

Introduction to Statistics Using SAS[®] : ANOVA, Linear Regression and Logistic Regression

Course Notes

Introduction to Statistics Using SAS[®] : ANOVA, Linear Regression and Logistic Regression Course Notes was developed by Gemma Robson. Additional contributions were made by Richard Hunt, Marc Huber, Catherine Truxillo, Azhar Nizam, Mike Patetta, Dan Kelly, Bob Lucas, Jill Tao, Paul Marovich, and artwork by Stanley Goldman. 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.

Introduction to Statistics Using SAS[®] : ANOVA, Linear Regression and Logistic Regression Course Notes

Copyright © 2012 SAS Institute Inc. Cary, NC, 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 NULL, course code ST092_UK, prepared date 05Jan2012.

ST092_UK_001

ISBN NULL

Table of Contents

Course Description	vii
Prerequisites	viii
Chapter 1 Course Logistics.....	1-1
1.1 Course Specifics	1-3
Chapter 2 Introduction to Statistics.....	2-1
2.1 Fundamental Statistical Concepts.....	2-3
Demonstration: Descriptive Statistics	2-15
2.2 Picturing Distributions.....	2-18
Demonstration: Examining Distributions	2-32
2.3 Confidence Intervals for the Mean	2-38
Chapter 3 Testing Business Questions	3-1
3.1 Hypothesis Testing.....	3-3
Demonstration: One-Sample <i>t</i> -Test.....	3-19
Chapter 4 Testing Two or More Groups.....	4-1
4.1 Comparing Two Groups.....	4-3
Demonstration: Two-Sample <i>t</i> -Test	4-11
4.2 One-Way ANOVA.....	4-16
Demonstration: The GLM Procedure- comparing two groups	4-25
Demonstration: The GLM Procedure- more than two groups	4-32
Demonstration: Post Hoc Pairwise Comparison	4-45
Chapter 5 Exploratory Data Analysis	5-1
5.1 Exploratory Data Analysis	5-3

Demonstration: Data Exploration, Correlations, and Scatter Plots	5-17
Chapter 6 Simple Linear Regression.....	6-1
6.1 Simple Linear Regression.....	6-3
Demonstration: Performing Simple Linear Regression	6-14
Demonstration: Confidence and Predicted Limits.....	6-19
Demonstration: Producing Predicted Values.....	6-22
6.2 Examining Assumptions	6-25
Demonstration: Residual Plots.....	6-33
Chapter 7 Multiple Linear Regression	7-1
7.1 Concepts of Multiple Regression	7-3
Demonstration: Fitting a Linear Regression Model with Two Predictor Variables.....	7-9
Demonstration: Fitting a Multiple Linear Regression Model.....	7-16
7.2 Model Building and Interpretation	7-18
Demonstration: Stepwise Regression.....	7-24
Demonstration: Examining Assumptions.....	7-35
Chapter 8 Categorical Data Analysis	8-1
8.1 Describing Categorical Data.....	8-3
Demonstration: Examining Distributions	8-14
Demonstration: Ordering Values in a Frequency Table	8-23
8.2 Tests of Association	8-26
Demonstration: Chi-Square Test	8-33
Demonstration: Detecting Ordinal Associations.....	8-41
Chapter 9 Logistic Regression.....	9-1
9.1 Introduction to Logistic Regression.....	9-3
Demonstration: Binary Logistic Regression	9-15

Chapter 10 Multiple Logistic Regression 10-1

10.1	Multiple Logistic Regression.....	10-3
	Demonstration: Multiple Logistic Regression.....	10-7
10.2	Multiple Logistic Regression with Interactions (Optional).....	10-15
	Demonstration: Multiple Logistic Regression with Interactions	10-19
10.3	Logit Plots (Self-Study)	10-30
	Demonstration: Plotting Estimated Logits	10-34

Appendix A Additional Topics A-1

A.1	ODS Statistical Graphics	A-3
	Demonstration: ODS Statistical Graphics Using PROC CORR.....	A-6
	Demonstration: ODS HTML Output Using the STYLE=STATISTICAL Option.....	A-9
	Demonstration: Using the ODS Graphics Editor.....	A-12
A.2	Paired <i>t</i> -Tests	A-17
	Demonstration: Paired <i>t</i> -Test	A-19
A.3	Fishers Exact p-values	A-21
	Demonstration: Exact <i>p</i> -Values for the Pearson Chi-Square Test.....	A-27
A.4	Nonparametric ANOVA	A-30
	Demonstration: The NPAR1WAY Procedure	A-36
A.5	Partial Leverage Plots	A-49
	Demonstration: Partial Leverage Plots	A-52

Chapter 2 Advanced Programs 2-1

2.1	Interaction Plot.....	2-3
-----	-----------------------	-----

Chapter 3 Additional Resources 3-1

3.1	References.....	3-3
-----	-----------------	-----

Appendix D Exercises.....D-1

Chapter 2	D-3
Chapter 3	D-6
Chapter 4	D-7
Chapter 5	D-11
Chapter 6	D-15
Chapter 7	D-18
Chapter 8	D-21
Chapter 9	D-24
Chapter 10	D-25

Appendix E SolutionsE-1

Chapter 2	E-3
Chapter 3	E-11
Chapter 4	E-13
Chapter 5	E-24
Chapter 6	E-31
Chapter 7	E-36
Chapter 8	E-47
Chapter 9	E-52
Chapter 10	E-55

Course Description

This course is for SAS software users who perform statistical analysis using SAS/STAT software. The focus is on t-tests, ANOVA, linear regression and logistic regression. This course (or equivalent knowledge) is a prerequisite to many of the courses in the statistical analysis curriculum.

To learn more...



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. You can also find this information on the Web at support.sas.com/training/ as well as in the Training Course Catalog.



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. Customers outside the USA, please contact your local SAS office.

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

Prerequisites

Before selecting this course, you should

be able to execute SAS programs and create SAS data sets. You can gain this experience by completing the SAS® Programming 1: Essentials course.

No statistical knowledge is necessary although knowledge of hypothesis testing and p-values is advantageous.

x

For Your Information

Chapter 1 Course Logistics

1.1 Course Specifics.....1-3

1.1 Course Specifics

Objectives

- Explain the Learner Response System.
- Explain the naming convention that is used for the course files.
- Discuss the Confidence and Relevance Learning Points.
- Describe at a high level some of the data used.

2

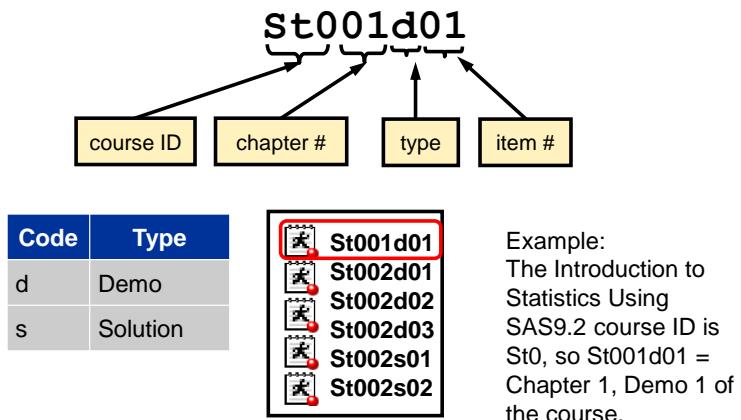
Learner Response System

The Learner Response System (LRS) is a UK initiative that:

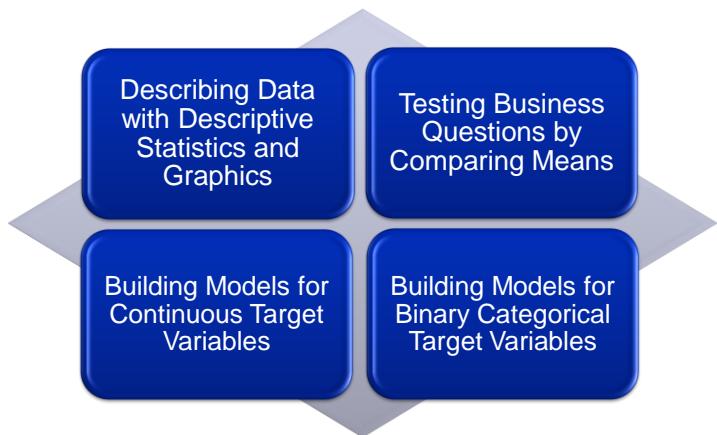
- Aids interaction and engagement to promote learning.
- Helps monitor achievement throughout the course by giving instant feedback.
- Used to create an Individual Learner Response Report to customise learning recommendations. This also includes a personal action plan to enhance and develop learning.

3

Filename Conventions



Confidence and Relevance Learning Points



6

Describing Data with Descriptive Statistics and Graphics includes:

- Fundamental Statistical Concepts
- Picturing Distributions
- Confidence Intervals for the Mean
- Describing Categorical Data

Testing business questions by comparing means includes:

- Hypothesis Testing
- Comparing Groups
- One-Way ANOVA

Building Models for Continuous Target Variables includes:

- Exploratory Data Analysis
- Simple Linear Regression
- Examining Assumptions
- Concepts of Multiple Linear Regression
- Model Building and Interpretation

Building Models for Binary Categorical Target Variables includes:

- Tests of Association
- Introduction to Logistic Regression
- Multiple Logistic Regression
- Multiple Logistic Regression with Interactions

Purchasing Example

Who is more likely to make a purchase?

- What are their characteristics?
- Are males more likely to make a purchase than females?
- Does income level effect purchasing?

7

Fitness Example

What are the characteristics of 'healthy' individuals?

- Are older individuals, on average, less fit than younger ones?
- Can an individuals' pulse rate at the end of a run be used to predict fitness?
- Can the time taken to complete the run be used to predict fitness?

8

Chapter 2 Introduction to Statistics

2.1 Fundamental Statistical Concepts	2-3
Demonstration: Descriptive Statistics.....	2-15
2.2 Picturing Distributions	2-18
Demonstration: Examining Distributions	2-32
2.3 Confidence Intervals for the Mean.....	2-38

2.1 Fundamental Statistical Concepts

Objectives

- Explain the purpose of statistics
- Understand the process of data analysis
- Decide what tasks to complete before you analyse your data.
- Use the MEANS procedure to produce descriptive statistics.

3

Purpose of Statistics

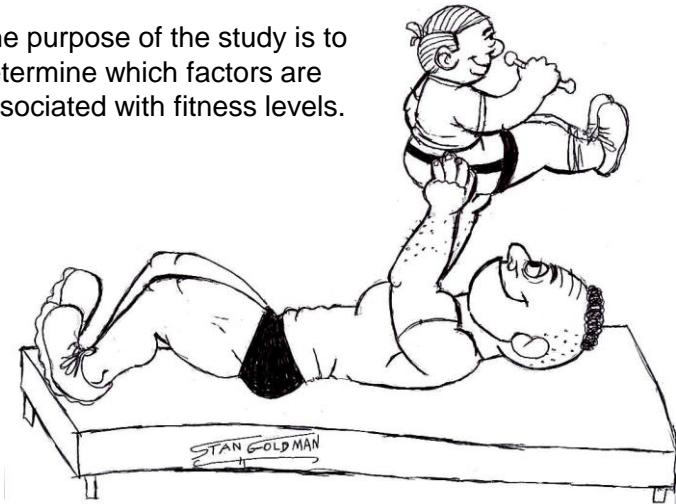
Obtain information from data to help answer questions.

4

One purpose of statistics is to help make sense of your data. Statistics provide information about your data, which you can use to help answer business questions and make informed decisions.

Defining the Problem- The Fitness Example

The purpose of the study is to determine which factors are associated with fitness levels.



5

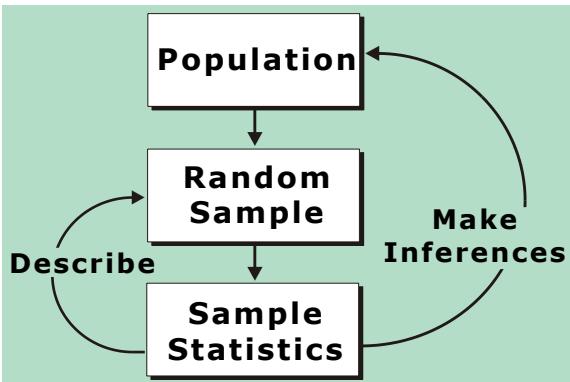
In exercise physiology, an objective measure of aerobic fitness is how fast the body can absorb and use oxygen (oxygen consumption). Subjects participated in a predetermined exercise run of 1.5 miles. Measurements of oxygen consumption as well as several other continuous measurements such as age, pulse, 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.

Define Study Questions.

- Are older individuals, on average, less fit than younger ones?
- Can an individuals' pulse rate at the end of a run be used to predict fitness?
- Can the time taken to complete the run be used to predict fitness?

6

Process of Statistical Analysis



7

Four processes are involved in statistical analysis:

1. Identify the population of interest, e.g. our customer database.
2. Draw a random sample because we are unlikely to need or have the entire database to be able to draw some conclusions.
3. Compute statistics to describe the sample.
4. Use the information from this sample to make inferences about the population of interest.

In reality this process may not happen, but the statistical methods have been developed assuming these processes and when we interpret the results of the methods we need to know what assumptions these are based on.

Defining the Problem

- Outline the purpose of the study
- Document the study questions
- Define the population of interest
- Determine the need for sampling
- Define the data collection protocol

8

When you have a particular business problem that you want to solve by examining some data, the first thing you must do is define the problem properly. The five-stage process shown in the slide is a useful guide for statistical analysis.

Marketing Example:

1. The purpose of the study is to examine our customer database to better target our products and services.
2. a) What are the characteristics of people who purchase our products?
b) Can we predict which customers will make a purchase?
c) Can we predict which customers will respond to a mailing?
3. Current customers in our database and prospective new customers.
4. We will take a random sample from our customer database and we assume this is representative of both our current customers and also prospective new customers.
5. The data have already been collected, but we need to do a lot of checking on the data before we can run any statistical analysis.

Samples

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.

9

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

A *sample* is a subset of the population. The sample should be *representative* of the population, meaning that the sample characteristics are similar to the population's characteristics.

Simple random sampling, a technique in which each member of the population has an equal probability of being selected. Random sampling can help to ensure that the sample is representative of the population.

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

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

Convenience sampling is a technique in which you select a sample that is easily available to you. This can lead to biased samples. A *biased* sample is one that is not representative of the population from which it is drawn.

Types of Data

There are two main types of data:

- Continuous
- Categorical

The type of data will determine how best to examine and summarise it.

10

In order to use the appropriate method of data summarisation and analysis, it is important to recognise the level of measurement of your data.

On a continuous scale,

- The variable has an unlimited number of possible values within a given range.
- The values are numeric only.

The variables age, weight, and annual income can all be continuous.

 Continuous data is also called interval data.

On a categorical scale,

- The variable usually has a small number of distinct values within a given range.
- The values can be character or numeric.

The variables gender, make a purchase (Yes/No) and colour are all categorical.

 Categorical data is also called Discrete data.

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

The first thing you should always do when you obtain your data is to explore and describe it. You need to check to make sure your data is error-free before proceeding with any analysis. Summary Statistics and Graphical Analysis are the easiest ways to explore and check your data.

Summary Statistics

The types of statistics we can use to describe our data are:

- Measures of central tendency
- Measures of variability
- Measures of shape

12

Measures of Central Tendency

Small sample: 1, 5, 3, 8, 11, 3, 4

$$\text{mean} = \bar{x} = \frac{1+5+3+8+11+3+4}{7} = 5$$

$$\text{median} = 4 (1, 3, 3, 4, 5, 8, 11)$$

$$\text{mode} = 3$$

13

These are three measures of central tendency or average. The *mean*, often denoted as \bar{x} is the most commonly used measure of average. The mean can only be calculated on continuous data.

For the *median* we need to rank the values from lowest to highest and select the middle value, in this example we can see the median is 4. If there is an even number of observations then we use the mean of the two middle values. We can calculate the median on continuous data and also on discrete data if there is some natural ordering to the values.

The *mode* is the most common value, so in this example we can see the mode is 3. The mode can be obtained for discrete and continuous data.

Percentiles

98
95
<u>92</u> 75th Percentile=91
90
85
<u>81</u> 50th Percentile=80
79
70
<u>63</u> 25th Percentile=59
55
47
42

third quartile

Quartiles break your data up into quarters.

first quartile

14

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

Commonly reported percentile values are

- the 25th percentile, also called the first quartile
- the 50th percentile, also called the median
- the 75th percentile, also called the third quartile.

Measures of Variability

Small sample: 1, 3, 3, 4, 5, 8, 11

- Range = 11 - 1 = 10
- Interquartile Range = 8 - 3 = 5
- Variance = 11.7
- Standard Deviation = 3.4

15

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

The *range* is the easiest to calculate and is simply the largest value minus the smallest. However, this measure of variability only depends on the two extreme values and is rarely used in practice.

The *interquartile range* (3rd quartile minus 1st quartile) is more often used as it is the range of the middle 50% of the data and is therefore less sensitive to extreme data values.

The *variance* is calculated as $\sum \frac{(x_i - \bar{x})^2}{n-1}$ where x_i is each individual value and \bar{x} is the mean. The variance is a measure of the average squared difference between each point and the mean, but because we have a squared term the variance is difficult to interpret.

The *standard deviation* is the square root of the variance, which means it is easier for us to interpret. A standard deviation of 3.4 means the average variability about the mean is about 3.4 units.

The standard deviation is dependent on the scale on which the data are measured, so the standard deviation of weight measured in kg will be different to the standard deviation of weight measured in pounds. If you want to compare the variability of variables across different measurement scales you can use the *coefficient of variation*.

The coefficient of variation is calculated as:

$$\frac{\text{standard deviation}}{\text{mean}} * 100\% = \frac{3.4}{5} * 100\% = 68\%$$

Example: Customer Call Centre

Average time it takes for call centre to answer the phone.

- Service level agreement: Average time should be 60 seconds
- Newly formed team: Are they achieving minimum required standard?

The MEANS Procedure

General form of the MEANS procedure:

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

17

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** in the command box. This opens the Help window, which accesses the entire SAS documentation. After locating the appropriate procedure, select **syntax** to see all options available for that procedure.



Descriptive Statistics

The following is a summary of what you will accomplish in this demonstration:

- use the PRINT procedure to familiarise yourself with the data
- use the default MEANS Procedure to produce simple descriptive statistics.
- Dictate which statistics the MEANS Procedure produces.
- Navigate the SAS Help.

```
/*st002d01.sas*/
options nodate nonumber ls=95 ps=80;
proc print data=st092.phone_new (obs=10);
  title 'Listing of the Phone_new Data Set';
run;
```

PROC PRINT output.

Listing of the Phone_new Data Set

Obs	team	time
1	New	66
2	New	80
3	New	62
4	New	52
5	New	36
6	New	58
7	New	70
8	New	54
9	New	53
10	New	114

```
/*st002d01.sas*/
proc means data=st092.phone_new maxdec=2 fw=10;
  var time;
  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 variables.

FW= specifies the field width for all columns.

PROC MEANS Output.

Descriptive Statistics Using PROC MEANS				
The MEANS Procedure				
Analysis Variable : time				
N	Mean	Std Dev	Minimum	Maximum
100	64.38	17.92	19.00	114.00

By default, Proc MEANS prints the number of non-missing observations for our analysis variable, the mean, the standard deviation, the minimum value, and the maximum value.

```
/*st002d01.sas*/
proc means data=st092.phone_new
            maxdec=2 fw=10
            n mean median std var q1 q3;
    var time;
    title 'Selected Descriptive Statistics for Time to Answer
the Phone';
run;
```

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

PROC MEANS Output

Selected Descriptive Statistics for Time to Answer the Phone						
The MEANS Procedure						
Analysis Variable : time						
N	Mean	Median	Std Dev	Variance	Lower Quartile	Upper Quartile
100	64.38	62.00	17.92	321.23	53.00	78.00

The Output indicates that:

- The mean of the data is 64.38 seconds.
- The standard deviation is 17.92, which means that the average variability around the mean is approximately 18 seconds.

The SAS Help holds a wealth of information on all procedures and functionality of SAS.

For more information on statistics available on the PROC MEANS statement, use the SAS Help.

Go to Help>SAS Help and Documentation.

Then follow the path on the contents tab for a list of statistic- keywords..

SAS Products

Base SAS

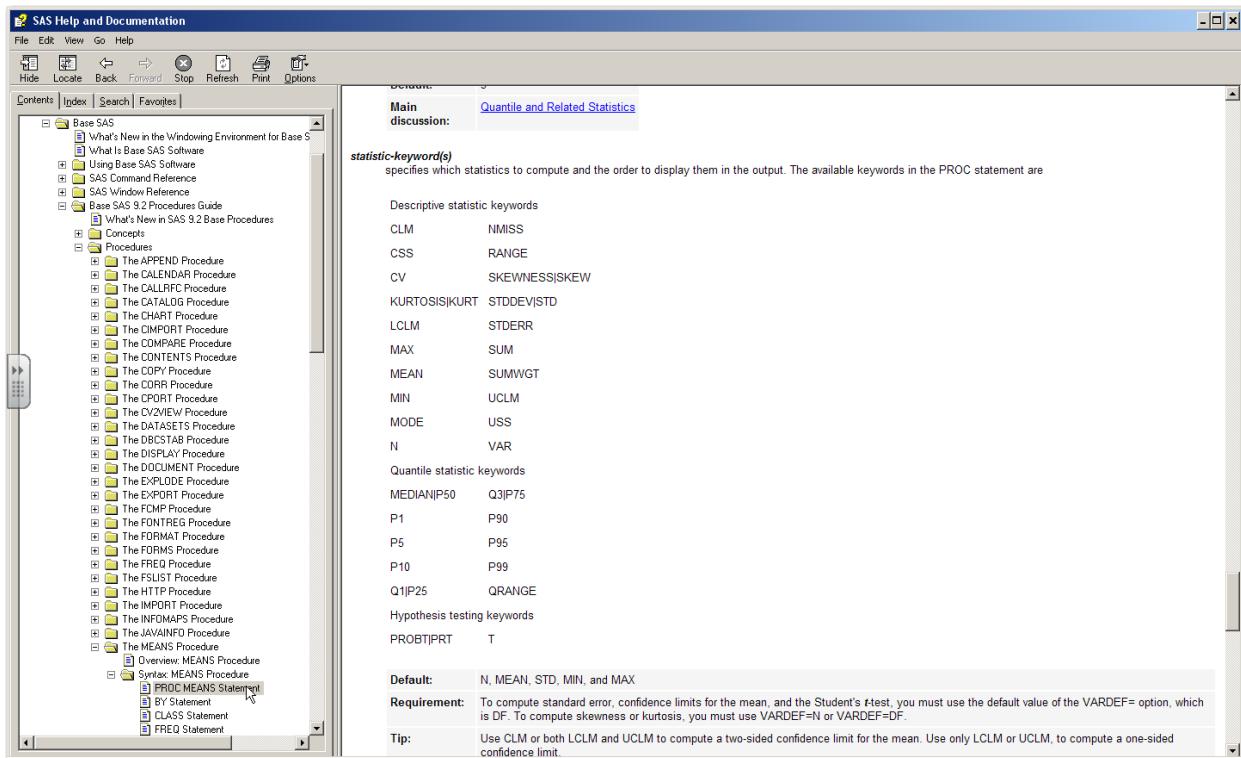
Base SAS 9.2 Procedures Guide

Procedures

The MEANS Procedure

Syntax: MEANS Procedure

PROC MEANS Statement.



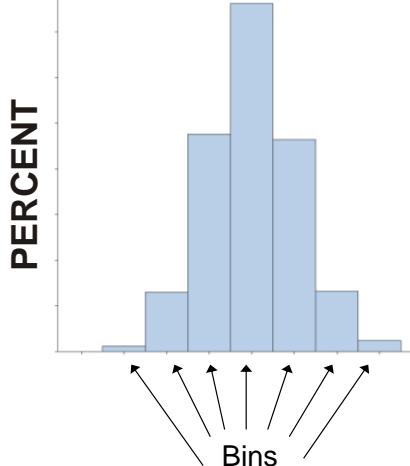
2.2 Picturing Distributions

Objectives

- Look at distributions of continuous variables.
- Describe the normal distribution.
- Use the UNIVARIATE procedure to generate histograms and normal probability plots and to produce descriptive statistics.

20

Picturing Distributions: Histogram



- Each bar in the histogram represents a group of values (a *bin*).
- The height of the bar is the percent of values in the bin- if the bins are equal.
- SAS determines the width and number of bins automatically, or you can specify them.

21

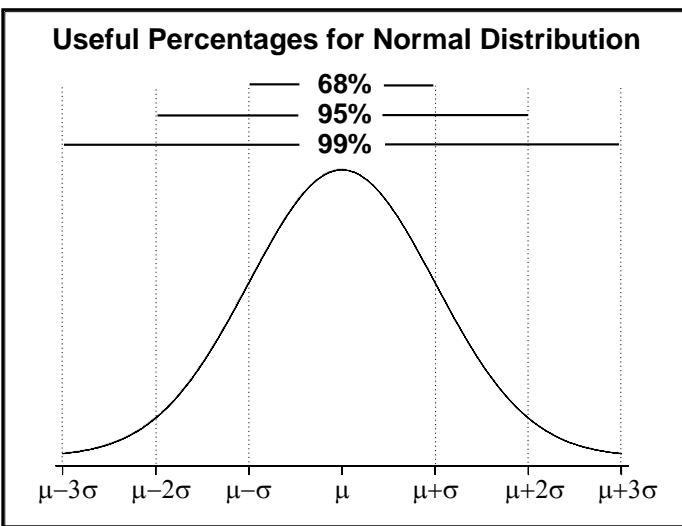
Most parametric (those in which parameters are to be estimated) statistical procedures assume an underlying distribution. It is a good idea to look at your data to see if the distribution of your sample data can reasonably be assumed to come from a population with that distribution. A histogram is a good way to get an idea of what the population distribution is shaped like.

Normal Distributions

A normal distribution:

- is a very useful distribution in statistics
- 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.
- A perfect normal distribution has mean = median = mode.

Normal Distributions, continued



23

Quite often, although not always, a normal distribution is assumed.

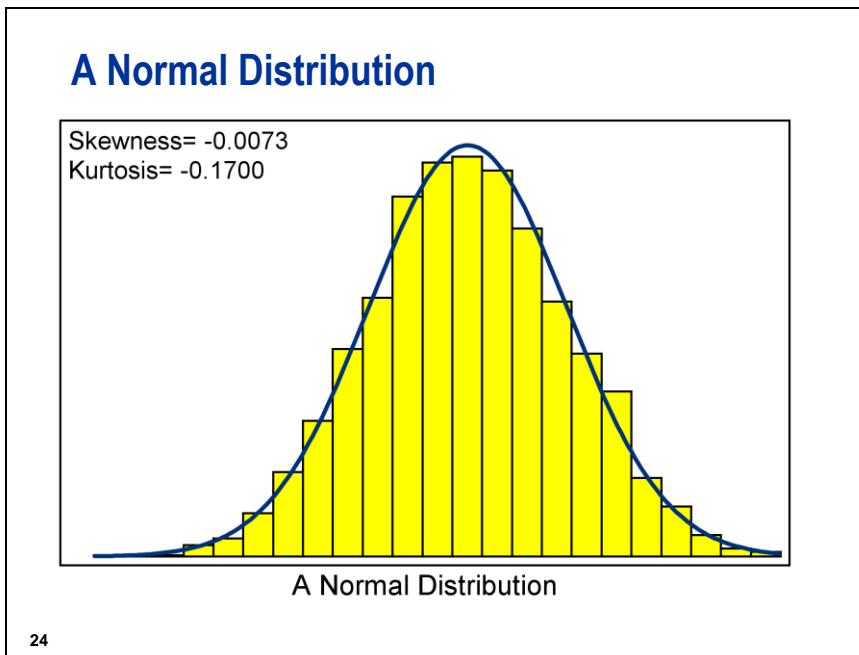
The normal distribution is a mathematical function. The height of the function at any point on the horizontal axis is the “probability density” at that point. Normal distribution probabilities (which can be thought of as the proportion of the area under the curve) tend to be higher near the middle. The center of the distribution is the population mean (μ). The standard deviation (σ) describes how variable the distribution is about μ . A larger standard deviation implies a wider normal distribution. The mean locates the distribution (sets its middle) and the standard deviation scales it.

An observation value is considered unusual if it is far away from the mean. How far is far? You may use the mathematical properties of the normal probability distribution function (PDF) to determine that. If a population follows a normal distribution, then 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
- 99.7% of the data falls within 3 standard deviations of the mean.

Often, values that are more than 2 standard deviations from the mean are regarded as unusual. Now you can see why. Only about 5% of all values are as far away from the mean as that. (Sometimes, only values more than 3 standard deviations away from the mean are closely examined as unusual.)

You will also use this information later when talking about the concepts of confidence intervals and hypothesis tests.



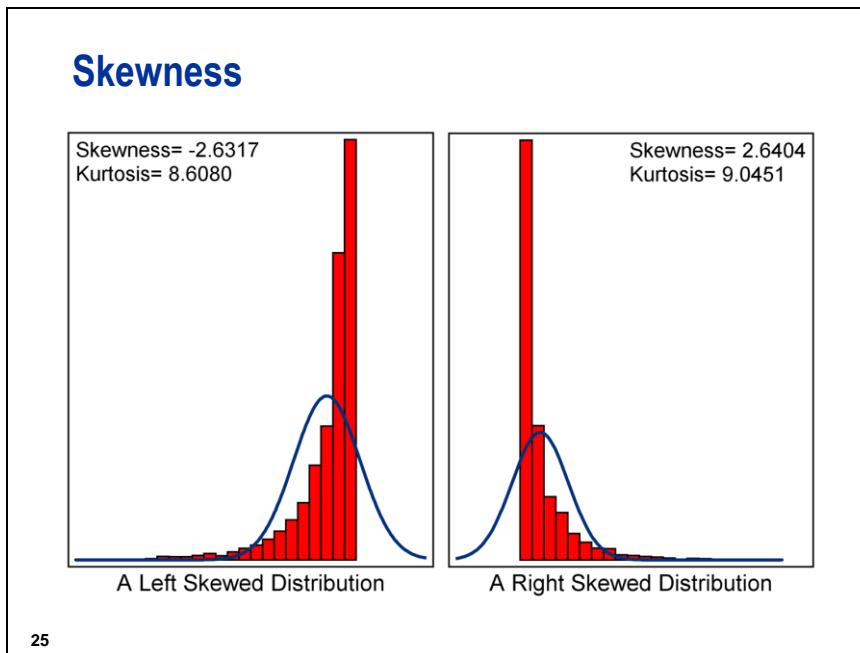
24

The distribution of your data might not look normal. There are infinitely different ways that a population can be distributed. When you look at your own data, you might take note of features of the distribution that indicate similarity or difference from the normal distribution.

In evaluating distributions, it is useful to look at statistical measures of the shape of the sample distribution compared to the normal.

Two such measures are skewness and kurtosis, which are defined over the next few pages.

A histogram of data from a sample drawn from a normal population will generally show values of skewness and kurtosis near 0 in SAS output.

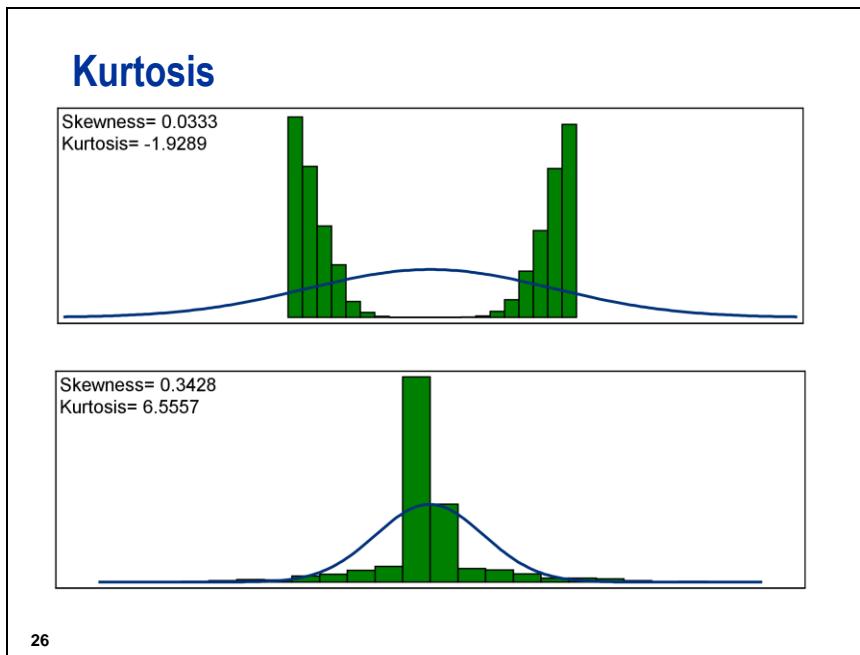


25

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.



26

Kurtosis is often very difficult to assess visually. 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 in SAS.

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*. Rectangular, bimodal, and multimodal distributions tend to have low values of kurtosis.

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.

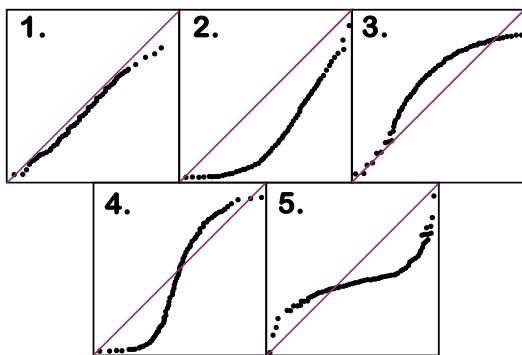
 The normal distribution actually has a kurtosis value of 3, but SAS subtracts a constant of 3 from all reported values of kurtosis, making the constant-modified value for the normal distribution 0 in SAS output. That is the value against which to compare a sample kurtosis value in SAS when assessing normality.

Graphical Displays of Distributions

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

- histograms
- normal probability plots
- box-and-whisker plots

Normal Probability Plots

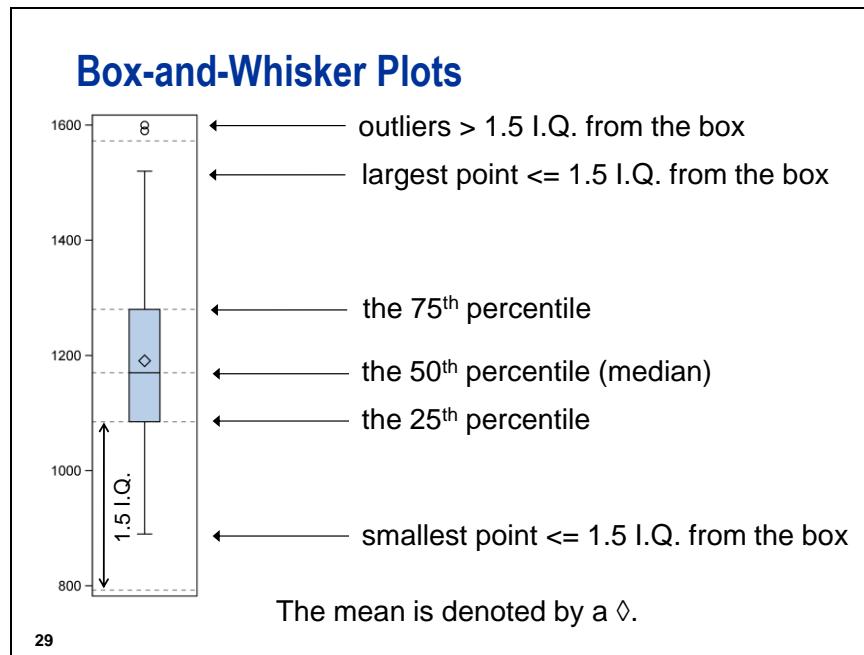


28

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.



Box plots provide information about the variability of data and the extreme data values. The box represents the middle 50% of your data (between the 25th and 75th percentile values), and you get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as assessing the symmetry of the box and whiskers around the median line. The whiskers extend from the box as far as the data extends, to a distance of, at most, 1.5 interquartile units (I.Q.). If any values lie more than 1.5 I.Q. from either end of the box, these values are represented in SAS by square symbols.

The plot above shows that the data is approximately symmetric.

General Syntax of ODS Graphics

