	63.7142857	3	
4	68.8571429	4	

Least Squares Means for effect paint				
Pr >  t  for H0: LSMean(i)=LSMean(j)				
Dependent Variable: bright				
i/j	1	2	3	4
1		<.0001	0.0294	0.0010
2	<.0001		0.0321	0.4302
3	0.0294	0.0321		0.4985
4	0.0010	0.4302	0.4985	

The significant differences using Tukey's method are the same as those with Bonferroni's method in this case. This might not always be true. Notice that the *p*-values with the Tukey adjustment method are smaller than the *p*-values with the Bonferroni method. This is because Tukey is a less conservative test.

### Assigning Treatments within Blocks



51

An experienced road paint expert might anticipate that there would be so much variability in brightness caused by the nuisance factors that the statistical test would not detect differences caused by paint formulas alone.

In order to estimate the model properly, you would need at least one stripe for each paint/road combination.

An experimental design like this is often referred to as a *randomized block design*, where **road** is the blocking factor. The variable **road** is included in the model, but you are not interested in the effect of **road**, only in controlling the variation it represents. By including **road** in the model, you could account for a nuisance factor.

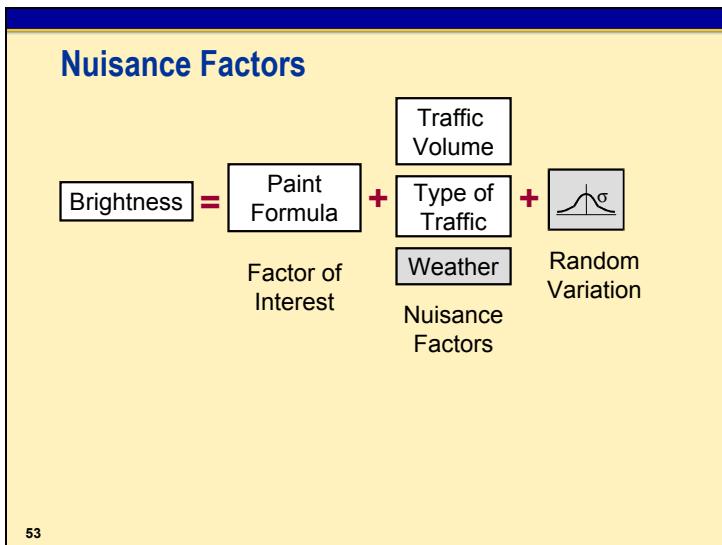
Blocking is a restriction on randomization.

### Including a Blocking Factor in the Model

$$\text{Brightness} = \text{Base Level} + \text{Road} + \text{Paint Formula} + \text{Unaccounted for Variation}$$

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$$

52



- ✍ Weather is just one of many possible nuisance factors.

## Including a Blocking Factor in the Model

Additional assumptions are as follows:

- Treatments are randomly assigned within each block.
  - The effects of the treatment factor are constant across the levels of the blocking factor.
-  In the paint example, the design is balanced, which means that there is the same number of paint stripes for every paint/road combination.

55

If the effects of the treatment factor are not constant across the levels of the blocking factor, then this condition is called *interaction*. You can still analyze the data, but be sure to collect enough data to include the interaction term in the model. Interactions are discussed in the next section.

In most randomized block designs, the blocking factor is treated as a *random effect*. Treating an effect as random changes how standard errors are calculated and can give different answers from treating it as a fixed effect (as in the example).

In this example, you have the same number of paint stripes for every paint/road combination. This is a balanced design. When treatment groups are going to be compared to each other (in other words, not to 0 or some other specified value), the results from treating the block as a fixed or random effect are exactly the same.

A model that includes both random and fixed effects is called a *mixed model* and can be analyzed with the MIXED procedure. The SAS class *Mixed Models Analyses Using the SAS® System* focuses on analyzing mixed models. *Statistics II: ANOVA and Regression* has more detail about how to analyze nonbalanced designs and data that does not meet the ANOVA assumptions, and it is a prerequisite for *Mixed Model Analyses Using the SAS® System*.

For more information on mixed models in SAS, you can also consult the SAS online documentation or the SAS Books by Users book *SAS® System for Mixed Models*, which also goes into detail about the statistical assumptions for mixed models. You can learn more about different types of experimental designs by taking the SAS class *Design of Experiments Using the ADX Interface*.



## Two-Way ANOVA

Example: The data set **sasuser.b\_roads1** is a fabricated example of data collected from randomly assigning paints within each road. Notice that each paint formula appears exactly once on each road.

```
/* c2demo08 */
proc print data=sasuser.b_roads1;
  title 'Paint Data: Including Road';
run;
```

Paint Data: Including Road			
Obs	road	paint	bright
1	Broadway	1	48
2	Main St.	1	49
3	Center St.	1	49
4	Center St.	3	56
5	Elm St.	1	57
6	Main St.	3	57
7	Station Rd.	1	58
8	Broadway	3	59
9	Beech St.	1	60
10	Park Dr.	1	61
11	Broadway	2	62
12	Center St.	4	62
13	Main St.	4	63
14	Station Rd.	3	65
15	Main St.	2	66
16	Broadway	4	66
17	Center St.	2	68
18	Elm St.	3	68
19	Beech St.	3	69
20	Park Dr.	3	70
21	Station Rd.	2	72
22	Beech St.	2	73
23	Park Dr.	2	73
24	Elm St.	4	73
25	Station Rd.	4	73
26	Elm St.	2	74
27	Beech St.	4	75
28	Park Dr.	4	78

Example: To include **road** in the model, add the variable name to the CLASS and MODEL statements.

```
/* c2demo09 */
proc glm data=sasuser.b_roads1;
  class paint road;
  model bright=paint road;
  lsmeans paint / pdiff=all adjust=tukey;
  title 'Paint Data: Multiple Comparisons Including Road';
run;
quit;
```

```
Paint Data: Multiple Comparisons Including Road
```

```
The GLM Procedure
```

```
Class Level Information
```

Class	Levels	Values
-------	--------	--------

paint	4	1 2 3 4
-------	---	---------

road	7	Beech St. Broadway Center St. Elm St. Main St. Park Dr. Station Rd.
------	---	--

```
Number of Observations Read 28
```

```
Number of Observations Used 28
```

Paint Data: Multiple Comparisons Including Road						
The GLM Procedure						
Dependent Variable: bright						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	9	1804.857143	200.539683	60.16	<.0001	
Error	18	60.000000	3.333333			
Corrected Total	27	1864.857143				
R-Square	Coeff Var	Root MSE	bright	Mean		
0.967826	2.833746	1.825742		64.42857		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
paint	3	1100.000000	366.666667	110.00	<.0001	
road	6	704.857143	117.476190	35.24	<.0001	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
paint	3	1100.000000	366.666667	110.00	<.0001	
road	6	704.857143	117.476190	35.24	<.0001	

As expected, the overall *F* test indicates that there are significant differences between the means of the different types of paint formula.

What have you gained by including **road** in the model? If you compare the estimate of the experimental error variance (MSE), you note this has decreased compared to the model that included paint only (3.33333 versus 45.78571). Depending on the magnitude of the decrease, this could affect the comparisons between the treatment means by finding more significant differences than the paint-only model.

The GLM Procedure  
 Least Squares Means  
 Adjustment for Multiple Comparisons: Tukey

paint	bright	LSMEAN	Number
	LSMEAN		
1	54.5714286	1	
2	69.7142857	2	
3	63.4285714	3	
4	70.0000000	4	

Least Squares Means for effect paint  
 $\text{Pr} > |t| \text{ for } H_0: \text{LSMean}(i) = \text{LSMean}(j)$

Dependent Variable: bright

i/j	1	2	3	4
1		<.0001	<.0001	<.0001
2	<.0001		<.0001	0.9910
3	<.0001	<.0001		<.0001
4	<.0001	0.9910	<.0001	

In this case, with the blocking factor in the model, paint formulas 2 and 4 are the only ones found not to be significantly different. Also note that all of the  $p$ -values have decreased.

In determining the usefulness of having a blocking factor (**road**) included in the model, you can consider the  $F$  value for the block. Some statisticians suggest that if this ratio is greater than 1, then the blocking factor is useful. But if the ratio is less than 1, then adding the variable is detrimental to the analysis. If you find that including the blocking factor is detrimental to the analysis, then you can exclude it from future studies, but it must be included in ANOVA models calculated with the sample that you have already collected.



Refer to Exercise 2 for Chapter 2 in Appendix A.

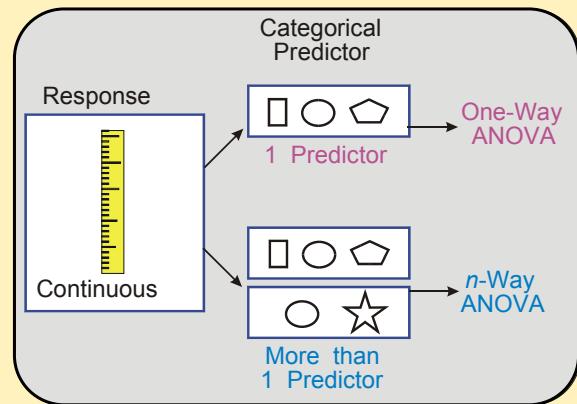
## 2.3 Two-Way ANOVA with Interactions

### Objectives

- Fit a two-way ANOVA model.
- Detect interactions between factors.
- Analyze the treatments when there is a significant interaction.

62

### *n*-Way ANOVA



63

In the previous section, you considered the case where you had one categorical predictor and a blocking variable. In this section, consider a case with two categorical predictors. In general, any time you have more than one categorical predictor variable and a continuous response variable, it is called *n*-way ANOVA. The *n* can be replaced with the number of categorical predictor variables.

The analysis for a randomized complete block design is actually a special type of *n*-way ANOVA.

## Drug Example



**DRUG** Level of drug

**DISEASE** Disease category

**BLOODP** Blood pressure

64

Data was collected in an effort to determine whether different levels of a given drug have an effect on blood pressure for people with a given disease.

## The Model

BloodP = Base Level + Disease + Drug + Drug and Disease + Unaccounted for Variation

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

65

$Y_{ijk}$  the observed **BloodP** for each subject

$\mu$  the overall population mean of the response, **BloodP**

$\alpha_i$  the effect of the  $i^{\text{th}}$  **Disease**

$\beta_j$  the effect of the  $j^{\text{th}}$  **Drug**

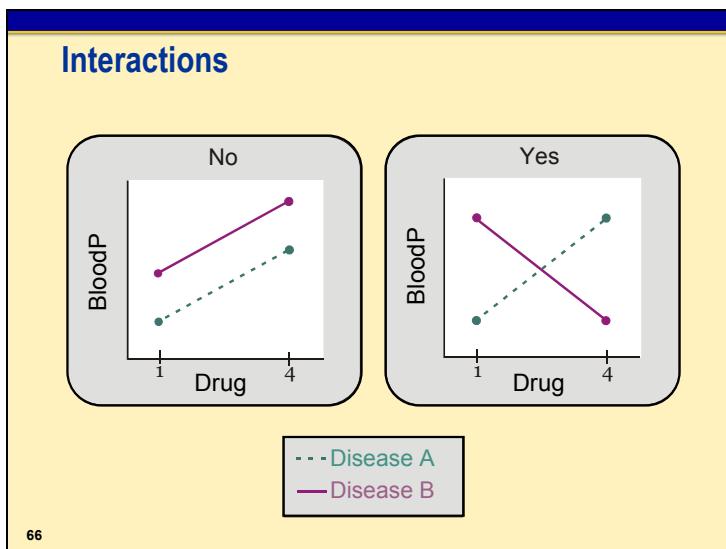
$(\alpha\beta)_{ij}$  the effect of the interaction between the  $i^{\text{th}}$  **Disease** and the  $j^{\text{th}}$  **Drug**

$\varepsilon_{ijk}$  error term, or residual

In the model it is assumed that the

- observations are independent
- data is normal for each treatment
- variances are approximately equal for each treatment.

 Verifying ANOVA assumptions with more than two variables is covered in *Statistics II: ANOVA and Regression*.



An interaction occurs when changing the level of one factor results in changing the difference between levels of the other factor.

The plots displayed above are called means plots. The average blood pressure over different levels of the drug were plotted and then connected for disease A and B.

In the left plot above, different types of disease show the same change across different levels of drug.

In the right plot, however, as the drug level increases, disease A average blood pressure **increases** and disease B **decreases**. This indicates an interaction between the variables **Drug** and **Disease**.

When you analyze an  $n$ -way ANOVA, first look at the test for interaction in the analysis of variance output to decide whether there is interaction between the factors.

If there is no interaction between the factors, the tests for the individual factor effects can be considered in the output to determine the significance/nonsignificance of these factors.

If there is interaction between the factors, the tests for the individual factor effects might be misleading due to masking of these effects by the interaction.

In the previous section, you used a block variable and a categorical predictor as effects in the model. It is generally assumed that blocks do not interact with other factors.

## Nonsignificant Interaction

Analyze the main effects with the interaction in the model.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

...or...

Delete the interaction from the model, and then analyze the main effects.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

67

When the interaction is not statistically significant, the main effects can be analyzed with the model as originally written. This is generally the method used when analyzing designed experiments.

However, even when analyzing designed experiments, some statisticians suggest that if the interaction is nonsignificant, then the interaction effect can be deleted from the model and then the main effects are analyzed. This increases the power of the main effects tests.

Neter, Kutner, Wasserman, and Nachtsheim (1996) suggest the following guidelines for when to delete the interaction from the model:

- there are fewer than 5 degrees of freedom for the error, **and**
- the mean square for the interaction divided by the error mean square is less than 2.



When you analyze data from an observational study, it is more common to delete the nonsignificant interaction from the model and then analyze the main effects.



## Two-Way ANOVA with Interactions

The data set **sasuser.b\_drug** contains the following variables:

<b>Drug</b>	level of drug
<b>Disease</b>	disease category
<b>BloodP</b>	blood pressure

 Before conducting an analysis of variance, you should explore the data.

```
options nodate nonumber;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=sasuser.b_drug;
  class drug;
  var bloodp;
  histogram / normal;
  probplot / normal(mu=est sigma=est color=red w=2);
  title 'explore b_drug, CLASS drug';
run;

proc univariate data=sasuser.b_drug;
  class disease;
  var bloodp;
  histogram / normal;
  probplot / normal(mu=est sigma=est color=red w=2);
  title 'explore b_drug, CLASS disease';
run;
```

Presume that the initial data exploration was completed (output not shown here) and that no particular concerns were noted about unusual data values or the distribution of the data. During this exploration, you determine that the sample sizes for all treatments are equal.

```
/* c2demo10 */
proc print data=sasuser.b_drug(obs=10);
  title 'Listing of sasuser.b_drug';
run;
```

Partial PROC PRINT Output

Listing of sasuser.b_drug			
Obs	Drug	Disease	BloodP
1	1	A	119.701
2	1	A	121.362
3	1	A	119.692
4	1	A	119.602
5	1	A	120.966
6	1	A	119.190
7	1	A	120.041
8	1	A	120.649
9	1	A	121.397
10	1	A	121.294

```
/* c2demo11 */
proc means data=sasuser.b_drug
  mean var std;
  class disease drug;
  var bloodp;
  title 'Selected Descriptive Statistics for sasuser.b_drug';
run;
```

Selected MEANS procedure statement:

CLASS        produces separate statistics for each combination of values in the CLASS statement.

## PROC MEANS Output

Descriptive Statistics on sasuser.b_drug Data Set					
The MEANS Procedure					
Analysis Variable : Bloodp					
Disease	Drug	N Obs	Mean	Variance	Std Dev
A	1	10	120.3892949	0.7021575	0.8379484
	2	10	135.3892949	0.7021575	0.8379484
	3	10	139.7062141	0.3932652	0.6271086
	4	10	149.9168917	1.1352520	1.0654820
B	1	10	159.8746313	2.1599596	1.4696801
	2	10	149.7218911	0.7185501	0.8476733
	3	10	140.0273514	0.2963634	0.5443927
	4	10	130.1873041	0.5856099	0.7652515
C	1	10	124.8205390	1.2613509	1.1230988
	2	10	124.8345403	1.0326229	1.0161806
	3	10	124.6400672	0.5892006	0.7675940
	4	10	125.1261637	0.7672723	0.8759408

To further explore the numerous treatments, examine the PROC MEANS output. The variable **BloodP** might increase, decrease, or stay the same for the four levels of the variable **Drug** as seen above.

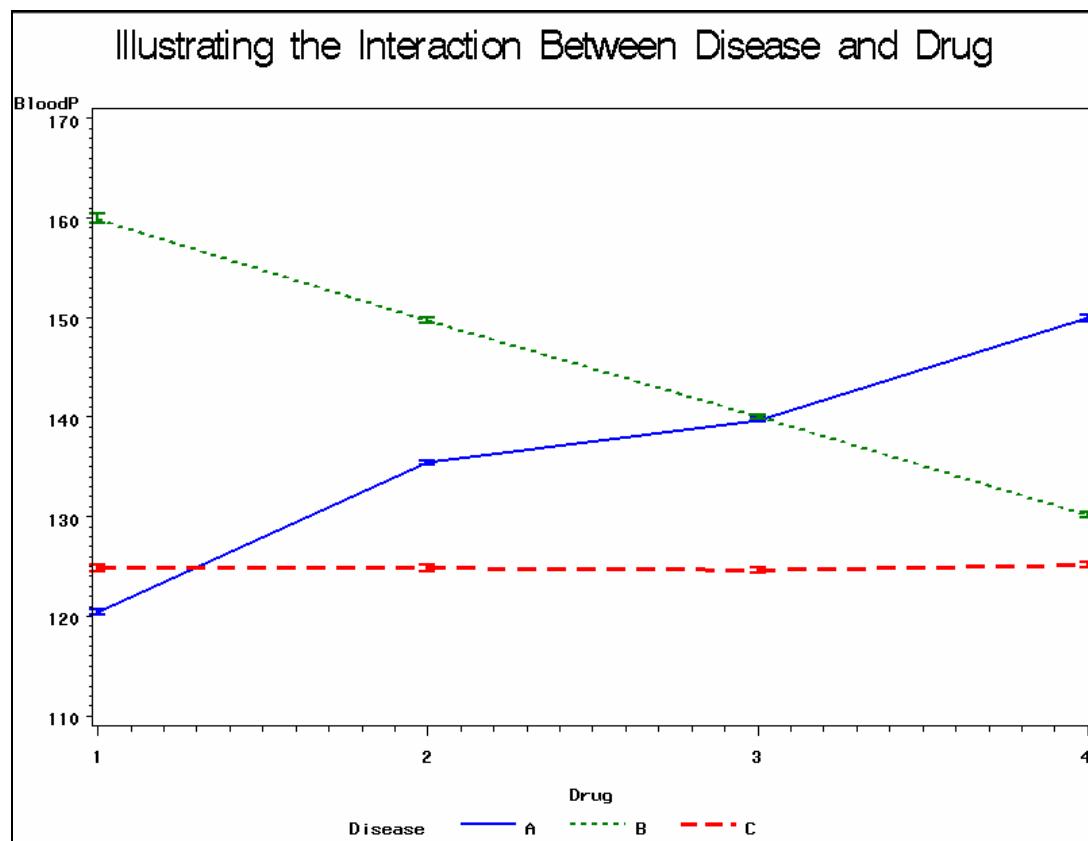
A means plot can help illustrate these relationships.

```
proc gplot data=sasuser.b_drug;
  symbol c=blue w=2 interpol=stdlmtj line=1;
  symbol2 c=green w=2 interpol=stdlmtj line=2;
  symbol3 c=red w=2 interpol=stdlmtj line=3;
  plot bloodp*drug=disease;
  title 'Illustrating the Interaction Between Disease and Drug';
run;
quit;
```

Selected SYMBOL statement options:

- INTERPOL= sets how the plotted points are to be related to each other visually. There are many different kinds of interpolation methods, including regression lines, needle plots, and joining with straight lines. See the SAS online documentation for more information.
- STD*i* specifies that SAS is to plot the means of the Y variable for each grouping of the X variable with a standard error bar above or below the point. The *i* indicates the number of standard errors wide the error bar should be. The default is 2.
- M indicates that the standard error used for the standard error bars should be the standard error of the mean.
- J connects the means with a straight line.
- T adds tops and bottoms to each line.

#### SAS/GPGRAPH Output



From the graph, the relationship is clearer. For disease type A, blood pressure rises as the drug level increases. For disease type B, blood pressure falls as the drug level increases. For disease type C, blood pressure is relatively unchanged for different drug levels.

You can use the GLM procedure to discover whether these differences and their interactions are statistically significant.

```
/* c2demo12 */
proc glm data=sasuser.b_drug;
  class disease drug;
  model bloodp=disease drug disease*drug;
  title 'Analyze the Effects of Drug and Disease';
  title2 'Including Interaction';
run;
quit;
```

 As seen in the MODEL statement, the interaction term can be added to the model by using a \* to separate the two main effects. It does **not** need to be created in a DATA step.

#### PROC GLM Output

```
Analyze the Effects of Drug and Disease
Including Interaction

The GLM Procedure

Class Level Information

      Class      Levels      Values
      Disease        3      A B C
      Drug          4      1 2 3 4

      Number of Observations Read      120
      Number of Observations Used      120
```

The first page of the output specifies the number of levels and the values of the class variables in the model, as well as the number of observations. You can verify that your variables and their levels were specified correctly.

The next part of the output, below, shows the source table with the *F* test for the overall model. This tests the null hypothesis that none of the effects in the model are statistically different, in other words, that there is no difference between the group means.

Analyze the Effects of Drug and Disease Including Interaction					
The GLM Procedure					
Dependent Variable: BloodP					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	17521.93235	1592.90294	1847.96	<.0001
Error	108	93.09386	0.86198		
Corrected Total	119	17615.02621			
R-Square	Coeff Var	Root MSE	BloodP Mean		
0.994715	0.685763	0.928429	135.3862		

The descriptive statistics indicate that the average blood pressure for all observations is 135.3862. The  $R^2$  for this model is 0.994715.

The *p*-value given is <0.0001. Presuming an alpha equal to 0.05, you reject the null hypothesis and conclude that at least one treatment mean is different from one other treatment mean. Which factor(s) cause this difference?

The next part of the output shows tests of the main effects and the interaction.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Disease	2	8133.949263	4066.974632	4718.18	<.0001
Drug	3	65.146099	21.715366	25.19	<.0001
Disease*Drug	6	9322.836990	1553.806165	1802.60	<.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Disease	2	8133.949263	4066.974632	4718.18	<.0001
Drug	3	65.146099	21.715366	25.19	<.0001
Disease*Drug	6	9322.836990	1553.806165	1802.60	<.0001

The sums of squares are used to test the null hypothesis that the effect of the individual terms in the model is insignificant. You should consider the test for the interaction first. The *p*-value is <0.0001. Presuming an alpha of 0.05, you reject the null hypothesis. You have sufficient evidence to conclude that there is an interaction between the two factors. As shown in the graph, the effect of the level of drug changes for different disease types.

Because of the interaction, you do not want to test the factors separately for differences between the means. Instead, specify that differences across treatment groups are supposed to be tested for both factors simultaneously by specifying them both in the LSMEANS statement separated by an asterisk.

```
proc glm data=sasuser.b_drug;
  class disease drug;
  model bloodp=drug disease drug*disease;
  lsmeans disease*drug / adjust=tukey pdiff=all;
  title 'Multiple Comparisons Tests for Drug and Disease';
run;
quit;
```

#### Partial PROC GLM Output

Multiple Comparisons Tests for Drug and Disease

The GLM Procedure  
Least Squares Means  
Adjustment for Multiple Comparisons: Tukey

Disease	Drug	BloodP LSMEAN	LSMEAN Number
A	1	120.389295	1
A	2	135.389295	2
A	3	139.706214	3
A	4	149.916892	4
B	1	159.874631	5
B	2	149.721891	6
B	3	140.027351	7
B	4	130.187304	8
C	1	124.820539	9
C	2	124.834540	10
C	3	124.640067	11
C	4	125.126164	12

The first table in the output assigns a number to each possible grouping by **Disease** and **Drug**.

The second table assigns *p*-values to each comparison.

Least Squares Means for effect Disease*Drug Pr >  t  for H0: LSMean(i)=LSMean(j)						
Dependent Variable: BloodP						
i/j	1	2	3	4	5	6
1		<.0001	<.0001	<.0001	<.0001	<.0001
2	<.0001		<.0001	<.0001	<.0001	<.0001
3	<.0001	<.0001		<.0001	<.0001	<.0001
4	<.0001	<.0001	<.0001		<.0001	1.0000
5	<.0001	<.0001	<.0001	<.0001		<.0001
6	<.0001	<.0001	<.0001	1.0000	<.0001	
7	<.0001	<.0001	0.9998	<.0001	<.0001	<.0001
8	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
9	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
10	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
11	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
12	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

Least Squares Means for effect Disease*Drug Pr >  t  for H0: LSMean(i)=LSMean(j)						
Dependent Variable: BloodP						
i/j	7	8	9	10	11	12
1	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
2	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
3	0.9998	<.0001	<.0001	<.0001	<.0001	<.0001
4	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
5	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
6	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
7		<.0001	<.0001	<.0001	<.0001	<.0001
8	<.0001		<.0001	<.0001	<.0001	<.0001
9	<.0001	<.0001		1.0000	1.0000	0.9999
10	<.0001	<.0001	1.0000		1.0000	0.9999
11	<.0001	<.0001	1.0000	1.0000		0.9901
12	<.0001	<.0001	0.9999	0.9999	0.9901	

How do you interpret this table of *p*-values? Presuming an alpha equal to 0.05, the following interpretation and conclusions can be drawn:

- Row 4 in each of the two tables represents **Disease=A**, **Drug=4**.
- The *p*-values compare people with disease A and drug dosage level 4 to all other combinations of factors. Because 10 of the 11 pairs are less than the alpha value, you can assume that there is a statistically significant difference between that combination and the other combinations of factors. The only group that is not statistically different is **Disease=B**, **Drug=2** (LSMEAN Number 6), *p*-value=1.0000.
- Rows 9-12 compare all six pairs of means for **Disease=C**. These means are not statistically significant with *p*-values between 0.9901 and 1.0000. These results are consistent with what is seen in the graph.



Refer to Exercise 3 for Chapter 2 in Appendix A.

## 2.4 Chapter Summary

An analysis of variance (ANOVA) is used to determine whether the means of a continuous measurement for two or more groups are equal. The response variable, or dependent variable, is of primary interest and is a continuous variable. The predictor variable, or independent variable, is a categorical variable. A one-way ANOVA has one independent, or grouping, variable.

Three analyses were discussed: completely randomized, randomized block, and two-way ANOVA.

If the result of an analysis of variance is to reject the null hypothesis and conclude that there are differences between the population group means, then multiple comparison tests are used to determine which pairs of means are different. The least significant difference test controls only the comparisonwise error rate. There are many multiple comparison techniques that control the experimentwise error rate.

The assumptions of an analysis of variance are

- observations are independent.
- pooled residuals are approximately normal.
- all groups have approximately equal response variances.

These assumptions can be verified using a combination of statements and options from the GLM and GPLOT (or PLOT) procedures.

- Examine the residuals plot. Look for a random scatter within each group.
- Examine the distribution of the residuals using PROC UNIVARIATE output. Look for values for skewness and kurtosis close to zero, a symmetric box-and-whisker plot, nonsignificant measures for the normality statistics, and a normal appearing normal probability plot.
- Use the MEANS statement HOVTEST option in PROC GLM, and compare the *p*-value with alpha; the null hypothesis for this test is that the variances are approximately equal. If you reject the null hypothesis, then you have sufficient evidence to conclude that the variances are not equal.

If these assumptions are not met, the probability of drawing incorrect conclusions from the analysis might be increased. Some alternative analysis techniques are to transform the response variable or generate a Welch ANOVA.

When you analyze an *n*-way ANOVA, the first consideration must be whether there is interaction between the factors. This is done by looking at the test for interaction in the ANOVA output. If there is no interaction between the factors, then the tests for the individual factor effects can be considered in the table to determine the significance/nonsignificance of these factors. If there is interaction between the factors, then the tests for the individual factor effects might be misleading due to masking of these effects by the interaction. In the case of a significant interaction, PROC GLM can be used to perform multiple comparison tests to compare treatment means.

```
PROC GLM DATA= SAS-data-set;
  CLASS variables;
  MODEL dependents=independents </ options>;
  MEANS effects </ options>;
  OUTPUT OUT=SAS-data-set keyword=variable...;
RUN;
```

```
PROC GPLOT DATA = SAS-data-set;
  PLOT vertical-variable*horizontal-variable </options>;
  SYMBOL <options>;
  AXISn <options>;
RUN;
```

# Chapter 3 Regression

3.1 Exploratory Data Analysis .....	3-2
3.2 Simple Linear Regression .....	3-27
3.3 Concepts of Multiple Regression.....	3-47
3.4 Model Building and Interpretation .....	3-61
3.5 Chapter Summary.....	3-83

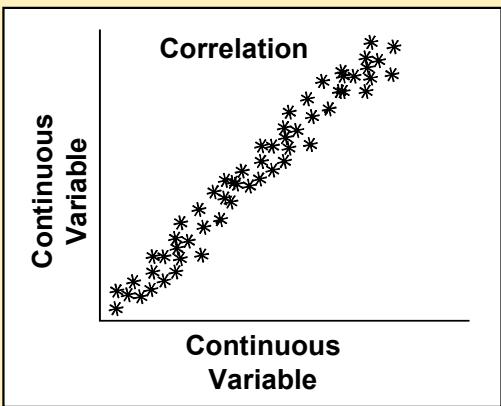
## 3.1 Exploratory Data Analysis

### Objectives

- Examine the relationship between two continuous variables using a scatter plot.
- Quantify the degree of linearity between two continuous variables using correlation statistics.
- Understand potential misuses of the correlation coefficient.
- Obtain Pearson correlation coefficients using the CORR procedure.

3

### Overview



4

In the previous chapter, you learned that when you have a discrete predictor variable and a continuous outcome variable you use ANOVA to analyze your data. In this section, you have two continuous variables.

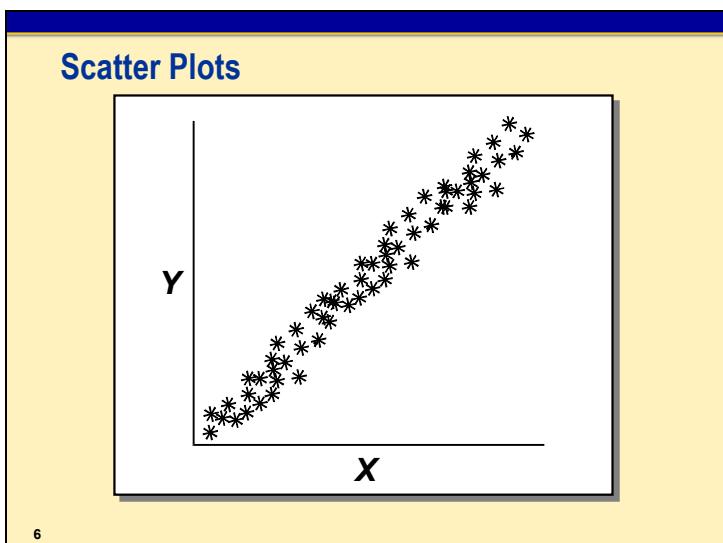
You use correlation analysis to examine and describe the relationship between two continuous variables. However, before you use correlation analysis, it is important to view the relationship between two continuous variables using a scatter plot.

### Example of Two Continuous Variables

The diagram shows two human figures. The figure on the left is standing on a blue rectangular base labeled "lb." (pounds). A red arrow points from this figure to the figure on the right, which is labeled "in." (inches) above its head, indicating height. Below the figures is a graph with a horizontal axis labeled "Height" and a vertical axis labeled "Weight". A blue line with a positive slope starts at the origin and ends at a question mark, representing the relationship between height and weight.

5

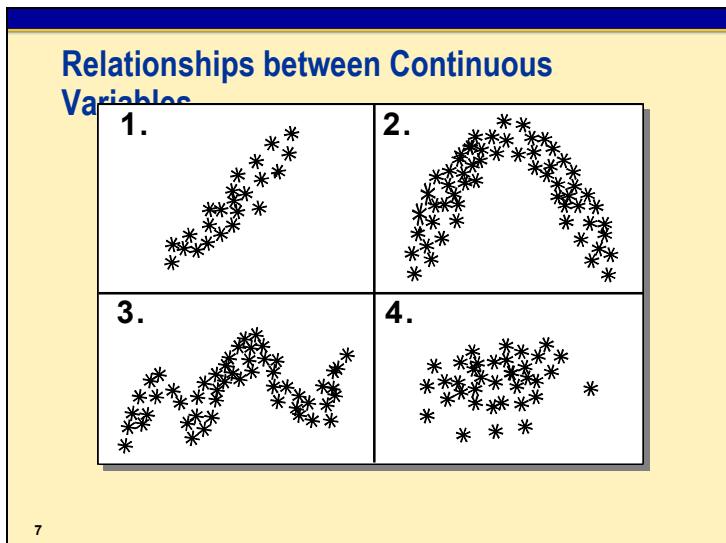
Example: A random sample of high school students is selected to determine the relationship between a person's height and weight. Height and weight are measured on a numeric scale. They have a large, potentially infinite number of possible values, rather than a few categories such as short, medium, and tall. Therefore, these variables are considered to be continuous.



*Scatter plots* are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.

Scatter plots are useful to

- explore the relationships between two variables
- locate outlying or unusual values
- identify possible trends
- identify a basic range of Y and X values
- communicate data analysis results.



Describing the relationship between two continuous variables is an important first step in any statistical analysis. The scatter plot is the most important tool you have in describing these relationships. The diagrams above illustrate some possible relationships.

1. A straight line describes the relationship.
2. Curvature is present in the relationship.
3. There could be a cyclical pattern in the relationship. You might see this when the predictor is time.
4. There is no clear relationship between the variables.

**Fitness Example**

The image shows four black silhouettes of a person running, each enclosed in a small rectangular frame. Above each silhouette is a small clock icon with three dots around it, suggesting time or measurement. The silhouettes are arranged in a staggered, non-linear path across the frame.

8

In exercise physiology, an object measure of aerobic fitness is how fast the body can absorb and use oxygen (oxygen consumption). Subjects participated in a predetermined exercise run of 1.5 miles. Measurements of oxygen consumption as well as several other continuous measurements such as age, pulse, and weight were recorded. The researchers are interested in determining whether any of these other variables can help predict oxygen consumption. This data is found in Rawlings (1998) but certain values of **Maximum\_Pulse** and **Run\_Pulse** were changed for illustration. **Name**, **Gender**, and **Performance** were also contrived for illustration.

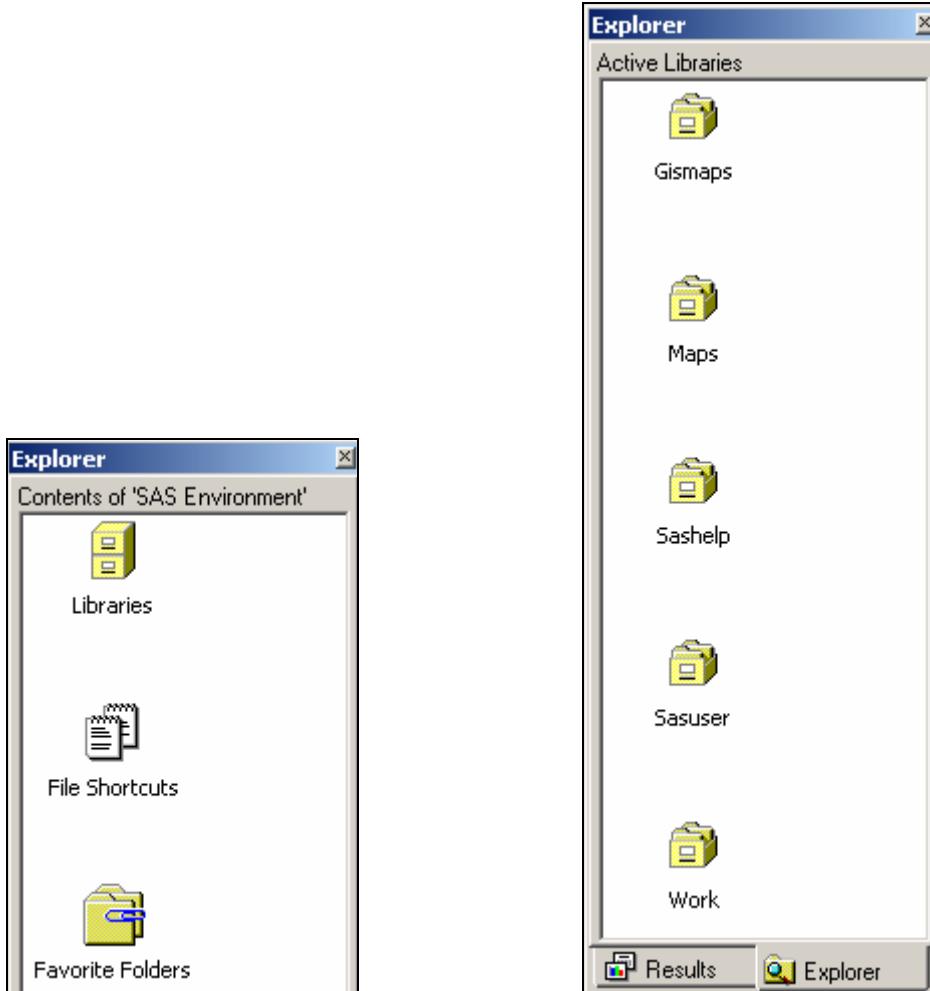
The data set **sasuser.b\_fitness** contains these variables:

<b>Name</b>	name of the member
<b>Gender</b>	gender of the member
<b>Runtime</b>	time to run 1.5 miles (in minutes)
<b>Age</b>	age of the member (in years)
<b>Weight</b>	weight of the member (in kilograms)
<b>Oxygen_Consumption</b>	a measure of the ability to use oxygen in the blood stream
<b>Run_Pulse</b>	pulse rate at the end of the run
<b>Rest_Pulse</b>	resting pulse rate
<b>Maximum_Pulse</b>	maximum pulse rate during the run
<b>Performance</b>	a measure of overall fitness

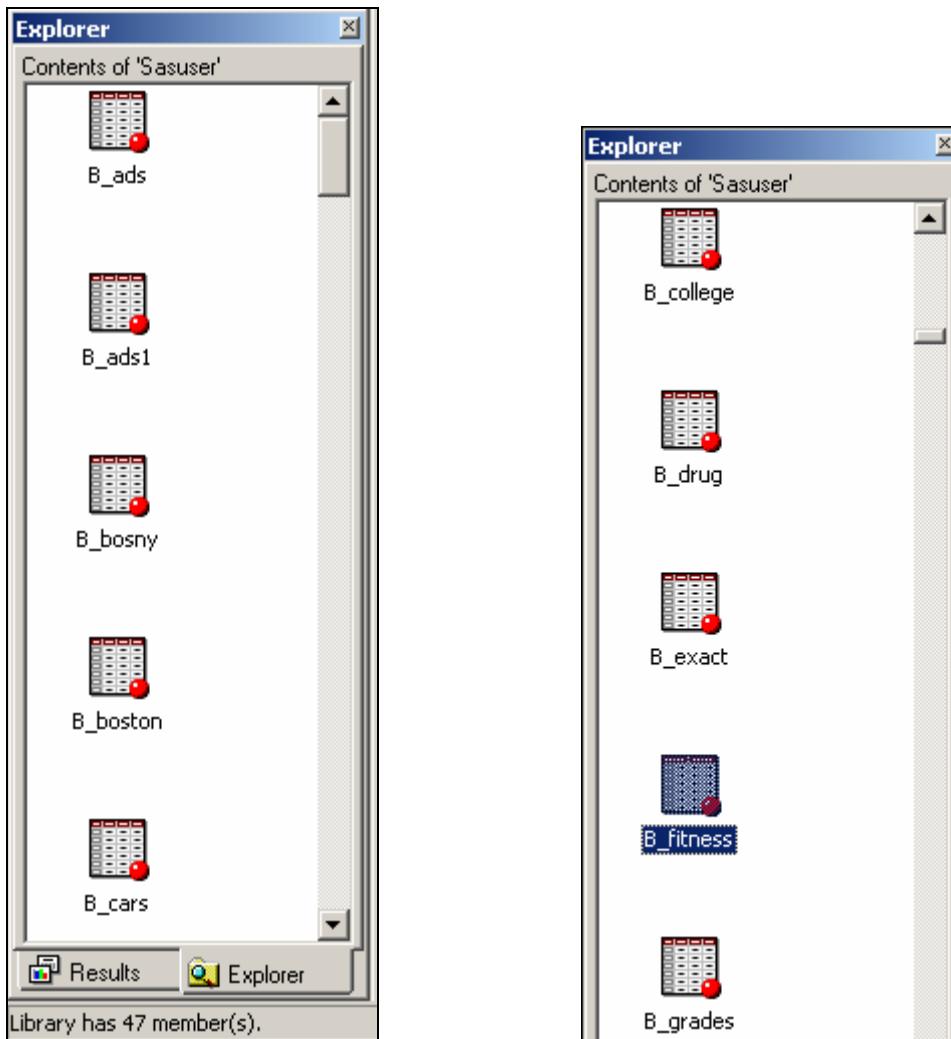


## Data Exploration

You can view the contents of any data set by using the Explorer window. Select **Libraries** from the Contents of ‘SAS Environment’ window. The Explorer window shows all currently defined SAS libraries.

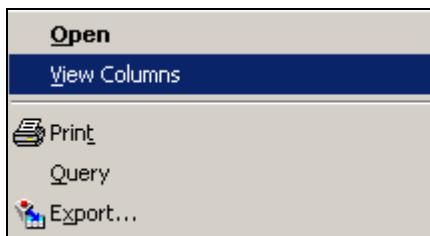


In the Explorer window, double-click on **Sasuser**. The tables in this library are displayed.

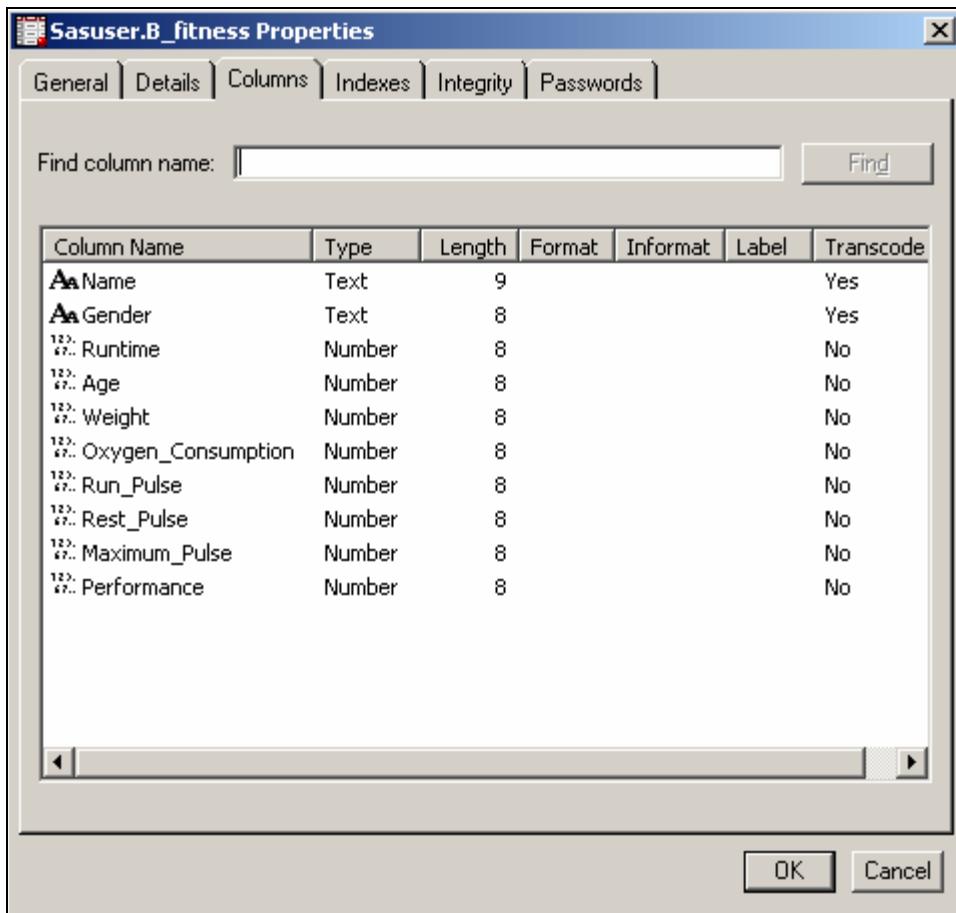


Scroll down, select **B\_fitness** and right-click.

Left-click **View Columns**.



This provides a list of the columns in the data table and their properties.



Select **OK**.

You can look at the data in the table by double-clicking on the data set name.

VIEWTABLE: Sasuser.B_fitness											
	Name	Gender	Runtime	Age	Weight	Oxygen_Consumption	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance	
1	Donna	F	8.17	42	68.15	59.57	166	40	172	14	
2	Gracie	F	8.63	38	81.87	60.06	170	48	186	13	
3	Luanne	F	8.65	43	85.84	54.3	156	45	168	13	
4	Mimi	F	8.92	50	70.87	54.63	146	48	155	11	
5	Chris	M	8.95	49	81.42	49.16	180	44	185	11	
6	Allen	M	9.22	38	89.02	49.87	178	55	180	12	
7	Nancy	F	9.4	49	76.32	48.67	186	56	188	10	
8	Patty	F	9.63	52	76.32	45.44	164	48	166	10	
9	Suzanne	F	9.93	57	59.08	50.55	148	49	155	9	
10	Teresa	F	10	51	77.91	46.67	162	48	168	9	
11	Bob	M	10.07	40	75.07	45.31	185	62	185	9	
12	Harriett	F	10.08	49	73.37	50.39	168	67	168	9	
13	Jane	F	10.13	44	73.03	50.54	168	45	168	9	
14	Harold	M	10.25	48	91.63	46.77	162	48	164	9	
15	Sammy	M	10.33	54	83.12	51.85	166	50	170	8	
16	Buffy	F	10.47	52	73.71	45.79	186	59	188	8	



You could also look at the data using the PRINT procedure.

```
/* c3demo01_p */
proc print data=sasuser.b_fitness;
  title 'Printout of sasuser.b_fitness';
run;
```

You should also investigate the univariate statistics of continuous variables in the data set; in this program, you are storing the results as a file, using the HTML style. The PATH and GPATH options ensure the specific location of the output and graphs.

```
/* c3demo01_u */
ods listing close;

ods html path='c:\'  (url=none)
      gpath='c:\' (url=none)
      file = 'fitness_unistats.html';

options ps=50 ls=76;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=sasuser.b_fitness;
  var Runtime -- Performance;
  histogram Runtime -- Performance / normal;
  probplot Runtime -- Performance
    / normal (mu=est sigma=est color=red w=2);
  title 'Univariate Statistics of sasuser.b_fitness';
run;
ods html close;
ods listing;
```

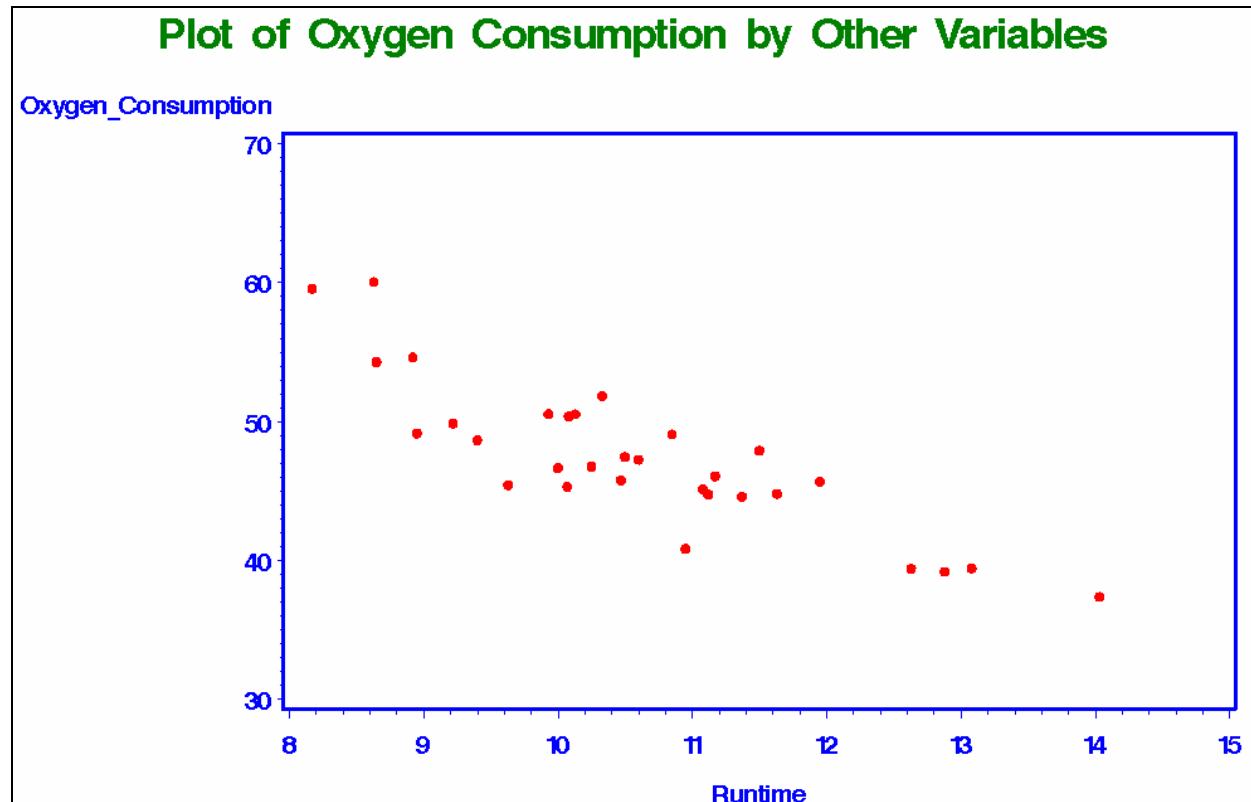
Examine the relationships between **Oxygen\_Consumption** and the continuous predictor variables in the data set using the GPLOT procedure.

```
/* c3demo02 */
options ps=50 ls=64;
goptions reset=all gunit=pct border
    fontres=presentation ftext=swissb;

axis1 length=70 w=3 color=blue label=(h=3) value=(h=3);
axis2 length=70 w=3 color=blue label=(h=3) value=(h=3);

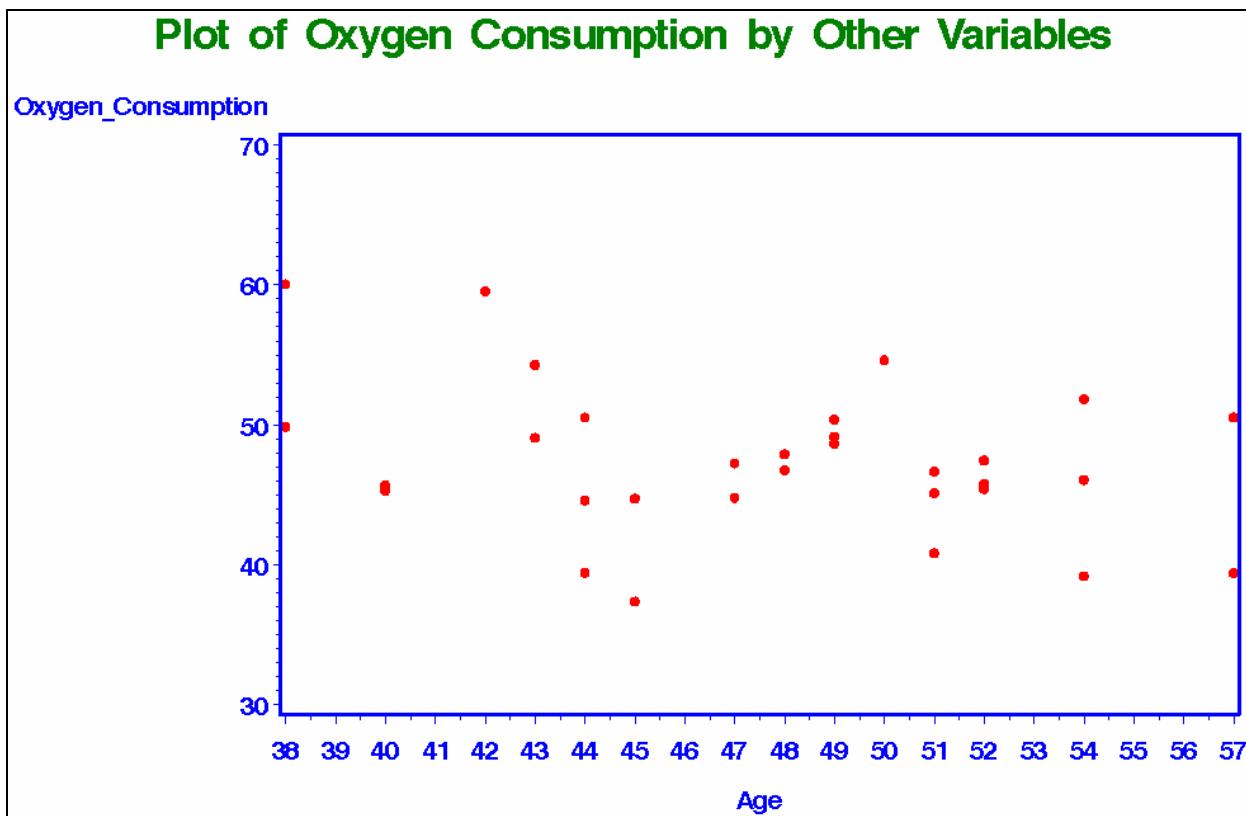
proc gplot data=sasuser.b_fitness;
    plot oxygen_consumption * (runtime age weight run_pulse
        rest_pulse maximum_pulse performance)
        / vaxis=axis1 haxis=axis2;
    symbol1 v=dot h=2 w=4 color=red;
    title h=3 color=green
        'Plot of Oxygen Consumption by Other Variables';
run;
quit;
```

PROC GPLOT Output



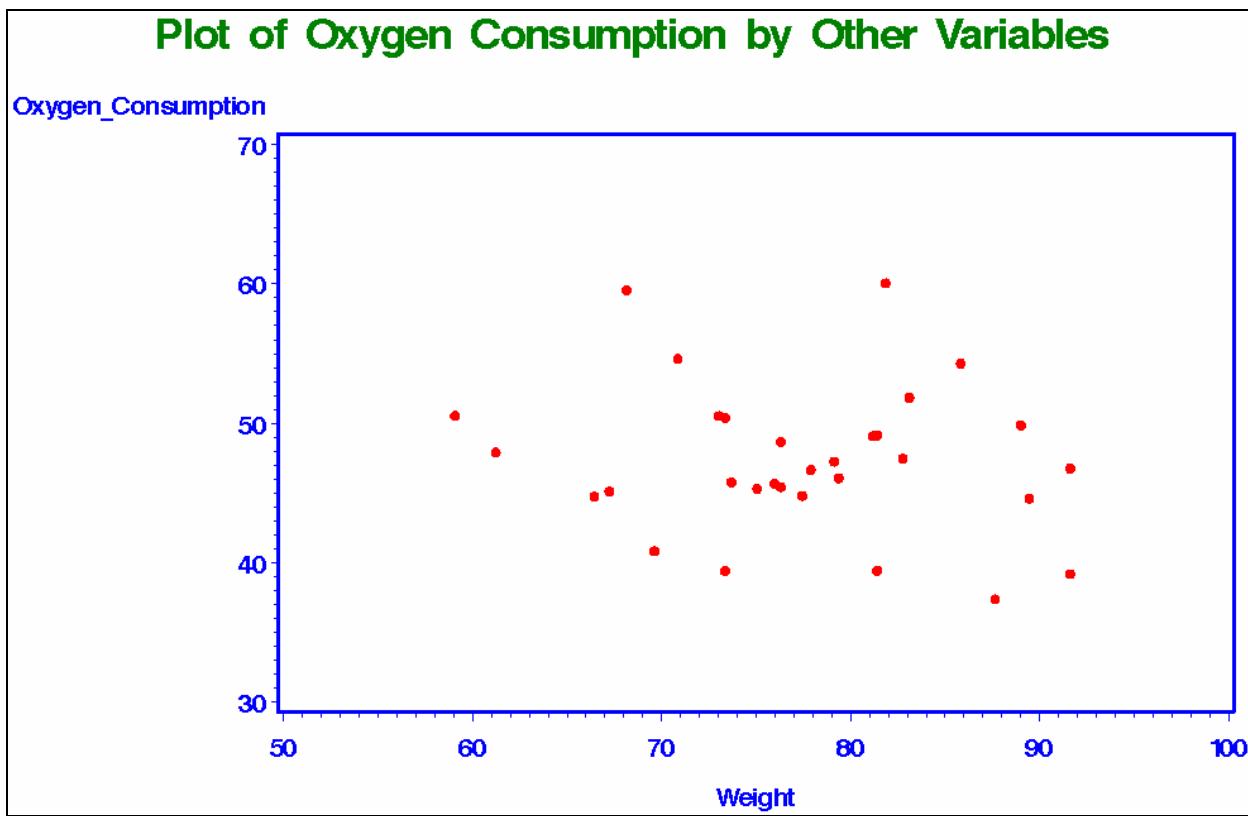
This plot suggests a linear relationship between **Oxygen\_Consumption** and **Runtime**. As **Runtime** increases, **Oxygen\_Consumption** decreases.

PROC GPLOT Output (continued)



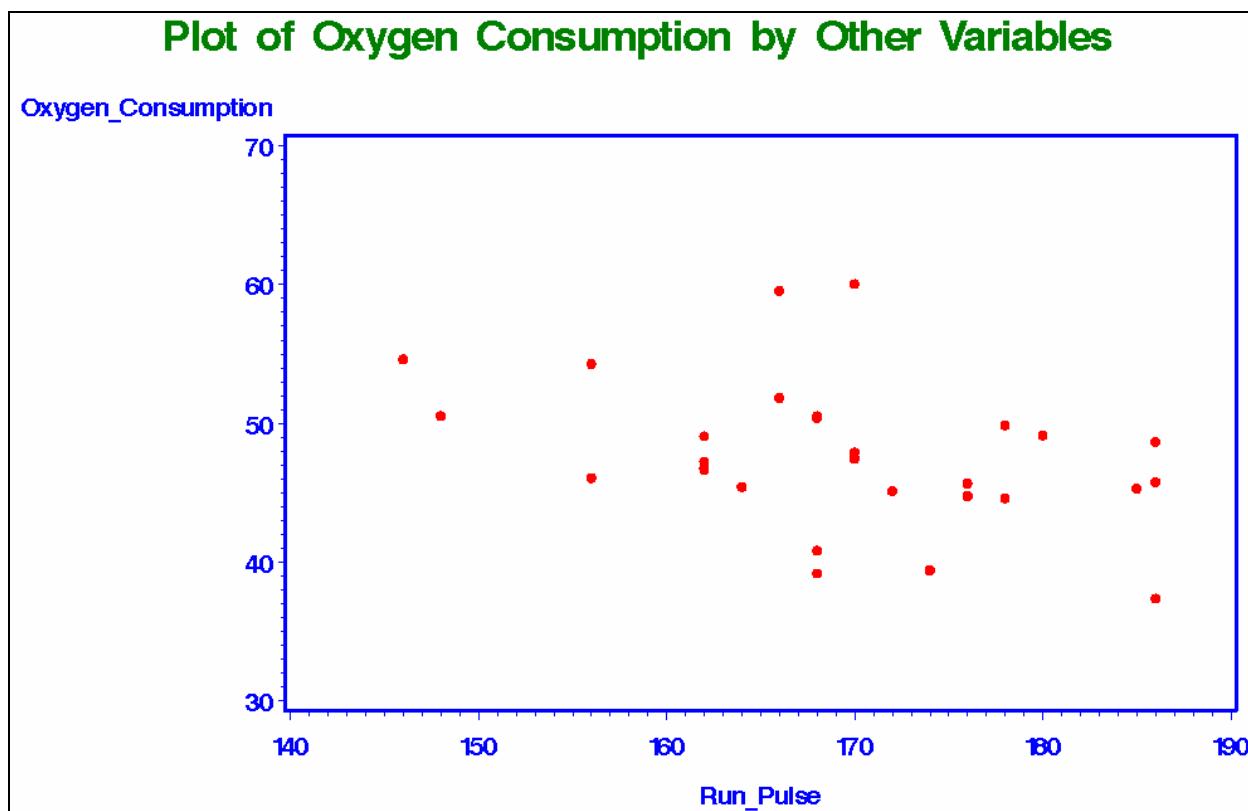
There appears to be a possible weak linear relationship between **Oxygen\_Consumption** and **Age**.

PROC GPLOT Output (continued)



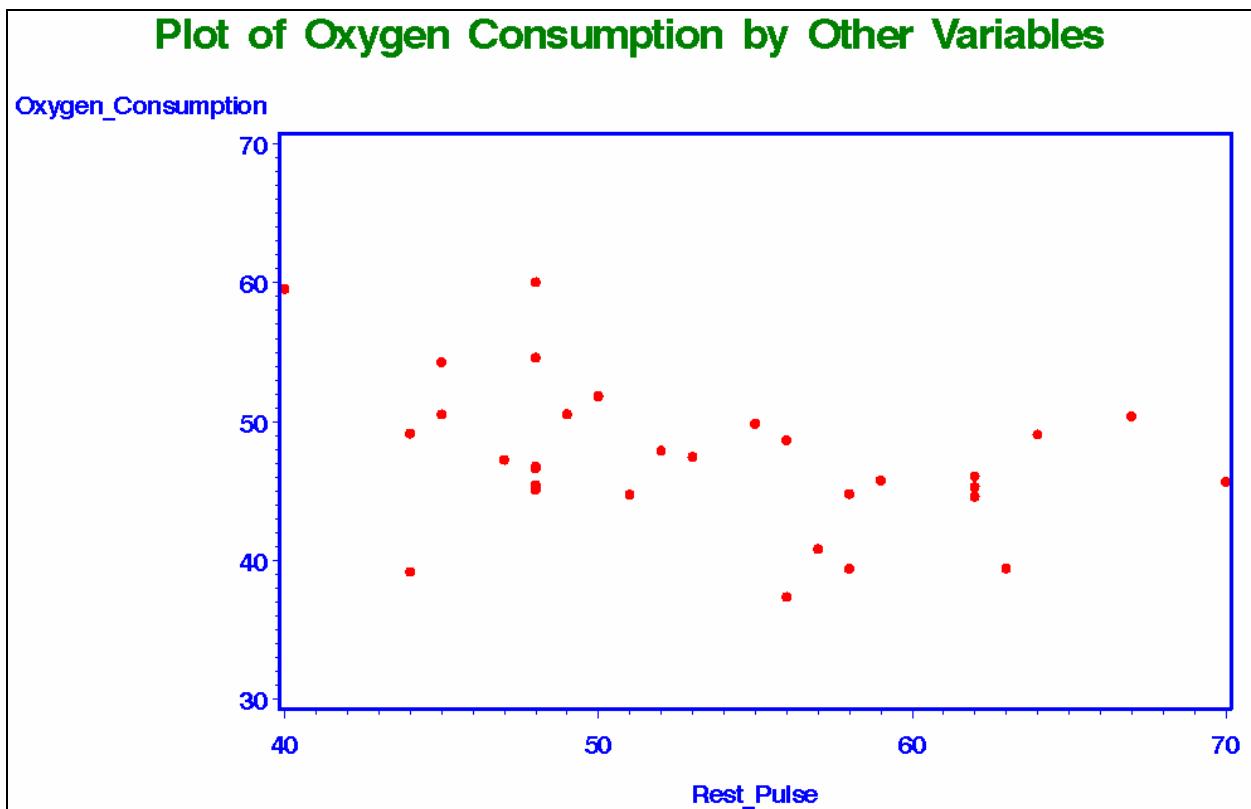
There does **not** appear to be a linear relationship between **Oxygen\_Consumption** and **Weight**.

PROC GPLOT Output (continued)



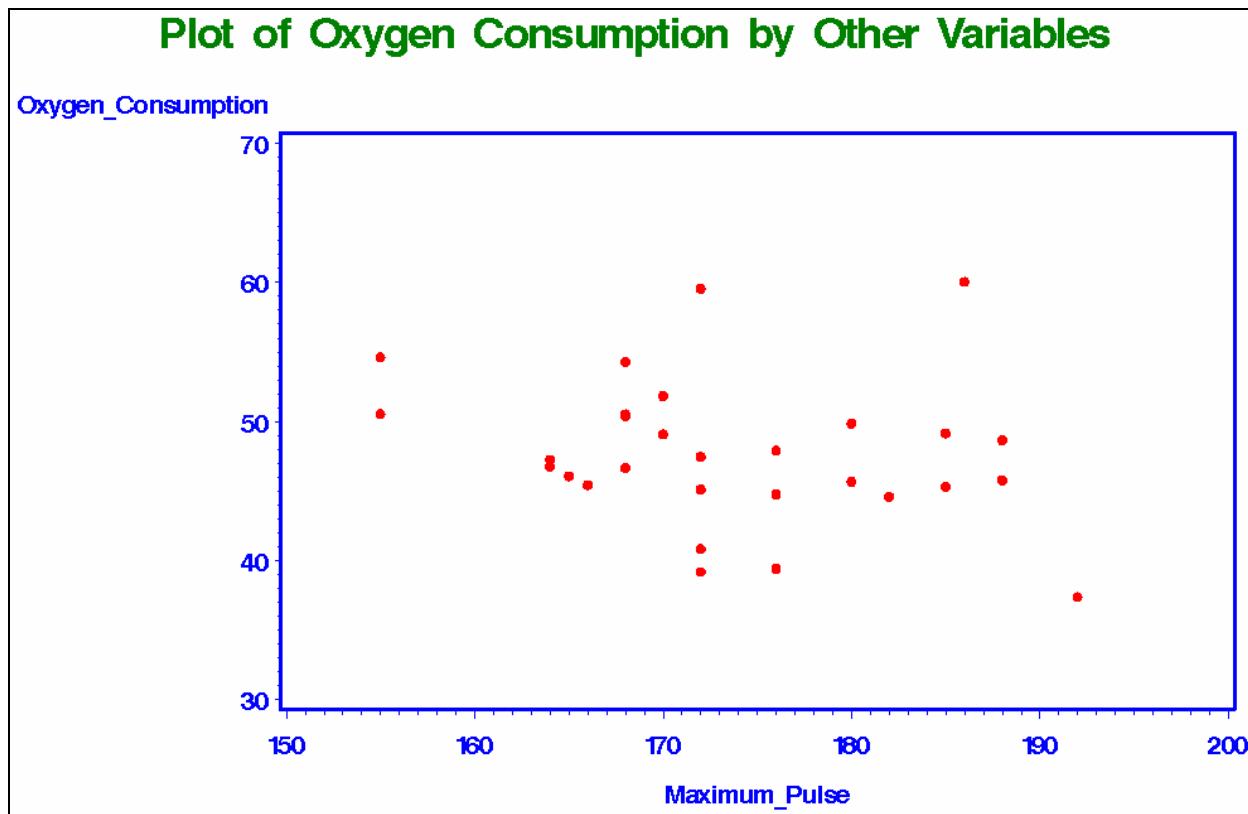
There appears to be a possible weak linear relationship between Oxygen\_Consumption and Run\_Pulse.

PROC GPLOT Output (continued)



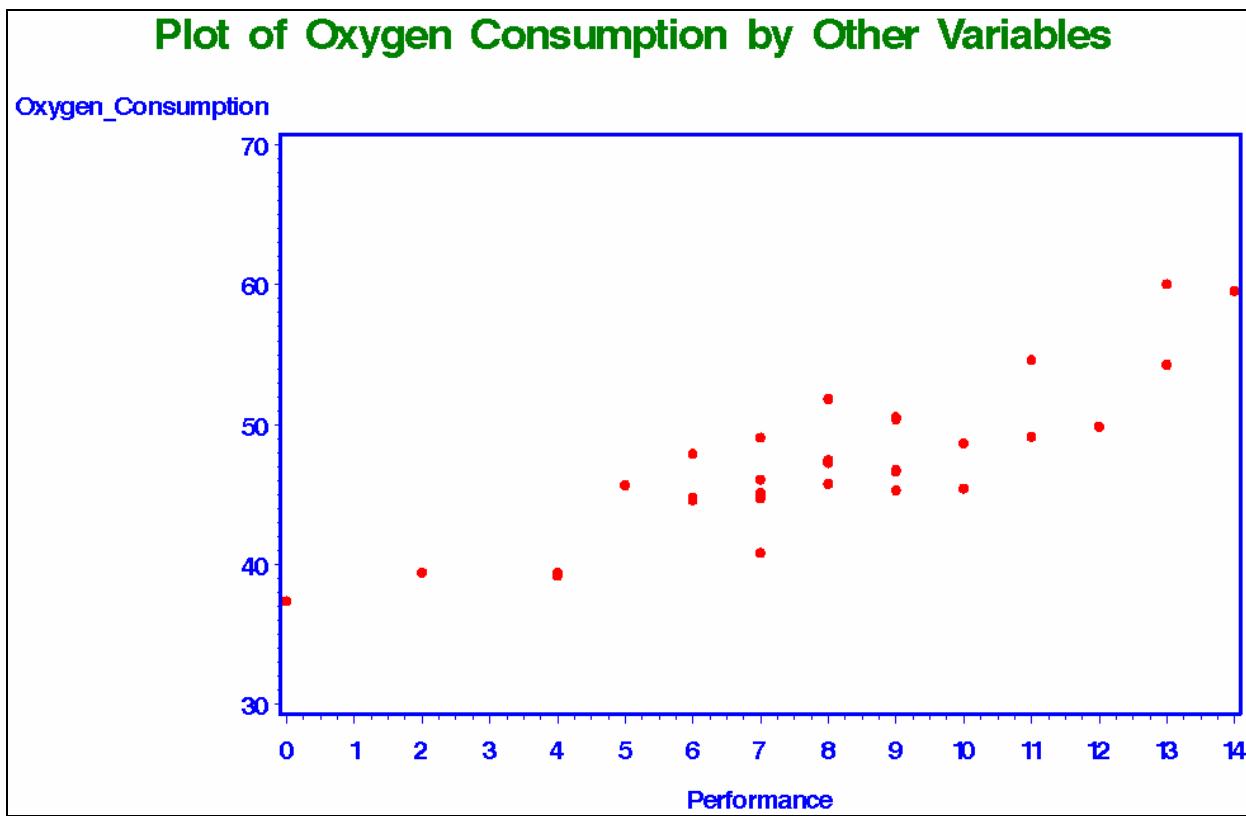
There might be a slight negative relationship between **Oxygen\_Consumption** and **Rest\_Pulse**.

PROC GPLOT Output (continued)

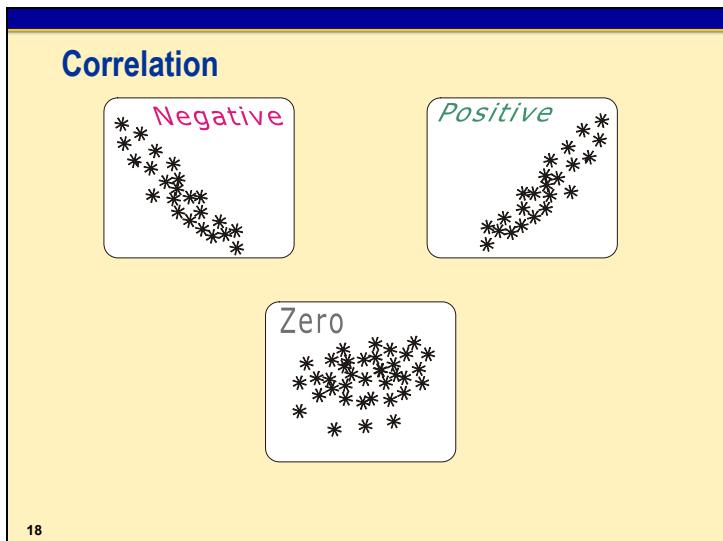


There does **not** appear to be a linear relationship between **Oxygen\_Consumption** and **Maximum\_Pulse**.

PROC GPLOT Output (continued)



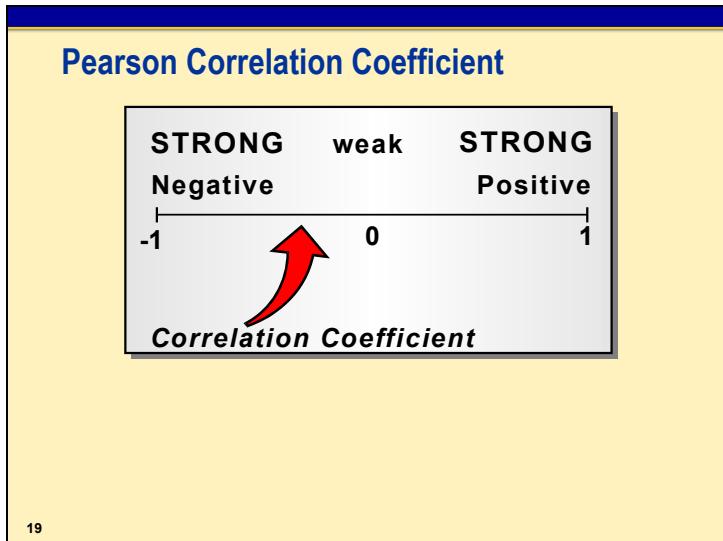
As **Performance** increases, **Oxygen\_Consumption** appears to increase slightly.



After you examine the scatter plot, you can quantify the relationship between two variables with correlation statistics. Two variables are correlated if there is a **linear** relationship between them. If not, the variables are uncorrelated.

You can classify correlated variables according to the type of correlation:

- |          |  |
|----------|--|
| positive | one variable tends to increase in value as the other variable increases in value |
| negative | one variable tends to decrease in value as the other variable increases in value |
| zero     | no linear relationship between the two variables (uncorrelated)                  |

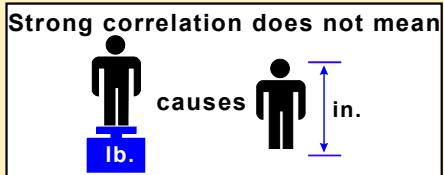


19 Correlation statistics measure the degree of linear relationship between two variables. A common correlation statistic used for continuous variables is the Pearson correlation coefficient. Values of correlation statistics are

- between  $-1$  and  $1$
- closer to either extreme if there is a high degree of linear relationship between the two variables
- close to  $0$  if there is no linear relationship between the two variables
- greater than  $0$  if there is a positive linear relationship
- less than  $0$  if there is a negative linear relationship.

 The magnitude of the relationship is based on sample size.

## Misuses of the Correlation Coefficient

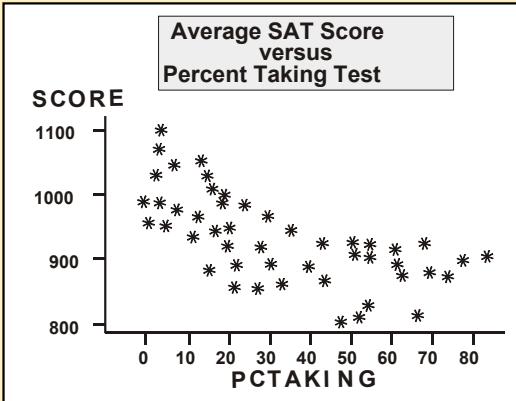


20

Common errors can be made when interpreting the correlation between variables. One example of this is using correlation coefficients to conclude a cause-and-effect relationship.

- A strong correlation between two variables does not mean change in one variable causes the other variable to change, or vice versa.
- Sample correlation coefficients can be large because of chance or because both variables are affected by other variables.