```
ODS GRAPHICS ON;  
  statistical procedure code  
ODS GRAPHICS OFF;
```

30

In this course, you will be making use of ODS Statistical Graphics for graphical analysis of data. ODS Statistical Graphics were first made available in SAS 9.2. In order to produce statistical graphics from SAS statistical procedures, the ODS GRAPHICS statement (or ODS GRAPHICS ON) must be submitted. This statement need only be submitted once within an interactive SAS session (or batch job) and will remain in effect until the ODS GRAPHICS OFF statement is submitted. An exception to this requirement is in the case of the new statistical graphics procedures, PROC SGLOT, PROC SGSCATTER and PROC SGRENDER. These procedures will produce graphics whether or not the ODS GRAPHICS statement is first submitted.

The SAS 9.2 documentation lists the graphics available within the description of the SAS procedure.

ODS Graphics Output

- Some graphs are created by default.
- Procedure options (such as PLOTS=) are used to specify which graphs to create.
- You can specify where you want your graphs displayed by using ODS destination statements (for example, LISTING, HTML, RTF).
- ODS SELECT and ODS EXCLUDE statements can be used to select and exclude graphs from your output.

31

ODS templates can be used modify the layout and details of each graph.

Some Recommended ODS Styles

Style	Description
DEFAULT	Color style intended for general-purpose work. This is the default for the HTML destination.
STATISTICAL	Color style recommended for output in Web pages or color print media. This is the style used in the SAS/STAT 9.2 documentation.
ANALYSIS	Color style with a somewhat different appearance from STATISTICAL.
JOURNAL and JOURNAL2	Gray-scale and pure black-and-white styles, respectively. Recommended for graphs in black-and-white publications.
RTF	Used to produce graphs to insert into a Microsoft Word document or a Microsoft PowerPoint slide.

32

ODS styles are used to control the general appearance and consistency of all graphs and tables. You will use a variety of styles and destinations throughout this course.

Statistical Graphics Procedures in SAS

- PROC SGSCATTER creates single-cell and multi-cell scatter plots and scatter plot matrices with optional fits and ellipses.
- PROC SGLOT creates single-cell plots with a variety of plot and chart types.
- PROC SGPANEL creates single-page or multi-page panels of plots and charts conditional on classification variables.
- PROC SGRENDER provides a way to create plots from graph templates that you have modified or written yourself.

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>;
  INSET keywords </ options>;
RUN;
```

34

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.
INSET	places a box or table of summary statistics, called an inset, directly in a graph created with a CDFPLOT, HISTOGRAM, PPLOT, PROBPLOT, or QQPLOT statement. The INSET statement must follow the plot statement that creates the plot that you want to augment.

Selected option for HISTOGRAM and PROBPLOT

NORMAL<(options)> creates a normal probability plot. Options (MU= SIGMA=) determine the mean and standard deviation of the normal distribution used to create reference lines (normal curve overlay in HISTOGRAM and diagonal reference line in PROBPLOT).

 ODS Statistical Graphics are experimental in PROC UNIVARIATE in SAS 9.2.

The SGPlot Procedure

General form of the SGPLOT procedure:

```
PROC SGPLOT <option(s)>;
  DOT category-variable </option(s)>;
  HBAR category-variable </option(s)>;
  HBOX response-variable </option(s)>;
  HISTOGRAM response-variable </option(s)>;
  NEEDLE X= variable Y= numeric-variable </option(s)>;
  REG X= numeric-variable Y= numeric-variable
    </option(s)>;
  SCATTER X= variable Y= variable </option(s)>;
  VBAR category-variable </option(s)>;
  VBOX response-variable </option(s)>;
RUN;
```

35

The SGPlot procedure creates one or more plots and overlays them on a single set of axes. You can use the SGPlot procedure to create statistical graphics such as histograms and regression plots, in addition to simple graphics such as box and whisker plots, scatter plots and line plots.

Selected SGPlot procedure statements:

VBOX creates a vertical box plot that shows the distribution of your data.

REFLINE creates a horizontal or vertical reference line.



Examining Distributions

The following is a summary of what you will accomplish in this demonstration:

- Use the UNIVARIATE Procedure to:
 - Generate Descriptive Statistics
 - Generate a histogram
 - Generate a Probability Plot
- Use the SGLOT Procedure to:
 - Generate a Box and Whisker Plot

```
/*st002d02*/
ods graphics on;
ods select Moments Quantiles ExtremeObs Histogram ProbPlot;
proc univariate data=st092.phone_new;
  var time;
  histogram time / normal(mu=est sigma=est);
  inset skewness kurtosis / position=ne;
  probplot time / normal(mu=est sigma=est);
  inset skewness kurtosis;
  title 'Descriptive Statistics Using PROC UNIVARIATE';
run;
ods graphics off;
```

Selected ODS statements:

ODS LISTING(*action*) opens, manages, or closes the LISTING destination.
*GPATH=**file-specification* <(url='Uniform-Resource-Locator' | NONE)>
 specifies the location for all graphics output that is generated while the destination is open.

 By default, output will go to the Listing destination. Other options are RTF, HTML and PDF destinations, which can also be opened, managed and closed by ODS RTF, ODS HTML and ODS PDF, respectively. If graphical output is requested for either HTML or LISTING destinations, it will be sent to the user's default location. You can select a different location with the GPATH= option.

ODS SELECT specifies output objects for ODS destinations.

Descriptive Statistics Using PROC UNIVARIATE

The UNIVARIATE Procedure
 Variable: time

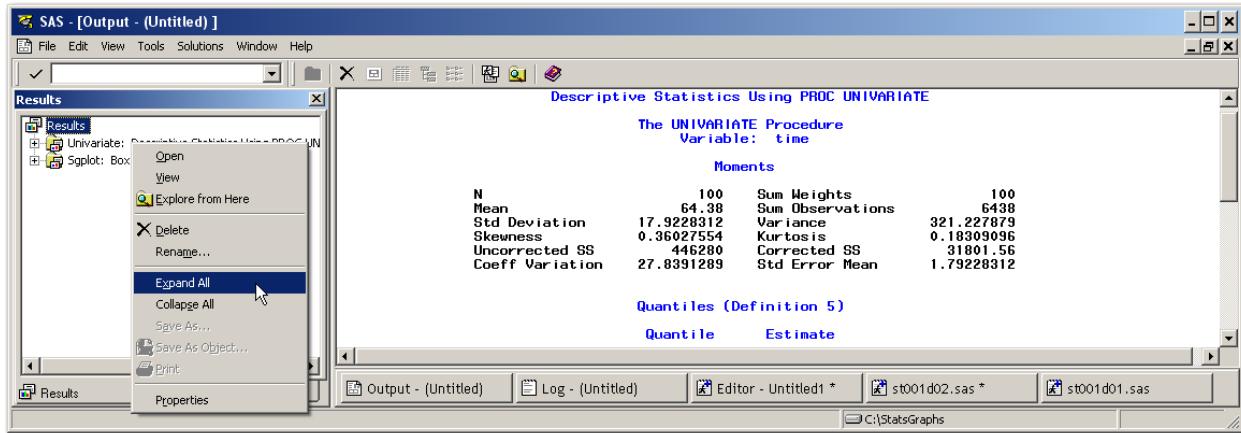
Moments			
N	100	Sum Weights	100
Mean	64.38	Sum Observations	6438
Std Deviation	17.9228312	Variance	321.227879
Skewness	0.36027554	Kurtosis	0.18309096
Uncorrected SS	446280	Corrected SS	31801.56
Coeff Variation	27.8391289	Std Error Mean	1.79228312
Quantiles (Definition 5)			
Quantile	Estimate		
100% Max	114.0		
99%	113.0		
95%	96.0		
90%	87.0		
75% Q3	78.0		
50% Median	62.0		
25% Q1	53.0		
10%	42.5		
5%	35.5		
1%	25.0		
0% Min	19.0		
Extreme Observations			
----Lowest----		----Highest---	
Value	Obs	Value	Obs
19	60	96	67
31	42	100	74
33	82	101	79
34	75	112	71
35	19	114	10

The tabular output indicates that

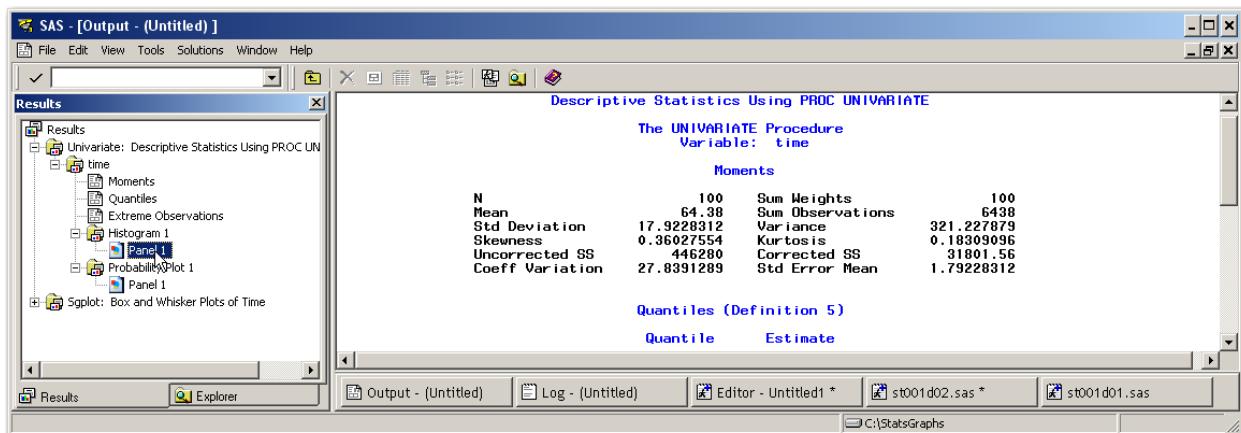
- the mean of the data is 64.38. This is approximately equal to the median (62.00), which indicates the distribution is fairly symmetric.
- the standard deviation is 17.9228312, which means that the average variability around the mean is approximately 18 seconds.
- the distribution is **slightly** skewed to the right.
- the distribution has **slightly** heavier tails than the normal distribution.
- the operator with the lowest score is observation 60, with a score of 19 seconds. The operator with the highest score is observation 10, with a score of 114 seconds

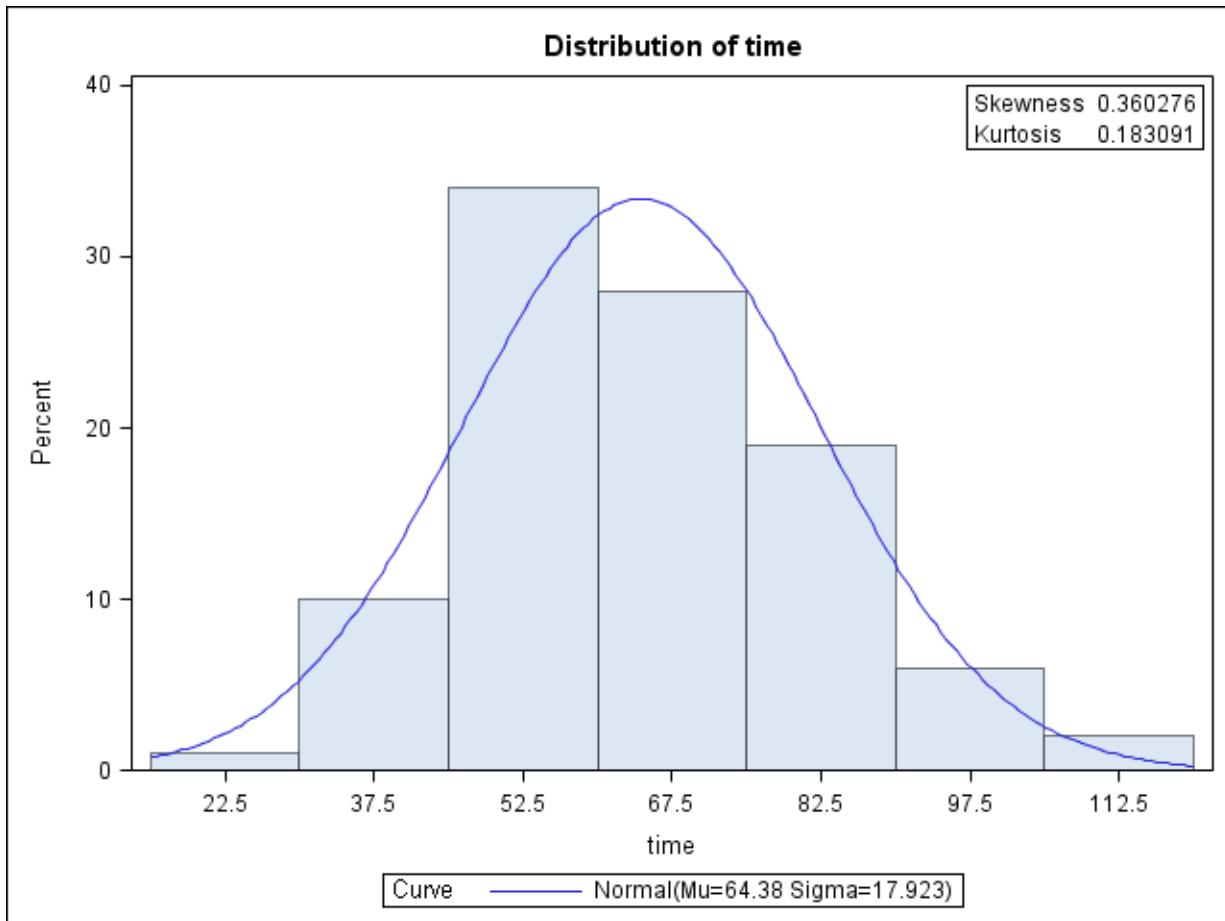
In order to view the graphical output follow these steps:

1. Expand the output from the Results window by right-clicking on the name of the procedure in the Results window and selecting Expand All in the drop-down menu.



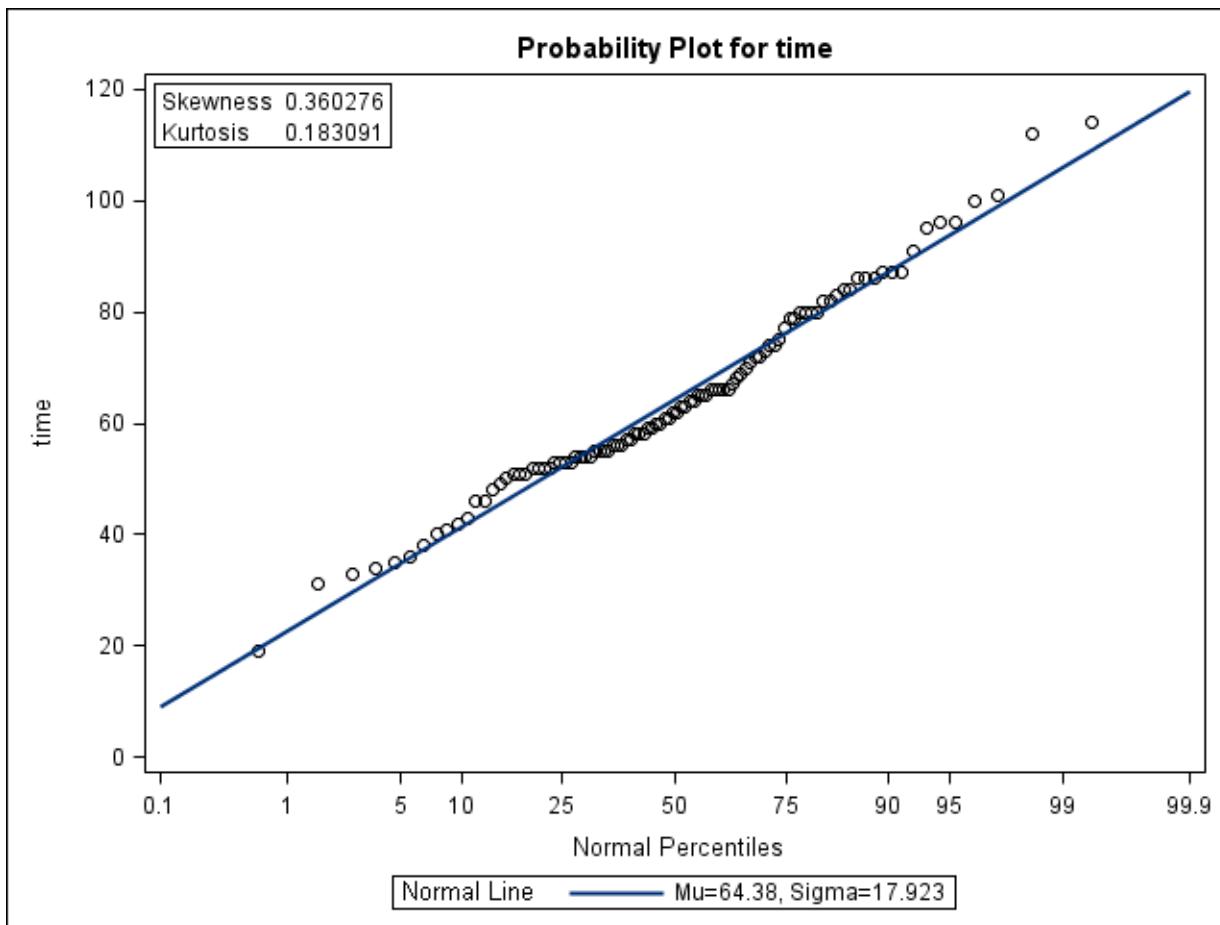
2. Double Click on the image icon to open the image in the user's default graphics software window.





The bin identified with the midpoint of 37.5 has approximately 10% of the values. The skewness and kurtosis values are reported in the inset.

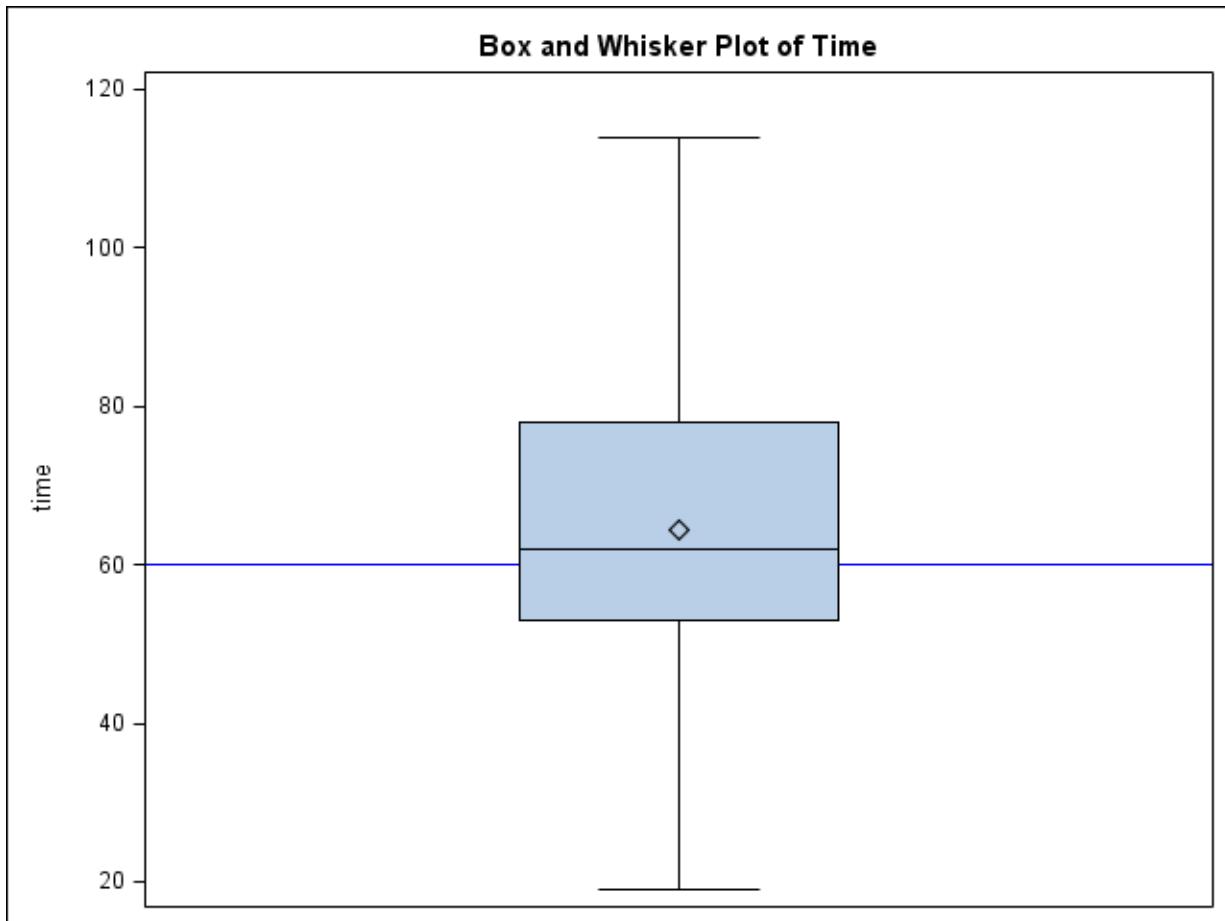
The mean and the standard deviation are also reported in the legend.



The normal probability plot above, shows a 45-degree line. This represents where the data values would fall if they came from a normal distribution. The circles represent the observed data values. Because the circles follow the 45-degree line in the graph, you can conclude that there does not appear to be any severe departure from the normality.

```
/*st002d02.sas*/
ods graphics on;
proc sgplot data=st092.phone_new;
    refline 60 / axis=y lineattrs=(color=blue);
    vbox time ;
    title "Box and Whisker Plot of Time";
run;
ods graphics off;
```

A reference line is requested at 60 on axis y. Because this is a vertical box plot, the y-axis is the Time variable.