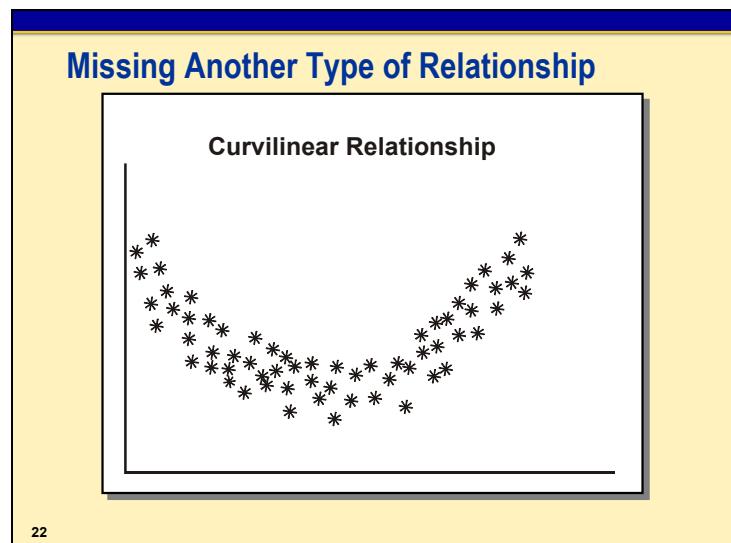
## SAT Example



21

An example of improperly concluding a cause-and-effect relationship is illustrated using data from the Scholastic Aptitude Test (SAT) from 1989. The scatter plot shown above plots each state's average total SAT score (**score**) versus the percent of eligible students in the state who took the SAT (**pctaking**). The correlation between **score** and **pctaking** is  $-0.86867$ . Looking at the plot and at this statistic, an eligible student for the next year can conclude, "If I am the only student in my state to take the SAT, I am guaranteed a good score."

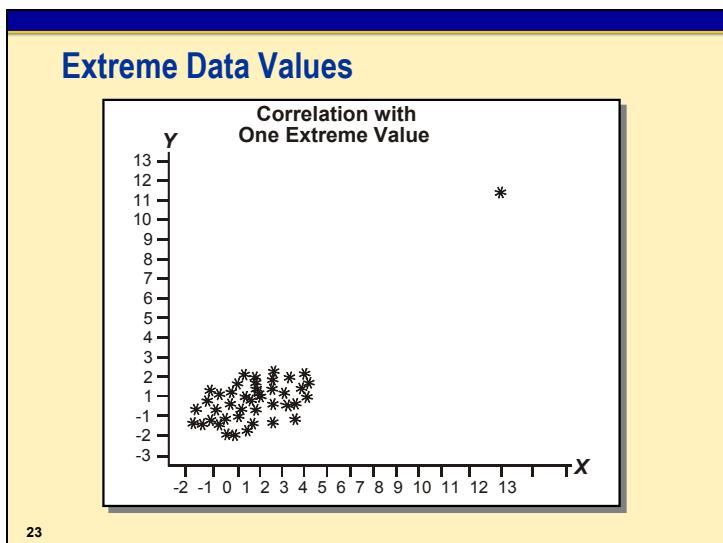
Clearly this type of thinking is faulty. Can you think of possible explanations for this relationship?



In the scatter plot above, the variables have a fairly low Pearson correlation coefficient. Why?

- Correlation coefficients measure linear relationships.
- A correlation coefficient close to 0 indicates that there is not a strong linear relationship between two variables.
- A correlation coefficient close to 0 does not mean there is no relationship of any kind between the two variables.

In this example, there is a curvilinear relationship between the two variables.



Correlation coefficients are highly affected by a few extreme values of either variable. The scatter plot above shows that the degree of linear relationship is mainly determined by one point. If you delete the unusual point from the data, the correlation is close to 0.

In this situation, follow these steps:

1. Investigate the unusual data point to make sure it is valid.
2. If the data point is valid, collect more data between the unusual data point and the group of data points to see whether a linear relationship unfolds.
3. Try to replicate the unusual data point by collecting data at a fixed value of  $x$  (in this case,  $x=13$ ). This determines whether the data point is unusual.
4. Compute two correlation coefficients, one with the unusual data point and one without it. This shows how influential the unusual data point is in the analysis.

## The CORR Procedure

General form of the CORR procedure:

```
PROC CORR DATA=SAS-data-set <options>;
  VAR variables;
  WITH variables;
RUN;
```

24

You can use the CORR procedure to produce correlation statistics for your data. By default, PROC CORR produces Pearson correlation statistics and corresponding *p*-values.

Selected CORR procedure statements:

VAR specifies variables for which to produce correlations. If a WITH statement is not specified, correlations are produced for each pair of variables in the VAR statement. If the WITH statement is specified, the VAR statement specifies the column variables in the correlation matrix.

WITH produces correlations for each variable in the VAR statement with all variables in the WITH statement. The WITH statement specifies the row variables in the correlation matrix.



## Generating Correlation Coefficients

Use PROC CORR to produce a Pearson correlation coefficient for **Oxygen\_Consumption** with the other continuous predictor variables.

```
/* c3demo03 */
proc corr data=sasuser.b_fitness rank;
  var runtime age weight run_pulse rest_pulse maximum_pulse
    performance;
  with oxygen_consumption;
  title 'PROC CORR: oxygen_consumption with predictor variables';
run;
```

Selected PROC CORR statement option:

RANK      orders the correlations from highest to lowest in absolute value.

The output from PROC CORR is shown below. By default, the analysis generates univariate statistics for the analysis variables and correlations.

### PROC CORR Output

PROC CORR: oxygen_consumption with predictor variables						
The CORR Procedure						
1 With Variables:	Oxygen_Consumption					
7      Variables:	Runtime	Age	Weight	Maximum_Pulse	Performance	Run_Pulse
	Rest_Pulse					
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Oxygen_Consumption	31	47.37581	5.32777	1469	37.39000	60.06000
Runtime	31	10.58613	1.38741	328.17000	8.17000	14.03000
Age	31	47.67742	5.26236	1478	38.00000	57.00000
Weight	31	77.44452	8.32857	2401	59.08000	91.63000
Run_Pulse	31	169.64516	10.25199	5259	146.00000	186.00000
Rest_Pulse	31	53.45161	7.61944	1657	40.00000	70.00000
Maximum_Pulse	31	173.77419	9.16410	5387	155.00000	192.00000
Performance	31	8.00000	3.11983	248.00000	0	14.00000

## PROC CORR Output (continued)

Pearson Correlation Coefficients, N = 31				
Prob >  r  under H0: Rho=0				
Oxygen_Consumption	Performance	Runtime	Rest_Pulse	Run_Pulse
0.86377		-0.86219	-0.39935	-0.39808
<.0001		<.0001	0.0260	0.0266

Pearson Correlation Coefficients, N = 31				
Prob >  r  under H0: Rho=0				
Oxygen_Consumption	Age	Maximum_Pulse	Weight	
-0.31162		-0.23677	-0.16289	
0.0879		0.1997	0.3813	

The correlation coefficient between **Oxygen\_Consumption** and **Performance** is 0.86377. The *p*-value is small, which indicates that the population correlation coefficient (Rho) is significantly different from 0. The second largest correlation coefficient, in absolute value, is **Runtime**, -0.86219.

The correlation analysis indicates that several variables might be good predictors for **Oxygen\_Consumption**.

When you prepare to conduct a regression analysis, it is always good practice to examine the correlations between the potential predictor variables. PROC CORR can be used to generate a matrix of correlation coefficients.

```
/* c3demo04 */
proc corr data=sasuser.b_fitness nosimple;
  var runtime age weight run_pulse rest_pulse maximum_pulse
    performance;
  title 'correlations among predictor variables';
run;
title;
```

Selected PROC CORR statement option:

NOSIMPLE suppresses printing simple descriptive statistics for each variable.

## PROC CORR Output

correlations among predictor variables							
The CORR Procedure							
7 Variables:	Runtime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
Maximum_Pulse Performance							
Pearson Correlation Coefficients, N = 31 Prob >  r  under H0: Rho=0							
	Runtime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
Runtime	1.00000	0.19523 0.2926	0.14351 0.4412	0.31365 0.0858	0.45038 0.0110	0.22610 0.2213	-0.98841 <.0001
Age	0.19523 0.2926	1.00000	-0.24050 0.1925	-0.31607 0.0832	-0.15087 0.4178	-0.41490 0.0203	-0.22943 0.2144
Weight	0.14351 0.4412	-0.24050 0.1925	1.00000	0.18152 0.3284	0.04397 0.8143	0.24938 0.1761	-0.10544 0.5724
Run_Pulse	0.31365 0.0858	-0.31607 0.0832	0.18152 0.3284	1.00000	0.35246 0.0518	0.92975 <.0001	-0.31369 0.0857
Rest_Pulse	0.45038 0.0110	-0.15087 0.4178	0.04397 0.8143	0.35246 0.0518	1.00000	0.30512 0.0951	-0.47957 0.0063
Maximum_Pulse	0.22610 0.2213	-0.41490 0.0203	0.24938 0.1761	0.92975 <.0001	0.30512 0.0951	1.00000	-0.22035 0.2336
Performance	-0.98841 <.0001	-0.22943 0.2144	-0.10544 0.5724	-0.31369 0.0857	-0.47957 0.0063	-0.22035 0.2336	1.00000

There are strong correlations between **Runtime** and **Performance** (-0.98841) and between **Run\_Pulse** and **Maximum\_Pulse** (0.92975).

The following correlation table was created from the matrix by choosing small *p*-values. The table is in descending order, based on the absolute value of the correlation. It provides a summary of the correlation analysis of the independent variables.

Row Variable	Column Variable	Pearson's r	Prob >  r
<b>Runtime</b>	<b>Performance</b>	-0.98841	<.0001
<b>Run_Pulse</b>	<b>Maximum_Pulse</b>	0.92975	<.0001
<b>Rest_Pulse</b>	<b>Performance</b>	-0.47957	0.0063
<b>Runtime</b>	<b>Rest_Pulse</b>	0.45038	0.0110
<b>Age</b>	<b>Maximum_Pulse</b>	-0.41490	0.0203
<b>Run_Pulse</b>	<b>Rest_Pulse</b>	0.35246	0.0518



Refer to Exercise 1 for Chapter 3 in Appendix A.

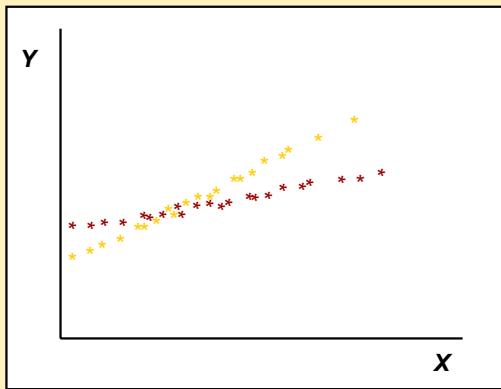
## 3.2 Simple Linear Regression

### Objectives

- Explain the concepts of simple linear regression.
- Fit a simple linear regression using the REG procedure.
- Produce predicted values and confidence intervals.

31

### Overview



32

In the last section, you used correlation analysis to quantify the linear relationships between continuous response variables. Two pairs of variables can have the same correlation statistic, but the linear relationship can be different. In this section, you use simple linear regression to define the linear relationship between a response variable and a predictor variable.

The *response variable* is the variable of primary interest.

The *predictor variable* is used to explain the variability in the response variable.

## Simple Linear Regression Analysis

The objectives of simple linear regression are to

- assess the significance of the predictor variable in explaining the variability or behavior of the response variable
- predict the values of the response variable given the values of the predictor variable.

33

In simple linear regression, the values of the predictor variable are assumed fixed. Thus, you try to explain the variability of the response variable given the values of the predictor variable.

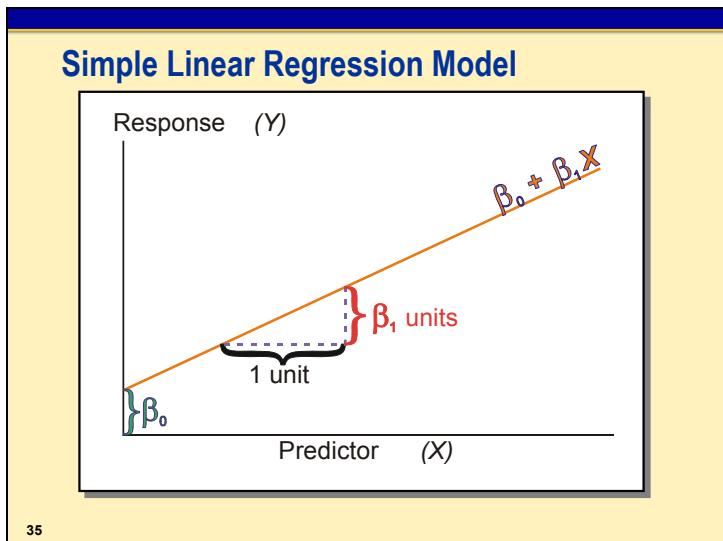
## Fitness Example



34

The analyst noted that the performance measure has the highest correlation with the oxygen consumption capacity of the club members. Consequently, he wants to further explore the relationship between **Oxygen\_Consumption** and **Performance**.

The analyst decides to run a simple linear regression of **Oxygen\_Consumption** versus **Performance**.



The relationship between the response variable and the predictor variable can be characterized by the equation  $Y = \beta_0 + \beta_1 X + \varepsilon$

where

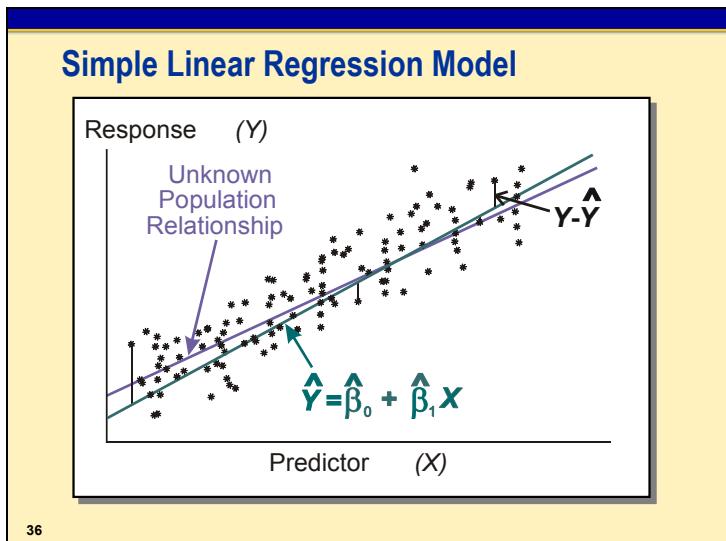
$Y$  response variable

$X$  predictor variable

$\beta_0$  intercept parameter, which corresponds to the value of the response variable when the predictor is 0

$\beta_1$  slope parameter, which corresponds to the magnitude of change in the response variable given a one unit change in the predictor variable

$\varepsilon$  error term representing deviations of  $Y$  about  $\beta_0 + \beta_1 X$ .



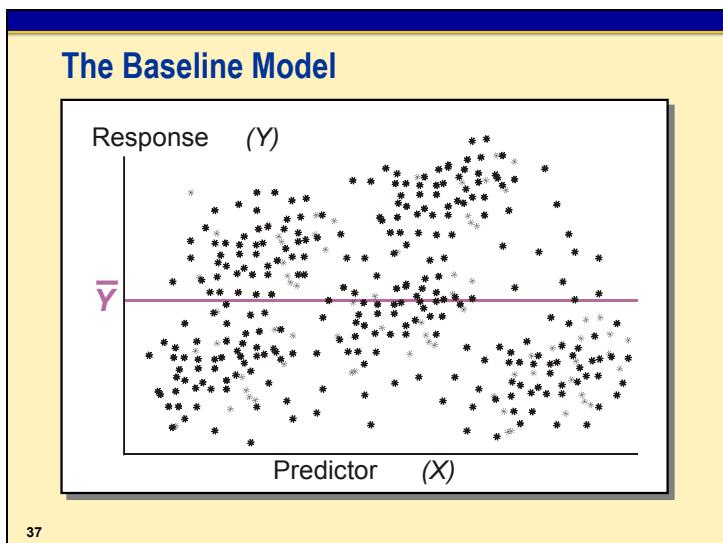
Because your goal in simple linear regression is usually to characterize the relationship between the response and predictor variables in your population, you begin with a sample of data. From this sample, you estimate the unknown population parameters ( $\beta_0, \beta_1$ ) that define the assumed relationship between your response and predictor variables.

Estimates of the unknown population parameters  $\beta_0$  and  $\beta_1$  are obtained by the *method of least squares*. This method provides the estimates by determining the line that minimizes the sum of the squared vertical distances between the observations and the fitted line. In other words, the fitted or regression line is as close as possible to all the data points.

The method of least squares produces parameter estimates with certain optimum properties. If the assumptions of simple linear regression are valid, the least squares estimates are unbiased estimates of the population parameters and have minimum variance. The least squares estimators are often called BLUE (Best Linear Unbiased Estimators). The term *best* is used because of the minimum variance property.

Because of these optimum properties, the method of least squares is used by many data analysts to investigate the relationship between continuous predictor and response variables.

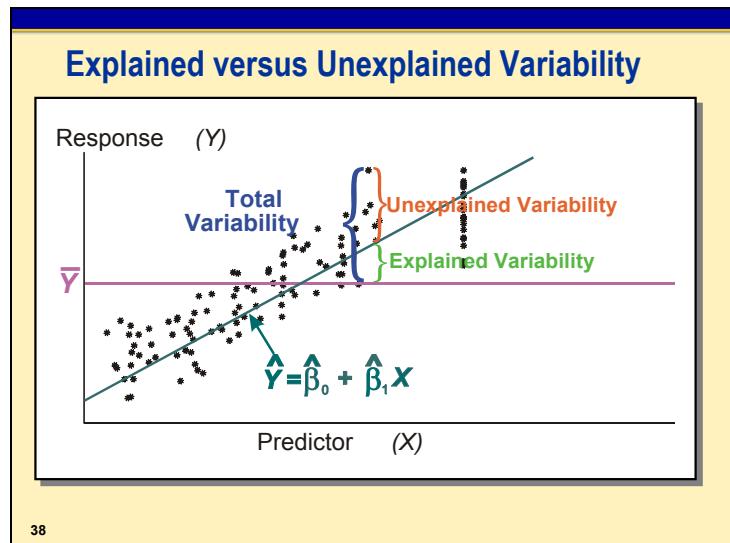
With a large and representative sample, the fitted regression line should be a good approximation of the relationship between the response and predictor variables in the population. The estimated parameters obtained using the method of least squares should be good approximations of the true population parameters.



37

To determine whether the predictor variable explains a significant amount of variability in the response variable, the simple linear regression model is compared to the baseline model. The fitted regression line in a baseline model is a horizontal line across all values of the predictor variable. The slope of the regression line is 0 and the intercept is the sample mean of the response variable, ( $\bar{Y}$ ).

In a baseline model, there is no association between the response variable and the predictor variable. Knowing the mean of the response variable is as good in predicting values in the response variable as knowing the values of the predictor variable.



38

To determine whether a simple linear regression model is better than the baseline model, compare the explained variability to the unexplained variability.

Explained variability is related to the difference between the regression line and the mean of the response variable. The model sum of squares (SSM) is the amount of variability explained by your model. The model sum of squares is equal to  $\sum(\hat{Y}_i - \bar{Y})^2$ .

Unexplained variability is related to the difference between the observed values and the regression line. The error sum of squares (SSE) is the amount of variability unexplained by your model. The error sum of squares is equal to  $\sum(Y_i - \hat{Y}_i)^2$ .

Total variability is related to the difference between the observed values and the mean of the response variable. The corrected total sum of squares is the sum of the explained and unexplained variability. The corrected total sum of squares is equal to  $\sum(Y_i - \bar{Y})^2$ .

## Model Hypothesis Test

### Null Hypothesis:

- The simple linear regression model does not fit the data better than the baseline model.
- $\beta_1 = 0$

### Alternative Hypothesis:

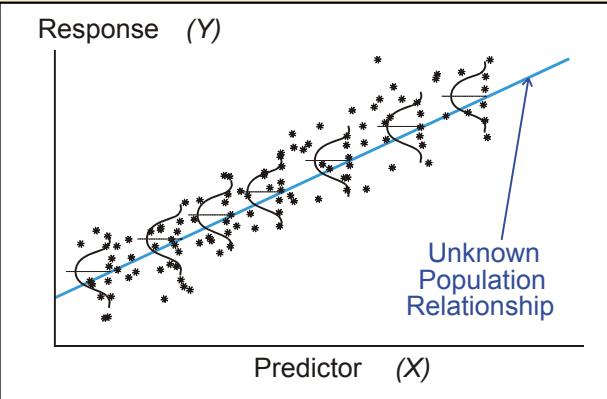
- The simple linear regression model does fit the data better than the baseline model.
- $\beta_1 \neq 0$

39

If the estimated simple linear regression model does **not** fit the data better than the baseline model, you fail to reject the null hypothesis. Thus, you **do not** have enough evidence to say that the slope of the regression line in the population is **not** 0 and that the predictor variable explains a significant amount of variability in the response variable.

If the estimated simple linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that the slope of the regression line in the population is **not** 0 and that the predictor variable explains a significant amount of variability in the response variable.

## Assumptions of Simple Linear Regression



40

One of the assumptions of simple linear regression is that the mean of the response variable is linearly related to the value of the predictor variable. In other words, a straight line connects the means of the response variable at each value of the predictor variable.

The other assumptions are the same as the assumptions for ANOVA: the error terms are normally distributed, have equal variances, and are independent at each value of the predictor variable.



The verification of these assumptions is discussed in a later chapter.

## The REG Procedure

General form of the REG procedure:

```
PROC REG DATA=SAS-data-set <options>;
   MODEL dependent(s)=regressor(s) </ options>;
RUN;
```

41

The REG procedure enables you to fit regression models to your data.

Selected REG procedure statement:

MODEL specifies the response and predictor variables. The variables must be numeric.

-  PROC REG supports RUN-group processing, which means that the procedure stays active until a PROC, DATA, or QUIT statement is encountered. This enables you to submit additional statements followed by another RUN statement without resubmitting the PROC statement.



## Performing Simple Linear Regression

Example: Because there is an apparent linear relationship between **Oxygen\_Consumption** and **Performance**, perform a simple linear regression analysis with **Oxygen\_Consumption** as the response variable.

```
/* c3demo05 */
proc reg data=sasuser.b_fitness;
  model oxygen_consumption=performance;
  title 'Simple Linear Regression of Oxygen Consumption '
    'and Performance';
run;
quit;
```

PROC REG Output

Simple Linear Regression of Oxygen Consumption and Performance

The REG Procedure

Model: MODEL1

Dependent Variable: Oxygen\_Consumption

Number of Observations Read	31
Number of Observations Used	31

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.34150	635.34150	85.22	<.0001
Error	29	216.21305	7.45562		
Corrected Total	30	851.55455			

Root MSE	2.73050	R-Square	0.7461
Dependent Mean	47.37581	Adj R-Sq	0.7373
Coeff Var	5.76349		

### Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	35.57526	1.36917	25.98	<.0001
Performance	1	1.47507	0.15979	9.23	<.0001

The Number of Observations Read and the Number of Observations Used are the same, indicating that no missing values were detected for **Oxygen\_Consumption** and **Performance**.

The Analysis of Variance (ANOVA) table provides an analysis of the variability observed in the data and the variability explained by the regression line.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.34150	635.34150	85.22	<.0001
Error	29	216.21305	7.45562		
Corrected Total	30	851.55455			

The ANOVA table for simple linear regression is divided into six columns.

- Source                labels the source of variability.
  - Model                is the variability explained by your model.
  - Error                is the variability unexplained by your model.
  - Corrected Total     is the total variability in the data.
- DF                    is the degrees of freedom associated with each source of variability.
- Sum of Squares      is the amount of variability associated with each source of variability.
- Mean Square          is the ratio of the sum of squares and the degrees of freedom. This value corresponds to the amount of variability associated with each degree of freedom for each source of variation.
- F Value              is the ratio of the mean square for the model and the mean square for the error. This ratio compares the variability explained by the regression line to the variability unexplained by the regression line.
- Pr > F              is the *p*-value associated with the *F* value.

The *F* value is testing whether the slope of the predictor variable is equal to 0. The *p*-value is small (less than .05), so you have enough evidence at the .05 significance level to reject the null hypothesis. Thus, you can conclude that the simple linear regression model fits the data better than the baseline model. In other words, **Performance** explains a significant amount of variability of **Oxygen\_Consumption**.

The second part of the output provides summary measures of fit for the model.

Root MSE	2.73050	R-Square	0.7461
Dependent Mean	47.37581	Adj R-Sq	0.7373
Coeff Var	5.76349		

- R-Square      the coefficient of determination, usually referred to as the  $R^2$  value. This value is
- between 0 and 1.
  - the proportion of variability observed in the data explained by the regression line. In this example, the value is 0.7461, which means that the regression line explains 75% of the total variation in the response values.
  - the square of the Pearson correlation coefficient.
- Root MSE      the root mean square error is an estimate of the standard deviation of the response variable at each value of the predictor variable. It is the square root of the MSE.
- Dependent Mean      the overall mean of the response variable,  $\bar{Y}$ .
- Coeff Var      the coefficient of variation is the size of the standard deviation relative to the mean. The coefficient of variation is
- calculated as  $\left( \frac{\text{RootMSE}}{\bar{Y}} \right) * 100$
  - a unitless measure, so it can be used to compare data that has different units of measurement or different magnitudes of measurement.
- Adj R-Sq      the adjusted  $R^2$  is the  $R^2$  that is adjusted for the number of parameters in the model. This statistic is useful in multiple regression and is discussed in a later section.

The Parameter Estimates table defines the model for your data.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	35.57526	1.36917	25.98	<.0001
Performance	1	1.47507	0.15979	9.23	<.0001

- DF represents the degrees of freedom associated with each term in the model.
- Parameter Estimate is the estimated value of the parameters associated with each term in the model.
- Standard Error is the standard error of each parameter estimate.
- t Value is the *t* statistic, which is calculated by dividing the parameters by their corresponding standard errors.
- Pr > |t| is the *p*-value associated with the *t* statistic. It tests whether the parameter associated with each term in the model is different from 0. For this example, the slope for the predictor variable is statistically different from 0. Thus, you can conclude that the predictor variable explains a significant portion of variability in the response variable.

Because the estimate of  $\beta_0=35.58$  and  $\beta_1=1.48$ , the estimated regression equation is given by Predicted **Oxygen\_Consumption** = 35.58 + 1.48(**Performance**).

The model indicates that an increase of one unit for **Performance** amounts to a 1.48 increase in **Oxygen\_Consumption**. However, this equation is appropriate only in the range of values you observed for the variable **Performance**.

The parameter estimates table also shows that the intercept parameter is not equal to 0. However, the test for the intercept parameter only has practical significance when the range of values for the predictor variable includes 0. In this example, the test could have practical significance because **Performance**=0 is inside the range of values you are considering (**Performance** ranges from 0 to 14).

## Producing Predicted Values

What is `Oxygen_Consumption` when  
`Performance` is 3 or 6 or 9?

45

One objective in regression analysis is to predict values of the response variable given values of the predictor variables. You can obviously use the estimated regression equation to produce predicted values, but if you want a large number of predictions, this can be cumbersome.

To produce predicted values in PROC REG, follow these steps:

1. Create a data set with the values of the independent variable for which you want to make predictions.
2. Concatenate the data in the step above with the original data set.
3. Fit a simple linear regression model to the new data set and specify the P option in the MODEL statement. Because the observations added in the previous step contain missing values for the response variable, PROC REG does not include these observations when fitting the regression model. However, PROC REG does produce predicted values for these observations.



## Producing Predicted Values

Example: Produce predicted values of **Oxygen\_Consumption** when **Performance** is 0, 3, 6, 9, and 12.

```
/* c3demo06 */
data need_predictions;
  input performance @@;
  datalines;
0 3 6 9 12
;
run;

data predoxy;
  set sasuser.b_fitness
    need_predictions;
run;

proc reg data=predoxy;
  model oxygen_consumption=performance / p;
  id performance;
  title 'Oxygen_Consumption=Performance with Predicted Values';
run;
quit;
```

Selected REG procedure statement:

ID specifies a variable to label observations in the output produced by certain MODEL statement options.

Selected MODEL statement option:

P prints the values of the response variable, the predicted values, and the residual values.

Partial PROC REG Output

```
Oxygen_Consumption=Performance with Predicted Values
```

```
The REG Procedure
```

```
Model: MODEL1
```

```
Dependent Variable: Oxygen_Consumption
```

Number of Observations Read	36
Number of Observations Used	31
Number of Observations with Missing Values	5

Notice that 36 observations were read; 31 were used and 5 had missing values. The observations in **need\_predictions** had missing values for **Oxygen\_Consumption**, so they were eliminated from the analysis.

## Partial PROC REG Output

Obs	Performance	Dependent Variable	Predicted Value	Residual
.	.	.	.	.
32	0	.	35.5753	.
33	3	.	40.0005	.
34	6	.	44.4257	.
35	9	.	48.8509	.
36	12	.	53.2761	.

Because you specified **Performance** in the ID statement, the values of this variable appear in the first column.

The output shows that the estimated value of **Oxygen\_Consumption** is 35.58 when **Performance** equals 0. However, when **Performance** is 12, the predicted **Oxygen\_Consumption** is 53.28.

If you have a large data set and have already fitted the regression model, a more efficient way to produce predicted values is in a DATA step. You can either write the parameter estimates in the DATA step or use the OUTTEST= option in PROC REG. Here is an example program.

```
/* c3demo06_n */
data _null_;
  input performance @@;
  oxygen_consumption=35.57526+1.47507*performance;
  put performance= oxygen_consumption=;
  datalines;
0 3 6 9 12
;
run;
```

## SAS Log

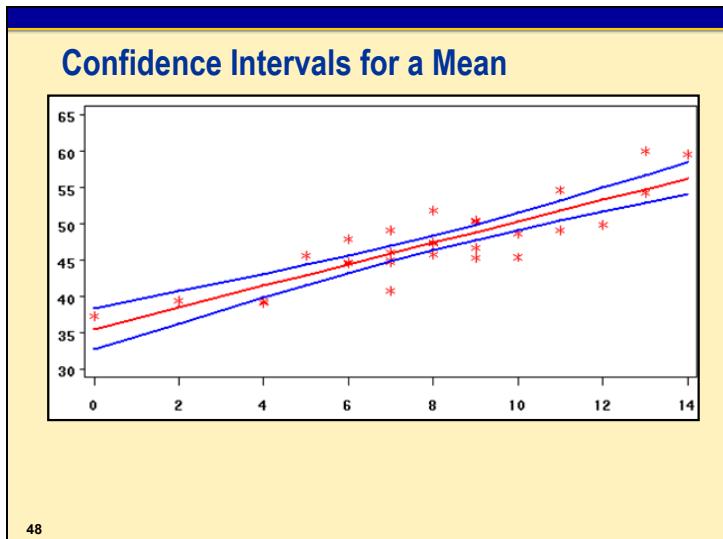
```
47  /* c3demo06_n.sas */
48  data _null_;
49    input performance @@;
50    oxygen_consumption=35.57526+1.47507*performance;
51    put performance= oxygen_consumption=;
52    datalines;

performance=0 oxygen_consumption=35.57526
performance=3 oxygen_consumption=40.00047
performance=6 oxygen_consumption=44.42568
performance=9 oxygen_consumption=48.85089
performance=12 oxygen_consumption=53.2761
NOTE: SAS went to a new line when INPUT statement reached past the end of a line.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.01 seconds

54  ;
55  run;
```



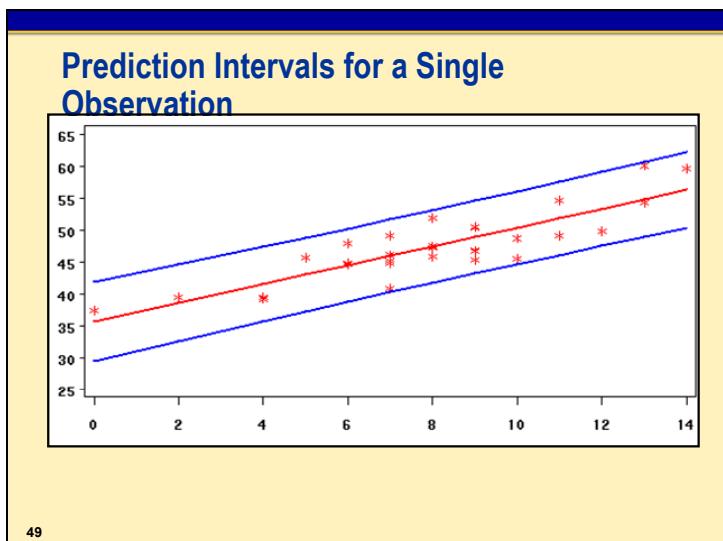
Choose only values within or near the range of the predictor variable when you are predicting new values for the response variable. For this example, the values of the variable **Performance** range from 0 to 14. Therefore, it is unwise to predict the value of **Oxygen\_Consumption** for a **Performance** rating of 100. The reason is that the relationship between the predictor variable and the response variable might be different beyond the range of your data.



48

To assess the level of precision around the mean estimates of **Oxygen\_Consumption**, you can produce confidence intervals around the means.

- A 95% confidence interval for the mean says that you are 95% confident your interval contains the population mean of Y for a particular X.
- Confidence intervals become wider as you move away from the mean of the independent variable. This reflects the fact that your estimates become more variable as you move away from the means of X and Y.



49

Suppose that the mean **Oxygen\_Consumption** at a fixed value of **Performance** is not the focus. If you are interested in establishing an inference on a future single observation, you need a prediction interval.

- A 95% prediction interval is one that you are 95% confident will contain a new observation.
- Prediction intervals are wider than confidence intervals because single observations have more variability than sample means.



## Producing Confidence Intervals

Example: Invoke PROC REG and produce confidence intervals for the mean and individual values of **Performance**.

```
/* c3demo07 */
options ps=50 ls=76;
options reset=all fontres=presentation ftext=swissb htext=1.5;

proc reg data=predoxy;
  model oxygen_consumption=performance / clm cli alpha=.05;
  id name performance;
  plot oxygen_consumption*performance / conf pred;
  symbol1 c=red v=dot;
  symbol2 c=red;
  symbol3 c=blue;
  symbol4 c=blue;
  symbol5 c=green;
  symbol6 c=green;
  title;
run;
quit;
```

Selected REG procedure statement:

PLOT prints scatter plots with y-variables on the vertical axis and x-variables on the horizontal axis.

Selected PROC REG statement option:

LINEPRINTER creates plots requested as line printer plots. If you do **not** specify this option, requested plots are created on a high-resolution graphics device. This option is required if plots are requested and you do not have SAS/GRAF software.

Selected MODEL statement options:

CLM produces all P option output, plus standard errors of the predicted values, and upper and lower 95% confidence bounds for the mean at each value of the predictor variable.

CLI produces all P option output, plus standard errors of the predicted values, and upper and lower 95% prediction bounds at each value of the predictor variable.

ALPHA= sets the significance level used for the construction of confidence intervals.

Selected PLOT statement options:

CONF requests overlaid plots of confidence intervals.

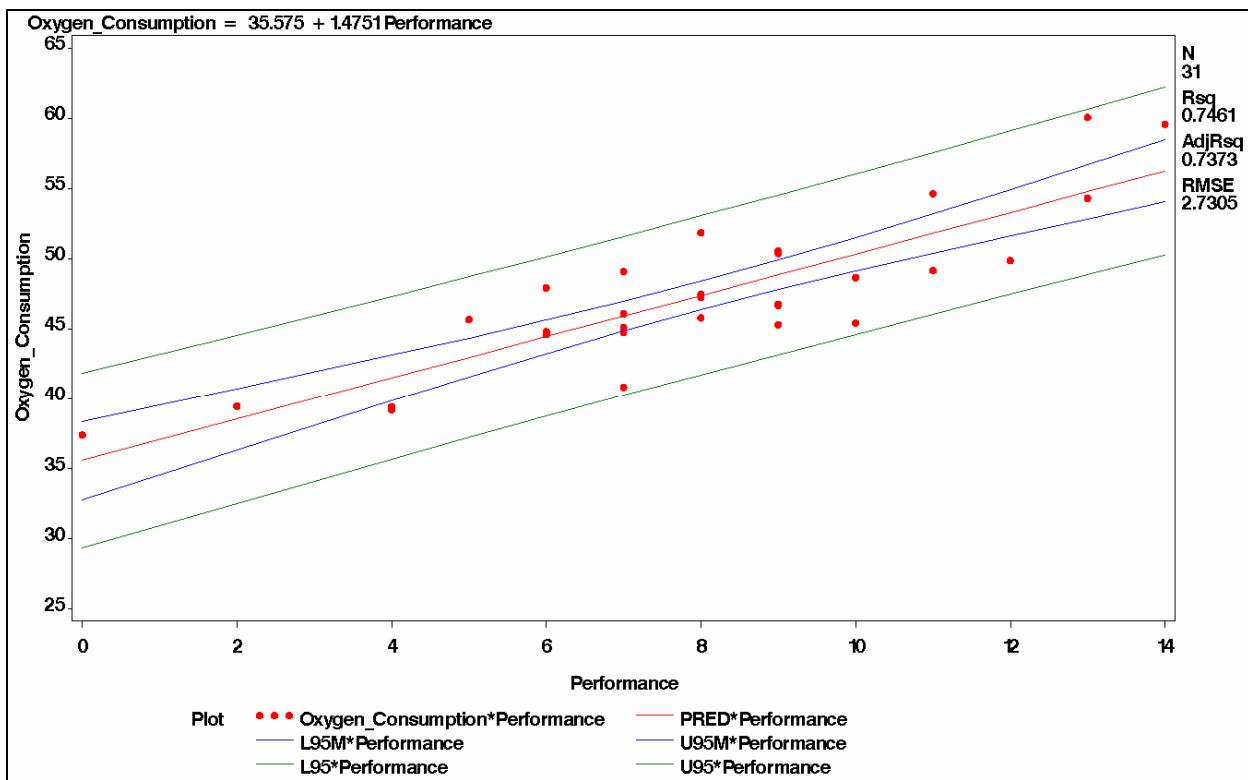
PRED requests overlaid plots of prediction intervals.

## Partial PROC REG Output

The REG Procedure							
Model: MODEL1							
Dependent Variable: Oxygen_Consumption							
Output Statistics							
Obs	Name	Performance	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL	Mean
32		0	.	35.5753	1.3692	32.7750	38.3755
33		3	.	40.0005	0.9375	38.0831	41.9178
34		6	.	44.4257	0.5854	43.2285	45.6228
35		9	.	48.8509	0.5158	47.7960	49.9058
36		12	.	53.2761	0.8056	51.6284	54.9238
Output Statistics							
Obs	Name	Performance	95% CL	Predict	Residual		
32		0	29.3280	41.8225	.		
33		3	34.0960	45.9049	.		
34		6	38.7143	50.1370	.		
35		9	43.1676	54.5341	.		
36		12	47.4536	59.0986	.		

When **Performance** is 6,

- the confidence interval for the mean of **Oxygen\_Consumption** is (43.23, 45.62)
- the prediction interval for **Oxygen\_Consumption** is (38.71, 50.14).



The data, regression line, confidence intervals, and predictions intervals are plotted in the graph above.



**Refer to Exercise 2 for Chapter 3 in Appendix A.**

### 3.3 Concepts of Multiple Regression

#### Objectives

- Explain the mathematical model for multiple regression.
- Describe the main advantage of multiple regression versus simple linear regression.
- Explain the standard output from the REG procedure.
- Describe common pitfalls of multiple linear regression.

56

#### Multiple Linear Regression with Two Variables

**Variables:** the two-variable model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

$Y$  is the dependent variable.

$X_1$  and  $X_2$  are the independent or predictor variables.

$\varepsilon$  is the error term.

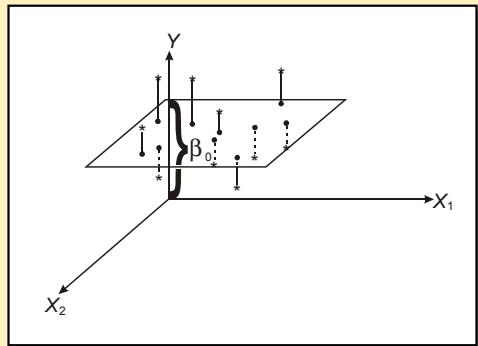
$\beta_0$ ,  $\beta_1$ , and  $\beta_2$  are unknown parameters.

57

In simple linear regression, you can model the relationship between the two variables (two dimensions) with a line (one dimension).

For the two-variable model, you can model the relationship of three variables (three dimensions) with a plane (two dimensions).

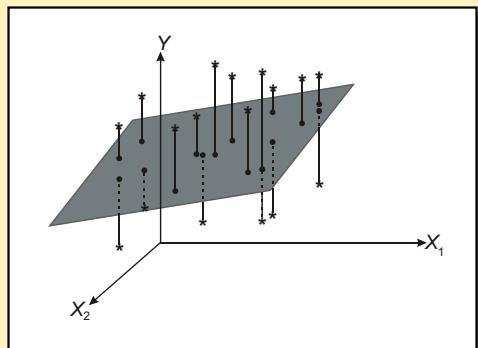
### Picturing the Model: No Relationship



58

If there is no relationship among Y and X<sub>1</sub> and X<sub>2</sub>, the model is a horizontal plane passing through the point (Y =  $\beta_0$ , X<sub>1</sub> = 0, X<sub>2</sub> = 0).

### Picturing the Model: A Relationship



59

If there is a relationship among Y and X<sub>1</sub> and X<sub>2</sub>, the model is a sloping plane passing through three points:

- (Y =  $\beta_0$ , X<sub>1</sub> = 0, X<sub>2</sub> = 0)
- (Y =  $\beta_0 + \beta_1$ , X<sub>1</sub> = 1, X<sub>2</sub> = 0)
- (Y =  $\beta_0 + \beta_2$ , X<sub>1</sub> = 0, X<sub>2</sub> = 1)

## The Multiple Linear Regression Model

In general, you model the dependent variable  $Y$  as a linear function of  $k$  independent variables, (the  $X$ s) as

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

60

You investigate the relationship of  $k + 1$  variables ( $k + 1$  dimensions) with a  $k$ -dimensional surface.

The multiple general linear model is not restricted to modeling only planes. By using higher order terms, such as quadratic or cubic powers of the  $X$ s or cross products of one  $X$  with another, more complex surfaces than planes can be modeled.

In the examples, the models are limited to relatively simple surfaces, such as planes.

-  The model has  $p = k + 1$  parameters (the  $\beta$ s) because of the intercept,  $\beta_0$ .

## Model Hypothesis Test

### Null Hypothesis:

- The regression model does not fit the data better than the baseline model.
- $\beta_1 = \beta_2 = \dots = \beta_k = 0$

### Alternative Hypothesis:

- The regression model does fit the data better than the baseline model.
- Not all  $\beta$ s equal zero.

61

If the estimated linear regression model does **not** fit the data better than the baseline model, you fail to reject the null hypothesis. Thus, you do **not** have enough evidence to say that all of the slopes of the regression in the population are **not** 0 and that the predictor variables explain a significant amount of variability in the response variable.

If the estimated linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that at least one slope of the regression in the population is **not** 0 and that at least one predictor variable explains a significant amount of variability in the response variable.

## Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a linear function of the Xs.
- The random error term,  $\varepsilon$ , is assumed to have a normal distribution with a mean of zero.
- The random error term,  $\varepsilon$ , is assumed to have a constant variance,  $\sigma^2$ .
- The errors are independent.

63

Techniques to evaluate the validity of these assumptions are discussed in a later chapter.

Because of the central limit theorem, the assumption that the errors are normally distributed is not as restrictive as you might think.

 You also estimate  $\sigma^2$  from the data.

## Multiple Linear Regression versus Simple Linear Regression

### Main Advantage

Multiple linear regression enables you to investigate the relationship among Y and several independent variables simultaneously.

### Main Disadvantages

Increased complexity makes it more difficult to

- ascertain which model is “best”
- interpret the models.

64

The advantages far outweigh the disadvantages. In practice, many responses depend on multiple factors that might interact in some way.

SAS tools help you decide upon a “best” model, a choice that can depend upon the purposes of the analysis as well as subject matter expertise.

## Common Applications

Multiple linear regression is a powerful tool for the following:

- Prediction – to develop a model to predict future values of a response variable (Y) based on its relationships with other predictor variables (Xs)
- Analytical or Explanatory Analysis – to develop an understanding of the relationships between the response variable and predictor variables.

65

Even though multiple linear regression enables you to analyze many different experimental designs, ranging from simple to complex, you will focus on applications for analytical studies and predictive modeling. Other SAS procedures are better suited for analyzing experimental data.

The distinction between using multiple regression for an analytic analysis and prediction modeling is somewhat artificial. A model developed for prediction will probably be a good analytic model. Conversely, a model developed for an analytic study will probably be a good prediction model.

Myers (1999) actually refers to four applications of regression: prediction, variable screening, model specifications, and parameter estimation. The term *analytical analysis* is similar to Myers' parameter estimation application and variable screening.

## Prediction

The terms in the model, the values of their coefficients, and their statistical significance are of secondary importance.

The focus is on producing a model that is the best at predicting future values of Y as a function of the Xs. The predicted value of Y is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_k X_k$$

66

Most investigators do not ignore the terms in the model (the Xs), the values of their coefficients (the  $\beta$ s), or their statistical significance (the  $p$ -values). They use these statistics to help choose among models with different numbers of terms and predictive capabilities.

### Analytical or Explanatory Analysis

The focus is on understanding the relationship between the dependent variable and the independent variables.

Consequently, the statistical significance of the coefficients is important as well as the magnitudes and signs of the coefficients.

67

### Analytic Analysis Example

#### PREDICTORS

**Performance**

**Runtim**

**Age**

**Weight**

**Run\_Pulse**

**Rest\_Pulse**

**Maximum\_Pulse**

#### RESPONSE

**Oxygen\_Consumption**



68

An analyst knows from doing a simple linear regression that the measure of performance is an important variable in explaining the oxygen consumption capability of a club member.

The analyst is interested in investigating other information to ascertain whether other variables are important in explaining the oxygen consumption capability.

Recall that you did a simple linear regression on **Oxygen\_Consumption** with **Performance** as the independent variable.

The  $R^2$  for this model was 0.7461, which suggests that more of the variation in the oxygen consumption is still unexplained.

Consequently, adding other variables to the model, such as **Runtim** or **Age**, might provide a significantly better model.



## Fitting a Multiple Linear Regression Model

Example: Invoke PROC REG and perform multiple linear regression analysis of **Oxygen\_Consumption** on **Performance** and **Runtime**. Interpret the output for the two-variable model.

```
/* c3demo08 */
proc reg data=sasuser.b_fitness;
  model oxygen_consumption=performance runtime;
  title 'Multiple Linear Regression for b_fitness Data';
run;
quit;
```

The only required statement for PROC REG is the MODEL statement. The syntax for the MODEL statement is

**MODEL Y = X1 X2 ... Xk;**

where

Y is the dependent variable

X1 X2 ... Xk is a list of the independent variables that will be included in the model.

## PROC REG Output

Multiple Linear Regression for b_fitness Data					
The REG Procedure					
Model: MODEL1					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read 31					
Number of Observations Used 31					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	637.96565	318.98283	41.82	<.0001
Error	28	213.58890	7.62818		
Corrected Total	30	851.55455			
Root MSE 2.76192 R-Square 0.7492					
Dependent Mean 47.37581 Adj R-Sq 0.7313					
Coeff Var 5.82980					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	55.37940	33.79380	1.64	0.1125
Performance	1	0.85780	1.06475	0.81	0.4272
Runtime	1	-1.40429	2.39427	-0.59	0.5622

Examine the sections of the output separately.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	637.96565	318.98283	41.82	<.0001
Error	28	213.58890	7.62818		
Corrected Total	30	851.55455			

- Model DF is 2, the number of parameters minus 1.
- Error DF is 28, the total numbers of observations (31) minus the number of parameters in the model (3).
- Corrected Total DF is 30, the number of observations minus 1.
- Model Sum of Squares is the total variation in the Y explained by the model.
- Error Sum of Squares is the variation in the Y **not** explained by the model.
- Corrected Total Sum of Squares is the total variation in the Y.
- Model Mean Square is the Model Sum of Squares divided by the Model DF.
- Mean Square Error is the Error Sum of Squares divided by the Error DF and is an estimate of  $\sigma^2$ , the variance of the random error term.
- F Value is the (Mean Square Model)/(Mean Square Error).
- Pr > F is small; therefore, you reject  $H_0: \beta_1 = \beta_2 = 0$  and conclude that at least one  $\beta_i \neq 0$ .

The  $R^2$  for this model, 0.7492, is only slightly larger than the  $R^2$  for the model in which **Performance** is the only predictor variable, 0.7461.

The  $R^2$  always increases as you include more terms in the model. However, choosing the “best” model is not as simple as just making the  $R^2$  as large as possible.

The adjusted  $R^2$  is a measure similar to  $R^2$ , but it takes into account the number of terms in the model.

The adjusted  $R^2$  for this model is 0.7313, smaller than the adjusted  $R^2$  of 0.7373 for the **Performance** only model. This strongly suggests that the variable **Runtime** does not explain the oxygen consumption capacity if you know **Performance**.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	55.37940	33.79380	1.64	0.1125
Performance	1	0.85780	1.06475	0.81	0.4272
Runtime	1	-1.40429	2.39427	-0.59	0.5622

Using the estimates for  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  above, this model can be written as

$$\text{Oxygen_Consumption} = 55.3794 + 0.8578 * \text{Performance} - 1.40429 * \text{Runtime}$$

Both the *p*-values for **Performance** and **Runtime** are large, which suggests that neither slope is significantly different from 0.

The reason is that the test for  $\beta_i=0$  is conditioned on the other terms in the model. So the test for  $\beta_1=0$  is conditional on or adjusted for  $X_2$  (**Runtime**). Similarly, the test for  $\beta_2=0$  is conditional on  $X_1$  (**Performance**).

**Performance** was significant when it was the only term in the model, but is not significant when **Runtime** is included. This implies that the variables are correlated with each other.

The significance level of the test does **not** depend on the order in which you list the independent variables in the MODEL statement, but it does depend upon the variables included in the MODEL statement.

## Common Problems

Four common problems with regression are

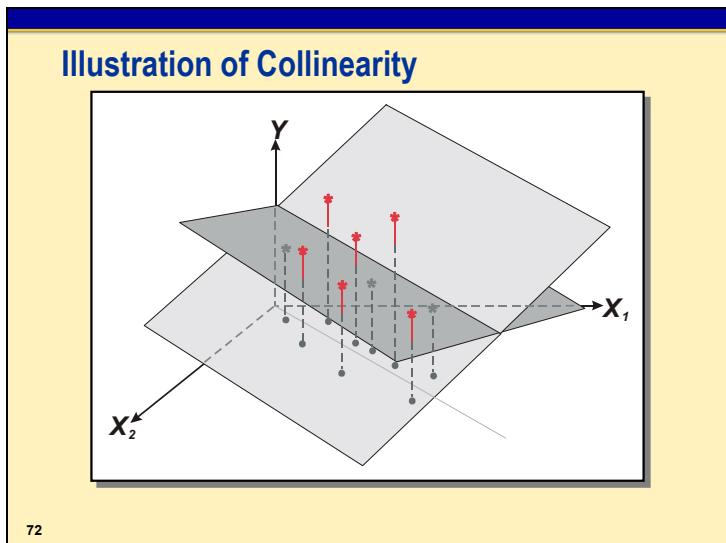
- nonconstant variance
- correlated errors
- influential observations
- collinearity.

71

The first three problems can arise in simple linear regression or multiple regression. The first two problems are always violations of the assumptions. The third can be a violation of the assumptions, but not always.

The fourth problem, however, is unique to multiple linear regression. *Collinearity* is redundant information among the independent variables. Collinearity is **not** a violation of assumptions of multiple regression.

When the number of potential Xs is large, the likelihood of collinearity becoming a problem increases.



$X_1$  and  $X_2$  almost follow a straight line  $X_1 = X_2$  in the  $(X_1, X_2)$  plane.

Consequently, one variable provides nearly as much information as the other does. They are redundant.

Why is this a problem? Two reasons exist.

1. Neither can appear to be significant when both are in the model; however, both can be significant when only one is in the model. Thus, collinearity can hide significant variables.
2. Collinearity also increases the variance of the parameter estimates and consequently increases prediction error.

When collinearity is a problem, the estimates of the coefficients are unstable. This means that they have a large variance. Consequently, the true relationship between Y and the Xs might be quite different from that suggested by the magnitude and sign of the coefficients.

The following three slides using Venn diagrams show another visualization of collinearity.

### Collinear Predictors in Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Model  $R^2=.25$   
Effect of  $X_1$ :  $p\text{-value}=.001$   
 $r_{y1}=.50$

$\text{continued...}$

73

The Venn diagram shows the variability of X and Y, and the extent to which variation in X explains variation in Y. The coefficient  $r_{y1}$  represents the correlation between Y and  $X_1$ . Consider that the simple linear regression of Y on  $X_1$ .  $X_1$  accounts for 25% of the variance in Y, as shown by the dark blue area of overlap.

### Collinear Predictors in Multiple Regression

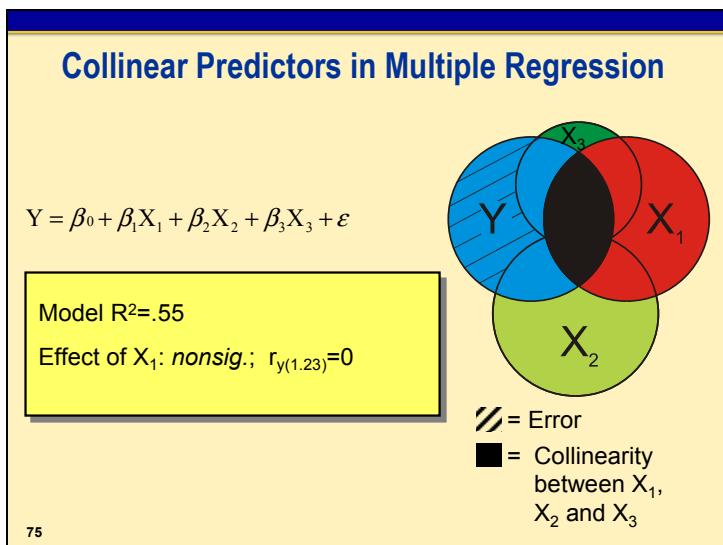
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Model  $R^2=.40$   
Effect of  $X_1$ :  $p\text{-value}=.01$   
 $r_{y(1,2)}=.25$

$\text{continued...}$

74

You suspect that  $X_2$  is associated with Y and add it to the multiple regression model. However,  $X_1$  and  $X_2$  are correlated with one another. The coefficient  $r_{y(1,2)}$  reflects the correlation of Y with  $X_1$ , controlling for the variable  $X_2$ .  $R^2$  increases when  $X_2$  is added to the model, but the individual effects of  $X_1$  and  $X_2$  appear smaller because the effect tests are based on partial correlation. In other words, only the unique variance accounted for by each variable is reflected in the effect tests.



Add one more variable to the model,  $X_3$ , that is correlated with  $X_1$ . The coefficient  $r_{y(1.23)}$  reflects the correlation between  $Y$  and  $X_1$  controlling for the variables  $X_2$  and  $X_3$ . Notice that the independent effect of  $X_1$  is no longer statistically significant, as all the variance in  $Y$  accounted for by  $X_1$  is also accounted for by other predictors in the model. The  $R^2$  for this model has increased with each new term in the model, but the individual effects have decreased as terms are added to the model.

This example is extreme, but it illustrates the importance of planning your model carefully and checking for collinearity among predictors.



**Refer to Exercises 3, 4, and 5 for Chapter 3 in Appendix A.**

## 3.4 Model Building and Interpretation

### Objectives

- Explain the REG procedure options for model selection.
- Describe model selection options and interpret output to evaluate the fit of several models.

81

### Model Selection

Eliminating one variable at a time manually for

- small data sets is a reasonable approach
- large data sets can take an extreme amount of time.

82

The exercises are designed to walk you through a model selection process. You start with all the variables in the **b\_fitness** data set and eliminate the least significant terms.

For this small example, a model can be developed in a reasonable amount of time. If you start with a large model, however, eliminating one variable at a time can take an extreme amount of time.

You continue this process until only terms with  $p$ -values less than some number, such as 0.10 or 0.05, remain.

## Model Selection Options

The SELECTION= option in the MODEL statement of PROC REG supports these model selection techniques:

### All-possible regressions ranked using

- RSQUARE, ADJRSQ, or CP

### Stepwise selection methods

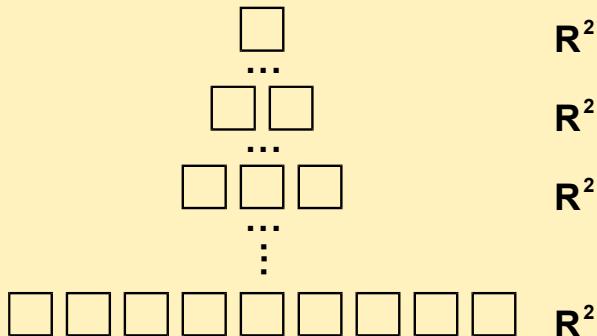
- STEPWISE, FORWARD, or BACKWARD

SELECTION=NONE is the default.

83

## RSQUARE Selection Option

### $R^2$ Selection



84

In the **b\_fitness** data set, there are 7 possible independent variables. Therefore, there are  $2^7 - 1 = 127$  possible regression models. There are 7 possible one-variable models, 21 possible two-variable models, 35 possible three-variable models, and so on.

You will only look at the best four as measured by the model  $R^2$  for  $k=1, 2, 3, \dots, 7$ . The BEST= option only reduces the output. All regressions are still calculated.

If there were 20 possible independent variables, there would be over 1,000,000 models. In a later demonstration, you see another technique that does not have to examine all the models to help you choose a set of candidate models.

### Mallows' $C_p$

- Mallows'  $C_p$  is a simple indicator of model bias.  
Models with a large  $C_p$  are biased.
- Look for models with  $C_p \leq p$ , where  $p$  equals the number of parameters in the model, including the intercept.

Mallows recommends choosing the first model where  $C_p$  approaches  $p$ .

85

$$\text{Mallows' } C_p \text{ (1973) is estimated by } C_p = p + \frac{(MSE_p - MSE_{\text{full}})(n - p)}{MSE_{\text{full}}}$$

where

$MSE_p$  is the mean squared error for the model with  $p$  parameters

$MSE_{\text{full}}$  is the mean squared error for the full model used to estimate the true residual variance

$n$  is the number of observations.

Bias in this context refers to the model underfitting the sample. In other words, important variables are left out of the model.

### Hocking's Criterion

Hocking (1976) suggests selecting a model based on the following:

- $C_p \leq p$  for prediction
- $C_p \leq 2p - p_{\text{full}} + 1$  for parameter estimation

86



## Automatic Model Selection

Example: Invoke PROC REG to produce a regression of **Oxygen\_Consumption** on all the other variables in the **sasuser.b\_fitness** data set.

```
/* c3demo09 */
options ps=50 ls=97;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc reg data=sasuser.b_fitness;
  ALL_REG: model oxygen_consumption
    = performance runtime age weight
      run_pulse rest_pulse maximum_pulse
    / selection=rsquare adjrsq cp best=4;
  plot cp.*np. /
    nomodel nostat
    vaxis=0 to 30 by 5
    haxis=2 to 7 by 1 /* p=0,p=1 do not add information */
    cmallows=red
    chocking=blue;
  symbol v=plus color=green h=2;
  title h=2 'Best=4 Models Using All Regression Option';
run;
quit;
```

Selected MODEL statement options:

**SELECTION=** enables you to choose the different selection methods.

Selected **SELECTION=** option methods:

**RSQUARE** tells PROC REG to use the model R<sup>2</sup> to rank the model from best to worst for a given number of variables.

**ADJRSQ** prints the adjusted R<sup>2</sup> for each model.

**CP** prints Mallows' C<sub>p</sub> statistic for each model.

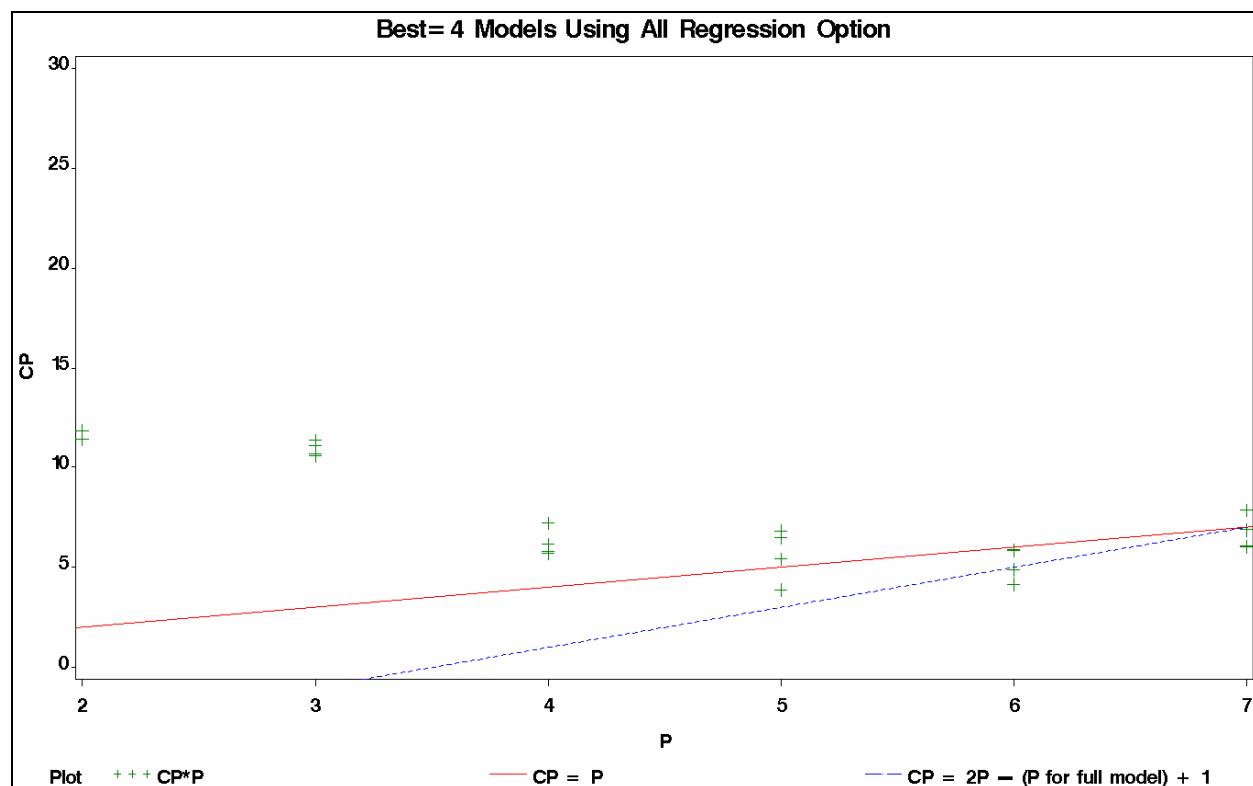
**BEST=n** limits the output to only the best *n* models for a fixed number of variables.

The PLOT statement specifies that the values of the Mallows'  $C_p$  statistic (cp.) be plotted using the vertical axis and that the number of terms in the model (np.) be plotted using the horizontal axis.

Selected PLOT statement options:

- NOMODEL suppresses the model from the graph.
- NOSTAT suppresses n, Rsq, Adjrsq, and RMSE from the graph
- VAXIS= specifies the range for the vertical axis.
- HAXIS= specifies the range for the horizontal axis. The default is the range of the data.
- CMALLOWS= requests a  $C_p = p$  reference line and specifies a color.
- CHOCKING= requests a  $2p - p_{\text{full}} + 1$  reference line in addition to the CMALLOWS reference line and specifies a color.

PROC REG Output



The line  $C_p = p$  is plotted to help you identify models that satisfy the criterion  $C_p \leq p$  for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion  $C_p \leq 2p - p_{\text{full}} + 1$  for parameter estimation.

Use the graph and review the output to select a relatively short list of models that satisfy the criterion appropriate for your objective. The first model to fall below the line for Mallows' criterion has five parameters. The first model to fall below Hocking's criterion has six parameters.

The models are ranked by their  $R^2$ .

## PROC REG Output

Best=4 Models Using All Regression Option				
The REG Procedure				
Model: ALL_REG				
Dependent Variable: Oxygen_Consumption				
R-Square Selection Method				
Number of Observations Read 31				
Number of Observations Used 31				
Number in Model      Adjusted R-Square      R-Square      C(p)      Variables in Model				
1	0.7461	0.7373	11.3942	Performance
1	0.7434	0.7345	11.8074	Runtime
1	0.1595	0.1305	100.1000	Rest_Pulse
1	0.1585	0.1294	100.2529	Run_Pulse
-----				
2	0.7647	0.7479	10.5794	Runtime Age
2	0.7640	0.7472	10.6839	Performance Run_Pulse
2	0.7614	0.7444	11.0743	Runtime Run_Pulse
2	0.7597	0.7425	11.3400	Performance Age
-----				
3	0.8101	0.7890	5.7169	Runtime Run_Pulse Maximum_Pulse
3	0.8096	0.7884	5.7963	Runtime Age Run_Pulse
3	0.8072	0.7858	6.1523	Performance Run_Pulse Maximum_Pulse
3	0.8003	0.7781	7.2046	Performance Age Run_Pulse
-----				
4	0.8355	0.8102	3.8790	Runtime Age Run_Pulse Maximum_Pulse
4	0.8253	0.7984	5.4191	Performance Age Run_Pulse Maximum_Pulse
4	0.8181	0.7901	6.5036	Performance Weight Run_Pulse Maximum_Pulse
4	0.8160	0.7877	6.8265	Runtime Weight Run_Pulse Maximum_Pulse
-----				
5	0.8469	0.8163	4.1469	Runtime Age Weight Run_Pulse Maximum_Pulse
5	0.8421	0.8105	4.8787	Performance Age Weight Run_Pulse Maximum_Pulse
5	0.8356	0.8027	5.8571	Runtime Age Run_Pulse Rest_Pulse Maximum_Pulse
5	0.8355	0.8026	5.8738	Performance Runtime Age Run_Pulse Maximum_Pulse
-----				
6	0.8476	0.8096	6.0381	Performance Runtime Age Weight Run_Pulse Maximum_Pulse
6	0.8475	0.8094	6.0633	Runtime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse
6	0.8421	0.8026	6.8779	Performance Age Weight Run_Pulse Rest_Pulse Maximum_Pulse
6	0.8356	0.7945	7.8565	Performance Runtime Age Run_Pulse Rest_Pulse Maximum_Pulse
-----				
7	0.8479	0.8016	8.0000	Performance Runtime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse

In this example  $p_{\text{full}}$  equals 8—that is, 7 variables plus the intercept.

For  $p = 5$  (Number in Model = 4), the "best" predictive model has a  $C_p = 3.879 < 4$ , satisfying Mallows' criterion (**Oxygen\_Consumption = Runtime Age Weight Run\_Pulse Maximum\_Pulse**).

To determine the best explanatory model based on Hocking's criterion, the following table was created:

<b>p=# terms in the current model, including the intercept</b>	<b>Number in Model (k)</b>	<b>Minimum <math>C_p</math> with <math>p</math> terms</b>	<b>Hocking's Criterion : <math>2*p - 8 + 1 =</math> <math>2*p - 7</math></b>	<b><math>C_p &lt;</math> Hocking's Criterion?</b>
2	1	11.3942	$2*2 - 7 = -3$	Not Applicable
3	2	10.5794	$2*3 - 7 = -1$	Not Applicable
4	3	5.7169	$2*4 - 7 = 1$	No
5	4	3.8790	$2*5 - 7 = 3$	No
6	5	4.1469	$2*6 - 7 = 5$	Yes

When  $p=6$ , the model

**Oxygen\_Consumption = Runtime Age Weight Run\_Pulse Maximum\_Pulse**

had the smallest  $C_p$  and will be considered the “best” explanatory model.

For  $p = 6$  (Number in Model=5), four models satisfy Mallows' criterion, but only two models also satisfy Hocking's criterion.

## All Possible Regression Models

The two best candidate models for  $p = 5$  and  $6$  include these independent variables:

$p = 5$  and  $C_p = 3.88$ :      **Runtime, Age,**  
                                 **Run\_Pulse,**  
                                 **Maximum\_Pulse**

$p = 6$  and  $C_p = 4.15$ :      **Runtime, Age,**  
                                 **Weight,**  
                                 **Run\_Pulse,**  
                                 **Maximum\_Pulse**

93

In practice, you might not want to limit your subsequent investigation to only the best model for a given number of terms. Some models might be essentially equivalent based on their  $R^2$  or other measures.

A limitation of the evaluation you have done thus far is that you do not know the magnitude and signs of the coefficients of the candidate models or their statistical significance.



## Estimating and Testing the Coefficients for the Selected Models

Example: Invoke PROC REG to compare the ANOVA tables and parameter estimates for the two-candidate models in the **sasuser.b\_fitness** data set.

```
/* c3demo10 */
proc reg data=sasuser.b_fitness;
  PREDICT: model oxygen_consumption
            = runtime age run_pulse maximum_pulse;
  EXPLAIN: model oxygen_consumption
            = runtime age weight run_pulse maximum_pulse;
  title 'Check "Best" Two Candidate Models';
run;
quit;
```

PROC REG can have more than one MODEL statement. You can assign a label to each MODEL statement to identify the output generated for each model.

## Output for the PREDICT Model

Check "Best" Two Candidate Models					
The REG Procedure Model: PREDICT					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read 31					
Number of Observations Used 31					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	711.45087	177.86272	33.01	<.0001
Error	26	140.10368	5.38860		
Corrected Total	30	851.55455			
Root MSE 2.32134 R-Square 0.8355					
Dependent Mean 47.37581 Adj R-Sq 0.8102					
Coeff Var 4.89984					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	97.16952	11.65703	8.34	<.0001
Runtime	1	-2.77576	0.34159	-8.13	<.0001
Age	1	-0.18903	0.09439	-2.00	0.0557
Run_Pulse	1	-0.34568	0.11820	-2.92	0.0071
Maximum_Pulse	1	0.27188	0.13438	2.02	0.0534

The  $R^2$  and adjusted  $R^2$  are the same as calculated during the model selection program. If there are missing values in the data set, however, this might not be true.

The model  $F$  is large and highly significant. **Age** and **Maximum\_Pulse** are not significant at the 0.05 level of significance. However, all terms are significant at alpha=0.10.

The adjusted  $R^2$  is close to the  $R^2$ , which suggests that there are not too many variables in the model.

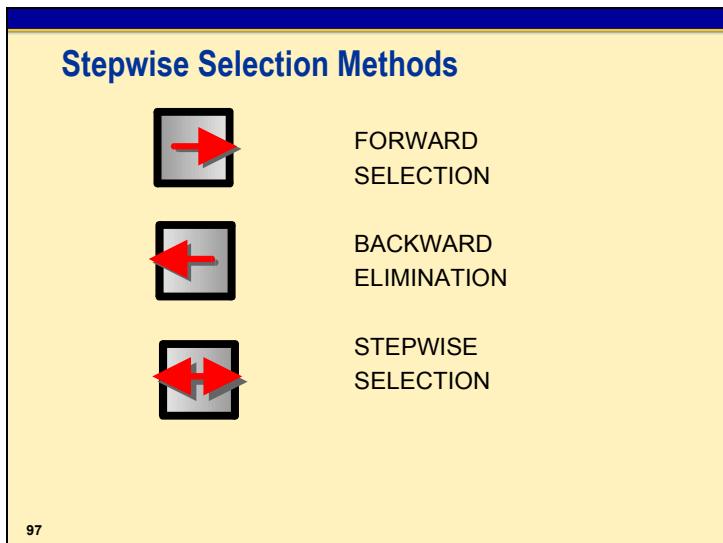
## Output for the EXPLAIN Model

Check "Best" Two Candidate Models					
The REG Procedure Model: EXPLAIN					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read 31					
Number of Observations Used 31					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.20532	144.24106	27.66	<.0001
Error	25	130.34923	5.21397		
Corrected Total	30	851.55455			
Root MSE 2.28341 R-Square 0.8469					
Dependent Mean 47.37581 Adj R-Sq 0.8163					
Coeff Var 4.81978					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	101.33835	11.86474	8.54	<.0001
Runtime	1	-2.68846	0.34202	-7.86	<.0001
Age	1	-0.21217	0.09437	-2.25	0.0336
Weight	1	-0.07332	0.05360	-1.37	0.1836
Run_Pulse	1	-0.37071	0.11770	-3.15	0.0042
Maximum_Pulse	1	0.30603	0.13452	2.28	0.0317

The adjusted R<sup>2</sup> is slightly larger than in the PREDICT model and very close to the R<sup>2</sup>.

The model *F* is large, but smaller than in the PREDICT model. However, it is still highly significant. All terms included in the model are significant except **Weight**. Note that the *p*-values for **age**, **Run\_Pulse**, and **Maximum\_Pulse** are smaller in this model than they were in the PREDICT model.

Including the additional variable in the model changes the coefficients of the other terms and changes the *t* statistics for all.



The all-possible regression technique that was discussed can be computer intensive, especially if there are a large number of potential independent variables.

PROC REG also offers the following stepwise SELECTION= options:

- |          |  |
|----------|--|
| FORWARD  | first selects the best one-variable model. Then it selects the best two variables among those that contain the first selected variable. FORWARD continues this process, but stops when it reaches the point where no additional variables have a <i>p</i> -value level < 0.50. |
| BACKWARD | starts with the full model. Next, the variable that is least significant, given the other variables, is removed from the model. BACKWARD continues this process until all of the remaining variables have a <i>p</i> -value < 0.10.  |
| STEPWISE | works like a combination of the two. The default entry <i>p</i> -value is 0.15 and the default stay <i>p</i> -value is also 0.15.  |

The SLENTRY= and SLSTAY= options can be used to change the default values.



## Stepwise Regression

Example: Select a model for predicting **Oxygen\_Consumption** in the **sasuser.b\_fitness** data set by using the FORWARD, BACKWARD and STEPWISE methods.

```
/* c3demo11 */
proc reg data=sasuser.b_fitness;
  FORWARD: model oxygen_consumption
    = performance runtime age weight
      run_pulse rest_pulse maximum_pulse
    / selection=forward;
  BACKWARD: model oxygen_consumption
    = performance runtime age weight
      run_pulse rest_pulse maximum_pulse
    / selection=backward;
  STEPWISE: model oxygen_consumption
    = performance runtime age weight
      run_pulse rest_pulse maximum_pulse
    / selection=stepwise;
  title 'Stepwise Regression Methods';
run;
quit;
```

## Partial PROC REG Output

Stepwise Regression Methods					
The REG Procedure					
Model: FORWARD					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read 31					
Number of Observations Used 31					
Forward Selection: Step 1					
Variable Performance Entered: R-Square = 0.7461 and C(p) = 11.3942					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.34150	635.34150	85.22	<.0001
Error	29	216.21305	7.45562		
Corrected Total	30	851.55455			
Parameter Standard					
Variable	Estimate	Error	Type II SS	F Value	Pr > F
Intercept	35.57526	1.36917	5033.48080	675.13	<.0001
Performance	1.47507	0.15979	635.34150	85.22	<.0001
Bounds on condition number: 1, 1					

## Partial PROC REG Output (continued)

Stepwise Regression Methods					
The REG Procedure					
Model: FORWARD					
Dependent Variable: Oxygen_Consumption					
Forward Selection: Step 2					
Variable Run_Pulse Entered: R-Square = 0.7640 and C(p) = 10.6839					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	650.60420	325.30210	45.33	<.0001
Error	28	200.95035	7.17680		
Corrected Total	30	851.55455			
Parameter Standard					
Variable	Estimate	Error	Type II SS	F Value	Pr > F
Intercept	48.60983	9.03851	207.58002	28.92	<.0001
Performance	1.39954	0.16511	515.66060	71.85	<.0001
Run_Pulse	-0.07327	0.05024	15.26270	2.13	0.1559
Bounds on condition number: 1.1091, 4.4366					
-----					
Forward Selection: Step 3					
...					