2.3 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.

38

Parameters and Statistics

Summary statistics are used to approximate population parameters.

	Population Parameters	Sample Statistics
Mean	μ	\bar{x}
Variance	σ^2	s^2
Standard Deviation	σ	s

39

Point Estimates

\bar{x} estimates μ

s estimates σ

40

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

- An estimate of the average **time** is 67.38, and an estimate of the standard deviation is 17.92283
- Because you only have an estimate of the unknown population mean, you need to know the variability of your estimate.

How good is the sample mean?

How close the sample mean is to the population mean will depend on:

- Variability of data
- Size of the sample
- Is the sample representative?

41

Variability among samples



→ 64.38



→ 66.10

▪
▪
▪

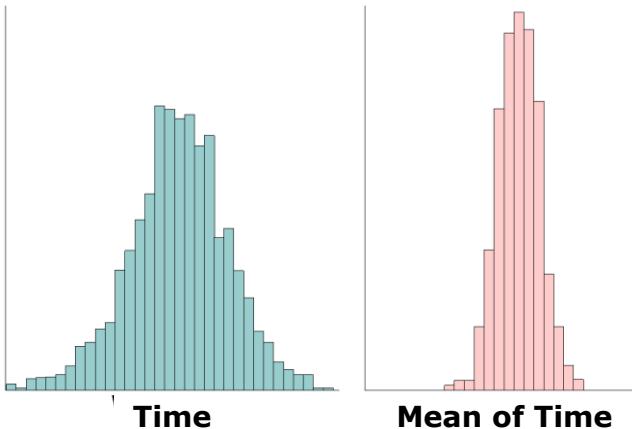
▪
▪
▪

42

The sample mean is only an estimate of the population mean. If you collected another sample of calls, you would likely obtain another estimate of the mean.

Different samples yield different estimates of the mean for the same population. How close on average these sample means are to one another is the variability of the estimate of the population mean.

Distribution of Sample Means



43

What is a distribution of sample means? It is just that. It is a distribution of many mean values, each of a common sample size.

Suppose 1000 random samples, all with the same sample size of 10, are taken from an identified population.

- The left histogram shows the distribution of all 5000 **observations**.
- The right histogram, however, represents the distribution of the 1000 **sample means**.

The variability of the distribution of sample means is smaller than the variability of the distribution of the 5000 observations.

 The samples in the 1000 are assumed to be taken with replacement, meaning that after 10 student values are taken, all ten of those students can be chosen again in subsequent samples.

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.

$$- \text{ Standard error of the mean} = \frac{s}{\sqrt{n}} = S_{\bar{x}}$$

44

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.

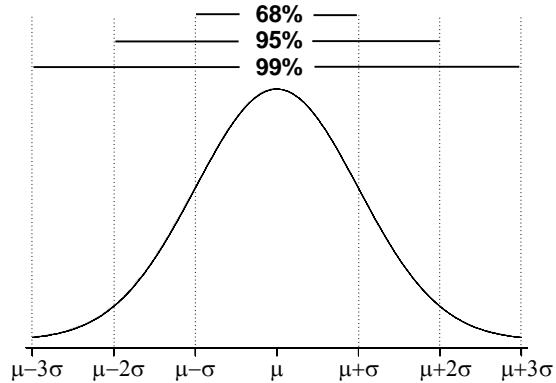
The standard error of the mean for the variable **Time** is $17.923 / \sqrt{100}$, or approximately 1.79. This is a measure of how much variability of sample means there is around the population mean. The smaller the standard error, the more precise your sample estimate is.



You can improve the precision of an estimate by increasing the sample size.

Normal Distribution for the Mean

Useful Distribution Revisited



45

For purposes of finding confidence limits for parameters (such as a mean), you might make assumptions about a theoretical population distribution. You might, for instance, assume normality of sample means. The σ above refers to the standard error of the mean.

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.

46

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 customer contact centre example, you interpret a 95% confidence interval by stating that you are 95% confident that the interval contains the mean time it takes the phone to be answered 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.

Details

In any normal distribution of sample means with parameters μ and σ , over samples of size n, the probability is 0.95 for

$$-1.96\sigma_{\bar{x}} \leq \bar{x} - \mu \leq 1.96\sigma_{\bar{x}}$$

This is the basis of confidence intervals for the mean. If you rearrange the terms above and replace the known $\sigma_{\bar{x}}$ with the estimated standard error, $s_{\bar{x}}$, the probability is 0.95 for

$$\bar{x} - 1.96s_{\bar{x}} \leq \mu \leq \bar{x} + 1.96s_{\bar{x}}$$

95% Confidence Interval

If the data follow a normal distribution, approximately 95% of observations lie within two standard deviations of the mean.

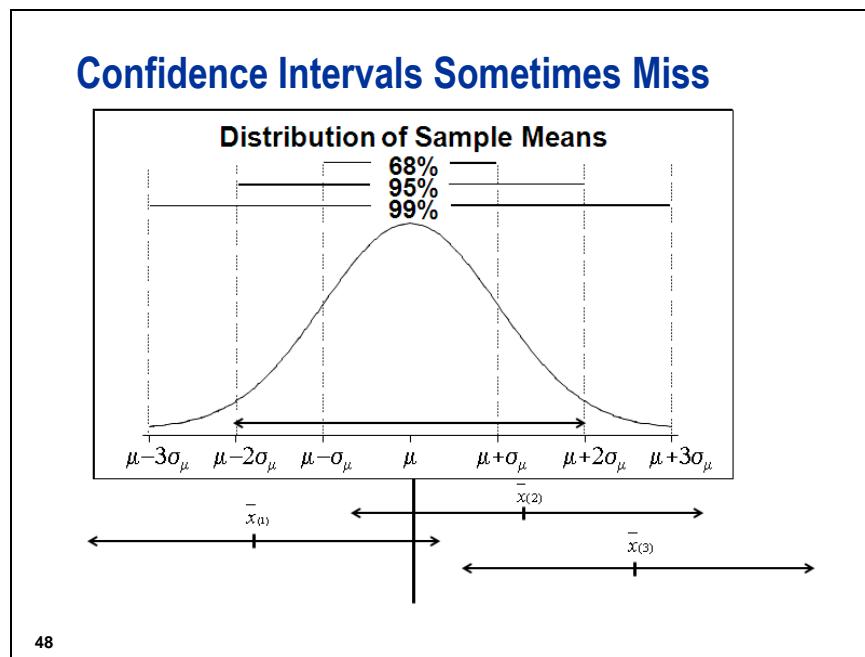
Using a similar idea, if the sample means follow a normal distribution then 95% of sample means will lie within two standard errors of the population mean

Therefore:

$$95\% \text{ CI} = \bar{x} \pm 2 \cdot s_{\bar{x}} \quad \text{or} \quad (\bar{x} - 2 \cdot s_{\bar{x}}, \bar{x} + 2 \cdot s_{\bar{x}})$$

47

Student's t distribution arises when you are making inferences about a population mean and (as in nearly all practical statistical work) the population standard deviation (and therefore, standard error) is unknown and has to be estimated from the data. It is approximately normal as the sample size grows larger. The t in the equation above refers to the number of standard deviation (or standard error) units away from the mean required to get a desired confidence in a confidence interval. That value will vary not only with the confidence that you choose, but also with the sample size. For 95% confidence, that t value will usually be approximately 2, because, as you have seen, 2 standard errors below to 2 standard errors above a mean will give you about 95% of the area under a normal distribution curve.



48

The graph above is the distribution of sample means. You typically take only one sample from that distribution, but in this picture you see that 3 researchers have each taken a sample from the same population. Each sample had a different mean. The standard errors were all about the same and about the same as the population standard error.

The double-headed arrows around each of the means (for researcher 1, 2, and 3) measure about 2 standard errors to each side of the each sample mean (t is about 2 for these researchers). The sample means for researcher 1 and 2 fell within 2 standard errors away from the (unknown) population mean, just by good luck. Actually, 95% of all researchers should have equivalent “luck”. Researcher number 3 was in the unlucky 5%. He did his work just as well and hard and blissfully reported his sample mean and confidence interval, but because his sample mean was more than 2 standard errors from the population mean, his confidence interval did not extend far enough to include the true mean.

If the confidence interval is faithfully calculated using the formula shown earlier and assumptions are met, 95% of the time they will include the true mean. Unfortunately, there is no way to know if yours happens to be in the 95% or the 5% group.

 The actual observed value of t (the number of standard errors your observed mean is away from a hypothesized mean) is related to a specific probability, known in statistics as a p -value, which will be described in the next section.

Normality and the Central Limit Theorem

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

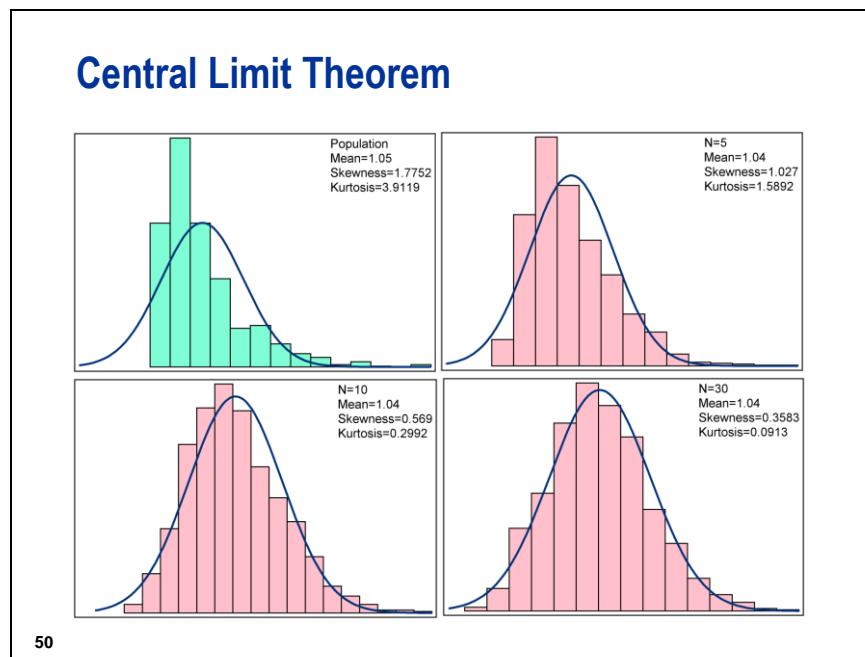
To satisfy the assumption of normality, you can either

- verify that the population distribution is approximately normal, or
- apply the **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.)

49

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 customer contact centre is 100, you can apply the central limit theorem and satisfy the assumption of normality for the confidence intervals



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. The number 30 is not a magic number, but a common rule of thumb.



Confidence Intervals

The following is a summary of what you will accomplish in this demonstration:

- Use the MEANS Procedure to generate a 95% confidence Interval for the mean of TIME.

```
/*st002d03.sas*/
proc means data=st092.phone_new maxdec=4
            n mean stderr clm;
var Time;
title '95% Confidence Interval for Time';
run;
```

Selected PROC MEANS statement options:

CLM calculates confidence limits for the mean.

STDERR calculates the standard error of the mean.

The output is shown below.

95% Confidence Interval for Time				
The MEANS Procedure				
Analysis Variable : time				
N	Mean	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
100	64.3800	1.7923	60.8237	67.9363

In the call centre example, you are 95% confident that the population mean is contained in the interval 60.82 and 67.94.

- How do you increase the precision 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 precise.
- You can use the ALPHA= option in the PROC MEANS statement to construct confidence intervals with a different confidence level. Choose (1.00-Confidence/100) as your ALPHA level. By default, ALPHA=0.05 (1.00 – 95/100).

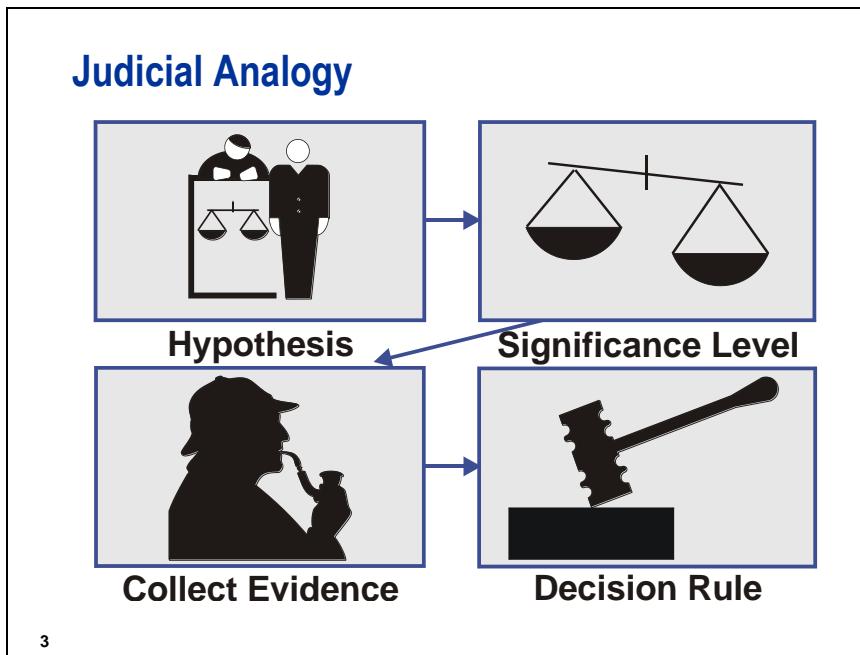
Chapter 3 Testing Business Questions

3.1 Hypothesis Testing	3-3
Demonstration: One-Sample <i>t</i> -Test.....	3-19

3.1 Hypothesis Testing

Objectives

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



3

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”.

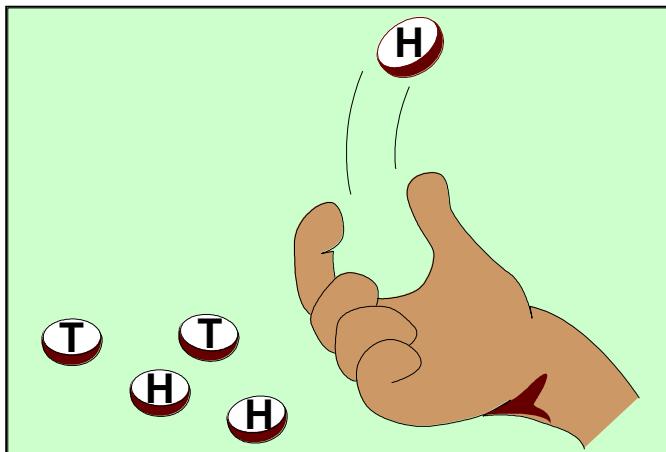
Collect evidence.

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.

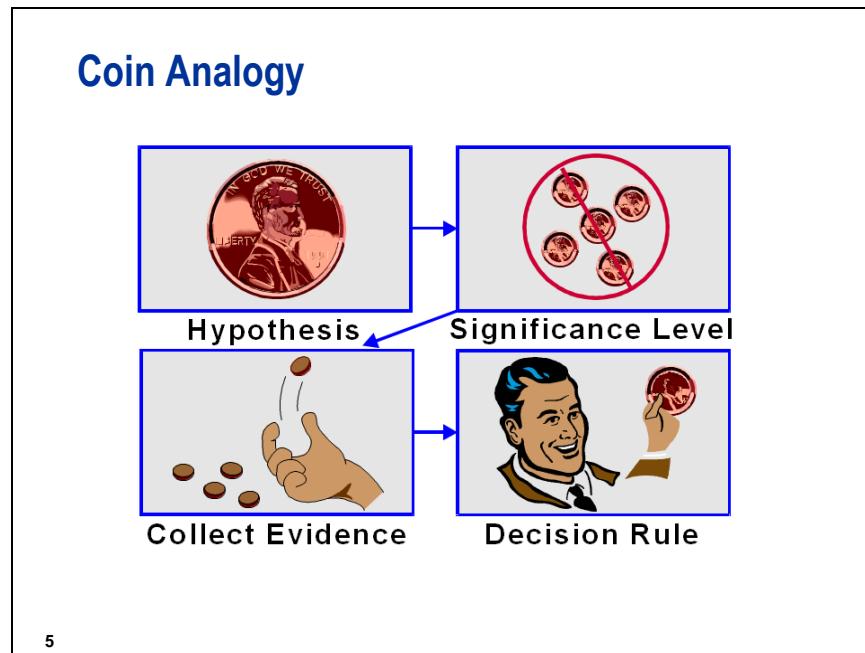
Statistical hypothesis testing follows this same basic path.

Coin Example



4

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.



5

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.

The Decision Rule

In general, you compare α and the p -Value in order to:

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

6

It is important to clarify that

- α , 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 α and the associated p -value to make a decision.

Remember, α is set ahead of time based on the circumstances of the experiment. The level of α 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 customer contact centre example, α was set to 0.05, based on the consequences of making a Type I error (the error of concluding that the mean time it takes to answer the phone is not 60 when it really is 60 seconds). If making a Type I error is especially egregious, you might consider lowering your significance level.

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

7

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

The probability of a Type I error, often denoted α , 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.

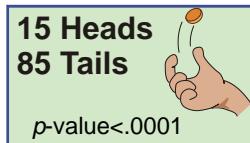
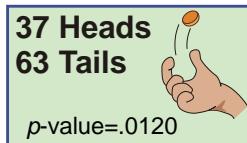
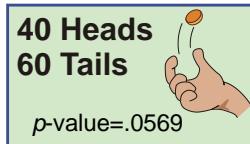
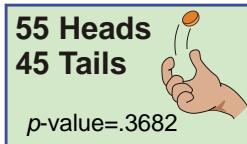
The probability of a Type II error, often denoted β , is the probability that you fail to reject the null hypothesis when it is false. In the

- 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 β is the Type II error rate. This is the probability that you correctly reject the null hypothesis.

Coin Experiment – Effect Size Influence

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



8

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.

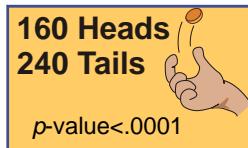
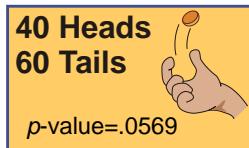
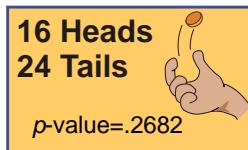
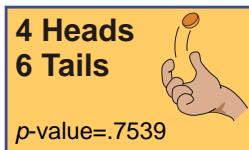
In this situation, the greater the difference between the number of heads and tails, the more evidence you have that the coin is not fair.

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.

Coin Experiment – Sample Size Influence

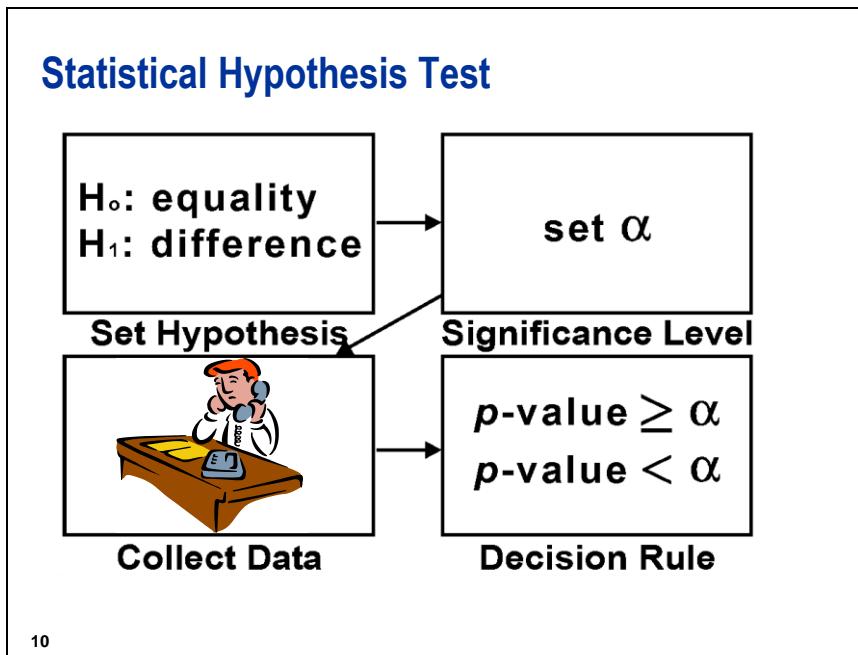
Flip a coin and get 40% heads and decide if it is fair.



9

A p -value is not only affected by the magnitude of the test statistic (in this case, the test statistic is the observed proportion of heads in k flips). It is also affected by the sample size (number of coin flips, k).

For a fair coin, you would expect 50% of k flips to turn up heads. In this example, in each case, the observed proportion of heads from k flips was 0.4. This value is different from the 0.5 you would expect under H_0 . The evidence is stronger, the greater the number of trials (k) the proportion is based on. As you saw in the section on confidence intervals, the variability around a mean estimate is smaller, the larger the sample size. For larger sample sizes, you can measure means more precisely. Therefore, 40% heads out of 400 flips would make you surer that this was not just a chance difference from 50% than would 40% out of 10 flips. The smaller p -values reflect this confidence. The p -value here is assessing the probability that this difference from 50% occurred purely by chance.



10

In statistics,

1. the null hypothesis, denoted H_0 , is your initial assumption and is usually one of equality or no relationship. For the customer contact centre example, H_0 is that the average time it takes for a phone call to be answered is 60 seconds.
2. the significance level is usually denoted by α , 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 α
 - reject the null hypothesis if the p -value is less than α .

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

Is the Mean Different to 60 Seconds?

- We want to compare the sample mean to the hypothesised value (60 seconds)
- Is the difference real or just due to chance?
- What statistics do we need to calculate?

11

Performing a Test of Hypothesis

To test the hypothesis that the population mean equals an hypothesised value we calculate the t statistic

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

Where μ_0 = hypothesised value.

12

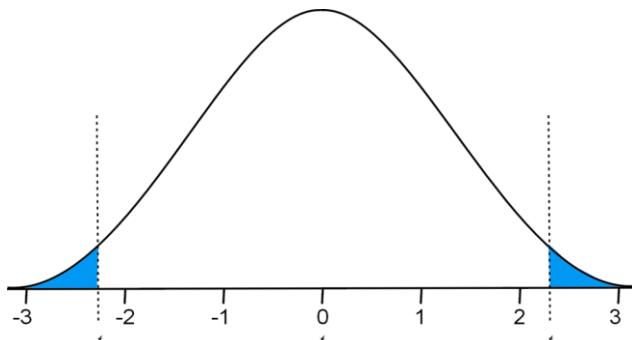
For the customer contact centre example, μ_0 is the hypothesized value of 60, \bar{x} is the sample mean time taken to answer the phone, and $s_{\bar{x}}$ is the standard error of the mean.

- 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 t-test

If null hypothesis is true (mean=60) then t statistic will follow a t-distribution



The t statistic can be positive or negative.

13

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

The area in each of the tails corresponds to $\alpha/2$ or 2.5%. The sum of the areas under the tails is 5%, which is alpha.



The alpha and t -distribution mentioned here are the same as those in the section on confidence intervals. In fact, there is a direct relationship. The rejection region based on α begins at the point where the $(1.00-\alpha)$ confidence interval will no longer include the true value of μ_0 .

To Obtain p-value

- Chance of obtaining this t-value (or more extreme) assuming the null hypothesis is true, i.e. mean is 60 seconds
 - OR
- Chance of getting this difference due to chance

14

Assumptions

- Sample mean is normally distributed
 - Data are normal, or
 - Central Limit Theorem if $n > 30$
- Independent observations

15

In our example the sample size is 100 so we can apply the central limit theorem.

As we collected a random sample and our data is representative of the population, the independent observations assumption is validated.

Two-Sided Test of Hypothesis

The test of hypothesis is two-sided if the null is rejected when the actual value of interest is either less than or greater than the hypothesised value.

$$H_0: \mu = 60$$

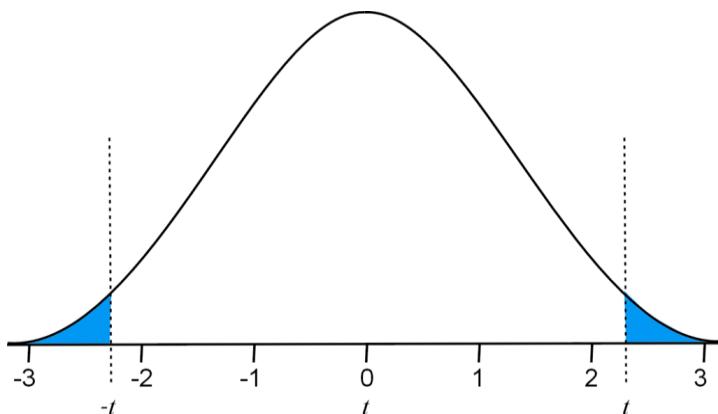
$$H_1: \mu \neq 60$$

16

For the phone example, if discrepancies in either direction (above 60 seconds or below 60 seconds) are expected we would conduct a two-sided test of hypothesis.

 SAS, by default, reports p-values for two sided tests.

Two-Sided Test of Hypothesis



17

For a two-sided test of hypothesis, the rejection region is contained in both tails of the distribution. If the t statistic falls in the rejection region 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%

One-Sided Test of Hypothesis

In some situations, you only expect a difference in one direction. Perhaps you only want evidence that the mean is significantly higher than 60.

For example, instead of testing

$$H_0: \mu = 60 \text{ versus } H_1: \mu \neq 60$$

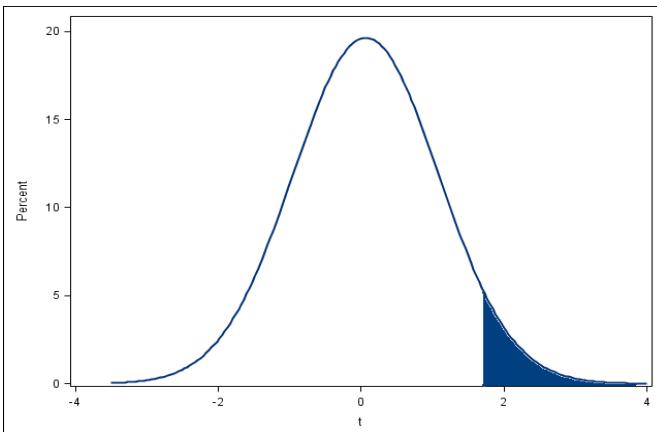
you test

$$H_0: \mu \leq 60 \text{ versus } H_1: \mu > 60$$

18

If we expect that the mean waiting time is greater than 60 seconds, a one-sided test may be more appropriate.

One-sided Test of Hypothesis



19

For a one-sided test of hypothesis, the entire rejection region falls in one tail of the distribution. Therefore, if the t statistic falls in the rejection region, you reject the null hypothesis. Otherwise you fail to reject the null hypothesis.

The area in the tail corresponds to α or 5%.

 The decision to conduct a one-sided or two-sided test of hypothesis should be made prior to examining the data. SAS, by default, provides p -values from two-sided tests.

To obtain the p -value from a one-sided test you can use the SIDES= option on the PROC TTest statement.

- Tests and Confidence Intervals produced in PROC TTEST using:
 - SIDES=U for Upper Tail Tests ($\mu_0 \leq k$) and CIs
 - SIDES=L for Lower Tail Tests ($\mu_0 \geq k$) and CIs

See the SAS/STAT Procedure guide for more information.

The TTEST Procedure

General form of the TTEST procedure:

```
PROC TTEST DATA=SAS-data-set;
  CLASS variable;
  VAR variables;
RUN;
```

20

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.



One-Sample t-Test

The following is a summary of what you will accomplish in this demonstration:

- Verify the assumption of a One-Sample T-test.
- Use the TTEST procedure to test the hypothesis that the mean time for the new team in the customer contact centre is equal to 60 seconds.