## Partial PROC REG Output (continued)

.. .							
<hr/>							
No other variable met the 0.5000 significance level for entry into the model.							
Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Performance	1	0.7461	0.7461	11.3942	85.22	<.0001
2	Run_Pulse	2	0.0179	0.7640	10.6839	2.13	0.1559
3	Maximum_Pulse	3	0.0432	0.8072	6.1523	6.05	0.0206
4	Age	4	0.0181	0.8253	5.4191	2.69	0.1130
5	Weight	5	0.0168	0.8421	4.8787	2.66	0.1155
6	Runtime	6	0.0056	0.8476	6.0381	0.88	0.3587

The model selected at each step is printed and a summary of the sequence of steps is given at the end of the output. In the summary, the variables are listed in the order in which they were selected. The partial R<sup>2</sup> shows the increase in the model R<sup>2</sup> as each term was added.

The model that FORWARD selected has more variables than the models chosen using the all-regressions techniques.

In this example, only one variable was eliminated from the model. Remember that this is not always the case.

## Partial PROC REG Output (continued)

Stepwise Regression Methods					
The REG Procedure					
Model: BACKWARD					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read 31					
Number of Observations Used 31					
Backward Elimination: Step 0					
All Variables Entered: R-Square = 0.8479 and C(p) = 8.0000					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.03251	103.14750	18.32	<.0001
Error	23	129.52204	5.63139		
Corrected Total	30	851.55455			
Parameter Estimates					
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	93.33753	36.49782	36.82939	6.54	0.0176
Performance	0.25756	1.02373	0.35646	0.06	0.8036
Runtime	-2.08804	2.22856	4.94363	0.88	0.3585
Age	-0.21066	0.10519	22.58631	4.01	0.0571
Weight	-0.07741	0.05681	10.45445	1.86	0.1862
Run_Pulse	-0.36618	0.12299	49.91978	8.86	0.0067
Rest_Pulse	-0.01389	0.07114	0.21460	0.04	0.8469
Maximum_Pulse	0.30490	0.13990	26.74945	4.75	0.0398
Bounds on condition number: 54.342, 888.21					

## Stepwise Regression Methods

The REG Procedure

Model: BACKWARD

Dependent Variable: Oxygen\_Consumption

Backward Elimination: Step 1

Variable Rest\_Pulse Removed: R-Square = 0.8476 and C(p) = 6.0381

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	721.81791	120.30298	22.25	<.0001
Error	24	129.73665	5.40569		
Corrected Total	30	851.55455			

Variable	Parameter Estimate	Standard Error	Type I	II	SS	F Value	Pr > F
Intercept	90.83022	33.47159	39.80699		7.36	0.0121	
Performance	0.32048	0.95201	0.61258		0.11	0.7393	
Runtime	-1.98433	2.12049	4.73376		0.88	0.3587	
Age	-0.20470	0.09862	23.28867		4.31	0.0488	
Weight	-0.07689	0.05560	10.33766		1.91	0.1794	
Run_Pulse	-0.36818	0.12008	50.81482		9.40	0.0053	
Maximum_Pulse	0.30593	0.13697	26.96687		4.99	0.0351	

Bounds on condition number: 48.957, 700.99

## Partial PROC REG Output (continued)

...							
-----							
Backward Elimination: Step 3							
Variable Weight Removed: R-Square = 0.8355 and C(p) = 3.8790							
Analysis of Variance							
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F		
Model	4	711.45087	177.86272	33.01	<.0001		
Error	26	140.10368	5.38860				
Corrected Total	30	851.55455					
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F		
Intercept	97.16952	11.65703	374.42127	69.48	<.0001		
Runtime	-2.77576	0.34159	355.82682	66.03	<.0001		
Age	-0.18903	0.09439	21.61272	4.01	0.0557		
Run_Pulse	-0.34568	0.11820	46.08558	8.55	0.0071		
Maximum_Pulse	0.27188	0.13438	22.05933	4.09	0.0534		
	Bounds on condition number: 8.4426, 76.969						
-----	-----	-----	-----	-----	-----	-----	
	Backward Elimination: Step 3						
	All variables left in the model are significant at the 0.1000 level.						
	Summary of Backward Elimination						
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Rest_Pulse	6	0.0003	0.8476	6.0381	0.04	0.8469
2	Performance	5	0.0007	0.8469	4.1469	0.11	0.7393
3	Weight	4	0.0115	0.8355	3.8790	1.87	0.1836

Using the BACKWARD elimination option and the default  $p$ -value, three independent variables were eliminated.

## Partial PROC REG Output (continued)

Stepwise Regression Methods						
The REG Procedure						
Model: STEPWISE						
Dependent Variable: Oxygen_Consumption						
Number of Observations Read 31						
Number of Observations Used 31						
Stepwise Selection: Step 1						
Variable Performance Entered: R-Square = 0.7461 and C(p) = 11.3942						
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	635.34150	635.34150	85.22	<.0001	
Error	29	216.21305	7.45562			
Corrected Total	30	851.55455				
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F	
Intercept	35.57526	1.36917	5033.48080	675.13	<.0001	
Performance	1.47507	0.15979	635.34150	85.22	<.0001	
Bounds on condition number: 1, 1						
-----						
All variables left in the model are significant at the 0.1500 level.						
No other variable met the 0.1500 significance level for entry into the model.						
Summary of Stepwise Selection						
Step Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value Pr > F
1 Performance		1	0.7461	0.7461	11.3942	85.22 <.0001

Using the STEPWISE option and the default  $p$ -value, only one independent variable was included in the model.

## Stepwise Regression Models

FORWARD

**Performance,**  
**Runtim**e, **Age**, **Weight**,  
**Run\_Pulse**,  
**Maximum\_Pulse**

BACKWARD

**Runtim**e, **Age**,  
**Run\_Pulse**,  
**Maximum\_Pulse**

STEPWISE

**Performance**

100

## Comparison of Selection Methods

Stepwise regression uses fewer computer resources.

All-possible regression generates more candidate models that might have nearly equal R<sup>2</sup> statistics and C<sub>p</sub> statistics.

101

The stepwise regression methods have an advantage when there is a large number of independent variables.

With the all-possible regressions techniques, you can compare essentially equivalent models and use your knowledge of the data set and subject area to select a model that is more easily interpreted.

### Chapter 3 Modeling Summary

Technique	Model	R-Square	Adj R-Square	MSE
Simple Linear Regression	<b>Performance</b>	0.7461	0.7373	7.45562
Stepwise: FORWARD	<b>Performance</b> <b>Runtim</b> e <b>Age</b> <b>Weight</b>	0.8476	0.8096	5.40569

	<b>Run_pulse Maximum_pulse</b>			
Stepwise: BACKWARD	<b>Runtime Age Run_pulse Maximum_pulse</b>	0.8355	0.8102	5.38860
Stepwise: STEPWISE	<b>Performance</b>	0.7461	0.7373	7.45562
Mallows (Prediction) PREDICT	<b>Runtime Age Run_pulse Maximum_pulse</b>	0.8355	0.8102	5.38860
Hocking (Explanatory) EXPLAIN	<b>Runtime Age Weight Run_pulse Maximum_pulse</b>	0.8469	0.8163	5.21397



Refer to Exercise 6 for Chapter 3 in Appendix A.

## 3.5 Chapter Summary

Before performing an analysis, it is important to examine scatter plots and calculate correlation statistics. Scatter plots describe the relationship between two continuous variables. The Pearson correlation statistic measures the degree of linear relationship between two variables.

Simple linear regression defines the linear relationship between a continuous response variable and a continuous predictor variable. The assumptions of a linear regression analysis are

- the mean of the response variable is linearly related to the value of the predictor variable
- the observations are independent
- the error terms for each value of the predictor variable are normally distributed
- the error variances for each value of the predictor variable are equal.

You can verify these assumptions by

- examining a plot of the residuals versus the predicted values
- checking the residuals for normality.

When you perform a simple linear regression, the null hypothesis is that the simple linear regression does not fit the data better than the baseline model ( $\beta_1 = 0$ ). The alternative hypothesis is that the simple linear regression model does fit the data better than the baseline model ( $\beta_1 \neq 0$ ).

Multiple regression enables you to investigate the relationship between a response variable and several predictor variables simultaneously. The null hypothesis is that the slopes for all of the predictor variables are equal to zero ( $\beta_1 = \beta_2 = \dots = \beta_k = 0$ ). The alternative hypothesis is that at least one slope is not equal to zero. If you reject the null hypothesis, you must determine which of the independent variables have non-zero slopes and are, therefore, useful in the model.

The tests of the parameter estimates help you determine which slopes are non-zero, but they must be considered carefully. They test the significance of each variable when it is added to a model that already contains all of the other independent variables. Therefore, if independent variables in the model are correlated with one another, the significance of both variables can be hidden in these tests.

There are different model selection options. They can generally be divided into two types: all-possible regression options and stepwise options. With the all-possible regression options, regressions using all possible combinations of variables are calculated. All of the regressions are then ranked either by  $R^2$ , adjusted  $R^2$ , or Mallows'  $C_p$ . All-possible regression techniques can be computer intensive, especially if there are a large number of potential independent variables. Stepwise selection procedures help choose the independent variables that are most useful in explaining or predicting your dependent variable. Some of the stepwise selection methods are FORWARD, BACKWARD, and STEPWISE.

Four common problems with regression are nonconstant variance, correlated errors, influential observations, and collinearity. Collinearity is a problem unique to multiple regression. It can hide significant variables and increase the variance of the parameter estimates resulting in an unstable model.

```
PROC CORR DATA=SAS-data-set <options>;
  VAR variables;
  WITH variables;
RUN;
```

```
PROC REG DATA=SAS-data-set <options>;
  MODEL dependent(s)=regressor(s) </ options>;
  ID variable;
  PLOT y-variable*x-variable </ options>;
RUN;
QUIT;
```

# Chapter 4 Regression Diagnostics

<b>4.1 Examining Residuals .....</b>	<b>4-2</b>
<b>4.2 Influential Observations.....</b>	<b>4-14</b>
<b>4.3 Collinearity .....</b>	<b>4-24</b>
<b>4.4 Chapter Summary.....</b>	<b>4-43</b>

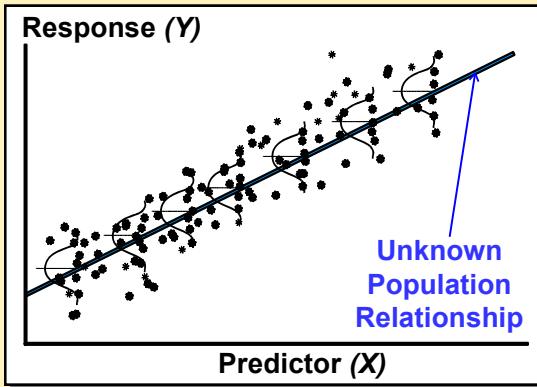
## 4.1 Examining Residuals

### Objectives

- Review the assumptions of linear regression.
- Examine the assumptions with scatter plots and residual plots.

3

### Assumptions for Regression

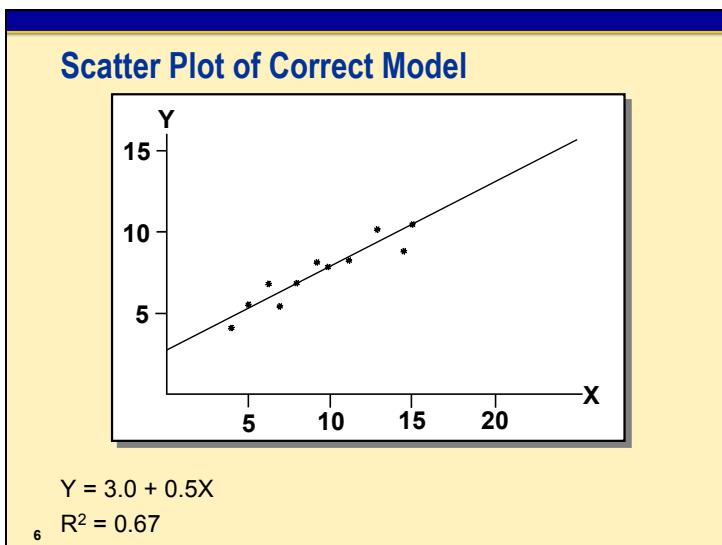


4

Recall that the model for the linear regression has the form  $Y = \beta_0 + \beta_1 X + \epsilon$ . When you perform a regression analysis, several assumptions about the error terms must be met to provide valid tests of hypothesis and confidence intervals. The assumptions are that the error terms

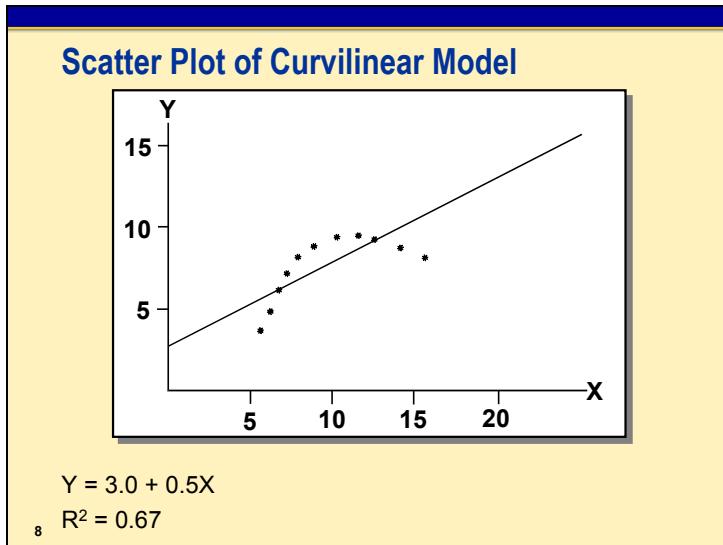
- have a mean of 0 at each value of the predictor variable
- are normally distributed at each value of the predictor variable
- have the same variance at each value of the predictor variable
- are independent.

You can use scatter plots and residual plots to help verify some of these assumptions.

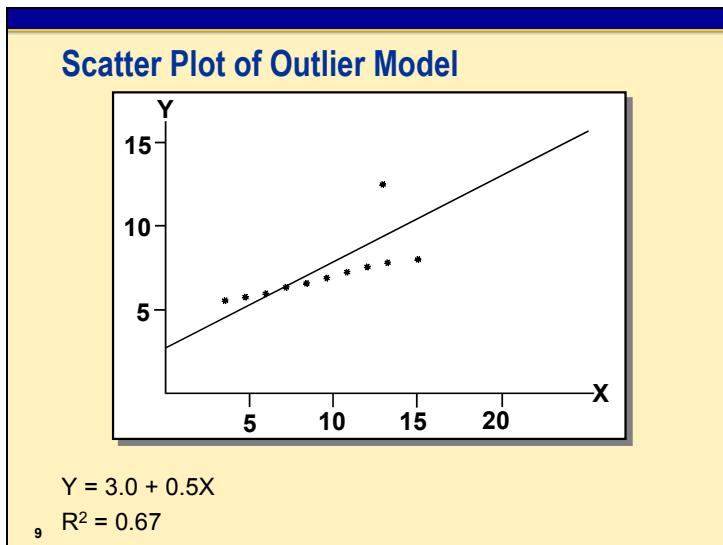


To illustrate the importance of plotting data, four examples were developed by Anscombe (1973). In each example, the scatter plot of the data values is different. However, the regression equation and the R<sup>2</sup> statistic are the same.

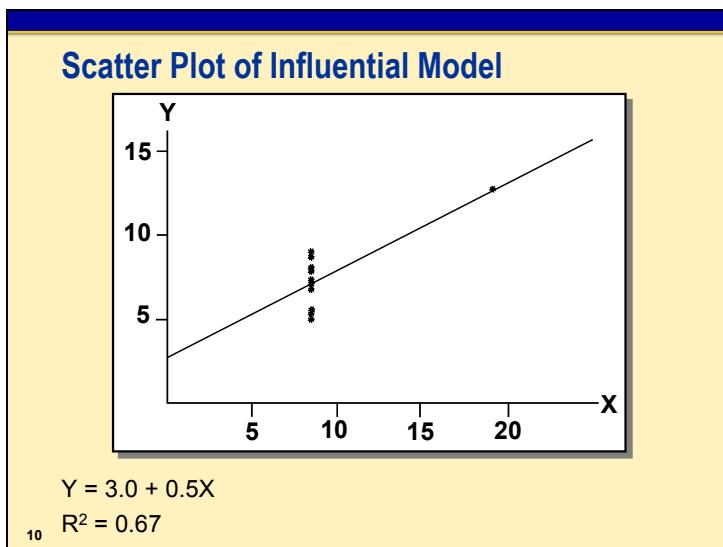
In the first plot, a regression line adequately describes the data.



In the second plot, a simple linear regression model is not appropriate because you are fitting a straight line through a curvilinear relationship.

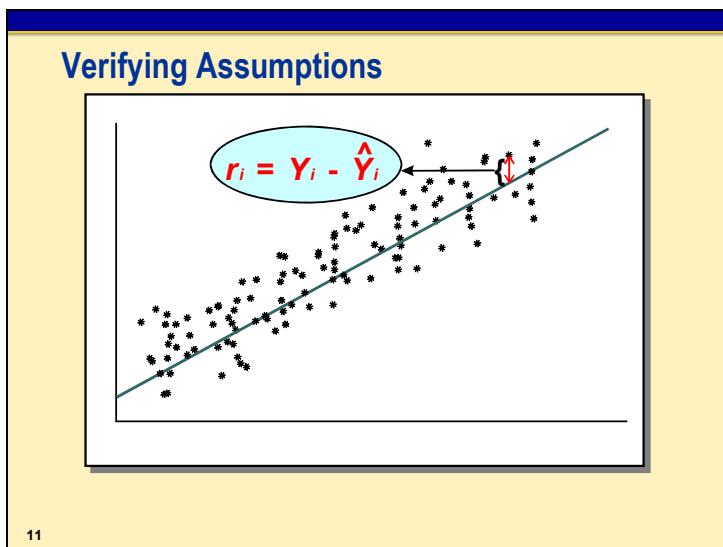


In the third plot, there seems to be an outlying data value that is affecting the regression line. This outlier is an influential data value in that it is substantially changing the fit of the regression line.



In the fourth plot, the outlying data point dramatically changes the fit of the regression line. In fact the slope would be undefined without the outlier.

The four plots illustrate that relying on the regression output to describe the relationship between your variables can be misleading. The regression equations and the  $R^2$  statistics are the same even though the relationships between the two variables are different. Always produce a scatter plot before you conduct a regression analysis.



11

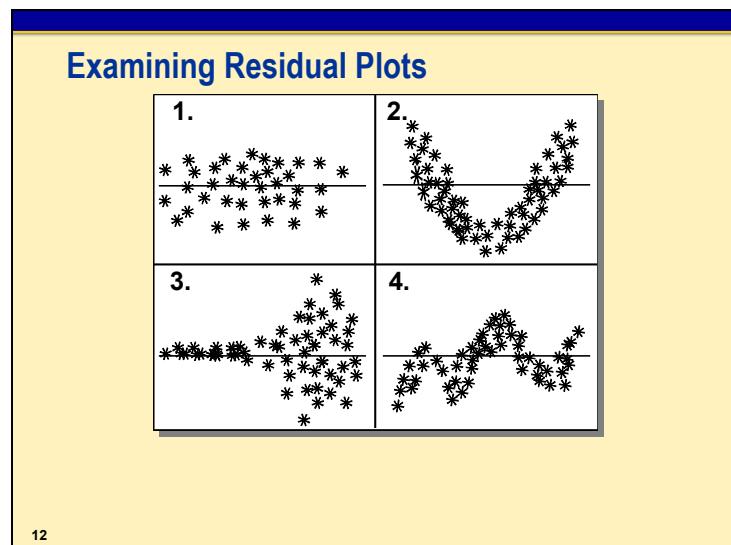
To verify the assumptions for regression, you can use the residual values from the regression analysis. Residuals are defined as

$$r_i = Y_i - \hat{Y}_i$$

where  $\hat{Y}_i$  is the predicted value for the  $i^{\text{th}}$  value of the dependent variable.

You can examine two types of plots when verifying assumptions:

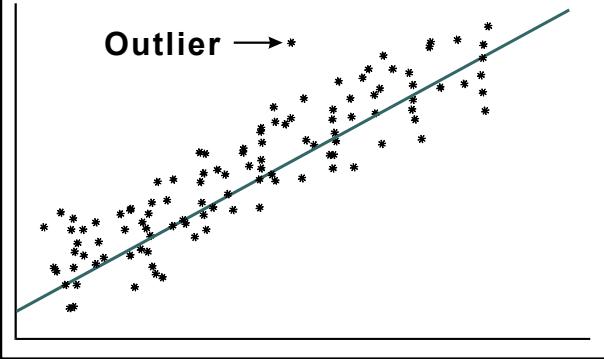
- the residuals versus the predicted values
- the residuals versus the values of the independent variables



The graphs above are plots of residual values versus predicted values or predictor variable values for four models fit to different sets of data. If model assumptions are valid, then the residual values should be randomly scattered about a reference line at 0. Any patterns or trends in the residuals might indicate problems in the model.

1. The model form appears to be adequate because the residuals are randomly scattered about a reference line at 0 and no patterns appear in the residual values.
2. The model form is incorrect. The plot indicates that the model should take into account curvature in the data. One possible solution is to add a quadratic term as one of the predictor variables.
3. The variance is not constant. As you move from left to right, the variance increases. One possible solution is to transform your dependent variable.
4. The observations are not independent. For this graph, the residuals tend to be followed by residuals with the same sign, which is called *autocorrelation*. This problem can occur when you have observations that have been collected over time. A possible solution is to use the AUTOREG procedure in SAS/ETS software.

### Detecting Outliers



A scatter plot showing data points as small asterisks. A solid blue line represents the regression fit. One data point is highlighted with a large asterisk and an arrow pointing to it, labeled 'Outlier'.

13

Besides verifying assumptions, it is also important to check for outliers. Observations that are outliers are far away from the bulk of your data. These observations are often data errors or reflect unusual circumstances. In either case, it is good statistical practice to detect these outliers and find out why they have occurred.

### Studentized Residual

Studentized residuals (SR) are obtained by dividing the residuals by their standard errors.

Suggested cutoffs are as follows:

- $|SR| > 2$  for data sets with a relatively small number of observations
- $|SR| > 3$  for data sets with a relatively large number of observations

14

One way to check for outliers is to use the studentized residuals. These are calculated by dividing the residual values by their standard errors. For a model that fits the data well and has no outliers, most of the studentized residuals should be close to 0. In general, studentized residuals that have an absolute value less than 2.0 could have easily occurred by chance. Studentized residuals that are between an absolute value of 2.0 to 3.0 occur infrequently and could be outliers. Studentized residuals that are larger than an absolute value of 3.0 occur rarely by chance alone and should be investigated



There is a difference between the labels used in SAS and in SAS Enterprise Guide.

SAS		SAS Enterprise Guide
Studentized residuals	⇒	Standardized residuals



## Residual and Scatter Plots

Example: Invoke the REG procedure and use a PLOT statement to produce high-resolution residual plots and diagnostic plots for the PREDICT model generated in the previous chapter.

```
/* c4demo01 */
options ps=50 ls=97;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc reg data=sasuser.b_fitness;
  PREDICT: model oxygen_consumption
            = runtime age run_pulse maximum_pulse;
  plot r.*(p. runtime age run_pulse maximum_pulse);
  plot student.*obs. / vref=3 2 -2 -3
            haxis=0 to 32 by 1;
  plot student.*nqq. ;
  symbol v=dot;
  title 'PREDICT Model - Plots of Diagnostic Statistics';
run;
quit;
```

Selected REG procedure statement:

PLOT produces plots of variables from the input data set and statistics from the analysis. The statistics you plot can be any that are available in the OUTPUT data set. To plot a statistic from the analysis, follow the keyword with a period to indicate that it is not a variable from the input data set.

Selected PLOT statement options:

VREF specifies where reference lines perpendicular to the vertical axis are to appear.

HAXIS specifies range and tick marks for the horizontal axis.

Selected keywords for the PLOT statement:

R. residuals

P. predicted values

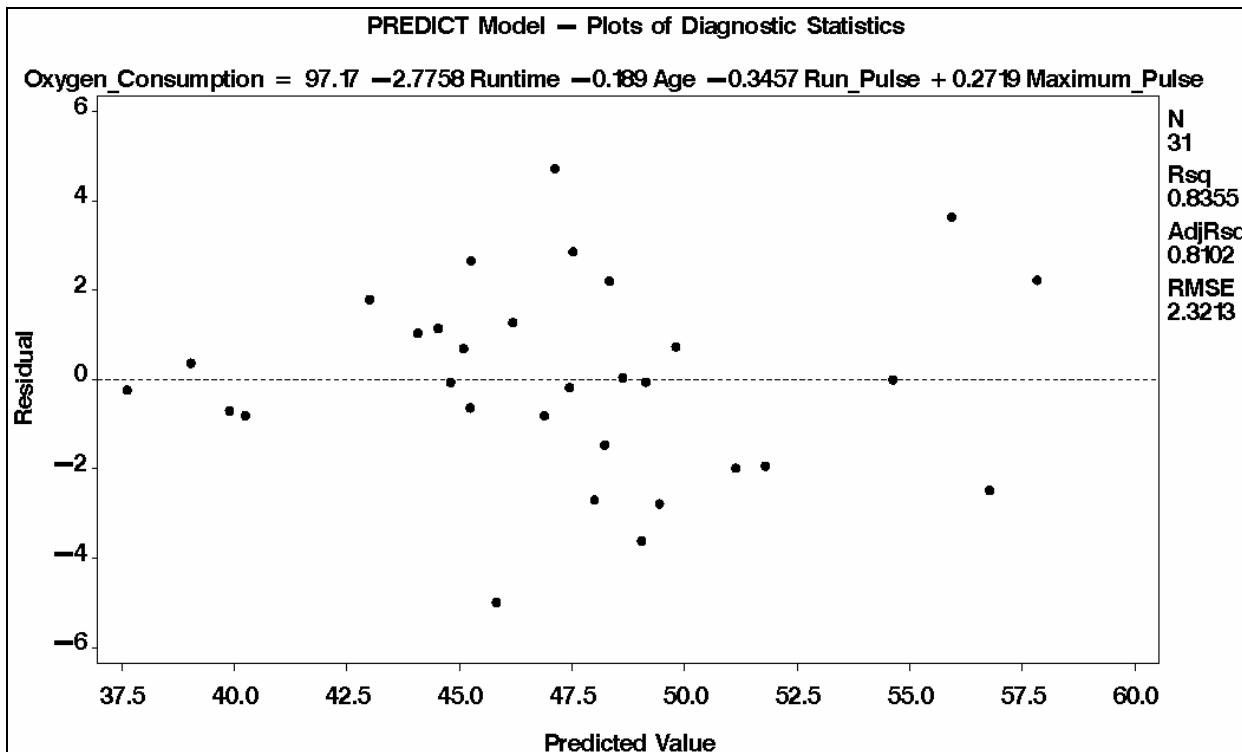
STUDENT. student residuals

NQQ. normal quantile values

OBS. observation number in the data set.

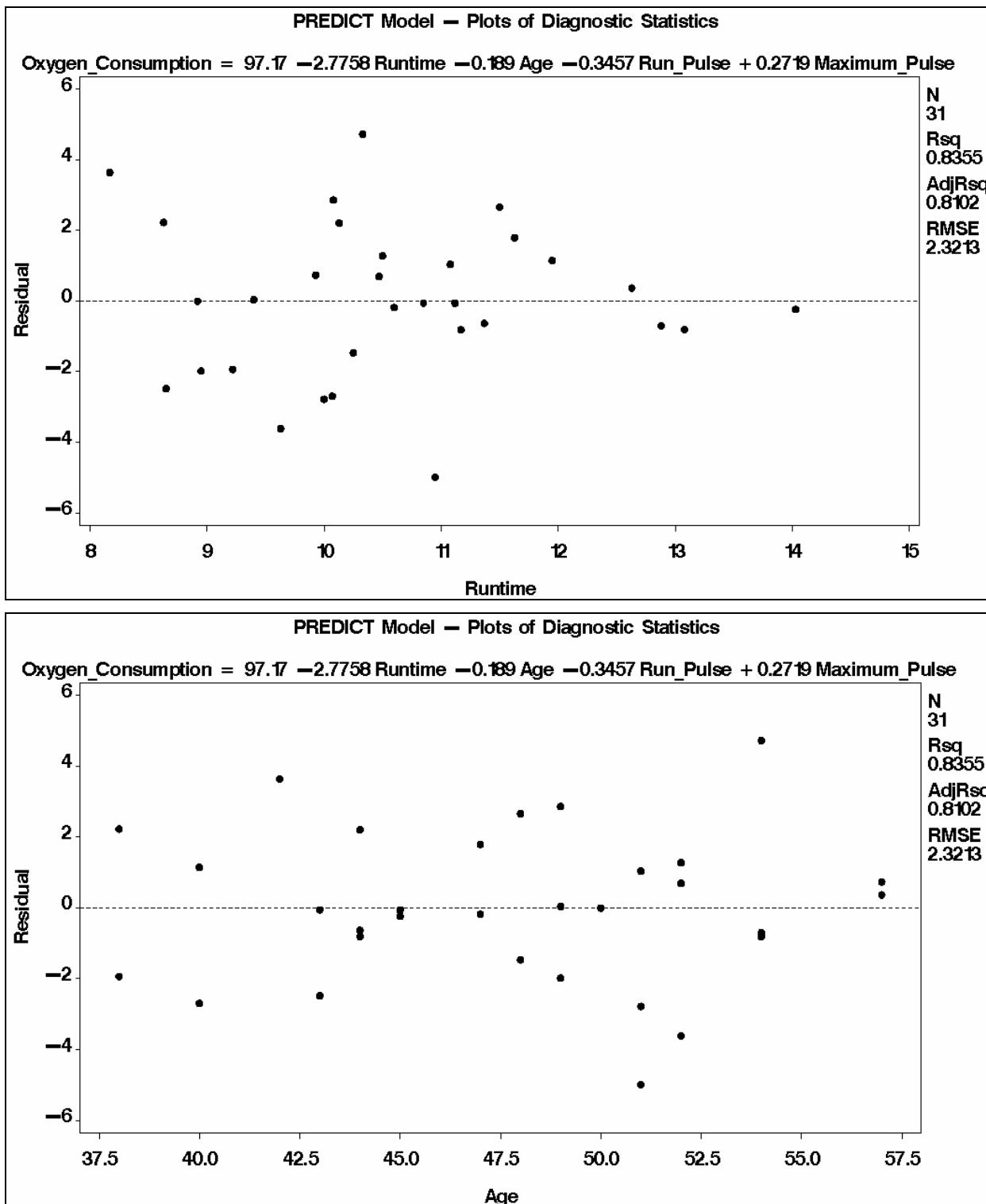
The normal quantile-quantile plot helps to indicate whether the residuals are normally distributed. The assumption of normality should be verified, but it is not as important as the other regression assumptions.

The plot of the residuals by predicted values of **Oxygen\_Consumption** is shown below. The residual values appear to be randomly scattered about the reference line at 0. There are no apparent trends or patterns in the residuals.

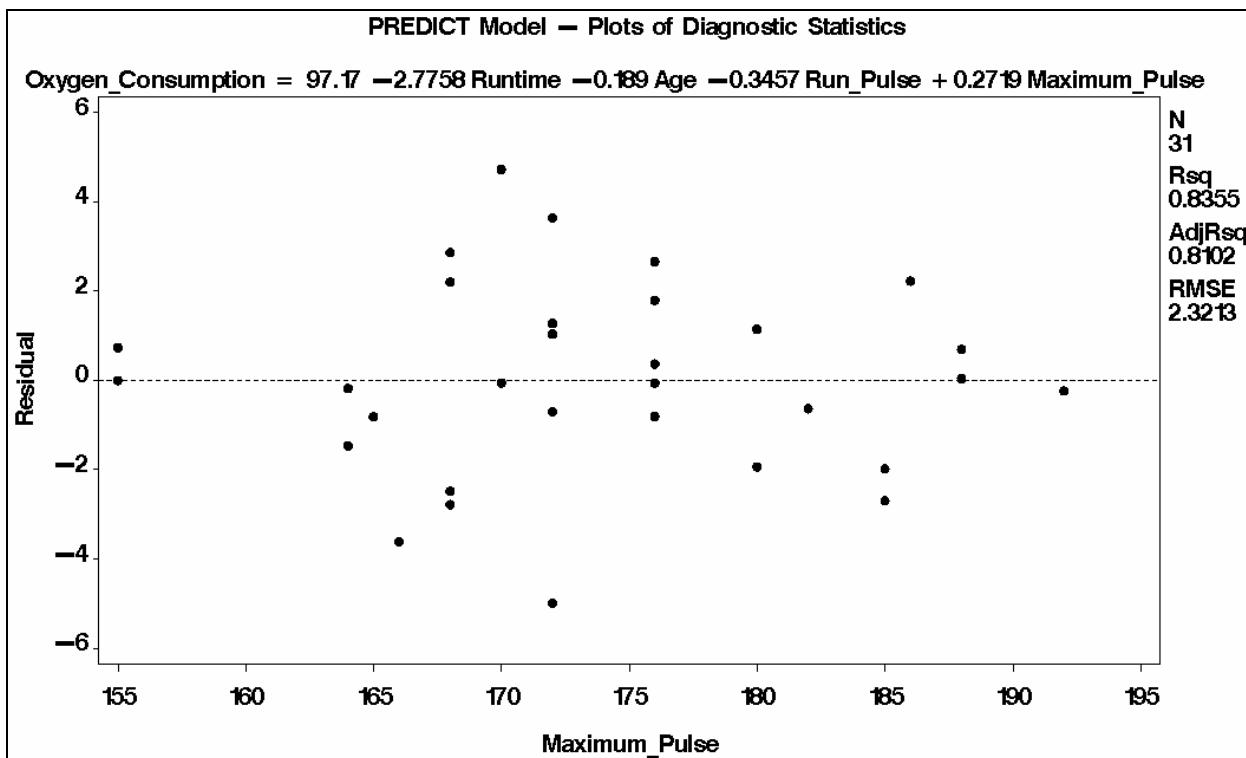
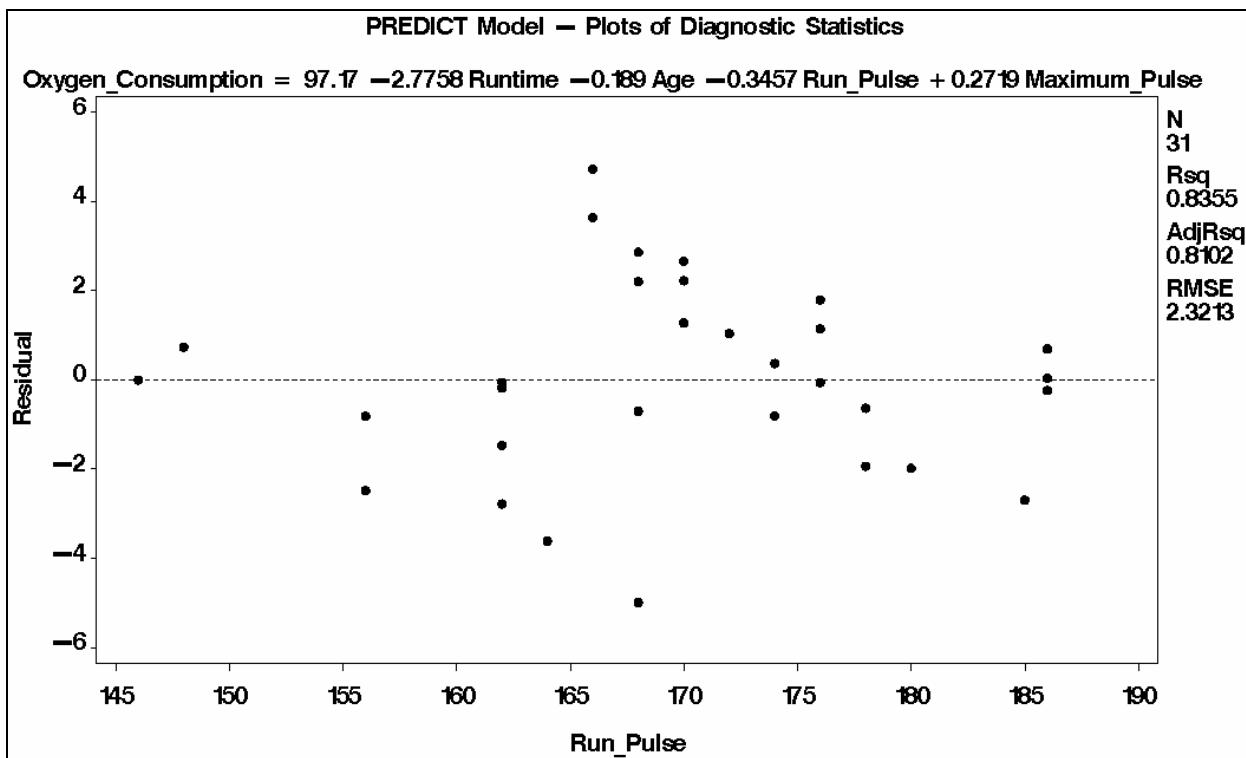


The statistics printed on the side are the same as those found in the PROC REG output.

The plot of the residuals versus the values of the independent variables, **Runtim**, **Age**, **Run\_Pulse**, and **Maximum\_Pulse** are shown below. There is also no apparent trend or pattern in the residuals.

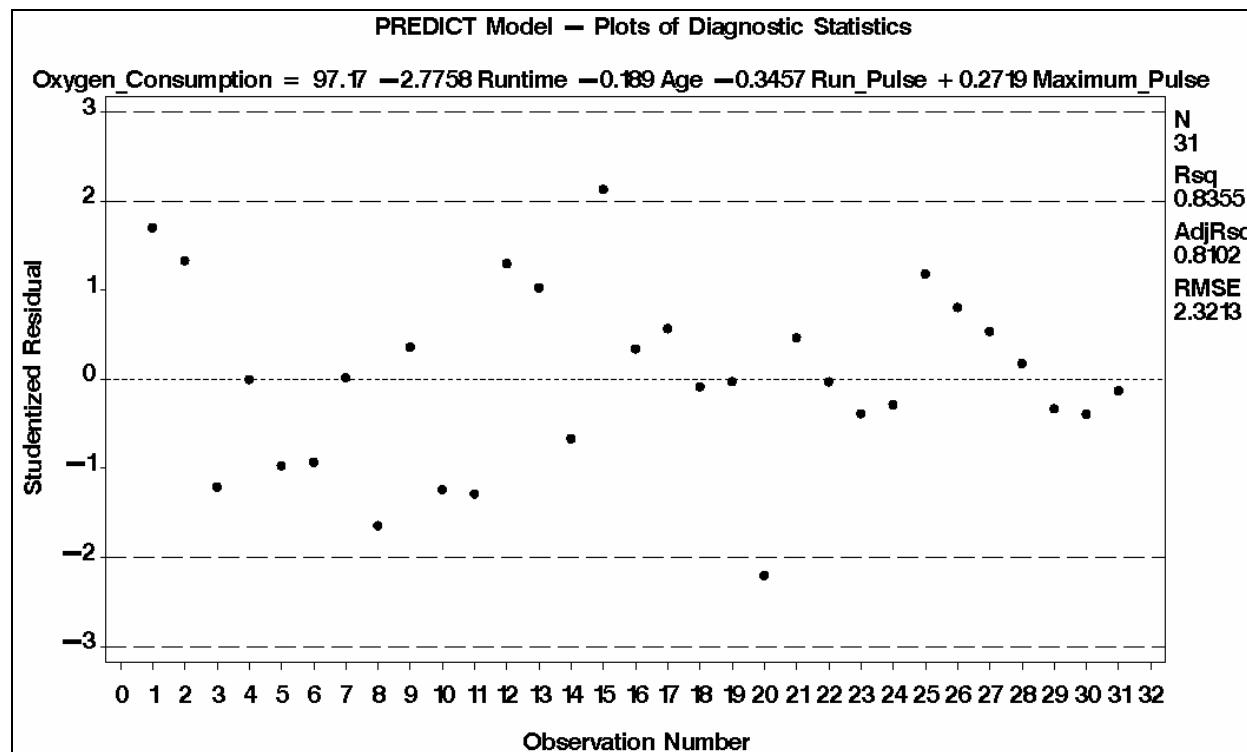


PROC REG Output (continued)



The plot of the student residuals by observation number is shown below. Reference lines are drawn on the student residual axis at 3.0, 2.0, -2.0, and -3.0. Two observations, number 15 (Sammy) and number 20 (Jack), are potential outliers.

PROC REG Output (continued)

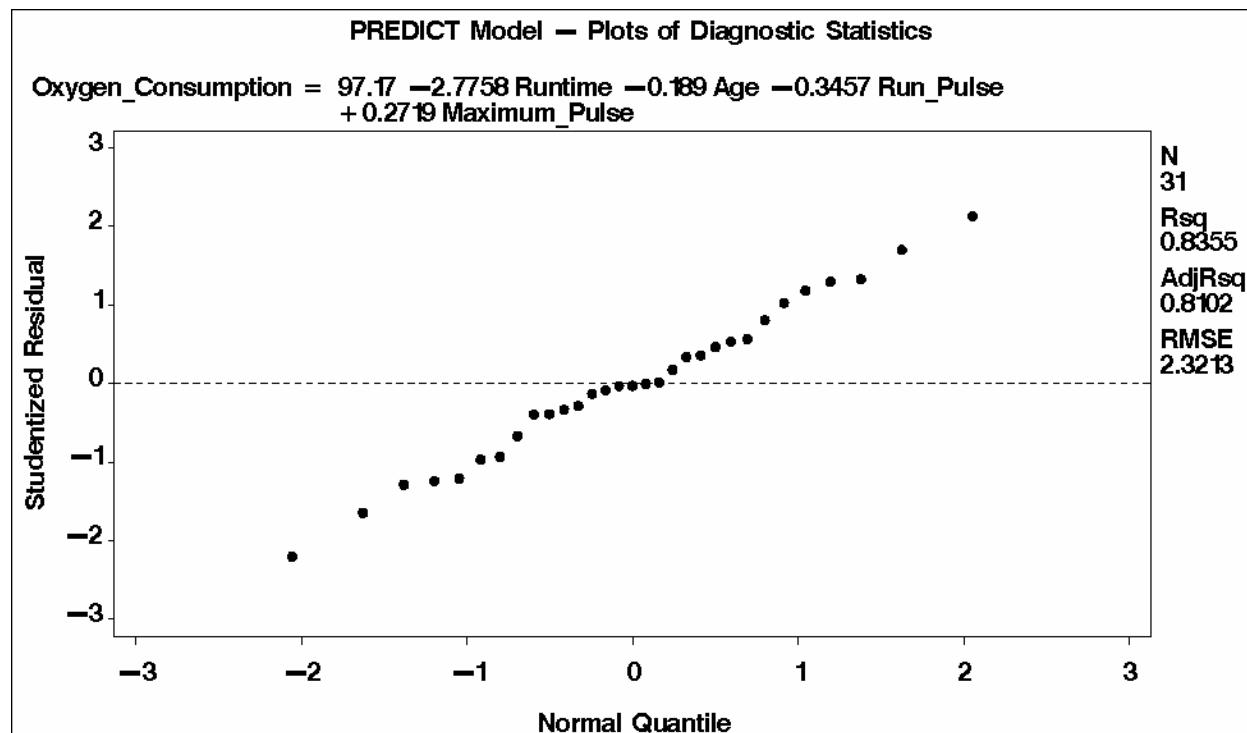


You can also use the R option in the MODEL statement of PROC REG to obtain residual diagnostics. Output from the R option includes the values of the response variable, the predicted values of the response variable, the standard error of the predicted values, the residuals, the standard error of the residuals, the student residuals, and a plot of the student residuals in tabular rather than graphic form. The R option is used in the next section.

The plot of the normal quantiles versus the student residuals is shown below. The plot is obtained by plotting the student residuals against their expected quantiles if the residuals come from a normal distribution. If the residuals are normally distributed, the plot should appear to be a straight line with a slope of about 1. If the plot deviates substantially from the ideal, then there is evidence against normality.

The plot below shows no deviation from the expected pattern. Thus, you can conclude that the residuals do not significantly violate the normality assumption. If the residuals did violate the normality assumption, then a transformation of the response variable might be warranted.

PROC REG Output (continued)



You can use the NORMAL option in the UNIVARIATE procedure to generate a hypothesis test on whether the residuals are normally distributed. This could be necessary if you feel the plot above shows a violation of the normality assumption. First you must create an output data set with the residuals in PROC REG using an OUTPUT statement (as shown in Chapter 2 with an OUTPUT statement in the GLM procedure) or the Output Delivery System. Then use that data set as the input data set in PROC UNIVARIATE.



**Refer to Exercise 1 for Chapter 4 in Appendix A.**

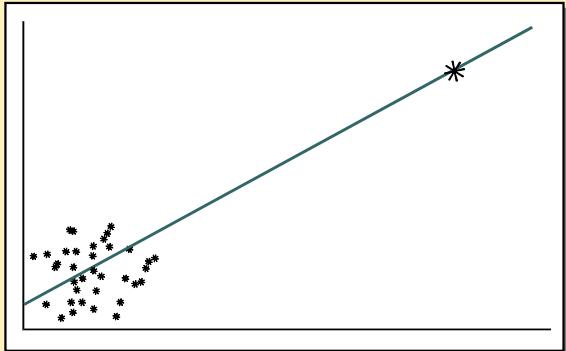
## 4.2 Influential Observations

### Objectives

- Use statistics to identify potential influential observations.

21

### Influential Observations



22

Recall in the previous section that you saw examples of data sets where the simple linear regression model fits were essentially the same. However, plotting the data revealed that the model fits were different.

One of the examples showed a highly influential observation like the example above.

Identifying influential observations in multiple linear regression is more complex because you have more predictors to consider.

The REG procedure has options to calculate statistics to identify influential observations.

## Diagnostic Statistics

Four statistics that help identify influential observations are

- STUDENT residual
- Cook's D
- RSTUDENT residual
- DFFITS.

23

The R option in the MODEL statement prints the first two statistics, as well as several others discussed previously. The INFLUENCE option in the MODEL statement prints the RSTUDENT and DFFITS statistics, as well as several others that are not discussed, such as the Hat Diagonal, Covariance Ratio, and DFBETAS.

## Cook's D Statistic

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis.

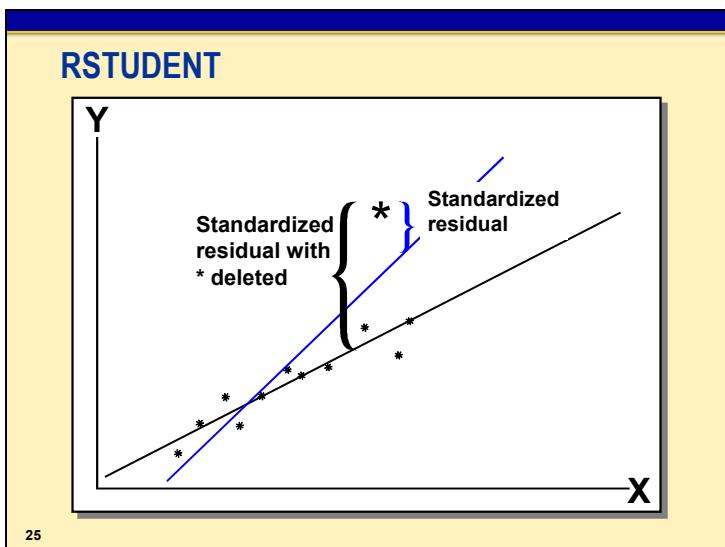
A suggested cutoff is  $D_i > \frac{4}{n}$ , where  $n$  is the sample size.

If the above condition is true, then the observation might have an adverse effect on the analysis.

24

To detect influential observations, you can use Cook's D statistic. This statistic measures the change in the parameter estimates that results from deleting each observation.

Identify observations above the cutoff and investigate the reasons they occurred.



Recall that STUDENT residuals are the ordinary residuals divided by their standard errors. The RSTUDENT residuals are similar to the STUDENT residuals except that they are calculated after deleting the  $i^{\text{th}}$  observation. In other words, the RSTUDENT is the difference between the observed Y and the predicted value of Y excluding this observation from the regression.

If the RSTUDENT is different from the STUDENT residual for a specific observation, that observation is likely to be influential

- There is a difference between the labels used in SAS and in SAS Enterprise Guide.

SAS		SAS Enterprise Guide
Studentized residuals	⇒	Standardized residuals
Rstudent residuals (studentized residual with the $i^{\text{th}}$ observation removed)	⇒	Studentized residuals

## DFFITS

DFFITS<sub>i</sub> measures the impact that the *i*<sup>th</sup> observation has on the predicted value.

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$$

$\hat{Y}_i$  is the *i*<sup>th</sup> predicted value.

$\hat{Y}_{(i)}$  is the *i*<sup>th</sup> predicted value when the *i*<sup>th</sup> observation is deleted.

$s(\hat{Y}_i)$  is the standard error of the *i*<sup>th</sup> predicted value.

26

Belsey, Kuh, and Welsch (1980) provide this suggested cutoff:  $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$ , where  $p$  is the number of terms in the current model, including the intercept, and  $n$  is the sample size.



## Looking for Influential Observations

Example: Generate the RSTUDENT and DFFITS influence statistics for the PREDICT variable model. Save the statistics to an output data set and create a data set with only observations that exceed the suggested cutoffs of the influence statistics.

```
/* c4demo02a */
goptions reset=all;
proc reg data=sasuser.b_fitness;
  PREDICT: model oxygen_consumption
    =runtime age run_pulse maximum_pulse
    / r influence;
  id name;
  output out=ck4outliers
    rstudent=rstud dffits=dfits cookd=cooks;
  title;
run;
quit;
```

Selected REG procedure statement:

OUTPUT creates a new SAS data set that saves the diagnostic statistics calculated after fitting the model.

Selected keywords for the OUTPUT statement:

COOKD= requests the Cook's D statistic.

DFFITS= requests the DFFITS statistic.

RSTUDENT= requests the RSTUDENT statistic.

Selected MODEL statement option:

INFLUENCE requests the diagnostics be printed.

## Partial PROC REG Output

The REG Procedure					
Model: PREDICT					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read					31
Number of Observations Used					31
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	711.45087	177.86272	33.01	<.0001
Error	26	140.10368	5.38860		
Corrected Total	30	851.55455			
Root MSE					
		2.32134	R-Square	0.8355	
Dependent Mean					
		47.37581	Adj R-Sq	0.8102	
Coeff Var					
		4.89984			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	97.16952	11.65703	8.34	<.0001
Runtime	1	-2.77576	0.34159	-8.13	<.0001
Age	1	-0.18903	0.09439	-2.00	0.0557
Run_Pulse	1	-0.34568	0.11820	-2.92	0.0071
Maximum_Pulse	1	0.27188	0.13438	2.02	0.0534

The ANOVA table and the Parameter Estimates table are identical to the previous example.

## Partial PROC REG Output (continued)

The REG Procedure							
Model: PREDICT							
Dependent Variable: Oxygen_Consumption							
Output Statistics							
Obs	Name	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	Donna	59.5700	55.9333	0.9104	3.6367	2.135	1.703
2	Gracie	60.0600	57.8362	1.6123	2.2238	1.670	1.332
3	Luanne	54.3000	56.7812	1.0775	-2.4812	2.056	-1.207
4	Mimi	54.6300	54.6309	1.0870	-0.000855	2.051	-0.0004
5	Chris	49.1600	51.1400	1.0944	-1.9800	2.047	-0.967

Output Statistics							
Obs	Name	Residual	Std Error Residual	Student Residual	-2	-1	0
1	Donna	3.6367	2.135	1.703		***	
2	Gracie	2.2238	1.670	1.332		**	
3	Luanne	-2.4812	2.056	-1.207		**	
4	Mimi	-0.000855	2.051	-0.0004			
5	Chris	-1.9800	2.047	-0.967		*	

Output Statistics							
Obs	Name	-2	-1	0	1	2	Cook's D
1	Donna		***		0.105	1.7718	0.1538
2	Gracie		**		0.331	1.3526	0.4824
3	Luanne		**		0.080	-1.2179	0.2155
4	Mimi				0.000	-0.000409	0.2193
5	Chris		*		0.053	-0.9659	0.2223

DFBETAS							
Obs	Name	Intercept	Runtime	Age	Run_Pulse	Pulse	Maximum_Pulse
1	Donna	0.3224	-0.4897	-0.2658	0.0429	-0.0645	
2	Gracie	-0.2501	-0.2278	-0.1814	-0.9617	1.0269	
3	Luanne	-0.2127	0.1280	0.1711	0.4084	-0.3017	
4	Mimi	-0.0001	0.0000	0.0000	0.0001	-0.0000	
5	Chris	0.3170	0.3586	-0.2798	0.0185	-0.1792	..

These statistics were requested by the INFLUENCE option.

#### Partial PROC REG Output (continued)

Sum of Residuals	0
Sum of Squared Residuals	140.10368
Predicted Residual SS (PRESS)	190.90531

The PRESS statistic is the sum of the PRESS residuals. These measure the deviation of the  $i^{\text{th}}$  observation about the regression line formed when that observation is deleted from the analysis. In other words, it measures how well the regression model predicts the  $i^{\text{th}}$  observation as though it were a new observation.

When the PRESS statistic is large compared to the Sum of the Squared Residuals, it indicates the presence of influential observations. The PRESS statistic is most useful when comparing several candidate models, such as comparing the PREDICT and EXPLAIN models that were examined earlier.

Use the following program to search for possible influential observations.

```
/* c4demo02b */
/* set the values of these macro variables, */
/* based on your data and model.          */
%let numparms=5; /* # of predictor variables + 1 */
%let numobs=31; /* # of observations */
%let idvars=name; /* relevant identification variable(s) */

data influential;
  set ck4outliers;

  cutdfits=2*(sqrt(&numparms/&numobs));
  cutcookd=4/&numobs;

  rstud_i=(abs(rstud)>3);
  dfits_i=(abs(dfits)>cutdfits);
  cookd_i=(cooks>cutcookd);
  sum_i=rstud_i + dfits_i + cookd_i;
  if sum_i > 0;
run;
```

An expression enclosed in parentheses is a logical operator that returns the value 1 if the expression is true and 0 if the expression is false.

The DATA step sets 0/1 indicator variables (**rstud\_i**, **dfits\_i**, and **cookd\_i**) for the diagnostic statistics using the suggested cutoffs. The **sum\_i** variable is the total number of diagnostic statistics that exceed the cutoffs for the observation. The last line subsets the file so that the data set **influential** includes only those observations that have at least one statistic that exceeds the cutoff. If the number of influential observations is large, you might not have the proper model.

```
/* c4demo02c */
proc print data=influential;
  var sum_i &idvars cooksd rstud dfits cutcookd cutdfits
    cookd_i rstud_i dfits_i;
  title 'Observations that Exceed Suggested Cutoffs';
run;
```

## PROC PRINT Output

Observations that Exceed Suggested Cutoffs										
				c	c					
				u	u	c	r	d		
				t	t	o	s	f		
		c		c	d	o	t	i		
s	u	N	o	s	f	o	f	k	u	t
o	m	a	k	t	i	o	i	d	d	s
b	—	m	s	u	t	k	t	—	—	—
s	i	e	d	d	s	d	s	i	i	i
1	2	Gracie	0.33051	1.35265	1.30587	0.12903	0.80322	1	0	1

## How to Handle Influential Observations

1. Recheck the data to ensure that no transcription or data entry errors have occurred.
2. If the data is valid, one possible explanation is that the model is not adequate.
  - A model with higher order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.

29

If the unusual data is erroneous, correct the errors and reanalyze the data.

In this course, time does not permit discussion of higher order models in any depth.

Another possibility is that the observation, although valid, could be unusual. If you had a larger sample size, there might be more observations like the unusual ones.

You might have to collect more data to confirm the relationship suggested by the influential observation.

In general, do not exclude data. In many circumstances, some of the unusual observations contain important information.

If you do choose to exclude some observations, include a description of the types of observations you exclude and provide an explanation. Also discuss the limitation of your conclusions, given the exclusions, as part of your report or presentation.



**Refer to Exercise 2 for Chapter 4 in Appendix A.**

## 4.3 Collinearity

### Objectives

- Determine if collinearity exists in a model.
- Generate output to evaluate the strength of the collinearity and what variables are involved in the collinearity.
- Determine methods to minimize collinearity in a model.

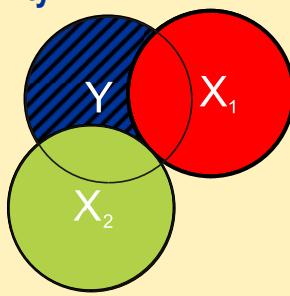
34

### A Model with No Collinearity

Model  $R^2=0.37$

$X_1$ :  $p$ -value < 0.0001

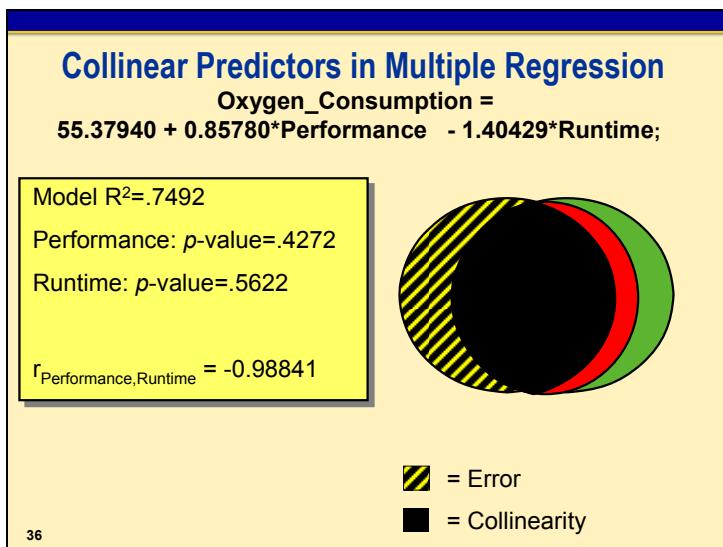
$X_2$ :  $p$ -value < 0.0001



= Error

= Collinearity

35



Recall that collinearity arises when the Xs contain redundant information; for example, **Performance** and **Runtime** are highly correlated with each other.

Collinearity can cause these problems in your model:

- truly significant terms can be hidden
- the variances of the coefficients are increased, which results in less precise estimates of the parameters and the predicted values

Collinearity is **not** a violation of the assumptions.



## Example of Collinearity

Example: Generate a regression with **Oxygen\_Consumption** as the dependent variable and **Performance, Runtime, Age, Weight, Run\_Pulse, Rest\_Pulse**, and **Maximum\_Pulse** as the independent variables. Compare this model with the PREDICT model from the previous section.

```
/* c4demo03 */
proc reg data=sasuser.b_fitness;
  FULLMODL;
  model oxygen_consumption
    = performance runtime age weight
      run_pulse rest_pulse maximum_pulse;
  title 'Collinearity -- Full Model';
run;
quit;
```

### PROC REG Output

Collinearity -- Full Model					
The REG Procedure					
Model: FULLMODL					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read		31			
Number of Observations Used		31			
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.03251	103.14750	18.32	<.0001
Error	23	129.52204	5.63139		
Corrected Total	30	851.55455			
Root MSE		2.37306	R-Square	0.8479	
Dependent Mean		47.37581	Adj R-Sq	0.8016	
Coeff Var		5.00900			

## PROC REG Output (continued)

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	93.33753	36.49782	2.56	0.0176
Performance	1	0.25756	1.02373	0.25	0.8036
Runtime	1	-2.08804	2.22856	-0.94	0.3585
Age	1	-0.21066	0.10519	-2.00	0.0571
Weight	1	-0.07741	0.05681	-1.36	0.1862
Run_Pulse	1	-0.36618	0.12299	-2.98	0.0067
Rest_Pulse	1	-0.01389	0.07114	-0.20	0.8469
Maximum_Pulse	1	0.30490	0.13990	2.18	0.0398

The Model  $F$  is highly significant and the  $R^2$  is large. These statistics suggest that the model fits the data well.

However, when you examine the  $p$ -values of the parameters, only **Run\_Pulse** and **Maximum\_Pulse** are statistically significant.

Recall that the PREDICT model included **Runtime**; however, in the full model, this same variable is not statistically significant ( $p$ -value=0.3585).

Including all the terms in the model hid at least one significant term.

When you have a significant Model  $F$  but no highly significant terms, collinearity is a likely problem.

## Collinearity Diagnostics

PROC REG offers these tools that help quantify the magnitude of the collinearity problems and identify the subset of Xs that is collinear:

- VIF
- COLLIN
- COLLINOINT

39

Selected MODEL statement options:

VIF	provides a measure of the magnitude of the collinearity (Variance Inflation Factor).
COLLIN	includes the intercept vector when analyzing the X'X matrix for collinearity.
COLLINOINT	excludes the intercept vector.

Two options, COLLIN and COLLINOINT, also provide a measure of the magnitude of the problem as well as give information that can be used to identify the sets of Xs that are the source of the problem.

## Variance Inflation Factor (VIF)

The VIF is a relative measure of the increase in the variance because of collinearity. It can be thought of as the ratio:

$$VIF_i = \frac{1}{1 - R_i^2}$$

A  $VIF_i > 10$  indicates that collinearity is a problem.

40

You can calculate a VIF for each term in the model.

Marquardt (1990) suggests that a  $VIF > 10$  indicates the presence of strong collinearity in the model.

$VIF_i = 1/(1 - R_i^2)$ , where  $R_i^2$  is the  $R^2$  of  $X_i$ , regressed on all the other Xs in the model.

For example, if the model is  $Y = X1 X2 X3 X4$ ,  $i = 1$  to 4.

To calculate the  $R^2$  for  $X3$ , fit the model  $X3 = X1 X2 X4$ . Take the  $R^2$  from the model with  $X3$  as the dependent variable and replace it in the formula  $VIF_3 = 1/(1 - R_3^2)$ . If  $VIF_3$  is greater than 10,  $X3$  is possibly involved in collinearity.

## COLLIN and COLLINPOINT Options

Both options generate condition indices and variance proportion statistics.

- The COLLIN option includes the intercept.
- The COLLINPOINT is adjusted for the intercept.

41

The COLLIN and COLLINPOINT options calculate these types of statistics:

- eigenvalues
- condition indices
- variance proportions.

*Eigenvalues* are also called characteristic roots. Eigenvalues near zero indicate strong collinearity. A value  $\lambda$  is called an eigenvalue if there exists a nonzero vector  $z$  such that  $(X'X)z = \lambda z$ . The *condition index*,  $\eta_i$ , is the square root of the largest eigenvalue divided by  $\lambda_i$ .

*Variance proportions* used in combination with the condition index can be used to identify the sets of Xs that are collinear. Variance proportions greater than 0.50 indicate which terms are correlated. Variance proportions are calculated for each term in the model.

The variance proportions for each term sum to 1.

## COLLIN Guidelines

Condition index values

- between 10 and 30 suggest weak dependencies
- between 30 and 100 indicate moderate dependencies
- greater than 100 indicate strong collinearity.

42

## Variance Proportions

Those predictors with variance proportions greater than 0.50 associated with a large condition index identify subsets of collinear predictors.

43

## COLLINOINT Guidelines

There are no published guidelines for the COLLINOINT option statistics.

However, using the COLLIN guidelines in conjunction with the COLLINOINT statistics enables you to evaluate the severity of the collinearity adjusting out the intercept.

44

## Using COLLIN and COLLINNOINT Statistics

Start on the the table  
that includes the intercept.

Is the Condition Index on  
the last row > 100?

No



Yes

Go to Variance Proportion  
on the intercept

45

Is the Variance Proportion  
on the intercept > 0.50?

Yes

Go to bottom row  
of the intercept  
adjusted table.

No

Stay on current row.

Find the variables with  
Variance Proportions > 0.50.

Rerun the REG  
procedure code and  
start again at the top of  
the flow chart.

Eliminate one of  
the variables.

46



## Collinearity Diagnostics

Example: Invoke PROC REG and use the VIF, COLLIN, and COLLINOINT options to assess the magnitude of the collinearity problem and identify the terms involved in the problem.

```
/* c4demo04 */
proc reg data=sasuser.b_fitness;
  FULLMODL:
  model oxygen_consumption
    = performance runtime age weight
      run_pulse rest_pulse maximum_pulse
    / vif collin collinoint;
  title 'Collinearity -- Full Model';
run;
quit;
```

## Partial PROC REG Output

Collinearity -- Full Model					
The REG Procedure					
Model: FULLMODL					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read				31	
Number of Observations Used				31	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.03251	103.14750	18.32	<.0001
Error	23	129.52204	5.63139		
Corrected Total	30	851.55455			
Root MSE		2.37306	R-Square	0.8479	
Dependent Mean		47.37581	Adj R-Sq	0.8016	
Coeff Var		5.00900			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	93.33753	36.49782	2.56	0.0176
Performance	1	0.25756	1.02373	0.25	0.8036
Runtime	1	-2.08804	2.22856	-0.94	0.3585
Age	1	-0.21066	0.10519	-2.00	0.0571
Weight	1	-0.07741	0.05681	-1.36	0.1862
Run_Pulse	1	-0.36618	0.12299	-2.98	0.0067
Rest_Pulse	1	-0.01389	0.07114	-0.20	0.8469
Maximum_Pulse	1	0.30490	0.13990	2.18	0.0398

Some of the VIFs are much larger than 10. A severe collinearity problem is present.

## Partial COLLIN Option Output

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	Performance	Runtime	Age
1	7.81224	1.00000	0.00000223	0.00003396	0.00000516	0.00011543
2	0.14978	7.22204	4.610439E-7	0.01283	0.00026016	0.00032355
3	0.01739	21.19723	0.00006157	0.00023609	0.00028745	0.24299
4	0.01246	25.03710	0.00000120	0.00120	0.00016004	0.05498
5	0.00606	35.90012	0.00027949	0.00007171	0.00149	0.09288
6	0.00179	66.03652	0.01276	0.03405	0.07620	0.38685
7	0.00018592	204.98810	0.00326	0.03584	0.02721	0.01651
8	0.00009415	288.05165	0.98363	0.91573	0.89439	0.20535
Proportion of Variation						
Number	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse		
				Run_Pulse	Rest_Pulse	Maximum_Pulse
1	0.00015063	0.00000679	0.00019829	0.00000501		
2	0.00018997	0.00001537	0.00374	0.00000627		
3	0.00908	0.00032301	0.24059	0.00022961		
4	0.39864	0.00016217	0.33791	0.00022890		
5	0.45536	0.01695	0.29325	0.00969		
6	0.10219	0.04272	0.01670	0.01335		
7	0.01929	0.92679	0.00001297	0.92625		
8	0.01510	0.01303	0.10759	0.05024		

Two condition indices are well above 100. For the largest, the variance proportions for the **Intercept**, **Performance**, and **Runtime** are greater than 0.50.

## COLLINOINT Option Output

Collinearity Diagnostics (intercept adjusted)						
Number	Eigenvalue	Condition Index	Proportion of Variation-----			
			Performance	Runtime	Age	Weight
1	2.92687	1.00000	0.00124	0.00133	0.00328	0.00953
2	1.87356	1.24988	0.00196	0.00194	0.10087	0.01834
3	0.94035	1.76424	0.00014220	0.00035679	0.00167	0.74750
4	0.74998	1.97550	0.00001910	0.00003187	0.20986	0.00001480
5	0.43947	2.58069	0.00329	0.00519	0.57367	0.16190
6	0.06022	6.97181	0.00019461	0.00012410	0.03802	0.02856
7	0.00955	17.50829	0.99315	0.99103	0.07263	0.03416

Collinearity Diagnostics (intercept adjusted)			
Number	-----Proportion of Variation-----		
	Run_Pulse	Rest_Pulse	Maximum_Pulse
1	0.00870	0.03205	0.00750
2	0.00620	0.00309	0.00967
3	0.00695	0.03473	0.00343
4	0.02020	0.43182	0.01612
5	0.00433	0.41363	0.00220
6	0.95340	0.00431	0.96071
7	0.00023243	0.08038	0.00036791

A similar pattern of collinearity appears when using the COLLINOINT option. Examining the last row of the above table reveals that **Performance** (0.99315) and **Runtime** (0.99103) possess variance proportions greater than 0.50. You can conclude that these two variables are involved in the collinearity.

Begin the process of eliminating collinear terms by returning to the Parameter Estimates table and recording the *p*-values of the identified subset of the independent variables:

**Performance**      *p*-value=0.8036

**Runtime**      *p*-value=0.3585

With this subset of variables, eliminate **Performance** from the model. Note that this variable also has a high VIF.

```
/* c4demo05 */
proc reg data=sasuser.b_fitness;
  NOPERF:
  model oxygen_consumption
    = runtime age weight
      run_pulse rest_pulse maximum_pulse
    / vif collin collinoint;
  title 'Collinearity -- Performance Removed';
run;
quit;
```

## Partial PROC REG Output

Collinearity -- Performance Removed					
The REG Procedure					
Model: NOPERF					
Dependent Variable: Oxygen_Consumption					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	721.67605	120.27934	22.23	<.0001
Error	24	129.87851	5.41160		
Corrected Total	30	851.55455			
Root MSE					
Dependent Mean					
Coeff Var					
Analysis of Variance					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	101.96313	12.27174	8.31	<.0001
Runtime	1	-2.63994	0.38532	-6.85	<.0001
Age	1	-0.21848	0.09850	-2.22	0.0363
Weight	1	-0.07503	0.05492	-1.37	0.1845
Run_Pulse	1	-0.36721	0.12050	-3.05	0.0055
Rest_Pulse	1	-0.01952	0.06619	-0.29	0.7706
Maximum_Pulse	1	0.30457	0.13714	2.22	0.0360
Variance Inflation					

**Run\_Pulse** and **Maximum\_Pulse** are significant in this model, as they were in the previous model, but now both **Runtime** and **Age** are significant in this model.

Note that the VIFs are now all less than 10.

Partial PROC REG Output (continued)

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	Runtime	Age	Weight
1	6.94983	1.00000	0.00002395	0.00021174	0.00015997	0.00019576
2	0.01856	19.35297	0.00224	0.02439	0.15550	0.00878
3	0.01521	21.37532	0.00069190	0.12332	0.15174	0.23637
4	0.00914	27.57505	0.00635	0.61945	0.03075	0.17375
5	0.00603	33.94799	0.00139	0.12581	0.11951	0.45090
6	0.00105	81.17086	0.79602	0.09233	0.47800	0.10834
7	0.00017900	197.04044	0.19329	0.01449	0.06435	0.02167

Collinearity Diagnostics			
Number	Proportion of Variation		
	Run_Pulse	Rest_Pulse	Maximum_Pulse
1	0.00000860	0.00027961	0.00000633
2	0.00000185	0.39351	0.00000723
3	0.00113	0.03259	0.00121
4	0.00152	0.19195	0.00125
5	0.01510	0.35859	0.00840
6	0.06682	0.01756	0.00556
7	0.91542	0.00552	0.98356

The largest condition index is still greater than 100, indicating that there is still collinearity in this model. For the largest condition index, the variance proportions for **Run\_Pulse** (0.91542) and **Maximum\_Pulse** (0.98356) are greater than 0.5. Note that the intercept is not involved in collinearity, so there is no need to examine the COLLINOINT output.

Because the variable **Maximum\_Pulse** (0.0360) has a higher *p*-value than **Run\_Pulse** (0.0055), generate another model and eliminate the variable **Maximum\_Pulse** from the MODEL statement.

```
/* c4demo06 */
proc reg data=sasuser.b_fitness;
  NOPRFMAX:
  model oxygen_consumption
    = runtime age weight
      run_pulse rest_pulse
    / vif collin collinoint;
  title 'Collinearity -- Performance and Maximum Pulse Removed';
run;
quit;
```

## PROC REG Output

Collinearity -- Performance and Maximum Pulse Removed					
The REG Procedure					
Model: NOPRFMAX					
Dependent Variable: Oxygen_Consumption					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	694.98323	138.99665	22.19	<.0001
Error	25	156.57132	6.26285		
Corrected Total	30	851.55455			
Analysis of Variance					
Root MSE                  R-Square          0.8161					
Dependent Mean          Adj R-Sq          0.7794					
Coeff Var                5.28238					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t  Variance Inflation
Intercept	1	115.46115	11.46893	10.07	<.0001 0
Runtime	1	-2.71594	0.41288	-6.58	<.0001 1.57183
Age	1	-0.27650	0.10217	-2.71	0.0121 1.38477
Weight	1	-0.05300	0.05811	-0.91	0.3704 1.12190
Run_Pulse	1	-0.12213	0.05207	-2.35	0.0272 1.36493
Rest_Pulse	1	-0.02485	0.07116	-0.35	0.7298 1.40819

The variables **Weight** and **Rest\_Pulse** are not statistically significant, indicating that they might be removed from the model. All VIFs are relatively small.

## PROC REG Output (continued)

Collinearity Diagnostics		
Number	Eigenvalue	Condition Index
1	5.95261	1.00000
2	0.01855	17.91390
3	0.01434	20.37297
4	0.00882	25.97155
5	0.00465	35.78017
6	0.00102	76.21454

Collinearity Diagnostics						
Number	Proportion of Variation					
	Intercept	Runtime	Age	Weight	Run_Pulse	Rest_Pulse
1	0.00004324	0.00029113	0.00023471	0.00027579	0.00007258	0.00038178
2	0.00296	0.02190	0.17447	0.00826	0.00002193	0.38754
3	0.00139	0.09587	0.14694	0.36846	0.00674	0.02990
4	0.01086	0.75407	0.04148	0.06095	0.00710	0.27246
5	0.02723	0.02828	0.18069	0.46144	0.26977	0.29881
6	0.95752	0.09958	0.45619	0.10061	0.71629	0.01090

The largest condition index is now approximately 76. This indicates that there are some moderate dependencies between the predictor variables in this model. Examination of the variance proportions indicates that **Intercept** and **Run\_Pulse** are involved in collinearity.

## PROC REG Output (continued)

Collinearity Diagnostics (intercept adjusted)		
Number	Eigenvalue	Condition Index
1	1.86111	1.00000
2	1.28404	1.20392
3	0.89216	1.44433
4	0.59808	1.76404
5	0.36462	2.25927

Collinearity Diagnostics (intercept adjusted)					
Number	Proportion of Variation				
	Runtime	Age	Weight	Run_Pulse	Rest_Pulse
1	0.07701	0.02981	0.04373	0.12341	0.11184
2	0.14039	0.27964	0.09841	0.01037	0.03290
3	0.04970	0.07934	0.68711	0.03614	0.08003
4	0.00449	0.05979	0.03567	0.66283	0.45266
5	0.72841	0.55142	0.13508	0.16726	0.32257

Using the COLLINPOINT output, **Runtime** (variance proportion=0.72841) and **Age** (variance proportion=0.55142) are involved in collinearity.

Now return to the Parameter Estimates table and record the *p*-values of **Runtime** (<0.0001) and **Age** (0.0121).

Options include the following:

- Accept the current model without deleting any more variables because **Runtime** and **Age** are both statistically significant. Furthermore, remember that the COLLIN Condition Index is approximately 76 for this model and that falls into the moderate range of collinearity.
- As noted earlier, the variables **Weight** (*p*-value=0.3704) and **Rest\_Pulse** (*p*-value=0.7298) are not statistically significant; you might want to eliminate **Rest\_Pulse** from the model and re-execute the reduced model.

### Guidelines for Eliminating Terms

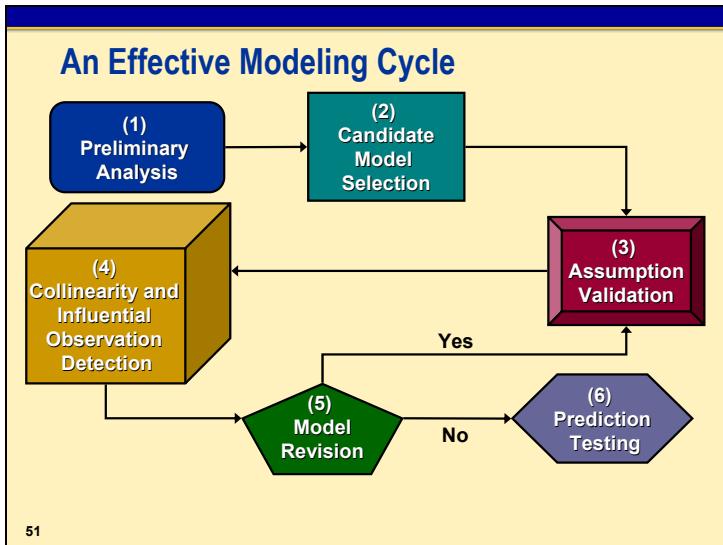
1. Determine the set of Xs involved in collinearity using the variance proportions associated with the largest condition index (if it is greater than 100).
2. Drop the variable among the set with the largest  $p$ -value that also has a large VIF.
3. Rerun the regression and repeat, if necessary.

49

In the previous demonstration you saw how to identify the sets of Xs that were collinear.

The natural question is, “Which terms should be dropped?” Subject matter expertise should be used as well as the suggested guidelines above.

There are other approaches to dealing with collinearity. Two techniques are ridge regression and principle components regression. In addition, recentering the predictor variables can sometimes eliminate collinearity problems, especially in a polynomial regression.



- (1) **Preliminary Analysis** This step includes the use of descriptive statistics, graphs, and correlation analysis.
- (2) **Candidate Model Selection** This step uses the numerous selection options in PROC REG to identify one or more candidate models.
- (3) **Assumption Validation** This step includes the plots of residuals and graphs of the residuals versus the predicted values. It also includes a test for equal variances.
- (4) **Collinearity and Influential Observation Detection.** The former includes the use of the VIF statistic, condition indices, and variation proportions; the latter includes the examination of Rstudent residuals, Cook's D statistic, and DFFITS statistics.
- (5) **Model Revision.** If steps (3) and (4) indicate the need for model revision, generate a new model by returning to these two steps.
- (6) **Prediction Testing.** If possible, validate the model with data not used to build the model.



Refer to Exercise 3 for Chapter 4 in Appendix A.

## 4.4 Chapter Summary

The four assumptions of linear regression analysis are

- the mean of the response variable is linearly related to the value of the predictor variable(s)
- the observations are independent
- the error terms for each value of the predictor variable are normally distributed
- the error variances for each value of the predictor variable are equal.

If these assumptions are not valid, the probability of drawing incorrect conclusions from the analysis might be increased.

It is important to be aware of influential observations in any regression model even though their existence does not violate the regression assumptions. For multiple regression, scatter plots do not necessarily identify influential observations. However, some statistics that can help identify influential observations are studentized residuals, RSTUDENT residuals, Cook's D, and DFFITS.

If more than one percent of the observations are identified as influential observations, it is possible that you do not have an adequate model; you may want to add higher-level terms, such as polynomial and interaction terms. In general, do not exclude data.

Collinearity is a problem unique to multiple regression. It can hide significant variables and increase the variance of the parameter estimates, resulting in an unstable model. Statistics useful in identifying collinearity are the variance inflation factor (VIF) and condition indices combined with variance proportions. From a statistical perspective, after you have identified the subset of independent variables that are collinear, one solution is to remove variable(s), only one at a time, from the model to eliminate the collinearity.

```
PROC REG DATA=SAS-data-set <options>;
  MODEL response=predictor </ options>;
  ID variable;
  PLOT y-variable*x-variable </ options>;
  OUTPUT OUT= SAS-data-set keyword=names;
RUN;
QUIT;
```

# Chapter 5 Categorical Data Analysis

<b>5.1 Describing Categorical Data.....</b>	<b>5-2</b>
<b>5.2 Tests of Association.....</b>	<b>5-17</b>
<b>5.3 Introduction to Logistic Regression.....</b>	<b>5-36</b>
<b>5.4 Multiple Logistic Regression.....</b>	<b>5-58</b>
<b>5.5 Logit Plots (Self-Study).....</b>	<b>5-78</b>
<b>5.6 Chapter Summary.....</b>	<b>5-84</b>

## 5.1 Describing Categorical Data

### Objectives

- Recognize the differences between categorical data and continuous data.
- Identify a variable's scale of measurement.
- Examine the distribution of categorical variables.
- Do preliminary examinations of associations between variables.

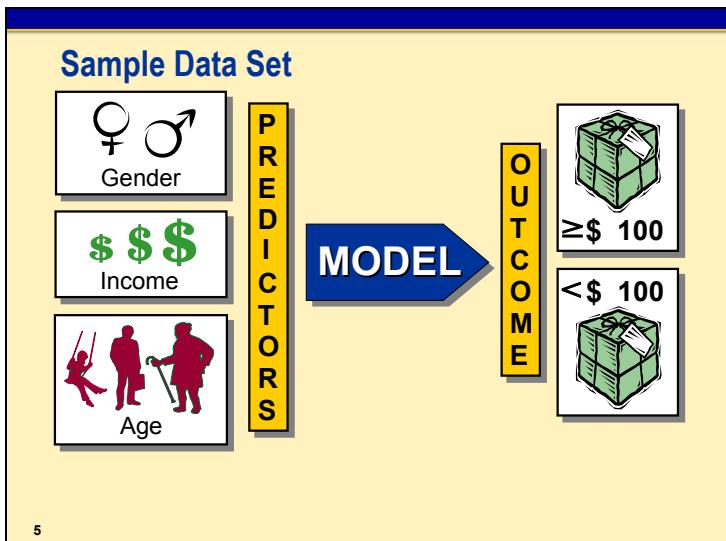
3

### Overview

Type of Response	Type of Predictors		
	Categorical	Continuous	Categorical and Continuous
Continuous	Analysis of Variance	Linear Regression	Analysis of Covariance (Regression with dummy variables)
Categorical	Logistic Regression or Contingency Tables	Logistic Regression	Logistic Regression

4

*Categorical data analysis* is concerned with categorical responses, regardless of whether the predictor variables are categorical or continuous. Categorical responses have a measurement scale consisting of a set of categories. *Continuous data analysis* is concerned with the analysis of continuous responses, regardless of whether the predictor variables are categorical or continuous.



Example: A company that sells its products via a catalog wants to identify those customers to whom advertising efforts should be directed. It has been decided that customers who spend 100 dollars or more are the target group. Based on the orders received over the last six months, the company wants to characterize this group of customers. The data is stored in the **sasuser.b\_sales** data set.

The variables in the data set are

**purchase** purchase price (1=100 dollars or more, 0=Under 100 dollars)  
**age** age of customers in years  
**gender** gender of customer (Male, Female)  
**income** annual income (Low, Middle, High).

This is a hypothetical data set.

## Identifying the Scale of Measurement

**Variable**

Agree  
No Opinion  
Disagree

Before analyzing, identify the measurement scale for each variable.

6

There are a variety of statistical methods for analyzing categorical data. To choose the appropriate method, you must determine the scale of measurement for your response variable.

## Nominal Variables

Variable:  
Kind of Beverage

Order any way you please!

7

*Nominal variables* have values with no logical ordering. In the **sasuser.b\_sales** data set, **gender** is a nominal variable.

## Ordinal Variables

Variable: Size of Beverage



8

*Ordinal variables* have values with a logical order. However, the relative distances between the values are not clear. In the `sasuser.b_sales` data set, `income` is an ordinal variable. Binary variables can also be considered ordinal variables.

After you choose the appropriate scale of measurement, you can describe the relationship between categorical variables with the use of mosaic plots and frequency tables.

## Examining Categorical Variables

By examining the distribution of categorical variables, you can

- screen for unusual data values
- determine the frequency of data values
- recognize possible associations among variables.

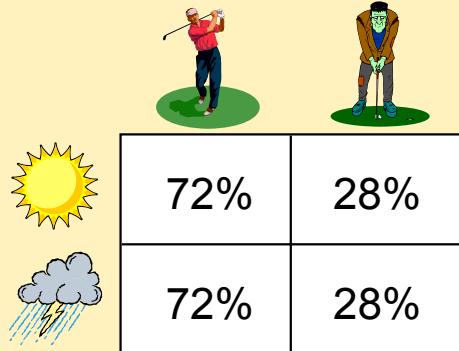
9

## Association

- An association exists between two variables if the distribution of one variable changes when the level (or value) of the other variable changes.
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

10

## No Association



Is your manager's mood associated  
with the weather?

11

There appears to be no association here because the **same** in each column.

## Association

82%	18%
60%	40%

Is your manager's mood associated with the weather?

There appears to be an association here because the row percentages are **different** in each column.

# Frequency Tables

A frequency table shows the number of observations that fall in certain categories or intervals. A one-way frequency table examines one variable.

Typically, there are four types of frequency measures included in a frequency table:

**frequency** is the number of times the value appears in the data set.

percent is 100 times the relative frequency. This represents the percentage of the data that has this value.

cumulative frequency accumulates the frequency of each of the values by adding the second frequency to the first and so on.

cumulative percent accumulates the percentage by adding the second percentage to the first and so on.

## Crosstabulation Tables

A *crosstabulation* table shows the number of observations for each combination of the row and column variables.

	column 1	column 2	...	column c
row 1	cell <sub>11</sub>	cell <sub>12</sub>	...	cell <sub>1c</sub>
row 2	cell <sub>21</sub>	cell <sub>22</sub>	...	cell <sub>2c</sub>
...	...	...	...	...
row r	cell <sub>r1</sub>	cell <sub>r2</sub>	...	cell <sub>rc</sub>

14

By default, a crosstabulation table has four measures in each cell:

- |           |   |
|-----------|---|
| frequency | number of observations falling into a category formed by the row variable value and the column variable value |
| percent   | number of observations in each cell as a percentage of the total number of observations                       |
| row pct   | number of observations in each cell as a percentage of the total number of observations in that row           |
| col pct   | number of observations in each cell as a percentage of the total number of observations in that column        |

## The FREQ Procedure

General form of the FREQ procedure:

```
PROC FREQ DATA=SAS-data-set;
  TABLES table-requests </ options>;
RUN;
```

15

Selected FREQ procedure statement:

**TABLES** requests tables and specifies options for producing tests. The general form of a table request is *variable1\*variable2\*...*, where any number of these requests can be made in a single TABLES statement. For two-way crosstabulation tables, the first variable represents the rows and the second variable represents the columns.

 PROC FREQ can generate large volumes of output as the number of variables or the number of variable levels (or both) increases.



## Examining Distributions

Example: Invoke PROC FREQ and create one-way frequency tables for the variables **gender**, **age**, **income**, and **purchase** and create two-way frequency tables for the variables **purchase** and **gender**, and **purchase** and **income**. Also use the FORMAT procedure to format the values of **purchase**.

```
/* c5demo01 */
proc format;
  value purfmt 1="$100 +"
            0("< $100"
            ;
run;

proc freq data=sasuser.b_sales;
  tables purchase gender income age
  gender*purchase income*purchase;
  format purchase purfmt.;
run;
```

### PROC FREQ Output

The FREQ Procedure					
	Frequency	Percent	Cumulative		Cumulative Percent
			Frequency	Percent	
< \$100	269	62.41	269	62.41	
\$100 +	162	37.59	431	100.00	
gender	Frequency	Percent	Cumulative		Cumulative Percent
			Frequency	Percent	
Female	240	55.68	240	55.68	
Male	191	44.32	431	100.00	
income	Frequency	Percent	Cumulative		Cumulative Percent
			Frequency	Percent	
High	155	35.96	155	35.96	
Low	132	30.63	287	66.59	
Medium	144	33.41	431	100.00	

PROC FREQ is an excellent tool for determining any miscoding in your data. There seem to be no unusual data values that could be due to coding errors for any of the categorical variables.

## PROC FREQ Output (continued)

age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
23	1	0.23	1	0.23
24	1	0.23	2	0.46
25	2	0.46	4	0.93
26	5	1.16	9	2.09
28	3	0.70	12	2.78
29	6	1.39	18	4.18
30	6	1.39	24	5.57
31	11	2.55	35	8.12
32	11	2.55	46	10.67
33	25	5.80	71	16.47
34	23	5.34	94	21.81
35	28	6.50	122	28.31
36	19	4.41	141	32.71
37	29	6.73	170	39.44
38	37	8.58	207	48.03
39	30	6.96	237	54.99
40	31	7.19	268	62.18
41	35	8.12	303	70.30
42	19	4.41	322	74.71
43	18	4.18	340	78.89
44	19	4.41	359	83.29
45	17	3.94	376	87.24
46	12	2.78	388	90.02
47	13	3.02	401	93.04
48	8	1.86	409	94.90
49	7	1.62	416	96.52
50	5	1.16	421	97.68
51	4	0.93	425	98.61
52	2	0.46	427	99.07
55	2	0.46	429	99.54
56	1	0.23	430	99.77
58	1	0.23	431	100.00



If a continuous variable does not have a lot of values, as is the case with **age** in this data, then it is acceptable to use PROC FREQ. However, if **age** had numerous values, it would be better to use the UNIVARIATE procedure to explore this variable.

The requested two-way frequency tables are shown below. You can get a preliminary idea whether there are associations between the outcome variable, **purchase**, and the predictor variables, **gender** and **income**, by examining the distribution of **purchase** for each value of the predictors.

#### PROC FREQ Output (continued)

Table of gender by purchase				
gender	purchase			
	Frequency	Percent	Row Pct	Col Pct
Female	139	32.25	57.92	51.67
	101	23.43	42.08	62.35
Male	130	30.16	68.06	48.33
	61	14.15	31.94	37.65
Total	269	62.41	162	37.59
				431
				100.00

By examining the row percentages, you see that **purchase** is associated with **gender**.

## PROC FREQ Output (continued)

Table of income by purchase				
	income	purchase		
	Frequency			
	Percent			
	Row Pct			
	Col Pct	< \$100	\$100 +	Total
High	81	74		155
	18.79	17.17		35.96
	52.26	47.74		
	30.11	45.68		
Low	90	42		132
	20.88	9.74		30.63
	68.18	31.82		
	33.46	25.93		
Medium	98	46		144
	22.74	10.67		33.41
	68.06	31.94		
	36.43	28.40		
Total		269	162	431
		62.41	37.59	100.00

## Ordering Values

When you have an ordinal variable such as `income`, it is important to put the values in logical order for analysis purposes.

Present Order	Logical Order
High	Low
Low	Medium
Medium	High

19

Treating an ordinal variable as nominal can reduce the power of your statistical tests. In other words, statistical tests that detect linear associations have more power than statistical tests that detect general associations.



## Ordering Values in the Frequency Table

Example: Obtain a logical order in a frequency table for the values in the variable **income**.

1. Create a new variable named **inclevel** so that the sort order corresponds to its logical order.

```
/* c5demo02 */
data sasuser.b_sales_inc;
  set sasuser.b_sales;
  inclevel=1*(income='Low') + 2*(income='Medium')
            + 3*(income='High');
run;
```

2. Use PROC FORMAT to create user-defined formats.

```
proc format;
  value incfmt 1='Low Income'
            2='Medium Income'
            3='High Income';
run;
```

3. Use PROC FREQ with a FORMAT statement.

```
proc freq data=sasuser.b_sales_inc;
  tables inclevel*purchase;
  format inclevel incfmt. purchase purfmt.;
  title1 'Create variable INCLEVEL to correct INCOME';
run;
```



If your data is in a logical order in a data set, you can use the ORDER=DATA option in PROC FREQ.

The crosstabulation of **inclevel\*purchase** is shown below. The values of **inclevel** are now in a logical order.

Create variable INCLEVEL to correct INCOME				
The FREQ Procedure				
		Table of inclevel by purchase		
inclevel		purchase		
Frequency				Total
Low Income		< \$100	\$100 +	
		90	42	132
		20.88	9.74	30.63
		68.18	31.82	
		33.46	25.93	
Medium Income		98	46	144
		22.74	10.67	33.41
		68.06	31.94	
		36.43	28.40	
High Income		81	74	155
		18.79	17.17	35.96
		52.26	47.74	
		30.11	45.68	
Total		269	162	431
		62.41	37.59	100.00

By examining the row percentages, you see that **purchase** is associated with **income**. For example, 48% of the high-income customers made purchases of 100 dollars or more compared to 32% of the low-income customers and 32% of the medium-income customers.

## 5.2 Tests of Association

### Objectives

- Perform a chi-square test for association.
- Examine the strength of the association.
- Produce exact  $p$ -values for the chi-square test for association.
- Perform a Mantel-Haenszel chi-square test.

24

### Introduction

		purchase	
		< \$100	\$100 +
gender	Female	0.58	0.42
	Male	0.68	0.32

Row probabilities of **gender** by **purchase**

25

There appears to be an association between **gender** and **purchase** because the row probabilities are different in each column. To test for this association, you are assessing whether the probability of females purchasing items of 100 dollars or more (0.42) is significantly different from the probability of males purchasing items of 100 dollars or more (0.32).

### Null Hypothesis

- There is no association between **gender** and **purchase**.
- The probability of purchasing items of 100 dollars or more is the same whether you are male or female.

26

### Alternative Hypothesis

- There is an association between **gender** and **purchase**.
- The probability of purchasing items over 100 dollars is different between males and females.

27

## Chi-Square Test

### NO ASSOCIATION

observed frequencies = expected frequencies

### ASSOCIATION

observed frequencies ≠ expected frequencies

28

A commonly used test that examines whether there is an association between two categorical variables is the Pearson chi-square test. The chi-square test measures the difference between the observed cell frequencies and the cell frequencies that are expected if there is no association between the variables. If you have a significant chi-square statistic, there is strong evidence that an association exists between your variables.



The expected frequencies are calculated by the formula (row total \* column total) / sample size.

## p-Value for Chi-Square Test

The *p*-value is the

- probability of observing a chi-square statistic at least as large as the one actually observed, given that there is no association between the variables
- probability of the association you observe in the data occurring by chance.

29

In general, the larger the chi-square values, the smaller the *p*-value, which means that you have more evidence against the null hypothesis.

## Chi-Square Tests

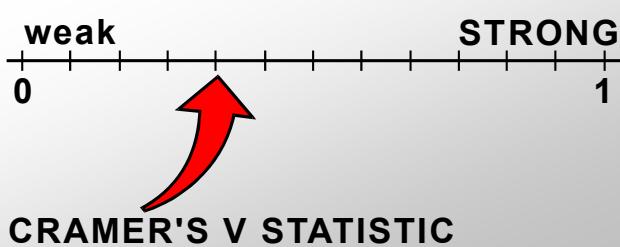
Chi-square tests and the corresponding  $p$ -values

- determine whether an association exists
- do not measure the strength of an association
- depend on and reflect the sample size.

30

If you double the size of your sample by duplicating each observation, you double the chi-square statistic even though the strength of the association does not change.

## Measures of Association



31

One measure of the strength of the association between two nominal variables is Cramer's V statistic. It is in the range of  $-1$  to  $1$  for 2-by-2 tables and  $0$  to  $1$  for larger tables. Values further away from  $0$  indicate the presence of a relatively strong association.

Cramer's V statistic is derived from the Pearson chi-square statistic.



## Chi-Square Test

Example: Use the FREQ procedure to test for an association between the variables **gender** and **purchase**. Also generate the expected cell frequencies and the cell's contribution to the total chi-square statistic.

```
/* c5demo03 */
proc freq data=sasuser.b_sales_inc;
  tables gender*purchase
    / chisq expected cellchi2 nocol nopercnt;
  format purchase purfmt.;
  title1 'Association between GENDER and PURCHASE';
run;
```

Selected TABLES statement options:

CHISQ	produces the chi-square test of association and the measures of association based upon the chi-square statistic.
EXPECTED	prints the expected cell frequencies under the hypothesis of no association.
CELLCHI2	prints each cell's contribution to the total chi-square statistic.
NOCOL	suppresses printing the column percentages.
NOPERCENT	suppresses printing the cell percentages.

The frequency table is shown below.

Association between GENDER and PURCHASE			
The FREQ Procedure			
Table of gender by purchase			
gender		purchase	
Frequency			
Expected			
Cell Chi-Square			
Row Pct	< \$100	\$100 +	Total
Female	139 149.79 0.7774 57.92	101 90.209 1.2909 42.08	240
Male	130 119.21 0.9769 68.06	61 71.791 1.6221 31.94	191
Total	269	162	431

It appears that the cell for **purchase** = 1 (100 dollars or more) and **gender** = Male contributes the most to the chi-square statistic.



The cell chi-square is calculated using the formula  
 $(\text{observed frequency} - \text{expected frequency})^2 / \text{expected frequency}$ .

The overall chi-square statistic is calculated by adding up the cell chi-square values over all rows and columns:  $\sum ((\text{observed} - \text{expected})^2 / \text{expected})$ .

Below is the table that shows the chi-square test and Cramer's V.

Statistics for Table of gender by purchase			
Statistic	DF	Value	Prob
Chi-Square	1	4.6672	0.0307
Likelihood Ratio Chi-Square	1	4.6978	0.0302
Continuity Adj. Chi-Square	1	4.2447	0.0394
Mantel-Haenszel Chi-Square	1	4.6564	0.0309
Phi Coefficient		-0.1041	
Contingency Coefficient		0.1035	
Cramer's V		-0.1041	

Fisher's Exact Test			
Cell (1,1) Frequency (F)	139	Left-sided Pr <= F	0.0195
Right-sided Pr >= F	0.9883		
Table Probability (P)	0.0078		
Two-sided Pr <= P	0.0355		

Sample Size = 431

Because the *p*-value for the chi-square statistic is 0.0307, which is below .05, you reject the null hypothesis at the 0.05 level and conclude there is evidence of an association between **gender** and **purchase**. However, Cramer's V indicates that the association detected with the chi-square test is relatively weak. This means that the association was detected because of the large sample size, not because of its strength.

### When Not to Use the Chi-Square Test

*Expected*

11	4
6	3

When more than 20% of cells have expected counts less than five

35

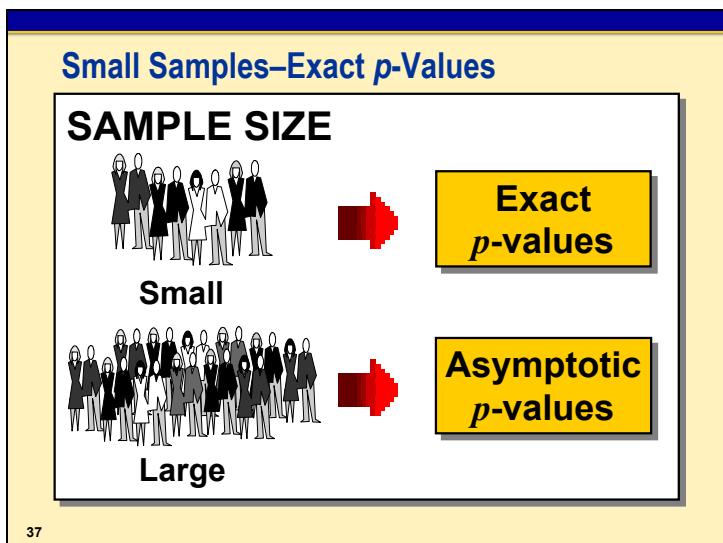
There are times when the chi-square test might not be appropriate. In fact, when more than 20% of the cells have expected cell frequencies of less than 5, the chi-square test might not be valid. This is because the  $p$ -values are based on the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large. Therefore, when the sample sizes are small, the asymptotic (large sample)  $p$ -values might not be valid.

### Observed versus Expected Values

Observed Values			Expected Values		
1	5	8	3.43	4.57	6.00
5	6	7	4.41	5.88	7.71
6	5	6	4.16	5.55	7.29

36

The criterion for the chi-square test is based on the expected values, not the observed values. In the slide above, 1 out of 9, or 11% of the cells, have observed values less than 5. However, 4 out of 9, or 44%, of the cells have expected values less than 5. Therefore, the chi-square test might not be valid.



The EXACT statement provides exact  $p$ -values for many tests in the FREQ procedure. Exact  $p$ -values are useful when the sample size is small, in which case the asymptotic  $p$ -values might not be useful.

However, large data sets (in terms of sample size, number of rows, and number of columns) can require a prohibitive amount of time and memory for computing exact  $p$ -values. For large data sets, consider whether exact  $p$ -values are needed or whether asymptotic  $p$ -values might be quite close to the exact  $p$ -values.

**Exact  $p$ -Values for Pearson Chi-Square**

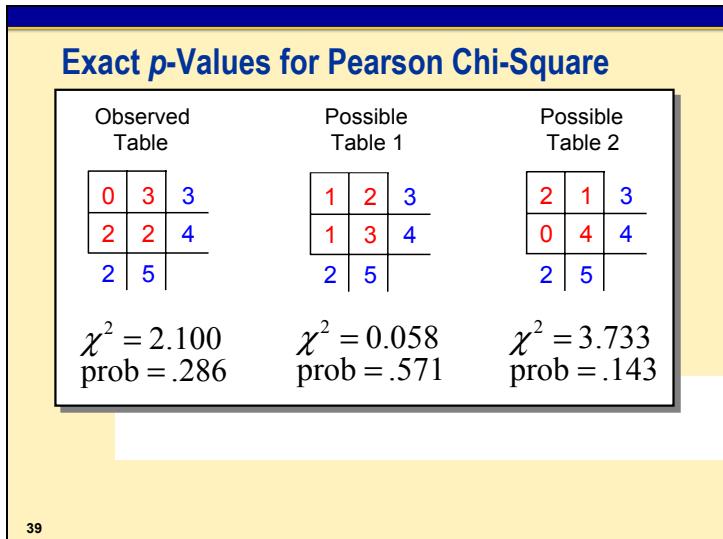
The table is titled 'Observed Table' and contains the following data:

		0	3	3
2	2	2	4	
	2	5		

38

Exact  $p$ -values reflect the probability of observing a table with at least as much evidence of an association as the one actually observed, given there is no association between the variables. If your significance level is .05, exact  $p$ -values below .05 reflect significant associations.

For example, consider the table above. With such a small sample size, the asymptotic  $p$ -values would not be valid.



A key assumption behind the computation of exact  $p$ -values is that the column totals and row totals are fixed. Thus, there are a total of three possible tables.

To compute an exact  $p$ -value for this example, examine the chi-square value for each table and the probability that the table occurs given the three tables (the probabilities add up to 1). The Observed Table has a chi-square value of 2.100, so any table with a chi-square value of 2.100 or higher would be used to compute the exact  $p$ -value. Thus, the exact  $p$ -value would be 0.286 (Observed Table) + 0.143 (Possible Table 2) = .429. This means you have a 43% chance of obtaining a table with at least as much of an association as the observed table simply by random chance.



## Exact *p*-Values for the Pearson Chi-Square Test

Example: Invoke PROC FREQ and produce exact *p*-values for the Pearson chi-square test. Use the data set **sasuser.b\_exact**, which has the data from the previous example.

```
/* c5demo04 */
proc freq data=sasuser.b_exact;
  tables a*b;
  exact pchi;
run;
```

Selected FREQ procedure statement:

EXACT produces exact *p*-values for the statistics listed as keywords. If you use only one TABLES statement, you do not need to specify options in the TABLES statement to perform the analyses that the EXACT statement requests.

Selected EXACT statement option:

PCHI requests exact *p*-values for the chi-square statistics. It also produces Cramer's V and other related statistics.

 If you use multiple TABLES statements and want exact computations, you must specify options in the TABLES statement to compute the desired statistics.

The frequency table is shown below.

Association using EXACT PCHI statement				
The FREQ Procedure				
		Table of a by b		
a	b			
Frequency		1	2	Total
Percent		0.00	42.86	3
Row Pct		0.00	100.00	42.86
Col Pct		0.00	60.00	
1		3		
2		2	2	4
		28.57	28.57	57.14
		50.00	50.00	
		100.00	40.00	
Total		2	5	7
		28.57	71.43	100.00

This is the observed table from the previous example.

The Pearson Chi-Square Test table contains the Exact Pr  $\geq$  ChiSq value of 0.4286 and is shown below.

Statistics for Table of a by b			
Statistic	DF	Value	Prob
Chi-Square	1	2.1000	0.1473
Likelihood Ratio Chi-Square	1	2.8306	0.0925
Continuity Adj. Chi-Square	1	0.3646	0.5460
Mantel-Haenszel Chi-Square	1	1.8000	0.1797
Phi Coefficient		-0.5477	
Contingency Coefficient		0.4804	
Cramer's V		-0.5477	

WARNING: 100% of the cells have expected counts less than 5.  
 (Asymptotic) Chi-Square may not be a valid test.

Pearson Chi-Square Test			
Chi-Square		2.1000	
DF		1	
Asymptotic Pr > ChiSq		0.1473	
Exact Pr $\geq$ ChiSq		0.4286	

Notice the difference between the exact  $p$ -value (0.4286) and the asymptotic  $p$ -value (0.1473) in the Pearson Chi-Square Test table. Exact  $p$ -values tend to be larger than asymptotic  $p$ -values because the exact tests are more conservative.

The warning message informs you that because of the small sample size, the asymptotic chi-square might not be a valid test.

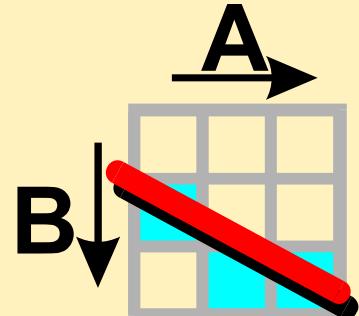
### Association among Ordinal Variables

Is  associated with  ?

? 42

You have already seen that **purchase** and **gender** have a significant association. Another question you can ask is whether **purchase** and **income** have a significant association. You can use the chi-square test, but because **income** is ordinal and **purchase** can be considered ordinal, you might want to test for an ordinal association. The appropriate test for ordinal associations is the Mantel-Haenszel chi-square test.

### Mantel-Haenszel Chi-Square Test



Test Ordinal Association 43

The Mantel-Haenszel chi-square test is particularly sensitive to ordinal associations. An *ordinal association* implies that as one variable increases, the other variable tends to increase or decrease. For the test results to be meaningful when there are variables with more than two levels, the levels must be in a logical order.

Null hypothesis: There is no ordinal association between the row and column variables.

Alternative hypothesis: There is an ordinal association between the row and column variables.

## Mantel-Haenszel Chi-Square Test

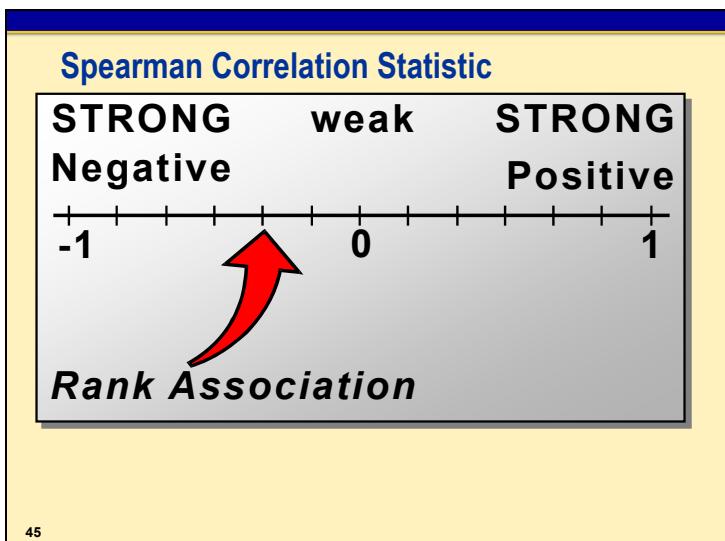
The Mantel-Haenszel chi-square test

- determines whether an ordinal association exists
- does not measure the strength of the ordinal association
- depends upon and reflects the sample size.

44

The Mantel-Haenszel chi-square statistic is more powerful than the general association chi-square statistic for detecting an ordinal association. The reasons are that

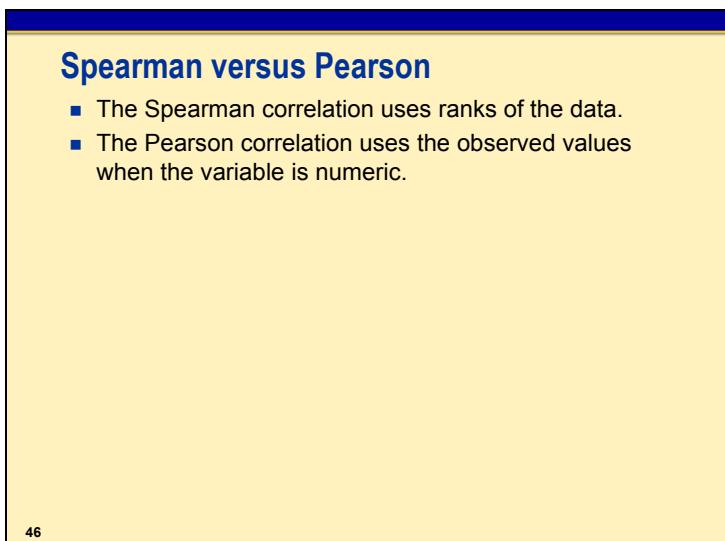
- all of the Mantel-Haenszel statistic's power is concentrated toward that objective
- the power of the general association statistic is dispersed over a greater number of alternatives.



45

To measure the strength of the ordinal association, you can use the Spearman correlation statistic. This statistic

- has a range between  $-1$  and  $1$
- has values close to  $1$  if there is a relatively high degree of positive correlation
- has values close to  $-1$  if there is a relatively high degree of negative correlation
- is appropriate only if both variables are ordinally scaled and the values are in a logical order.



46

The Spearman statistic can be interpreted as the Pearson correlation between the ranks on variable X and the ranks on variable Y.

For character values, SAS assigns by default a 1 to column 1, a 2 to column 2, and so on. You can change the default with the SCORES= option in the TABLES statement.



## Detecting Ordinal Associations

Example: Use PROC FREQ to test whether an ordinal association exists between **purchase** and **income**. Use the variable **inclevel** and the appropriate format to ensure that the income levels are in a logical order.

```
/* c5demo05 */
proc freq data=sasuser.b_sales_inc;
  tables inclevel*purchase / chisq measures cl;
  format inclevel incfmt. purchase purfmt. ;
  title1 'Ordinal Association between INCLEVEL and PURCHASE?';
run;
```

Selected TABLES statement options:

- |          |  |
|----------|--|
| CHISQ    | produces the Pearson chi-square, the likelihood-ratio chi-square, and the Mantel-Haenszel chi-square. It also produces measures of association based on chi-square such as the phi coefficient, the contingency coefficient, and Cramer's V. |
| MEASURES | produces the Spearman correlation statistic along with other measures of association.  |
| CL       | produces confidence bounds for the MEASURES statistics.  |

The crosstabulation is shown below.

Ordinal Association between INCLEVEL and PURCHASE?			
The FREQ Procedure			
Table of inclevel by purchase			
inclevel	purchase		
Frequency	< \$100	\$100 +	Total
Percent			
Row Pct			
Col Pct			
Low Income	90 20.88 68.18 33.46	42 9.74 31.82 25.93	132 30.63
Medium Income	98 22.74 68.06 36.43	46 10.67 31.94 28.40	144 33.41
High Income	81 18.79 52.26 30.11	74 17.17 47.74 45.68	155 35.96
Total	269 62.41	162 37.59	431 100.00

The results of the Mantel-Haenszel chi-square test are shown below.

Statistics for Table of inclevel by purchase			
Statistic	DF	Value	Prob
Chi-Square	2	10.6404	0.0049
Likelihood Ratio Chi-Square	2	10.5425	0.0051
Mantel-Haenszel Chi-Square	1	8.1174	0.0044
Phi Coefficient		0.1571	
Contingency Coefficient		0.1552	
Cramer's V		0.1571	

Because the *p*-value of the Mantel-Haenszel chi-square is 0.0044, you can conclude at the 0.05 significance level that there is evidence of an ordinal association between **purchase** and **income**.

The Spearman correlation statistic and the 95% confidence bounds are shown below.

Ordinal Association between INCLEVEL and PURCHASE?				
The FREQ Procedure				
Statistics for Table of inclevel by purchase				
Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.2324	0.0789	0.0777	0.3871
Kendall's Tau-b	0.1312	0.0454	0.0423	0.2201
Stuart's Tau-c	0.1466	0.0508	0.0471	0.2461
Somers' D C R	0.1102	0.0382	0.0353	0.1850
Somers' D R C	0.1562	0.0540	0.0505	0.2620
Pearson Correlation	0.1374	0.0480	0.0433	0.2315
Spearman Correlation	0.1391	0.0481	0.0449	0.2334
Lambda Asymmetric C R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R C	0.0616	0.0470	0.0000	0.1536
Lambda Symmetric	0.0388	0.0300	0.0000	0.0976
Uncertainty Coefficient C R	0.0185	0.0114	0.0000	0.0408
Uncertainty Coefficient R C	0.0112	0.0069	0.0000	0.0246
Uncertainty Coefficient Symmetric	0.0139	0.0086	0.0000	0.0307
Sample Size = 431				

The Spearman correlation statistic (0.1391) indicates that there is a relatively small positive ordinal relationship between **income** and **purchase** (as **income** levels increase, **purchase** levels increase).

The ASE is the asymptotic standard error (0.0481), which is an appropriate measure of the standard error for larger samples.

Because the 95% confidence interval (0.0449, 0.2334) for the Spearman correlation statistic does not contain 0, the relationship is significant at the 0.05 significance level.

The confidence bounds are valid only if your sample size is large. A general guideline is to have a sample size of at least 25 for each degree of freedom in the Pearson chi-square statistic.



Refer to Exercise 1 for Chapter 5 in Appendix A.

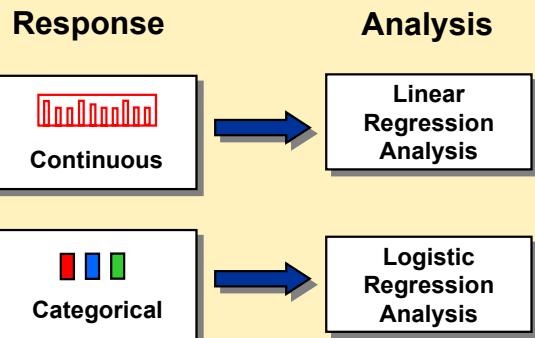
## 5.3 Introduction to Logistic Regression

### Objectives

- Explain the concepts of logistic regression.
- Fit a binary logistic regression model using the LOGISTIC procedure.
- Explain effect and reference cell coding.
- Define and explain the odds ratio.
- Explain the standard output from the LOGISTIC procedure.

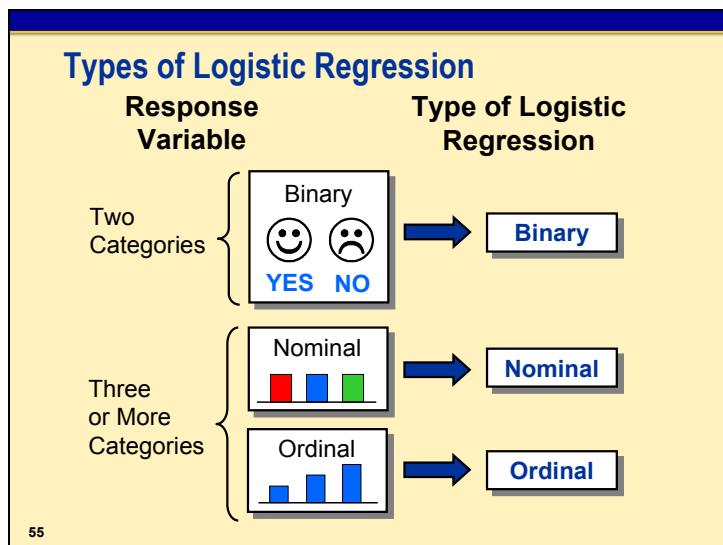
53

### Overview



54

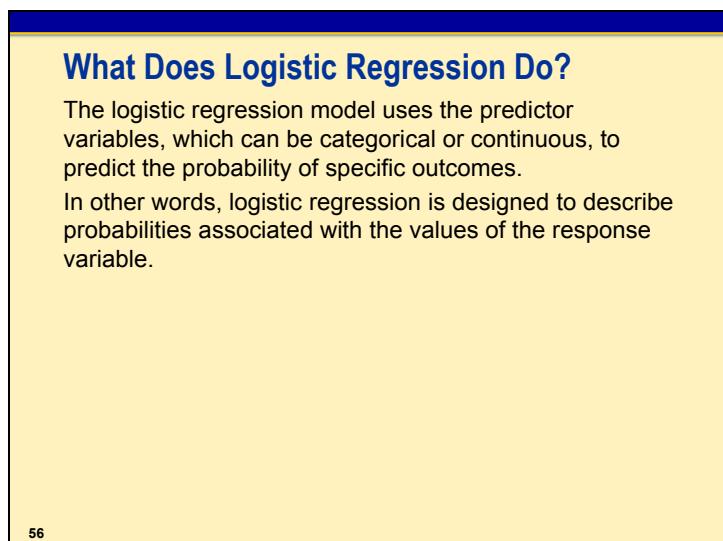
*Regression analysis* enables you to characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous. In *logistic regression*, the response variable is categorical.



If the response variable is dichotomous (two categories), the appropriate logistic regression model is binary logistic regression.

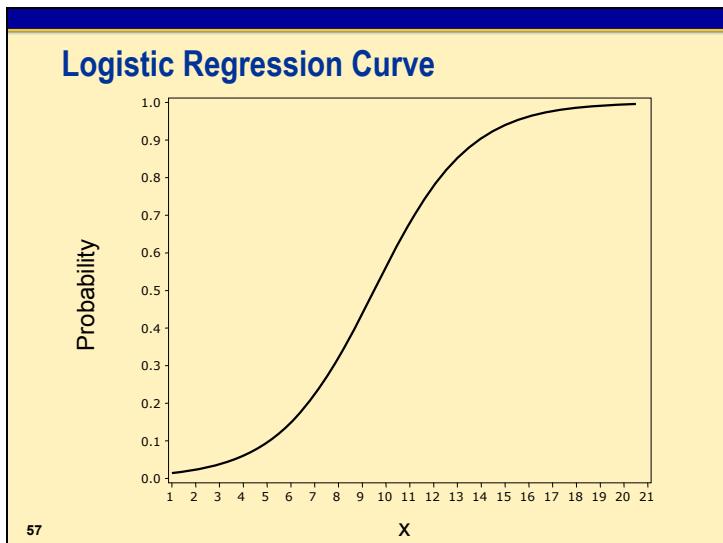
If you have more than two categories (levels) within the response variable, then there are two possible logistic regression models:

1. If the response variable is nominal, you fit a nominal logistic regression model.
2. If the response variable is ordinal, you fit an ordinal logistic regression model.



Because you are modeling probabilities, a continuous linear regression model would not be appropriate. One problem is that the predicted values from a linear model can assume, theoretically, any value. However, probabilities are by definition bounded between 0 and 1. Logistic regression models ensure that the estimated probabilities are between 0 and 1.

Another problem is that the relationship between the probability of the outcome and a predictor variable is usually nonlinear rather than linear. In fact, the relationship often resembles an S-shaped curve.



The nonlinear relationship between the probability of the outcome and the predictor variables is solely due to the constrained scale of the probabilities. Furthermore, the relationship is fairly linear in the middle of the range of the probabilities (.20 to .80) and fairly nonlinear at the end of the range (0 to .20 and .80 to 1).

The parameter estimate of this curve determines the rate of increase or decrease of the estimated curve. When the parameter estimate is greater than 0, the probability of the outcome increases as the predictor variable values increase. When the parameter estimate is less than 0, the probability decreases as the predictor variable values increase. As the absolute value of the parameter estimate increases, the curve has a steeper rate of change. When the parameter estimate is equal to 0, the curve resembles a straight line.

### Logit Transformation

Logistic regression models transform probabilities called logits.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

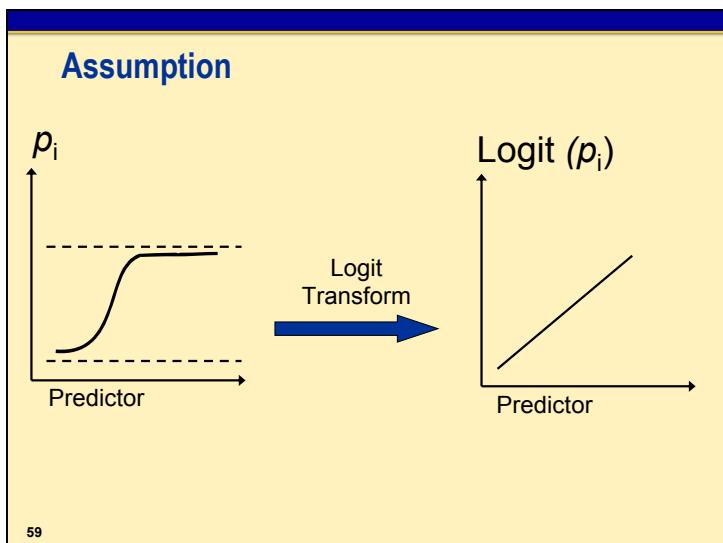
where

- $i$  indexes all cases (observations)
- $p_i$  is the probability the event (a sale, for example) occurs in the  $i$ th case
- log is the natural log (to the base e).

58

A logistic regression model applies a transformation to the probabilities. The probabilities are transformed because the relationship between the probabilities and the predictor variable is nonlinear.

The logit transformation ensures that the model generates estimated probabilities between 0 and 1.



Assumption in logistic regression:

The logit transformation of the probabilities results in a linear relationship with the predictor variables.

To verify this assumption, it would be useful to plot the logits by the predictor variable. Logit plots are illustrated in a later section.

### Logistic Regression Model

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

where

$\text{logit}(p_i)$  logit transformation of the probability of the event

$\beta_0$  intercept of the regression line

$\beta_1$  slope of the regression line

$\varepsilon_i$  error (residual) associated with each observation.

60

For a binary outcome variable, the linear logistic model with one predictor variable has the form above.

Unlike linear regression, the logit is not normally distributed and the variance is not constant. Also, logistic regression usually requires a more complex estimation method called maximum likelihood to estimate the parameters than linear regression. This method finds the parameter estimates that are most likely to occur given the data. This is accomplished by maximizing the likelihood function that expresses the probability of the observed data as a function of the unknown parameters.

## LOGISTIC Procedure

General form of the LOGISTIC procedure:

```
PROC LOGISTIC DATA=SAS-data-set <options>;
  CLASS variables </ options>;
  MODEL response=predictors </ options>;
  OUTPUT OUT=SAS-data-set keyword=name
    </ options>;
RUN;
```

62

Selected LOGISTIC procedure statements:

- CLASS        names the classification variables to be used in the analysis. The CLASS statement must precede the MODEL statement.
- MODEL        specifies the response variable and the predictor variables.
- OUTPUT       creates an output data set containing all the variables from the input data set and any requested statistics.

## Effect Coding: Two Levels

Design Variables

<u>Class</u>	<u>Value</u>	<u>1</u>
gender	Female	1
	Male	-1

63

### Effect Coding: Three Levels

Design Variables

<u>Class</u>	<u>Value</u>	<u>Label</u>	<u>1</u>	<u>2</u>
inclevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	-1	-1

64

For *effect coding* (also called *deviation from the mean coding*), the number of design variables created is the number of levels of the CLASS variable minus 1. For example, because the variable **inclevel** has three levels, two design variables were created. For the last level of the CLASS variable (High), all the design variables have a value of -1. Parameter estimates of the CLASS main effects using this coding scheme estimate the difference between the effect of each level and the average effect over all levels.

### Effect Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

$\beta_0$  = the average value of the logit across all categories

$\beta_1$  = the difference between the logit for Low income and the average logit

$\beta_2$  = the difference between the logit for Medium income and the average logit

$-(\beta_1 + \beta_2)$  = the difference between the average logit and the logit for High income

65

Because the sum of the deviations around the mean must equal zero, the effect for High income must be the negative of the sum of the effects for Low and Medium income.

### Reference Cell Coding: Two Levels

Design Variables

<u>Class</u>	<u>Value</u>	<u>1</u>
gender	Female	1
	Male	0

66

### Reference Cell Coding: Three Levels

Design Variables

<u>Class</u>	<u>Value</u>	<u>Label</u>	<u>1</u>	<u>2</u>
inclevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	0	0

67

For *reference cell coding*, parameter estimates of the CLASS main effects estimate the difference between the effect of each level and the last level, called the *reference level*. For example, the effect for the level Low estimates the difference between Low and High. You can choose the reference level in the CLASS statement.

### Reference Cell Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

$\beta_0$  = the value of the logit when income is High

$\beta_1$  = the difference between the logits for Low and High income

$\beta_2$  = the difference between the logits for Medium and High income



## Binary Logistic Regression

Example: Fit a binary logistic regression model in PROC LOGISTIC. Select **purchase** as the outcome variable and **gender** as the predictor variable. Specify reference cell coding and specify **Male** as the reference group. Also use the **EVENT=** option to model the probability of spending 100 dollars or more and request profile likelihood confidence intervals around the estimated odds ratios.

```
/* c5demo06 */
proc logistic data=sasuser.b_sales_inc;
  class gender (param=ref ref='Male');
  model purchase(event='1')=gender / clodds=pl;
  title1 'LOGISTIC MODEL (1): purchase=gender';
run;
```

Selected MODEL statement option:

**EVENT=** specifies the event category for the binary response model. PROC LOGISTIC models the probability of the event category. The **EVENT=** option has no effect when there are more than two response categories. You can specify the value (formatted if a format is applied) of the event category in quotes or you can specify one of the following keywords. The default is **EVENT=FIRST**.

**FIRST** designates the first ordered category as the event.

**LAST** designates the last ordered category as the event.

Selected CLASS statement options:

**PARAM=** specifies the parameterization method for the classification variable or variables. Design matrix columns are created from CLASS variables according to the following coding schemes. There are several codes that can be used, but two are listed below:

**EFFECT** specifies effect coding (default).

**REFERENCE | REF** specifies reference cell coding.

**REF=** specifies the reference level for **PARAM=EFFECT** or **PARAM=REFERENCE**.

Selected MODEL statement option:

**CLODDS=PL** requests profile likelihood confidence intervals for the odds ratios of all predictor variables, which are desirable for small sample sizes.

 If there are numerous levels in the CLASS variable, you might want to reduce the number of levels using subject matter knowledge. This is especially important when the levels have few or no observations.

## LOGISTIC MODEL (1): purchase=gender

## The LOGISTIC Procedure

## Model Information

Data Set SASUSER.B\_SALES\_INC  
Response Variable purchase  
Number of Response Levels 2  
Model binary logit  
Optimization Technique Fisher's scoring

Number of Observations Read 431  
Number of Observations Used 431

## Response Profile

Ordered Value	purchase	Total Frequency
1	0	269
2	1	162

Probability modeled is purchase=1.

## Class Level Information

Class	Value	Design Variables
gender	Female	1
	Male	0

The Model Information table describes the data set, the response variable, the number of response levels, the type of model, the algorithm used to obtain the parameter estimates, and the number of observations read and used.

The Response Profile table shows the response variable values listed according to their ordered values. By default, PROC LOGISTIC orders the response variable alphanumerically so that it bases the logistic regression model on the probability of the smallest value. Because you used the EVENT=option, in this example, the model is based on the probability of purchasing items of 100 dollars or more (**purchase**=1).

The Response Profile table also shows the value of the response variable and the frequency.

The Class Level Information table includes the predictor variable in the CLASS statement. Because you used the PARAM=REF and REF='Male' options, this table reflects your choice of **gender**=Male as the reference level. The design variable is 1 when **gender**=Female and 0 when **gender**=Male.

Model Convergence Status		
Convergence criterion (GCONV=1E-8) satisfied.		
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	572.649	569.951
SC	576.715	578.084
-2 Log L	570.649	565.951

The Model Convergence Status simply informs you that the convergence criterion was met. There are a number of options to control the convergence criterion, but the default is the gradient convergence criterion with a default value of 1E-8 (0.00000001).

The Model Fit Statistics provides three tests: AIC is Akaike's 'A' information criterion, SC is the Schwarz criterion, and  $-2\log L$  is the  $-2$  log likelihood. AIC and SC are goodness-of-fit measures you can use to compare one model to another. Lower values indicate a more desirable model. AIC adjusts for the number of predictor variables, and SCs adjust for the number of predictor variables and the number of observations. SC uses a bigger penalty for extra variables and therefore favors more parsimonious models.



A reference for AIC can be found in Findley and Parzen (1995).

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.6978	1	0.0302
Score	4.6672	1	0.0307
Wald	4.6436	1	0.0312

The Testing Global Null Hypothesis: BETA=0 table provides three statistics to test the null hypothesis that all regression coefficients of the model are 0.

Using the Likelihood Ratio test, a significant  $p$ -value for the Likelihood Ratio test provides evidence that at least one of the regression coefficients for an explanatory variable is nonzero (in this example the  $p$ -value is 0.0302, which is significant at the .05 level). This statistic is similar to the overall  $F$  test in linear regression. The Score and Wald tests are also used to test whether all the regression coefficients are 0. The likelihood ratio test is the most reliable, especially for small sample sizes (Agresti 1996).

## Type 3 Analysis of Effects

Effect	DF	Chi-Square	Wald	Pr > ChiSq
gender	1	4.6436		0.0312

The Type 3 Analysis of Effects table is generated when a predictor variable is used in the CLASS statement. The listed effect (variable) is tested using the Wald Chi-Square statistic (in this example, 4.6436 with a *p*-value of 0.0312). This analysis is similar to the individual *t*-test in the REG procedure. Because **gender** is the only variable in the model, the value listed in the table will be identical to the Wald test in the Testing Global Null Hypothesis table.

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7566	0.1552	23.7700	<.0001
gender Female	1	0.4373	0.2029	4.6436	0.0312

The Analysis of Maximum Likelihood Estimates table lists the estimated model parameters, their standard errors, Wald tests, and odds ratios.

The parameter estimates are the estimated coefficients of the fitted logistic regression model. The logistic regression equation is  $\text{logit}(\hat{p}) = -0.7566 + 0.4373 * \text{gender}$ , for this example.

The Wald chi-square, and its associated *p*-value, tests whether the parameter estimate is significantly different from 0. For this example, both the *p*-values for the intercept and the variable **gender** are significant at the 0.05 significance level.

## What Is an Odds Ratio?

An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Example: How much more likely are females to purchase 100 dollars or more in items compared to males?

72

## Probability of Outcome

	Outcome		Total
	Yes	No	
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Probability of a **Yes outcome**  
in Group A =  $20/80$  (**0.25**)

Probability of a **No outcome**  
in Group A =  $60/80$  (**0.75**)

73

You have a 25% chance of getting the outcome in group A.

What is the chance of getting the outcome in group B?

**Odds**

**Odds of Outcome in Group A**

$$\frac{\text{Probability of a Yes outcome in Group A}}{\text{Probability of a No outcome in Group A}}$$

$$0.25 \div 0.75 = 0.33$$

74

	Outcome		<b>Total</b>
	<b>YES</b>	<b>NO</b>	
<b>Group A</b>	20	60	80
<b>Group B</b>	10	90	100
	30	150	180

The odds of an outcome is the ratio of the expected number of times that the outcome will occur to the expected number of times the outcome will **not** occur. In other words, the odds is simply the ratio of the probability of the outcome to the probability of no outcome. The odds for group A equals 0.33 indicating that you expect only 1/3 as many occurrences as non-occurrences in group A.

What is the odds of getting the outcome in group B?

**Odds Ratio**

**Odds Ratio of Group A to Group B**

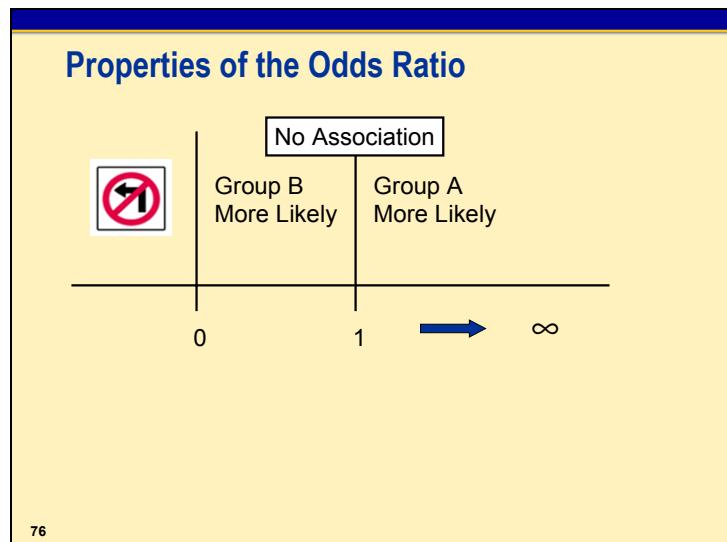
$$\frac{\text{Odds of outcome in Group A}}{\text{Odds of outcome in Group B}}$$

$0.33 \div 0.11 = 3$

75

	Outcome		<b>Total</b>
	<b>YES</b>	<b>NO</b>	
<b>Group A</b>	20	60	80
<b>Group B</b>	10	90	100
	30	150	180

The odds ratio of group A to B equals 3, indicating that the odds of getting the outcome in group A are 3 times those in group B.



The odds ratio shows the strength of the association between the predictor variable and the outcome variable. If the odds ratio is 1, then there is no association between the predictor variable and the outcome. If the odds ratio is greater than 1, then group A is more likely to have the outcome. If the odds ratio is less than 1, then group B is more likely to have the outcome.

## Odds Ratio Calculation from the Current Logistic Regression Model

Logistic regression model:

$$\text{logit}(\hat{p}) = \log(\text{odds}) = \beta_0 + \beta_1 * (\text{gender})$$

Odds ratio (females to males):

$$\text{odds}_{\text{females}} = e^{\beta_0 + \beta_1}$$

$$\text{odds}_{\text{males}} = e^{\beta_0}$$

$$\text{odds ratio} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

77

The odds ratio is computed by exponentiating the parameter estimate for the predictor variable. For this example, the odds ratio for **gender** (coded 1 for females and 0 for males) compares the predicted odds of females to purchase 100 dollars or more in items compared to males.

**Odds Ratio for Categorical Predictor**

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gender Female vs Male	1.549	1.040	2.305

Profile Likelihood Confidence Interval for Adjusted Odds Ratios			
Effect	Unit	Estimate	95% Confidence Limits
gender	1.0000	1.549	1.043 2.312

78

The odds ratio indicates that females are 1.55 times more likely to purchase 100 dollars or more than males.

The 95% confidence limits indicate that you are 95% confident that the true odds ratio is between 1.04 and 2.31. Because the 95% confidence interval does not include 1.00, the odds ratio is significant at the .05 significance level.

 If you want a different significance level for the confidence intervals, you can use the ALPHA= option in the MODEL statement. The value must be between 0 and 1. The default value of .05 results in the calculation of a 95% confidence interval.

The profile likelihood confidence intervals are different from the Wald-based confidence intervals. This difference is because the Wald confidence intervals use a normal approximation, whereas the profile likelihood confidence intervals are based on the value of the log-likelihood. These likelihood-ratio confidence intervals require much more computation but are generally preferred to the Wald confidence intervals, especially for sample sizes less than 50 (Allison 1999).

**Odds Ratio for Continuous Predictor**

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.020	0.985	1.056

Profile Likelihood Confidence Interval for Adjusted Odds Ratios			
Effect	Unit	Estimate	95% Confidence Limits
Age	1.0000	1.020	0.985 1.056

79

For a continuous predictor variable, the odds ratio measures the increase or decrease in odds associated with a one-unit difference on the predictor variable. For example, **age** shows an odds ratio of 1.020, which means that a person who is one year older has 2%  $((1.020 - 1.000) * 100)$  greater odds of purchasing \$100 or more of items from the catalog than the younger person. The model assumes that this odds ratio is the same across all ages, so it does not matter if you compare a 21-year-old with a 20-year-old or a 35-year-old with a 34-year old. Notice that the confidence interval for age includes 1, which corroborates the conclusion of nonsignificance from the  $p$ -value.

### Model Assessment: Comparing Pairs

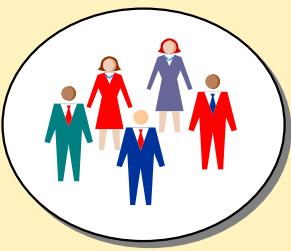
- Counting concordant, discordant, and tied pairs is a way to assess how well the model predicts its own data and therefore how well the model fits.
- In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs.

80

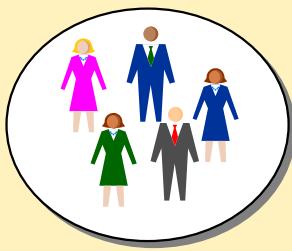
### Comparing Pairs

To find concordant, discordant, and tied pairs, compare everyone who had the outcome of interest against everyone who did not.

< \$100



\$100 +



81

### Concordant Pair

Compare a woman who bought more than \$100 worth of goods from the catalog and a man who did not.



$$\text{P}(100+) = .32$$



$$\text{P}(100+) = .42$$

The actual sorting agrees with the model.  
This is a **concordant** pair.

82

For all pairs of observations with different values of the response variable, a pair is *concordant* if the observation with the outcome has a **higher** predicted outcome probability (based on the model) than the observation without the outcome.

### Discordant Pair

Compare a man who bought more than \$100 worth of goods from the catalog and a woman who did not.



$$\text{P}(100+) = .42$$



$$\text{P}(100+) = .32$$

The actual sorting disagrees with the model.  
This is a **discordant** pair.

83

A pair is *discordant* if the observation with the outcome has a **lower** predicted outcome probability than the observation without the outcome.

## Tied Pair

Compare two women. One bought more than \$100 worth of goods from the catalog, and the other did not.

< \$100



$$P(100+) = .42$$

\$100 +



$$P(100+) = .42$$

The model cannot distinguish between the two.  
This is a **tied** pair.

84

A pair is *tied* if it is neither concordant nor discordant (the probabilities are the same).

## Concordant versus Discordant

Customer Purchasing Over \$100			
Customer Purchasing Less Than \$100	Predicted Outcome Probability	Females (0.42)	Males (0.32)
	Females (0.42)	Tie	Discordant Pair
	Males (0.32)	Concordant Pair	Tie

85

This table is a summary of discordant, concordant, and tied pairs. Because the predictor variable (**gender**) has only two levels, there are only two predicted outcome probabilities for purchasing items of 100 dollars or more (Female=0.42 and Male=0.32). For all pairs of observations with different outcomes (making purchases of 100 dollars or more versus making purchases of less than 100 dollars), a comparison is made of the predicted outcome probabilities. If the observation with the outcome (in this case making purchases of 100 dollars or more) has a higher predicted outcome probability compared to an observation without the outcome, the pair is concordant. However, if the observation with the outcome has a lower predicted outcome probability compared to the predicted outcome probability of an observation without the outcome, the pair is discordant. If the predicted outcome probabilities are tied, then the pair is tied.

In more complex models, there are more than two predicted outcome probabilities. However, the same comparisons are made across all pairs of observations with different outcomes.

**Model: Concordant, Discordant, and Tied Pairs**

Association of Predicted Probabilities and Observed Responses		
Percent Concordant	30.1	Somers' D 0.107
Percent Discordant	19.5	Gamma 0.215
Percent Tied	50.4	Tau-a 0.050
Pairs	43578	c 0.553

86

The Association of Predicted Probabilities and Observed Responses table lists several measures of association to help you assess the predictive ability of the logistic model.

Concordant represents the percentage of concordant pairs of observations. For all pairs of observations with different values of the response variable, a pair is concordant if the observation with the outcome has a higher predicted outcome probability (based on the model) than the observation without the outcome.

Discordant represents the percentage of discordant pairs of observations. A pair is discordant if the observation with the outcome has a lower predicted outcome probability than the observation without the outcome.

Tied represents the percentage of tied pairs of observations. A pair is tied if it is neither concordant nor discordant.

You can use these percentages as goodness-of-fit measures to compare one model to another. In general, higher percentages of concordant pairs and lower percentages of discordant pairs indicate a more desirable model.

The Association of Predicted Probabilities and Observed Responses table also shows the number of observation pairs upon which the percentages are based. For this example, there are 162 observations with an outcome of 100 dollars or more and 269 observations with an outcome of Under 100 dollars. This creates  $162 \times 269 = 43578$  pairs of observations with different outcome values.

The four rank correlation indices (Somer's D, Gamma, Tau-a, and *c*) are computed from the numbers of concordant, discordant, and tied pairs of observations. In general, a model with higher values for these indices has better predictive ability than a model with lower values for these indices. The *c* statistic estimates the probability of an observation with the outcome having a higher predicted probability than an observation without the outcome.



Refer to Exercise 2 for Chapter 5 in Appendix A.

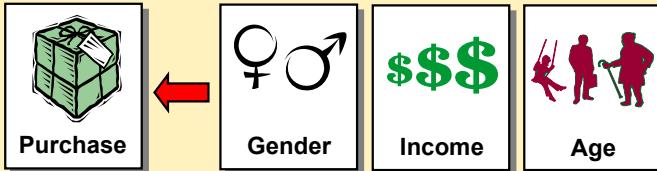
## 5.4 Multiple Logistic Regression

### Objectives

- Define and explain the adjusted odds ratio.
- Fit a multiple logistic regression model using the backward elimination method.
- Fit a multiple logistic regression model with interactions.

91

### Multiple Logistic Regression

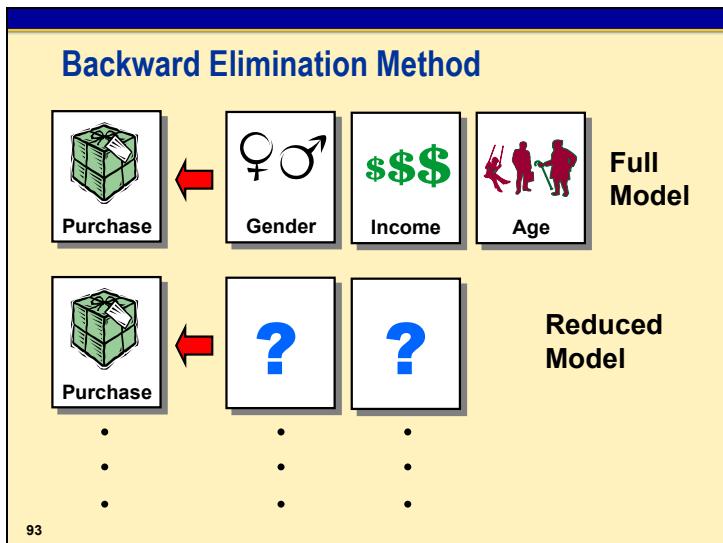


$$\text{logit } (\rho_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

92

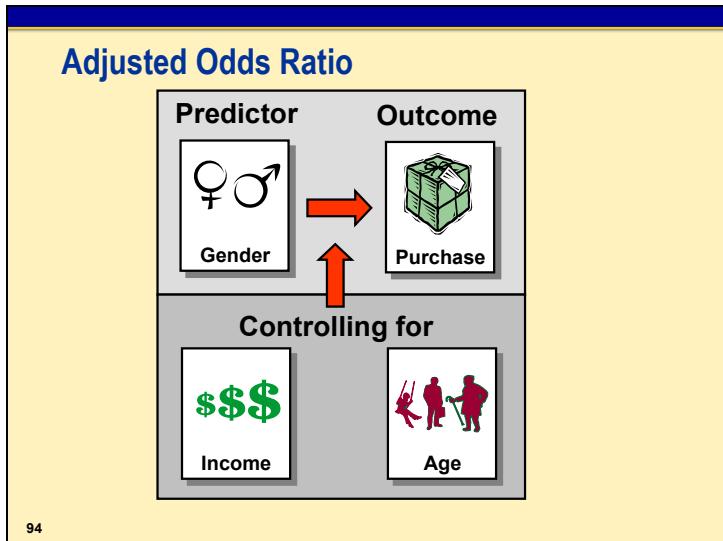
In multiple logistic regression models, several continuous or categorical predictor variables are trying to explain the variability of the response variable. The goal in multiple logistic regression is similar to that in linear multiple regression. Find the best subset of variables by eliminating unnecessary ones. Models that are parsimonious, or simple, are more likely to be numerically stable and easier to generalize.

If you have a large number of variables, you might need to try a variable reduction method such as variable clustering.



One way to eliminate unnecessary terms in a model is the *backward elimination method*. PROC LOGISTIC begins by fitting the full model with all the main effects. It then eliminates the nonsignificant parameter estimates one at a time, starting with the least significant term (the one with the largest *p*-value). The final model should only have significant main effects.

The significance level you choose depends on how much evidence you need in the significance of the predictor variables. The smaller your significance level, the more evidence you need to keep the predictor variable. In other words, the smaller your significance level, the smaller the *p*-value has to be to keep the predictor variable.



94

One major difference between a model with one predictor variable and a model with more than one predictor variable is that the reported odds ratios are now adjusted odds ratios.

*Adjusted odds ratios* measure the effect between a predictor variable and a response variable while holding all the other predictor variables constant.

For example, the odds ratio for the variable **gender** would measure the effect of **gender** on **purchase** while holding **income** and **age** constant.

The assumption is that the odds ratio for **gender** is the same regardless of the level of **income** or **age**. If that assumption is not true, you have an interaction. This is discussed later in the chapter.



## Multiple Logistic Regression

Example: Fit a multiple logistic regression model using the backward elimination method. The full model should include all the main effects.

```
/* c5demo07 */
proc logistic data=sasuser.b_sales_inc;
  class gender (param=ref ref='Male')
    income (param=ref ref='Low');
  model purchase(event='1')=gender age income / selection=backward;
  title1 'LOGISTIC MODEL (2): purchase=gender age income';
run;
```

Because **income** is a character variable, it has been added to the CLASS statement using the PARAM=REF and REF='Low' options to choose Low as the reference group.

Selected MODEL statement option:

**SELECTION=** specifies the method to select the variables in the model. BACKWARD requests backward elimination, FORWARD requests forward selection, NONE fits the complete model specified in the MODEL statement, STEPWISE requests stepwise selection, and SCORE requests best subset selection. The default is NONE.

The default significance level for the backward elimination method is .05. If you want to change the significance level, you can use the SLSTAY= option in the MODEL statement. Values must be between 0 and 1

The Model Information and Response Profile of the PROC LOGISTIC output is the same as the first model, but the title has been changed to reflect the new model.

LOGISTIC MODEL (2): purchase=gender age income

The LOGISTIC Procedure

Model Information

Data Set	SASUSER.B_SALES_INC
Response Variable	purchase
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	431
Number of Observations Used	431

Response Profile

Ordered Value	purchase	Total Frequency
1	0	269
2	1	162

Probability modeled is purchase=1.

PROC LOGISTIC identifies the chosen BACKWARD selection method and then provides a Class Level Information table.

Backward Elimination Procedure			
Class Level Information			
		Design Variables	
Class	Value	1	0
gender	Female	1	0
	Male	0	1
income	High	1	0
	Low	0	0
	Medium	0	1

The variable **income** has been added to this table, and because there are three levels, two design variable columns are displayed. You have chosen `Low` as the reference value using the `PARAM=REF` and `REF='Low'` options in the `CLASS` statement. PROC LOGISTIC has generated two Design Variables for the three levels of **income**. Design Variable 1 will be 1 when **income**=High, and will be 0 when **income**=Low or **income**=Medium. Design variable 2 will be 1 when **income**=Medium, and 0 when **income**=High or **income**=Low.

The next part of the output shows the backward elimination process in PROC LOGISTIC. At Step 0, the intercept and three predictor variables are entered into the model. The Model Fit Statistics and Testing Global Null Hypothesis tables are presented for this step.

Step 0. The following effects were entered:

```
Intercept gender age income
```

#### Model Convergence Status

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

#### Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	572.649	562.208
SC	576.715	582.539
-2 Log L	570.649	552.208

#### Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.4410	4	0.0010
Score	18.2729	4	0.0011
Wald	17.6172	4	0.0015

At Step 1, the variable **age** was removed from the model and the Model Fit Statistics and Testing Global Null Hypothesis tables are updated.

Step 1. Effect age is removed:

```

Model Convergence Status

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

Model Fit Statistics

          Intercept
          Intercept and
Criterion      Only      Covariates

AIC           572.649    562.190
SC            576.715    578.454
-2 Log L     570.649    554.190

Testing Global Null Hypothesis: BETA=0

Test          Chi-Square   DF   Pr > ChiSq

Likelihood Ratio   16.4592   3    0.0009
Score            16.3718   3    0.0010
Wald             15.8824   3    0.0012

Residual Chi-Square Test

Chi-Square   DF   Pr > ChiSq

1.9836       1    0.1590

```

The Residual Chi-Square Test table displays the joint significance of the variables not in the model (in this case, **age**). This score chi-square statistic has an asymptotic chi-squared distribution with the degrees of freedom being the difference between the full and reduced models.

When the selection process is complete, a note states that no additional variables met the specified significance level for removal from the model, and a Summary of Backward Elimination table is generated.

NOTE: No (additional) effects met the 0.05 significance level for removal from the model.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	age	1	2	1.9729	0.1601

Type 3 Analysis of Effects

Effect	DF	Chi-Square	Wald	Pr > ChiSq
gender	1	5.8211		0.0158
income	2	11.6669		0.0029

In this part of the output, the Summary of Backward Elimination table lists the step number, the name of each predictor variable (effect) that is removed from the model at each step, degrees of freedom, the number of the predictor variable in the MODEL statement, the Wald Chi-Square statistic for each variable, and the corresponding *p*-value upon which each variable's removal from the model is based.

The Type 3 Analysis of Effects table for this model indicates that the coefficients for **gender** and **income** are statistically different from 0 at the 0.05 level of significance. Note that **income** has two degrees of freedom because it has three levels.

The Analysis of Maximum Likelihood Estimates table is now examined. The *p*-value for **gender**=Female (0.0158) indicates that its coefficient is statistically different from 0 at the 0.05 level of significance. In addition, you can also state that females and males are statistically different from one another in terms of purchasing 100 dollars or more.

Analysis of Maximum Likelihood Estimates					
Parameter		DF	Estimate	Standard Error	Wald Chi-Square
Intercept		1	-1.1125	0.2403	21.4255
gender	Female	1	0.5040	0.2089	5.8211
income	High	1	0.7605	0.2515	9.1447
income	Medium	1	0.0963	0.2628	0.1342
					0.7141

The coefficient for **income**=High is also statistically different from 0, based on its *p*-value (0.0025). Because **income**=Low is the reference group, you can state that high- and low-income people are statistically different from one another with respect to purchasing 100 dollars or more. When examining **income**=Medium, the *p*-value of 0.7141 indicates that this coefficient is not different from 0. Again, because Low is the reference group, you can state that medium- and low-income people are not statistically different and have similar purchasing trends. This result is not surprising given the income level and purchase crosstabulation table.

- ✍ What action can you take at this point? If your analysis goal is building predictive models, you can write a DATA step to, in essence, collapse the Low and Medium observations into a single group. The new variable (**highinc**) would be equal to High when **income**=High, or Low/Medium otherwise. You would then replace **income** in the MODEL statement with **highinc**, and execute PROC LOGISTIC again. Remember to correctly interpret the coefficient of **highinc**.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gender Female vs Male	1.655	1.099	2.493
income High vs Low	2.139	1.307	3.502
income Medium vs Low	1.101	0.658	1.843

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	54.0	Somers' D	0.246
Percent Discordant	29.4	Gamma	0.295
Percent Tied	16.6	Tau-a	0.116
Pairs	43578	c	0.623

The last part of the output provides the Odds Ratio Estimates table as well as the Association of Predicted Probabilities and Observed Responses table.