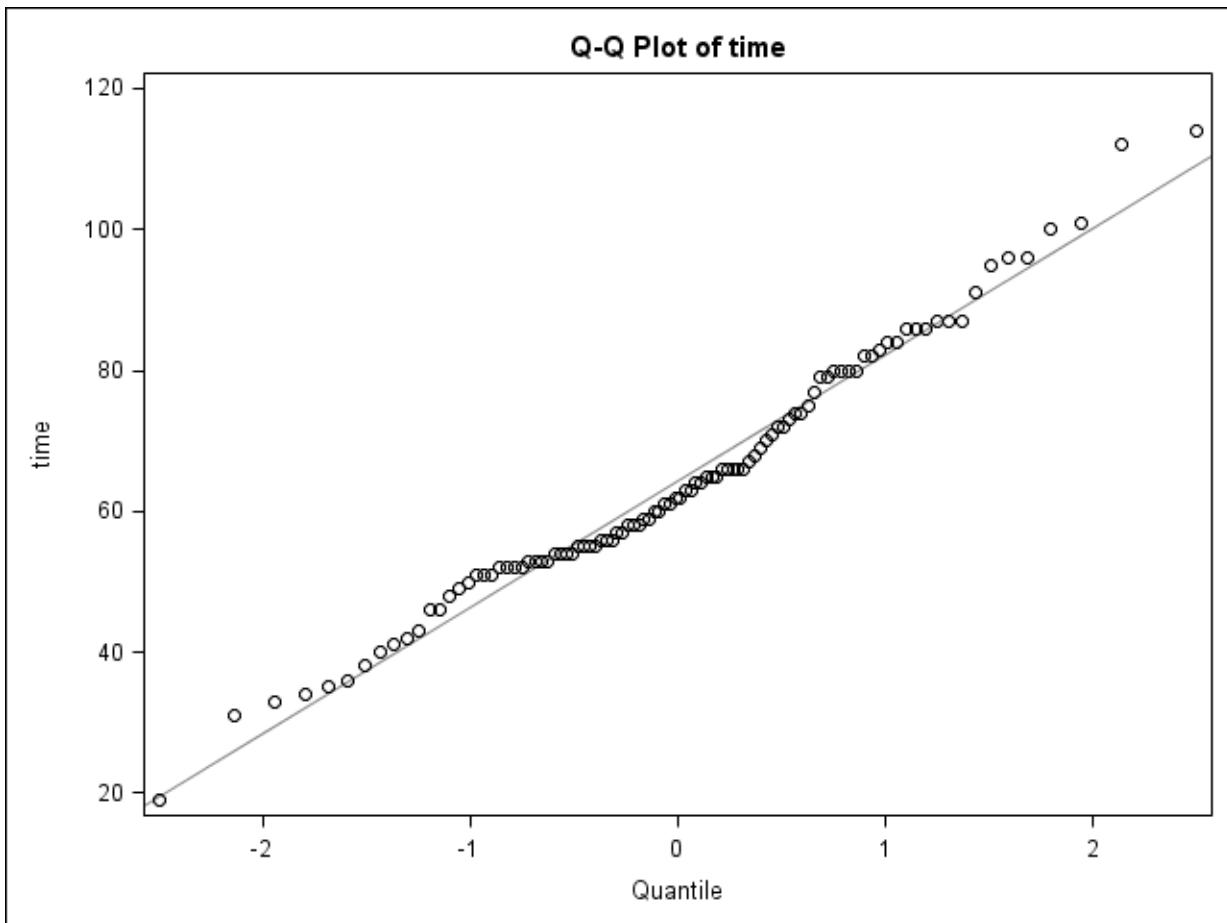
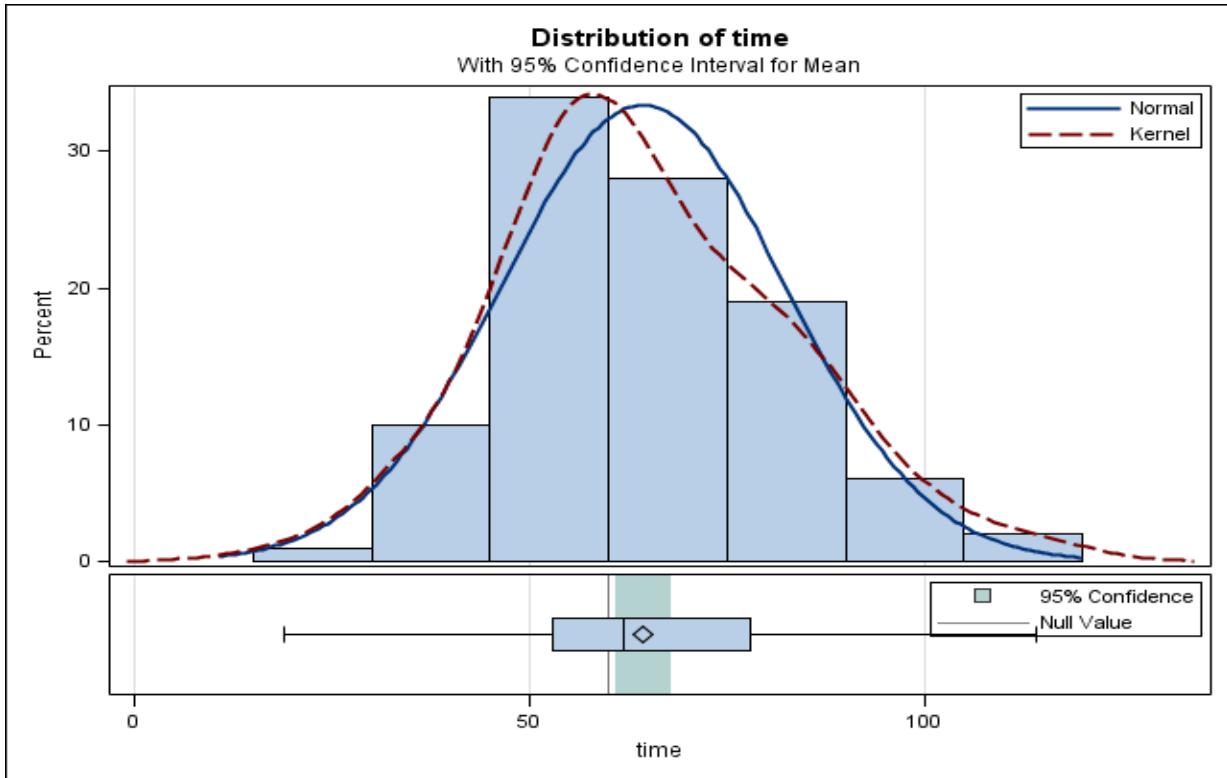
```
/*st003d01*/
ods graphics on;
proc ttest data=st092.phone_new h0=60 plots(shownull)=interval;
  var time;
title 'One-Sample t-test - testing the mean time is 60 seconds';
run;

ods graphics off;
```

First it is advisable to verify the assumptions of the one-sample t-test. The normality assumption can be verified by invoking the central limit theorem (the number of observations is 100) or by viewing the Summary Panel and the Q-Q plot.

Selected PLOTS()= options in the PROC TTEST statement:

PLOTS(SHOWNULL)=INTERVAL includes a plot of confidence intervals of the difference between groups. SHOWNULL places a vertical reference line at the mean value of the null hypothesis ($H_0=0$ by default).



The Q-Q Plot (Quantile-Quantile Plot) is similar to the Normal Probability plot you saw earlier. The x-axis for this plot is just scaled as quantiles, rather than probabilities. It seems that the data approximate a normal distribution. There does seem to be an outlier- a phone call answered in 19 seconds.

 The statistical tables for the TTEST procedure are displayed below.

One-Sample t-test \bar{x} testing the mean time is 60 seconds The TTEST Procedure						
	N	Mean	Std Dev	Std Err	Minimum	Maximum
①	100	64.3800	17.9228	1.7923	19.0000	114.0
		Mean	95% CL Mean	Std Dev	95% CL Std Dev	
②		64.3800	60.8237 67.9363	17.9228	15.7364 20.8205	
			DF	t Value	Pr > t	
③			99	2.44	0.0163	

- ① In the Statistics table, examine the descriptive statistics.
- ② The confidence limits for the sample mean and sample standard deviation are also shown.
- ③ The t-statistic is 2.44 and the p-value is 0.0163.

When starting out, it is advisable to complete the 4 steps of Hypothesis testing.

Step 1- Set Hypothesis

$$H_0: \mu = 60$$

$$H_1: \mu \neq 60.$$

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$p\text{-value}=0.0163.$$

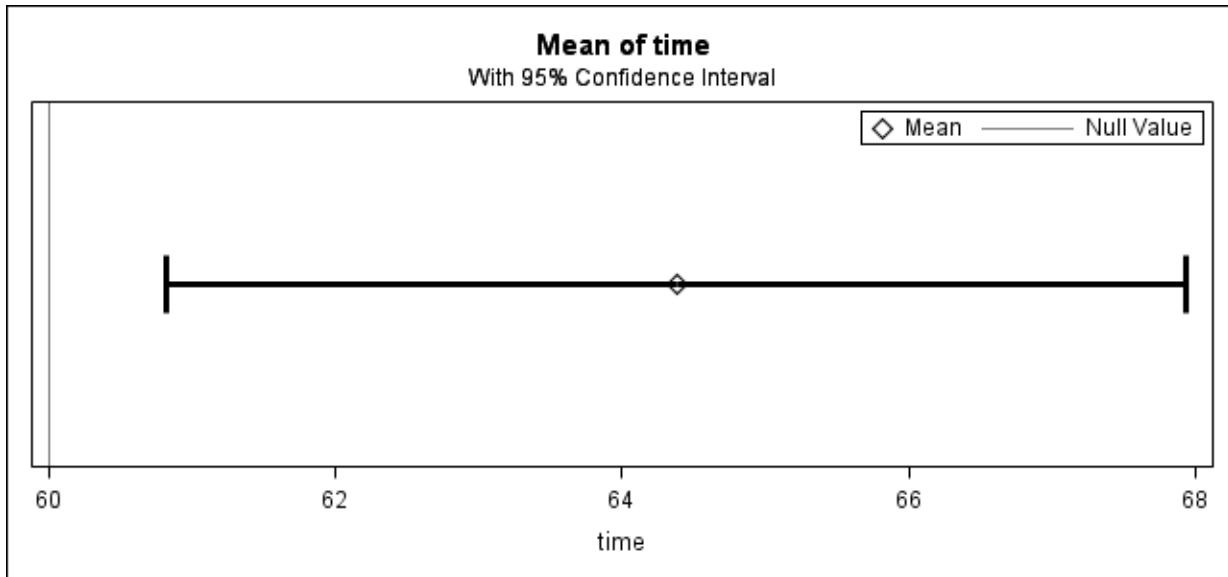
Step 4- Decision Rule.

The $p\text{-value} < \alpha$, Reject H_0 , therefore, the mean is not equal to 60 seconds.

Return your attention to the Statistics table. The mean is 64.38; you can conclude that the mean is significantly bigger than 60 seconds.

Also, the confidence interval for the mean is (60.8237 67.9363), this value does not include 60, therefore we are 95% confident that 60 is not a valid number for the population mean.

This can also be seen in the Interval Plot.



Chapter 4 Testing Two or More Groups

4.1 Comparing Two Groups.....	4-3
Demonstration: Two-Sample <i>t</i> -Test.....	4-11
4.2 One-Way ANOVA.....	4-16
Demonstration: The GLM Procedure- comparing two groups	4-25
Demonstration: The GLM Procedure- more than two groups.....	4-32
Demonstration: Post Hoc Pairwise Comparison	4-45

4.1 Comparing Two Groups

Objectives

- Analyse differences between two population means using the TTEST procedure.
- Recognise and verify the assumptions of a two-sample *t*-test.

Phone Example

- New team



- Existing team (A)



4

We want to compare the service provided by our newly formed team to that from an existing team, labeled A. A random sample of the time it took the two teams to answer the phone was taken, and the results are stored in SAS data set **st092.phone_new_and_a**.

The variables in the data set are:

team: two teams, New team and Existing team A

time: time taken in seconds to answer the phone

Performing a Two-Sample t-test

To test the hypothesis that the mean time for 'New' equals the mean time of 'Existing'

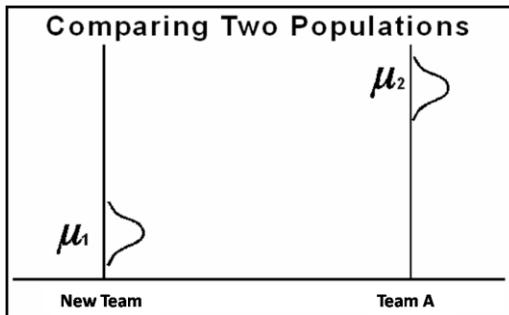
$$t = \frac{(\text{mean}_{\text{New}} - \text{mean}_{\text{Existing}})}{\text{combined standard error}}$$

5

We need to calculate the difference between the mean time for the New team and the mean time for the Existing team A.

To determine if the difference we obtain indicates a real difference in time between the two teams we need to include a measure of the variability of the means.

Assumptions



- independent observations
- normally distributed data for each group
- equal variances for each group

6

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.

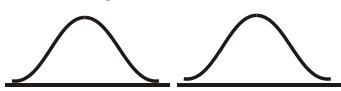
The assumption of normality can be relaxed if the data is approximately normally distributed or if enough data is collected. This assumption can be verified by examining plots of the data.

There are several tests for equal variances. If this assumption is not valid, an approximate *t*-test can be performed.

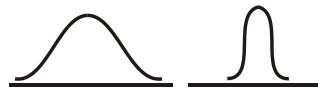
If these assumptions are **not** valid and no adjustments are made, the probability of drawing incorrect conclusions from the analysis could increase.

F-Test for Equality of Variances

$$H_0: \sigma_1^2 = \sigma_2^2$$



$$H_1: \sigma_1^2 \neq \sigma_2^2$$



$$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$$

7

To evaluate the assumption of equal variances in each group you can use graphics or the *F*-test for equality of variances. The null hypothesis for this test is that the variances are equal. 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.

If your data are not normally distributed, you can look at plots to help determine whether the variances are approximately equal.

Equal Variance t-Test and p-Values

t-Tests for Equal Means: $H_0: \mu_1 - \mu_2 = 0$

Equal Variance t-Test (Pooled):

T = 7.4017 DF = 6.0 Prob > |T| = 0.0003 

Unequal Variance t-Test (Satterthwaite):

T = 7.4017 DF = 5.8 Prob > |T| = 0.0004

F-Test for Equal Variances: $H_0: \sigma^2_1 = \sigma^2_2$

Equality of Variances Test (Folded F):

F' = 1.51 DF = (3,3) Prob > F' = 0.7446 

8

- ① 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. Therefore, ② 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.

Unequal Variance *t*-Test and *p*-Values

***t*-Tests for Equal Means:** $H_0: \mu_1 - \mu_2 = 0$

Equal Variance *t*-Test (Pooled):

$T = -1.7835$ DF = 13.0 Prob > | T | = 0.0979

Unequal Variance *t*-Test (Satterthwaite):

$T = -2.4518$ DF = 11.1 Prob > | T | = 0.0320 ②

***F*-Test for Equal Variances:** $H_0: \sigma_1^2 = \sigma_2^2$

Equality of Variances Test (Folded *F*):

$F' = 15.28$ DF = (9,4) Prob > F' = 0.0185 ①

9

- ① 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. Therefore, ② 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.

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.

The TTEST Procedure

General form of the TTEST procedure:

```
PROC TTEST DATA=SAS-data-set;
  CLASS variable;
  VAR variables;
  RUN;
```

10

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.



Two-Sample *t*-Test

The following is a summary of what you will accomplish in this demonstration:

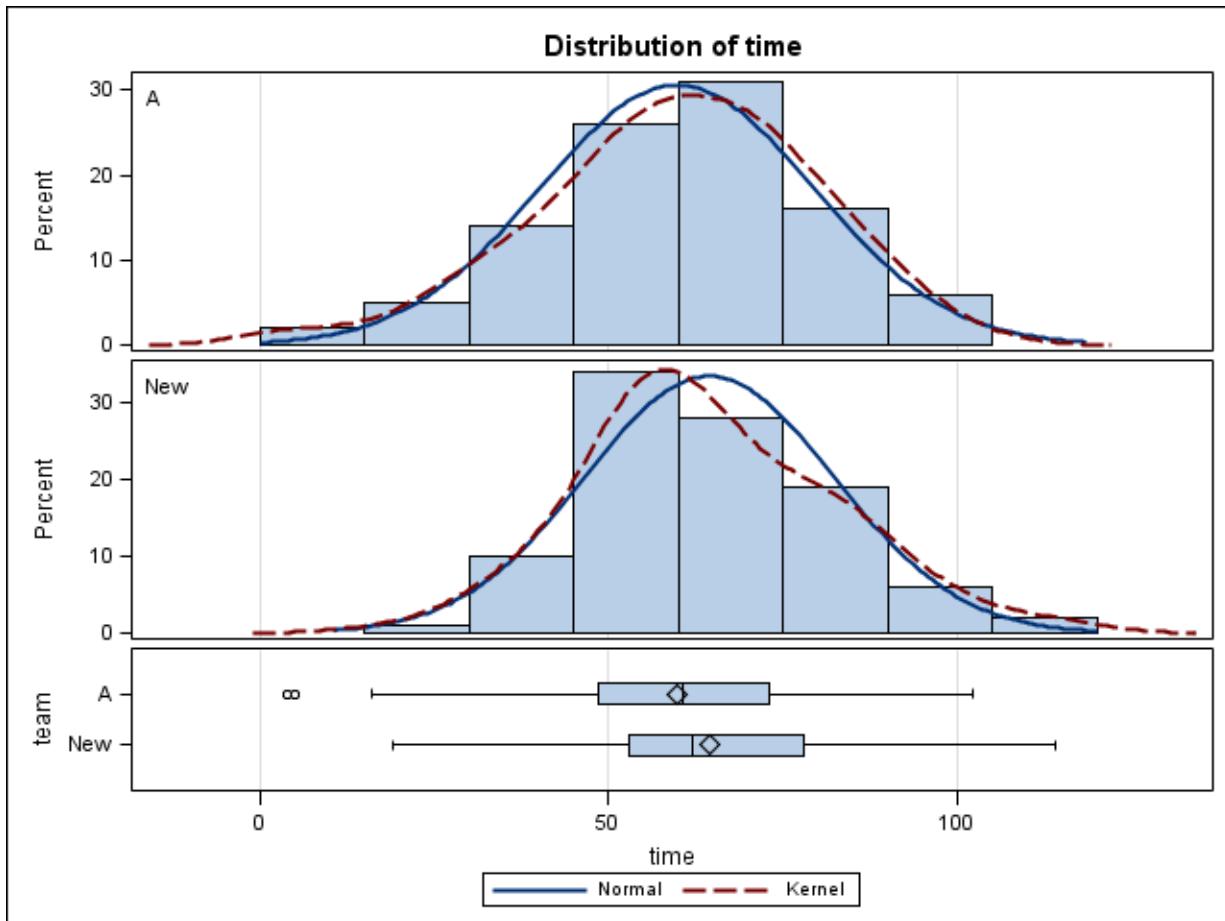
- Verify the assumptions of a Two-Sample T-test
- Use the TTEST procedure to test the hypothesis that the average time taken to answer the phone for the new team is equal to the average time taken to answer the phone for the A team.

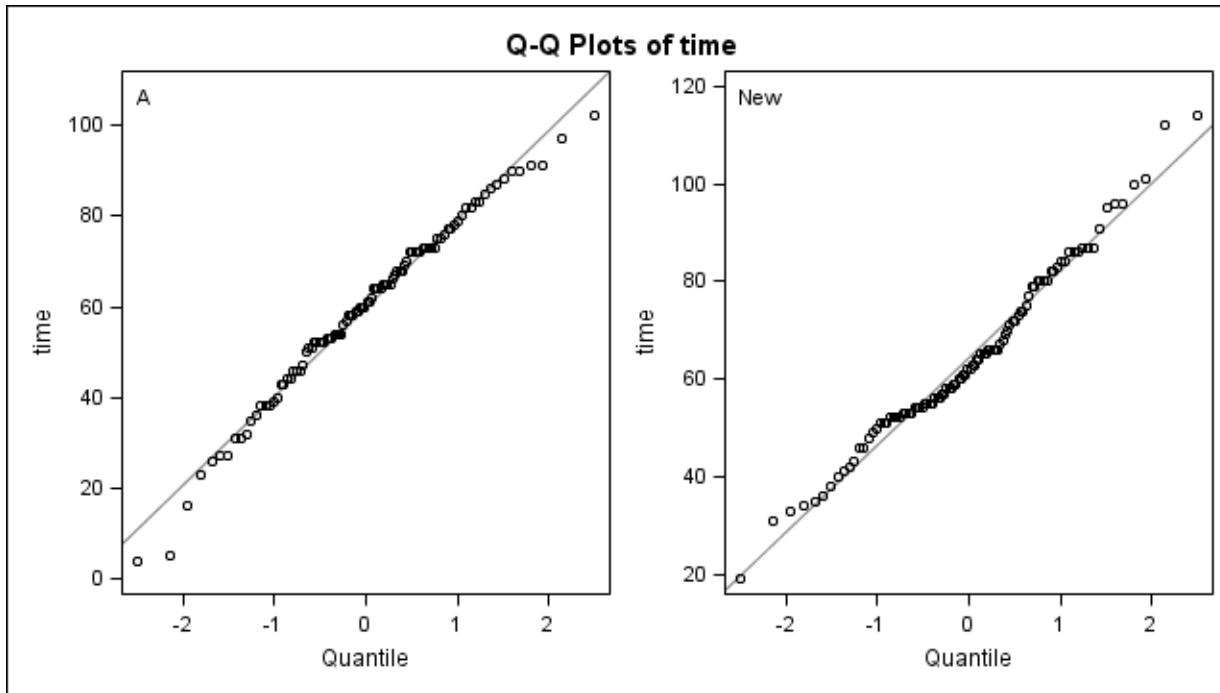
```
/*st004d01 */
ods graphics on;
proc ttest data=st092.phone_new_and_a plots(shownull)=interval;
  class team;
  var time;
  title "Two-Sample t-test Comparing New and A Team";
run;
ods graphics off;
```

Selected PLOTS()= options in the PROC TTEST statement:

PLOTS(SHOWNULL)=INTERVAL includes a plot of confidence intervals of the difference between groups. SHOWNULL places a vertical reference line at the mean value of the null hypothesis ($H_0=0$ by default).

First it is advisable to verify the assumptions of t -tests. There is an assumption of normality of the distribution of each group. This assumption can be verified with a quick check of the Summary Panel and the Q-Q Plot.





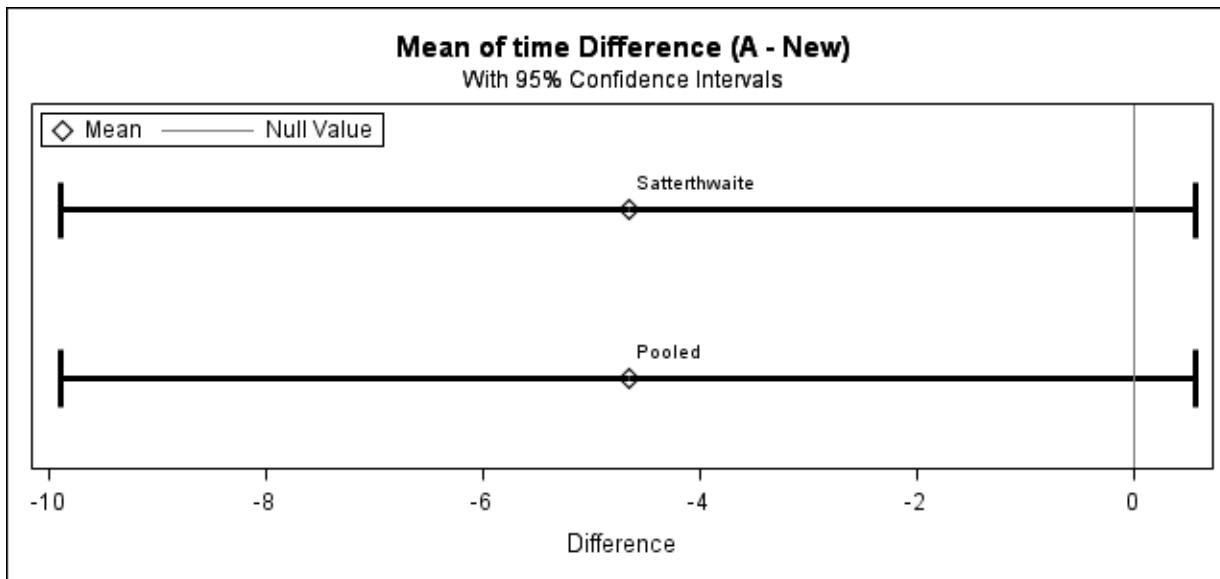
The Q-Q Plot (Quantile-Quantile Plot) is similar to the Normal Probability plot you saw earlier. The x-axis for this plot is just scaled as quantiles, rather than probabilities. For each group it seems that the data approximate a normal distribution.

The statistical tables for the TTEST procedure are displayed below.

Two-Sample t-test Comparing New and A Team							
The TTEST Procedure							
Variable: time							
team	N	Mean	Std Dev	Std Err	Minimum	Maximum	
A	100	59.7200	19.5319	1.9532	4.0000	102.0	
① New	100	64.3800	17.9228	1.7923	19.0000	114.0	
Diff (1-2)		-4.6600	18.7447	2.6509			
team	Method	Mean	95% CL Mean	Std Dev	95% CL	Std Dev	
A		59.7200	55.8444	63.5956	19.5319	17.1492	
New		64.3800	60.8237	67.9363	17.9228	15.7364	
Diff (1-2)	Pooled	-4.6600	-9.8876	0.5676	18.7447	17.0662	
Diff (1-2)	Satterthwaite	-4.6600	-9.8878	0.5678		20.7921	
	Method	Variances	DF	t Value	Pr > t		
③	Pooled	Equal	198	-1.76	0.0803		
	Satterthwaite	Unequal	196.55	-1.76	0.0803		
Equality of Variances							
Method	Num DF	Den DF	F Value	Pr > F			
②	Folded F	99	99	1.19	0.3938		

- 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 0.3938. 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 (Pooled) *t*-test, you do not reject the null hypothesis that the group means are equal. The mean difference between team A and New is -6.66 seconds. However, you conclude that there is no significant difference in the average amount of time it takes team A and team New to answer the phone.

Return your attention to the Statistics table. Because the confidence interval for the mean (-9.8878 0.5678) includes 0, you cannot even say with 95% confidence that the difference between team A and New is not zero. This is equivalent to the *p*-value being greater than 0.05.



Confidence intervals are shown in the output object titled Difference Interval Plot. Because the variances here are so similar between the two groups, the Pooled and Satterthwaite intervals (and p -values) are very similar. Notice that the lower bound of the Pooled interval extends past zero.

4.2 One-Way ANOVA

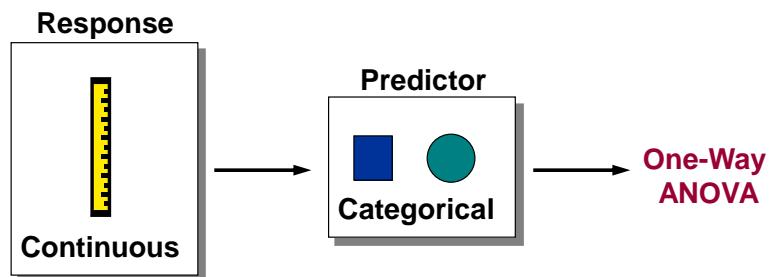
Objectives

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

13

Overview

Are there any differences among the population means?



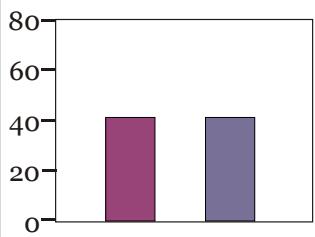
14

Analysis of variance (ANOVA) is a statistical technique used to compare the means of two or more groups of observations or treatments. For this type of problem, you have a

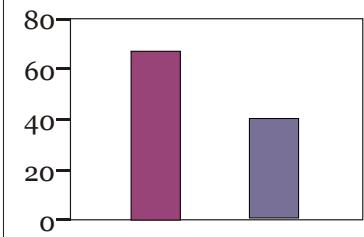
- continuous dependent variable, or *response* variable
- categorical independent variable also called a *predictor* or *explanatory* variable.