The effects for **gender** Female vs Male and **income** High vs Low both indicate that they are statistically significant at the 0.05 level because their 95% Wald Confidence Intervals do not include 1.000. Notice that the 95% confidence interval for **income** Medium vs Low does not imply significance. The interval (0.658, 1.843) includes 1.000.

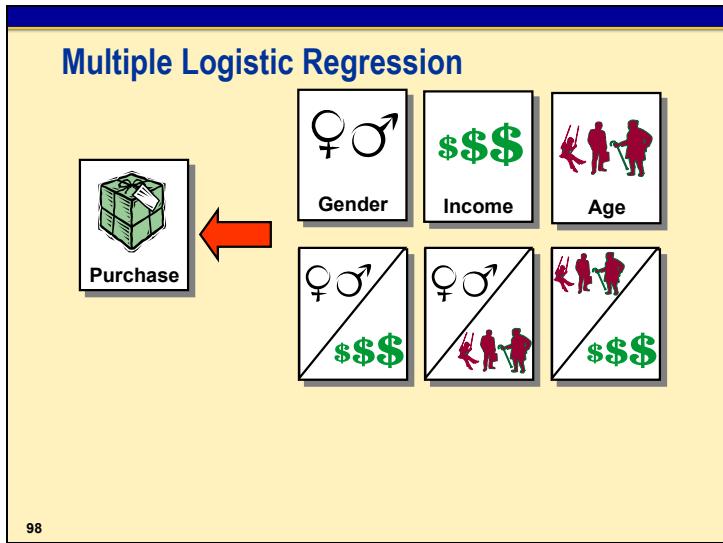
When you compare the percentages of this model with the previous model where **gender** was the only predictor variable, the concordant percentage increased (from 30.1 to 54.0), but the discordant percentage also increased (from 19.5 to 29.4). The tied percentage showed the most change, decreasing from 50.4 to 16.6.

The *c* statistic increased (0.553 to 0.623) from the simple **gender** model, which is desirable.

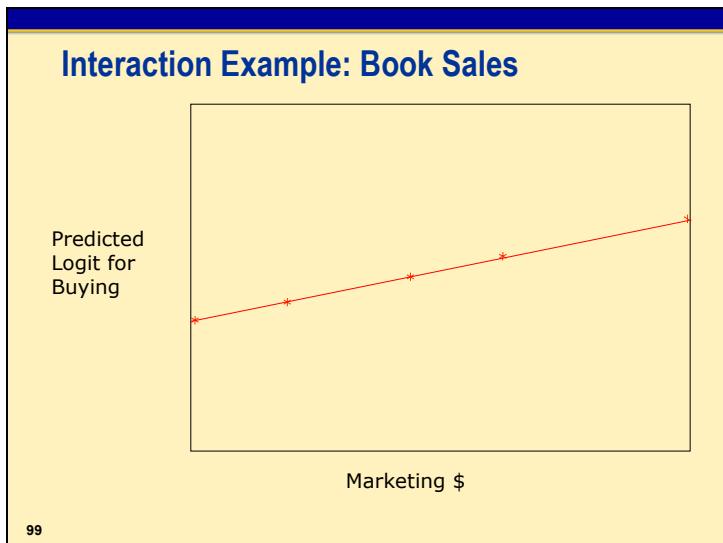
Comparing Models	
<b>Gender Only</b>	<b>Gender + Income</b>
<b>AIC</b>	569.951
<b>SC</b>	578.084
<b>-2 Log L</b>	565.951
<b>Conc.</b>	30.1%
<b>Disc.</b>	19.5%
<b>Ties</b>	50.4%
<b>c</b>	0.553
<b>AIC</b>	562.190
<b>SC</b>	578.454
<b>-2 Log L</b>	554.190
<b>Conc.</b>	54.0%
<b>Disc.</b>	29.4%
<b>Ties</b>	16.6%
<b>c</b>	0.623

97

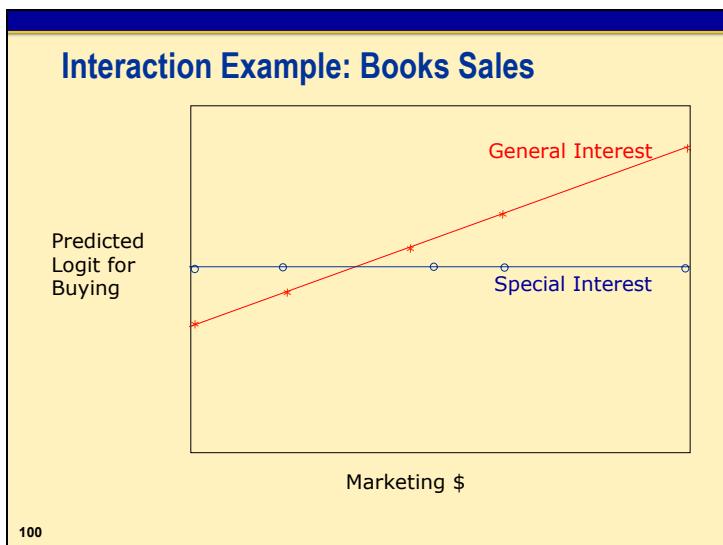
Adding **income** to the model decreased AIC and the percentage of tied pairs, and increased the percentage of concordant pairs and the *c* statistic. The SC increased slightly, and discordant pairs increased. Adding **income** improved the model.



In the last example, a multiple logistic regression model was fitted with only the main effects (just predictor variables are in the model). Thus, you are assuming that the effect of each variable on the outcome is the same regardless of the levels of the other variables. For example, you are assuming that the effect of **gender** (Female to Male) on the probability of making purchases of 100 dollars or more is the same regardless of **income** level. If this assumption is not correct, you might want to fit a more complex model that has interactions.



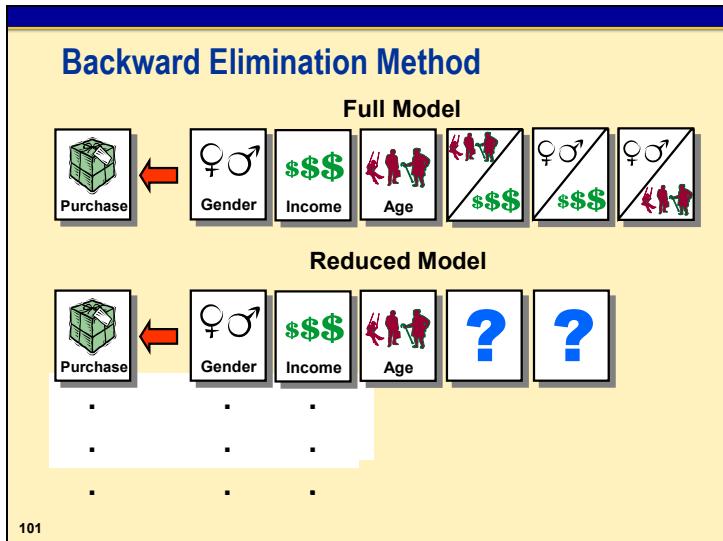
Assume that one dollar of marketing money has the same effect for all books.



However, if you consider the type of book to be sold, there seems to be a difference in the effect of marketing dollars on buying general interest books versus special interest. This is called an *interaction*. An interaction between two variables A and B is said to occur when the effect of A on the outcome depends on the observed level of B, or when the effect of B on the outcome depends on the observed level of A.

In the example above, the effect of **marketing** depends on the level of **booktype**. For **booktype**=General Interest, as **marketing** increases, the probability of buying increases. However, for **booktype**=Special Interest, as **marketing** increases, the probability of buying does not change.

Therefore, there is a **marketing** by **booktype** interaction.



When you use the backward elimination method with interactions in the model, PROC LOGISTIC begins by fitting the full model with all the main effects and interactions. PROC LOGISTIC then eliminates the nonsignificant interactions one at a time, starting with the least significant interaction (the one with the largest  $p$ -value). Next, PROC LOGISTIC eliminates the nonsignificant main effects not involved in any significant interactions. The final model should only have significant interactions, the main effects involved in the interactions, and the significant main effects.

For any effect that is in a model, all effects contained by that effect must also be in the model. This requirement is called *model hierarchy*. For example, if the interaction **gender\*income** is in the model, then the main effects **gender** and **income** must also be in the model. This ensures that you have a hierarchically well-formulated model.

- For a more customized analysis, the HIERARCHY= option specifies whether hierarchy is maintained and whether a single effect or multiple effects are allowed to enter or leave the model in one step for forward, backward, and stepwise selection. The default is HIERARCHY=SINGLE. You can change this option by inserting the HIERARCHY= option in the MODEL statement. See the SAS/STAT User's Guide in the SAS online documentation for more on using this option. In the LOGISTIC procedure, HIERARCHY=SINGLE is the default, meaning that SAS will not remove a main effect before first removing all interactions involving that main effect.



## Multiple Logistic Regression with Interactions

Example: Fit a multiple logistic regression model using the backward elimination method. In the MODEL statement, specify all the main effects and the two-factor interactions.

```
/* c5demo08 */
proc logistic data=sasuser.b_sales_inc;
  class gender (param=ref ref='Male')
    income (param=ref ref='Low');
  model purchase(event='1')=gender|age|income @2/ selection=backward;
  title1 'LOGISTIC MODEL (3): main effects and 2-way interactions';
  title2 '/ sel=backward';
run;
```



The bar notation with the @2 constructs a model with all the main effects and the two-factor interactions. If you increased it to @3, then you would construct a model with all of the main effects, the two-factor interactions, and the three-factor interaction. However, the three-factor interaction might be more difficult to interpret.

The Model Information, Response Profile, and Class Level Information tables have not changed.

```
LOGISTIC MODEL (3): main effects and 2-way interactions
/ sel=backward
```

#### The LOGISTIC Procedure

##### Model Information

Data Set	SASUSER.B_SALES_INC
Response Variable	purchase
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	431
Number of Observations Used	431

##### Response Profile

Ordered Value	purchase	Total Frequency
1	0	269
2	1	162

Probability modeled is purchase=1.

##### Backward Elimination Procedure

##### Class Level Information

Class	Value	Design Variables
gender	Female	1
	Male	0
income	High	1
	Low	0
	Medium	1

Step 0. The following effects were entered:

Intercept gender age age\*gender income gender\*income age\*income

##### Model Convergence Status

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

## PROC LOGISTIC Output (continued)

Model Fit Statistics					
Criterion	Intercept Only	Intercept and Covariates			
AIC	572.649	560.330			
SC	576.715	600.991			
-2 Log L	570.649	540.330			
Testing Global Null Hypothesis: BETA=0					
Test	Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio	30.3195	9	0.0004		
Score	28.9614	9	0.0007		
Wald	26.7755	9	0.0015		
Step 1. Effect age*income is removed:					
Model Convergence Status					
Convergence criterion (GCONV=1E-8) satisfied.					
Model Fit Statistics					
Criterion	Intercept Only	Intercept and Covariates			
AIC	572.649	557.936			
SC	576.715	590.465			
-2 Log L	570.649	541.936			
Testing Global Null Hypothesis: BETA=0					
Test	Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio	28.7135	7	0.0002		
Score	26.8148	7	0.0004		
Wald	24.7124	7	0.0009		
Residual Chi-Square Test					
Chi-Square	DF	Pr > ChiSq			
1.5966	2	0.4501			
Step 2. Effect age*gender is removed:					
Model Convergence Status					
Convergence criterion (GCONV=1E-8) satisfied.					

## PROC LOGISTIC Output (continued)

Model Fit Statistics			
Criterion	Intercept Only	Intercept and Covariates	
AIC	572.649	557.592	
SC	576.715	586.054	
-2 Log L	570.649	543.592	
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	27.0577	6	0.0001
Score	25.6386	6	0.0003
Wald	24.1104	6	0.0005
Residual Chi-Square Test			
Chi-Square	DF	Pr > ChiSq	
3.2232	3	0.3585	
Step 3. Effect age is removed:			
Model Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied.			
Model Fit Statistics			
Criterion	Intercept Only	Intercept and Covariates	
AIC	572.649	557.194	
SC	576.715	581.591	
-2 Log L	570.649	545.194	
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	25.4552	5	0.0001
Score	24.1139	5	0.0002
Wald	22.7265	5	0.0004
Residual Chi-Square Test			
Chi-Square	DF	Pr > ChiSq	
4.7980	4	0.3087	
NOTE: No (additional) effects met the 0.05 significance level for removal from the model.			

## PROC LOGISTIC Output (continued)

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	age*income	2	5	1.5891	0.4518
2	age*gender	1	4	1.6408	0.2002
3	age	1	3	1.5965	0.2064
Type 3 Analysis of Effects					
	Effect	DF	Chi-Square	Wald	Pr > ChiSq
	gender	1	4.9207	4.9207	0.0265
	income	2	18.8745	18.8745	<.0001
	gender*income	2	8.8363	8.8363	0.0121
Analysis of Maximum Likelihood Estimates					
Parameter		DF	Estimate	Standard Error	Wald Chi-Square
Intercept		1	-1.4759	0.3919	14.1841
gender	Female	1	0.9949	0.4485	4.9207
income	High	1	1.5026	0.4549	10.9113
income	Medium	1	0.1235	0.4873	0.0642
gender*income	Female High	1	-1.2223	0.5523	4.8979
gender*income	Female Medium	1	0.1026	0.5851	0.0307
Association of Predicted Probabilities and Observed Responses					
	Percent Concordant	54.8	Somers' D	0.261	
	Percent Discordant	28.6	Gamma	0.314	
	Percent Tied	16.6	Tau-a	0.123	
	Pairs	43578	c	0.631	

The interactions between **age\*income** and **age\*gender** were eliminated from the model because their *p*-values were greater than the default value of 0.05, as reported in the Summary of Backward Elimination table. However, because the interaction of **gender** and **income** was significant, the main effects **gender** and **income** must remain in the model. Because the main effect **age** was not involved in an interaction that was still in the model and it was not significant, the term was dropped from the model.

Comparing the goodness-of-fit statistics and the statistics that assess the predictive ability of the full model and the final model shows that the full model has better predictive ability (because of the higher  $c$  statistic) and the final model has better goodness-of-fit statistics (because of the lower AIC and SC statistics).

Statistic	Full Model <b>purchase=gender age income gender*age gender*income age*income</b>	Final Model <b>purchase=gender income gender*income</b>
AIC	560.330	557.194
SC	600.991	581.591
% Concordant	64.3	54.8
% Discordant	34.5	28.6
% Tied	1.1	16.6
$c$	0.649	0.631

## Comparing Models

Gender, Income Main Effects	
AIC	562.190
SC	578.454
-2 Log L	554.190
Conc.	54.0%
Disc.	29.4%
Ties	16.6%
c	0.623

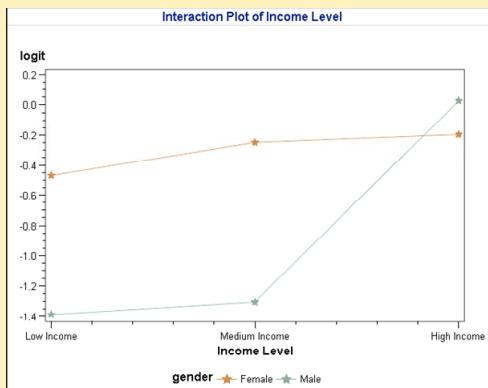
Main Effects + Interaction	
AIC	557.194
SC	581.591
-2 Log L	545.194
Conc.	54.8%
Disc.	28.6%
Ties	16.6%
c	0.631

104

AIC decreased (improved) for this model, but SC increased.

What is the purpose of your model? The “best” predictive model would include only the main effects based on the Schwarz criterion. However, using AIC, the best explanatory model would include the interaction term.

## Interaction Plot



105

To visualize the interaction between **gender** and **income**, you could produce an interaction plot. The plot would show two slopes for **income**, one for males and one for females. If there is no interaction between **gender** and **income**, then the slopes should be relatively parallel. However, the graph above shows that the slopes are not parallel. The reason for the interaction is that the probability of making purchases of 100 dollars or more is highly related to income for men but is weakly related to income for women.



The code for the interaction plot is shown in Appendix D, “Advanced Programs.”

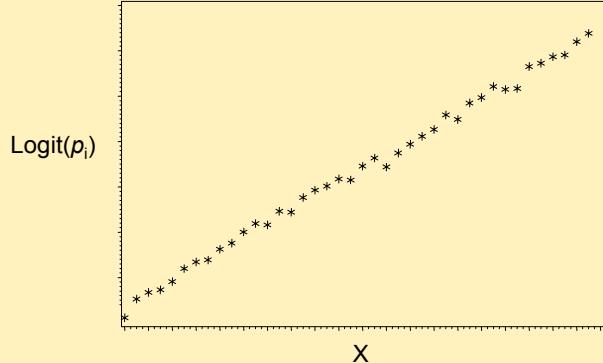
## 5.5 Logit Plots (Self-Study)

### Objectives

- Explain the concept of logit plots.
- Plot estimated logits for continuous and ordinal variables.

108

### Linear Logit Plot



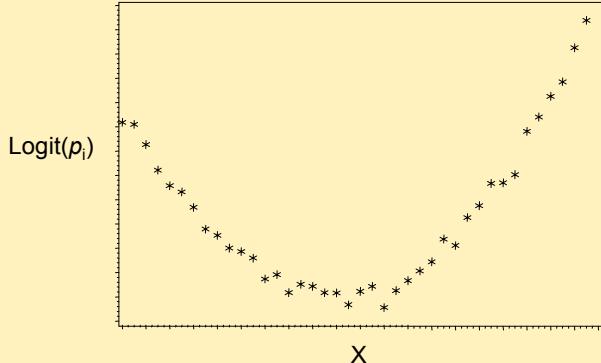
109

For continuous predictor variables with a large number of unique values, binning the data (collapsing data values into groups) is necessary to compute the logit. The bin size should have an adequate number of observations to reduce the sample variability of the logits. If the standard logistic regression model adequately fits the data, the logit plots should be fairly linear. The above graph shows a predictor variable that meets the assumption of linearity in the logit.



If the predictor variable is a nominal variable, then there is no need to create a logit plot.

### Quadratic Logit Plot



110

The logit plot can also show serious nonlinearities between the outcome variable and the predictor variable. The above graph reveals a quadratic relationship between the outcome and predictor variables. Adding a polynomial term or binning the predictor variable into three groups (two dummy variables would model the quadratic relationship) and treating it as a classification variable can improve the model fit.

### Estimated Logits

$$\ln\left(\frac{m_i + 1}{M_i - m_i + 1}\right)$$

where

$m_i$  = number of events

$M_i$  = number of cases

111

A common approach in computing logits is to take the log of the odds. The logit is undefined, however, for any bin in which the outcome rate is 100% or 0%. To eliminate this problem and reduce the variability of the logits, a common recommendation is to add a small constant to the numerator and denominator of the formula that computes the logit (Santner and Duffy 1989).



## Plotting Estimated Logits

Example: Plot the estimated logits of the outcome variable **purchase** versus the predictor variable **inclevel**. To construct the estimated logits, the number of customers who spend 100 dollars or more and the total number of customers by each level of **inclevel** must be computed.

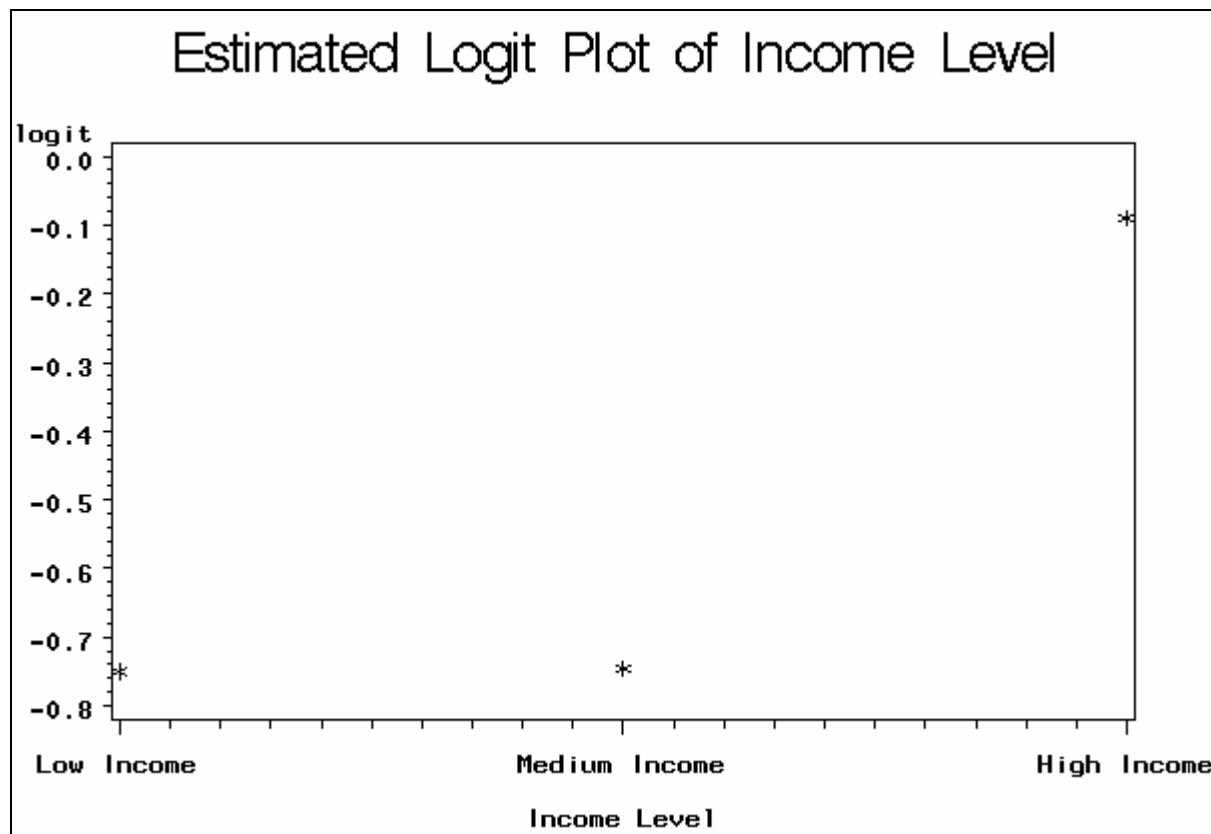
```
/* c5demo09 */
proc means data=sasuser.b_sales_inc noplay nway;
  class inclevel;
  var purchase;
  output out=bins sum(purchase)=purchase;
run;

data bins;
  set bins;
  logit=log((purchase+1)/(_freq_-purchase+1));
run;

proc gplot data=bins;
  plot logit*inclevel;
  symbol v=star i=none;
  format inclevel incfmt.;
  label inclevel='Income Level';
  title 'Estimated Logit Plot of Income Level';
run;
quit;
```

Selected PROC MEANS statement option:

NWAY causes the output data set to have only one observation for each level of the class variable.



The logit plot for this ordinal variable is not linear. The variable **inclevel** should be entered into the model as a CLASS variable. In addition, the graph indicates that low- and medium-income customers have approximately the same probability of spending 100 dollars or more. A possible recommendation is to combine the low- and medium-income groups into one group and make **income** a binary variable (high versus all other) in the model.

- If a linear pattern is detected in a logit plot, the ordinal variable should be removed from the CLASS statement, implying that it would be a considered continuous variable.

Example: Plot the estimated logits of the outcome variable **purchase** versus the predictor variable **age**. Because age is a continuous variable, bin the observations into 10 groups to ensure that an adequate number of observations is used to compute the estimated logit.

```
/* c5demo10 */
proc rank data=sasuser.b_sales_inc groups=10 out=ranks10;
  var age;
  ranks bin10;
run;

proc means data=ranks10 noprint nway;
  class bin10;
  var purchase age;
  output out=bins10 sum(purchase)=purchase mean(age)=age;
run;

data bins10;
  set bins10;
  logit=log((purchase+1)/(_freq_-purchase+1));
run;

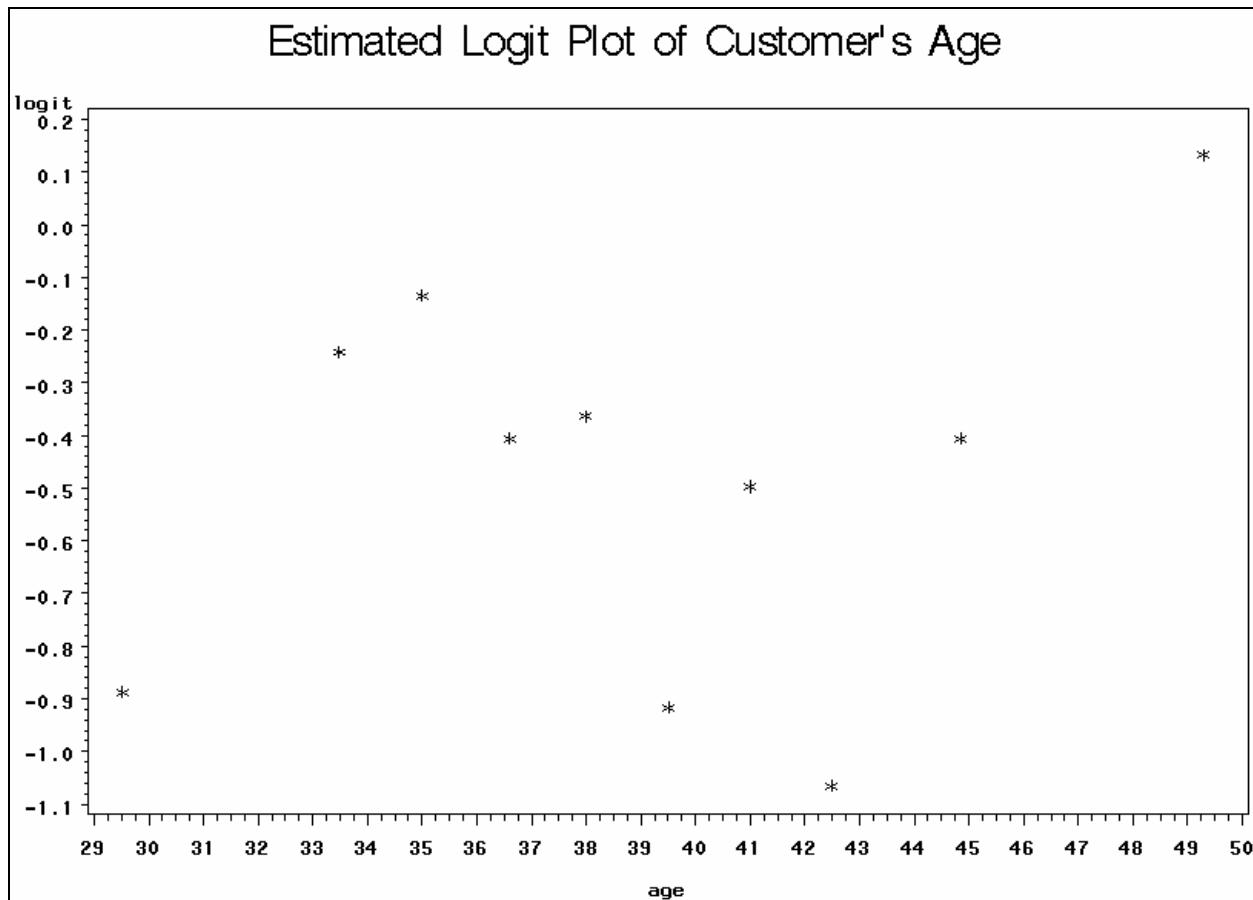
proc gplot data=bins10;
  plot logit*age;
  symbol v=star i=none;
  title "Estimated Logit Plot of Customer's Age";
run;
quit;
```

Selected PROC RANK statement option:

GROUPS=*n* bins the variables into *n* groups.

Selected RANK procedure statement:

RANKS names the group indicators in the OUT= data set. If the RANKS statement is omitted, then the group indicators replace the VAR variables in the OUT= data set.



The estimated logit plot shows no apparent pattern. Therefore, **age** can be entered into the model as a continuous variable because creating several groups will probably not improve the fit of the model. Although it seems that **age** is not an important predictor for **purchase**, the estimated logit plot is a univariate plot that can be misleading in the presence of interactions and partial associations (association between the response variable and the predictor variable changes with the addition of another predictor variable in the model). A model with two-factor interactions and main effects should be evaluated before **age** is eliminated. Estimated logit plots should never be used to eliminate variables.

## 5.6 Chapter Summary

Categorical data analysis deals with the analysis of categorical response variables, regardless of whether the explanatory variables are categorical or continuous. The scale of measurement of the variables is an important consideration when you decide the appropriate statistic to use. When you have two nominal variables, the Pearson chi-square statistic is appropriate. The strength of the association can be measured by Cramer's V. Because the Pearson chi-square statistic requires a large sample size, Fisher's exact test should be used to detect an association when you have a small sample size. When you have two ordinal variables, the Mantel-Haenszel chi-square statistic should be used to detect an ordinal association. The strength of the association can be measured by the Spearman correlation statistic.

Logistic regression uses the explanatory variables, which can be categorical or continuous, to predict the probability that the outcome or response variable takes on a given value. In other words, logistic regression is designed to describe probabilities associated with the values of the outcome variable. Probabilities are bounded by 0 and 1, so a linear model cannot be used because linear functions are inherently unbounded. The solution to this problem is to transform the probabilities to logits, which are unbounded, so that a linear regression model can be used because the logits are linear in the parameters.

The output from logistic regression shows the odds ratio, which is a measure of association between the explanatory variable and the outcome variable. The odds ratio compares the odds of an event in one group to the odds of an event in another group. The odds of an event is the ratio of the expected number of times that an event will occur to the expected number of times it will not occur.

The output also shows Akaike's 'A' Information Criteria (AIC) and Schwarz's Bayesian Criterion (SC), which are goodness-of-fit measures that adjust for the number of explanatory variables in the model. Lower values indicate a more desirable model. There are also four rank correlation indices that are computed from the numbers of concordant and discordant pairs of observations. In general, a model with higher values for these indices has better predictive ability than a model with lower values for these indices.

One model-building strategy is to assess all the two-factor interactions first and then assess the main effects. An interaction occurs when the effect of one variable on the outcome depends on the observed level of another variable. When a model has no interactions you are assuming that the effect of each variable on the outcome is the same regardless of the levels of the other variables. A common variable selection technique is the backward elimination method. One strategy is to eliminate the nonsignificant interactions one at a time, starting with the least significant interaction. Then eliminate the nonsignificant main effects not involved in any significant interactions.

```
PROC FREQ DATA=SAS-data-set;
  TABLES table-requests </ options>;
  EXACT statistic-keywords;
RUN;
```

```
PROC LOGISTIC DATA=SAS-data-set <options>;
  CLASS predictor-variables </ options>;
  MODEL response=predictor-variables
    </ options>;
  OUTPUT OUT= SAS-data-set </ options>;
RUN;
```

# Appendix A Exercises and Solutions

<b>Exercises .....</b>	<b>A-2</b>
Chapter 1.....	A-2
Chapter 2.....	A-4
Chapter 3.....	A-5
Chapter 4.....	A-8
Chapter 5.....	A-9
 <b>Solutions to Exercises .....</b>	 <b>A-11</b>
Chapter 1.....	A-11
Chapter 2.....	A-16
Chapter 3.....	A-36
Chapter 4.....	A-66
Chapter 5.....	A-77

## Exercises

### Chapter 1

#### 1. Producing Descriptive Statistics

A random sample of 50 observations pertaining to 50 male runners in the 1997 Boston Marathon was obtained. The data is in the data set **sasuser.b\_boston**. The data pertaining to the top 87 males who were in the top 100 (men and women) is in the data set **sasuser.b\_top100**. Both data sets have the following variables:

**age** runner's age in years

**tottime** total time in seconds it took the runner to complete the course

**halftime** time it took in seconds to complete the first half of the distance

- a. What are the minimum, the maximum, the mean, and the standard deviation for each of the variables in the data set **sasuser.b\_boston**? Do the variables appear to be normally distributed?

	<b>age</b>	<b>tottime</b>	<b>halftime</b>
Minimum			
Maximum			
Mean			
Standard Deviation			
Skewness			
Kurtosis			
Distribution: Normal	Yes/No	Yes/No	Yes/No

- b. What are the minimum, the maximum, the mean, and the standard deviation for each of the variables in the data set **sasuser.b\_top100**? Do the variables appear normally distributed?

	<b>age</b>	<b>tottime</b>	<b>halftime</b>
Minimum			
Maximum			
Mean			
Standard Deviation			
Skewness			
Kurtosis			
Distribution: Normal	Yes/No	Yes/No	Yes/No

## 2. Producing Confidence Intervals

- a. Generate the 95% confidence interval for the total time it takes for participants in the data set **sasuser.b\_boston** to complete the marathon.
- 1) Is it appropriate to obtain a confidence interval for this data?
  - 2) What is the confidence interval?
  - 3) How do you interpret this interval?

## 3. Performing a One-Sample *t*-Test

- a. Perform a one-sample *t*-test to determine whether the mean of the random sample of participants in the data set **sasuser.b\_boston** is significantly different from the average time it took for the top 87 male participants, 8891.37 seconds.
- 1) What is the value of the *t* statistic and the corresponding *p*-value?
  - 2) Do you reject or fail to reject the null hypothesis at the .05 level that the average time for the participants is 8891.37 seconds?
  - 3) Are the assumptions of the one-sample *t*-test validated in this example?

## Chapter 2

### 1. Analyzing Data in a Completely Randomized Design

Consider an experiment to study four types of advertising: local newspaper ads, local radio ads, in-store salespeople, and in-store displays. The country is divided into 144 locations, and 36 locations are randomly assigned to each type of advertising. The level of sales is measured for each region in thousands of dollars. You want to see whether the average sales are significantly different for various types of advertising. The data set **sasuser.b\_ads** contains data for these variables:

**ad** type of advertising

**sales** level of sales in thousands of dollars

- a. Examine the data using the UNIVARIATE and BOXPLOT procedures. What information can you obtain from looking at the data?
- b. Test the hypothesis that the means are equal. Be sure to check that the assumptions of the analysis method you choose are met. What conclusions can you reach at this point in your analysis?
- c. Conduct pairwise comparisons with an experimentwise error rate of  $\alpha=0.05$ . Which types of advertising are significantly different?

### 2. Analyzing Data in a Randomized Block Design

When you design the advertising experiment in the first question, you are concerned that there is variability caused by area of the country. You are not particularly interested in what differences are caused by **area**, but you are interested in isolating the variability due to this factor. The data set **sasuser.b\_ads1** contains data for these variables:

**ad** type of advertising

**area** area of the country

**sales** level of sales in thousands of dollars

- a. Test the hypothesis that the means are equal. Include all of the variables in your MODEL statement. What can you conclude from your analysis? Was adding the blocking factor **area** into the model detrimental to the analysis?
- b. Conduct pairwise comparisons with an experimentwise error rate of  $\alpha=0.05$ . Which types of advertising are significantly different?

### 3. Performing Two-Way ANOVA

Consider an experiment to test three different brands of cement and whether an additive makes the cement stronger. Thirty test plots are poured and the following features are recorded in the data set **sasuser.b\_cement**:

<b>strength</b>	the measured strength of a cement test plot
<b>additive</b>	whether an additive was used in the test plot
<b>brand</b>	the brand of cement being tested

- a. Examine the data using the MEANS, BOXPLOT, and GPLOT procedures. What information can you obtain from looking at the data?
- b. Test the hypothesis that the means are equal. What conclusions can you reach at this point in your analysis?
- c. Do the appropriate multiple comparisons test for statistically significant effects?

## Chapter 3

### 1. Describing the Relationship between Two Continuous Variables

The cost of tuition and graduation rates are recorded for the top 200 private and public colleges selected by *Money* magazine in 1991. Data is stored in the data set **sasuser.b\_colleg**. The data set contains information for these variables:

<b>name</b>	name of the college or university
<b>rate</b>	graduation rate, excluding transfer students
<b>region</b>	school's geographical region
<b>state</b>	state where the college or university is located
<b>tuition</b>	tuition rate
<b>type</b>	type of school, either private or public

- a. Use the UNIVARIATE procedure to examine the distribution of the variables **rate** and **tuition**.
  - 1) What conclusions can you draw about the distribution of these variables?
  - 2) Do there appear to be any unusual observations?
- b. Generate a scatter plot for the variables **rate** versus **tuition**.
  - 1) Can a straight line adequately describe the data?
  - 2) Are there any outliers you should investigate?

c. Generate a correlation coefficient for the variables **rate** and **tuition**.

- 1) What is the correlation coefficient for **rate** and **tuition**?
- 2) How would you interpret this coefficient?
- 3) What is the *p*-value for the coefficient?
- 4) Is it statistically significant at the 0.05 level?

## 2. Fitting a Simple Linear Regression

A college entrance exam is designed to predict freshman-year grade point averages. Twenty-five students take the exam, and the data is stored in a SAS data set named **sasuser.b\_grades**.

The variables in the data set are as follows:

**score** student's exam score

**gpa** grade point average at the end of the freshman year

a. Generate a scatter plot for the variables **gpa** versus **score**.

- 1) Can a straight line adequately describe the data?
- 2) What is the range of **score**?
- 3) Are there any outliers or influential observations you should investigate?

b. Perform a regression analysis by specifying **gpa** as the response variable and **score** as the predictor variable.

- 1) What is the value of the *F* statistic and the associated *p*-value? How would you interpret this with regards to the model?
- 2) Write out the predicted regression equation. How would you interpret this?
- 3) What is the value of the  $R^2$  statistic? How would you interpret this?
- 4) What is the parameter estimate for **score**? What is the interpretation of the estimate?

c. Produce predicted values for **gpa** when **score** is 40, 60, and 80.

- 1) What are the predicted values?
- 2) Is it appropriate to predict **gpa** when **score** is 200?

d. Produce confidence and prediction intervals around these predictions.

- 1) What is the 95% confidence interval for the predicted mean of **gpa** when **score** is 60? How would you interpret this?
- 2) What is the 95% prediction interval for the predicted value of **gpa** when **score** is 60? How would you interpret this?

### 3. Performing a Regression Using the REG Procedure

Using the **sasuser.b\_fitness** data set, run a regression of **Oxygen\_Consumption** on the variables **Performance**, **Runtime**, **Age**, **Weight**, **Run\_Pulse**, **Rest\_Pulse**, and **Maximum\_Pulse**.

- a. Compare the ANOVA table with the **Oxygen\_Consumption** and **Performance** regression ANOVA table in the demonstration. What is different?
- b. How do the  $R^2$  and the adjusted  $R^2$  compare with these statistics for the **Oxygen\_Consumption** and **Performance** regression demonstration?
- c. Did the estimate for the intercept change? Did the estimate for the slope of **Performance** change?

### 4. Simplifying the Model

- a. Rerun the model in Exercise 3, but eliminate the variable with the highest  $p$ -value. Compare the output with the Exercise 3 model.
- b. Did the  $p$ -value for the model change?
- c. Did the  $R^2$  and adjusted  $R^2$  change?
- d. Did the parameter estimates and their  $p$ -values change?

### 5. More Simplifying of the Model

- a. Rerun the model in Exercise 4, but drop the variable with the highest  $p$ -value.
- b. How did the output change from the previous model?
- c. Did the number of parameters with a  $p$ -value less than 0.05 change?

### 6. Using All-Regression Techniques

The data set **sasuser.b\_cars** contains information about the median price (**MidPrice**) of 92 different makes and models of cars. The data set also contains data about the cars' miles per gallon, city and highway (**CityMPG** and **HighwayMPG**), engine size and other characteristics (**Enginesize**, **HorsePower**, **RPM**, **Revolutions**), fuel tank capacity (**FuelTankSize**), and weight (**Weight**).

- a. Use an all-regressions technique to identify a set of candidate models using the CMALLWS and CHOCKING options that predict **MidPrice** as a function of the other variables.
- b. Use a stepwise regression method to select a candidate model; try STEPWISE and BACKWARD.
- c. Compare the selected candidate models and use the two different approaches.

## Chapter 4

### 1. Examining Residuals

- a. A college entrance exam is designed to predict freshman-year grade point averages. Twenty-five students take the exam, and the data is stored in the **sasuser.b\_grades** data set. Run a regression of **gpa** on **score**. Create residual plots of the residuals by **score** and by the predicted values, a plot of student residuals by observation number, and a normal quantile-quantile plot.
- 1) Do the residual plots indicate any problems with the model assumptions?
  - 2) Are there any outliers indicated by the studentized residuals?
  - 3) Does the quantile-quantile plot indicate any problems with the normality assumption?

### 2. Generating Potential Outliers

- a. Using the **sasuser.b\_cars** data set, run a regression of **Midprice** on **CityMPG**, **EngineSize**, **Horsepower**, and **Revolutions**.
- 1) Using the same model as above, create an output data set with the **RSTUDENT**, **COOKD**, and **DFFITS** diagnostic statistics. Use these statistics to identify potential influential observations based on the suggested cutoff values. (Hint: The data set has 92 observations. You might want to print only those that exceed the cutoff values.)

### 3. Ascertaining Collinearity

- a. Using the **sasuser.b\_cars** data set, run a regression of **Midprice** on all the other variables in the file.
- 1) Determine whether there is a collinearity problem.
  - 2) If so, identify the sets of Xs that are collinear with each other, and eliminate one term from the model. Reassess the need to continue this process.

## Chapter 5

### 1. Performing Tests and Measures of Association

An insurance company wants to relate the safety of vehicles to several other variables. A score has been given to each vehicle model, using the frequency of insurance claims as a basis. The data is in the **sasuser.b\_safety** data set.

The variables in the data set are as follows:

**safety** safety score (1=Below Average, 0=Average or Above)

**type** type of vehicle (Sports, Small, Medium, Large, and Sport/Utility)

**region** manufacturing region (Asia, N America)

**weight** weight of the vehicle in thousands of pounds

- a. Examine the **sasuser.b\_safety** data set using the PRINT procedure. Invoke the FREQ procedure and create one-way frequency tables for the variables **safety**, **type**, and **region**.

- 1) What is the measurement scale of each variable?

<u>Variable</u>	<u>Measurement Scale</u>
<b>safety</b>	_____
<b>type</b>	_____
<b>region</b>	_____
<b>weight</b>	_____

- 2) What is the proportion of cars made in North America?
  - 3) For the variables **safety**, **type**, and **region**, are there any unusual data values that warrant further investigation?
- b. Use PROC FREQ to examine the crosstabulation of the variables **region** by **safety**. Generate a temporary format to clearly identify the values of **safety**. Along with the default output, generate the expected frequencies and the cell chi-square.
- 1) For the cars made in Asia, what percentage had a below-average safety score?
  - 2) For the cars with an average or above safety score, what percentage was made in North America?
  - 3) Do you see any association between **region** and **safety**?
  - 4) What cell contributed the most to any possible association?
- c. Perform a chi-square test of association between **region** and **safety**.
- 1) Interpret the *p*-value from the test with respect to probability.
  - 2) Do you reject or fail to reject the null hypothesis at the 0.05 level?

- d. Create a new variable named **size**. Assign a 1 for **type** equal to Small or Sports, 2 for **type** equal to Medium, and 3 for **type** equal to Large or Sport/Utility. Examine the ordinal association between **size** and **safety** using PROC FREQ.
- 1) What statistic should you use to detect an ordinal association between **size** and **safety**?
  - 2) Do you reject or fail to reject the null hypothesis at the 0.05 level?
  - 3) What is the strength of the ordinal association between **size** and **safety**?
  - 4) What is the 95% confidence interval around that statistic?

## 2. Performing a Logistic Regression Analysis

- a. Fit a simple logistic regression model with **safety** as the outcome variable and **weight** as the predictor variable. Use the EVENT= option to model the probability of below-average safety scores.
- 1) Do you reject or fail to reject the null hypothesis that all regression coefficients of the model are 0?
  - 2) Write out the logistic regression equation.
  - 3) Interpret the odds ratio for **weight**.
  - 4) Interpret the 95% confidence interval for the odds ratio.
  - 5) Interpret the percentage of concordant observations.
- b. Fit a multiple logistic regression model with **safety** as the outcome variable and **weight** and **region** as the predictor variables. Use the EVENT= option to model the probability of below-average safety scores. Specify **region** as a classification variable using reference cell coding and specify Asia as the reference level. Also request the 95% profile likelihood confidence intervals.
- 1) Interpret the parameter estimate for **region**.
  - 2) Do you think this is a better model than the one fit with just **weight**?
  - 3) Why are the profile likelihood confidence intervals different than the Wald confidence intervals?
  - 4) Interpret the *c* statistic.

# Solutions to Exercises

## Chapter 1

### 1. Producing Descriptive Statistics

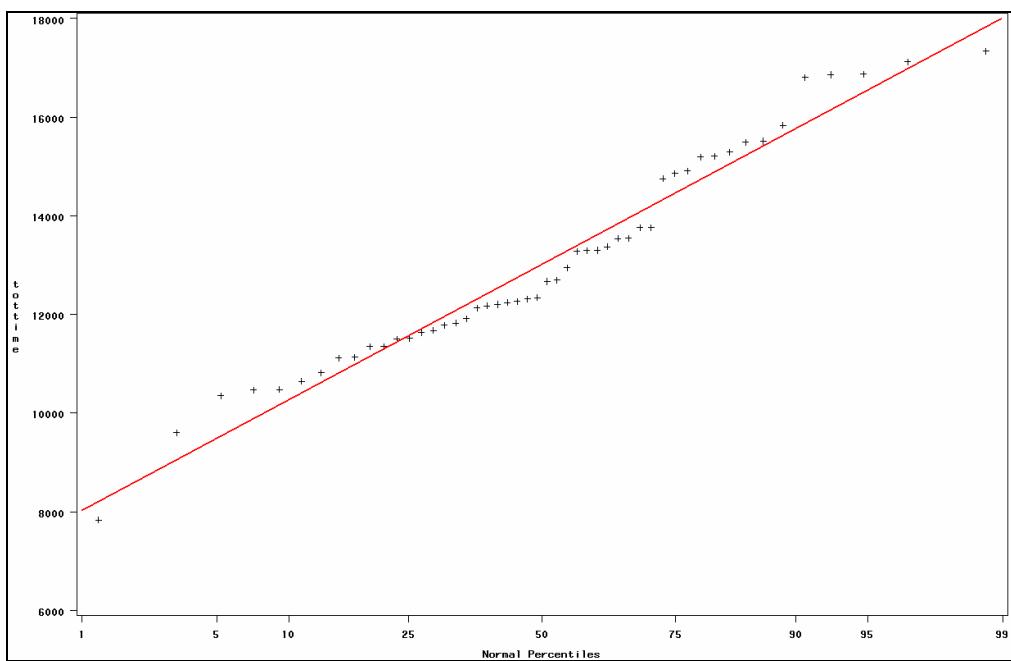
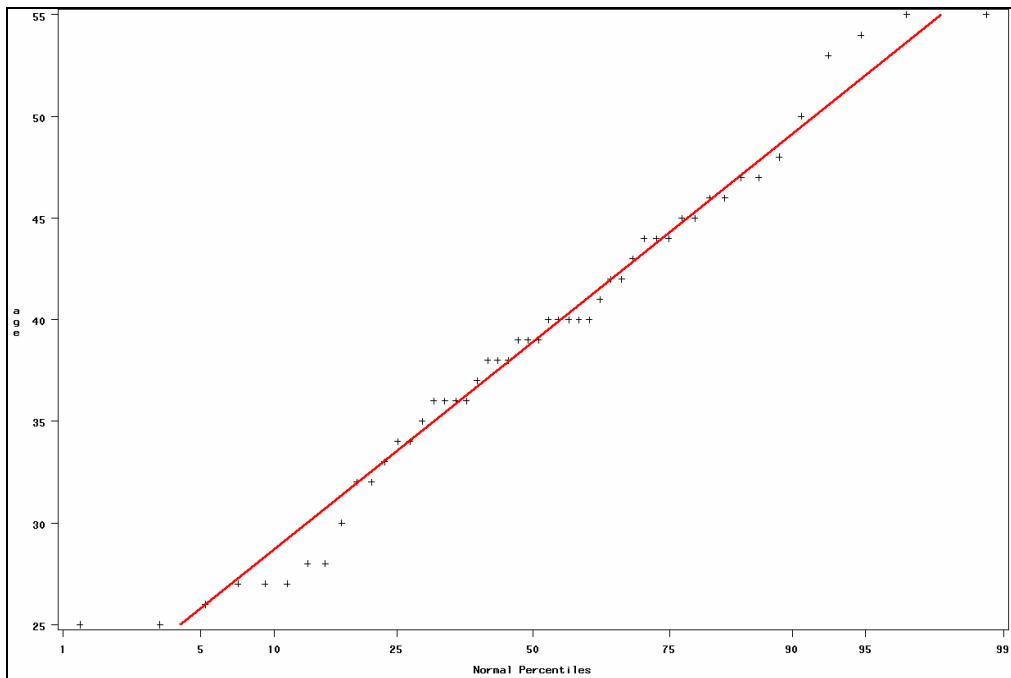
- a. Use the UNIVARIATE procedure to produce descriptive statistics.

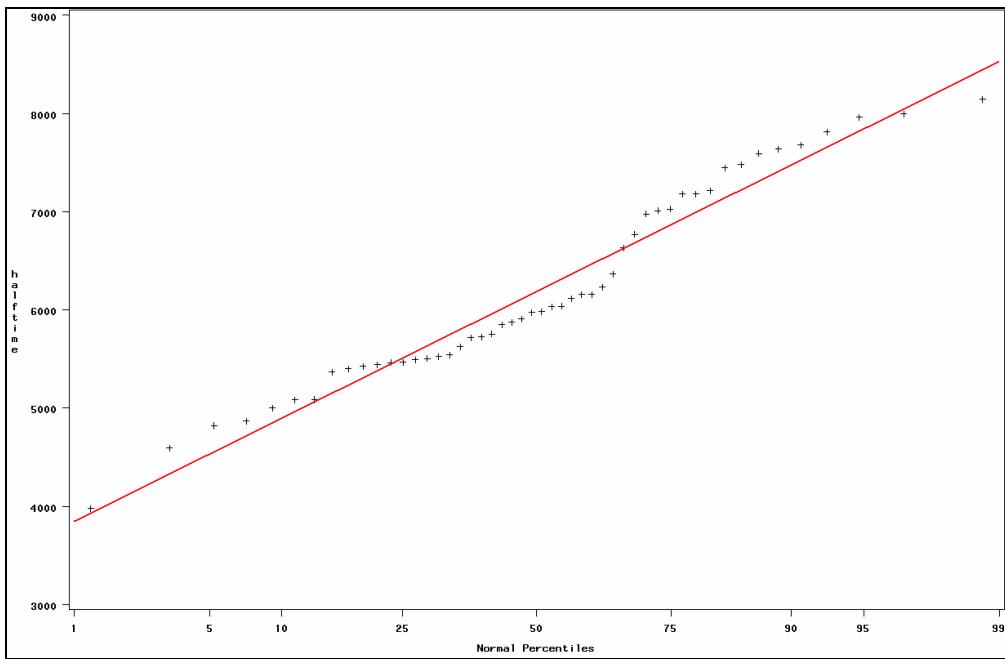
```
proc univariate data=sasuser.b_boston;
  var age tottime halftime;
  probplot age tottime halftime /
    normal (mu=est sigma=est color=red w=2);
run;
```

A summary of the output from PROC UNIVARIATE is shown below.

	<b>age</b>	<b>tottime</b>	<b>halftime</b>
Minimum	25.00	7834.00	3976.00
Maximum	55.00	17340.00	8146.00
Mean	38.92	13018.98	6187.00
Standard Deviation	7.97	2143.90	1006.07
Skewness	0.09	0.27	0.28
Kurtosis	-0.49	-0.31	-0.70
Distribution: Normal	Yes	Yes	Yes

An examination of high-resolution normal probability plots, combined with an interpretation of the skewness and kurtosis statistics, leads to the conclusion that all three variables are normally distributed. The normal probability plots have been included below in the following order: **age**, **tottime**, and **halftime**.





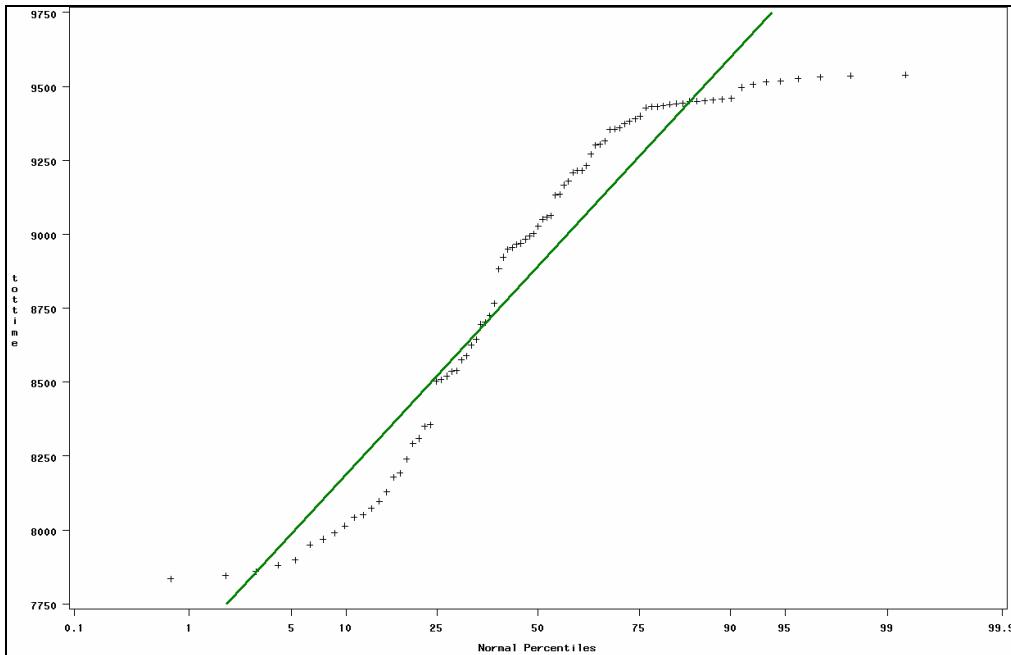
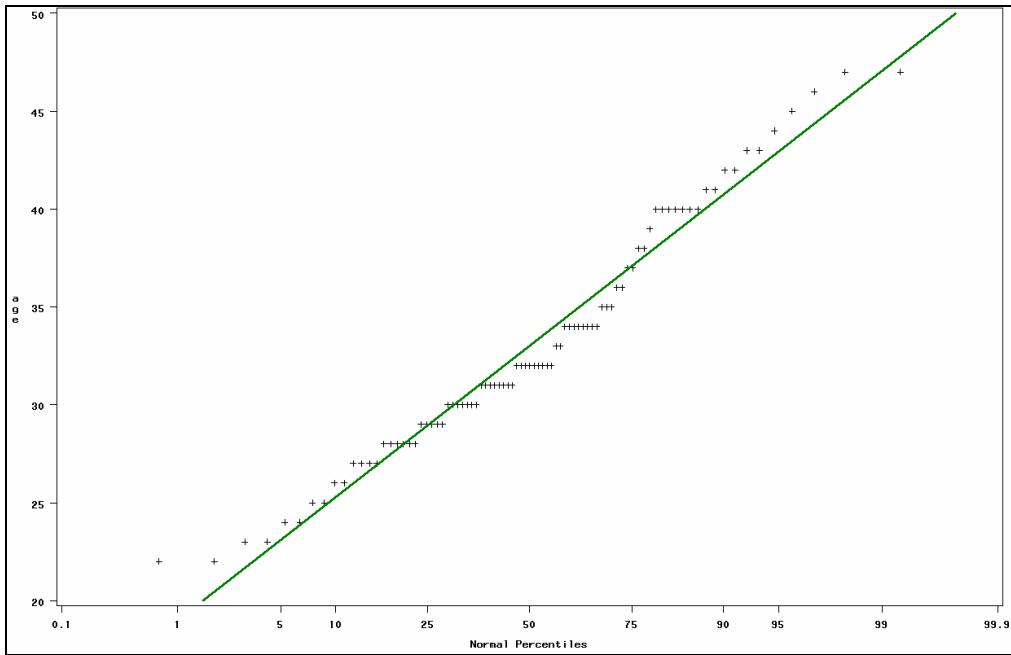
- b. Use PROC UNIVARIATE to produce descriptive statistics.

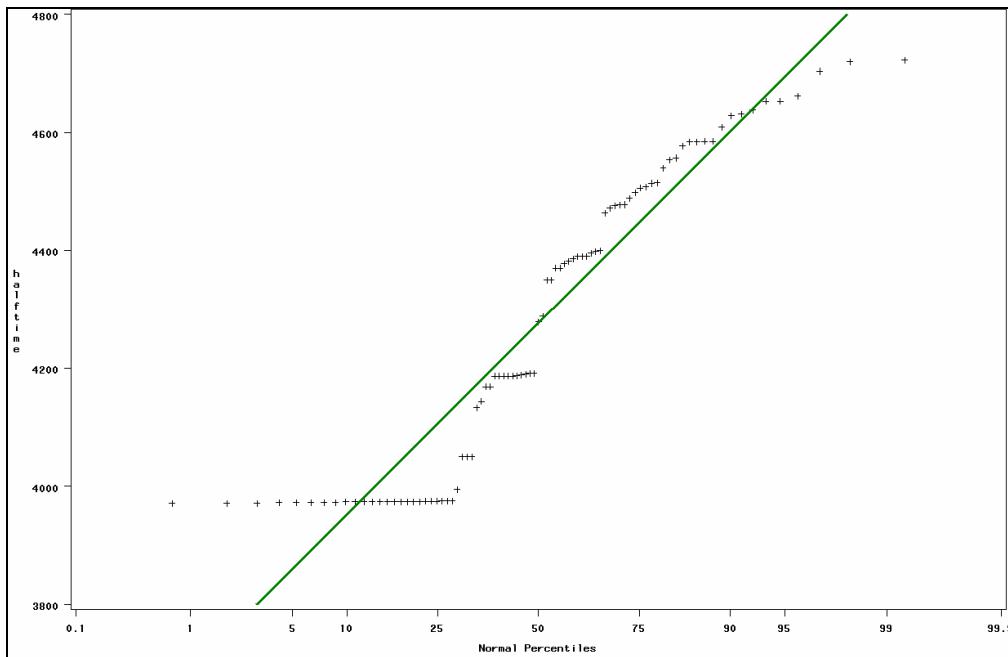
```
proc univariate data=sasuser.b_top100;
  var age tottime halftime;
  probplot age tottime halftime /
    normal (mu=est sigma=est color=green w=2);
run;
```

A summary of the PROC UNIVARIATE output is shown below.

	<b>age</b>	<b>tottime</b>	<b>halftime</b>
Minimum	22.00	7834.00	3971.00
Maximum	47.00	9539.00	4723.00
Mean	33.01	8891.38	4276.95
Standard Deviation	6.03	550.14	253.16
Skewness	0.44	-0.55	0.11
Kurtosis	-0.38	-1.09	-1.45
Distribution: Normal	Yes	No	No

A close examination of the high-resolution normal probability plots and an interpretation of the skewness and kurtosis statistics leads to the conclusion that **age** appears normally distributed, but **tottime** and **halftime** do not appear to be normal. The normal probability plots have been included below in the following order: **age**, **tottime**, and **halftime**.





## 2. Producing Confidence Intervals

- a. Use the MEANS procedure to produce confidence intervals.

```
proc means data=sasuser.b_boston n mean stderr clm;
  var tottime;
  run;
```

The MEANS Procedure				
Analysis Variable : tottime				
N	Mean	Std Error	Lower 95%	Upper 95%
50	13018.98	303.1929748	12409.69	13628.27

- 1) Because the sample size is large enough, the central limit theorem is invoked to validate the assumption of normality of the sample mean.
- 2) The 95% confidence interval for the total time it takes for participants in the data set **sasuser.b\_boston** to complete the marathon is between 12409.69 and 13628.27 seconds.
- 3) You have 95% confidence that the above interval includes the true mean total time it takes for participants to complete the marathon.

### 3. Performing a One-Sample *t*-Test

- a. Use PROC UNIVARIATE to perform a one-sample *t*-test.

```
proc univariate data=sasuser.b_boston mu0=8891.37;
  var tottime;
run;
```

Partial PROC UNIVARIATE Output

```
Tests for Location: Mu0=8891.4
```

Test	-Statistic-	-----	p Value-----
Student's t	t 13.6138	Pr >  t	<.0001
Sign	M 24	Pr >=  M	<.0001
Signed Rank	S 635.5	Pr >=  S	<.0001

- 1) The *t* statistic is 13.6138 and the corresponding *p*-value is less than 0.0001.
- 2) You reject the null hypothesis that the average time for the participants is 8897.13 seconds.
- 3) Because the values of **tottime** are normally distributed, the assumptions of the one-sample *t*-test are validated.

## Chapter 2

### 1. Analyzing Data in a Completely Randomized Design

- a. For each type of advertising, use PROC UNIVARIATE with the CLASS statement to generate descriptive statistics, histograms and normal probability plots.



The following descriptive statistics and their associated graphs have been formatted for ease of use.

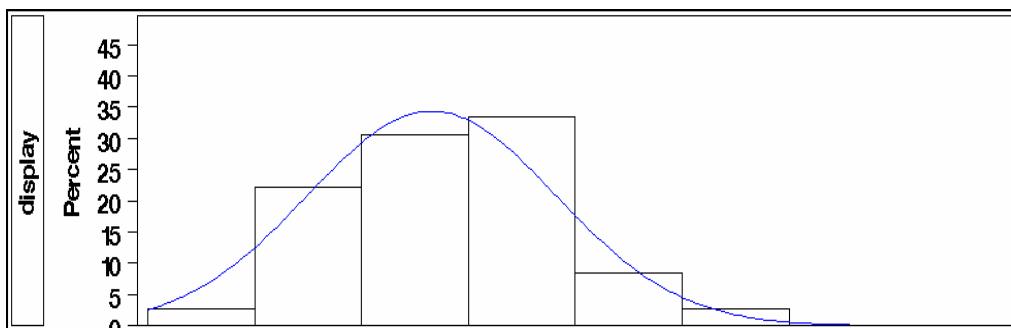
```
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=sasuser.b_ads;
  class ad;
  var sales;
  histogram / normal;
  probplot / normal (mu=est sigma=est color=red w=2);
run;
```

## Partial PROC UNIVARIATE Output for In-Store Displays

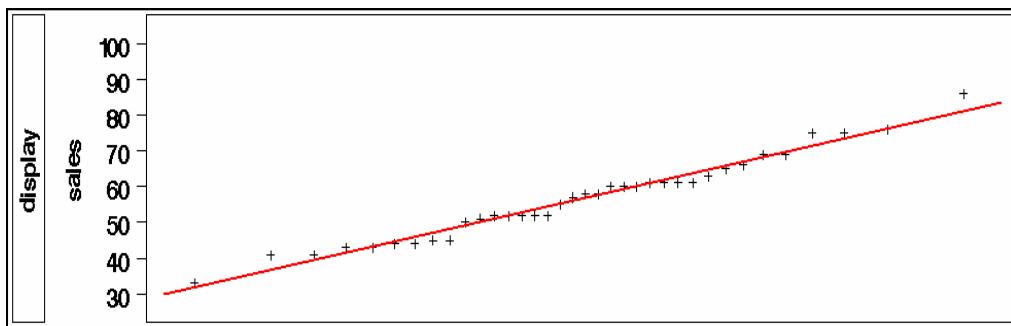
The UNIVARIATE Procedure																											
Variable: sales																											
ad = display																											
Moments																											
<table border="1" style="width:100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">N</td><td style="padding: 2px; text-align: right;">36</td><td style="padding: 2px;">Sum Weights</td><td style="padding: 2px; text-align: right;">36</td></tr> <tr> <td style="padding: 2px;">Mean</td><td style="padding: 2px; text-align: right;">56.5555556</td><td style="padding: 2px;">Sum Observations</td><td style="padding: 2px; text-align: right;">2036</td></tr> <tr> <td style="padding: 2px;">Std Deviation</td><td style="padding: 2px; text-align: right;">11.6188134</td><td style="padding: 2px;">Variance</td><td style="padding: 2px; text-align: right;">134.996825</td></tr> <tr> <td style="padding: 2px;">Skewness</td><td style="padding: 2px; text-align: right;">0.34564696</td><td style="padding: 2px;">Kurtosis</td><td style="padding: 2px; text-align: right;">0.0256814</td></tr> <tr> <td style="padding: 2px;">Uncorrected SS</td><td style="padding: 2px; text-align: right;">119872</td><td style="padding: 2px;">Corrected SS</td><td style="padding: 2px; text-align: right;">4724.88889</td></tr> <tr> <td style="padding: 2px;">Coeff Variation</td><td style="padding: 2px; text-align: right;">20.5440709</td><td style="padding: 2px;">Std Error Mean</td><td style="padding: 2px; text-align: right;">1.9364689</td></tr> </table>				N	36	Sum Weights	36	Mean	56.5555556	Sum Observations	2036	Std Deviation	11.6188134	Variance	134.996825	Skewness	0.34564696	Kurtosis	0.0256814	Uncorrected SS	119872	Corrected SS	4724.88889	Coeff Variation	20.5440709	Std Error Mean	1.9364689
N	36	Sum Weights	36																								
Mean	56.5555556	Sum Observations	2036																								
Std Deviation	11.6188134	Variance	134.996825																								
Skewness	0.34564696	Kurtosis	0.0256814																								
Uncorrected SS	119872	Corrected SS	4724.88889																								
Coeff Variation	20.5440709	Std Error Mean	1.9364689																								

The values of skewness and kurtosis for display are both close to 0.

## Partial PROC UNIVARIATE Output for In-Store Displays (continued)



## Partial PROC UNIVARIATE Output for In-Store Displays (continued)

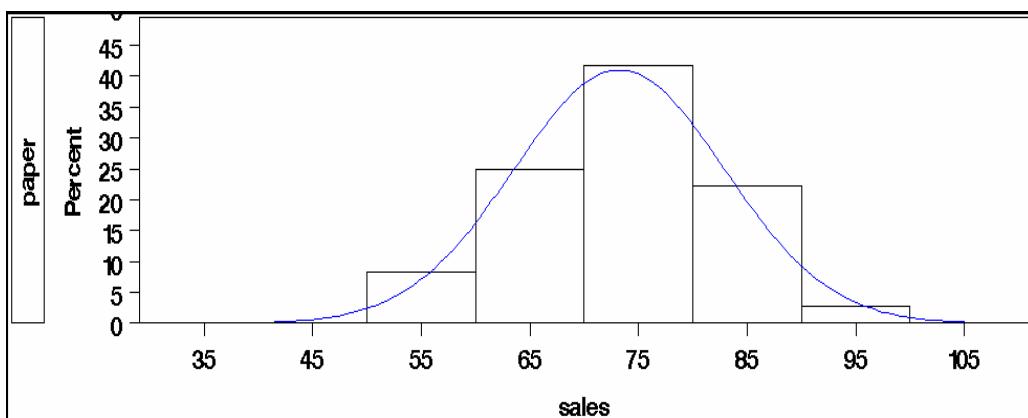


The histogram and normal probability plots for display indicate no patterns.

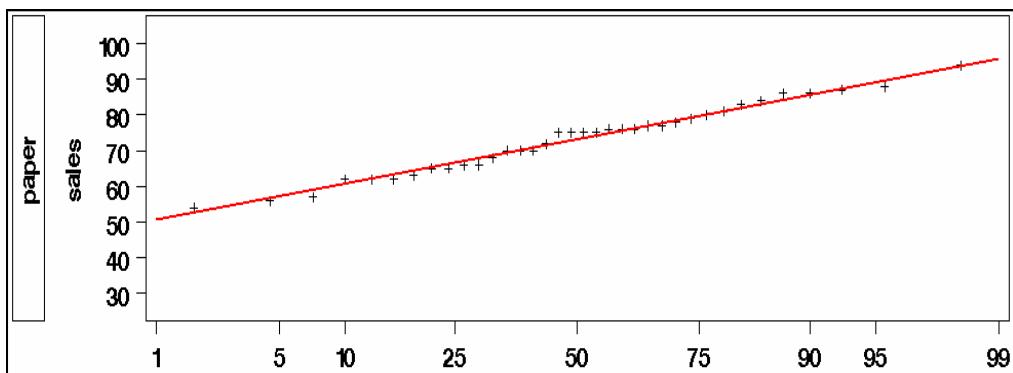
## Partial PROC UNIVARIATE Output for Local Newspaper Ads

The UNIVARIATE Procedure																											
Variable: sales																											
ad = paper																											
Moments																											
<table border="1" style="width:100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">N</td><td style="padding: 2px; text-align: right;">36</td><td style="padding: 2px;">Sum Weights</td><td style="padding: 2px; text-align: right;">36</td></tr> <tr> <td style="padding: 2px;">Mean</td><td style="padding: 2px; text-align: right;">73.2222222</td><td style="padding: 2px;">Sum Observations</td><td style="padding: 2px; text-align: right;">2636</td></tr> <tr> <td style="padding: 2px;">Std Deviation</td><td style="padding: 2px; text-align: right;">9.7339204</td><td style="padding: 2px;">Variance</td><td style="padding: 2px; text-align: right;">94.7492063</td></tr> <tr> <td style="padding: 2px;">Skewness</td><td style="padding: 2px; text-align: right;">-0.0474705</td><td style="padding: 2px;">Kurtosis</td><td style="padding: 2px; text-align: right;">-0.5475341</td></tr> <tr> <td style="padding: 2px;">Uncorrected SS</td><td style="padding: 2px; text-align: right;">196330</td><td style="padding: 2px;">Corrected SS</td><td style="padding: 2px; text-align: right;">3316.22222</td></tr> <tr> <td style="padding: 2px;">Coeff Variation</td><td style="padding: 2px; text-align: right;">13.2936697</td><td style="padding: 2px;">Std Error Mean</td><td style="padding: 2px; text-align: right;">1.62232007</td></tr> </table>				N	36	Sum Weights	36	Mean	73.2222222	Sum Observations	2636	Std Deviation	9.7339204	Variance	94.7492063	Skewness	-0.0474705	Kurtosis	-0.5475341	Uncorrected SS	196330	Corrected SS	3316.22222	Coeff Variation	13.2936697	Std Error Mean	1.62232007
N	36	Sum Weights	36																								
Mean	73.2222222	Sum Observations	2636																								
Std Deviation	9.7339204	Variance	94.7492063																								
Skewness	-0.0474705	Kurtosis	-0.5475341																								
Uncorrected SS	196330	Corrected SS	3316.22222																								
Coeff Variation	13.2936697	Std Error Mean	1.62232007																								

## Partial PROC UNIVARIATE Output for Local Newspaper Ads (continued)



## Partial PROC UNIVARIATE Output for Local Newspaper Ads (continued)



No patterns are seen in the above graphs for paper.

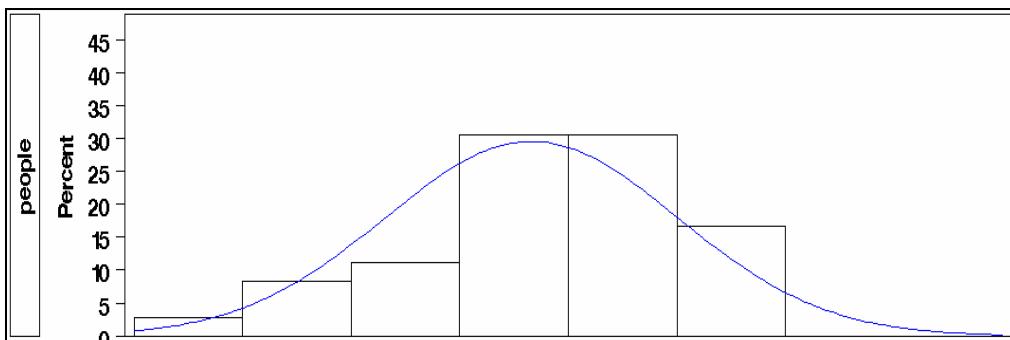
## Partial PROC UNIVARIATE Output for In-Store Salespeople

```
The UNIVARIATE Procedure
Variable: sales
ad = people

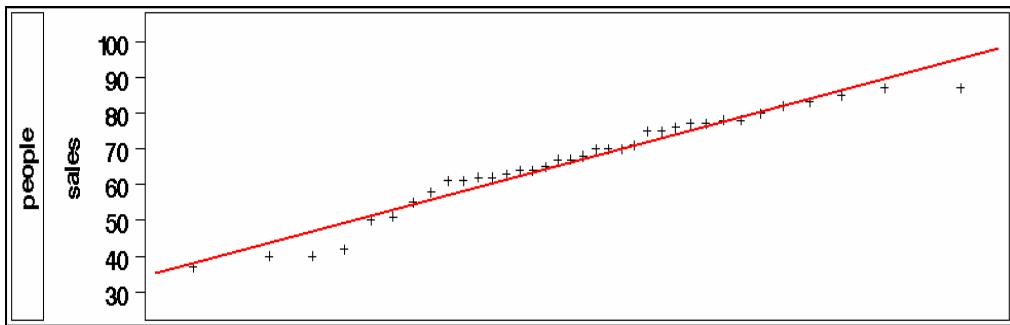
Moments

N           36   Sum Weights      36
Mean        66.6111111  Sum Observations 2398
Std Deviation 13.4976776  Variance       182.187302
Skewness     -0.5998808  Kurtosis       -0.2130516
Uncorrected SS    166110  Corrected SS    6376.55556
Coeff Variation  20.2634026  Std Error Mean 2.24961294
```

## Partial PROC UNIVARIATE Output for In-Store Salespeople (continued)



## Partial PROC UNIVARIATE Output for In-Store Salespeople (continued)



The histogram and normal probability plots for `people` do not indicate any patterns.

## Partial PROC UNIVARIATE Output for Local Radio Ads

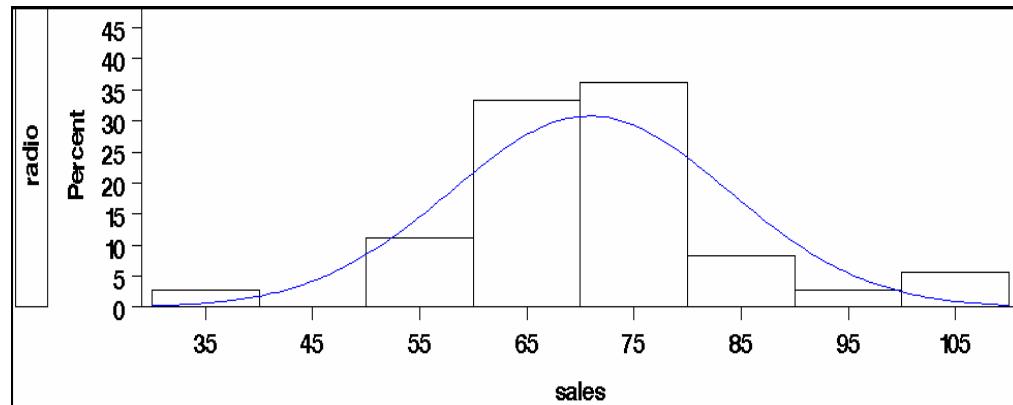
```
The UNIVARIATE Procedure
Variable: sales
ad = radio
```

## Moments

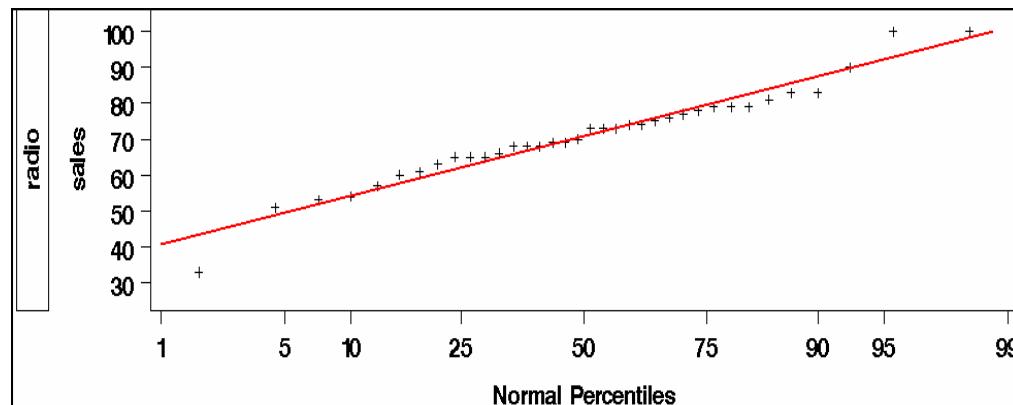
N	36	Sum Weights	36
Mean	70.8888889	Sum Observations	2552
Std Deviation	12.9676031	Variance	168.15873
Skewness	-0.2172278	Kurtosis	1.65652424
Uncorrected SS	186794	Corrected SS	5885.55556
Coeff Variation	18.292857	Std Error Mean	2.16126718

The relatively large value of kurtosis could indicate a potential outlier for radio.

## Partial PROC UNIVARIATE Output for Local Radio Ads (continued)



## Partial PROC UNIVARIATE Output for Local Radio Ads (continued)

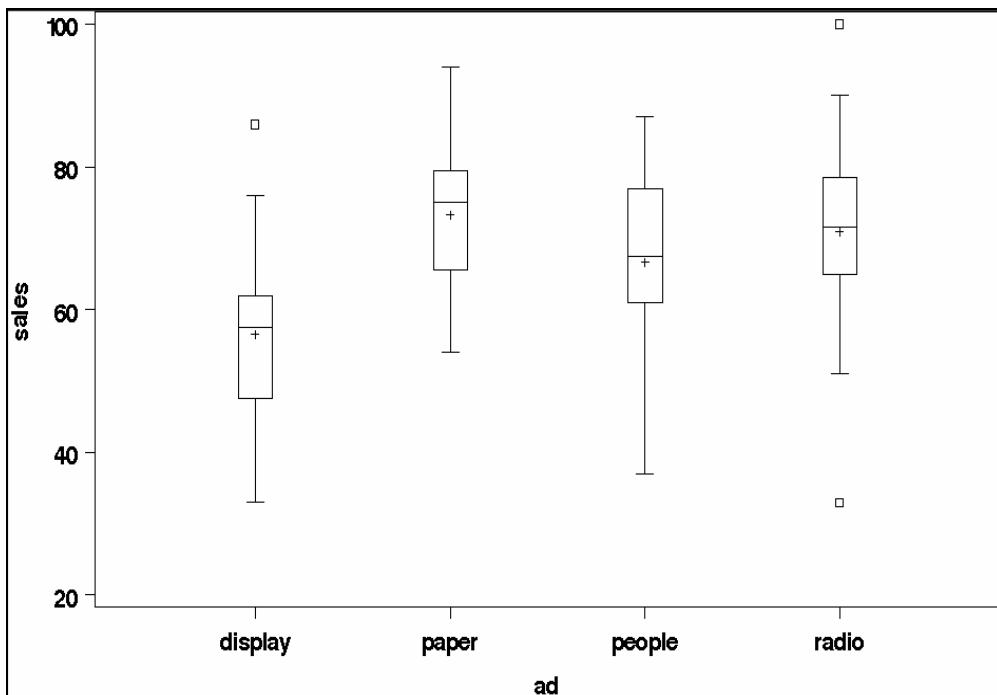


The histogram for radio does indicate some data in the tails, which agrees with the relatively large value of kurtosis.

In order to generate side-by-side box-and-whisker plots with PROC BOXPLOT, the data must be sorted by the horizontal (grouping) variable used in the PLOT statement.

```
proc sort data=sasuser.b_ads out=sorted;
  by ad;
run;

proc boxplot data=sorted;
  plot sales*ad / cboxes=black boxstyle=schematic;
run;
```



It appears that the in-store display mean is lower than the others. The value `display` has a positive outlier, while `radio` has outliers in both directions.

b. When you check the model assumptions, you find the following:

- Levene's test for equality of variance has a *p*-value of 0.3532. Therefore, you do not reject the null hypothesis that the variances are equal.
- The histogram, normal probability plot, and box-and-whisker plot indicate that there is no severe departure from the assumption that the residuals have a normal distribution.

```

proc glm data=sasuser.b_ads;
  class ad;
  model sales=ad;
  means ad / hovtest;
  output out=check r=resid p=pred;
run;
quit;

data check;
  set check;
  dummy = '1';
run;

proc univariate data=check;
  var resid;
  histogram / normal;
  probplot / normal (mu=est sigma=est color=red w=2);
run;

proc boxplot data=check;
  plot resid*dummy / cboxes=black
                    boxstyle=schematic;
run;

```

Partial PROC GLM Output

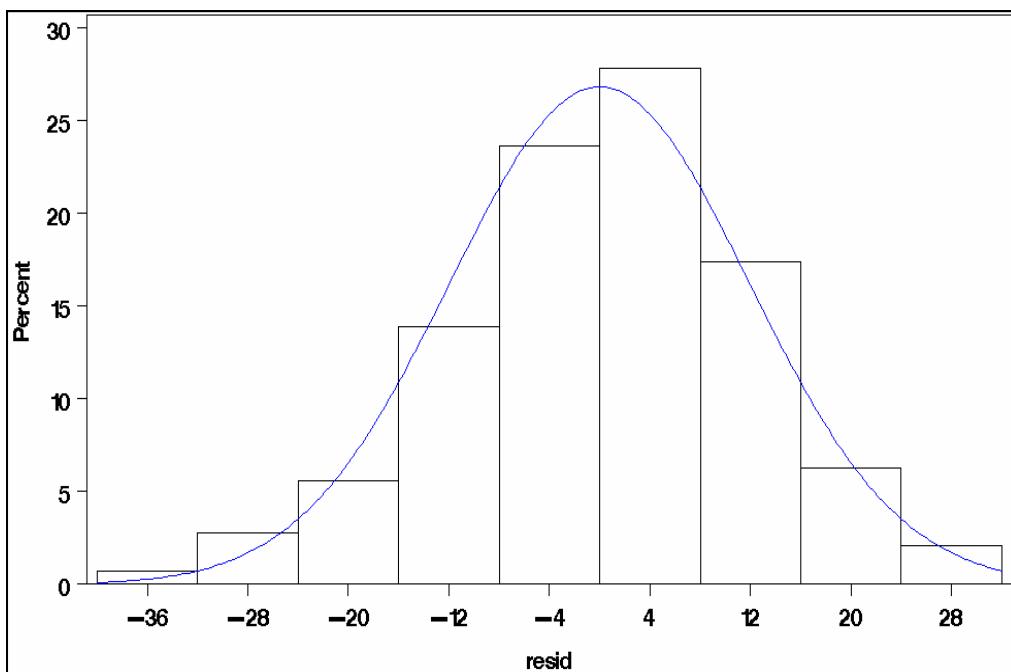
The GLM Procedure					
Levene's Test for Homogeneity of sales Variance					
ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ad	3	154637	51545.6	1.10	0.3532
Error	140	6586668	47047.6		

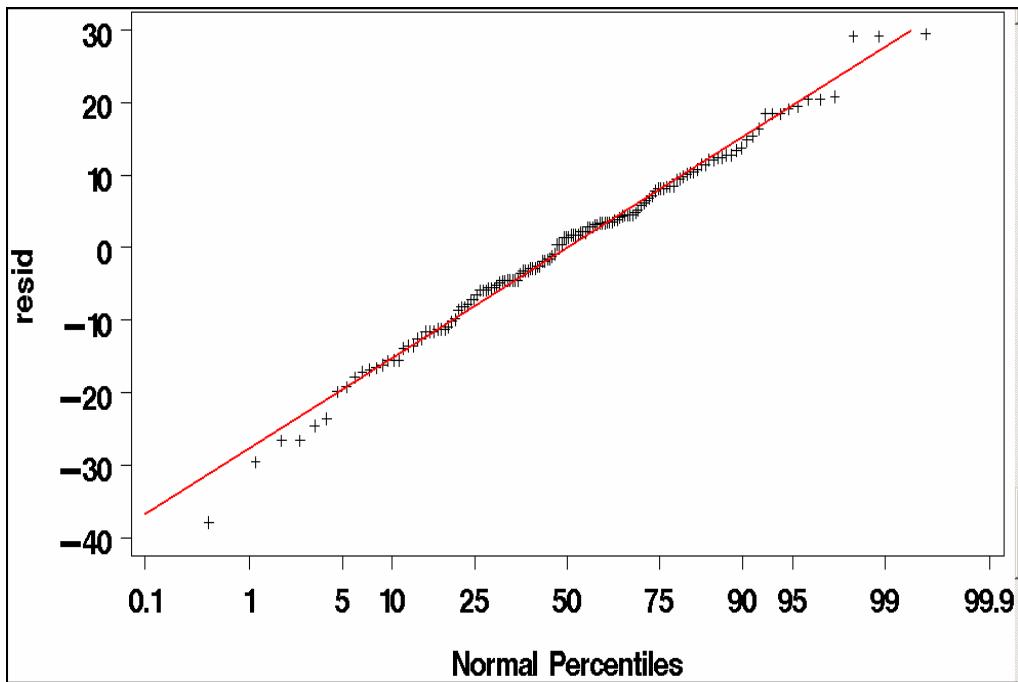
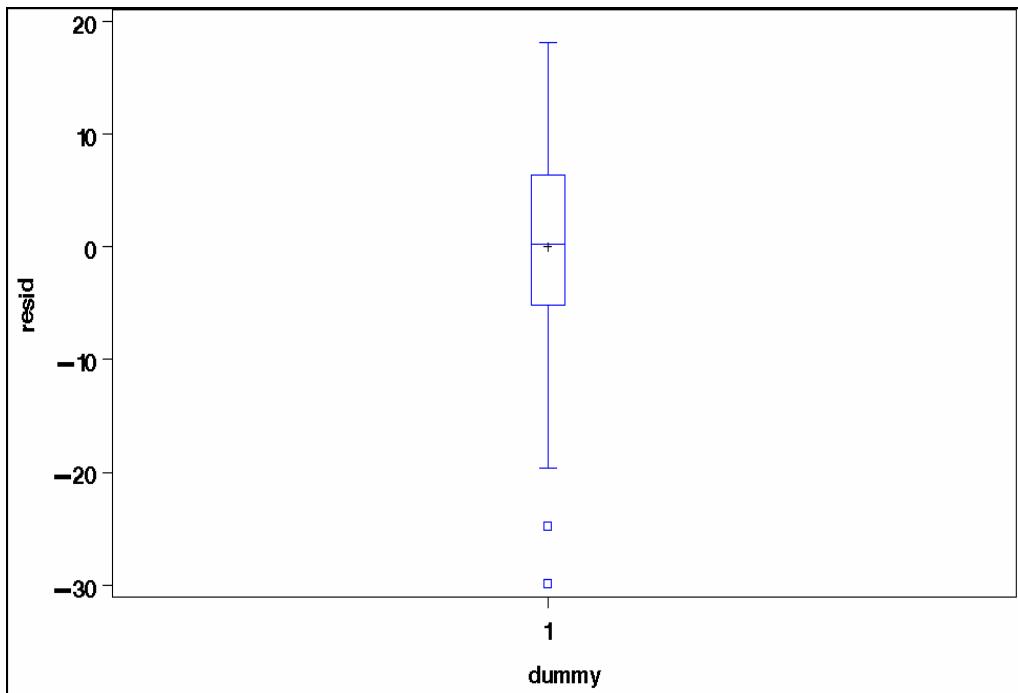
Partial PROC UNIVARIATE Output for **resid**

```
The UNIVARIATE Procedure
Variable: resid

Moments

N           144   Sum Weights      144
Mean         0     Sum Observations 0
Std Deviation 11.9155602 Variance    141.980575
Skewness     -0.200981 Kurtosis     0.41041374
Uncorrected SS 20303.2222 Corrected SS 20303.2222
Coeff Variation .     Std Error Mean 0.99296335
```

Partial PROC UNIVARIATE Output for **resid** (continued)

Partial PROC UNIVARIATE Output for **resid** (continued)PROC BOXPLOT Output for **resid**

## Partial PROC GLM Output

The GLM Procedure										
Class Level Information										
Class	Levels	Values								
ad	4	display paper people radio								
			Number of observations 144							
The GLM Procedure										
Dependent Variable: sales										
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F					
Model	3	5866.08333	1955.36111	13.48	<.0001					
Error	140	20303.22222	145.02302							
Corrected Total	143	26169.30556								
R-Square	Coeff Var	Root MSE	sales Mean							
0.224159	18.02252	12.04255	66.81944							
Source	DF	Type I SS	Mean Square	F Value	Pr > F					
ad	3	5866.083333	1955.361111	13.48	<.0001					
Source	DF	Type III SS	Mean Square	F Value	Pr > F					
ad	3	5866.083333	1955.361111	13.48	<.0001					

The overall  $F$  test from the analysis of variance table has a  $p$ -value less than or equal to .0001. Presuming that all assumptions of the model are valid, you know that at least one treatment mean is different from one other treatment mean. At this point you do not know which means are significantly different.

- c. Based on Tukey's multiple comparison method, using in-store displays is significantly different from all other types of advertising.

```
proc glm data=sasuser.b_ads;
  class ad;
  model sales=ad;
  lsmeans ad / pdiff=all;
  title 'Control Experimentwise Error Rate';
run;
quit;
```

Partial PROC GLM Output

Control Experimentwise Error Rate				
The GLM Procedure				
Least Squares Means				
Adjustment for Multiple Comparisons: Tukey				
ad	sales	LSMEAN Number		
display	56.5555556	1		
paper	73.2222222	2		
people	66.6111111	3		
radio	70.8888889	4		
Least Squares Means for effect ad				
Pr >  t  for H0: LSMean(i)=LSMean(j)				
Dependent Variable: sales				
i/j	1	2	3	
1		<.0001	0.0030	<.0001
2	<.0001		0.0964	0.8440
3	0.0030	0.0964		0.4360
4	<.0001	0.8440	0.4360	

## 2. Analyzing Data in a Randomized Block Design

- a. The overall  $F$  test with a  $p$ -value of less than or equal to 0.0001 means that you reject the null hypothesis that all treatment means are equal and conclude that at least one treatment mean is significantly different from one other treatment mean.
- Note that the  $F$  value for **area** is 6.07, indicating that the blocking factor was effective.
  - The MSE also decreased from 145.0230 to 89.73916 when the blocking factor was added to the analysis.

```
proc glm data=sasuser.b_ads1;
  class ad area;
  model sales = ad area;
run;
quit;
```

Partial PROC GLM Output

The GLM Procedure			
Class Level Information			
Class	Levels	Values	
ad	4	display paper people radio	
area	18	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18	
Number of Observations Read			144
Number of Observations Used			144

## Partial PROC GLM Output (continued)

The GLM Procedure						
Dependent Variable: sales						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	20	15131.38889	756.56944	8.43	<.0001	
Error	123	11037.91667	89.73916			
Corrected Total	143	26169.30556				
R-Square	Coeff Var	Root MSE	sales Mean			
0.578211	14.17712	9.473076	66.81944			
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
ad	3	5866.083333	1955.361111	21.79	<.0001	
area	17	9265.305556	545.017974	6.07	<.0001	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
ad	3	5866.083333	1955.361111	21.79	<.0001	
area	17	9265.305556	545.017974	6.07	<.0001	

- b. Including the blocking factor in the model and controlling the experimentwise error rate, in-store displays are still significantly different from all other types of advertising. Also, newspaper advertising is significantly different from in-store salespeople.

```
proc glm data=sasuser.b_ads1;
  class ad area;
  model sales=ad area;
  lsmeans ad / pdiff=all adjust=tukey;
run;
quit;
```

#### Partial PROC GLM Output

The GLM Procedure  
 Least Squares Means  
 Adjustment for Multiple Comparisons: Tukey

ad	sales	LSMEAN Number
display	56.5555556	1
paper	73.2222222	2
people	66.6111111	3
radio	70.8888889	4

Least Squares Means for effect ad  
 $\text{Pr} > |t| \text{ for } H_0: \text{LSMean}(i) = \text{LSMean}(j)$

Dependent Variable: sales

i/j	1	2	3	4
1		<.0001	<.0001	<.0001
2	<.0001		0.0190	0.7233
3	<.0001	0.0190		0.2268
4	<.0001	0.7233	0.2268	

### 3. Performing Two-Way ANOVA

- a. Use PROC MEANS and PROC BOXPLOT to explore **strength**, based on the levels of **brand** or **additive**. Use PROC GPLOT to examine possible interactions.

```
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
title;

proc means data=sasuser.b_cement /* by brand */
    mean var std;
class brand;
var strength;
title 'Descriptive Statistics: b_cement - by brand';
run;

proc sort data=sasuser.b_cement out=sort_b;
by brand;
run;
proc boxplot data=sort_b;
plot strength*brand / cboxes=black boxstyle=schematic;
run;

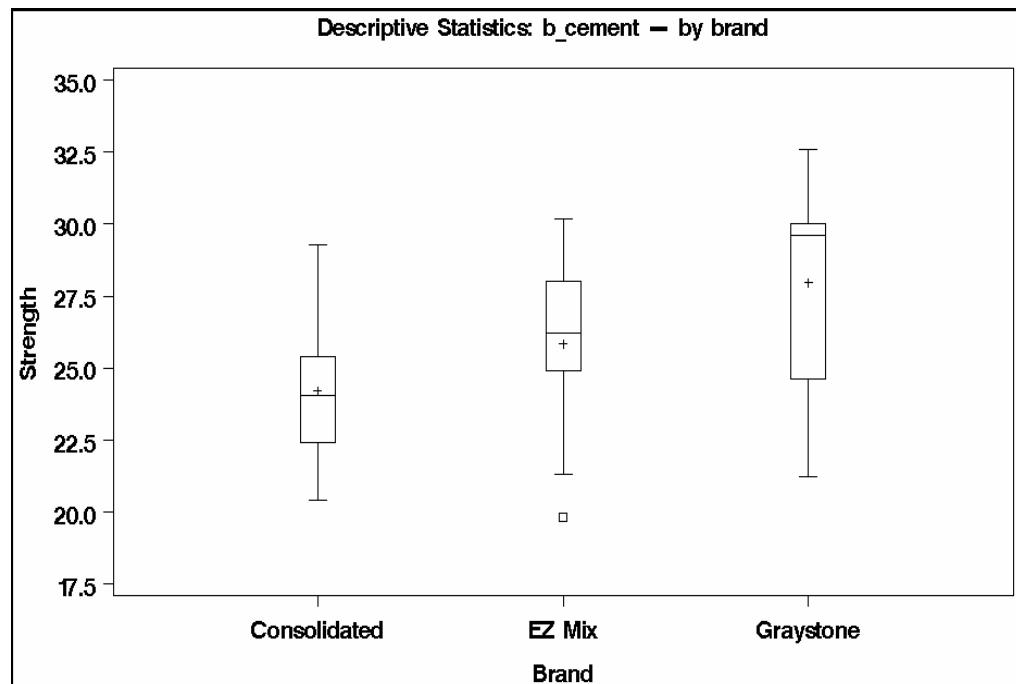
proc means data=sasuser.b_cement /* by strength */
    mean var std;
class additive;
var strength;
title 'Descriptive Statistics: b_cement - by additive';
run;

proc sort data=sasuser.b_cement out=sort_a;
by additive;
run;
proc boxplot data=sort_a;
plot strength*additive / cboxes=black boxstyle=schematic;
run;

proc gplot data=sasuser.b_cement;
symbol c=blue w=2 interpol=stdlmtj line=1;
symbol2 c=green w=2 interpol=stdlmtj line=2;
plot strength*brand=additive;
title 'Interactions?';
run;
quit;
```

PROC MEANS Output and PROC BOXPLOT Output by **brand**

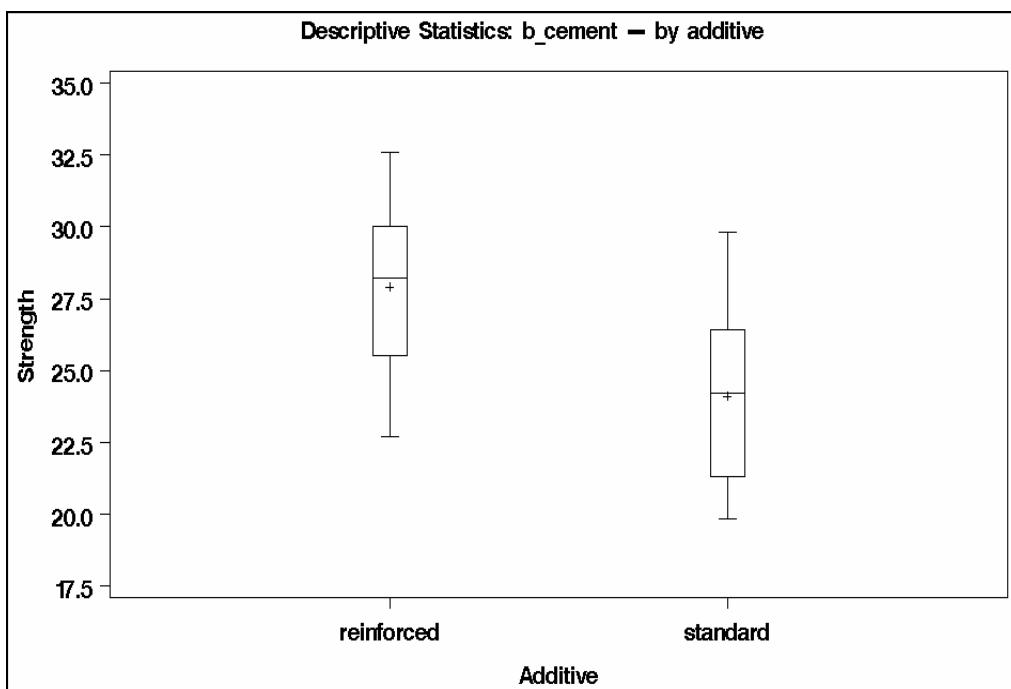
Descriptive Statistics: b_cement - by brand				
The MEANS Procedure				
Analysis Variable : Strength				
Brand	N	Mean	Variance	Std Dev
Obs				
Consolidated	10	24.200000	6.3888889	2.5276251
EZ Mix	10	25.8300000	10.3067778	3.2104171
Graystone	10	27.9700000	13.2334444	3.6377802



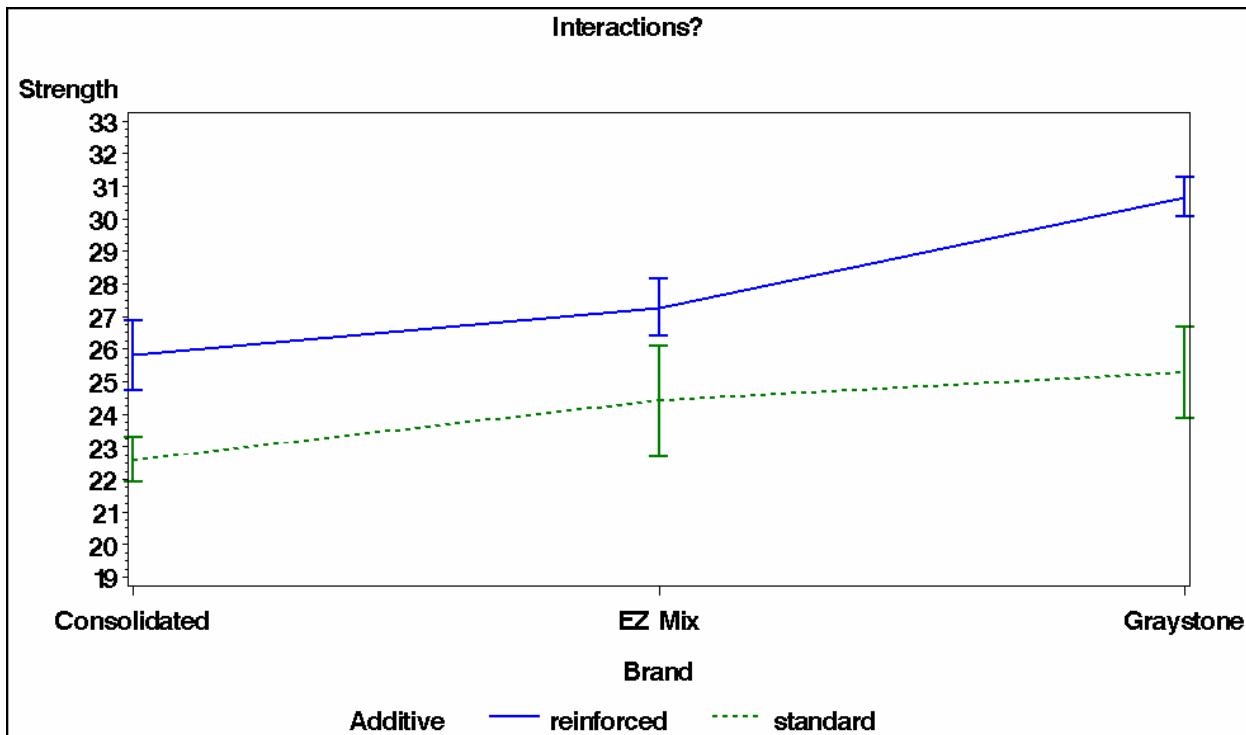
The data is balanced within **brand**. Graystone has the highest mean of **strength**, and EZ Mix has an outlier.

PROC MEANS Output and PROC BOXPLOT Output by **additive**

Descriptive Statistics: b_cement - by additive				
The MEANS Procedure				
Analysis Variable : Strength				
Additive	Obs	N	Mean	Variance
reinforced	15	15	27.9066667	7.6606667
standard	15	15	24.0933333	8.8963810



The data is balanced within **additive**. The value **reinforced** has the higher mean.



It does not appear that there is an interaction between **additive** and **brand**.

- b. The overall *F* test with a *p*-value of 0.0009 means that you reject the null hypothesis that **strength** is not affected by **brand** or **additive**. The Type I and Type III sums of squares are equal because the data is balanced. Both **brand** and **additive** are statistically significant, but the interaction is not significant (*p*-value = 0.4862). You can remove the interaction from the model and do pairwise tests separately for **brand** and **additive** if desired.

```
proc glm data=sasuser.b_cement;
  class brand additive;
  model strength=brand additive brand*additive;
  title 'Analysis of Concrete Brands';
run;
quit;
```

#### Partial PROC GLM Output

Analyze the Effects of brand and additive and their interaction					
The GLM Procedure					
Dependent Variable: Strength					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	189.9080000	37.9816000	6.04	0.0009
Error	24	150.9520000	6.2896667		
Corrected Total	29	340.8600000			
R-Square	Coeff Var	Root MSE	Strength Mean		
0.557144	9.645849	2.507921	26.00000		
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Brand	2	71.4980000	35.7490000	5.68	0.0095
Additive	1	109.0613333	109.0613333	17.34	0.0003
Brand*Additive	2	9.3486667	4.6743333	0.74	0.4862
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Brand	2	71.4980000	35.7490000	5.68	0.0095
Additive	1	109.0613333	109.0613333	17.34	0.0003
Brand*Additive	2	9.3486667	4.6743333	0.74	0.4862

- c. The elimination of the interaction term does change the overall F statistic ( $p$ -value = 0.0002) as well as the tests for **brand** and **additive**.

There is a significant difference due to **additive** (reinforced cement is stronger than standard). Pairwise comparisons are not necessary for **additive** because it only has two levels.

```
proc glm data=sasuser.b_cement;
  class brand additive;
  model strength=brand additive;
  lsmeans brand / pdiff=all;
  title 'strength=brand additive';
run;
quit;
```

Partial PROC GLM Output

strength=brand additive						
The GLM Procedure						
Dependent Variable: Strength						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	3	180.5593333	60.1864444	9.76	0.0002	
Error	26	160.3006667	6.1654103			
Corrected Total	29	340.8600000				
R-Square	Coeff Var	Root MSE	Strength Mean			
0.529717	9.550094	2.483024	26.00000			
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
Brand	2	71.4980000	35.7490000	5.80	0.0083	
Additive	1	109.0613333	109.0613333	17.69	0.0003	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
Brand	2	71.4980000	35.7490000	5.80	0.0083	
Additive	1	109.0613333	109.0613333	17.69	0.0003	

## Partial PROC GLM Output (continued)

```

strength=brand additive

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

      Strength      LSMEAN
Brand          LSMEAN    Number

Consolidated   24.2000000   1
EZ Mix         25.8300000   2
Graystone      27.9700000   3

Least Squares Means for effect Brand
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Strength

      i/j       1       2       3

      1           0.3224    0.0061
      2           0.3224    0.1512
      3           0.0061    0.1512

```

The tests show that there is a significant difference between the Consolidated and Graystone brands.

## Chapter 3

### 1. Describing the Relationship between Two Continuous Variables

- Use the UNIVARIATE procedure to examine the distribution of the variables **rate** and **tuition**.

```

options ps=50 ls=97;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=sasuser.b_colleg;
var rate tuition;
id name;
histogram rate tuition / normal;
probplot rate tuition / normal
(mu=est sigma=est color=blue w=2);
title;
run;

```

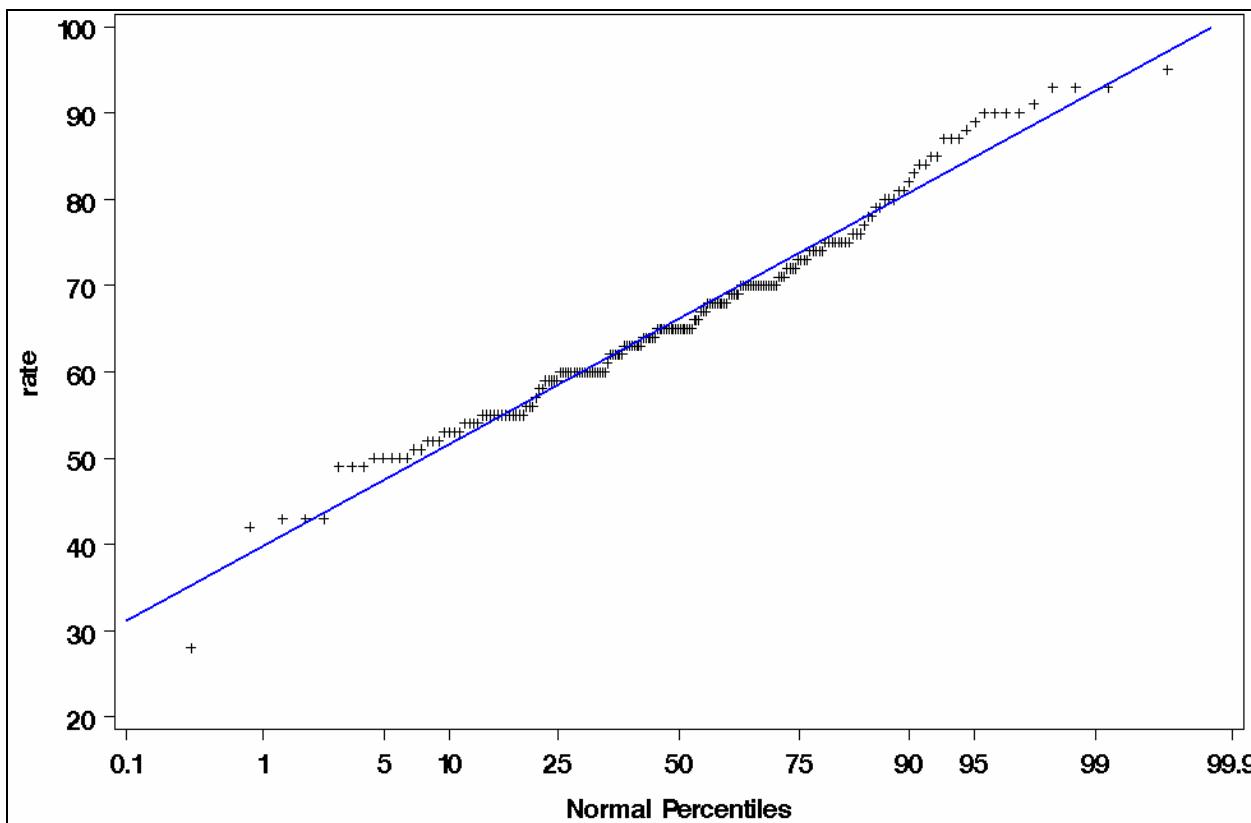
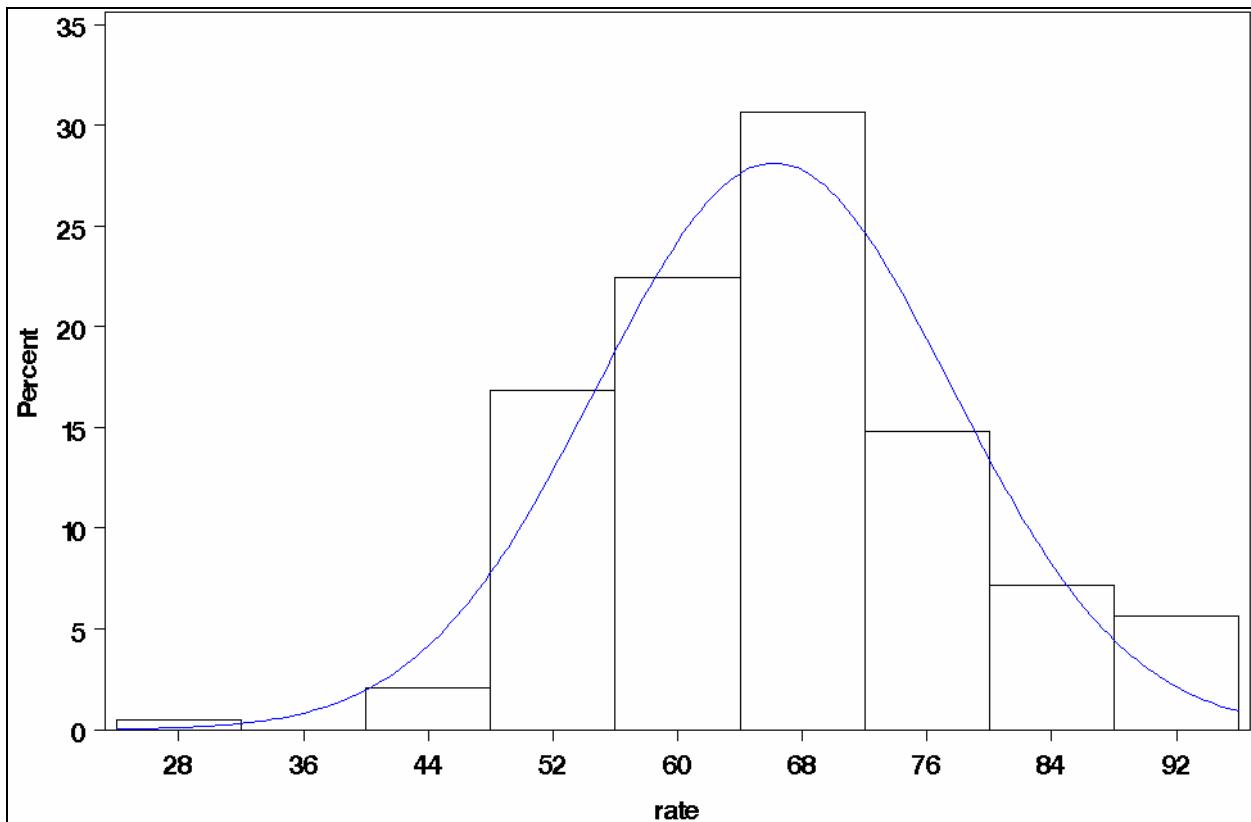
## Partial PROC UNIVARIATE Output

The UNIVARIATE Procedure			
Variable: rate			
Moments			
N	196	Sum Weights	196
Mean	66.1989796	Sum Observations	12975
Std Deviation	11.3506443	Variance	128.837127
Skewness	0.23195477	Kurtosis	0.34296403
Uncorrected SS	884055	Corrected SS	25123.2398
Coeff Variation	17.1462527	Std Error Mean	0.81076031
Basic Statistical Measures			
Location		Variability	
Mean	66.19898	Std Deviation	11.35064
Median	65.00000	Variance	128.83713
Mode	60.00000	Range	67.00000
		Interquartile Range	13.50000

## Partial PROC UNIVARIATE Output (continued)

The UNIVARIATE Procedure			
Variable: rate			
Extreme Observations			
-----Lowest-----			
Value	name		Obs
28	City College-City U. of N.Y.		25
42	U of Minnesota-Twin Cities		171
43	Washington State		193
43	U of Maryland-College Park		168
43	New College of U of South Fla		94
Extreme Observations			
-----Highest-----			
Value	name		Obs
91	Princeton		104
93	Columbia U.		29
93	Duke		36
93	U of Notre Dame		176
95	Harvard		54
Missing Values			
-----Percent Of-----			
Missing Value	Count	All Obs	Missing Obs
.	4	2.00	100.00

## Partial PROC UNIVARIATE Output (continued)



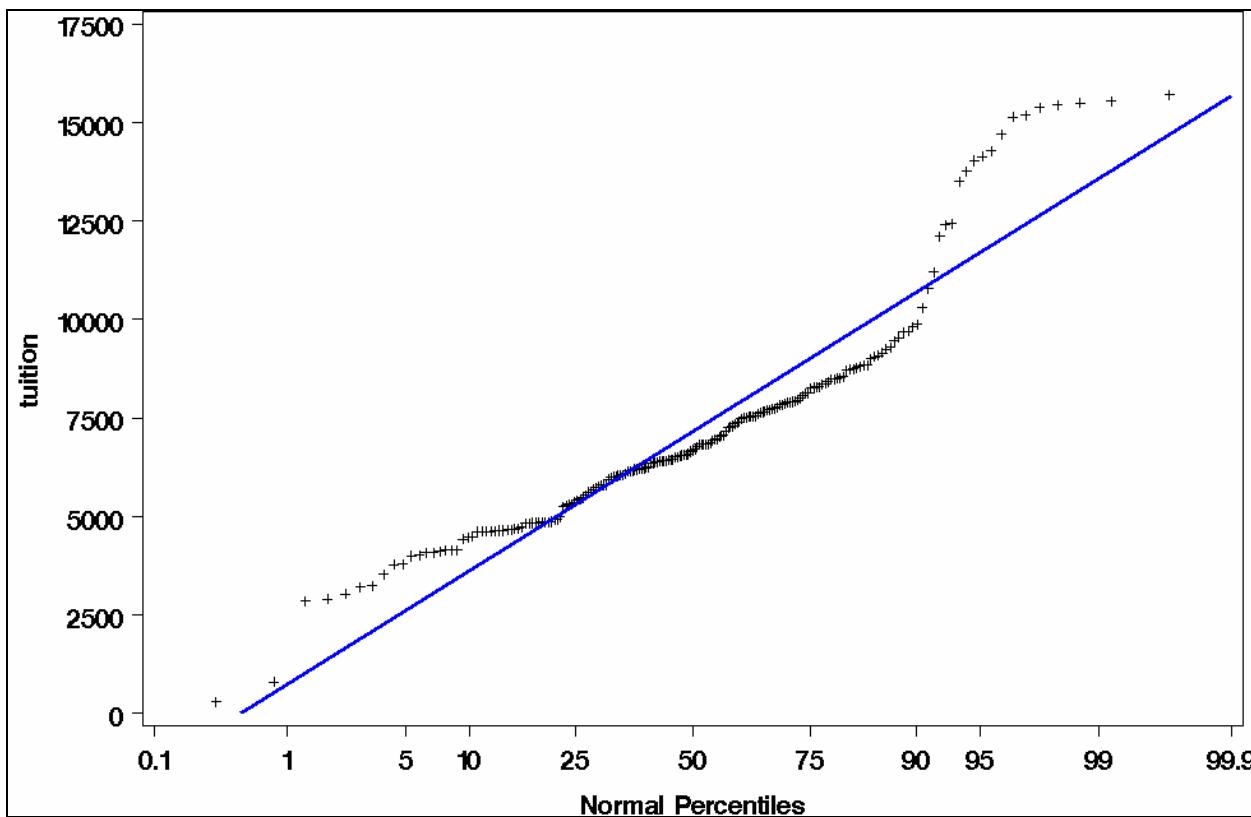
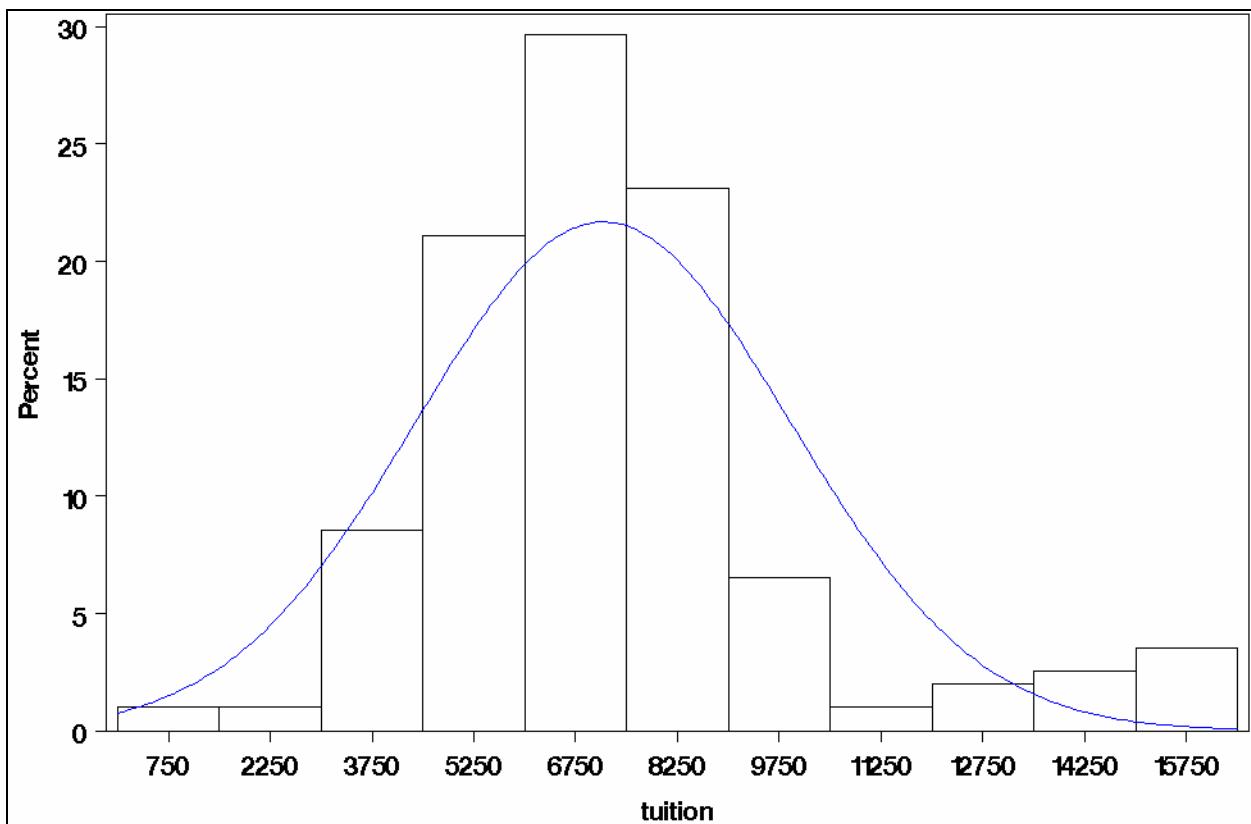
## Partial PROC UNIVARIATE Output (continued)

The UNIVARIATE Procedure			
Variable: tuition			
Moments			
N	199	Sum Weights	199
Mean	7152.34673	Sum Observations	1423317
Std Deviation	2762.1446	Variance	7629442.79
Skewness	1.195457	Kurtosis	2.06346036
Uncorrected SS	1.16907E10	Corrected SS	1510629673
Coeff Variation	38.6187178	Std Error Mean	195.803239
Basic Statistical Measures			
Location		Variability	
Mean	7152.347	Std Deviation	2762
Median	6685.000	Variance	7629443
Mode	4622.000	Range	15390
		Interquartile Range	2900

## Partial PROC UNIVARIATE Output (continued)

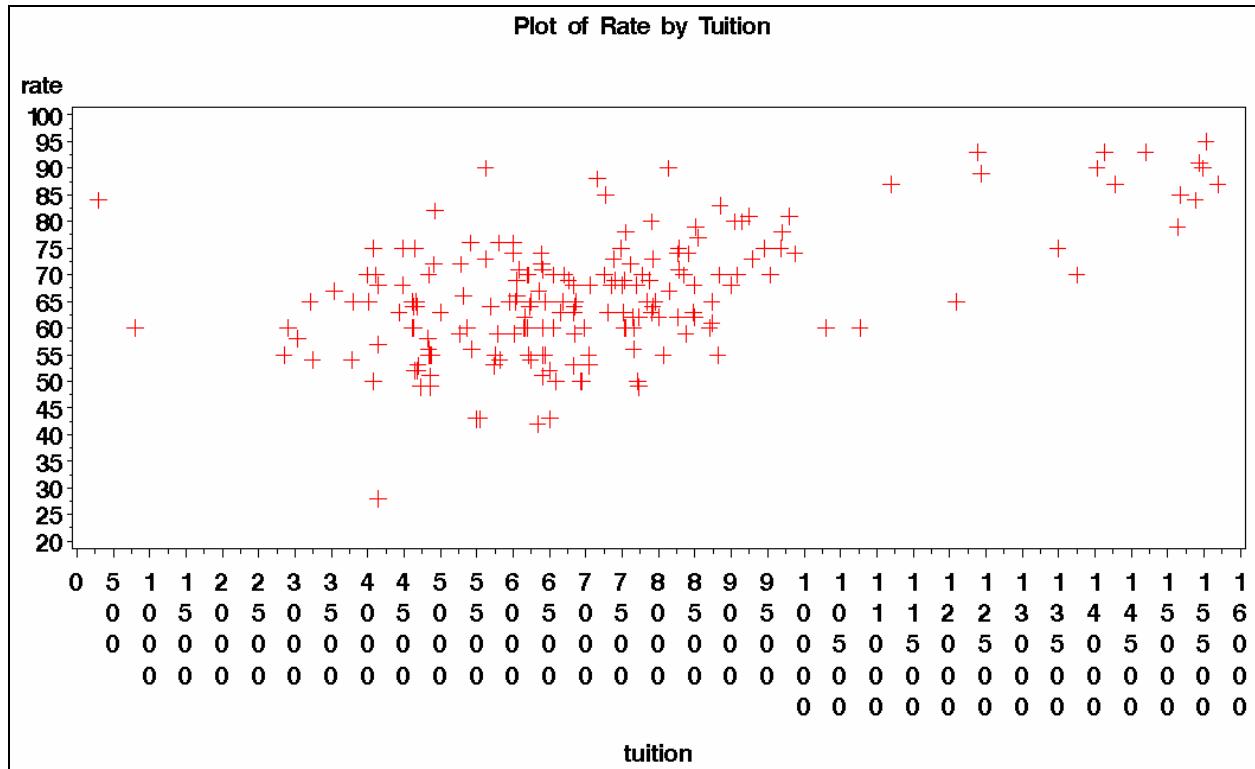
Variable: tuition			
Extreme Observations			
-----Lowest-----			
Value	name	Obs	
300	Cooper Union	31	
800	Emory and Henry	39	
2850	Brigham Young	16	
2903	Jersey City State	63	
3032	Montclair State	89	
Extreme Observations			
-----Highest-----			
Value	name	Obs	
15380	Johns Hopkins	65	
15440	Princeton	104	
15490	Swarthmore	138	
15530	Harvard	54	
15690	MIT	81	
Missing Values			
-----Percent Of-----			
Missing Value	Count	All Obs	Missing Obs
.	1	0.50	100.00

## Partial PROC UNIVARIATE Output (continued)



- 1) The variable **rate** is skewed slightly to the right with slightly heavy tails. The variable **tuition** is skewed to the right.
- 2) There are two unusual observations for the variable **rate**: City College with a graduation rate of 28% and Harvard with a graduation rate of 95%. There are several unusual observations for the variable **tuition**: those colleges with tuition below \$1000 and those with tuition above \$13,000.
- b. Use the GPLOT procedure to generate a scatter plot for the variables **rate** versus **tuition**.

```
proc gplot data=sasuser.b_colleg;
  plot rate*tuition
    / haxis = 0 to 16000 by 500
      vaxis = 20 to 100 by 5;
  symbol v=plus color=red h=2;
  title 'Plot of Rate by Tuition';
run;
quit;
```



- 1) A straight line can adequately describe the data.

- 2) If you examine the plot from left to right, there appear to be three outliers that warrant investigation:

- **tuition**=300 and **rate**=84
- **tuition**=800 and **rate**=60
- **tuition**=4100 and **rate**=28.

- c. Use the CORR procedure to generate a correlation coefficient for the variables **rate** and **tuition**.

```
proc corr data=sasuser.b_colleg nosimple;
  var rate tuition;
  title;
run;
```

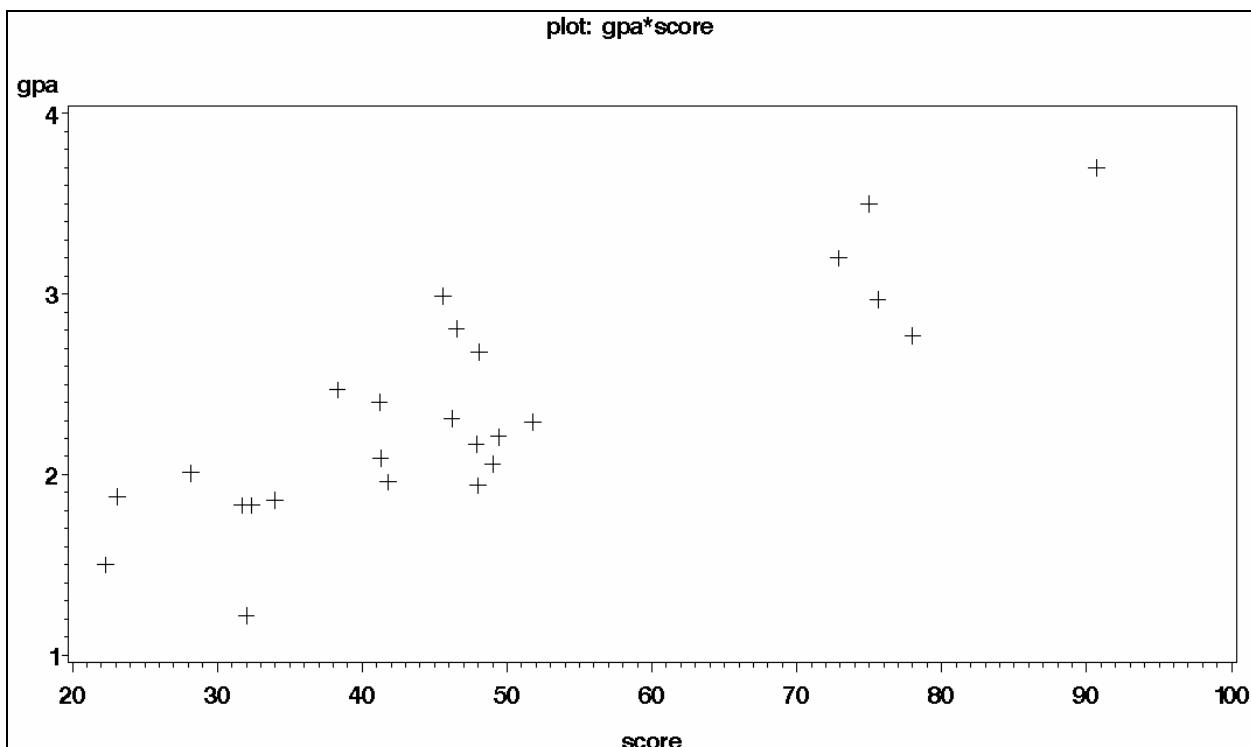
The CORR Procedure		
2 Variables: rate tuition		
Pearson Correlation Coefficients		
Prob >  r  under H0: Rho=0		
Number of Observations		
	rate	tuition
rate	1.00000	0.55385 <.0001 196
tuition	0.55385 <.0001 196	1.00000 199

- 1) The correlation coefficient is 0.55385.
- 2) The correlation indicates a moderately strong positive linear relationship between **rate** and **tuition**.
- 3) The *p*-value is less than 0.0001.
- 4) The correlation coefficient is statistically significant at the .05 level.

## 2. Fitting a Simple Linear Regression

- a. Use PROC GPLOT to generate a scatter plot for the variables **gpa** versus **score**.

```
options ps=50 ls=97;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc gplot data=sasuser.b_grades;
  plot gpa*score;
  symbol v=plus h=2;
  title 'plot: gpa*score';
run;
quit;
title;
```



- 1) A straight line can adequately describe the data.
- 2) The range of **score** is approximately 22 to 92.
- 3) There are no outliers that warrant investigation.

- b. Use the REG procedure to perform a regression analysis.

```
proc reg data=sasuser.b_grades;
  model gpa=score;
run;
quit;
```

### PROC REG Output

The REG Procedure Model: MODEL1 Dependent Variable: gpa					
Number of Observations Read			25		
Number of Observations Used			25		
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.33971	6.33971	58.77	<.0001
Error	23	2.48109	0.10787		
Corrected Total	24	8.82080			
Root MSE      R-Square      0.7187 Dependent Mean      2.34600      Adj R-Sq      0.7065 Coeff Var      14.00003					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	0.98051	0.18985	5.16	<.0001
score	1	0.02866	0.00374	7.67	<.0001

- 1) The value of the  $F$  statistic is 58.77, and the corresponding  $p$ -value is less than 0.0001. With this result you can reject the null hypothesis and conclude that at least one of the parameter estimates is not equal to 0. Because this is a simple linear regression, this means the parameter estimate for **score** is not equal to 0.
- 2) The predicted regression equation is  $\text{gpa}=0.98051 + 0.02866*\text{score}$ . The model indicates that the predicted grade point average at the end of the freshman year is equal to  $0.98051 + (0.02866 * \text{the student's exam score})$ .
- 3) The  $R^2$  statistic is 0.7187. This means that the regression line explains 71.87% of the total variation in the data.
- 4) The parameter estimate for **score** is 0.02866. This means that a 1-point increase in the exam score would amount to a 0.02866 increase in the grade point average at the end of the freshman year.

- c. Because the regression equation is known, use a DATA step to generate the predicted values of gpa. The results are in the log.

```
data _null_;
  input score @@;
  gpa = 0.98051 + 0.02866 * score;
  put score= gpa=;
  datalines;
40 60 80
;
run;
```

```
44  data _null_;
45    input score @@;
46    gpa = 0.98051 + 0.02866 * score;
47    put score= gpa=;
48    datalines;

score=40 gpa=2.12691
score=60 gpa=2.70011
score=80 gpa=3.27331
NOTE: SAS went to a new line when INPUT statement reached past the end of a line.
NOTE: DATA statement used (Total process time):
      real time          0.07 seconds
      cpu time          0.00 seconds

50  ;
51  run;
```

- 1) The predicted values are 2.12691, 2.70011, and 3.27331, respectively.
- 2) Because the values of **score** range from 22 to 92, it would be inappropriate to predict **gpa** when **score** is 200.

 There is an alternative way of generating these predicted values, using the OUTTEST= option in PROC REG. Notice that the predicted values are more precise because the values are stored with more digits and have not been rounded.

```
proc reg data=sasuser.b_grades
  noint
  outest=savedbetas (rename=(score=beta_score));
  model gpa=score;
run;
quit;

data _null_;
  input score @@;
  if _n_ = 1
    then set savedbetas;

  gpa = intercept + beta_score*score;
  put score= gpa=;
  datalines;
40 60 80
;
run;
```

Use VIEWTABLE to examine the values in **savedbetas**.

VIEWTABLE: Parameter Estimates and Statistics							
	Label of model	Type of statistics	Dependent variable	Root mean squared error	Intercept	beta_score	gpa
1	MODEL1	PARMS	gpa	0.3284407606	0.9805065248	0.0286627514	-1

#### SAS Log

```
4  proc reg data=sasuser.b_grades
5    noint
6    outest=savedbetas (rename=(score=beta_score));
7    model gpa=score;
8  run;
9
quit;

NOTE: The data set WORK.SAVEDBETAS has 1 observations and 7 variables.
NOTE: PROCEDURE REG used (Total process time):
      real time          0.10 seconds
      cpu time           0.10 seconds

10
```

## SAS Log (continued)

```

11  data _null_;
12    input score @@;
13    if _n_ = 1
14      then set savedbetas;
15
16    gpa = intercept + beta_score*score;
17    put score= gpa=;
18    datalines;

score=40 gpa=2.1270165795
score=60 gpa=2.7002716069
score=80 gpa=3.2735266343
NOTE: SAS went to a new line when INPUT statement reached past the end of a line.
NOTE: There were 1 observations read from the data set WORK.SAVEDBETAS.
NOTE: DATA statement used (Total process time):
      real time          0.04 seconds
      cpu time          0.03 seconds

20  ;
21  run;

```

- d. Create a data set that contains the **sasuser.b\_grades** data and the additional observations on which you want predictions. Then use PROC REG to produce confidence and prediction intervals around these predictions.

```

data need_predictions;
  input score @@;
  datalines;
40 60 80
;
run;

data sasuser.b_grade2;
  set sasuser.b_grades
    need_predictions;
run;

proc reg data=sasuser.b_grade2;
  model gpa=score / p clm cli;
  id score;
run;
quit;

```

## Partial PROC REG Output

Output Statistics						
Obs	score	Dependent Variable	Predicted Value	Std Error Mean	95% CL Predict	Mean
26	40	.	2.1270	0.0716	1.9788	2.2752
27	60	.	2.7003	0.0803	2.5341	2.8664
28	80	.	3.2735	0.1377	2.9887	3.5583
Obs	score	95% CL	Predict	Residual		
26	40	1.4316	2.8224	.		
27	60	2.0008	3.3997	.		
28	80	2.5368	4.0102	.		

- 1) The predicted confidence interval for the predicted mean of **gpa** when **score** is 60 is 2.5341 through 2.8664. This indicates that you are 95% confident that the population mean of **gpa** is between 2.5341 and 2.8664 when **score** is 60.
- 2) The predicted prediction interval for the predicted value of **gpa** when **score** is 60 is 2.0008 through 3.3997. This indicates that you are 95% confident that a new value of **gpa** falls between 2.0008 and 3.3997 when **score** is 60.

**3. Performing a Regression Using the REG Procedure**

This program runs a regression of **Oxygen\_Consumption** on the other continuous variables in the **b\_fitness** data set.

```
options ps=50 ls=97;
proc reg data=sasuser.b_fitness;
  model oxygen_consumption=performance runtime age
    weight run_pulse rest_pulse maximum_pulse;
  title 'Regression of Oxygen_Consumption on All '
    'Predictors';
run;
quit;
```

## PROC REG Output

Regression of Oxygen_Consumption on All Predictors					
The REG Procedure					
Model: MODEL1					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read		31			
Number of Observations Used		31			
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.03251	103.14750	18.32	<.0001
Error	23	129.52204	5.63139		
Corrected Total	30	851.55455			
Root MSE		2.37306	R-Square	0.8479	
Dependent Mean		47.37581	Adj R-Sq	0.8016	
Coeff Var		5.00900			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	93.33753	36.49782	2.56	0.0176
Performance	1	0.25756	1.02373	0.25	0.8036
Runtime	1	-2.08804	2.22856	-0.94	0.3585
Age	1	-0.21066	0.10519	-2.00	0.0571
Weight	1	-0.07741	0.05681	-1.36	0.1862
Run_Pulse	1	-0.36618	0.12299	-2.98	0.0067
Rest_Pulse	1	-0.01389	0.07114	-0.20	0.8469
Maximum_Pulse	1	0.30490	0.13990	2.18	0.0398

- a. There are key differences between the ANOVA table for this model and the Simple Linear Regression model.
  - The degrees of freedom for the model are much higher, 7 versus 1.
  - The Mean Square Model and the *F* ratio are much smaller.
- b. Both the  $R^2$  and adjusted  $R^2$  for the full models are larger than the simple linear regression. Consequently, the full model explains over 80 percent of the variation in the **Oxygen\_Consumption** variable versus only about 75 percent explained by the simple linear regression.
- c. Yes, including the other variables in the model changed both the estimate of the intercept and the slope for **Performance**. Also, the *p*-values for both changed dramatically. The slope of **Performance** is now not significantly different from zero.

#### 4. Simplifying the Model

- a. This program reruns the regression with **Rest\_Pulse** removed because it has the largest *p*-value (0.8469).

```
proc reg data=sasuser.b_fitness;
  REMOVE1: model oxygen_consumption=performance runtime age
             weight run_pulse maximum_pulse;
  title 'Remove Rest_Pulse';
run;
quit;
```

PROC REG Output

Remove Rest_Pulse					
The REG Procedure					
Model: REMOVE1					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read		31			
Number of Observations Used		31			
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	721.81791	120.30298	22.25	<.0001
Error	24	129.73665	5.40569		
Corrected Total	30	851.55455			
Root MSE		2.32501	R-Square	0.8476	
Dependent Mean		47.37581	Adj R-Sq	0.8096	
Coeff Var		4.90760			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	90.83022	33.47159	2.71	0.0121
Performance	1	0.32048	0.95201	0.34	0.7393
Runtime	1	-1.98433	2.12049	-0.94	0.3587
Age	1	-0.20470	0.09862	-2.08	0.0488
Weight	1	-0.07689	0.05560	-1.38	0.1794
Run_Pulse	1	-0.36818	0.12008	-3.07	0.0053
Maximum_Pulse	1	0.30593	0.13697	2.23	0.0351

- b. No, the *p*-value for the model did not change.

- c. The  $R^2$  only dropped by 0.0003, essentially no change. The adjusted  $R^2$  increased from .8016 to .8096. When an adjusted  $R^2$  increases by removing a variable from the models, it strongly implies that the removed variable was not necessary.

- d. All the parameter estimates and their  $p$ -values changed; some only a little.

## 5. More Simplifying of the Model

- a. This program reruns the regression with **Performance** removed, because it is the variable with the highest  $p$ -value in the Exercise 4 model.

```
proc reg data=sasuser.b_fitness;
  REMOVE2: model oxygen_consumption=runtime age weight
             run_pulse maximum_pulse;
  title 'Remove Rest_Pulse and Performance';
run;
quit;
```

PROC REG Output

Remove Rest_Pulse and Performance					
The REG Procedure					
Model: REMOVE2					
Dependent Variable: Oxygen_Consumption					
Number of Observations Read 31					
Number of Observations Used 31					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.20532	144.24106	27.66	<.0001
Error	25	130.34923	5.21397		
Corrected Total	30	851.55455			
Root MSE 2.28341 R-Square 0.8469					
Dependent Mean 47.37581 Adj R-Sq 0.8163					
Coeff Var 4.81978					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	101.33835	11.86474	8.54	<.0001
Runtime	1	-2.68846	0.34202	-7.86	<.0001
Age	1	-0.21217	0.09437	-2.25	0.0336
Weight	1	-0.07332	0.05360	-1.37	0.1836
Run_Pulse	1	-0.37071	0.11770	-3.15	0.0042
Maximum_Pulse	1	0.30603	0.13452	2.28	0.0317

- b. The ANOVA table did not change significantly. The  $R^2$  decreased slightly. The adjusted  $R^2$  increased again, confirming that the variable **Performance** did not contribute to explaining the variation in **Oxygen\_Consumption** when the other variables are in the model.
- c. The  $p$ -value for **Runtime** changed dramatically and is now less than 0.05. **Age**, **Run\_Pulse**, and **Maximum\_Pulse** also have  $p$ -values less than 0.05 as they did in the previous model.

## 6. Using All-Regression Techniques

- a. Here is the program for all regressions and the plot of Mallows'  $C_p=p$  and Hocking's reference line.

```
options ps=50 ls=97;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc reg data=sasuser.b_cars;
    ALL_REG: model midprice=citympg highwaympg enginesize
              horsepower rpm revolutions fueltanksize weight
              / selection=rsquare adjrsq cp best=5;
    plot cp.*np. /
          nomodel nostat
          vaxis=0 to 30 by 5
          haxis=2 to 8 by 1
          cmallows=red
          chocking=blue;
    symbol v=plus color=red h=2;
    title 'Best=5 Models Using All Regressions Option';
run;
quit;
```

## PROC REG Output

Best=5 Models Using All Regressions Option				
The REG Procedure				
Model: ALL_REG				
Dependent Variable: MidPrice				
R-Square Selection Method				
Number of Observations Read 92				
Number of Observations Used 92				
Number in Model      Adjusted R-Square      R-Square      C(p)      Variables in Model				
1	0.6735	0.6699	7.4674	HorsePower
1	0.4754	0.4696	65.4143	Weight
1	0.4468	0.4407	73.7725	FuelTankSize
1	0.4172	0.4107	82.4296	EngineSize
1	0.4075	0.4009	85.2687	CityMPG
-----				
2	0.6891	0.6821	4.9076	HorsePower Weight
2	0.6879	0.6809	5.2786	HorsePower FuelTankSize
2	0.6873	0.6803	5.4456	CityMPG HorsePower
2	0.6855	0.6785	5.9588	HighwayMPG HorsePower
2	0.6778	0.6705	8.2330	EngineSize HorsePower
Number in Model      Adjusted R-Square      R-Square      C(p)      Variables in Model				
3	0.7037	0.6936	2.6456	HorsePower Revolutions Weight
3	0.6984	0.6881	4.1941	CityMPG HorsePower Revolutions
3	0.6937	0.6833	5.5609	HorsePower Revolutions FuelTankSize
3	0.6914	0.6809	6.2491	HighwayMPG HorsePower Revolutions
3	0.6911	0.6806	6.3239	HorsePower RPM Weight
-----				
4	0.7095	0.6962	2.9396	CityMPG EngineSize HorsePower Revolutions
4	0.7082	0.6948	3.3271	CityMPG HorsePower Revolutions Weight
4	0.7063	0.6928	3.8831	EngineSize HorsePower Revolutions Weight
4	0.7047	0.6912	4.3478	HighwayMPG HorsePower Revolutions Weight
4	0.7039	0.6903	4.5910	HorsePower Revolutions FuelTankSize Weight
-----				
5	0.7125	0.6958	4.0724	CityMPG EngineSize HorsePower Revolutions Weight
5	0.7108	0.6940	4.5675	CityMPG EngineSize HorsePower RPM Revolutions
5	0.7105	0.6937	4.6468	CityMPG HighwayMPG HorsePower Revolutions Weight
5	0.7101	0.6933	4.7720	CityMPG HighwayMPG EngineSize HorsePower Revolutions
5	0.7099	0.6930	4.8448	CityMPG EngineSize HorsePower Revolutions FuelTankSize

## PROC REG Output (continued)

6	0.7143	0.6941	5.5443	CityMPG EngineSize HorsePower RPM Revolutions Weight
6	0.7141	0.6939	5.6047	CityMPG HighwayMPG EngineSize HorsePower Revolutions Weight
6	0.7127	0.6924	6.0158	CityMPG EngineSize HorsePower Revolutions FuelTankSize Weight
6	0.7114	0.6910	6.4109	CityMPG HighwayMPG EngineSize HorsePower RPM Revolutions
6	0.7112	0.6908	6.4592	CityMPG EngineSize HorsePower RPM Revolutions FuelTankSize
7	0.7159	0.6923	7.0665	CityMPG HighwayMPG EngineSize HorsePower RPM Revolutions Weight
7	0.7146	0.6908	7.4741	CityMPG EngineSize HorsePower RPM Revolutions FuelTankSize Weight
7	0.7143	0.6905	7.5515	CityMPG HighwayMPG EngineSize HorsePower Revolutions FuelTankSize Weight
7	0.7119	0.6879	8.2551	CityMPG HighwayMPG EngineSize HorsePower RPM Revolutions FuelTankSize
7	0.7106	0.6865	8.6284	CityMPG HighwayMPG HorsePower RPM Revolutions FuelTankSize Weight
8	0.7162	0.6888	9.0000	CityMPG HighwayMPG EngineSize HorsePower RPM Revolutions FuelTankSize Weight

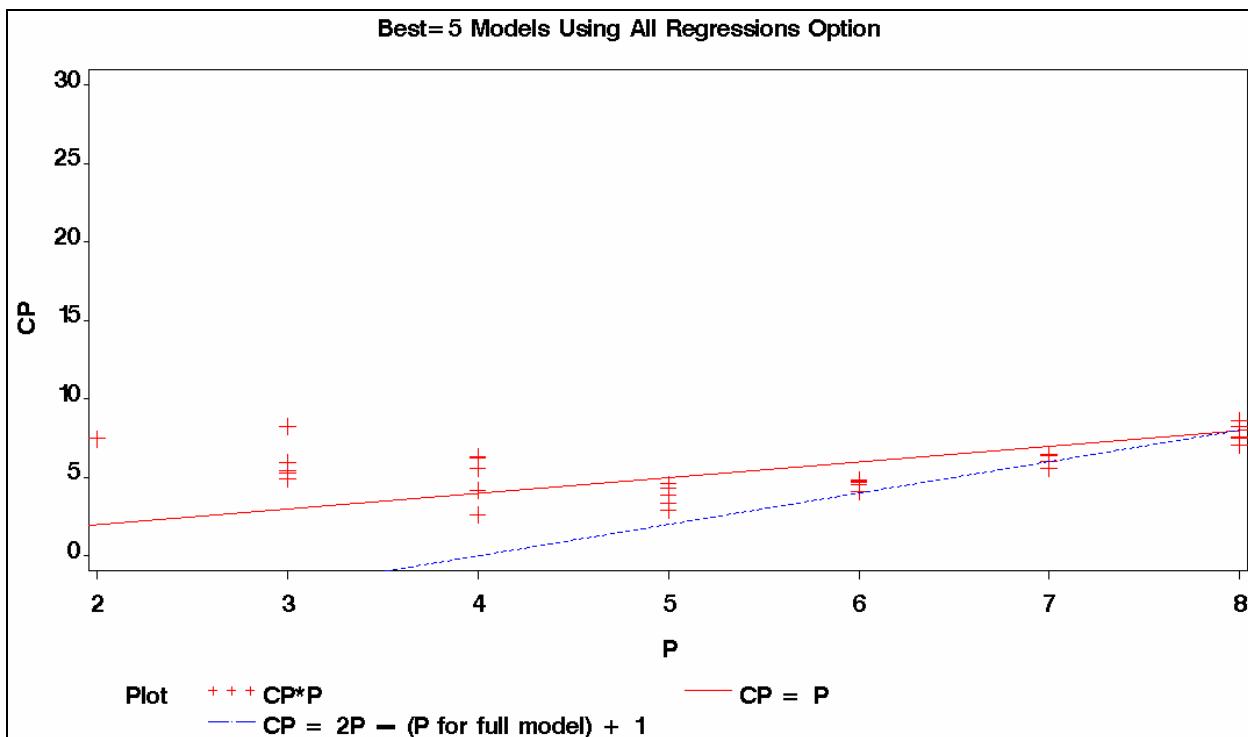
**MidPrice=HorsePower Revolutions Weight** is the smallest model where  $C_p < p$  ( $2.6456 < 4$ ), satisfying Mallow's  $C_p$ , identifying it as the "best" predictive model.

In this example  $p_{\text{full}}$  equals 9, 8 variables plus the intercept.

<b>p=# terms in the current model, including the intercept</b>	<b>Number in Model (k)</b>	<b>Minimum <math>C_p</math> with <math>p</math> terms</b>	<b>Hocking's Criterion : <math>2*p - 9 + 1 = 2*p - 8</math></b>	<b><math>C_p &lt;</math> Hocking's Criterion?</b>
4	3	2.6456	$2*4 - 8 = 0$	No
5	4	2.9396	$2*5 - 8 = 2$	No
6	5	4.0724	$2*6 - 8 = 4$	No
7	6	5.5443	$2*7 - 8 = 6$	Yes

**MidPrice=CityMPG EngineSize HorsePower RPM Revolutions Weight** is the smallest model that satisfies Hocking's criterion ( $5.5443 < 6$ ). This result is easier to see if you use the graph below.

## PROC GPLOT Output



- b. This program performs a stepwise regression.

```
options ps=50 ls=97;
proc reg data=sasuser.b_cars;
  STEPWISE: model midprice=citympg highwaympg enginesize
             horsepower rpm revolutions fueltanksize weight
             / selection=stepwise;
  title 'Selecting Models using SELECTION=STEPWISE';
run;
quit;
```

## PROC REG Output

Selecting Models using SELECTION=STEPWISE					
The REG Procedure					
Model: STEPWISE					
Dependent Variable: MidPrice					
Number of Observations Read 92					
Number of Observations Used 92					
Stepwise Selection: Step 1					
Variable HorsePower Entered: R-Square = 0.6735 and C(p) = 7.4674					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4558.23301	4558.23301	185.69	<.0001
Error	90	2209.31688	24.54797		
Corrected Total	91	6767.54989			
Parameter Standard					
Variable	Estimate	Error	Type II SS	F Value	Pr > F
Intercept	-0.38356	1.51673	1.56985	0.06	0.8009
HorsePower	0.13586	0.00997	4558.23301	185.69	<.0001
Bounds on condition number: 1, 1					
Stepwise Selection: Step 2					
Variable Weight Entered: R-Square = 0.6891 and C(p) = 4.9076					

## PROC REG Output (continued)

Selecting Models using SELECTION=STEPWISE					
The REG Procedure					
Model: STEPWISE					
Dependent Variable: MidPrice					
Stepwise Selection: Step 2					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4663.75669	2331.87835	98.65	<.0001
Error	89	2103.79320	23.63813		
Corrected Total	91	6767.54989			
Parameter Estimates					
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-5.42859	2.81367	87.99160	3.72	0.0569
HorsePower	0.11330	0.01448	1446.53699	61.20	<.0001
Weight	0.00270	0.00128	105.52368	4.46	0.0374
Bounds on condition number: 2.1914, 8.7656					
-----					
Stepwise Selection: Step 3					
Variable Revolutions Entered: R-Square = 0.7037 and C(p) = 2.6456					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4762.38708	1587.46236	69.67	<.0001
Error	88	2005.16281	22.78594		
Corrected Total	91	6767.54989			

## PROC REG Output (continued)

Selecting Models using SELECTION=STEPWISE									
The REG Procedure Model: STEPWISE Dependent Variable: MidPrice									
Stepwise Selection: Step 3									
<hr/>									
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F				
Intercept	-18.38406	6.81229	165.94481	7.28	0.0083				
HorsePower	0.11727	0.01435	1522.31498	66.81	<.0001				
Revolutions	0.00311	0.00149	98.63038	4.33	0.0404				
Weight	0.00437	0.00149	196.26395	8.61	0.0043				
<hr/>									
Bounds on condition number: 3.0942, 22.638									
<hr/>									
All variables left in the model are significant at the 0.1500 level.									
No other variable met the 0.1500 significance level for entry into the model.									
<hr/>									
Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value Pr > F		
1	HorsePower		1	0.6735	0.6735	7.4674	185.69 <.0001		
2	Weight		2	0.0156	0.6891	4.9076	4.46 0.0374		
3	Revolutions		3	0.0146	0.7037	2.6456	4.33 0.0404		

The output generated the model **MidPrice=HorsePower Weight Revolutions** using the SELECTION=STEPWISE option.