More than Two Groups

H_0 : All means equal

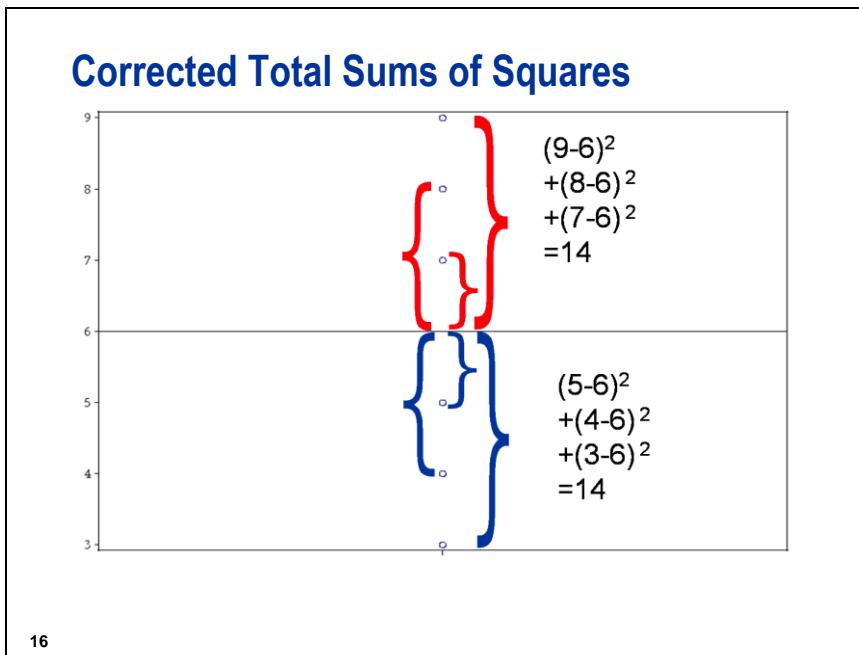


H_1 : At least one mean different



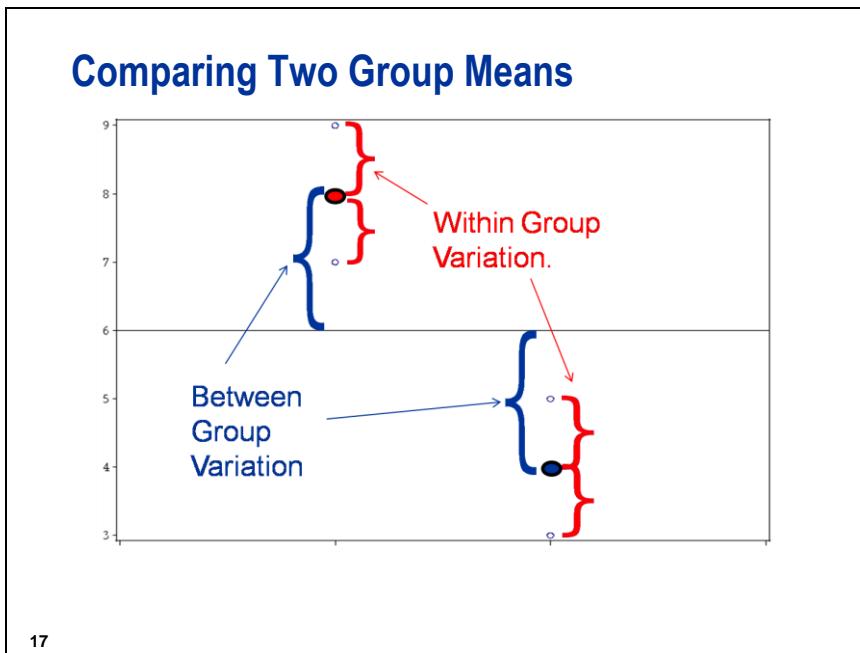
15

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?



Suppose you measured a target variable for two different groups.

To calculate the total amount of variability in the target variable, add up the squared distances to the overall mean. This has corrected the individual value with the overall mean, so is often called the *Corrected Total Sum of Squares*.

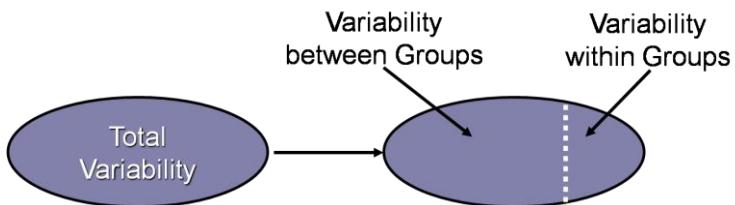


17

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

Partitioning Variability in ANOVA



18

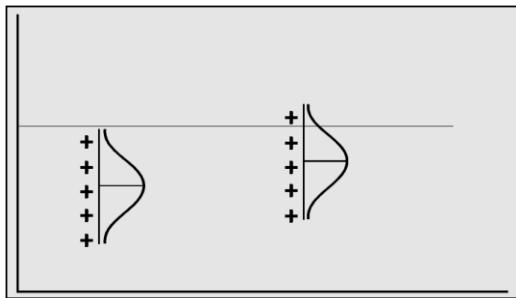
In ANOVA, the corrected total sum of squares is partitioned into two parts, the Model Sum of Squares and the Error Sum of Squares.

Model Sum of Squares (SSM) the variability explained by the independent variable and therefore represented by the **between** treatment sums of squares.

Error Sum of Squares (SSE) the variability not explained by the independent variable. Also referred to as **within** treatment variability or residual sum of squares.

Total Sum of Squares (SST) the **overall** variability in the response variable.
 $SST=SSM + SSE$.

Assumptions in ANOVA



- Observations are independent.
- Errors are normally distributed.
- All groups have equal response variances

19

Independence implies that the ϵ_{ij} s in the theoretical model are uncorrelated. The independence assumption should be verified with good data collection. In some cases, residuals can be used to verify this assumption.

The errors are assumed to be normally distributed for every group or treatment.

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) for one-way ANOVA.

Analysis of Two Groups

Verify Assumptions



Test Hypothesis

20

The validity of the p -values depends on the data meeting the assumptions for ANOVA. Therefore, it is good practice to verify those assumptions in the process of performing the analysis of group differences.

Assessing ANOVA Assumptions

- Good data collection methods help ensure the independence assumption.
- Diagnostic plots from PROC GLM can be used to verify the assumption that error is approximately normally distributed.
- The GLM procedure produces a hypothesis test with the HOVTEST option in the MEANS statement. H_0 for this hypothesis test is that the variances are equal for all populations.

21

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.

	team	time	resid	pred
99	New	53	-11.38	64.38
100	New	72	7.62	64.38
101	A	23	-36.72	59.72
102	A	83	23.28	59.72
103	A	76	16.28	59.72
104	A	73	13.28	59.72

22

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 and approximately equal groups (balanced design), 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 PLOTS=options;
  CLASS variables;
  MODEL dependents=independents </ options>;
  MEANS effects </ options>;
  LSMEANS effects </ options>;
  OUTPUT OUT=SAS-data-set keyword=variable...;
RUN;
QUIT;
```

23

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 GLM Procedure- comparing two groups

The following is a summary of what you will accomplish in this demonstration:

- Verify the assumptions of a One –Way ANOVA using the Diagnostics Plots.
- Use the GLM procedure to test the hypothesis that the average time taken to answer the phone for the new team is equal to the average time taken to answer the phone for the A team.

```
/*st004d02 */
ods graphics on;
proc glm data=st092.phone_new_and_a
            PLOTS(only)=diagnostics(unpack);
  class team;
  model time=team;
  means team / hovtest;
  title 'Testing for Equality of Means with PROC GLM';
run;
quit;

ods graphics off;
```

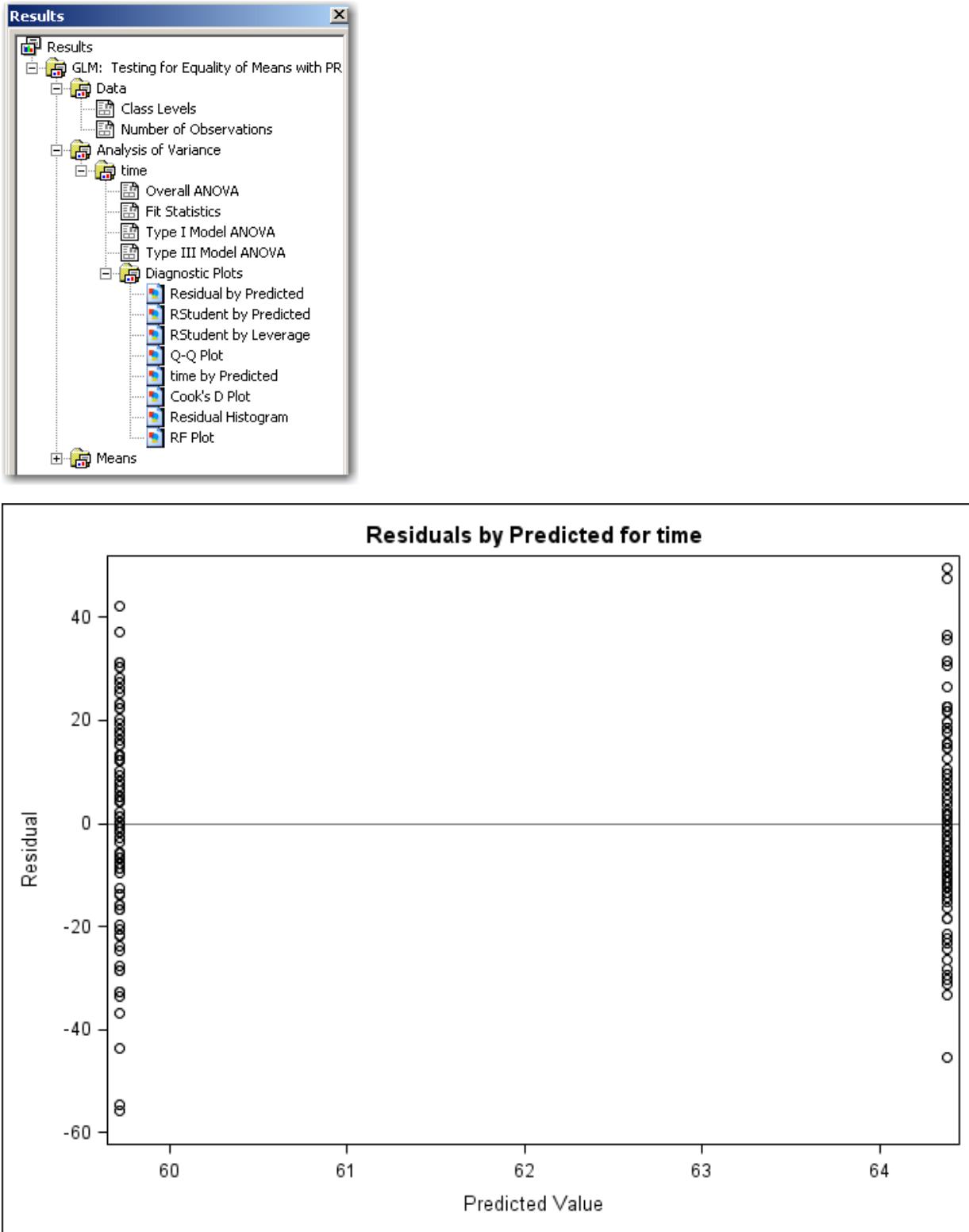
Selected MEANS statement option:

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.

Selected PLOTS option:

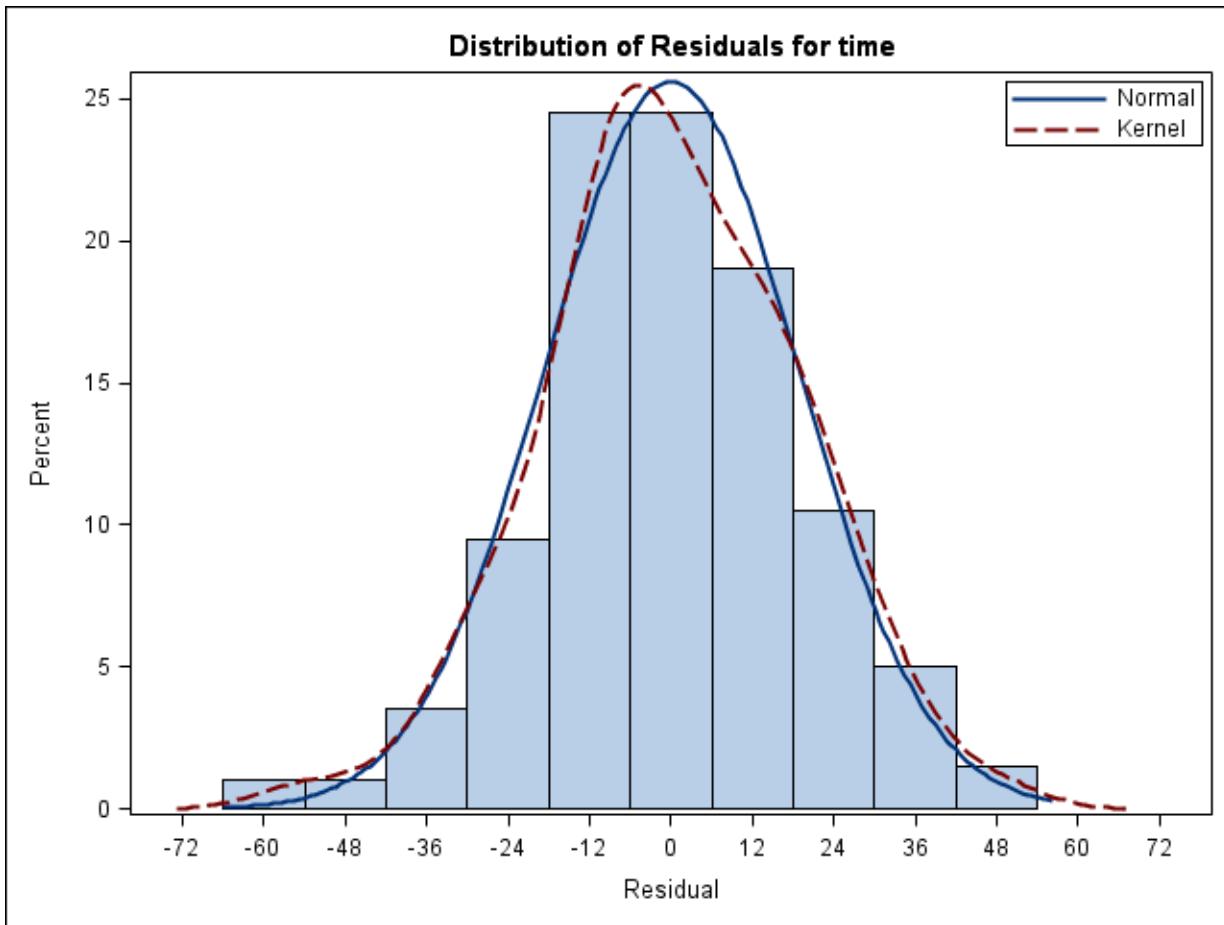
DIAGNOSTICS(UNPACK) produces a panel display of diagnostic plots for linear models. The UNPACK option unpacks the individual plots from the panel display.

It is good practice to look at your diagnostic plots to check for the validity of your ANOVA assumptions.. You might start by looking at the Residual by Predicted plot, which can be selected by expanding the GLM output in the Results frame:

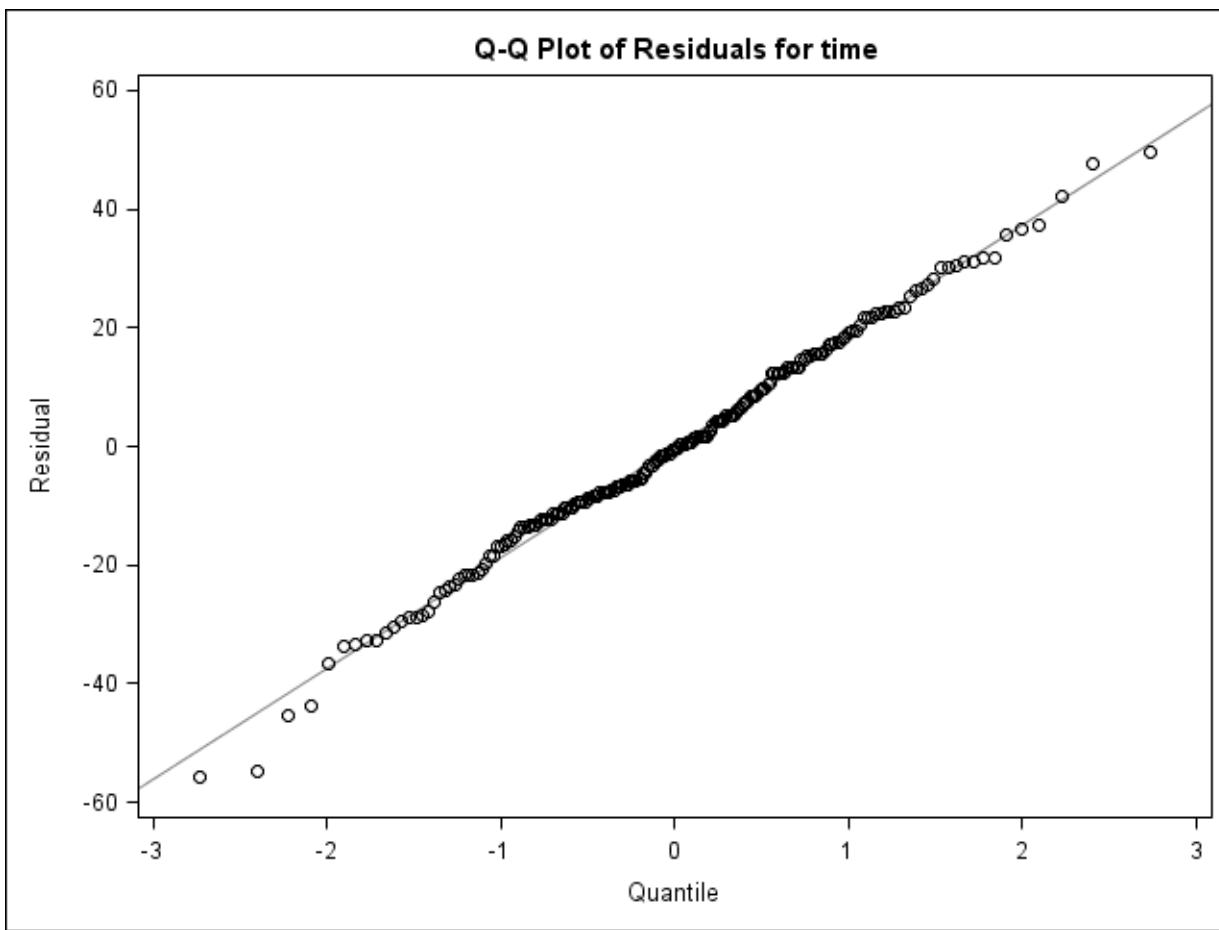


The graph above is a plot of the residuals versus the fitted values from the ANOVA model. Essentially, you are looking for a random scatter within each group. Any patterns or trends in this plot can indicate model assumption violations.

To check the normality assumption, open the Residual Histogram and Q-Q Plot.



This histogram looks normal.



The data values stay close to the diagonal reference line and give strong support to the assumption of normally distributed errors.

Near the end of the tabular output, you can check first the assumption of equal variances.

Testing for Equality of Means with PROC GLM					
The GLM Procedure					
Levene's Test for Homogeneity of time Variance					
ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
team	1	178002	178002	0.68	0.4107
Error	198	51857566	261907		

The output above 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 are equal over each **team** groups. The *p*-value of 0.4107 is not smaller than your typical alpha level of 0.05 and therefore the null hypothesis stands. The equal variance assumption is verified.

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.

Now, 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
team	2	A New
Number of Observations Read 200		
Number of Observations Used 200		

The second page of the output contains all of the information that is needed to test the equality of the treatment means. It 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	1085.78000	1085.78000	3.09	0.0803
Error	198	69569.72000	351.36222		
Corrected Total	199	70655.50000			

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.
- Error DF is the difference between the corrected total DF and the Model DF..

Mean squares are calculated by taking sums of squares and dividing by the corresponding degrees of freedom. They can be thought of as variances.

- Mean square for error (MSE) is an estimate of σ^2 , the constant variance assumed for all treatments.
- If $\mu_i = \mu_j$, for all $i \neq j$, then the mean square for the model (MSM) is also an estimate of σ^2 .

- If $\mu_i \neq \mu_j$, for any $i \neq j$, then MSM estimates σ^2 plus a positive constant.

$$\bullet \quad F = \frac{MSM}{MSE}.$$

 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 (μ), MSE equals the variance.

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 greater than 0.05, you conclude that there is **no** statistically significant difference between the means.

R-Square	Coeff Var	Root MSE	time Mean
0.015367	30.20896	18.74466	62.05000

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.

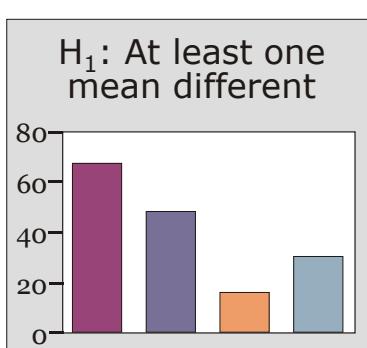
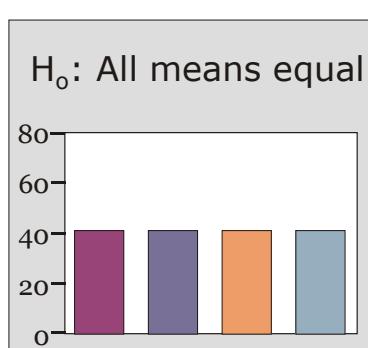
The time Mean is the mean of all of the data values in the variable **time** without regard to **team**.

Some interpret the R^2 value as the “proportion of variance accounted for by the model”. Therefore, one might say that in this model, **team** explains about 1.5% of the variability of **time**.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
team	1	1085.780000	1085.780000	3.09	0.0803
Source	DF	Type III SS	Mean Square	F Value	Pr > F
team	1	1085.780000	1085.780000	3.09	0.0803

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.

More than Two Groups



25

Recall that the objective is to determine whether there are differences between group means. Now, with more than two groups, you are testing:

- Null hypothesis: all means are equal
- Alternative hypothesis: at least one mean is different from one of the other means

Customer Call Centre

We want to compare the New team to three existing teams:

- New
- Existing A
- Existing B
- Existing C

26

In total there are four teams of customer call centre operators. We want to look at the differences between all four teams. We have already looked at differences between the New team and the existing A team. The data for all four teams is stored in SAS data set **st092.phone_all_groups**.



The GLM Procedure- more than two groups

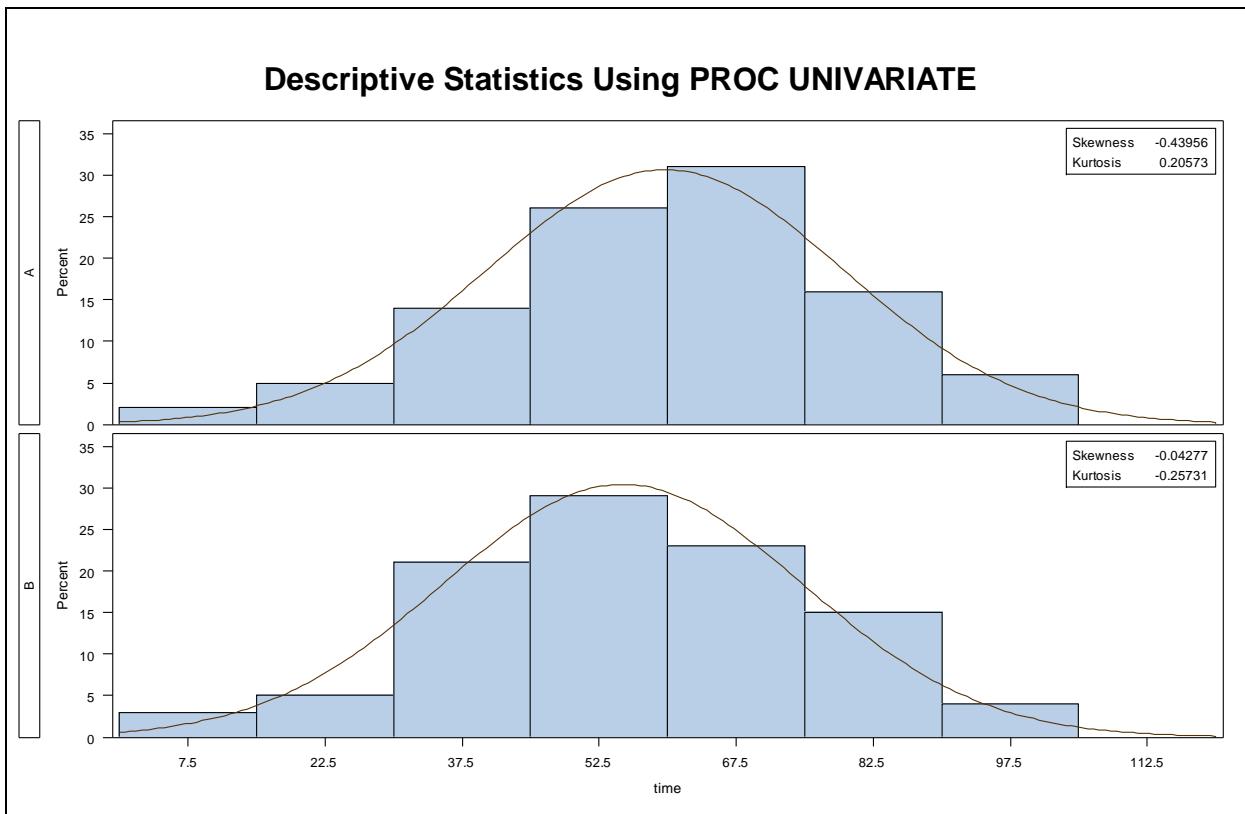
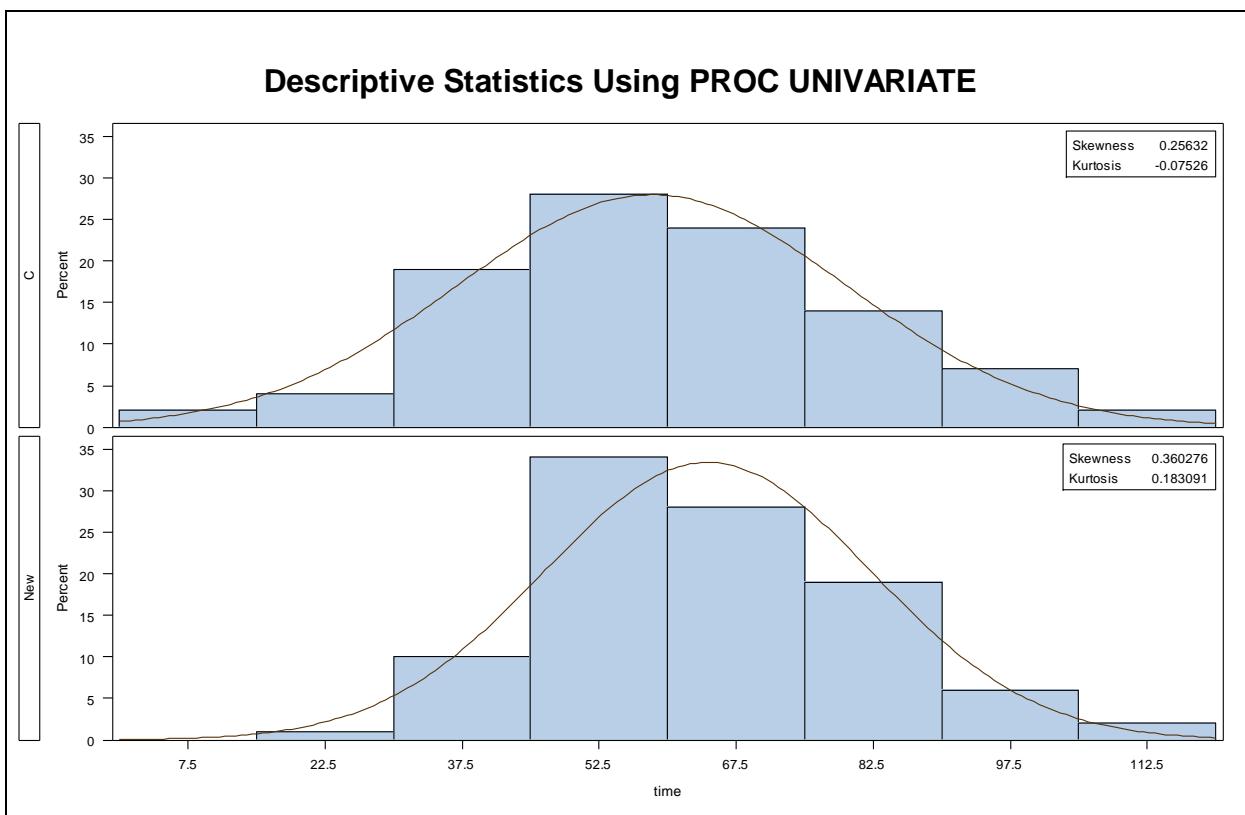
The following is a summary of what you will accomplish in this demonstration:

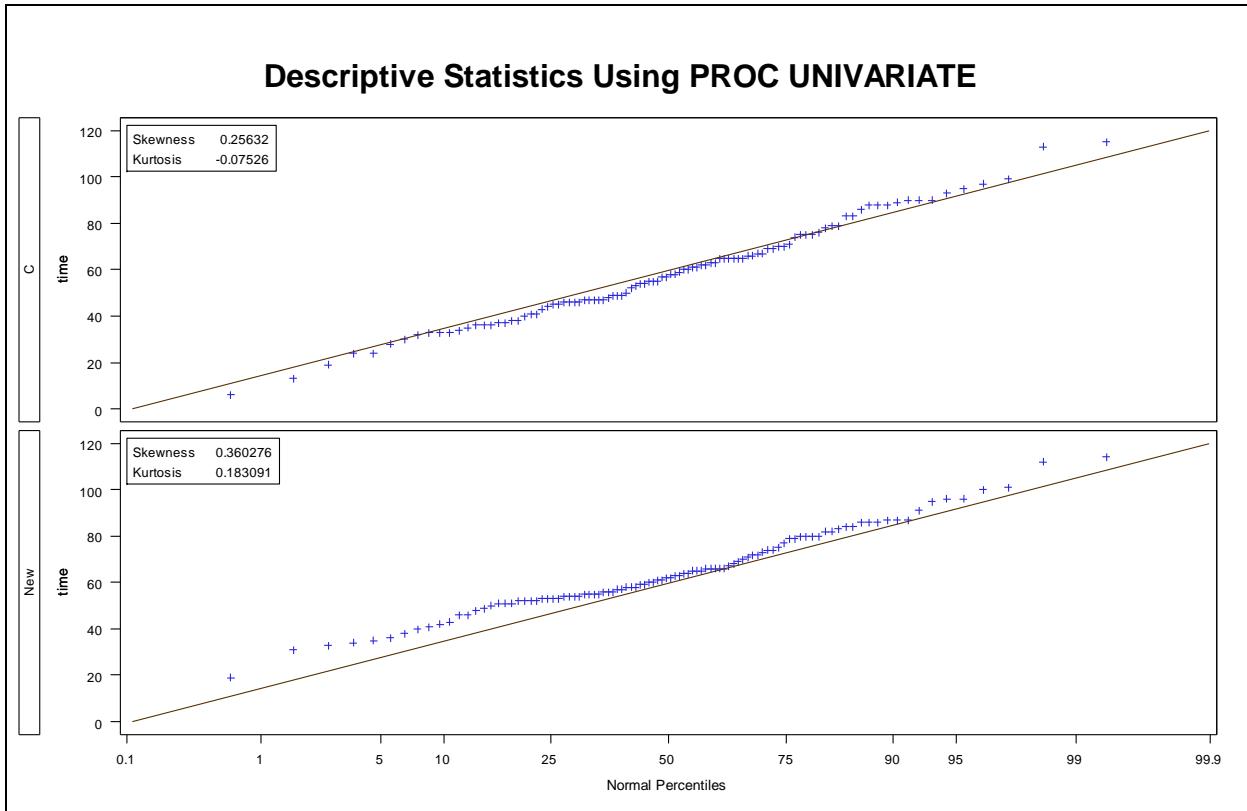
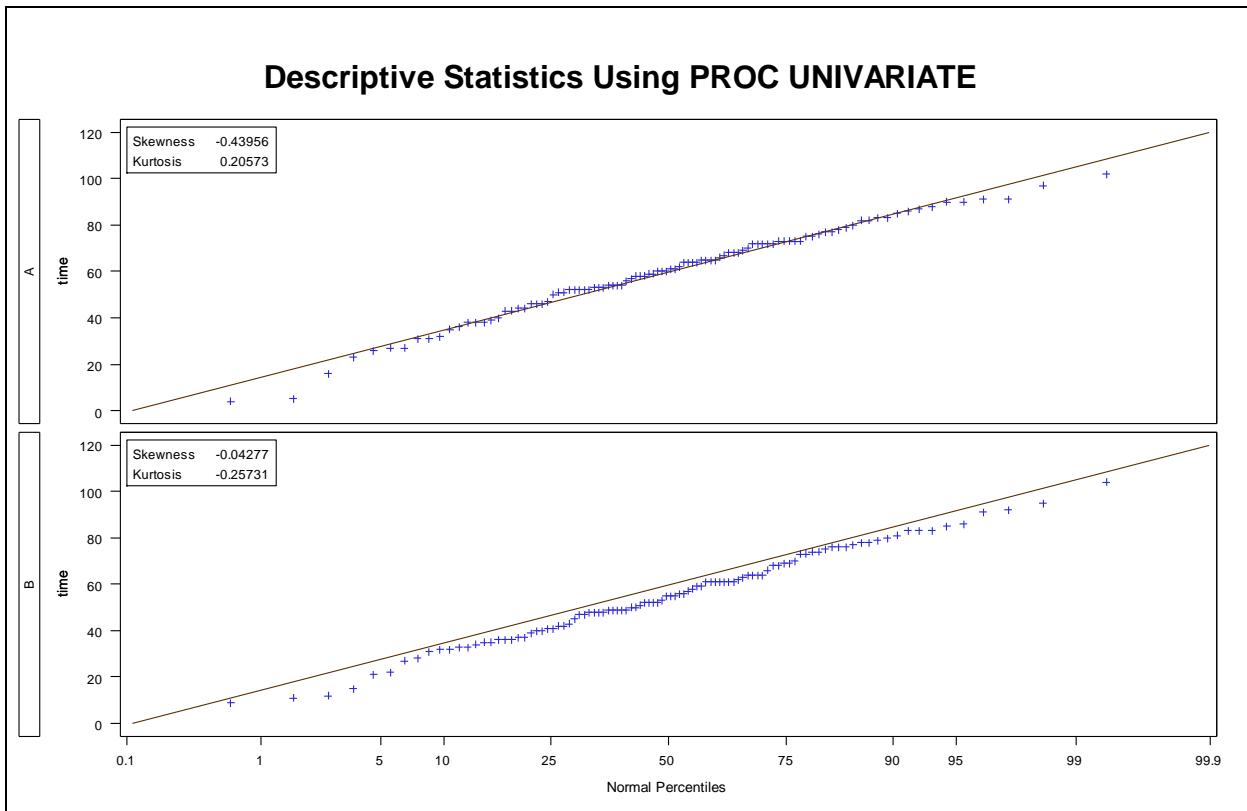
- Analyse the phone data stored in the **st092.phone_all_groups** data set. The variables in the data set are:
 - team** New, A, B and C
 - time** The time in seconds for operators to answer the phone.
- Verify the assumptions of the ONE-Way ANOVA
- Test the Hypothesis that the average time taken for each group to answer the phone is equal.

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, including histograms.

```
/*st004d03*/
ods graphics off;
proc univariate data=st092.phone_all_groups ;
  class team;
  var time;
  histogram time / normal(mu=est sigma=est);
  inset skewness kurtosis / position=ne;
  probplot time / normal(mu=est sigma=est);
  inset skewness kurtosis;
  title 'Descriptive Statistics Using PROC UNIVARIATE';
run;
```

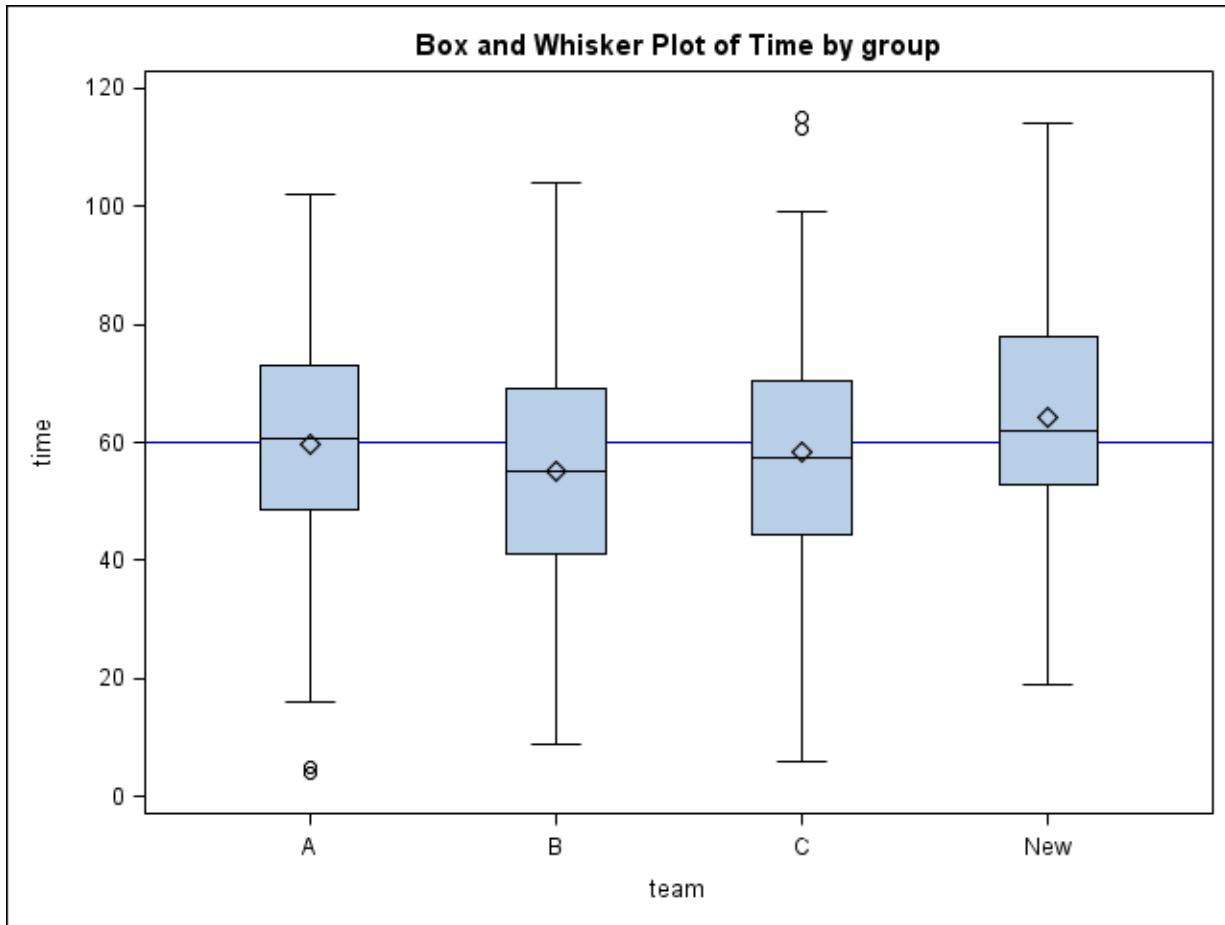
Check all histograms and the probability plots for normality.





For an over-all view of the data use the proc sgplot to create a side-by-side box plot.

```
/*st004d03*/
proc sgplot data=st092.phone_all_groups;
  refline 60 / axis=y lineattrs=(color=blue);
  vbox time/category=team ;
  title "Box and Whisker Plot of Time by group";
run;
```



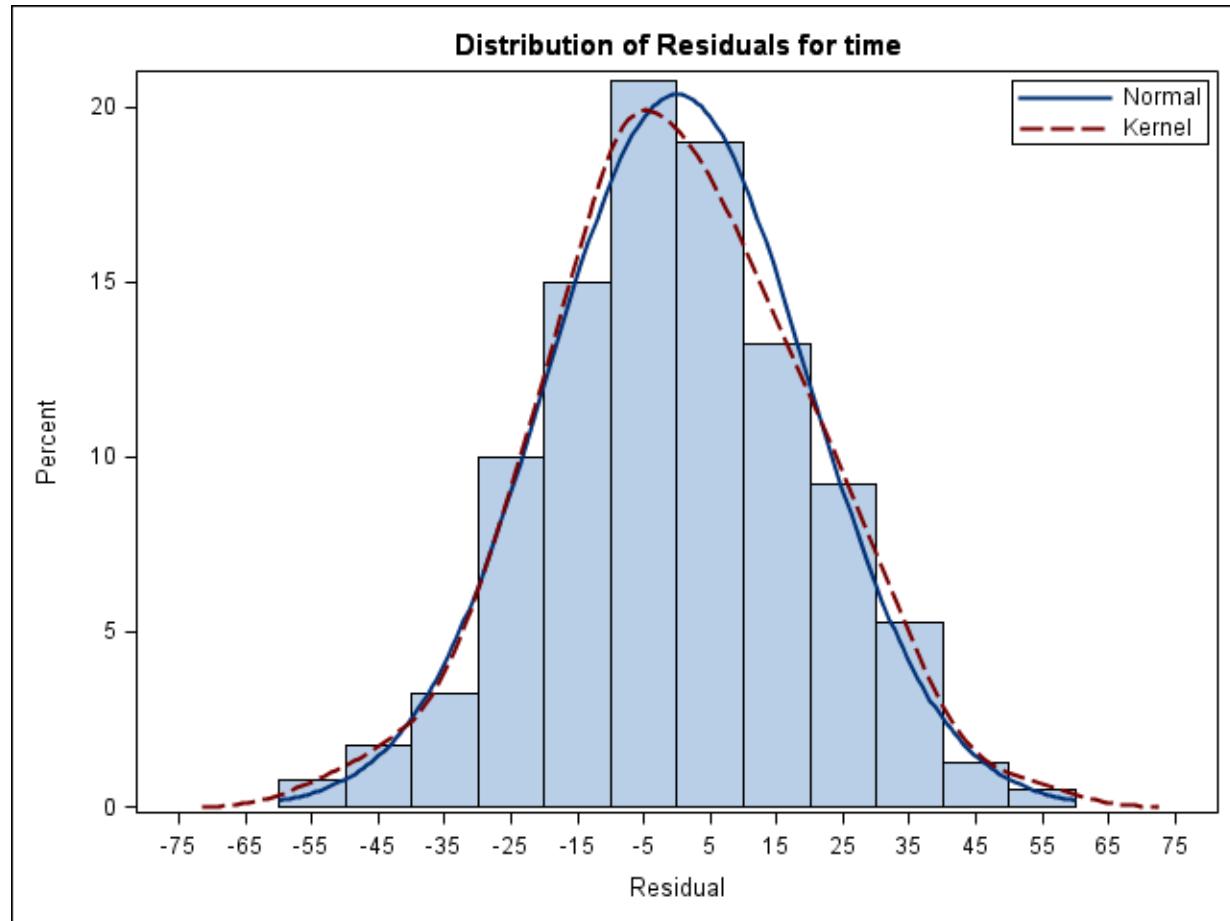
The New team does appear to have a slightly larger mean than the other teams and team B has the lowest mean. Could these differences reasonably occur by chance or are they statistically significant?

A boxplot can also be created in the PROC GLM by adding BoxPlot to the plots option in the PROC GLM Statement. The main difference is that more control is given when using a sgplot, for example, a reference line can be indicated.

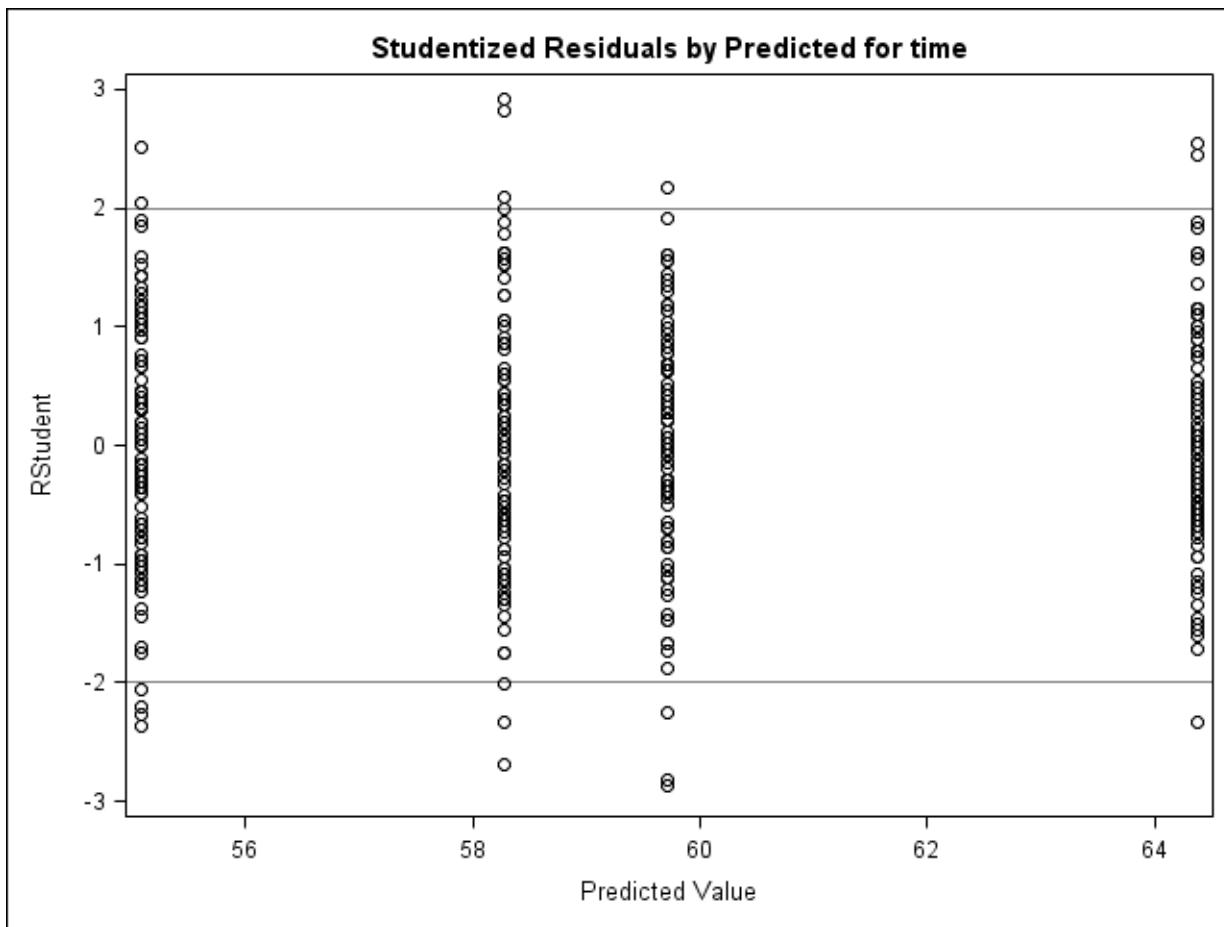
To answer that question, the GLM procedure to test the null hypothesis that the means are equal, needs to be used.

```
/*st004d03*/
proc glm data=st092.phone_all_groups
PLOTS(only)=diagnostics(unpack);
  class team;
  model time=team;
  means team / hovtest;
  title 'Testing for Equality of Means with PROC GLM';
run;
quit;
ods graphics off;
```

As always, verify the assumptions before analyzing the output.



The residuals are normally distributed.



There do not appear to be any violations of the assumptions, based on the plot of residuals versus predicted values.

Full output of the PROC GLM.

```

-----  

Testing for Equality of Means with PROC GLM  

The GLM Procedure  

Class Level Information  

Class      Levels   Values  

②        team       4     A B C New  

Number of Observations Read      400  

Number of Observations Used     400  

-----  

Testing for Equality of Means with PROC GLM

```

The GLM Procedure						
Dependent Variable: time						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
③ Model	3	4483.6475	1494.5492	3.86	0.0096	
	396	153252.7900	387.0020			
	399	157736.4375				
	R-Square	Coeff Var	Root MSE	time Mean		
	0.028425	33.13938	19.67237	59.36250		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
team	3	4483.647500	1494.549167	3.86	0.0096	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
team	3	4483.647500	1494.549167	3.86	0.0096	
<hr/>						
Testing for Equality of Means with PROC GLM						
The GLM Procedure						
Levene's Test for Homogeneity of time Variance						
ANOVA of Squared Deviations from Group Means						
①	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
	team	3	923238	307746	1.06	0.3670
	Error	396	1.1522E8	290962		
<hr/>						
Testing for Equality of Means with PROC GLM						
The GLM Procedure						
④	Level of team	N	-----time-----			
			Mean	Std Dev		
	A	100	59.7200000	19.5319371		
	B	100	55.0800000	19.6754885		
	C	100	58.2700000	21.4046417		
<hr/>						

① Firstly, check the homogeneity of variance assumption

Step 1- Set Hypothesis

$$H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4$$

H_1 : At least one variance is not equal to another/others.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$p\text{-value}=0.3670$$

Step 4- Decision Rule.

The $p\text{-value} > \alpha$, Fail to reject H_0 , therefore there is no evidence to say the variances are not equal.

② Turn your attention to the first page of the PROC GLM output, check the values of the class variable and the number of observations read versus the number of observations used. There are no missing information.

③ Now look at the analysis of variance table to test the hypothesis that the means are equal.

Step 1- Set Hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_1 : At least one mean is not equal to another/others.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$p\text{-value}=0.0096$$

Step 4- Decision Rule.

The $p\text{-value} < \alpha$, Reject H_0 , therefore at least one of the means is/are not equal to another.

④ Remember, the ANOVA table tests the assumption that the means are equal, it does not test which means are statistically significantly different from the other. Before we proceed to the statistical analysis that will test this hypothesis, let us examine the output from the MEANS statement in Proc GLM, that may provide some insight into where the differences lie.

Team B has the lowest mean time and the New team has the highest mean time. Multiple comparison techniques can be used to determine whether these are statistically significant differences.

Steps for ANOVA Summary

Null Hypothesis: All means are equal.

Alternative Hypothesis: At least one mean is different.

1. Produce descriptive statistics.
2. Verify assumptions.
 - Independence
 - Errors are normally distributed
 - Variances are equal for all groups
3. Examine the p -value on the ANOVA table. If the p -value is less than alpha, reject the null hypothesis.

My Groups Are Different. What Next?

- The p -value for **team** indicates you should reject the H_0 that all groups are the same.
- Which team is different from which other?
- Should I just go back and do a bunch of t -tests?

Multiple Comparison Methods

Comparisonwise Error Rate ($\alpha=0.05$)	Number of Comparisons	Experimentwise Error Rate ($\alpha=0.05$)
.05	1	.05
.05	3	.14
.05	6	.26
.05	10	.40

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

30

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.

Tukey's Multiple Comparison Method

This method is appropriate when considering pairwise comparisons only.

The experimentwise error rate is

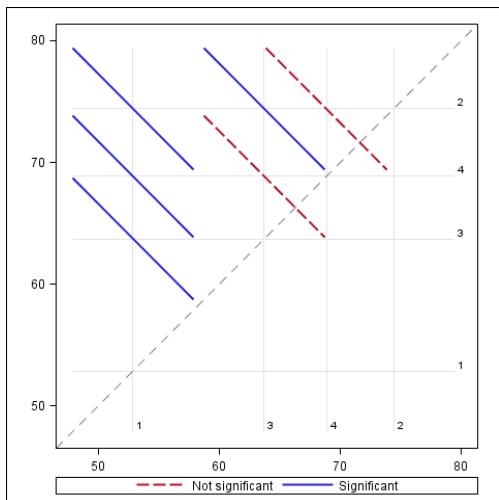
- equal to alpha when **all** pairwise comparisons are considered
- less than alpha when **fewer** than all pairwise comparisons are considered.

31

A pairwise comparison examines the difference between two treatment means. All pairwise comparisons means 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.

Diffograms



32

A diffogram can be used to quickly tell if two group means are statistically significant. The point estimates for the differences between pairs of group means can be found at the intersections of the gray lines. The colored diagonal lines show the confidence intervals. If the confidence intervals for the groups overlap at all, then the diagonal line for the pair will cross over the broken gray diagonal in the middle of the plot. In that case, the diagonal line for the pair will be broken and colored red. If the confidence intervals never do overlap, then the diagonal line for the pair will be solid and colored blue. These plots are automatically generated when using the PDIFF=ALL option in the LSMEANS statement.

Steps for ANOVA Summary

Null Hypothesis: All means are equal.

Alternative Hypothesis: At least one mean is different.

1. Produce descriptive statistics.
2. Verify assumptions.
 - Independence
 - Errors are normally distributed
 - Variances are equal for all groups
3. Examine the p -value on the ANOVA table. If the p -value is less than alpha, reject the null hypothesis.
4. If, and only if, the ANOVA table suggests there is a significant difference between the means, a comparison method is used to verify the significant differences

33



Post Hoc Pairwise Comparison

The following is a summary of what you will accomplish in this demonstration:

- As the ANOVA test has indicated that at least one group mean is different from one other group mean, perform a Post Hoc Pairwise Comparison to discover the differences.