This program performs a backward regression:

```
proc reg data=sasuser.b_cars;
  BACKWARD: model midprice=citympg highwaympg enginesize
              horsepower rpm revolutions fueltanksize weight
              / selection=backward;
  title 'Selecting Models using SELECTION=BACKWARD';
run;
quit;
```

## PROC REG Output

## Selecting Models using SELECTION=BACKWARD

## The REG Procedure

Model: BACKWARD

Dependent Variable: MidPrice

Number of Observations Read 92  
Number of Observations Used 92

### Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7162 and C(p) = 9.0000

## Analysis of Variance

Source	DF	Sum of Squares		Mean Square	F Value	Pr > F
Model	8	4846.75465		605.84433	26.18	<.0001
Error	83	1920.79524		23.14211		
Corrected Total	91	6767.54989				

Variable	Parameter	Standard	Type	II	SS	F Value	Pr > F
	Estimate	Error					
Intercept	-21.30023	13.86601		54.60955		2.36	0.1283
CityMPG	-0.46278	0.33213		44.93125		1.94	0.1672
HighwayMPG	0.21335	0.30985		10.97213		0.47	0.4930
EngineSize	1.96681	1.54127		37.68488		1.63	0.2055
HorsePower	0.09518	0.02534		326.62826		14.11	0.0003
RPM	0.00125	0.00168		12.76225		0.55	0.4598
Revolutions	0.00501	0.00199		146.56253		6.33	0.0138
FuelTankSize	-0.09563	0.37077		1.53946		0.07	0.7971
Weight	0.00292	0.00261		29.04458		1.26	0.2658

Bounds on condition number: 13.795, 517.18

### Backward Elimination: Step 1

Variable FuelTankSize Removed: R-Square = 0.7159 and C(p) = 7.0665

## PROC REG Output (continued)

Selecting Models using SELECTION=BACKWARD					
The REG Procedure					
Model: BACKWARD					
Dependent Variable: MidPrice					
Backward Elimination: Step 1					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	4845.21519	692.17360	30.25	<.0001
Error	84	1922.33471	22.88494		
Corrected Total	91	6767.54989			
Parameter Estimates					
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-21.54301	13.75694	56.12017	2.45	0.1211
CityMPG	-0.44933	0.32618	43.42806	1.90	0.1720
HighwayMPG	0.21416	0.30810	11.05737	0.48	0.4889
EngineSize	1.91594	1.52009	36.35611	1.59	0.2110
HorsePower	0.09486	0.02516	325.20857	14.21	0.0003
RPM	0.00123	0.00167	12.45546	0.54	0.4627
Revolutions	0.00487	0.00191	149.11596	6.52	0.0125
Weight	0.00257	0.00220	31.11140	1.36	0.2469
Bounds on condition number: 13.455, 387.6					
-----					
Backward Elimination: Step 2					
Variable HighwayMPG Removed: R-Square = 0.7143 and C(p) = 5.5443					

## PROC REG Output (continued)

The REG Procedure Model: BACKWARD Dependent Variable: MidPrice						
Backward Elimination: Step 2						
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	6	4834.15781	805.69297	35.42	<.0001	
Error	85	1933.39208	22.74579			
Corrected Total	91	6767.54989				
Parameter Standard						
Variable	Estimate	Error	Type II SS	F Value	Pr > F	
Intercept	-17.93412	12.70083	45.35203	1.99	0.1616	
CityMPG	-0.25861	0.17585	49.19572	2.16	0.1451	
EngineSize	2.01499	1.50878	40.56890	1.78	0.1853	
HorsePower	0.09549	0.02507	329.99320	14.51	0.0003	
RPM	0.00122	0.00167	12.22170	0.54	0.4656	
Revolutions	0.00458	0.00186	138.48985	6.09	0.0156	
Weight	0.00216	0.00211	23.67836	1.04	0.3105	
Bounds on condition number: 9.8797, 205.92						
-----						
Backward Elimination: Step 3						
Variable RPM Removed: R-Square = 0.7125 and C(p) = 4.0724						
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	4821.93612	964.38722	42.63	<.0001	
Error	86	1945.61377	22.62342			
Corrected Total	91	6767.54989				

## PROC REG Output (continued)

## Selecting Models using SELECTION=BACKWARD

## The REG Procedure

Model: BACKWARD

Dependent Variable: MidPrice

### Backward Elimination: Step 3

Variable	Parameter	Standard	Type I	SS	F Value	Pr > F
	Estimate	Error				
Intercept	-11.79750	9.52549		34.70274	1.53	0.2189
CityMPG	-0.23444	0.17226		41.90233	1.85	0.1771
EngineSize	1.29829	1.14602		29.03493	1.28	0.2604
HorsePower	0.11013	0.01511		1201.97296	53.13	<.0001
Revolutions	0.00465	0.00185		142.84618	6.31	0.0138
Weight	0.00197	0.00209		20.06682	0.89	0.3489

Bounds on condition number: 6.1645, 108.04

### Backward Elimination: Step 4

Variable Weight Removed: R-Square = 0.7095 and C(p) = 2.9396

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4801.86930	1200.46732	53.13	<.0001
Error	87	1965.68060	22.59403		
Corrected Total	91	6767.54989			

Variable	Parameter	Standard	Estimate			F Value	Pr > F
	Estimate	Error	Type I	II	SS		
Intercept	-5.51825	6.79871			14.88484	0.66	0.4192
CityMPG	-0.33106	0.13829			129.48763	5.73	0.0188
EngineSize	1.82506	0.99961			75.31613	3.33	0.0713
HorsePower	0.11240	0.01491			1284.67376	56.86	<.0001
Revolutions	0.00473	0.00185			148.79669	6.59	0.0120

Bounds on condition number: 4.3657, 50.637

## PROC REG Output (continued)

Selecting Models using SELECTION=BACKWARD							
The REG Procedure Model: BACKWARD Dependent Variable: MidPrice							
Backward Elimination: Step 4							
All variables left in the model are significant at the 0.1000 level.							
Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	FuelTankSize	7	0.0002	0.7159	7.0665	0.07	0.7971
2	HighwayMPG	6	0.0016	0.7143	5.5443	0.48	0.4889
3	RPM	5	0.0018	0.7125	4.0724	0.54	0.4656
4	Weight	4	0.0030	0.7095	2.9396	0.89	0.3489

Using the SELECTION=BACKWARD option, the following model was generated  
**MidPrice=CityMPG EngineSize HorsePower Revolutions.**

c.

Mallows (prediction)	<b>MidPrice=HorsePower Revolutions Weight</b>
Hocking (explanatory)	<b>MidPrice=</b> <b>CityMPG EngineSize HorsePower RPM Revolutions Weight</b>
STEPWISE	<b>MidPrice=HorsePower Weight Revolutions</b>
BACKWARD	<b>MidPrice= CityMPG EngineSize HorsePower Revolutions</b>

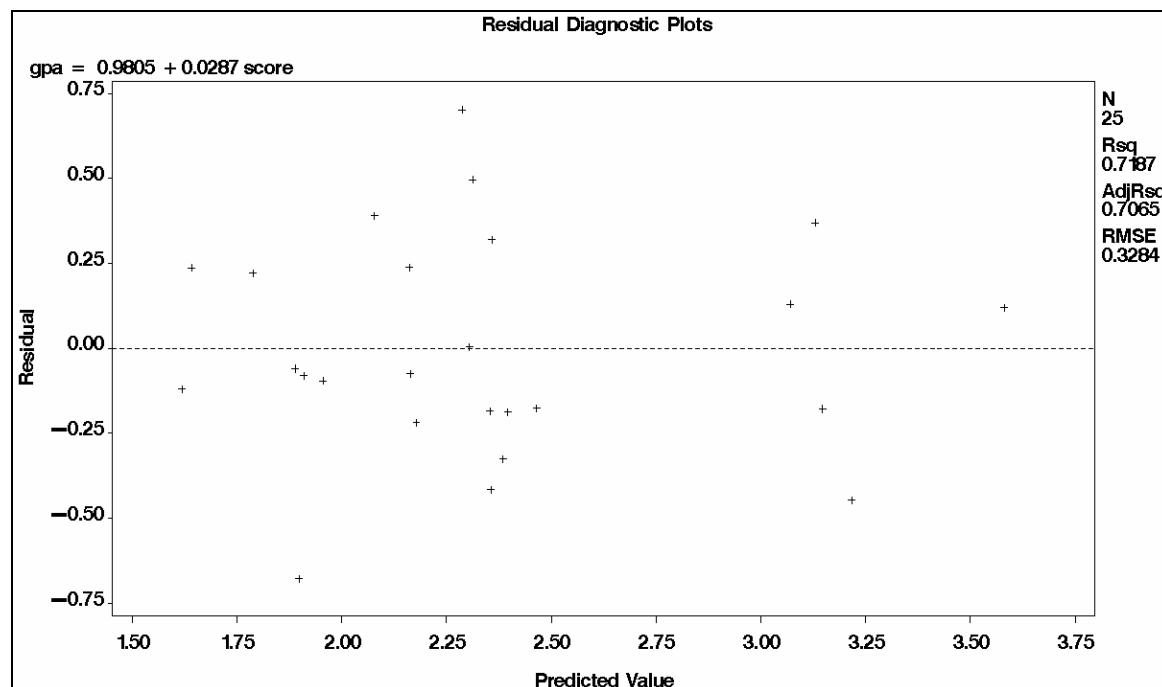
## Chapter 4

### 1. Examining Residuals

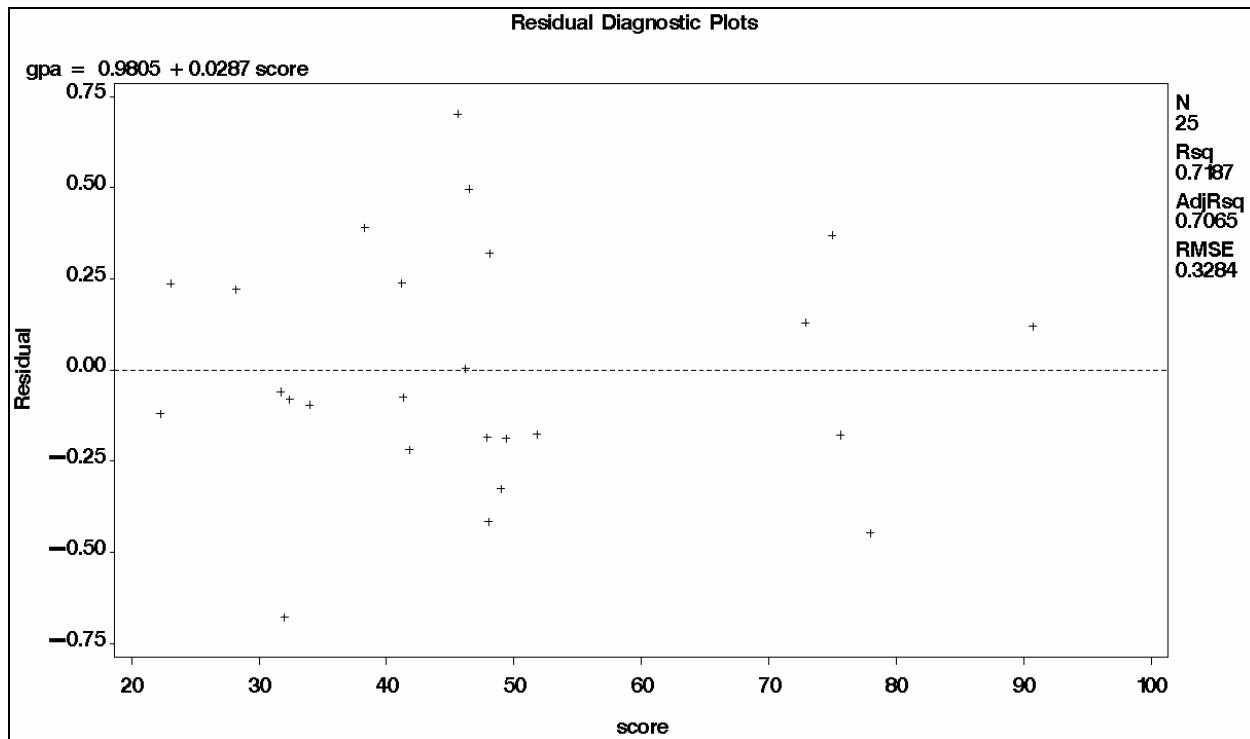
- a. Use the REG procedure with three PLOT statements to create an ordinary residual plot, a student residual plot, and a normal quantile-quantile plot.

```
options ps=50 ls=97;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc reg data=sasuser.b_grades;
  model gpa=score;
  plot r.*(p. score);
  plot student.*obs. / vref=3 2 -2 -3
    haxis=1 to 25 by 1;
  plot student.*nqq.;
  title 'Residual Diagnostic Plots';
run;
quit;
```

Partial PROC REG Output

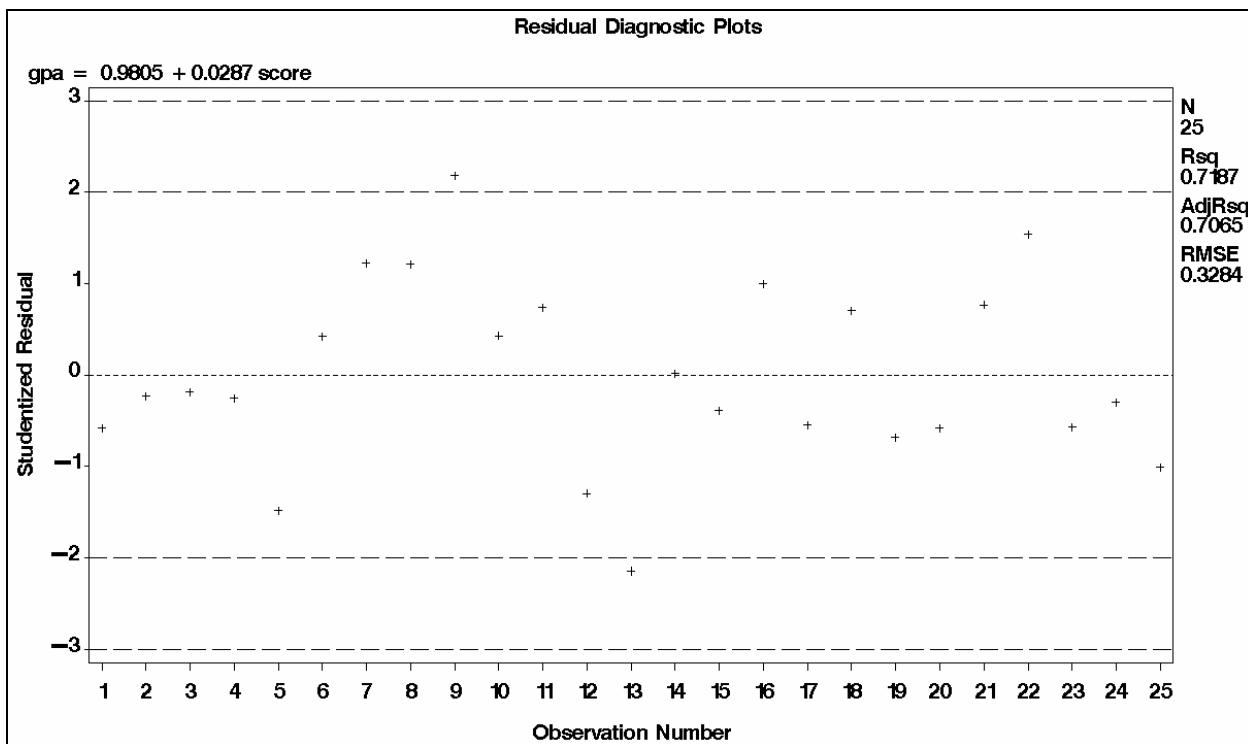


## Partial PROC REG Output (continued)



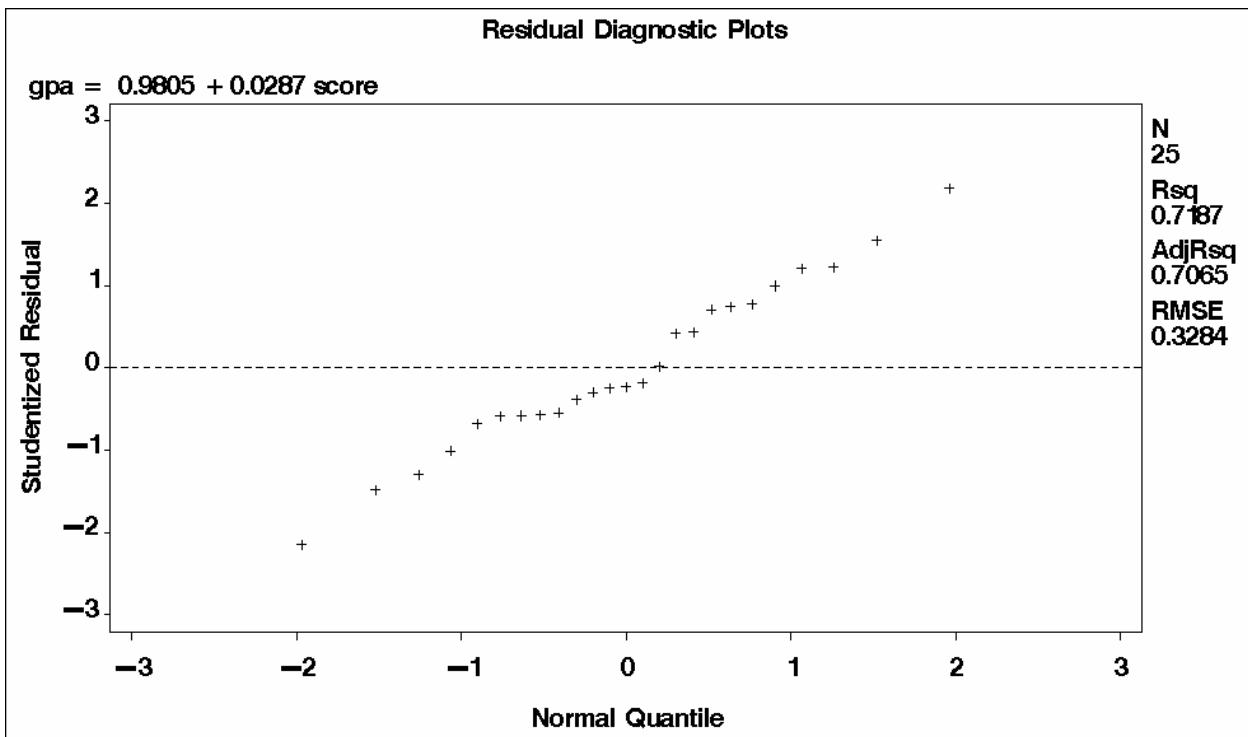
- 1) Inspection of the Residuals versus Predicted values and Residuals versus Independent Variable (**score**) plots shows no problems with the model assumptions.

## Partial PROC REG Output (continued)



- 2) Observation numbers 9 and 13 should be investigated as outliers.

## Partial PROC REG Output (continued)



- 3) The normal quantile-quantile plot shows no serious departures from normality.

## 2. Generating Potential Outliers

a.

- 1) Demonstrations **c4demo02a**, **c4demo02b**, and **c4demo02c** were copied and modified.

The data set **sasuser.b\_cars** was designated in the PROC REG statement. The NOPRINT option was also added to this statement. The macro variables **numparms**, **numobs** and **idvars** were changed to reflect **sasuser.b\_cars**.

PROC PRINT displays only the observations that exceed one or more of the cutoffs.

```
options reset=all;
proc reg data=sasuser.b_cars
  noprint;
model midprice=citympg enginesize horsepower revolutions
  / r influence;
output out=ck4outliers
  rstudent=rstud dfits=dfits cookd=cooksd;
title;
run;
quit;

/* set the values of these macro variables, */
/* based on your data and model.          */
%let numparms=5; /* 4 + 1 */
%let numobs=92; /* # of observations */
%let idvars=manufacturer model;
data influential;
  set ck4outliers;

  cutdfits=2*(sqrt(&numparms/&numobs));
  cutcookd=4/&numobs;

  rstud_i=(abs(rstud)>3);
  dfits_i=(abs(dfits)>cutdfits);
  cookd_i=(cooksd>cutcookd);
  sum_i=rstud_i + dfits_i + cookd_i;
  if sum_i > 0;
run;

proc print data=influential;
  var sum_i &idvars cooksd rstud dfits cutcookd cutdfits
    cookd_i rstud_i dfits_i;
  title 'Observations that Exceed Suggested Cutoffs';
run;
```

## PROC PRINT Output

Observations that Exceed Suggested Cutoffs												
M							c	c				
a							u	u	c	r	d	
n							t	t	o	s	f	
u							c	d	o	t	i	
f							o	f	k	u	t	
a							i	d	d	d	s	
c			c				t	t	-	-	-	
s t	M	o	r	d	c	d	o	t	i			
u u	o	o	s	f	o	f	k	u	t			
O m r	d	k	t	i	o	i	d	d	s			
b _ e	e	s	u	t	k	t	-	-	-			
s i r	l	d	d	s	d	s	i	i	i			
1 2 Audi	100	0.05300	2.91195	0.53643	0.043478	0.46625	1	0	1			
2 2 Dodge	Stealth	0.39757	-2.93386	-1.47027	0.043478	0.46625	1	0	1			
3 2 Geo	Metro	0.05203	0.82832	0.50913	0.043478	0.46625	1	0	1			
4 2 Honda	Civic	0.08233	1.15201	0.64279	0.043478	0.46625	1	0	1			
5 2 Infiniti	Q45	0.12779	2.36023	0.82009	0.043478	0.46625	1	0	1			
6 1 Lincoln	Continental	0.04265	2.61526	0.47703	0.043478	0.46625	0	0	1			
7 2 Mercedes-Benz	190E	0.04107	3.06459	0.47451	0.043478	0.46625	0	1	1			

Analysis of the output above reveals two basic types of cars that appear to be influential: most are luxury cars, such as the Mercedes-Benz, but a few are economy cars, such as the Geo Metro.

### 3. Ascertaining Collinearity

- a. The MODEL statement options VIF, COLLIN, and COLLINOINT produce the desired diagnostics. FULL43 is a label for this model.

```
proc reg data=sasuser.b_cars;
  FULL43: model midprice=citympg highwaympg enginesize
            horsepower RPM revolutions
            FuelTankSize weight
            / vif collin collinoint;
  title 'FULL43 - Collinearity Diagnostics for sasuser.b_cars';
run;
quit;
```

PROC REG Output

FULL43 - Collinearity Diagnostics for sasuser.b_cars					
The REG Procedure					
Model: FULL43					
Dependent Variable: MidPrice					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	4846.75465	605.84433	26.18	<.0001
Error	83	1920.79524	23.14211		
Corrected Total	91	6767.54989			
Analysis of Variance					
Root MSE                  4.81062      R-Square          0.7162					
Dependent Mean          19.04891      Adj R-Sq        0.6888					
Coeff Var                25.25406					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-21.30023	13.86601	-1.54	0.1283
CityMPG	1	-0.46278	0.33213	-1.39	0.1672
HighwayMPG	1	0.21335	0.30985	0.69	0.4930
EngineSize	1	1.96681	1.54127	1.28	0.2055
HorsePower	1	0.09518	0.02534	3.76	0.0003
RPM	1	0.00125	0.00168	0.74	0.4598
Revolutions	1	0.00501	0.00199	2.52	0.0138
FuelTankSize	1	-0.09563	0.37077	-0.26	0.7971
Weight	1	0.00292	0.00261	1.12	0.2658
Variance Inflation					

- 1) The variance inflation factors for **CityMPG**, **HighwayMPG**, and **EngineSize** all exceed 10, indicating the presence of multicollinearity.

## Partial PROC REG Output (continued)

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation-----			
			Intercept	CityMPG	HighwayMPG	EngineSize
1	8.56748	1.00000	0.00001775	0.00005398	0.00003896	0.00015826
2	0.34165	5.00766	0.00002040	0.00242	0.00092131	0.00981
3	0.04238	14.21803	0.00010723	0.00597	0.00440	0.10287
4	0.02959	17.01556	0.00094549	0.01676	0.00979	0.01302
5	0.00825	32.22765	0.00488	0.00056451	0.01389	0.21469
6	0.00453	43.50895	0.03842	0.10194	0.01032	0.16026
7	0.00314	52.20546	0.00151	0.09013	0.01363	0.19335
8	0.00206	64.55296	0.01999	0.67656	0.69449	0.12445
9	0.00091723	96.64673	0.93411	0.10559	0.25253	0.18141

Collinearity Diagnostics						
Number	HorsePower	RPM	Revolutions	Proportion of Variation-----		
				FuelTank	Size	Weight
1	0.00021121	0.00004222	0.00014346	0.00008444	0.00005034	
2	0.00891	0.00021349	0.00447	0.00086107	0.00057418	
3	0.14274	0.00911	0.01386	0.00006764	0.00054570	
4	0.14581	0.00012552	0.04576	0.04919	0.01662	
5	0.03055	0.04123	0.75147	0.02732	0.03674	
6	0.12733	0.16212	0.00664	0.56496	0.01328	
7	0.12317	0.14514	0.01017	0.33940	0.63917	
8	0.10002	0.18080	0.04987	0.01478	0.00292	
9	0.32126	0.46122	0.11761	0.00334	0.29009	

- 2) The largest condition index in the Collinearity Diagnostics table (96.64673) indicates moderate collinearity. You decide that value is considered unacceptable, based on your knowledge. Examining the Proportion of Variation columns associated with the largest condition index reveals that the intercept (0.93411) is involved in collinearity.

Therefore, turn your attention to the Collinearity Diagnostics (intercept adjusted) table.

Collinearity Diagnostics (intercept adjusted)						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			CityMPG	HighwayMPG	EngineSize	HorsePower
1	5.65062	1.00000	0.00191	0.00218	0.00252	0.00280
2	1.08877	2.27814	0.00056273	0.00157	0.00228	0.02868
3	0.59825	3.07332	0.01721	0.04653	0.02786	0.04183
4	0.32062	4.19809	0.02442	0.02101	0.00969	0.00275
5	0.15553	6.02764	0.00796	0.02836	0.06053	0.19789
6	0.08466	8.16953	0.03582	0.00864	0.00365	0.01637
7	0.05925	9.76544	0.00586	0.07711	0.86581	0.63997
8	0.04229	11.55856	0.90626	0.81460	0.02766	0.06970

Collinearity Diagnostics (intercept adjusted)				
Number	RPM	Revolutions	Proportion of Variation	
			FuelTank	Size
1	0.00177	0.00554	0.00433	0.00304
2	0.15485	0.01003	0.00259	0.00006437
3	0.00513	0.07715	0.00730	0.00015755
4	0.02437	0.30840	0.15850	0.02322
5	0.12978	0.38994	0.28363	0.03776
6	0.02221	0.01514	0.49895	0.85564
7	0.61231	0.03598	0.00109	0.02267
8	0.04957	0.15782	0.04361	0.05744

Look at this table's last line, and note that **CityMPG** (0.90626) and **HighwayMPG** (0.81460) are collinear.

Returning to the Parameter Estimates table, observe that **HighwayMPG** has a larger *p*-value (0.4930) than **CityMPG** (0.1672) but both variables had high variance inflation factors (13.79500 and 10.77958, respectively).

Based on the statistical methods you have studied thus far, generate another model, but remove **HighwayMPG**.

The reduced model is given the label NOHWYMPG to indicate that this model does not include the variable **HighwayMPG**.

```

proc reg data=sasuser.b_cars;
  NOHWYMPG: model midprice=citympg enginesize
              horsepower RPM revolutions
              fueltanksize weight
              / vif collin collinoint;
  title 'NOHWYMPG - Collinearity Diagnostics for sasuser.b_cars';
run;
quit;

```

## PROC REG Output

NOHWYMPG - Collinearity Diagnostics for sasuser.b_cars						
The REG Procedure						
Model: NOHWYMPG						
Dependent Variable: MidPrice						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	7	4835.78252	690.82607	30.04	<.0001	
Error	84	1931.76738	22.99723			
Corrected Total	91	6767.54989				
Analysis of Variance						
Root MSE                  4.79554          R-Square          0.7146						
Dependent Mean          19.04891          Adj R-Sq        0.6908						
Coeff Var                25.17489						
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	
Variance Inflation						
Intercept	1	-17.69884	12.80148	-1.38	0.1705	0
CityMPG	1	-0.27318	0.18511	-1.48	0.1438	4.31235
EngineSize	1	2.06686	1.52960	1.35	0.1802	10.04317
HorsePower	1	0.09582	0.02524	3.80	0.0003	6.84018
RPM	1	0.00124	0.00168	0.74	0.4624	4.00087
Revolutions	1	0.00472	0.00194	2.43	0.0171	3.70630
FuelTankSize	1	-0.09824	0.36959	-0.27	0.7910	5.85656
Weight	1	0.00253	0.00254	1.00	0.3223	8.89700

The variance inflation factor for **EngineSize** is the only one greater than 10.

PROC REG Output (continued)

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	CityMPG	EngineSize	HorsePower
1	7.61605	1.00000	0.00002612	0.00021510	0.00020476	0.00027108
2	0.30210	5.02100	0.00007627	0.01143	0.01113	0.00917
3	0.03995	13.80774	0.00055015	0.01974	0.09259	0.21419
4	0.02542	17.31005	0.00032418	0.17086	0.05873	0.07367
5	0.00784	31.15814	0.00879	0.05756	0.16307	0.03900
6	0.00445	41.35372	0.04200	0.59807	0.16037	0.11615
7	0.00311	49.46543	0.00118	0.09180	0.21365	0.13885
8	0.00107	84.35927	0.94706	0.05033	0.30025	0.40869

Collinearity Diagnostics					
Number	RPM	Revolutions	Proportion of Variation		
			FuelTank	Size	Weight
1	0.00005320	0.00018792	0.00010759	0.00006743	
2	0.00046346	0.00766	0.00067031	0.00049675	
3	0.00799	0.00253	0.00101	0.00245	
4	0.00000799	0.03733	0.06339	0.01849	
5	0.05858	0.84689	0.01110	0.03942	
6	0.14065	0.02284	0.54309	0.00868	
7	0.17198	0.02496	0.37842	0.68907	
8	0.62027	0.05760	0.00221	0.24133	

Note that the largest condition index is 84.35927, and according to the guidelines, this indicates a moderate amount of collinearity. However, **EngineSize**'s VIF is 10.04317, which exceeds the VIF cutoff of 10, and **Weight**'s VIF is close at 8.89700.

It was decided to continue the process to try to reduce the model. The values on the last line of the Proportion of Variation columns indicate that **Intercept** (0.94706) and **RPM** (0.62027) are collinear. Because the intercept is involved, the Collinearity Diagnostics (intercept adjusted) table is consulted.

## PROC REG Output (continued)

Collinearity Diagnostics (intercept adjusted)						
Number	Eigenvalue	Condition	Proportion of Variation-----			
		Index	CityMPG	EngineSize	HorsePower	RPM
1	4.94069	1.00000	0.00745	0.00352	0.00373	0.00250
2	1.07301	2.14582	0.00140	0.00112	0.03533	0.15661
3	0.43562	3.36776	0.07248	0.02392	0.02994	0.00748
4	0.26828	4.29143	0.37285	0.06201	0.03654	0.03214
5	0.14145	5.91010	0.35903	0.01946	0.18615	0.13785
6	0.08382	7.67742	0.01982	0.01150	0.00706	0.01057
7	0.05715	9.29827	0.16697	0.87847	0.70125	0.65287

Collinearity Diagnostics (intercept adjusted)			
Number	Revolutions	Proportion of Variation-----	
		FuelTank	Weight
1	0.00798	0.00558	0.00417
2	0.00677	0.00330	0.00022938
3	0.22143	0.09707	0.01337
4	0.23332	0.03586	0.00489
5	0.39102	0.28337	0.04411
6	0.02931	0.55418	0.93114
7	0.11017	0.02063	0.00209

Based on their Proportion of Variation statistics, **EngineSize** (0.87847), **HorsePower** (0.70125), and **RPM** (0.65287) are all involved in collinearity. Returning to the Parameter Estimates table, you record the *p*-value and variance inflation factor for each variable:

<u>Variable</u>	<u>p-value</u>	<u>VIF</u>
<b>EngineSize</b>	0.1802	10.04317
<b>HorsePower</b>	0.0003	6.84018
<b>RPM</b>	0.4624	4.00087

It is clear that **HorsePower** should stay in the model. It is your decision as to whether to delete **EngineSize** (large variance inflation factor) or **RPM** (largest *p*-value) from the model.

## Chapter 5

### 1. Performing Tests and Measures of Association

- a. Examine the **sasuser.b\_safety** data set using the PRINT procedure. Invoke the FREQ procedure and create one-way frequency tables for the variables **region**, **safety**, and **type**.

```
proc print data=sasuser.b_safety;
  var type region safety weight;
run;

proc freq data=sasuser.b_safety;
  tables region safety type;
run;
```

Partial PROC PRINT Output

Obs	type	region	safety	weight
1	Medium	N America	0	3.395
2	Sport/Utility	N America	0	4.180
3	Medium	N America	0	3.145
4	Small	N America	0	2.600
5	Medium	N America	0	3.085
6	Medium	N America	0	2.910
7	Sport/Utility	N America	0	4.180
8	Medium	Asia	0	3.415
9	Medium	N America	0	3.995
10	Small	N America	0	2.600
11	Small	N America	1	2.765
12	Small	Asia	0	2.665
13	Medium	N America	0	3.100
14	Medium	N America	0	3.455
15	Medium	N America	0	3.055
16	Large	N America	0	3.450
17	Large	N America	0	3.640
18	Large	N America	0	4.195
19	Large	N America	0	3.985
20	Large	N America	0	4.480

- 1) The measurement scale of each variable is

<u>Variable</u>	<u>Measurement Scale</u>
<b>safety</b>	<u>ORDINAL</u>
<b>type</b>	<u>NOMINAL</u>
<b>region</b>	<u>NOMINAL</u>
<b>weight</b>	<u>CONTINUOUS</u>

- 2) Examine the PROC FREQ output. The proportion of cars built in North America is 0.6354.

The FREQ Procedure				
region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	35	36.46	35	36.46
N America	61	63.54	96	100.00
safety	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	66	68.75	66	68.75
1	30	31.25	96	100.00
type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Large	16	16.67	16	16.67
Medium	29	30.21	45	46.88
Small	20	20.83	65	67.71
Sport/Utility	16	16.67	81	84.38
Sports	15	15.63	96	100.00

- 3) There are no unusual data values that warrant further attention.

- b. Use PROC FREQ to examine the crosstabulation of the variables **safety** by **region**. Generate a temporary format to clearly identify the values of **safety**.

```

proc format;
  value safdesc 0='Average or Above'
            1='Below Average';
run;

proc freq data=sasuser.b_safety;
  tables region*safety / expected cellchi2;
  format safety safdesc.;
run;

```

Table of region by safety			
region	safety		
	Frequency	Expected	
	Cell Chi-Square		
	Percent		
	Row Pct		
	Col Pct	Average or Above	Below Average
Asia	20 24.063 0.6859 20.83 57.14 30.30	15 10.938 1.5089 15.63 42.86 50.00	35 36.46
N America	46 41.938 0.3935 47.92 75.41 69.70	15 19.063 0.8658 15.63 24.59 50.00	61 63.54
Total	66 68.75	30 31.25	96 100.00

- 1) For the cars made in Asia, 42.86% had a below-average safety score.
- 2) For the cars with an average or above safety score, 69.70% were made in North America.
- 3) Yes, there seems to be an association between **region** and **safety**. A higher percentage (75.41 versus 57.14) of cars from North America had a higher safety rating.
- 4) The cell where **region** is Asia and **safety** is Below Average contributed the most to any possible association.

- c. Use PROC FREQ with the CHISQ option to perform a chi-square test of association between **region** and **safety**.

```
proc freq data=sasuser.b_safety;
  tables region*safety / chisq;
  format safety safdesc.;
run;
```

Partial PROC FREQ output.

Statistics for Table of region by safety			
Statistic	DF	Value	Prob
Chi-Square	1	3.4541	0.0631
Likelihood Ratio Chi-Square	1	3.3949	0.0654
Continuity Adj. Chi-Square	1	2.6562	0.1031
Mantel-Haenszel Chi-Square	1	3.4181	0.0645
Phi Coefficient		-0.1897	
Contingency Coefficient		0.1864	
Cramer's V		-0.1897	

- 1) The *p*-value represents the probability of observing a chi-square value at least as large as the one actually observed, given that the null hypothesis is true.
  - 2) You fail to reject the null hypothesis that there is **not** an association.
- d. Create a new variable named **size**. Assign a 1 for **type** equal to Small or Sports, 2 for **type** equal to Medium, and 3 for **type** equal to Large or Sport/Utility. Examine the ordinal association between **size** and **safety** using PROC FREQ.

```
data sasuser.b_safet2;
  set sasuser.b_safety;
  size=1*(type='Sports' or type='Small') +
    2*(type='Medium') +
    3*(type='Large' or type='Sport/Utility');
run;

proc format;
  value sizename 1='Small'
    2='Medium'
    3='Large';
run;

proc freq data=sasuser.b_safet2;
  tables size*safety / chisq measures cl;
  format safety safdesc.
    size sizename. ;
run;
```

Table of size by safety

size	safety		
	Frequency	Percent	Row Pct
	Average or Above	Below Average	Total
Small	12	23	35
	12.50	23.96	36.46
	34.29	65.71	
	18.18	76.67	
Medium	24	5	29
	25.00	5.21	30.21
	82.76	17.24	
	36.36	16.67	
Large	30	2	32
	31.25	2.08	33.33
	93.75	6.25	
	45.45	6.67	
Total	66	30	96
	68.75	31.25	100.00

Statistics for Table of size by safety

Statistic	DF	Value	Prob
Chi-Square	2	31.3081	<.0001
Likelihood Ratio Chi-Square	2	32.6199	<.0001
Mantel-Haenszel Chi-Square	1	27.7098	<.0001
Phi Coefficient		0.5711	
Contingency Coefficient		0.4959	
Cramer's V		0.5711	

Statistics for Table of size by safety				
Statistic	Value	ASE	95% Confidence Limits	
Gamma	-0.8268	0.0796	-0.9829	-0.6707
Kendall's Tau-b	-0.5116	0.0726	-0.6540	-0.3693
Stuart's Tau-c	-0.5469	0.0866	-0.7166	-0.3771
Somers' D C R	-0.4114	0.0660	-0.5408	-0.2819
Somers' D R C	-0.6364	0.0860	-0.8049	-0.4678
Pearson Correlation	-0.5401	0.0764	-0.6899	-0.3903
Spearman Correlation	-0.5425	0.0769	-0.6932	-0.3917
Lambda Asymmetric C R	0.3667	0.1569	0.0591	0.6743
Lambda Asymmetric R C	0.2951	0.0892	0.1203	0.4699
Lambda Symmetric	0.3187	0.0970	0.1286	0.5088
Uncertainty Coefficient C R	0.2735	0.0836	0.1096	0.4374
Uncertainty Coefficient R C	0.1551	0.0490	0.0590	0.2512
Uncertainty Coefficient Symmetric	0.1979	0.0615	0.0773	0.3186

Sample Size = 96

- 1) You should use the Mantel-Haenszel test to detect an ordinal association between **size** and **safety**.
- 2) You reject the null hypothesis that there is **not** an ordinal association.
- 3) The Spearman correlation statistic indicates that an ordinal association of moderate strength exists (-0.5425) between **size** and **safety**.
- 4) The 95% confidence interval around that statistic is (-0.6932, -0.3917).

## 2. Performing a Logistic Regression Analysis

- a. Fit a simple logistic regression model with **safety** as the outcome variable and **weight** as the predictor variable. Use the EVENT= option to model the probability of below-average safety scores.

```
proc logistic data=sasuser.b_safety;
  model safety(event='1')=weight;
run;
```

### PROC LOGISTIC Output

#### The LOGISTIC Procedure

##### Model Information

Data Set	SASUSER.B_SAFETY
Response Variable	safety
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	96
Number of Observations Used	96

##### Response Profile

Ordered Value	safety	Total Frequency
1	0	66
2	1	30

Probability modeled is safety=1.

##### Model Convergence Status

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

##### Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	121.249	94.614
SC	123.813	99.743
-2 Log L	119.249	90.614

##### Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.6344	1	<.0001
Score	21.3147	1	<.0001
Wald	16.9690	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	6.2896	1.6749	14.1015	0.0002
weight	1	-2.2942	0.5569	16.9690	<.0001
Odds Ratio Estimates					
Effect		Point Estimate	95% Confidence Limits	Wald	Pr > ChiSq
		weight	0.101	0.034	0.300
Association of Predicted Probabilities and Observed Responses					
Percent Concordant		83.6	Somers' D	0.674	
Percent Discordant		16.2	Gamma	0.675	
Percent Tied		0.2	Tau-a	0.293	
Pairs		1980	c	0.837	

- 1) You reject the null hypothesis that all regression coefficients of the model are 0 because the *p*-value of Likelihood Ratio statistic is less than 0.0001.
  - 2) The logistic regression equation is  $\text{logit}(\hat{p}) = 6.2896 - 2.2942 * \text{weight}$ , where  $\hat{p}$  is the predicted probability of having a below-average safety score.
  - 3) The odds ratio for **weight** means that vehicles 1,000 pounds heavier are 0.101 times more likely, with respect to odds, to have a below-average safety score compared to vehicles 1,000 pounds lighter.
  - 4) The 95% confidence interval indicates that you are 95% confident that the odds ratio for your population is within the interval 0.034 through 0.300.
  - 5) A concordant pair occurs when the observation with the outcome (in this case below-average safety score) has a higher predicted outcome probability (based on the model) than the observation without the outcome (average or above average safety scores). For all pairs of observations with different outcomes, 83.7% were concordant.
- b. Fit a multiple logistic regression model with **safety** as the outcome variable and **weight** and **region** as the predictor variables. Use the EVENT= option to model the probability of below-average safety scores. Specify **region** as a classification variable using reference cell coding and specify **Asia** as the reference level. Also request the 95% profile likelihood confidence intervals.

```
proc logistic data=sasuser.b_safety;
  class region (param=ref ref='Asia');
  model safety(event='1') = region weight / clodds=pl;
run;
```

## PROC LOGISTIC Output

## The LOGISTIC Procedure

## Model Information

Data Set	SASUSER.B_SAFETY
Response Variable	safety
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	96
Number of Observations Used	96

## Response Profile

Ordered Value	safety	Total Frequency
1	0	66
2	1	30

Probability modeled is safety=1.

## Class Level Information

Class	Value	Design Variables
region	Asia	0
	N America	1

## Model Convergence Status

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

## Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	121.249	95.712
SC	123.813	103.405
-2 Log L	119.249	89.712

## Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	29.5365	2	<.0001
Score	22.2730	2	<.0001
Wald	17.5015	2	0.0002

## Type 3 Analysis of Effects

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
region	1	0.9060	0.3412
weight	1	15.8831	<.0001

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard		Wald	
		Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	6.3808	1.6838	14.3599	0.0002
region N America	1	-0.5007	0.5260	0.9060	0.3412
weight	1	-2.2275	0.5589	15.8831	<.0001

## Odds Ratio Estimates

Effect		Point	95% Wald	
		Estimate	Confidence	Limits
region N America vs Asia		0.606	0.216	1.699
weight		0.108	0.036	0.322

## Association of Predicted Probabilities and Observed Responses

Percent Concordant	83.4	Somers' D	0.671
Percent Discordant	16.3	Gamma	0.673
Percent Tied	0.3	Tau-a	0.291
Pairs	1980	c	0.836

## Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
region N America vs Asia		1.0000	0.606	0.214
weight		1.0000	0.108	0.032

- 1) The parameter estimate for **region** compares the logit for North America to the logit for Asia controlling for **weight**. Because the *p*-value is above 0.05, you fail to reject the null hypothesis and state that there is no statistical evidence of a difference between North America and Asia with regard to below-average safety scores when controlling for **weight**.
- 2) Because the AIC and SC values increased for this model compared to the model with just **weight**, the addition of **region** did not improve the fit of the model. Therefore, the model with just **weight** is a better fitting model. However, the nonsignificance of **region** might have subject matter importance.
- 3) The profile likelihood confidence intervals are different than the Wald-based confidence intervals because the Wald confidence intervals use a normal approximation, whereas the profile likelihood confidence intervals are based on the value of the log-likelihood. The profile likelihood confidence intervals are generally preferred to the Wald confidence intervals, especially for sample sizes less than 50.
- 4) The *c* statistic estimates the probability of an observation with the outcome having a higher predicted probability than an observation without the outcome. With a *c* statistic of .836, you can state that you have a 0.836 probability that cars with the event (below-average safety scores) have a higher predicted probability compared to cars without the event.

# Appendix B Sampling from SAS Data Sets

<b>B.1 Random Samples .....</b>	<b>B-2</b>
---------------------------------	------------

## B.1 Random Samples

### Selecting Random Samples

The SURVEYSELECT procedure selects a random sample from a SAS data set.

```
PROC SURVEYSELECT DATA=name-of-SAS-data-set
                    OUT=name-of-output-data-set
                    METHOD=method-of-random-sampling
                    SEED=seed-value
                    SAMPSIZE=number of observations desired in
                               sample
                    ;
<STRATA stratification-variable(s)>;
RUN;
```

Selected PROC SURVEYSELECT statement options:

- DATA= identifies the data set to be selected from.
- OUT= indicates the name of the output data set.
- METHOD= specifies the random sampling method to be used. For simple random sampling without replacement, use METHOD=SRS. For simple random sampling with replacement, use METHOD=URS. For other selection methods and details on sampling algorithms, see the SAS online documentation for PROC SURVEYSELECT.
- SEED= specifies the initial seed for random number generation. If no SEED option is specified, SAS uses the system time as its seed value. This creates a different random sample every time the procedure is run.
- SAMPSIZE= indicates the number of observations to be included in the sample. To select a certain fraction of the original data set rather than a given number of observations, use the SAMPRATE= option.

Selected SURVEYSELECT procedure statement:

- STRATA enables the user to specify one or more stratification variables. If no STRATA statement is specified, no stratification takes place.

Other statements and options for the SURVEYSELECT procedure can be found in the SAS online documentation.

```

proc surveyselect
  data=sasuser.b_cars    /* sample from data table */
  seed=31475              /* recommended that you use this option */
  method=srs               /* simple random sample */
  sampsize=12              /* sample size */
  out=work.Carsample12    /* sample stored in this data set */
;
run;

```

-  If you do not provide a seed, you cannot reproduce the sample. It is recommended that you always include a seed when using PROC SURVEYSELECT.

The SAS System	
The SURVEYSELECT Procedure	
Selection Method Simple Random Sampling	
Input Data Set	B_CARS
Random Number Seed	31475
Sample Size	12
Selection Probability	0.130435
Sampling Weight	7.666667
Output Data Set	CARSAMPLE12

#### Partial VIEWTABLE

VIEWTABLE: Work.Carsample12					
	Manufacturer	Model	MidPrice	CityMPG	HighwayMPG
1	Acura	Integra	15.9	25	31
2	BMW	535i	30	22	30
3	Chevrolet	Astro	16.6	15	20
4	Dodge	Shadow	11.3	23	29
5	Ford	Crown_Victor	20.9	18	26
6	Geo	Storm	12.5	30	36
7	Hyundai	Elantra	10	22	29
8	Lexus	ES300	28	18	24
9	Mazda	RX-7	32.5	17	25
10	Pontiac	Bonneville	24.4	19	28
11	Volkswagen	Eurovan	19.7	17	21
12	Volkswagen	Passat	20	21	30

```

proc surveyselect
  data=sasuser.b_cars      /* sample from data table */
  seed=13094425            /* recommended that you use this option */
  method=srs                /* simple random sample */
  samprate=0.05              /* 0 < sampling rate < 1 */
  out=sasuser.b_cars12pc /* sample stored in this data set */
;
run;

```

The SURVEYSELECT Procedure

Selection Method Simple Random Sampling

Input Data Set	B_CARS
Random Number Seed	13094425
Sampling Rate	0.05
Sample Size	5
Selection Probability	0.054348
Sampling Weight	18.4
Output Data Set	B_CARS12PC

Partial VIEWTABLE

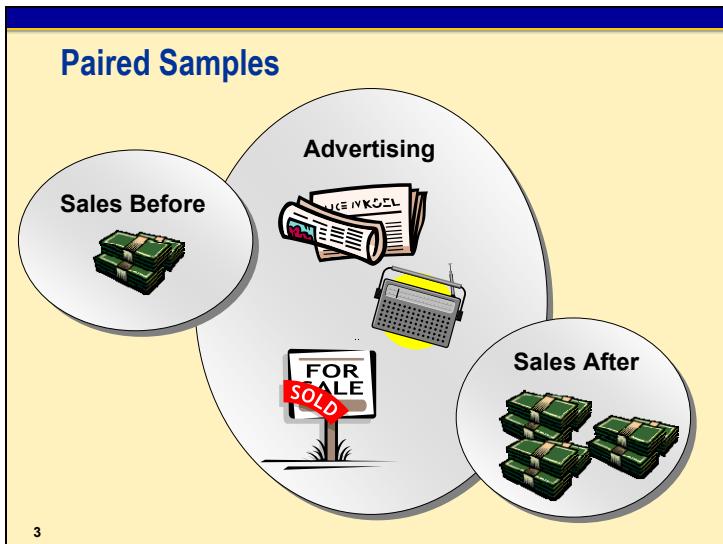
VIEWTABLE: Sasuser.B_cars12pc					
	Manufacturer	Model	MidPrice	CityMPG	HighwayMPG
1	Cadillac	DeVille	34.7	16	25
2	Ford	Aerostar	19.9	15	20
3	Hyundai	Excel	8	29	33
4	Saab	900	28.7	20	26
5	Toyota	Tercel	9.8	32	37

# Appendix C Additional Topics

C.1 Paired <i>t</i> -Tests .....	C-3
C.2 Two-Sample <i>t</i> -Tests .....	C-7
C.3 Output Delivery System .....	C-17
C.4 Nonparametric ANOVA .....	C-27
C.5 Partial Leverage Plots .....	C-40



## C.1 Paired t-Tests



For many types of data, repeat measurements are taken on the same subject throughout a study. The simplest form of this study is often referred to as the paired *t*-test.

In this study design,

- subjects are exposed to a treatment, for example, an advertising strategy
- a measurement is taken on the subjects before and after the treatment
- the subjects, on average, respond the same way to the treatment, although there might be differences among the subjects.

The assumptions of this test are that

- the subjects are selected randomly
- the distribution of the sample mean differences is normal. The central limit theorem can be applied for large samples.

The hypotheses of this test are

$$H_0: \mu_{\text{POST}} = \mu_{\text{PRE}}$$

$$H_1: \mu_{\text{POST}} \neq \mu_{\text{PRE}}$$

## The TTEST Procedure

General form of the TTEST procedure:

```
PROC TTEST DATA=SAS-data-set;
  CLASS variable;
  VAR variables;
  PAIRED variable*variable;
RUN;
```

4

Selected TTEST procedure statements:

- CLASS        specifies the two-level variable for the analysis. Only one variable is allowed in the CLASS statement.
- VAR        specifies numeric response variables for the analysis. If the VAR statement is not specified, PROC TTEST analyzes all numeric variables in the input data set that are not listed in a CLASS (or BY) statement.
- PAIRED    identifies the variables to be compared in paired comparisons. Variables are separated by an asterisk (\*). The asterisk requests comparisons between each variable on the left with each variable on the right. The differences are calculated by taking the variable on the left minus the variable on the right of the asterisk.



## Paired t-Test

Example: Dollar values of sales have been collected both before and after a particular advertising campaign. You are interested in determining the effect of the campaign on sales. You have collected data from 30 different randomly selected regions. The level of sales both before (**pre**) and after (**post**) the campaign were recorded and are shown below.

```
proc print data=sasuser.b_market (obs=20);  
  title;  
run;          /* ssdemo01 */
```

OBS	PRE	POST
1	9.52	10.28
2	9.63	10.45
3	7.71	8.51
4	7.83	8.62
5	8.97	10.03
6	8.62	9.45
7	10.11	9.68
8	9.96	9.62
9	8.50	11.84
10	9.62	11.95
11	10.29	10.52
12	10.13	10.67
13	9.11	11.03
14	8.95	10.53
15	10.86	10.70
16	9.31	10.24
17	9.59	10.82
18	9.27	10.16
19	11.86	12.12
20	10.15	11.28

The PAIRED statement used below is testing whether the mean of post-sales is significantly different from the mean of the presales because **post** is on the left of the asterisk and **pre** is on the right.

```
proc ttest data=sasuser.b_market;
  paired post*pre;
  title 'Testing the Difference Before and After a Sales Campaign';
run;          /* ssdemo02 */
```

Testing the Difference Before and After a Sales Campaign

The TTEST Procedure

Statistics

Difference	N	Lower CL		Upper CL		Lower CL		Upper CL	
		Mean	Mean	Mean	Std Dev	Std Dev	Std Dev	Std Dev	
post - pre	30	0.6001	0.9463	1.2925	0.7384	0.9271	1.2464		

Statistics

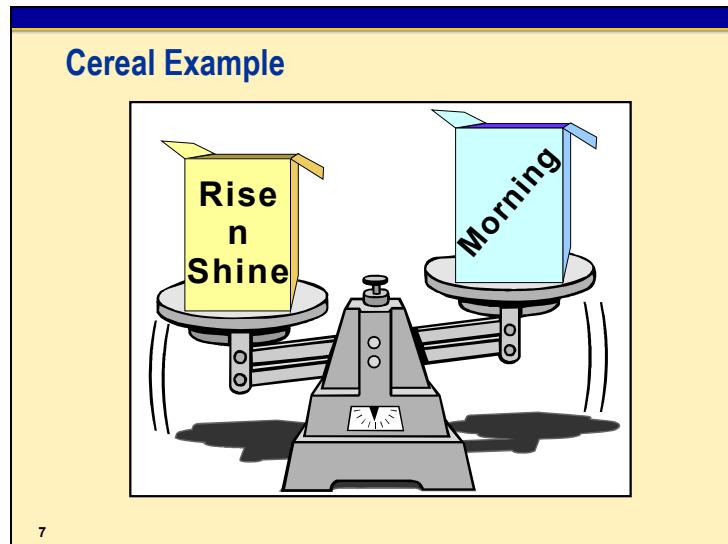
Difference	Std Err	Minimum	Maximum
post - pre	0.1693	-0.48	3.34

T-Tests

Difference	DF	t Value	Pr >  t
post - pre	29	5.59	<.0001

The T-Tests table provides the requested analysis. The *p*-value for the difference **post-pre** is less than 0.0001. Assuming that you want 0.01 level of significance, you reject the null hypothesis and conclude that there is a change in the average sales after the advertising campaign. Also, based on the fact that the mean is positive 0.9463, there appears to be an increase in the average sales after the advertising campaign.

## C.2 Two-Sample t-Tests



Example: A consumer advocacy group wants to determine whether two popular cereal brands, Rise n Shine and Morning, have the same amount of cereal. Both brands advertise that they have 15 ounces of cereal per box. A random sample of both brands is selected and the number of ounces of cereal is recorded. The data is stored in a data set named **sasuser.b\_cereal**.

The variables in the data set are as follows:

- brand** two groups, Rise n Shine and Morning, corresponding to the two brands  
**weight** weight of the cereal in ounces  
**idnumber** the identification number for each cereal box

## Assumptions

**Comparing Two Populations**

The diagram shows two vertical lines representing population distributions. The left line is labeled  $\mu_1$  and the right line is labeled  $\mu_2$ . Below the lines, the word "Morning" is written under the first line and "Rise n Shine" is written under the second line. The title "Comparing Two Populations" is centered above the lines.

- independent observations
- normally distributed data for each group
- equal variances for each group

8

Before you start the analysis, examine the data to verify that the assumptions are valid.

The assumption of independent observations means that no observations provide any information about any other observation you collect. For example, measurements are not repeated on the same subject. This assumption can be verified during the design stage.

The assumption of normality can be relaxed if the data is approximately normally distributed or if enough data is collected. This assumption can be verified by examining plots of the data.

There are several tests for equal variances. If this assumption is not valid, an approximate *t*-test can be performed.

If these assumptions are **not** valid, the probability of drawing incorrect conclusions from the analysis could increase.

### F-Test for Equality of Variances

$H_0: \sigma_1^2 = \sigma_2^2$        $H_1: \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$$

9

When performing this test, note that if the null hypothesis is true,  $F$  tends to be close to 1.

If you reject the null hypothesis, it is recommended that you use the unequal variance  $t$ -test in the TTEST procedure for testing the equality of group means.

This test is valid **only** for independent samples from normal distributions. Normality is required even for large sample sizes.

### Test Statistics and p-Values

**F-Test for Equal Variances:**  $H_0: \sigma_{12} = \sigma_{22}$

**Variance Test:**  $F' = 1.51$     DF = (3,3)    Prob > F' = 0.7446

**t-Tests for Equal Means:**  $H_0: \mu_1 = \mu_2$

**Unequal Variance t-Test:**  
 $T = 7.4017$     DF = 5.8    Prob > |T| = 0.0004

**Equal Variance t-Test:**  
 $T = 7.4017$     DF = 6.0    Prob > |T| = 0.0003

10

First, check the assumption for equal variances and then use the appropriate test for equal means. Because the  $p$ -value of the test  $F$ -statistic is 0.7446, there is not enough evidence to reject the null hypothesis of equal variances. Use the equal variance  $t$ -test line in the output to test whether the means of the two populations are equal.

The null hypothesis that the group means are equal is rejected at the 0.05 level. You conclude that there is a difference between the means of the groups.



The equal variance  $F$ -test is found at the bottom of the PROC TTEST output.

## Test Statistics and *p*-Values

**F-Test for Equal Variances:**  $H_0: \sigma_{12} = \sigma_{22}$

**Variance Test:**

$F' = 15.28$  DF = (9,4) Prob > F' = 0.0185

**t-Tests for Equal Means:**  $H_0: \mu_1 = \mu_2$

**Unequal Variance t-Test:**

$T = -2.4518$  DF = 11.1 Prob > |T| = 0.0320

**Equal Variance t-Test:**

$T = -1.7835$  DF = 13.0 Prob > |T| = 0.0979

11

Again, first check the assumption for equal variances and use the appropriate test for equal means. Because the *p*-value of the test *F*-statistic is less than alpha=0.05, there is enough evidence to reject the null hypothesis of equal variances. Use the unequal variance *t*-test line in the output to test whether the means of the two populations are equal.

The null hypothesis that the group means are equal is rejected at the .05 level.

However, notice that if you choose the equal variance *t*-test, you would not reject the null hypothesis at the .05 level. This shows the importance of choosing the appropriate *t*-test.



## Two-Sample *t*-Test

Example: Print the data in the **sasuser.b\_cereal** data set and do an initial check of the assumptions of the *t*-test and the *F*-test using the UNIVARIATE procedure. Then invoke PROC TTEST to test the hypothesis that the means are equal for the two groups.

```
proc print data=sasuser.b_cereal (obs=15);
  title 'Partial Listing of Cereal Data';
run;          /* ssdemo03 */
```

Part of the data is shown below.

Partial Listing of Cereal Data

OBS	BRAND	WEIGHT	ID
1	Morning	14.9982	61469897
2	Rise n Shine	15.0136	33081197
3	Morning	15.0100	68137597
4	Rise n Shine	14.9982	37070397
5	Morning	15.0052	64608797
6	Rise n Shine	14.9930	60714297
7	Morning	14.9733	16907997
8	Rise n Shine	15.0812	9589297
9	Morning	15.0037	93891897
10	Rise n Shine	15.0418	85859397
11	Morning	14.9957	38152597
12	Rise n Shine	15.0639	99108497
13	Morning	15.0099	59666697
14	Rise n Shine	15.0613	70847197
15	Morning	14.9943	47613397

```
proc sort data=sasuser.b_cereal out=sorted_cereal;
  by brand;
run;          /* ssdemo04 */

proc univariate data=sorted_cereal;
  by brand;
  var weight;
  histogram weight / normal;
  probplot weight / normal (mu=est sigma=est
                            color=blue w=1);
  title 'Univariate Analysis of the Cereal Data';
run;
```



In order to generate the analysis for each cereal brand, the data must be sorted by the variable **brand**. The SORT procedure step is needed before PROC UNIVARIATE, and the same BY variable used in PROC SORT is needed in PROC UNIVARIATE.

## Partial PROC UNIVARIATE Output

```

Univariate Analysis of the Cereal Data
----- brand=Morning -----
The UNIVARIATE Procedure
Variable: weight

Moments

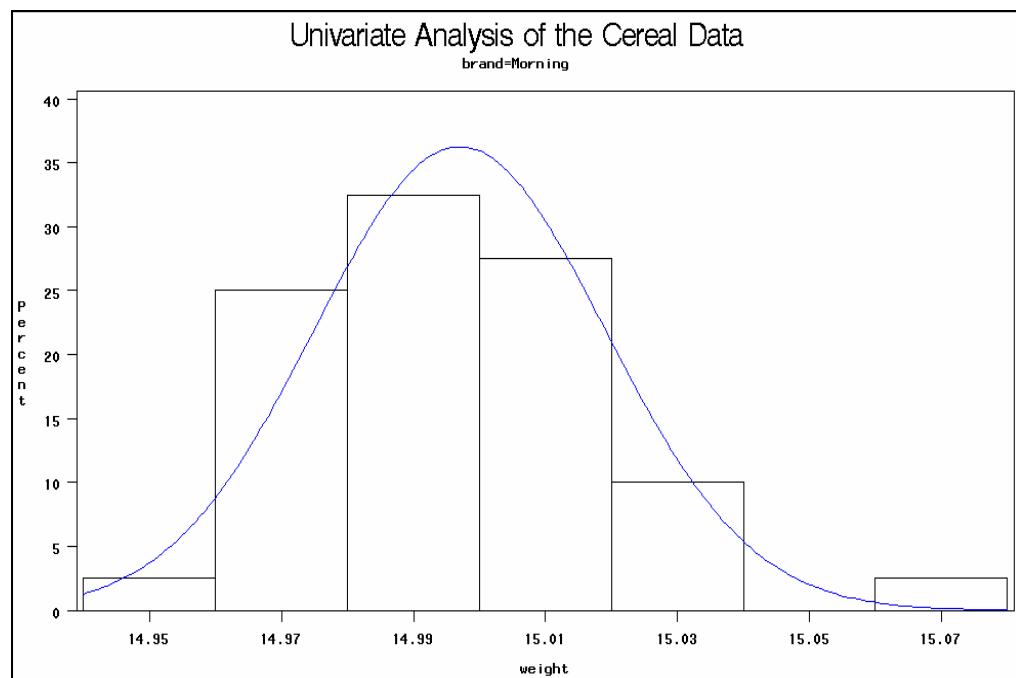
N           40   Sum Weights      40
Mean        14.9970125  Sum Observations  599.8805
Std Deviation 0.02201048  Variance       0.00048446
Skewness     0.87481049  Kurtosis       2.07993397
Uncorrected SS 8996.43425  Corrected SS  0.01889398
Coeff Variation 0.14676575  Std Error Mean 0.00348016

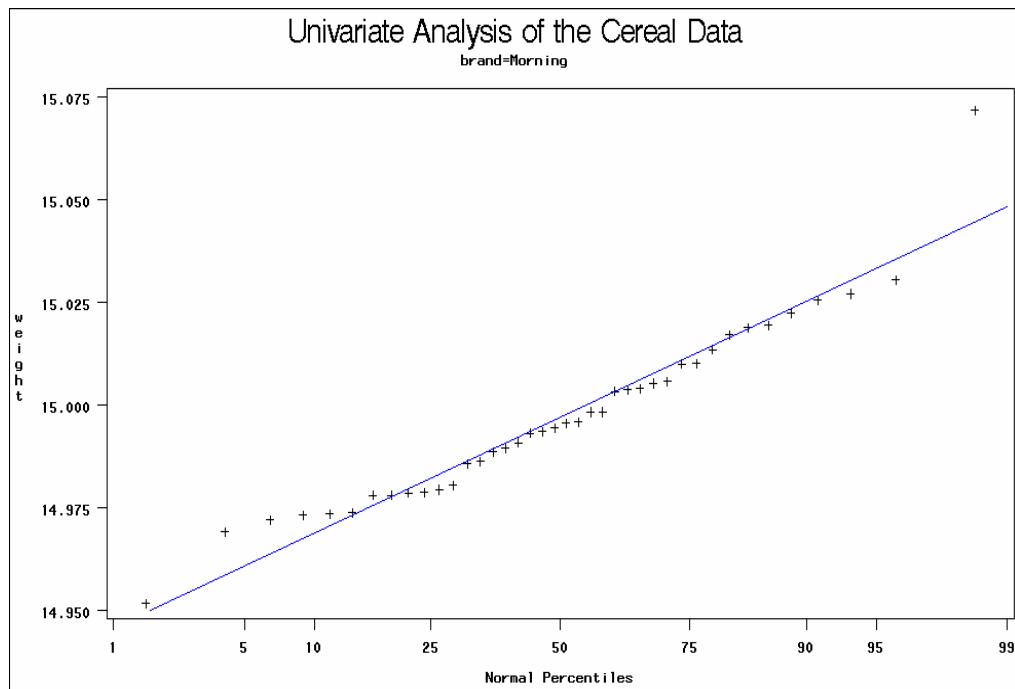
Basic Statistical Measures

Location          Variability
Mean      14.99701  Std Deviation      0.02201
Median    14.99490  Variance        0.0004845
Mode      14.97790  Range          0.12010
                  Interquartile Range 0.03095

NOTE: The mode displayed is the smallest of 2 modes with a count of 2.

```





The histogram and normal probability plots show one extreme value. Otherwise, the data for Morning appears to be symmetric. There appears to be no pattern for the data that reflects skewness or kurtosis.

## PROC UNIVARIATE Output (continued)

```

Univariate Analysis of the Cereal Data
----- brand=Rise n Shine -----
The UNIVARIATE Procedure
Variable: weight

Moments

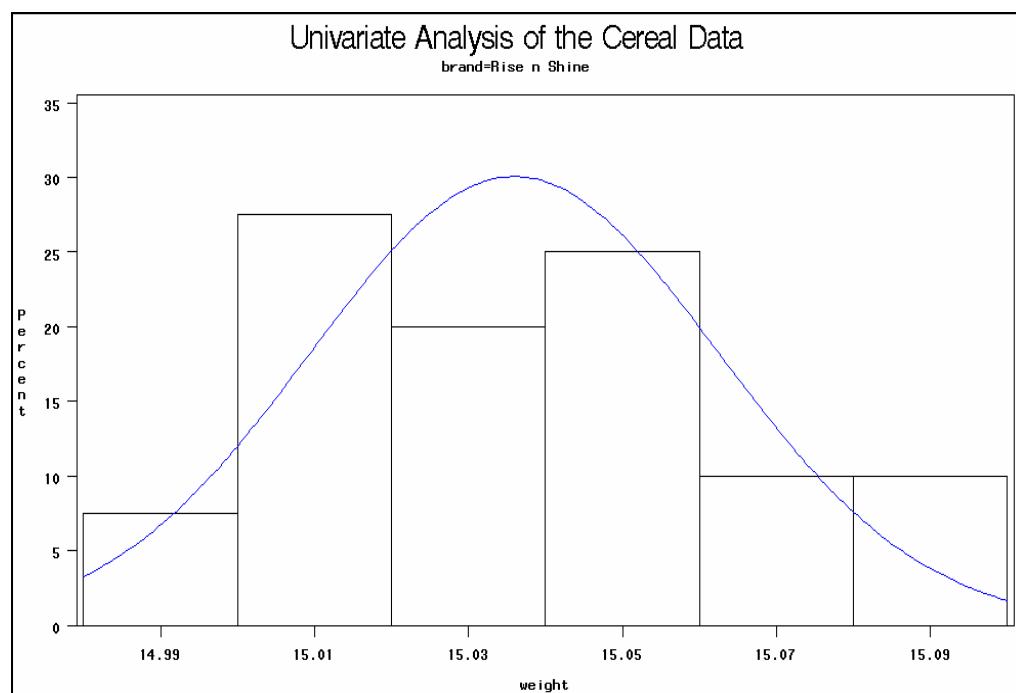
      N          40   Sum Weights        40
      Mean       15.03596  Sum Observations    601.4384
      Std Deviation  0.02654963  Variance      0.00070488
      Skewness      0.39889232  Kurtosis     -0.1975717
      Uncorrected SS 9043.23122  Corrected SS  0.02749044
      Coeff Variation 0.17657424  Std Error Mean 0.00419787

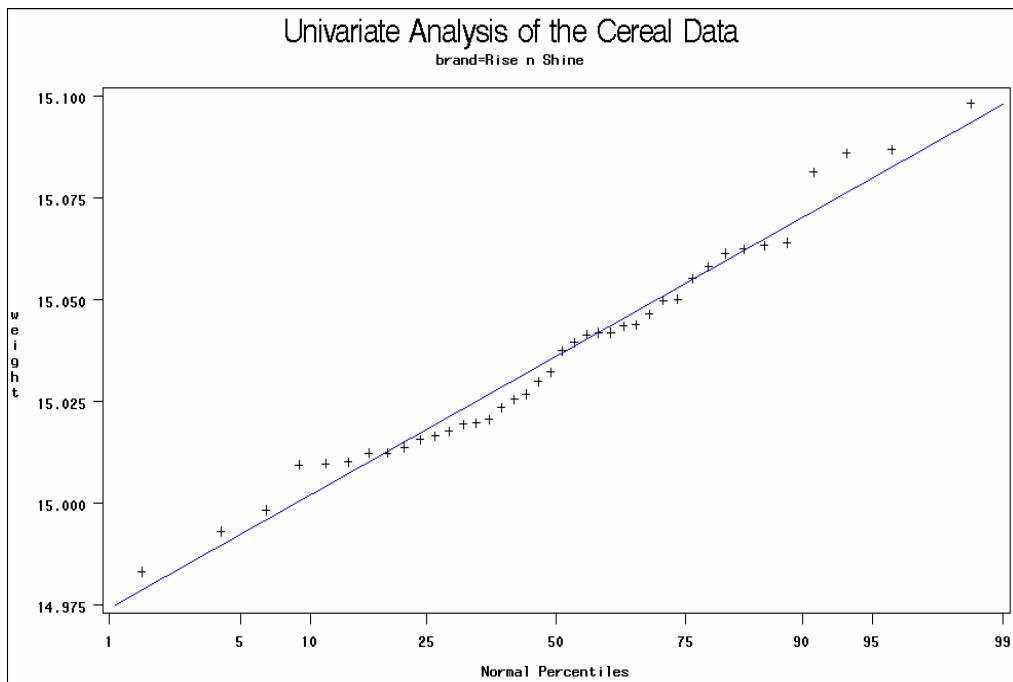
Basic Statistical Measures

      Location           Variability
      Mean      15.03596  Std Deviation      0.02655
      Median    15.03480  Variance        0.0007049
      Mode      15.01220  Range          0.11490
                           Interquartile Range 0.03650

NOTE: The mode displayed is the smallest of 2 modes with a count of 2.

```





The histogram demonstrates that the data is fairly symmetric. There are also no extreme values. The normal probability plot shows no serious departures from normality.

Because both brands have weights that are normally distributed, the assumptions of the *F*-test for equal variances are verified. The assumption of the *t*-test regarding the normality of the distribution of sample means is also satisfied. You could have used the central limit theorem to validate the assumption for the *t*-test because both brands have 40 observations.

Invoke the TTEST procedure and interpret the output.

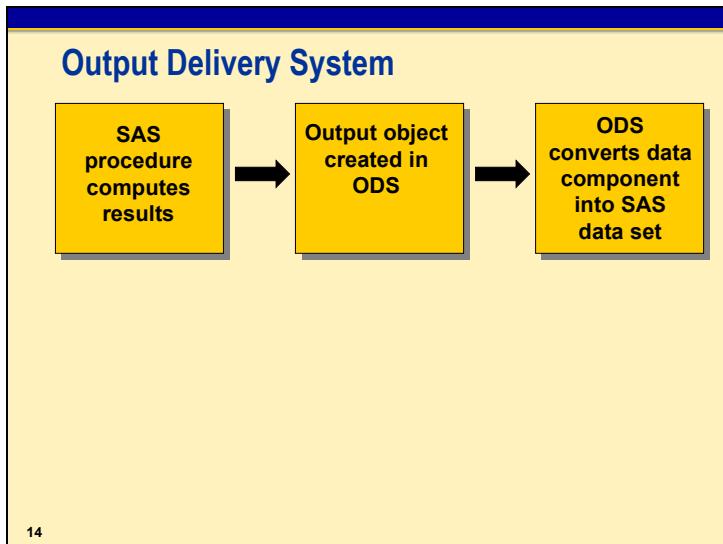
```
proc ttest data=sasuser.b_cereal;
  class brand;
  var weight;
  title 'Testing the Equality of Means for Two Cereal'
        ' Brands';
  run;          /* ssdemo05 */
```

Testing the Equality of Means for Two Cereal Brands							
The TTEST Procedure							
1 Statistics							
Variable	brand	N	Lower CL Mean	Upper CL Mean	Lower CL Mean	Std Dev	Std Dev
weight	Morning	40	14.99	14.997	15.004	0.018	0.022
weight	Rise n Shine	40	15.027	15.036	15.044	0.0217	0.0265
weight	Diff (1-2)		-0.05	-0.039	-0.028	0.0211	0.0244
Statistics							
Variable	brand	Upper CL Std Dev	Std Err	Minimum	Maximum		
weight	Morning	0.0283	0.0035	14.952	15.072		
weight	Rise n Shine	0.0341	0.0042	14.983	15.098		
weight	Diff (1-2)	0.0289	0.0055				
3 T-Tests							
Variable	Method	Variances		DF	t Value	Pr >  t	
weight	Pooled	Equal		78	-7.14	<.0001	
weight	Satterthwaite	Unequal		75.4	-7.14	<.0001	
Equality of Variances							
Variable	Method	Num DF	Den DF	F Value	Pr > F		
2 weight	Folded F	39	39	1.45	0.2460		

- ① In the Statistics table, examine the descriptive statistics for each group and their differences. The confidence limits for the sample mean and sample standard deviation are also shown.
- ② Look at the Equality of Variances table that appears at the bottom of the output. The *F*-test for equal variances has a *p*-value of 0.2460. In this case, do not reject the null hypothesis. Conclude that there is insufficient evidence to indicate that the variances are not equal.
- ③ Based on the *F*-test for equal variances, you then look in the T-Tests table at the *t*-test for the hypothesis of equal means. Using the equal variance *t*-test, you reject the null hypothesis that the group means are equal. Conclude that there is a difference in the average weight of the cereal between the Rise n Shine brand and the Morning brand.

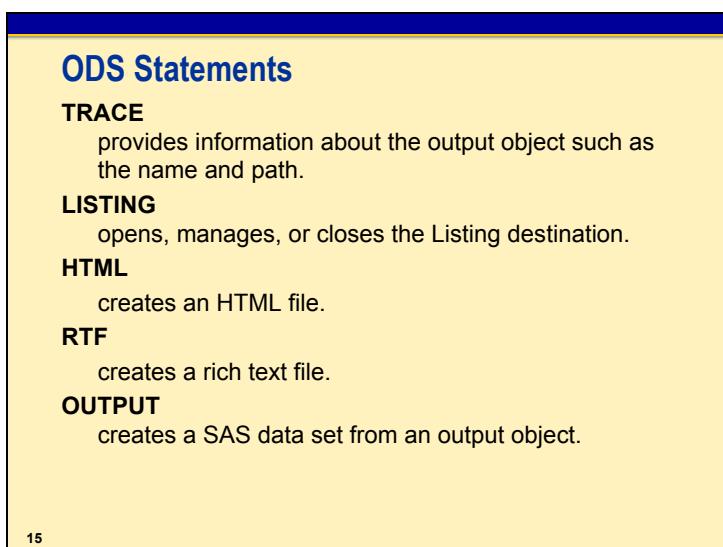
Return your attention to the Statistics table. Because the confidence interval for the mean (-0.05, -0.028) does not include 0, you can conclude that there is a significant difference between the two cereal means.