```
/*st004d04*/
ods graphics on;
ods select LSMeans Diff MeanPlot DiffPlot;
proc glm data=st092.phone_all_groups ;
  class team;
  model time=team;
  lsmeans team / pdiff=all adjust=tukey;
  title 'Testing for Equality of Means with PROC GLM';
run;
quit;
ods graphics off;
```

Multiple LSMEANS statements are permitted, although typically only one type of multiple comparison method would be used for each LSMEANS effect.

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 TUKEY option uses Tukey's adjustment 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.

The ODS Select Statement allows selection of output by name.

LSMeans lists the output showing the means for each group.

Diff shows p -values from pair-wise comparisons of all possible combinations of means.

MeanPlot graphically shows the Least Square Means.

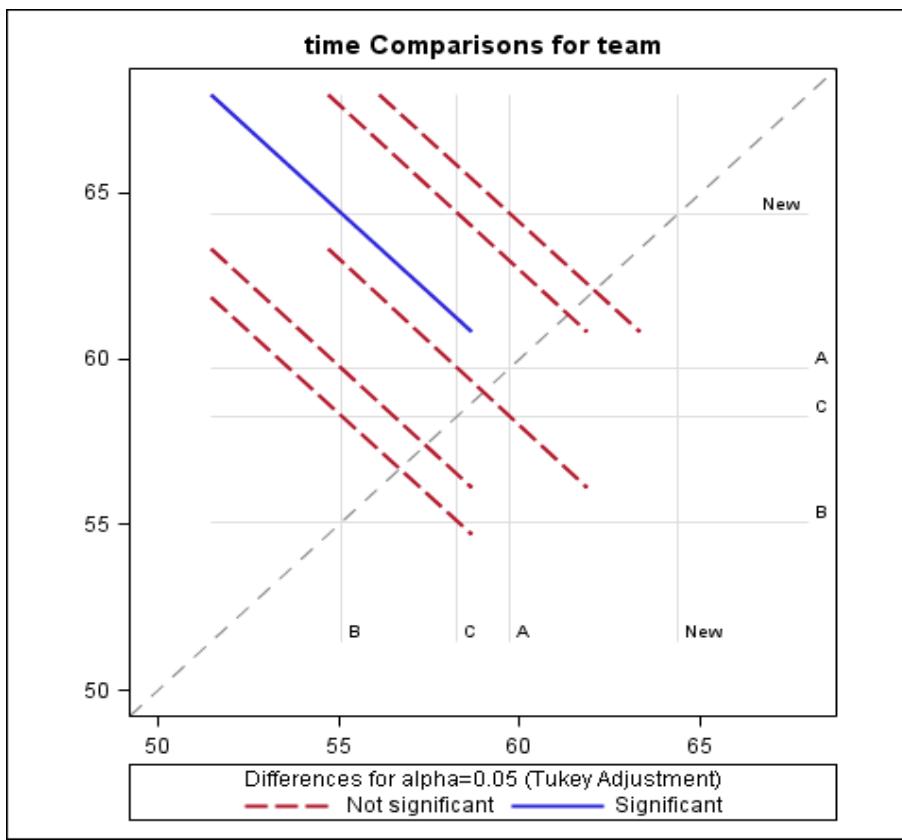
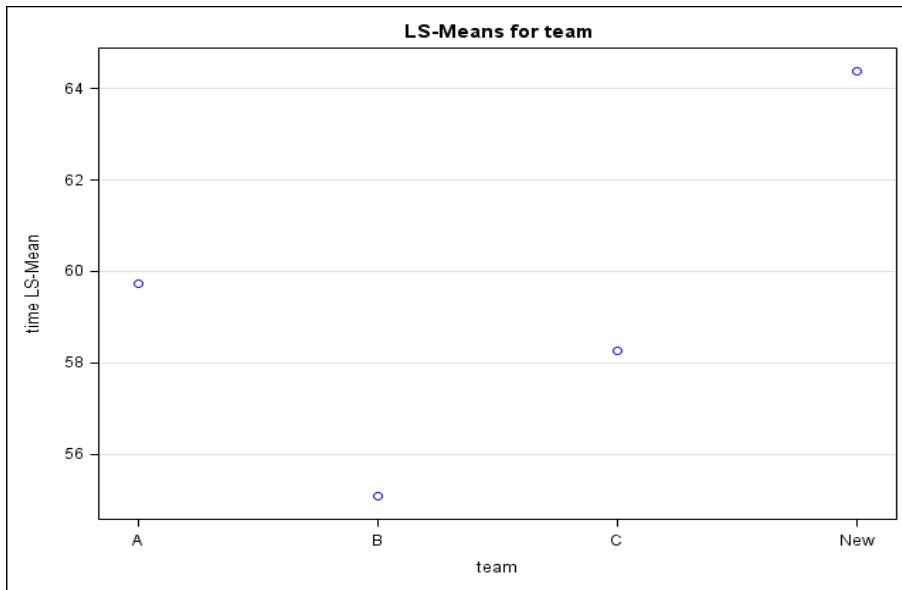
DiffPlot graphically shows the Tukey-adjusted differences among the LSMeans.

PROC GLM Output

Testing for Equality of Means with PROC GLM			
The GLM Procedure			
Least Squares Means			
Adjustment for Multiple Comparisons: Tukey			
team	time	LSMEAN Number	LSMEAN
A		59.7200000	1
B		55.0800000	2
C		58.2700000	3
New		64.3800000	4
 Least Squares Means for effect team			
Pr > t for H0: LSMean(i)=LSMean(j)			
 Dependent Variable: time			
i/j	1	2	3
1		0.3422	0.9540
2	0.3422		0.6608
3	0.9540	0.6608	
4	0.3383	0.0050	0.1260

The first part of the output shows the 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.

The only significant pairwise difference is between column 4 and row 2, which is the New team and existing team B.



The blue line shows the significant difference (the confidence limit for the difference does not cross the diagonal equivalence line) between team New and existing team B.

Chapter 5 Exploratory Data Analysis

5.1 Exploratory Data Analysis.....5-3

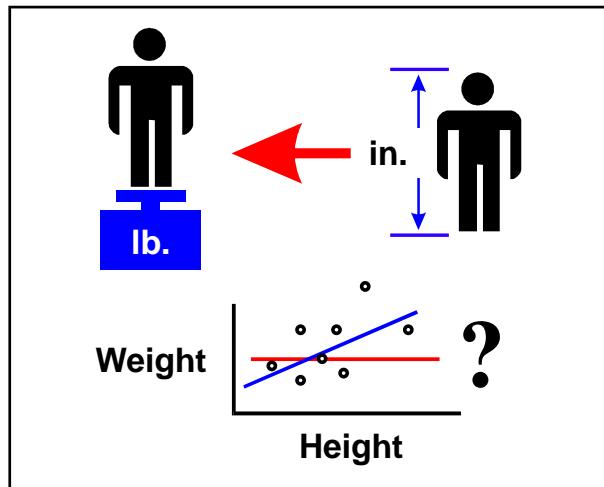
Demonstration: Data Exploration, Correlations, and Scatter Plots5-17

5.1 Exploratory Data Analysis

Objectives

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

Two Continuous Variables



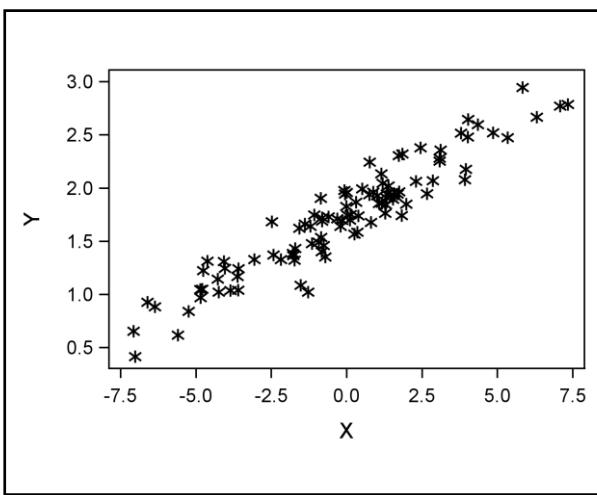
3

In the previous chapter, you learned that when you have a categorical 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: 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

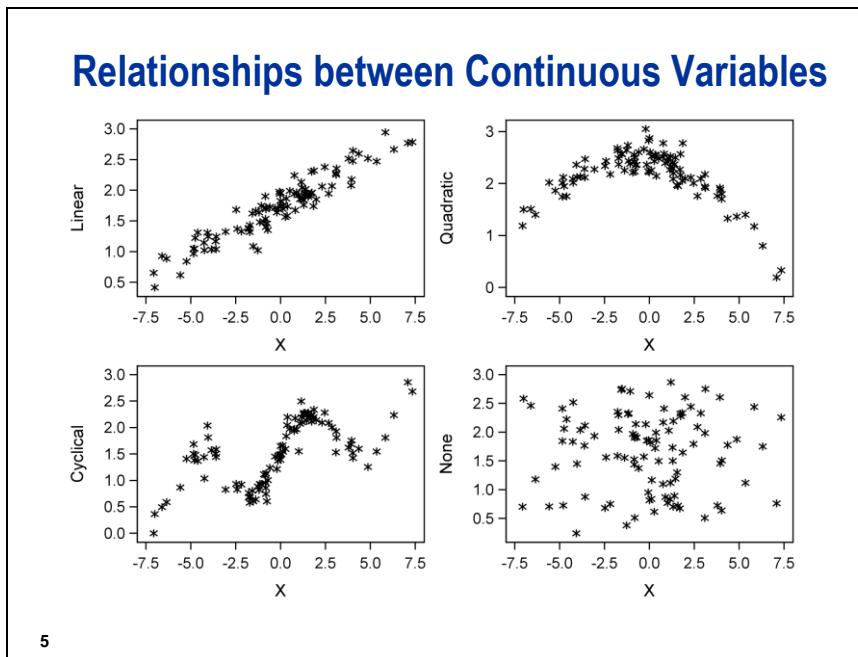


4

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

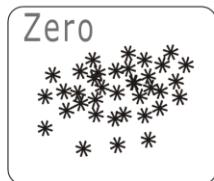
- explore the relationships between two variables
- locate outlying or unusual values
- identify possible trends
- identify a basic range of Y and X values
- communicate data analysis results.



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.

Correlation



6

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

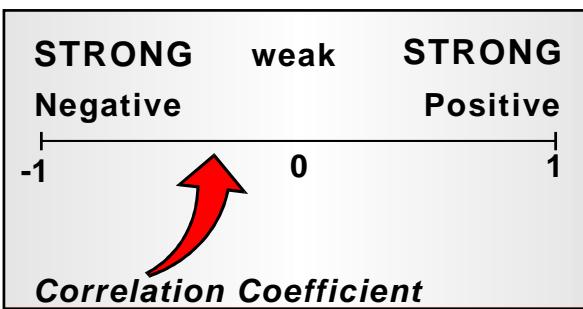
You can classify correlated variables according to the type of correlation:

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)

Pearson Correlation Coefficient



7

Correlation statistics measure the degree of linear association 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 association between the two variables
- close to 0 if there is no linear association between the two variables
- greater than 0 if there is a positive linear association
- less than 0 if there is a negative linear association.

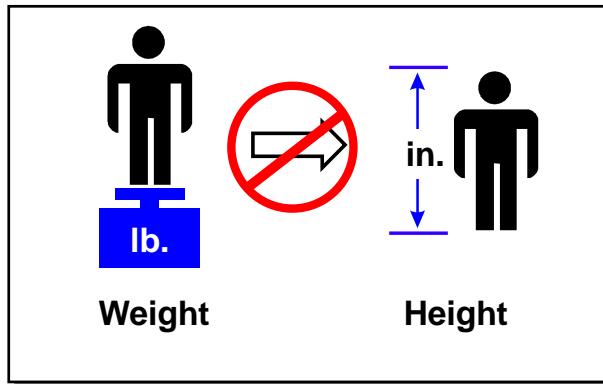
Hypothesis Test for a Correlation

- The parameter representing a correlation is ρ .
- ρ is estimated by the sample statistic r .
- $H_0: \rho=0$
- Rejecting H_0 indicates confidence that ρ is not exactly zero.
- A p -value does not measure the magnitude of the association.
- Sample size affects the p -value.

8

The null hypothesis for a test of a correlation coefficient is $\rho=0$. Rejecting the null hypothesis only means that you can be confident that the true population correlation is not 0. Small p -values can occur (as with many statistics) because of very large sample sizes. Even a correlation of 0.01 can be statistically significant with a large enough sample size. Therefore, it is important to also look at the value of r itself to see if it is a meaningfully large correlation.

Correlation versus Causation

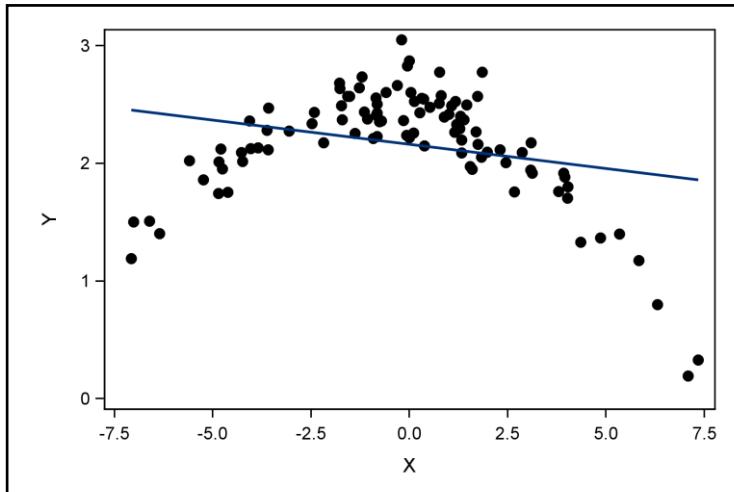


9

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.
- “Correlation does not imply causation.”

Missing Another Type of Relationship

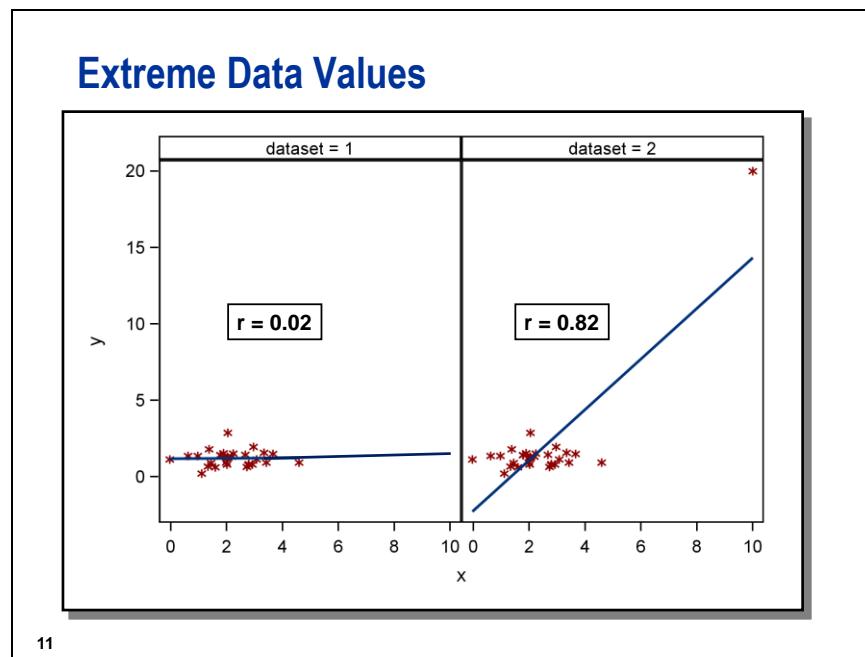


10

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

- Pearson correlation coefficients measure linear relationships.
- A Pearson correlation coefficient close to 0 indicates that there is not a strong linear relationship between two variables.
- A Pearson 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.



11

Correlation coefficients are highly affected by a few extreme values of either variable. The scatter plots above shows that the degree of linear relationship is mainly determined by one point. If you include the unusual point in the data set, the correlation is close to 1. If you do not include it, 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, $x=10$). 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. In this case, it is greatly influential.

The CORR Procedure

General form of the CORR procedure:

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

12

You can use the CORR procedure to produce correlation statistics and scatter plots 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.

ID The ID statement specifies one or more additional tip variables to identify observations in scatter plots and scatter plot matrix.

The CORR Procedure

- Scatter plots and scatter plot matrices are available through ODS Graphics.
- ID statement enables you to specify additional variables to identify observations in scatter plots and scatter plot matrices.

13

Exploratory analysis in preparation for multiple regression often involves looking at bivariate scatter plots and correlations between each of the predictor variables and the response variable. It is not suggested that exclusion or inclusion decisions be made on the basis of these analyses. The purpose is to explore the shape of the relationships (because linear regression assumes a linear shape to the relationship) and to screen for outliers. You will also want to check for multivariate outliers when you test your multiple regression models later.

PROC CORR provides bivariate correlation tables. These tables are accompanied by ODS Statistical Graphics. An ID statement in the procedure helps to identify outliers in the plots.

PROC CORR PLOTS Syntax and Selected Options

- **PLOTS <(ONLY)> <= plot-request>**
- **PLOTS <(ONLY)> <= (plot-request < plot-request >) >**
 - **ALL**
 - **MATRIX <(matrix-options)>**
 - **SCATTER <(scatter-options)>**
 - **HIST | HISTOGRAM**
 - **NVAR=ALL | n**
 - **ELLIPSE=PREDICTION | CONFIDENCE | NO**

14

Selected PLOTS= options:

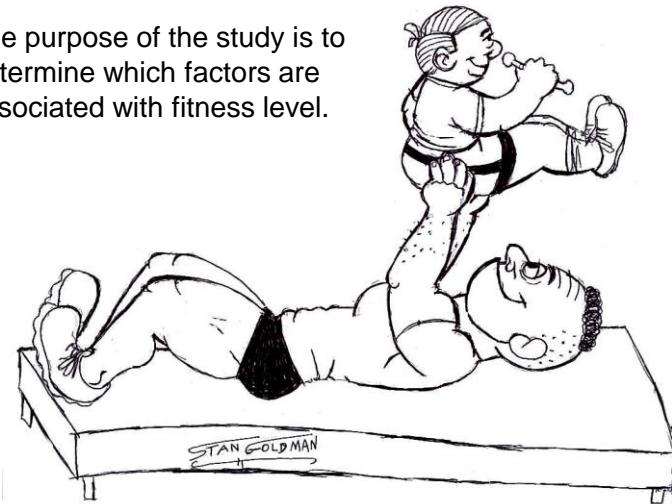
MATRIX <(matrix-options)>	requests a scatter plot matrix for variables.
SCATTER <(scatter-options)>	requests scatter plots for pairs of variables. When a scatter plot or a scatter plot matrix is requested, the Pearson correlations will also be displayed.

The available *matrix-options* are as follows:

HIST HISTOGRAM	displays histograms of variables in the VAR list in the scatter plot matrix.
NVAR=ALL n	specifies the maximum number of variables in the VAR list to be displayed in the scatter plot matrix. By default, NVAR=5.
ELLIPSE=	requests prediction ellipses for new observations (ELLIPSE=PREDICTION), confidence ellipses for the mean (ELLIPSE=CONFIDENCE), or no ellipses (ELLIPSE=NONE) to be created in the scatter plots. By default, ELLIPSE=PREDICTION.

Fitness Example

The purpose of the study is to determine which factors are associated with fitness level.



15

In exercise physiology, an objective 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 **st092.fitness** data set contains the following variables:

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

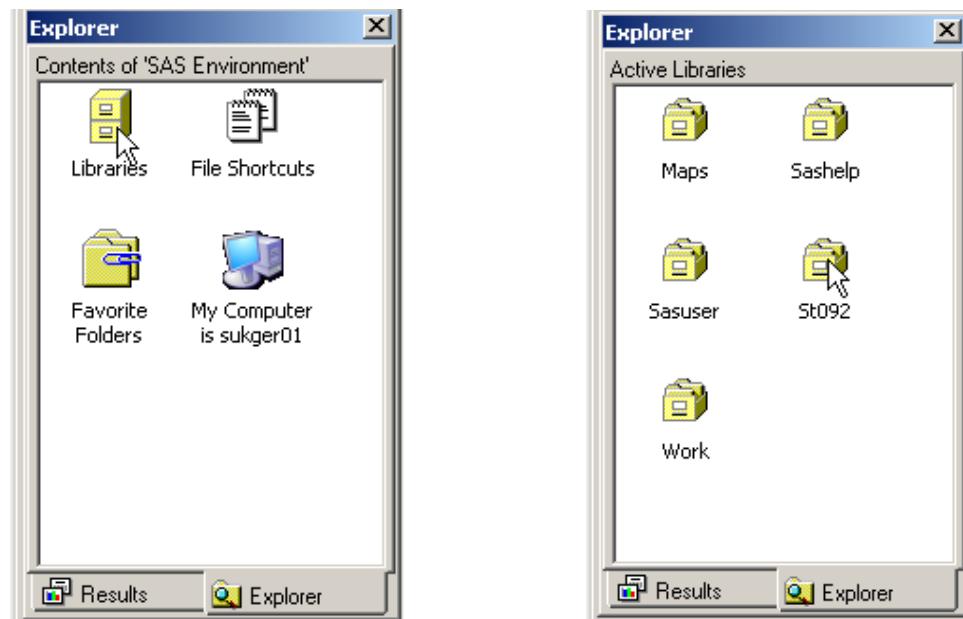


Data Exploration, Correlations, and Scatter Plots

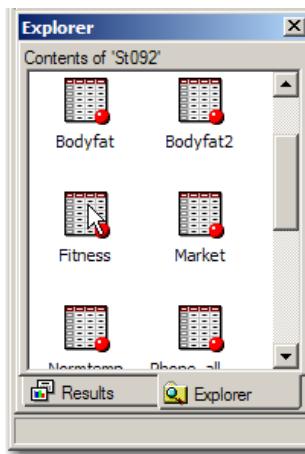
The following is a summary of what you will accomplish in this demonstration:

- View a SAS Data set and the data set properties using the Explorer window.
- Obtain scatter plots and correlation statistics to measure the strength of the linear relationship between the target variable vs predictor variables.
- Obtain scatter plots and correlation statistics to measure the strength of the linear relationship between predictor variables.

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.

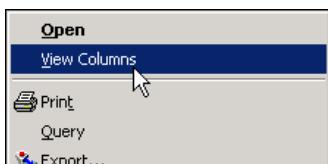


In the Explorer window, double-click on **St092**. The tables in this library are displayed.



Select **Fitness**, and right-click.

Left-click **View Columns**.



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

General Details Columns Indexes Integrity Passwords					
<input type="text" value="Find column name:"/> <input type="button" value="Find"/>					
Column Name	Type	Length	Format	Informat	Label
Name	Text	8			
Gender	Text	1			
RunTime	Number	8			
Age	Number	8			
Weight	Number	8			
Oxygen_Consum...	Number	8			
Run_Pulse	Number	8			
Rest_Pulse	Number	8			
Maximum_Pulse	Number	8			
Performance	Number	8			

Select **OK**.

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

VIEWTABLE: St092.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	90
2	Gracie	F	8.63	38	81.87	60.06	170	48	186	94
3	Luanne	F	8.65	43	85.84	54.3	156	45	168	83
4	Mimi	F	8.92	50	70.87	54.63	146	48	155	67
5	Chris	M	8.95	49	81.42	49.16	180	44	185	72
6	Allen	M	9.22	38	89.02	49.87	178	55	180	92
7	Nancy	F	9.4	49	76.32	48.67	186	56	188	64
8	Patty	F	9.63	52	76.32	45.44	164	48	166	56
9	Suzanne	F	9.93	57	59.08	50.55	148	49	155	43
10	Teresa	F	10	51	77.91	46.67	162	48	168	54
11	Bob	M	10.07	40	75.07	45.31	185	62	185	79
12	Harriett	F	10.08	49	73.37	50.39	168	67	168	57
13	Jane	F	10.13	44	73.03	50.54	168	45	168	67

- ✍ You could also look at the data using the PRINT procedure.
- ✍ You should also investigate the univariate statistics of continuous variables in the data set, just as you did in the earlier chapters, using PROC MEANS, PROC UNIVARIATE, and PROC SGLOT to explore distributions, measure central tendency and spread, and look for outliers.

Next, examine the relationships between **Oxygen_Consumption** and the continuous predictor variables in the data set using the CORR procedure. This program uses HTML output and the Statistical style, which is ideally suited for statistical presentations that use color but still print nicely in grayscale.

```
/*st005d01*/
ods graphics / imagemap=on;
ods listing close;
ods html file='corr.html'
      style=statistical;

proc corr data=st092.fitness rank
            plots(only)=scatter(nvar=all ellipse=none);
  var RunTime Age Weight Run_Pulse
    Rest_Pulse Maximum_Pulse Performance;
  with Oxygen_Consumption;
  id name;
  title "Correlations and Scatter Plots with
Oxygen_Consumption";
run;

ods html close;
ods listing;
ods graphics off;
```

Selected PROC CORR statement option:

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

 IMAGEMAP=ON in the ODS GRAPHICS statement will allow tooltips to be used in HTML output. Tooltips enable the user to identify data points by moving the mouse over observations in a plot. In PROC CORR, the variables used in the tooltips are the x-axis and y-axis variables, the observation number, and any variable in the ID statement.

ID when used in html output with imagemap on, adds the listed variables to the information available with tooltips.

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

PROC CORR Output

Correlations and Scatter Plots with Oxygen_Consumption

The CORR Procedure

1 With Variables: Oxygen_Consumption

7 Variables: RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

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	56.64516	18.32584	1756	20.00000	94.00000

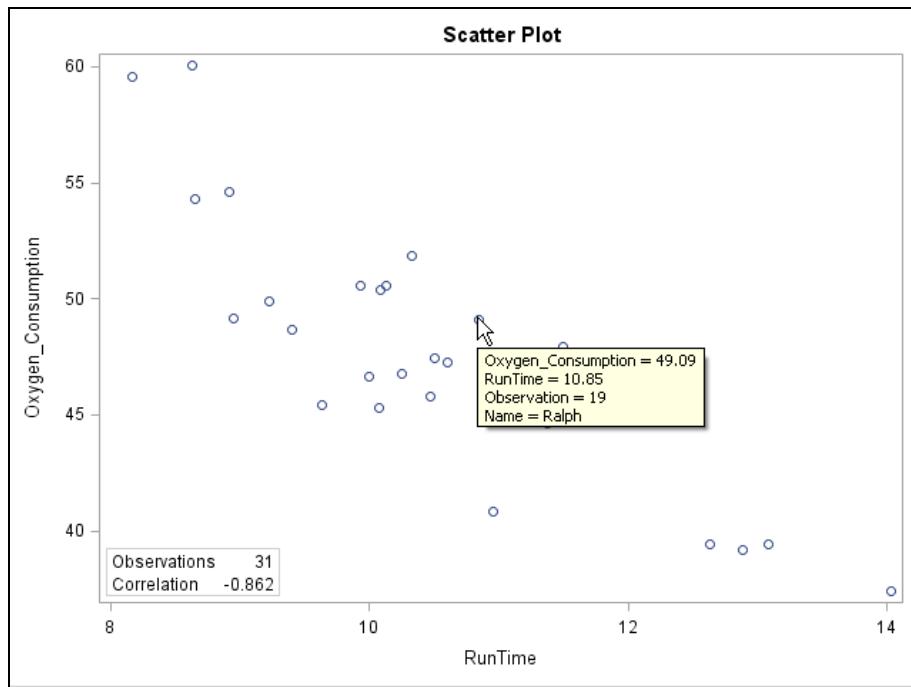
Pearson Correlation Coefficients, N = 31

Prob > |r| under H0: Rho=0

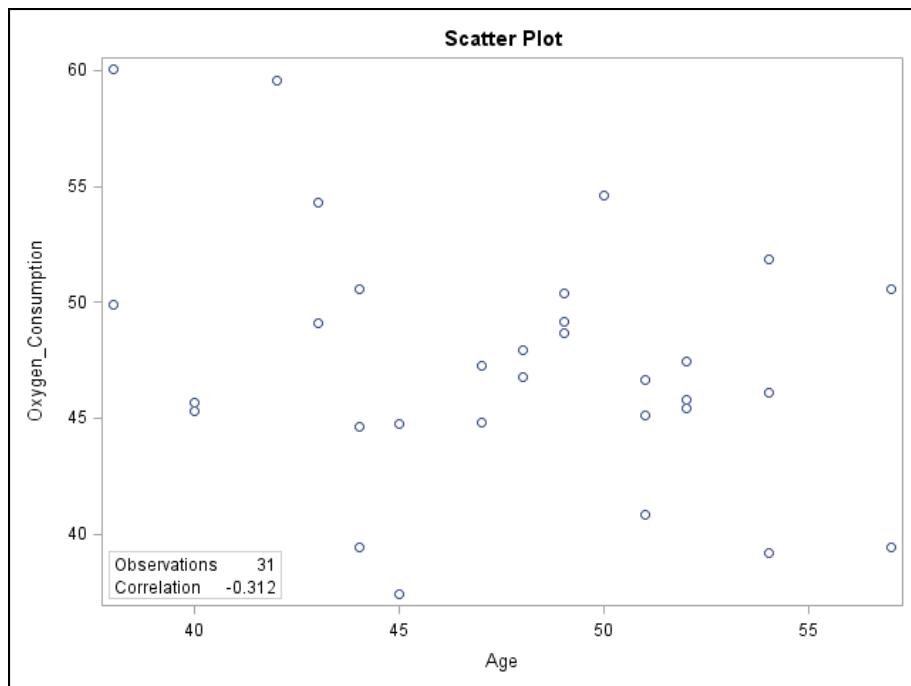
Oxygen_Consumption	RunTime	Performance	Rest_Pulse	Run_Pulse	Age	Maximum_Pulse	Weight
-0.86219	0.77890	-0.39935	-0.39808	-0.31162		-0.23677	-0.16289
<.0001	<.0001	0.0260	0.0266	0.0879		0.1997	0.3813

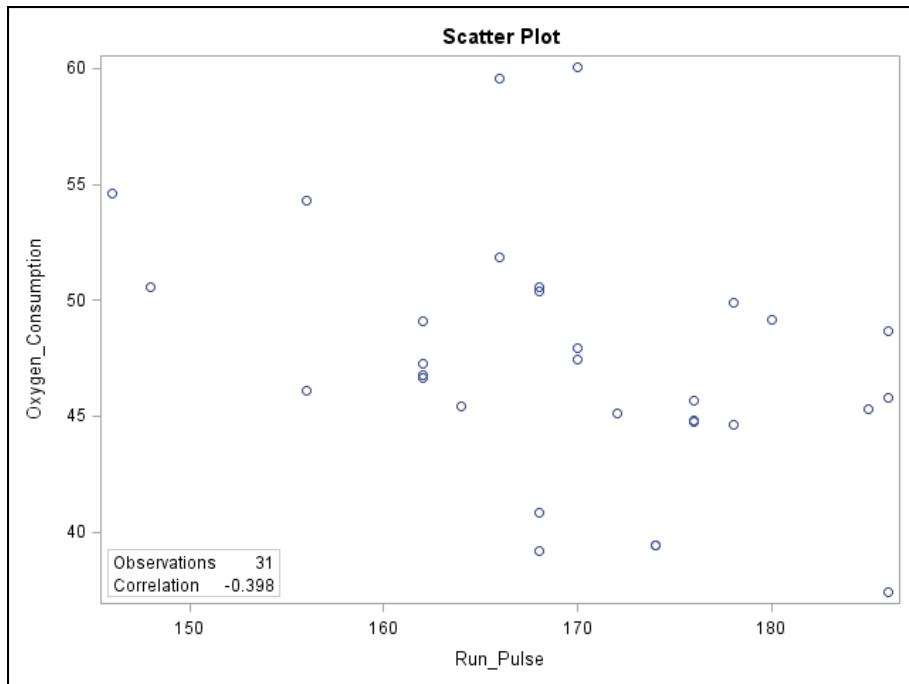
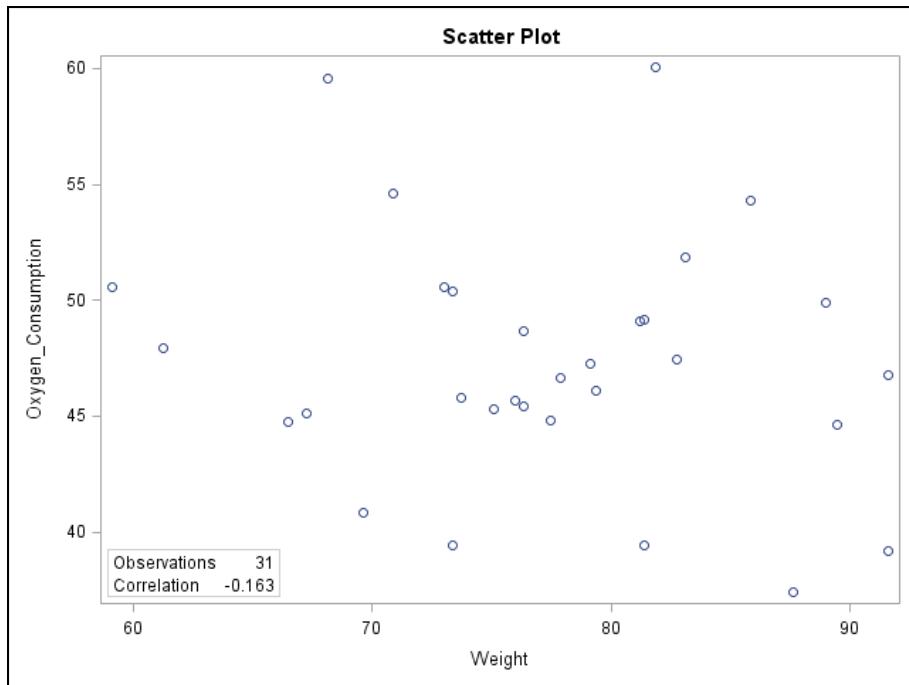
The correlation coefficient between **Oxygen_Consumption** and **RunTime** is -0.86219. The *p*-value is small, which indicates that the population correlation coefficient (Rho) is significantly different from 0. The second largest correlation coefficient, in absolute value, is **Performance**, at 0.77890.

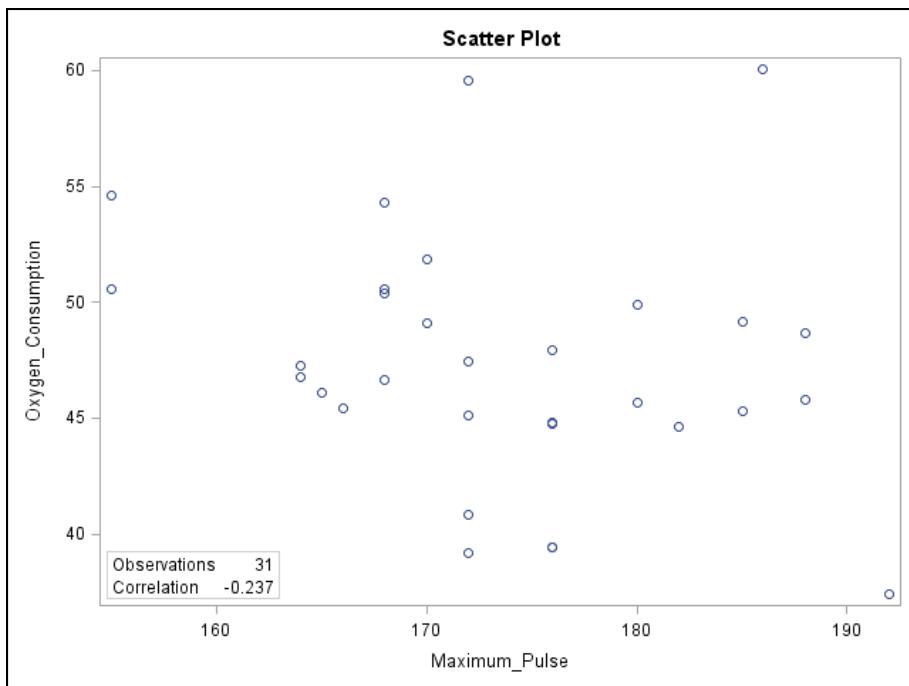
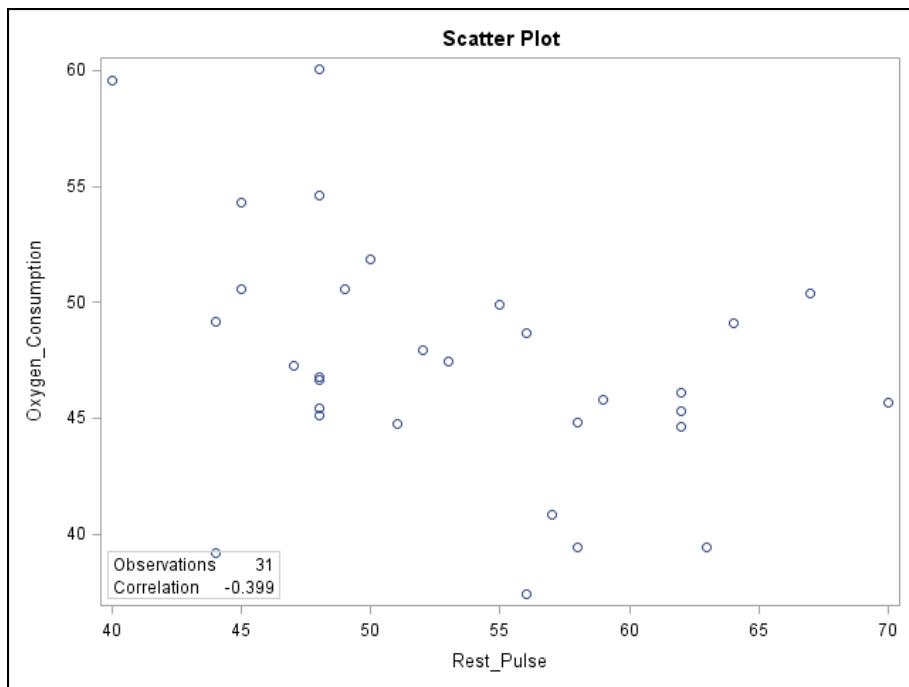
Scatter plots associated with these correlations are shown below.

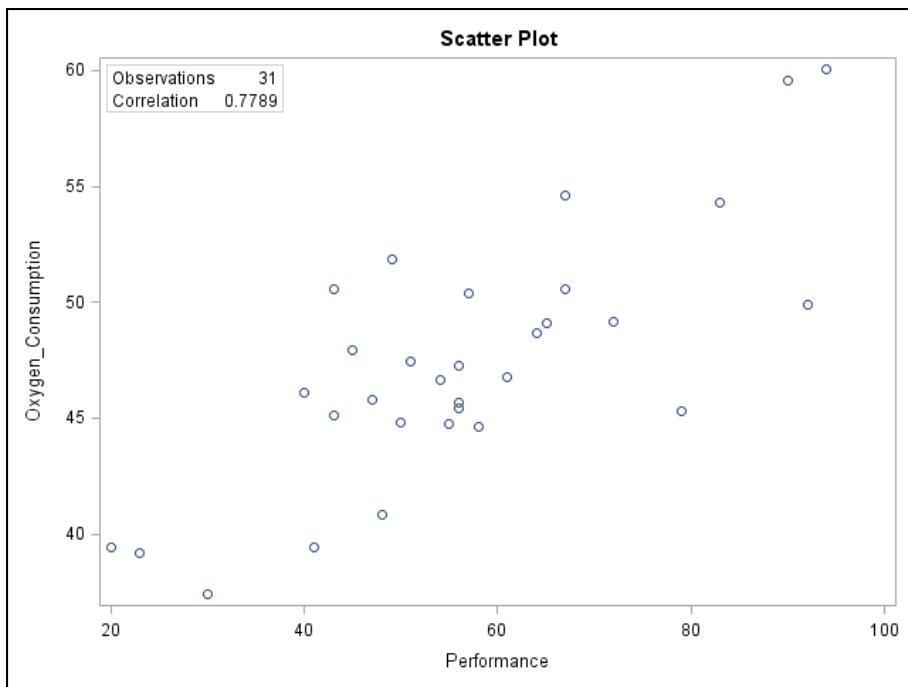


By moving the cursor over observations in the plot, the values of the x-axis, y-axis, observation number and any variables included in the ID statement, can be seen.









The correlation and scatter plot analyses indicate 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 among the potential predictor variables. PROC CORR can be used to generate a matrix of correlation coefficients. This example uses RTF output with the Journal style, which is ideally suited for high-quality grayscale printed graphics.