## C.3 Output Delivery System



14

The Output Delivery System (ODS) enables you to take output from a SAS procedure and convert it to a SAS data set. Instead of writing to the listing file directly, SAS procedures can now create an output object for each piece of output that is displayed. For example, each table produced in the UNIVARIATE procedure is now a separate entity in ODS. You can then take the data component of the output object and convert it to a SAS data set. This means that every number in every table of every procedure can be accessed via a data set.



15

The TRACE statement is used to obtain the name of the output object.



## Output Delivery System

Example: Examine some basic functionality of the Output Delivery System.

The ODS TRACE ON statement produces a trace record in the SAS Log window, including the name and label of each output object.

```
ods trace on;
/*-----*/
/* -generate and examine table definitions for UNIVARIATE */
/*-----*/
proc univariate data=sasuser.b_rise;
    var weight;
    histogram weight / normal;
    probplot weight / normal (mu=est sigma=est
                                color=blue w=1);
    title 'Univariate Analysis of sasuser.b_rise';
run;
ods trace off;          /* ssdemo06 */
```

### SAS Log

```
2   ods trace on;
3   /*-----*/
4   /* -generate and examine table definitions for UNIVARIATE */
5   /*-----*/
6   proc univariate data=sasuser.b_rise;
7       var weight;
8       histogram weight / normal;
9       probplot weight / normal (mu=est sigma=est
10                      color=blue w=1);
11      title 'Univariate Analysis of sasuser.b_rise';
12      run;
```

Output Added:

-----

Name: Moments  
Label: Moments  
Template: base.univariate.Moments  
Path: Univariate.weight.Moments

-----

Output Added:

-----

Name: BasicMeasures  
Label: Basic Measures of Location and Variability  
Template: base.univariate.Measures  
Path: Univariate.weight.BasicMeasures

-----

## SAS Log (continued)

```
Output Added:  
-----  
Name:      TestsForLocation  
Label:     Tests For Location  
Template:  base.univariate.Location  
Path:      Univariate.weight.TestsForLocation  
-----  
  
Output Added:  
-----  
Name:      Quantiles  
Label:     Quantiles  
Template:  base.univariate.Quantiles  
Path:      Univariate.weight.Quantiles  
-----  
  
Output Added:  
-----  
Name:      ExtremeObs  
Label:     Extreme Observations  
Template:  base.univariate.ExtObs  
Path:      Univariate.weight.ExtremeObs  
-----  
  
Output Added:  
-----  
Name:      Univar  
Data Name: GRSEG  
Path:      Univariate.weight.Univar  
-----  
  
Output Added:  
-----  
Name:      ParameterEstimates  
Label:     Parameter Estimates  
Template:  base.univariate.FitParms  
Path:      Univariate.weight.FittedDistributions.Normal.ParameterEstimates  
-----  
  
Output Added:  
-----  
Name:      GoodnessOfFit  
Label:     Goodness of Fit  
Template:  base.univariate.FitGood  
Path:      Univariate.weight.FittedDistributions.Normal.GoodnessOfFit  
-----  
  
Output Added:  
-----  
Name:      FitQuantiles  
Label:     Quantiles  
Template:  base.univariate.FitQuant  
Path:      Univariate.weight.FittedDistributions.Normal.FitQuantiles  
-----
```

## SAS Log (continued)

```
Output Added:  
-----  
Name:      Univar1  
Data Name: GRSEG  
Path:      Univariate.weight.Univar1  
-----  
NOTE: PROCEDURE UNIVARIATE used (Total process time):  
      real time      0.94 seconds  
      cpu time       0.37 seconds  
  
13   ods trace off;          /* ssdemo06 */
```

For each table, Name, Label, Template or Data Name, and Path are listed.

You can now select only the tables that are of interest to you.

```
ods select  
  Moments  
  BasicMeasures  
  GoodnessOfFit  
  ;  
ods listing;  
proc univariate data=sasuser.b_rise;  
  var weight;  
  histogram weight / normal;  
  probplot weight / normal (mu=est sigma=est  
                           color=blue w=1);  
  title1 'Selected Results Using ODS';  
run;           /* ssdemo07 */
```

## Selected Results Using ODS

The UNIVARIATE Procedure  
Variable: weight

## Moments

N	40	Sum Weights	40
Mean	15.03596	Sum Observations	601.4384
Std Deviation	0.02654963	Variance	0.00070488
Skewness	0.39889232	Kurtosis	-0.1975717
Uncorrected SS	9043.23122	Corrected SS	0.02749044
Coeff Variation	0.17657424	Std Error Mean	0.00419787

## Basic Statistical Measures

## Location Variability

Mean	15.03596	Std Deviation	0.02655
Median	15.03480	Variance	0.0007049
Mode	15.01220	Range	0.11490
		Interquartile Range	0.03650

NOTE: The mode displayed is the smallest of 2 modes with a count of 2.

## Selected Results using ODS

The UNIVARIATE Procedure  
Fitted Distribution for weight

## Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic---		----p Value----	
Kolmogorov-Smirnov	D	0.09608648	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05930447	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.38776343	Pr > A-Sq	>0.250

Although these reports are effective, in order to make them easier to distribute, use ODS to generate them in HTML format.

```
ods listing close;

ods html
  body='sel_u.htm';
ods select
  Moments
  BasicMeasures
  GoodnessOfFit
;
proc univariate data=sasuser.b_rise;
  var weight;
  histogram weight / normal;
  probplot weight / normal (mu=est sigma=est
                            color=blue w=1);
  title 'Selected Results in HTML format';
run;
ods html close;
ods listing;
title; /* ssdemo08 */
```

HTML Output

## ***Selected Results in HTML format***

### ***The UNIVARIATE Procedure***

***Variable: weight***

<b>Moments</b>			
<b>N</b>	40	<b>Sum Weights</b>	40
<b>Mean</b>	15.03596	<b>Sum Observations</b>	601.4384
<b>Std Deviation</b>	0.02654963	<b>Variance</b>	0.00070488
<b>Skewness</b>	0.39889232	<b>Kurtosis</b>	-0.1975717
<b>Uncorrected SS</b>	9043.23122	<b>Corrected SS</b>	0.02749044
<b>Coeff Variation</b>	0.17657424	<b>Std Error Mean</b>	0.00419787

### **Basic Statistical Measures**

<b>Location</b>		<b>Variability</b>	
<b>Mean</b>	15.03596	<b>Std Deviation</b>	0.02655
<b>Median</b>	15.03480	<b>Variance</b>	0.0007049
<b>Mode</b>	15.01220	<b>Range</b>	0.11490
		<b>Interquartile Range</b>	0.03650

HTML Output (continued)

<b>Selected Results in HTML format</b>				
<b>The UNIVARIATE Procedure</b>				
<b>Fitted Distribution for weight</b>				
<b>Goodness-of-Fit Tests for Normal Distribution</b>				
Test	<b>Statistic</b>		<b>p Value</b>	
Kolmogorov-Smirnov	D	0.09608648	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05930447	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.38776343	Pr > A-Sq	>0.250



The file containing this HTML, sel\_u.htm, is located in the root directory of where SAS resides in your environment.

If you are in the Windows environment, this HTML output is displayed immediately and is also available in the Results window.

You can also generate SAS data sets to extract specific values in later programming steps or for future analyses.

```
ods listing close;

ods output
  Moments=o_moments
  BasicMeasures=o_basic
  GoodnessOfFit=o_goodnessfit
  ;

proc univariate data=sasuser.b_rise;
  var weight;
  histogram weight / normal;
run;

ods listing;          /* ssdemo09 */
```

```
65  ods listing close;
66
67  ods output
68    Moments=o_moments
69    BasicMeasures=o_basic
70    GoodnessOfFit=o_goodnessfit
71    ;
72
73  proc univariate data=sasuser.b_rise;
74    var weight;
75    histogram weight / normal;
76  run;

NOTE: The data set WORK.O_GOODNESSFIT has 3 observations and 8 variables.
NOTE: The data set WORK.O_BASIC has 4 observations and 5 variables.
NOTE: The data set WORK.O_MOMENTS has 6 observations and 7 variables.
NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time          0.17 seconds
      cpu time          0.17 seconds
```

```
77
78  ods listing;          /* ssdemo09 */
```

 The SAS data sets that are generated with the OUTPUT statement are stored in the **work** library. To store them in a SAS data set, use a two-level SAS name.

	VarName	Distribution	Goodness-of-Fit Test	Label for Goodness-of-Fit Statistic	Value of Goodness-of-Fit Statistic	p-value Label	Sign of p-value	p-value
1	weight	Normal	Kolmogorov-Smirnov	D	0.09608648	Pr > D	>	0.150
2	weight	Normal	Cramer-von Mises	W-Sq	0.05930447	Pr > W-Sq	>	0.250
3	weight	Normal	Anderson-Darling	A-Sq	0.38776343	Pr > A-Sq	>	0.250

	VarName	LocMeasure	LocValue	VarMeasure	VarValue
1	weight	Mean	15.03596	Std Deviation	0.02655
2	weight	Median	15.03480	Variance	0.0007049
3	weight	Mode	15.01220	Range	0.11490
4	weight		_	Interquartile Range	0.03650

	VarName	Label1	cValue1	nValue1	Label2	cValue2	nValue2
1	weight	N	40	40.000000	Sum Weights	40	40.000000
2	weight	Mean	15.03596	15.035960	Sum Observations	601.4384	601.438400
3	weight	Std Deviation	0.02654963	0.026550	Variance	0.00070488	0.000705
4	weight	Skewness	0.39889232	0.398892	Kurtosis	-0.1975717	-0.197572
5	weight	Uncorrected SS	9043.23122	9043.231215	Corrected SS	0.02749044	0.027490
6	weight	Coeff Variation	0.17657424	0.176574	Std Error Mean	0.00419787	0.004198

## C.4 Nonparametric ANOVA

This section addresses nonparametric options within the NPAR1WAY procedure. Nonparametric one-sample tests are also available in the UNIVARIATE procedure.

### Nonparametric Analysis

*Nonparametric analyses* are those that rely only on the assumption that the observations are independent.

A nonparametric test is appropriate when

- the data contains valid outliers
- the data is skewed
- the response variable is ordinal and not continuous.

18

Nonparametric tests are most often used when the normality assumption required for analysis of variance is in question. Although ANOVA is robust against minor departures from its normality assumption, extreme departures from normality can make the test less sensitive to differences between means. Therefore, when the data is very skewed or there are extreme outliers, nonparametric methods might be more appropriate. In addition, when the data follows a count measurement scale instead of interval, nonparametric methods should be used.

 When the normality assumption is met, nonparametric tests are almost as good as parametric tests.

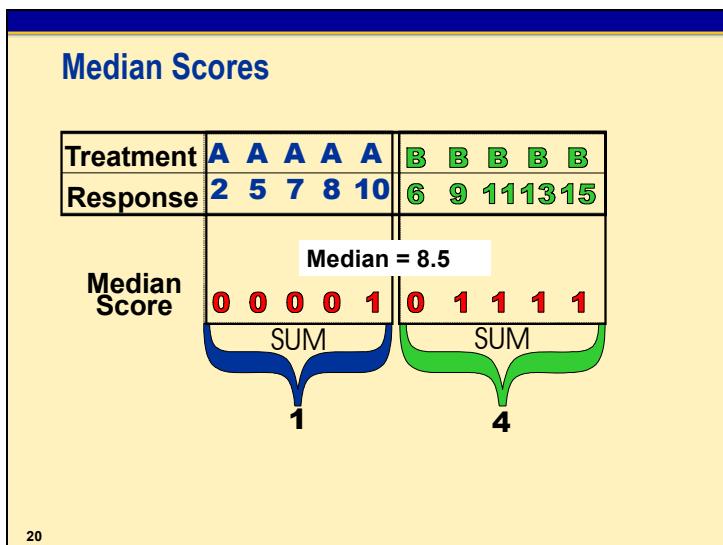
### Rank Scores

Treatment	A A A A A					B B B B B				
Response	2	5	7	8	10	6	9	11	13	15
Rank Score	1	2	4	5	7	3	6	8	9	10
	SUM				SUM					
	19				36					

19

In nonparametric analysis, the rank of each data point is used instead of the raw data.

The illustrated ranking system ranks the data from smallest to largest. In the case of ties, the ranks are averaged. The sums of the ranks for each of the treatments are used to test the hypothesis that the populations are identical. For two populations, the Wilcoxon rank-sum test is performed. For any number of populations, a Kruskal-Wallis test is used.

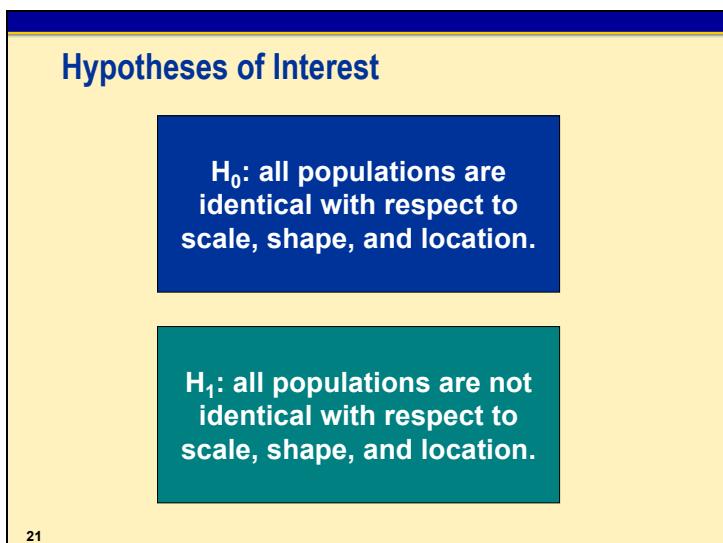


Recall that the median is the 50<sup>th</sup> percentile, which is the middle of your data values.

When calculating median scores, a score of

- 0 is assigned, if the data value is less than or equal to the median
  - 1 is assigned, if the data value is above the median.

The sums of the median scores are used to conduct the Median test for two populations or the Brown-Mood test for any number of populations.



Nonparametric tests compare the probability distributions of sampled populations rather than specific parameters of these populations.

In general, with no assumptions about the distributions of the data, you are testing these hypotheses:

- $H_0$ : all populations are identical with respect to shape and location
- $H_1$ : all populations are **not** identical with respect to shape and location.

Thus, if you reject the null hypothesis, you conclude that the population distributions are different, but you have not identified the reason for the difference. The difference could be because of different variances, skewness, kurtosis, or means.

## THE NPAR1WAY PROCEDURE

General form of the NPAR1WAY procedure:

```
PROC NPAR1WAY DATA=SAS-data-set <options>;
  CLASS variable;
  VAR variables;
RUN;
```

22

Selected NPAR1WAY procedure statements:

**CLASS** specifies a classification variable for the analysis. You must specify exactly one variable, although this variable can have any number of values.

**VAR** specifies numeric analysis variables.

## Hospice Example

Are there different effects of a marketing visit, in terms of increasing the number of referrals to the hospice, among the various specialties of physicians?



23

Consider a study done by Kathryn Skarzynski to determine whether there was a change in the number of referrals received from physicians after a visit by a hospice marketing nurse. One of her study questions was, “Are there different effects of the marketing visits, in terms of increasing the number of referrals, among the various specialties of physicians?”

### Veneer Example

Are there differences between the durability of brands of wood veneer?



24

Consider another experiment where the goal of the experiment is to compare the durability of three brands of synthetic wood veneer. This type of veneer is often used in office furniture and on kitchen countertops. To determine durability, four samples of each of three brands are subjected to a friction test. The amount of veneer material that is worn away due to the friction is measured. The resulting wear measurement is recorded for each sample. Brands that have a small wear measurement are desirable.



## The NPAR1WAY Procedure

Example: A portion of Ms. Skarzynski's data about the hospice marketing visits is in the data set **sasuser.b\_hosp**. The variables in the data set are as follows:

<b>id</b>	the ID number of the physician's office visited
<b>visit</b>	the type of visit, to the physician or to the physician's staff
<b>code</b>	the medical specialty of the physician
<b>ref3p</b>	the number of referrals three months prior to the visit
<b>ref2p</b>	the number of referrals two months prior to the visit
<b>ref1p</b>	the number of referrals one month prior to the visit
<b>ref3a</b>	the number of referrals three months after the visit
<b>ref2a</b>	the number of referrals two months after the visit
<b>ref1a</b>	the number of referrals one month after the visit

In addition, the following variables have been calculated:

<b>avgprior</b>	the average number of referrals per month for the three months prior to the visit
<b>diff1</b>	the difference between the number of referrals one month after the visit and the average number of referrals prior to the visit
<b>diff2</b>	the difference between the number of referrals two months after the visit and the average number of referrals prior to the visit
<b>diff3</b>	the difference between the number of referrals three months after the visit and the average number of referrals prior to the visit
<b>diffbys1</b>	the difference between the number of referrals one month after the visit and the number of referrals three months prior to the visit
<b>diffbys2</b>	the difference between the number of referrals two months after the visit and the number of referrals three months prior to the visit
<b>diffbys3</b>	the difference between the number of referrals three months after the visit and the number of referrals three months prior to the visit.

Print a subset of the variables for the first 10 observations in the data set.

```
proc print data=sasuser.b_hosp (obs=10);
  var visit code diffbys3;
run;          /* ssdemo10 */
```

Obs	visit	code	diffbys3
1	physician	family prac	0
2	physician	family prac	1
3	physician	oncologist	-1
4	physician	family prac	-3
5	physician	oncologist	1
6	physician	family prac	0
7	physician	oncologist	-1
8	physician	oncologist	-1
9	physician	internal med	1
10	physician	oncologist	1

One of the analyses to answer the research question is to compare **diffbys3** (the number of referrals three months after the visit minus the number three months before the visit) for the different specialties.

Initially, you want to examine the distribution of the data. The data has been sorted by **code**. A BY statement was used with PROC UNIVARIATE instead of a CLASS statement, conveniently grouping the histogram and normal probability plots for each level of **code**. The BOXPLOT procedure requires the data to be sorted.

```
options ps=50 ls=76;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc sort data=sasuser.b_hosp out=sorted_hosp;
  by code;
run;
proc univariate data=sorted_hosp;
  by code;
  var diffbys3;
  histogram diffbys3 / normal;
  probplot diffbys3 / normal (mu=est sigma=est);
  title 'Descriptive Statistics for Hospice Data';
run;

proc boxplot data=sorted_hosp;
  plot diffbys3*code / boxstyle=schematic cboxes=black;
run; /* ssdemo11 */
```

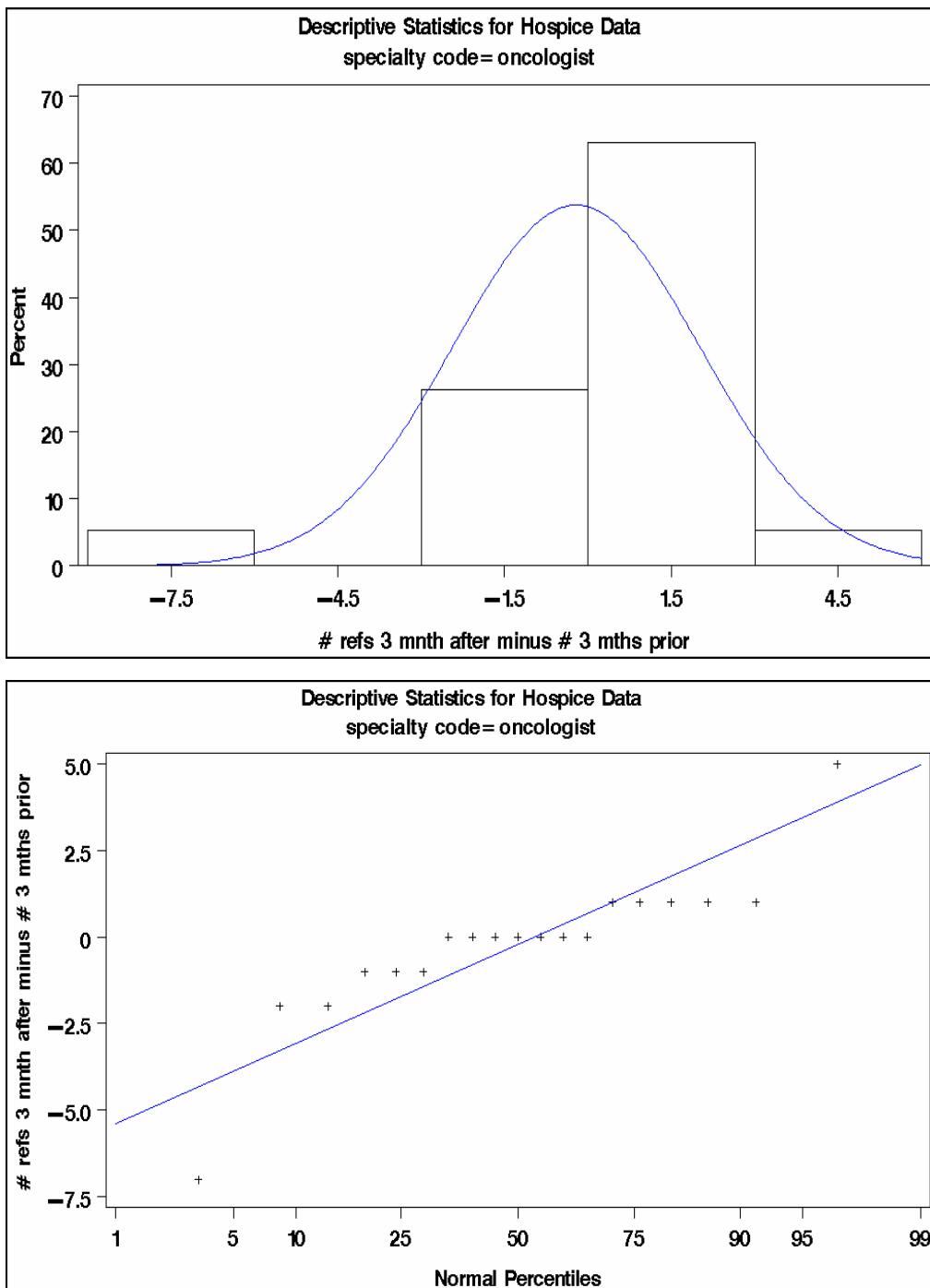
#### Selected PROC UNIVARIATE Output by **specialty code**

Group	Skewness	Kurtosis	Kolmogorov-Smirnov <i>p</i> -value	Cramer-von Mises <i>p</i> -value	Anderson-Darling <i>p</i> -value
oncologist	-0.988574	5.58306776	<0.010	<0.010	<0.005
internal med	0.94171457	-0.2843557	<0.010	<0.005	<0.005
family prac	-1.3336242	6.24954044	<0.010	<0.005	<0.005

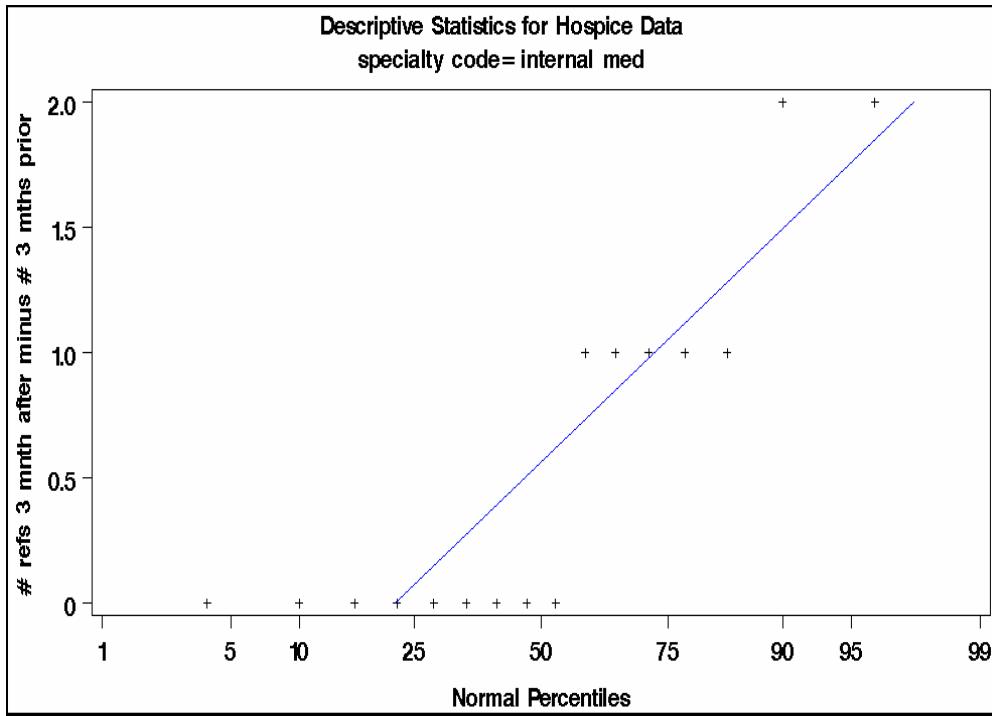
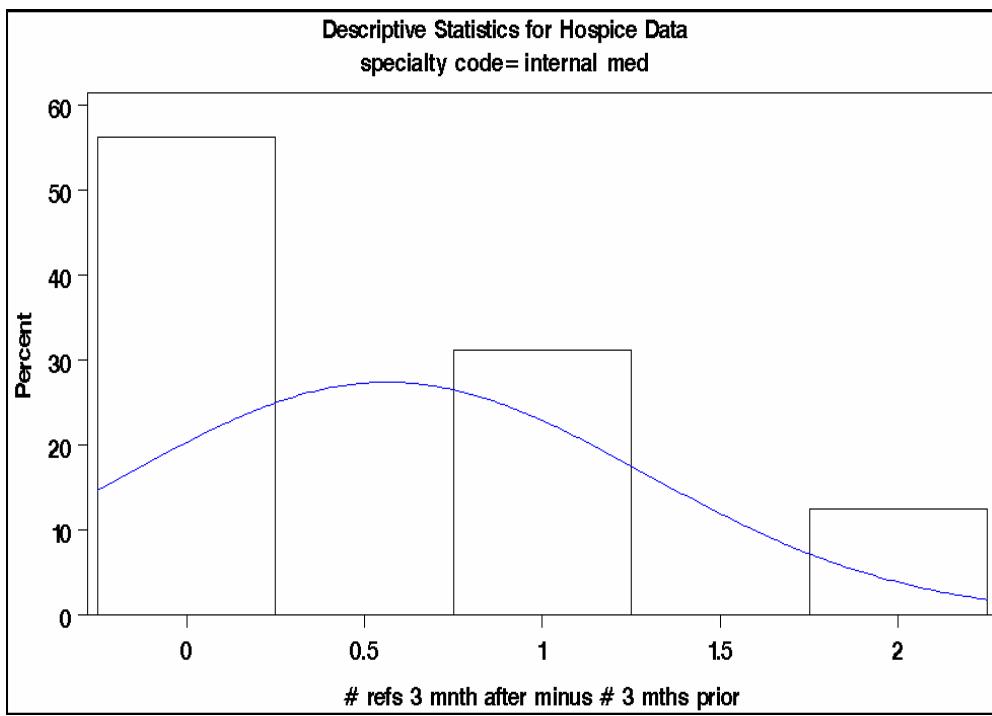
Based on skewness and kurtosis, the oncologists and family practice doctors might not be normal. All three goodness-of-fit tests reject the null hypothesis that the data is normal.

Now examine the histograms and normal probability plots for each group.

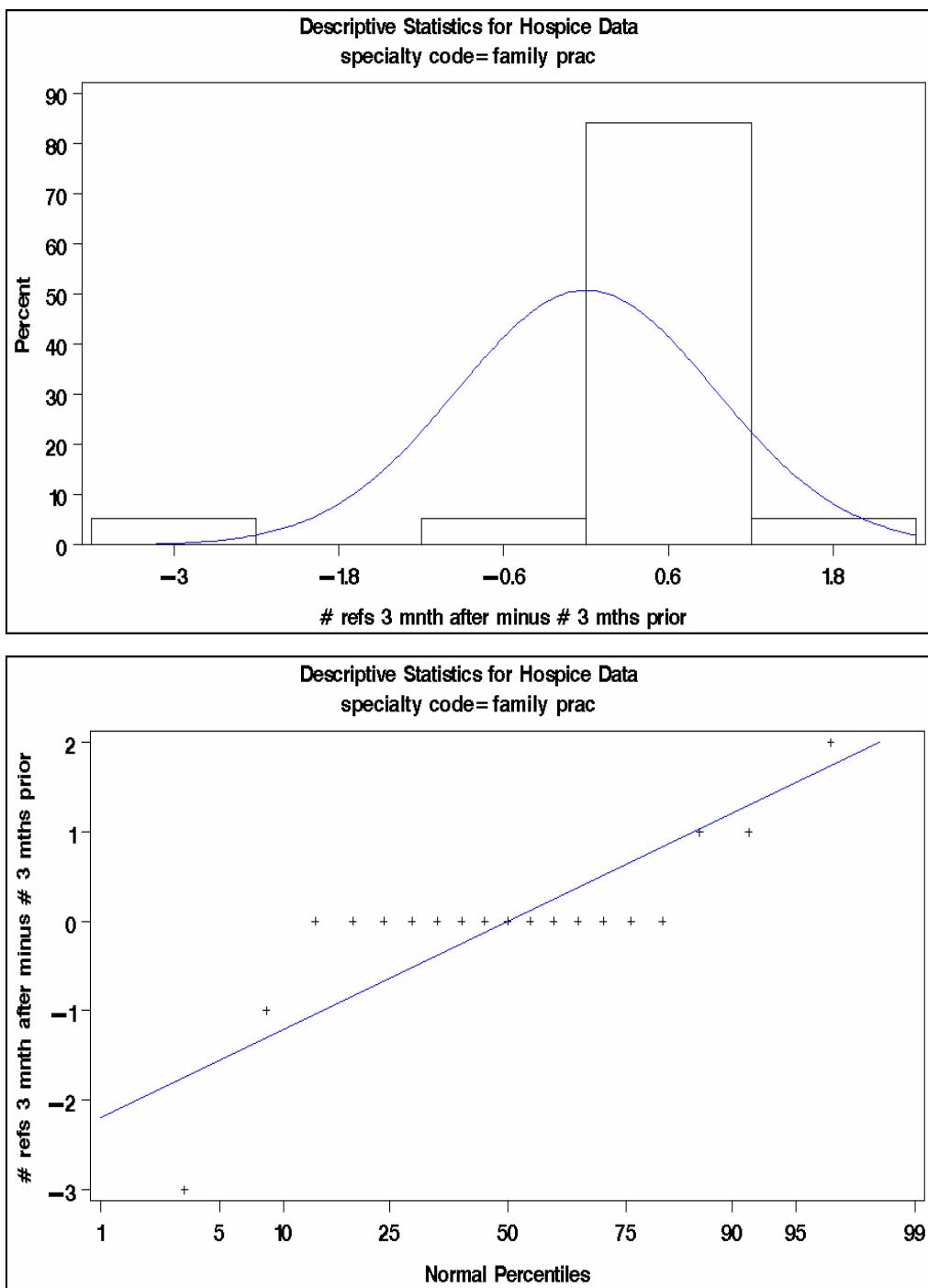
GRAPH Output (oncologists)



## GRAPH Output (internal medicine)

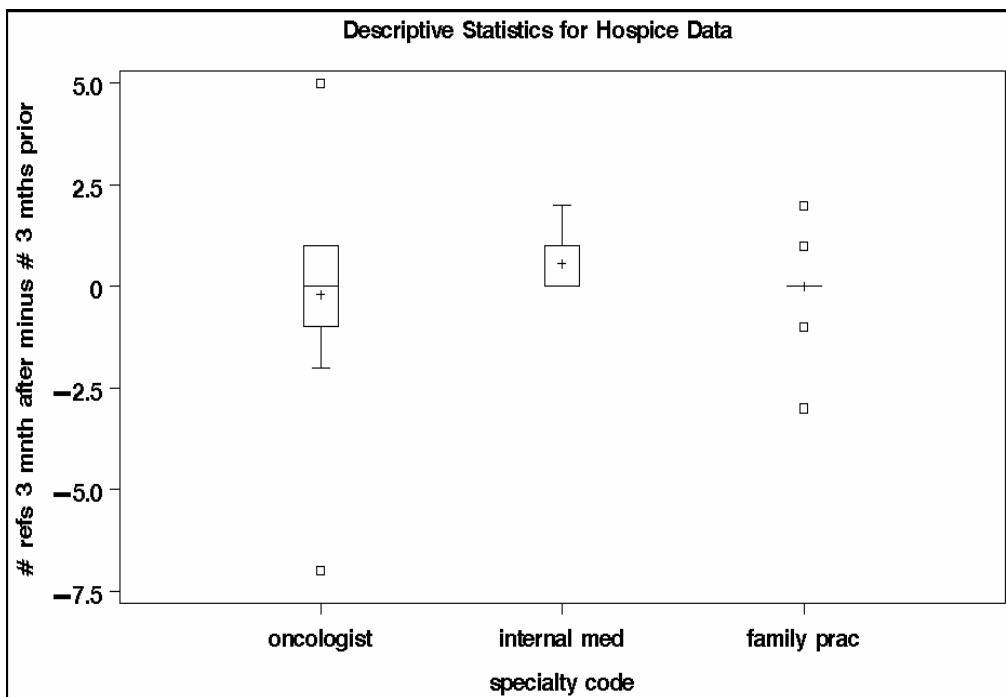


## GRAPH Output (family practice)



Family practice doctors appear to have outliers in the negative direction.

Now examine the PROC BOXPLOT output.



The box plots strongly support that the data is not normal. Remember that the data values of **diffbys3** are actually counts and therefore ordinal. This suggests that a nonparametric analysis would be more appropriate.

For illustrative purposes, use the WILCOXON option to perform a rank sum test and the MEDIAN option to perform the median test. This data was actually analyzed using the rank sum test. .

```
proc npar1way data=sorted_hosp wilcoxon median;
  class code;
  var diffbys3;
run;      /* ssdemo12 */
```

Selected PROC NPAR1WAY statement options:

WILCOXON requests an analysis of the rank scores. The output includes the Wilcoxon 2-sample test and the Kruskal-Wallis test for two or more populations.

MEDIAN requests an analysis of the median scores. The output includes the median 2-sample test and the median 1-way analysis test for two or more populations.

The NPAR1WAY Procedure					
Wilcoxon Scores (Rank Sums) for Variable diffbys3 Classified by Variable code					
code	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
oncologist	19	468.50	522.50	49.907208	24.657895
internal med	16	538.00	440.00	47.720418	33.625000
family prac	19	478.50	522.50	49.907208	25.184211

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	4.2304
DF	2
Pr > Chi-Square	0.1206

The PROC NPAR1WAY output from the WILCOXON option shows the actual sums of the rank scores and the expected sums of the rank scores if the null hypothesis is true. From the Kruskal-Wallis test (chi-square approximation), the *p*-value is .1206. Therefore, at the 5% level of significance, you do not reject the null hypothesis. There is not enough evidence to conclude that the distributions of change in hospice referrals for the different groups of physicians are significantly different

#### Partial PROC NPAR1WAY Output

The NPAR1WAY Procedure					
Median Scores (Number of Points Above Median) for Variable diffbys3 Classified by Variable code					
code	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
oncologist	19	8.566667	9.50	1.232093	0.450877
internal med	16	10.300000	8.00	1.178106	0.643750
family prac	19	8.133333	9.50	1.232093	0.428070

Average scores were used for ties.

Median One-Way Analysis

Chi-Square	3.8515
DF	2
Pr > Chi-Square	0.1458

Again, based on the *p*-value of .1458, at the 5% level of significance, you do not reject the null hypothesis. There is not enough evidence to conclude that there are differences between specialists.

Example: Recall the experiment to compare the durability of three brands of synthetic wood veneer. The data is stored in the **sasuser.b\_ven** data set.

```
proc print data=sasuser.b_ven;
  title 'Wood Veneer Wear Data';
run;          /* ssdemo13 */
```

#### Wood Veneer Wear Data

Obs	brand	wear
1	Acme	2.3
2	Acme	2.1
3	Acme	2.4
4	Acme	2.5
5	Champ	2.2
6	Champ	2.3
7	Champ	2.4
8	Champ	2.6
9	Ajax	2.2
10	Ajax	2.0
11	Ajax	1.9
12	Ajax	2.1

Because there is a sample size of only 4 for each brand of veneer, the usual PROC NPAR1WAY Wilcoxon test *p*-values might be inaccurate. Instead, the EXACT statement should be added to the PROC NPAR1WAY code. This provides exact *p*-values for the simple linear rank statistics based on the Wilcoxon scores rather than estimated *p*-values based on continuous approximations.

Exact analysis is available for both the WILCOXON and MEDIAN options in PROC NPAR1WAY. You can specify which of these scores you want to use to compute the exact *p*-values by adding either one or both of these options to the EXACT statement. If no options are listed in the EXACT statement, exact *p*-values are computed for all the linear rank statistics requested in the PROC NPAR1WAY statement.

You should exercise care when choosing to use the EXACT statement with PROC NPAR1WAY. Computational time can be prohibitive depending on the number of groups, the number of distinct response variables, the total sample size, and the speed and memory available on your computer. You can terminate exact computations and exit PROC NPAR1WAY at any time by pressing the system interrupt key and choosing to stop computations.

```
proc npar1way data=sasuser.b_ven wilcoxon;
  class brand;
  var wear;
  exact;
run;          /* ssdemo14 */
```

## Wood Veneer Wear Data

## The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable wear  
Classified by Variable brand

brand	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Acme	4	31.50	26.0	5.846522	7.8750
Champ	4	34.50	26.0	5.846522	8.6250
Ajax	4	12.00	26.0	5.846522	3.0000

Average scores were used for ties.

## Kruskal-Wallis Test

Chi-Square	5.8218
DF	2
Asymptotic Pr > Chi-Square	0.0544
Exact Pr >= Chi-Square	0.0480

In the PROC NPAR1WAY output shown above, the exact  $p$ -value is .0480, which is significant at  $\alpha=.05$ . Note the difference between the exact  $p$ -value and the  $p$ -value based on the chi-square approximation.

## C.5 Partial Leverage Plots

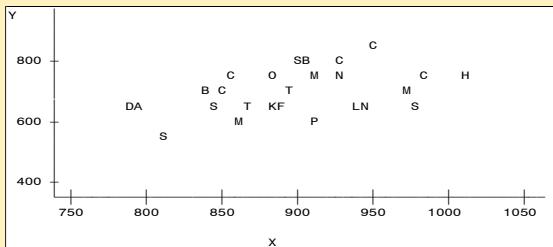
### Partial Leverage Plots

- Producing scatter plots of the response (Y) versus each of the possible predictor variables (the X's) is recommended.
- However, in the multiple regression situation, these plots can be somewhat misleading because Y might depend upon the other X's not accounted for in the plot.
- Partial leverage plots compensate for this limitation of the scatter plots.

27

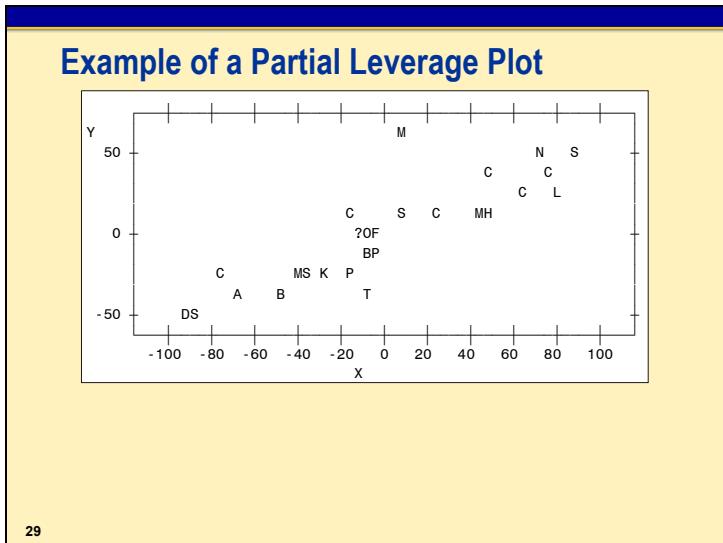
A *partial leverage plot* is a graphical method for visualizing the test of significance for the parameter estimates in the full model. The plot is basically a plot of the residuals from two partial regressions.

### Example of a Scatter Plot



28

In this scatter plot, there are no obvious influential observations.



29

In the partial leverage plot above, the observation labeled **M** stands out from the others. It did not stand out in the simple scatter plot.

The partial leverage plot revealed the outlying observation, but the scatter plot did not. This is because partial leverage plots are more sensitive to the influence of data points on individual parameter estimates.

Thus, partial regression leverage plots are graphical methods that enable you to see the effect of a single variable in a multiple regression setting.

### Partial Leverage Plots

Presume that you are performing a multiple linear regression with  $Y$  as the dependent variable and  $X_1$ ,  $X_2$ , and  $X_3$  as the independent variables.

To create a partial leverage plot for  $X_2$ :

- regress  $Y$  on  $X_1$  and  $X_3$ . These residuals are the vertical axis of the partial leverage plot.
- regress  $X_2$  on  $X_1$  and  $X_3$ . These residuals are the horizontal axis of the partial leverage plot.

30

In the example shown, there are three partial leverage plots, one for each independent variable.

In general terms, for a partial leverage plot of the independent variable  $X_r$ ,

- the vertical axis is the residuals from a regression of  $Y$  regressed on all  $X$ 's except  $X_r$
- the horizontal axis is the residuals from a regression of  $X_r$  regressed on all other  $X$ 's.



## Partial Leverage Plots

Example: Generate and interpret partial leverage plots for the BEST4 variable model.

```
options ps=35;

proc reg data=sasuser.b_fitness;
  best4: model oxygen_consumption=runtime age run_pulse
            maximum_pulse / partial;
  id name;
  title 'Producing Partial Leverage Plots';
run;
quit;

options ps=54;          /* ssdemo15 */
```

Selected MODEL statement option:

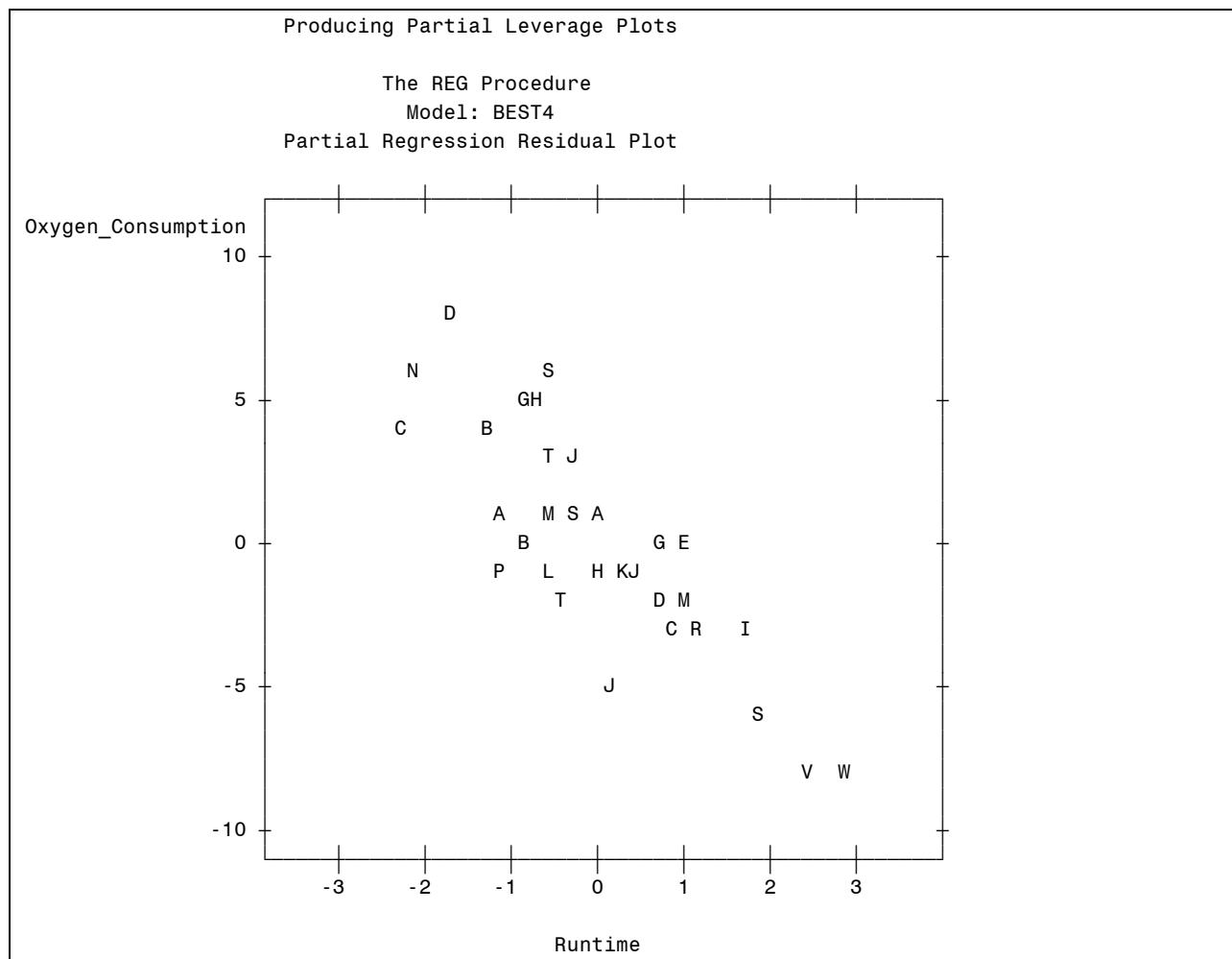
PARTIAL generates partial leverage plots for all predictor variables in the model.

When you use an ID statement with the PARTIAL option, the first nonblank character of the ID variable is used as the symbol in the plots. If two observations are too close together in a plot, a question mark is printed instead of their symbols.

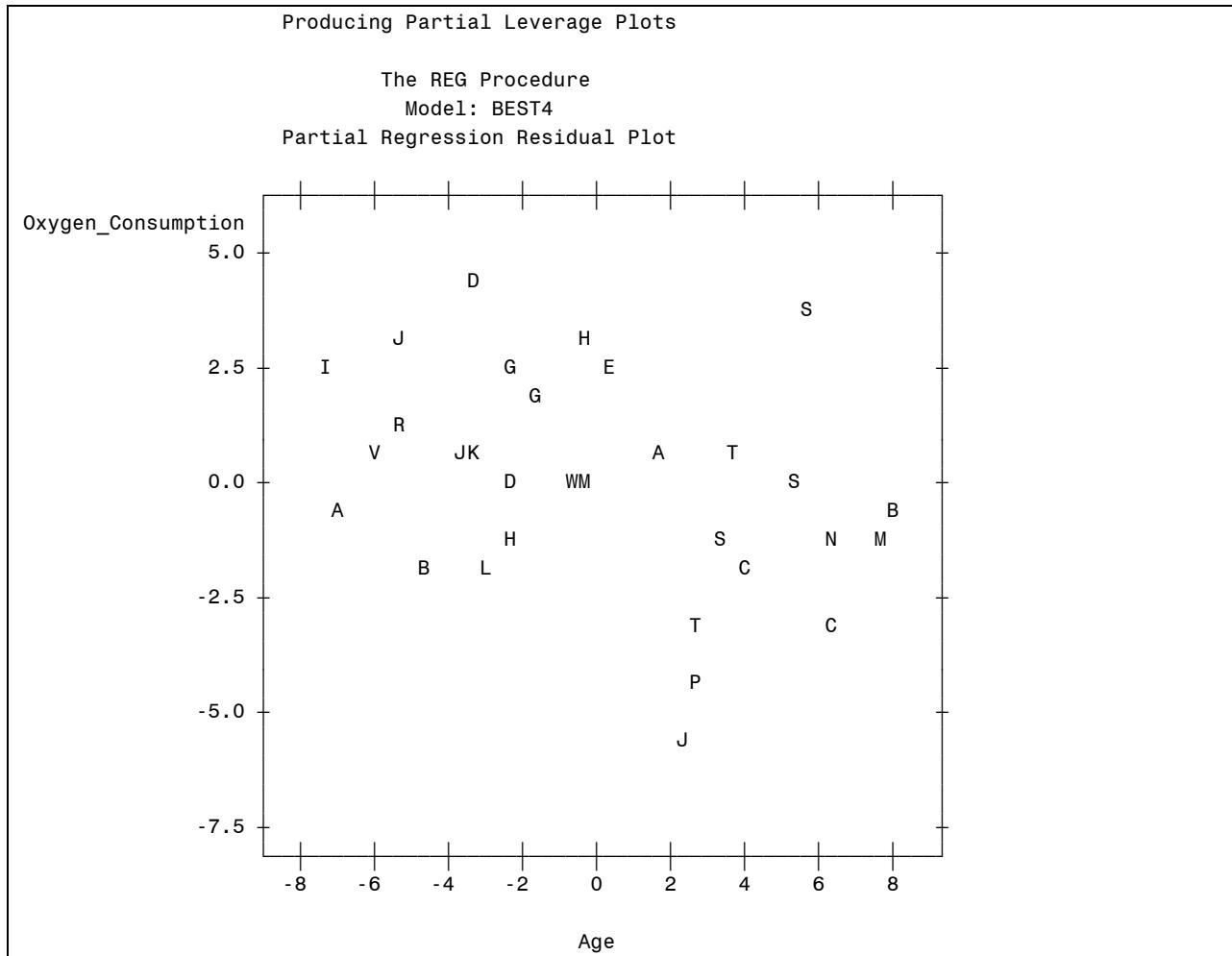
You usually do not look at the INTERCEPT plot.

### Partial PROC REG Output

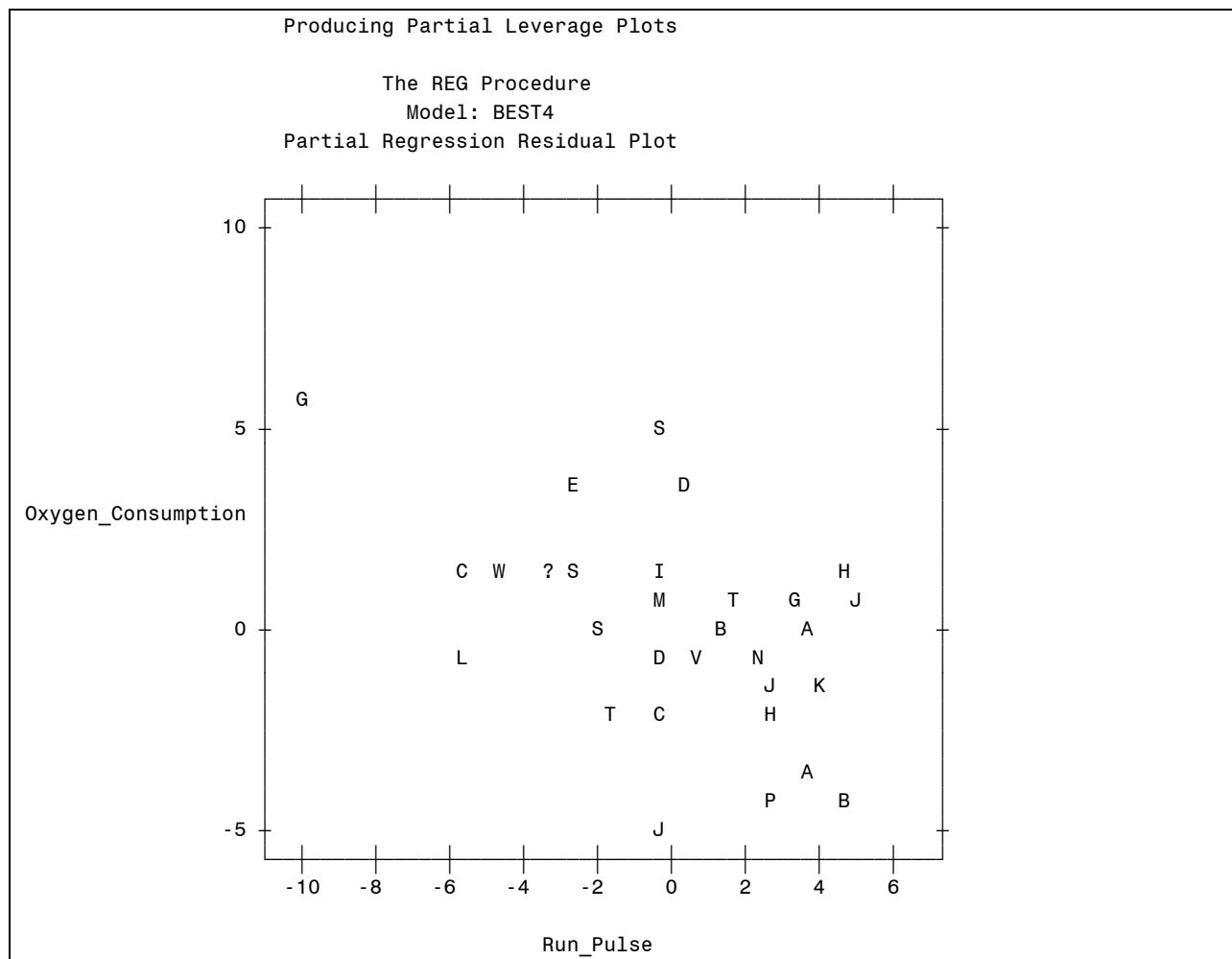
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	97.16952	11.65703	8.34	<.0001
Runtime	1	-2.77576	0.34159	-8.13	<.0001
Age	1	-0.18903	0.09439	-2.00	0.0557
Run_Pulse	1	-0.34568	0.11820	-2.92	0.0071
Maximum_Pulse	1	0.27188	0.13438	2.02	0.0534



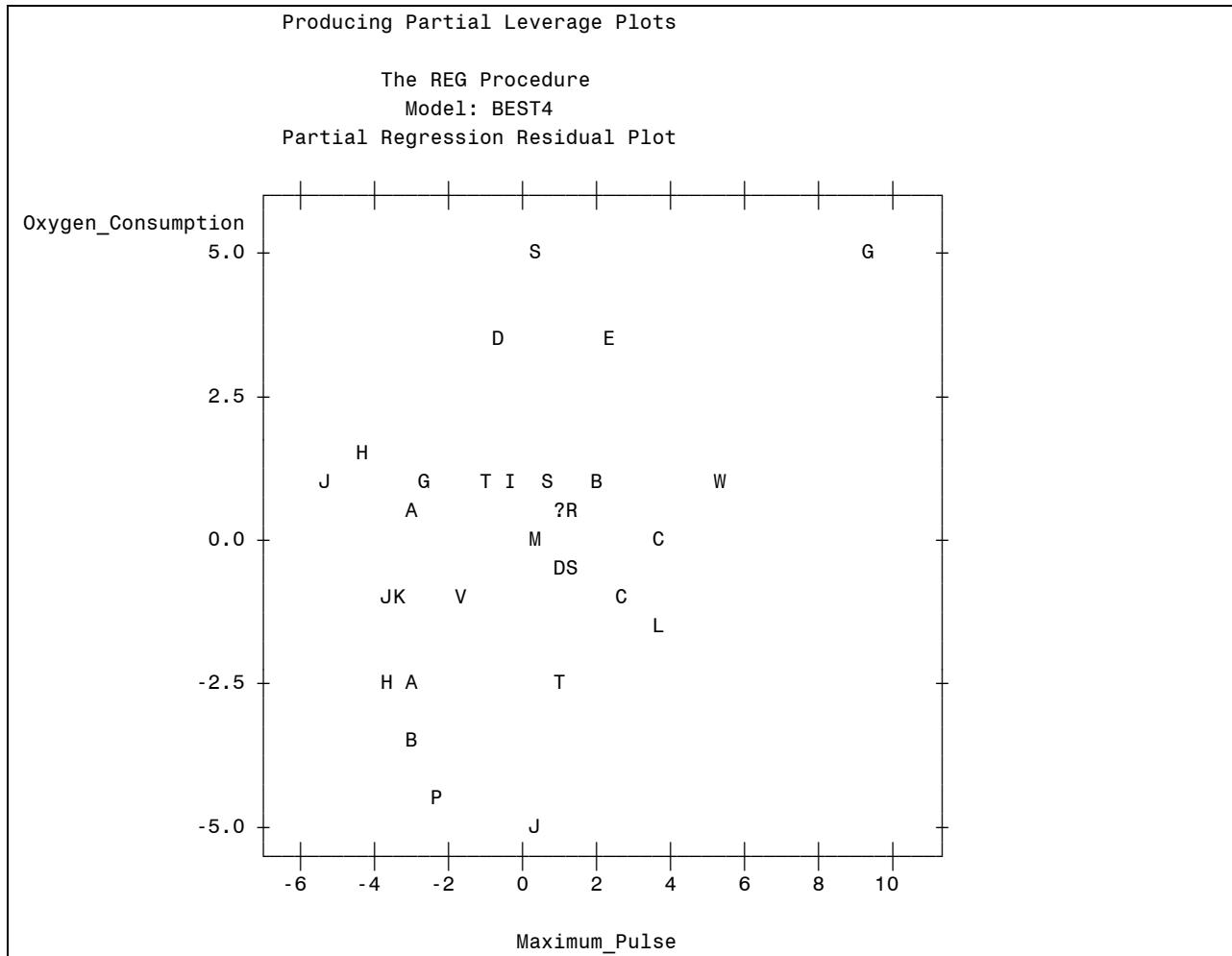
The slope of `runtime` is -2.77576. There do not appear to be any observations that stand out in the `runtime` plot.



The slope of **age** is -0.18903. None of the observations appear to have a dramatic influence on the slope of **age**. However, Sammy (the point indicated by the **S** at the top right of the graph) might have some influence.



The slope of `run_pulse` is -0.34568. Gracie appears to be influential in the slope of `run_pulse`.



The slope of **maximum\_pulse** is 0.27188. Gracie also appears to be influential on the slope of **maximum\_pulse**.



SAS/INSIGHT software produces high-resolution partial leverage plots.

### Summary of Partial Leverage Plots

No strong patterns are obvious in any of the plots.  
Consequently, it appears that the model fits the data well.  
Gracie appears to have some strong influence on the  
slopes of `run_pulse` and `maximum_pulse`.  
Sammy might have some influence on the slope of `age`.

32

For data sets that have a relatively small number of observations, such as the fitness example, identifying observations in partial leverage plots is not too much of a problem. However, for data sets with a large number of observations, it can be a problem to identify individual observations. Thus, conducting a numerical evaluation using the INFLUENCE option in the MODEL statement might be more appropriate.

# Appendix D Percentile Definitions

D.1 Calculating Percentiles .....	D-2
-----------------------------------	-----

## D.1 Calculating Percentiles

### Using the UNIVARIATE Procedure

Example: Calculate the 25th percentile for the following data using the five definitions available in PROC UNIVARIATE:

1        3        7        11        14

For all of these calculations (except definition 4), you use the value  $np=(5)(0.25)=1.25$ . This can be viewed as an observation number. However, there is obviously no observation 1.25.

**Definition 1** returns a weighted average. The value returned is 25% (25% is the fractional part of 1.25 expressed as a percentage) of the distance between observations 1 and 2:

$$\text{percentile} = 1 + (0.25)(3 - 1) = 1.5$$

**Definition 2** rounds to the nearest observation number. Thus, the value 1.25 is rounded to 1 and the first observation, 1, is taken as the 25th percentile. If  $np$  were 1.5, then the second observation is selected as the 25th percentile.

**Definition 3** always rounds up. Thus, 1.25 rounds up to 2 and the second data value, 3, is taken as the 25th percentile.

**Definition 4** is a weighted average similar to definition 1, except instead of using  $np$ , definition 4 uses  $(n+1)p=1.5$ .

$$\text{percentile} = 1 + (0.5)(3 - 1) = 2$$

**Definition 5** rounds up to the next observation number unless  $np$  is an integer, in which case an average of the observations represented by  $np$  and  $(np + 1)$  is calculated. In this example, definition 5 rounds up, and the 25th percentile is 3.

# Appendix E Advanced Programs

E.1 Interaction Plot .....	E-2
----------------------------	-----

## E.1 Interaction Plot

A DATA step with two DO loops is used to create a data set with plotting points. The data points include all possible combinations of **Gender** and **IncLevel**.

```
data plot;
  do Gender='Female','Male';
    do IncLevel=1,2,3;
      output;
    end;
  end;
run;

proc format;
  value incfmt 1='Low'
            2='Medium'
            3='High';
run;
```

To visualize the interaction, use the SCORE statement (new in SAS®9). The beta estimates from the model will be applied to the new observations, sorted in the DATA= option. New variables are created and stored in the OUT= option that represent the probabilities for each level of the outcome variable.

```
proc logistic data=sasuser.b_sales_inc
  noint;
  class gender (param=ref ref='Male')
    inclevel (param=ref ref='3');
  model purchase(event='1')=gender inclevel gender*inclevel;
  score data=plot out=scored;
run;
```

Selected LOGISTIC procedure statement:

**SCORE** creates a data set that contains all the data in the DATA= data set together with posterior probabilities and, optionally, prediction confidence intervals.

Selected SCORE statement options:

**DATA=** names the SAS data set that you want to score.

**SCORE=** enables you to score new data sets and output the scored values and, optionally, the corresponding confidence limits into a SAS data set.

VIEWTABLE: Posterior Probabilities for DATA=WORK.PLOT.					
	Gender	IncLevel	I_purchase	P_0	P_1
1	Female		1 0	0.6179775281	0.3820224719
2	Female		2 0	0.5633802817	0.4366197183
3	Female		3 0	0.55	0.45
4	Male		1 0	0.8139509384	0.1860490616
5	Male		2 0	0.7945198214	0.2054801786
6	Male		3 1	0.4933333333	0.5066666667

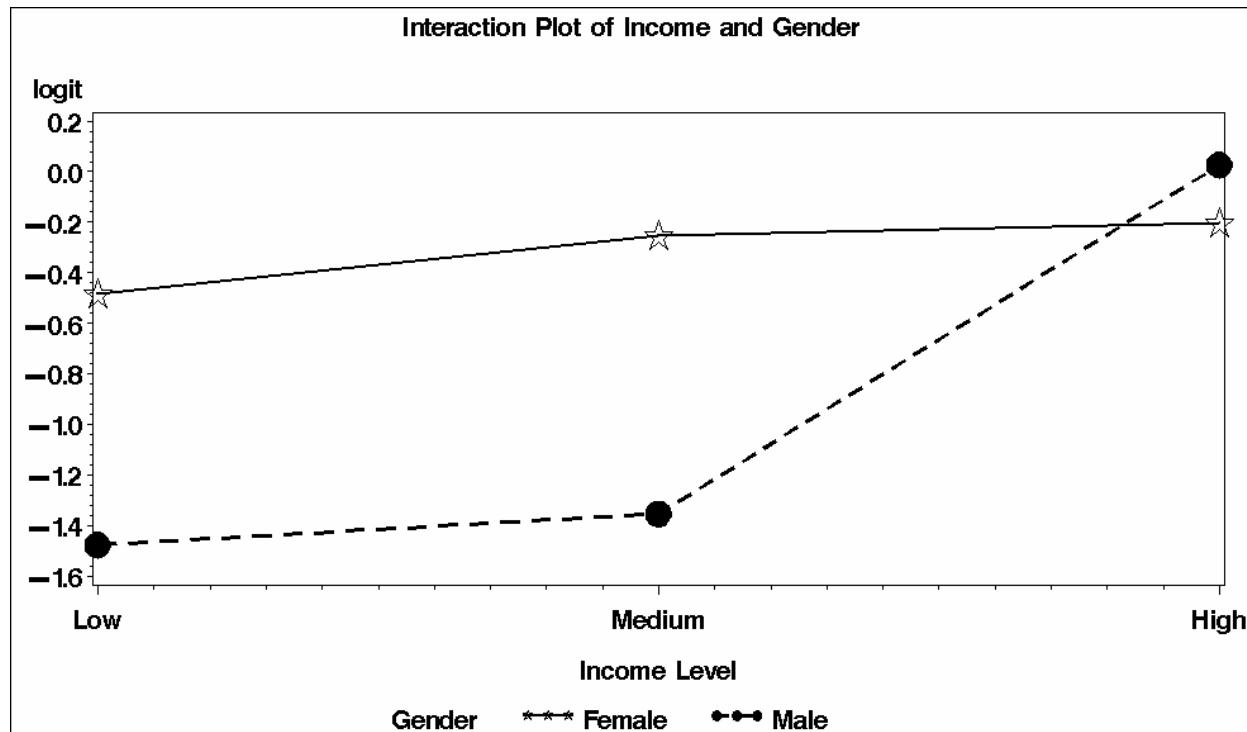
The logit for **purchase** is the natural log of the probability that **purchase**=1, divided by the probability that **purchase**=0. Use the variables created in the SCORE= data set, **p\_1** and **p\_0** respectively, to calculate **logit**.

```
data scored2;
  set scored;
  logit=log(p_1/p_0);
run;
```

The GPLOT procedure is used to create the interaction plot.

```
options ps=50 ls=64;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;

proc gplot data=scored2;
  plot logit*inclevel=gender;
  symbol1 c=black w=2 h=3 line=1 i=join v='=';
  symbol2 c=black w=2 h=3 line=3 i=join v=dot;
  format inclevel incfmt. ;
  label inclevel='Income Level';
  title 'Interaction Plot of Income and Gender';
run;
quit;
```



Now create the same plot using ActiveX.

```
ods listing close;

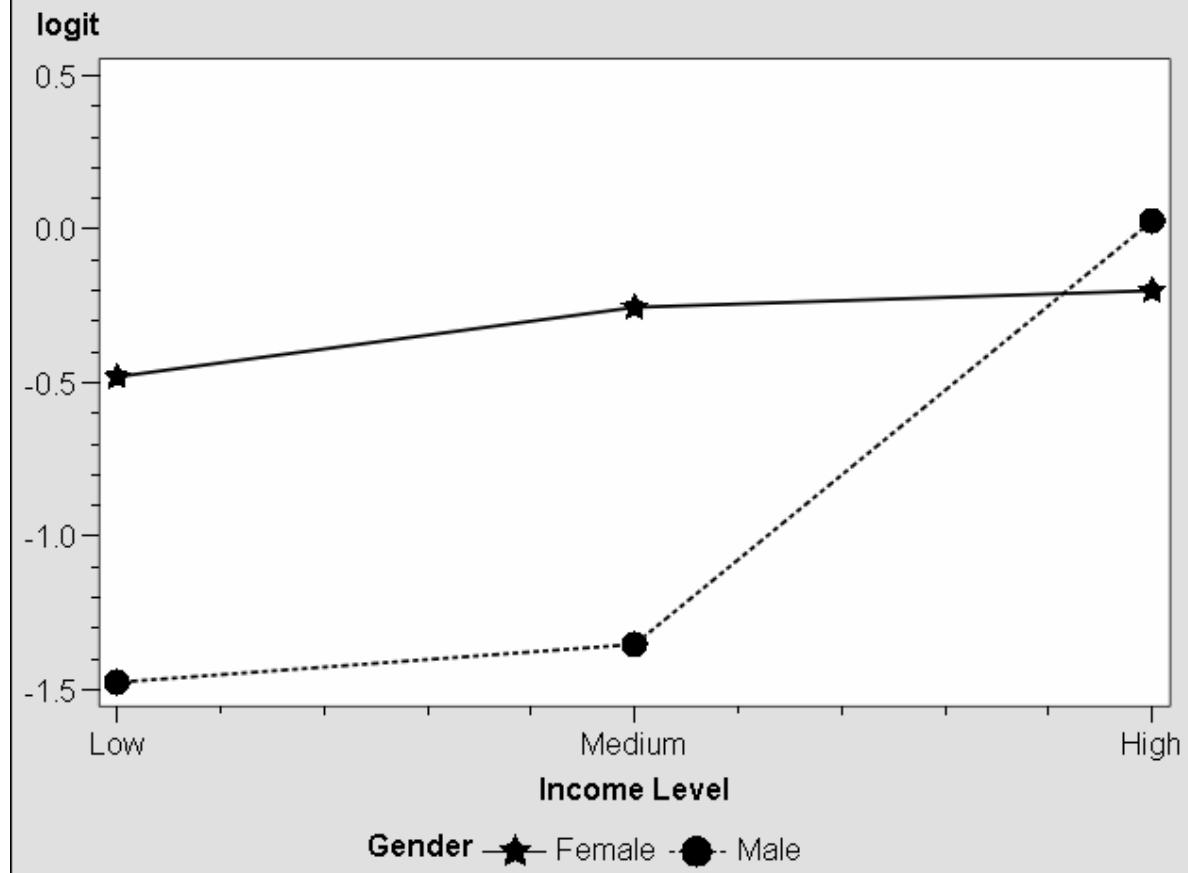
options device=activex;

ods html
  body='elogitplot.htm';

proc gplot data=scored2;
  plot logit*inclevel=gender;
  symbol1 c=black w=2 h=1.3 line=1 i=join v=star;
  symbol2 c=black w=2 h=1.3 line=3 i=join v=circle;
  format inclevel incfmt.;
  label inclevel='Income Level';
  title 'Interaction Plot of Income and Gender - Activex';
run;
quit;

ods html close;
ods listing;
```

### Interaction Plot of Income and Gender - Activex



# Appendix F Randomization Technique

F.1 Randomize Paints.....F-2

## F.1 Randomize Paints

A DATA step is used to generate the 28 observations for the completely randomized experiment. Each of the seven roads is given four stripe identification numbers. The variable **random** has been generated using a seed of 47, yet any positive integer would suffice. Selected variables of the data set **stripes** are printed for verification of the data.

```
options ls=75 ps=55  nodate nonumber;

/* associate a road with a number */
proc format;
  value roadid  1='Center'
            2='Broadway'
            3='Main'
            4='Elm'
            5='Station'
            6='Park'
            7='Beech'
            ;
run;

data stripes;

  stripe_id=0;
  do r=1 to 7; /* # of roads */

    road=put(r,$roadid.);
    do s=1 to 4; /* # of paints      */
      /* 7 * 4=28 obs.  */
      stripe_id=stripe_id + 1;
      random=ranuni(47);
      output;
    end; /* s */
  end; /* r */

  drop
    r s;
run;

proc print data=stripes;
  id road;
  var stripe_id;
  title 'Stripe-ID for each Road';
run;

proc sort data=stripes;
  by random;
run;
```

The data set **stripes** is now sorted by the variable **random**, and the four paints, identified with values Paint-1, Paint-2, Paint-3, and Paint-4 are assigned to each of the 28 stripes.

```
/* generate values for paint based on the MOD function, */
/* described below.                                         */
proc format;
  value paintid  0='Paint-4'
            1='Paint-2'
            2='Paint-1'
            3='Paint-3'
            ;
run;

/* associate the modular of 4 with a paint via the */
/* format PAINTID                                     */
data paints;
  set stripes;
  by random; /* NOTE: data is sorted by this variable */

  break=mod(_n_,4);/* _n_ is observation number.      */
    /* MOD computes the remainder of   */
    /* the first argument divided by */
    /* the second argument.          */

  select (break); /* use select instead of if-then-else */
  when (0) assigned_paint=put(break,$paintid.);
  when (1) assigned_paint=put(break,$paintid.);
  when (2) assigned_paint=put(break,$paintid.);
  when (3) assigned_paint=put(break,$paintid.);
  otherwise;
  end;

  drop
    break random;
run;

proc datasets library=work nolist;
  delete stripes;
run;
```

The data set **paints** is now sorted in two ways: by the paint that was assigned to each stripe and by the road/stripe combination. The latter is best used in the field.

```
proc sort data=paints out=grpdpaints;
  by assigned_paint;
run;

proc print data=grpdpaints;
  by assigned_paint;
  id assigned_paint;
  var road stripe_id;
  title 'Paint #(1,2,3 or 4) ... on Road/Stripe-ID';
run;

proc sort data=paints out=grpdpaints;
  by road stripe_id;
run;

proc print data=grpdpaints;
  by road;
  id road;
  var stripe_id assigned_paint;
  title 'On Road/Stripe-ID, Assign Paint #(1,2,3, or 4)';
run;
```

## Stripe-ID for each Road

road	stripe_id
Center	1
Center	2
Center	3
Center	4
Broadway	5
Broadway	6
Broadway	7
Broadway	8
Main	9
Main	10
Main	11
Main	12
Elm	13
Elm	14
Elm	15
Elm	16
Station	17
Station	18
Station	19
Station	20
Park	21
Park	22
Park	23
Park	24
Beech	25
Beech	26
Beech	27
Beech	28

Paint #(1,2,3 or 4) ... on Road/Stripe-ID

assigned_paint	road	stripe_id
----------------	------	-----------

Paint-1	Main	10
	Broadway	5
	Park	22
	Broadway	7
	Station	20
	Center	3
	Elm	16

Paint-2	Elm	13
	Park	23
	Beech	25
	Main	11
	Main	12
	Beech	28
	Station	19

Paint-3	Elm	14
	Main	9
	Station	18
	Broadway	6
	Center	1
	Station	17
	Elm	15

Paint-4	Center	4
	Park	21
	Park	24
	Center	2
	Beech	26
	Beech	27
	Broadway	8

On Road/Stripe-ID, Assign Paint #(1,2,3, or 4)

road	stripe_id	assigned_paint
Beech	25	Paint-2
	26	Paint-4
	27	Paint-4
	28	Paint-2
Broadway	5	Paint-1
	6	Paint-3
	7	Paint-1
	8	Paint-4
Center	1	Paint-3
	2	Paint-4
	3	Paint-1
	4	Paint-4
Elm	13	Paint-2
	14	Paint-3
	15	Paint-3
	16	Paint-1
Main	9	Paint-3
	10	Paint-1
	11	Paint-2
	12	Paint-2
Park	21	Paint-4
	22	Paint-1
	23	Paint-2
	24	Paint-4
Station	17	Paint-3
	18	Paint-3
	19	Paint-2
	20	Paint-1

# **Appendix G Basic Statistics Guidelines for Analysis**

**G.1 Guidelines for Analysis.....G-2**

## G.1 Guidelines for Analysis

Basic Statistics Guidelines for Analysis			
Predictor (X, Independent, Regressor, Effect, Explanatory)	Categorical	Continuous	Categorical and Continuous
Response (Y, Dependent, Target)			
<b>Continuous</b>	Analysis of Variance (Chapter 2)	Regression (Chapters 3 and 4)	Analysis of Covariance or Regression with Dummy Variables (Statistics II)
<b>Categorical</b>	Crosstabulation/ Contingency Table or Logistic Regression (Chapter 5)	Logistic Regression (Chapter 5)	Logistic Regression (Chapter 5)

# Appendix H Additional Resources

H.1 References .....	H-2
----------------------	-----

## H.1 References

- Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.
- Allison, P. 1999. *Logistic Regression Using the SAS® System: Theory and Application*. Cary, N.C.: SAS Institute Inc.
- Anscombe, F. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27:17-21.
- Belsey, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.
- Findley, D.F. and E. Parzen. 1995. "A Conversation with Hirotugu Akaike." *Statistical Science* Vol. 10, No. 1:104-117.
- Hocking, R. R. 1976. "The Analysis and Selection of Variables in Linear Regression." *Biometrics* 32:1-49
- Mallows, C. L. 1973. "Some Comments on  $C_p$ ." *Technometrics* 15:661-675.
- Marquardt, D. W. 1980. "You Should Standardize the Predictor Variables in Your Regression Models." *Journal of the American Statistical Association* 75:74-103.
- Myers, R. H. 1990. *Classical and Modern Regression with Applications, Second Edition*. Boston: Duxbury Press.
- Neter, J., M. H. Kutner, W. Wasserman, and C. J. Nachtsheim. 1996. *Applied Linear Statistical Models*, Fourth Edition. New York: WCB McGraw Hill.
- Rawlings, J. O. 1988. *Applied Regression Analysis: A Research Tool*. Pacific Grove, CA: Wadsworth & Brooks.
- Santner, T.J. and D. E. Duffy. 1989. *The Statistical Analysis of Discrete Data*. New York: Springer-Verlag.
- Welch, B. L. 1951. "On the Comparison of Several Mean Values: An Alternative Approach." *Biometrika* 38:330-336.