```
/*st005d01*/
ods graphics on;
ods listing close;
ods rtf file='corr.rtf'
    style=journal;
proc corr data=st092.fitness nosimple
    plots=matrix(nvar=all histogram);
var RunTime Age Weight Run_Pulse
    Rest_Pulse Maximum_Pulse Performance;
title "Correlations and Scatter Plot Matrix of Fitness
Predictors";
run;

ods rtf close;
ods listing;
ods graphics off;
```

Selected PROC CORR statement options:

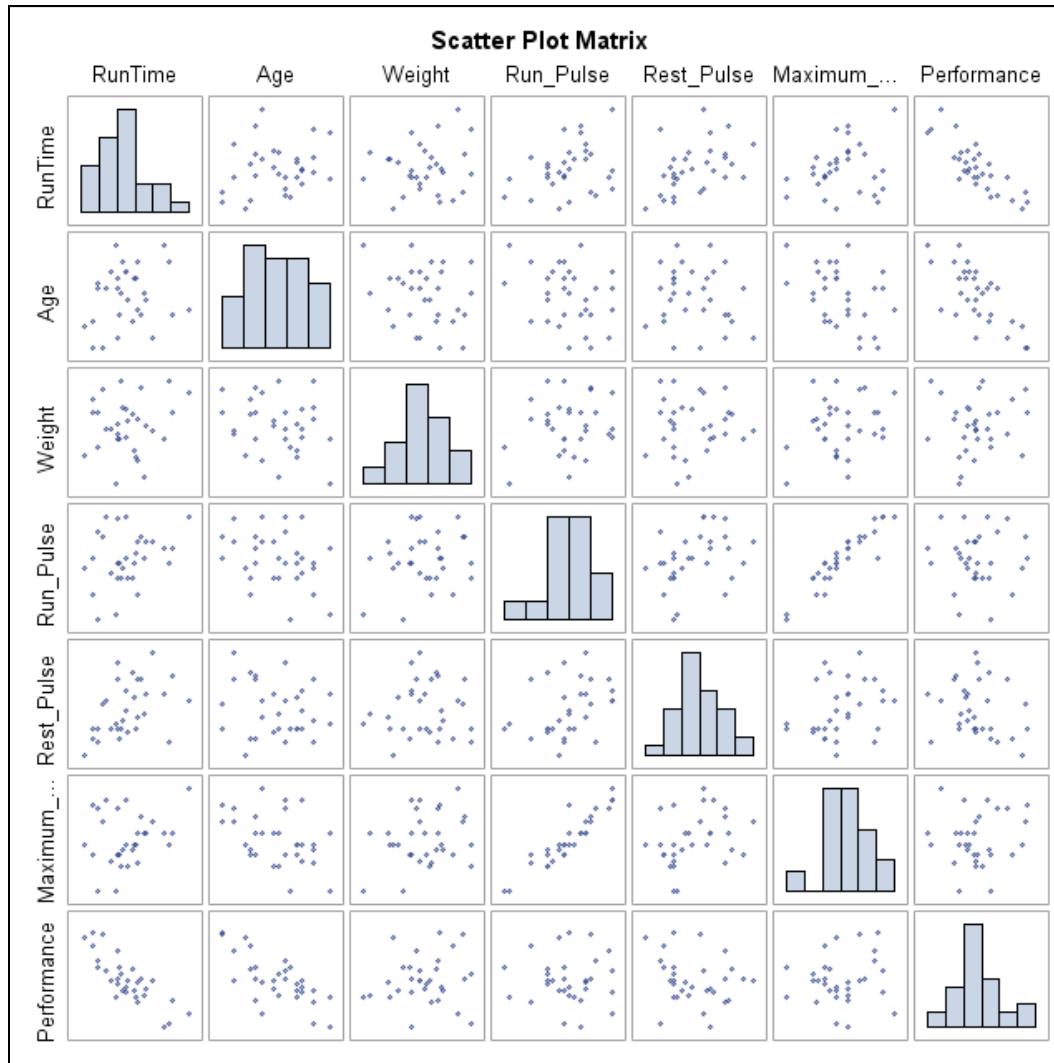
NOSIMPLE suppresses printing simple descriptive statistics for each variable.

PROC CORR Output

Correlations and Scatter Plot Matrix of Fitness Predictors***The CORR Procedure***

7 Variables: RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance							
Pearson Correlation Coefficients, N = 31 Prob > r under H0: Rho=0							
	RunTime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
<i>RunTime</i>	1.00000 0.2926	0.19523 0.4412	0.14351 0.0858	0.31365 0.0832	0.45038 0.0110	0.22610 0.2213	-0.82049 <.0001
<i>Age</i>	0.19523 0.2926	1.00000 0.1925	-0.24050 0.1925	-0.31607 0.0832	-0.15087 0.4178	-0.41490 0.0203	-0.71257 <.0001
<i>Weight</i>	0.14351 0.4412	-0.24050 0.1925	1.00000 0.3284	0.18152 0.3284	0.04397 0.8143	0.24938 0.1761	0.08974 0.6312
<i>Run_Pulse</i>	0.31365 0.0858	-0.31607 0.0832	0.18152 0.3284	1.00000 0.0518	0.35246 0.0518	0.92975 <.0001	-0.02943 0.8751
<i>Rest_Pulse</i>	0.45038 0.0110	-0.15087 0.4178	0.04397 0.8143	0.35246 0.0518	1.00000 0.0951	0.30512 0.0951	-0.22560 0.2224
<i>Maximum_Pulse</i>	0.22610 0.2213	-0.41490 0.0203	0.24938 0.1761	0.92975 <.0001	0.30512 0.0951	1.00000 0.09002	0.09002 0.6301
<i>Performance</i>	-0.82049 <.0001	-0.71257 <.0001	0.08974 0.6312	-0.02943 0.8751	-0.22560 0.2224	0.09002 0.6301	1.00000

There are strong correlations between **Run_Pulse** and **Maximum_Pulse** (0.92975) and between **Runtime** and **Performance** (-0.82049). These associations are seen in more detail in the matrix of scatter plots.



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.

<u>Row Variable</u>	<u>Column Variable</u>	<u>Pearson's r</u>	<u>Prob > r </u>
Run_Pulse	Maximum_Pulse	0.92975	<.0001
Runtime	Performance	-0.82049	<.0001
Performance	Age	-0.71257	<.0001
Runtime	Rest_Pulse	0.45038	0.0110
Age	Maximum_Pulse	-0.41490	0.0203
Run_Pulse	Rest_Pulse	0.35246	0.0518

Chapter 6 Simple Linear Regression

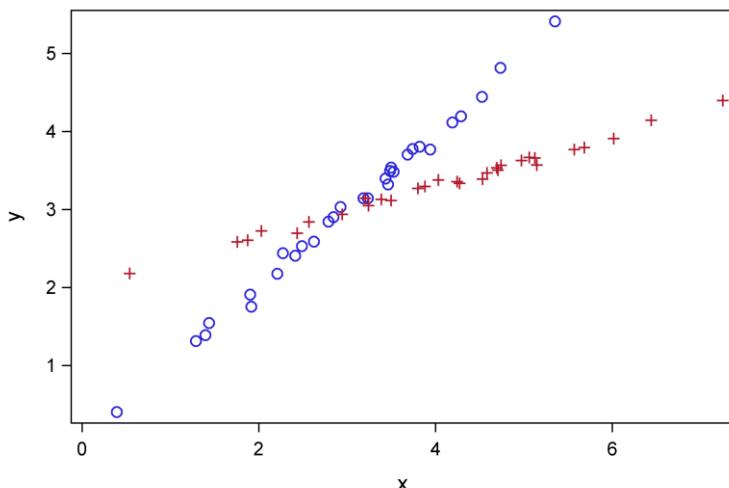
6.1 Simple Linear Regression	6-3
Demonstration: Performing Simple Linear Regression.....	6-14
Demonstration: Confidence and Predicted Limits.....	6-19
Demonstration: Producing Predicted Values	6-22
6.2 Examining Assumptions	6-25
Demonstration: Residual Plots.....	6-33

6.1 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.

Overview



4

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, but very different linear relationships. 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.

5

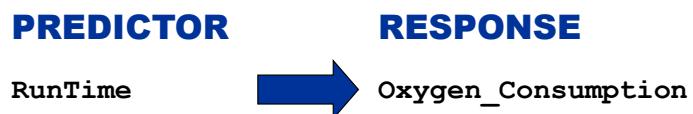
In simple linear regression, the values of the predictor variable are assumed fixed. Thus, you try to explain the variability of the response variable given the values of the predictor variable.

Terminology

Response	Predictor
Dependent	Independent
Analysis	Explanatory
Target	Input

6

Fitness Example

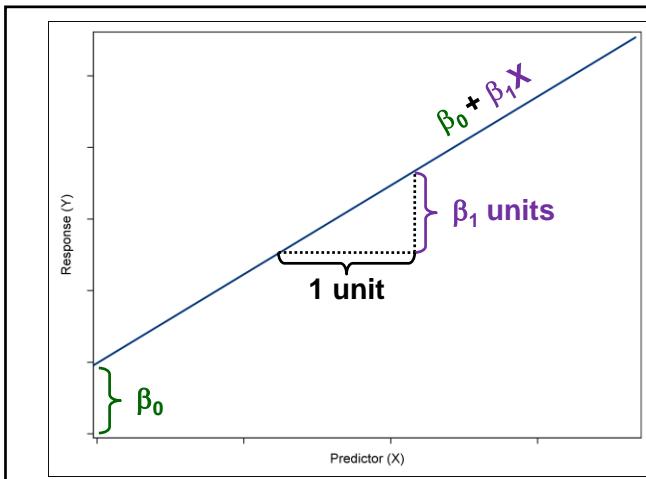


7

The analyst noted that the running time 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 **RunTime**.

She decides to run a simple linear regression of **Oxygen_Consumption** versus **RunTime**.

Simple Linear Regression Model



8

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

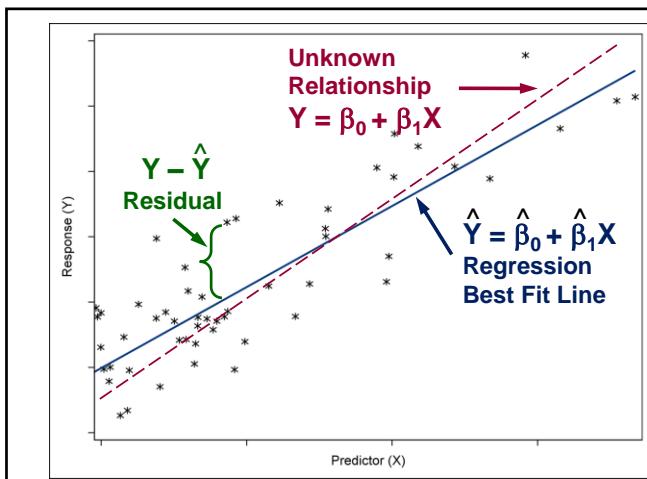
X predictor variable

β_0 intercept parameter, which corresponds to the value of the response variable when the predictor is 0

β_1 slope parameter, which corresponds to the magnitude of change in the response variable given a one unit change in the predictor variable

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

Simple Linear Regression Model



9

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 (β_0, β_1) that define the assumed relationship between your response and predictor variables.

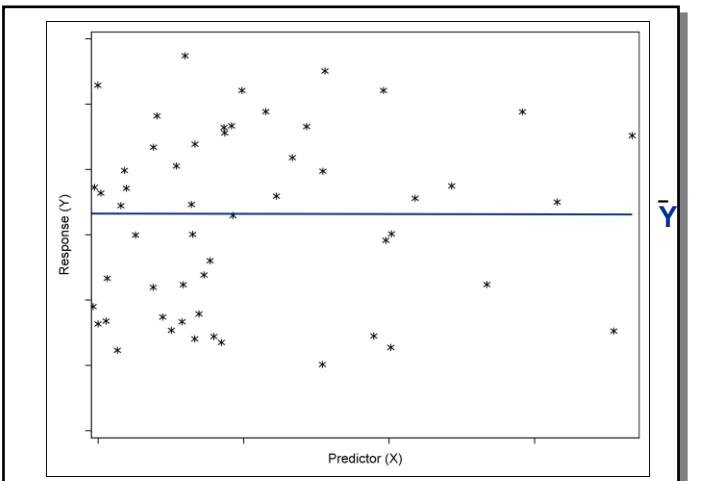
Estimates of the unknown population parameters β_0 and β_1 are obtained by the *method of least squares*. This method provides the estimates by determining the line that minimizes the sum of the squared vertical distances between the observations and the fitted line. In other words, the fitted or regression 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 (efficiency). 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.

The Baseline Model



10

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. Therefore, knowing the value of the predictor variable does not improve predictions of the response over simply using the mean of the response variable for everyone.

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:

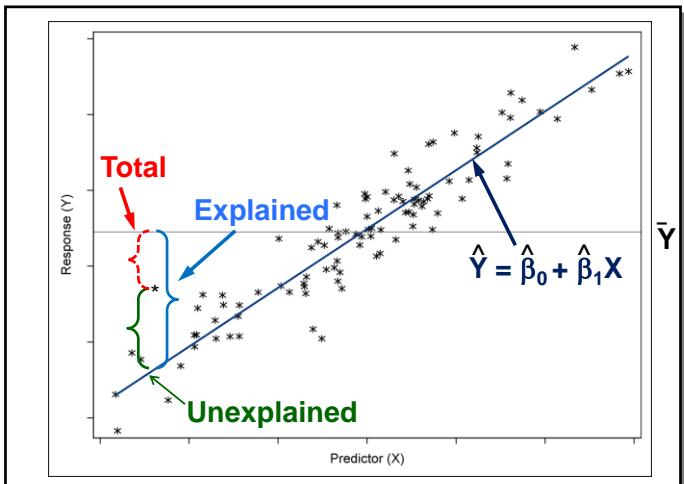
- The simple linear regression model does fit the data better than the baseline model.
- $\beta_1 \neq 0$

11

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.

Explained versus Unexplained Variability



12

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

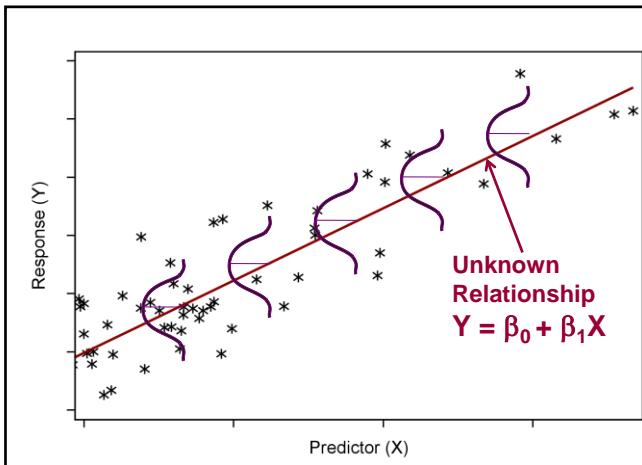
Explained variability is related to the difference between the regression line and the mean of the response variable. The model sum of squares (SSM) is the amount of variability explained by your model. The model sum of squares is equal to $\sum(\hat{Y}_i - \bar{Y})^2$

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 $\sum(Y_i - \bar{Y})^2$.

 The plot shows a seemingly contradictory relationship between explained, unexplained and total variability. Contribution to total variability for the data point is smaller than contribution to explained and unexplained variability. Remember that the relationship of total=unexplained + explained holds for sums of squares over all observations and not necessarily for any individual observation.

Assumptions of Simple Linear Regression



13

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.



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;
QUIT;
```

14

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.
- ☞ When ODS Graphics are turned on, default graphics are produced.



Performing Simple Linear Regression

The following is a summary of what you will accomplish in this demonstration:

- Perform and analyse a Simple Linear Regression.

As there is an apparent linear relationship between **Oxygen_Consumption** and **RunTime**, perform a simple linear regression analysis with **Oxygen_Consumption** as the response variable.

```
/*st006d01*/
ods graphics on;
proc reg data=st092.fitness;
  model Oxygen_Consumption = RunTime;
  title 'Predicting Oxygen_Consumption from RunTime';
run;
quit;
ods graphics off;
```

PROC REG Output

```
Predicting Oxygen_Consumption from RunTime

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen_Consumption

Number of Observations Read      31
Number of Observations Used     31
```

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 **RunTime**.

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	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			

① 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 (Between Group).
Error	is the variability unexplained by your model (Within Group).
Corrected Total	is the total variability in the data (Total).
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 <i>p</i> -value associated with the <i>F</i> value.

The *F* value tests whether the slope of the predictor variable is equal to 0.

Step 1- Set Hypothesis

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0.$$

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$p\text{-value} < 0.0001.$$

Step 4- Decision Rule.

The *p*-value $< \alpha$, Reject H_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, **RunTime** explains a significant amount of variability of **Oxygen_Consumption**.

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

Root MSE	2.74515	R-Square	0.7434	②
Dependent Mean	47.37581	Adj R-Sq	0.7345	
Coeff Var	5.79442			

- ② R-Square** the coefficient of determination also 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 0.7434, which means that the regression line explains 74% of the total variation in the response values.
 - the square of the multiple correlation between y and the x 's.



Notice that the R-square is the squared value of the correlation you saw earlier between **RunTime** and **Oxygen_Consumption** (0.86219). This is no coincidence. For simple regression, the R-square value will be the square of the value 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
Adj R-Sq	the adjusted R^2 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	Standard		
		Estimate	Error	t Value	Pr > t
Intercept	1	82.42494	3.85582	21.38	<.0001
RunTime	1	-3.31085	0.36124	-9.17	<.0001

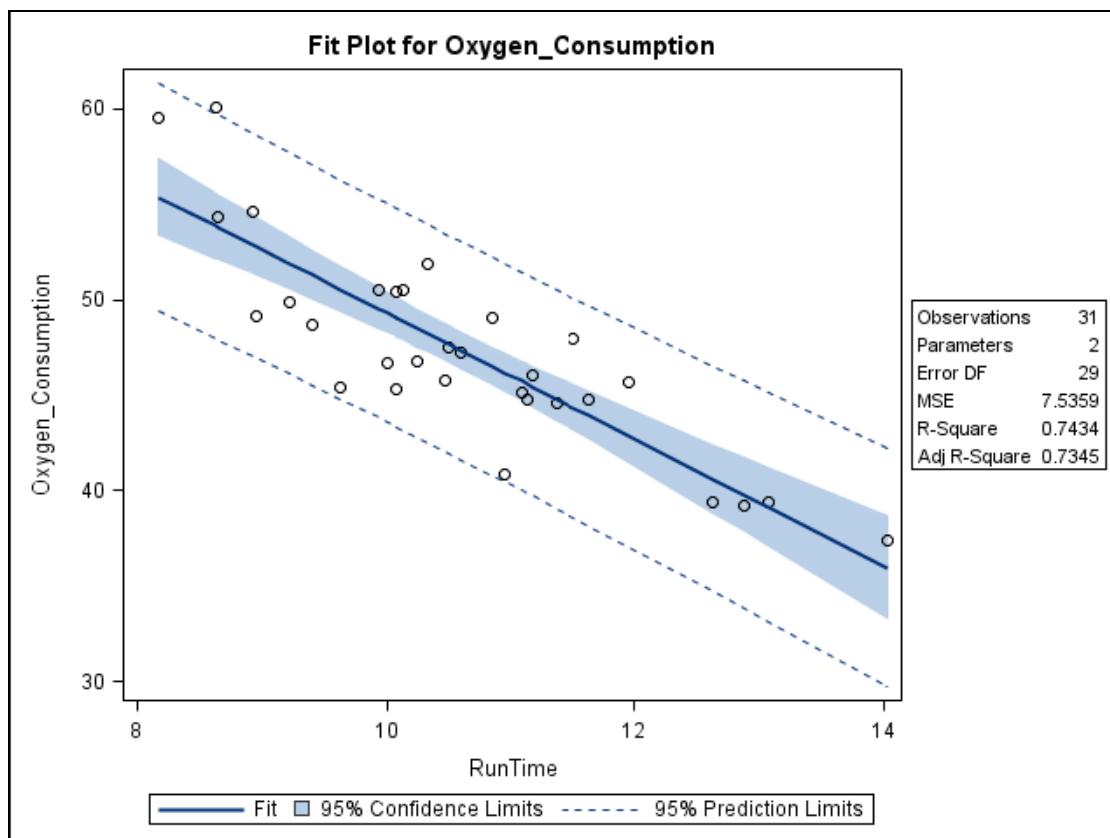
③ DF represents the degrees of freedom associated with each term in the model.

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 parameter estimates by their corresponding standard error estimates.
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.

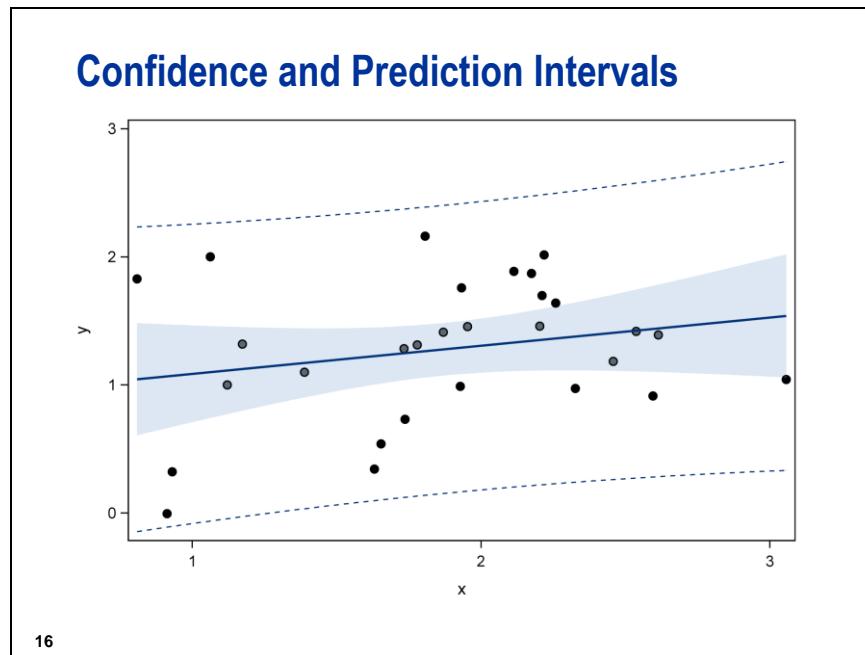
Because the estimate of $\beta_0=82.42494$ and $\beta_1=-3.31085$, the estimated regression equation is given by Predicted Oxygen_Consumption = 82.42494 - 3.31085 *(RunTime).

The model indicates that an increase of one unit for **Runtime** amounts to a 3.31085 decrease in **Oxygen_Consumption**. However, this equation is appropriate only in the range of values you observed for the variable **RunTime**.

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 not have practical significance because **RunTime**=0 (running at the speed of light) is not inside the range of observed values.



The Fit Plot produced by ODS Graphics shows the predicted regression line superimposed over a scatter plot of the data. You will learn more about this plot next.



To assess the level of precision around the mean estimates of **Oxygen_Consumption**, you can produce **confidence intervals around the means**. This is represented in the shaded area in the plot.

- 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 mean of X.

Suppose that the mean **Oxygen_Consumption** at a fixed value of **RunTime** is not the focus. If you are interested in establishing an inference on a future single observation, you need a **prediction interval around the individual observations**. This is represented by the area between the broken lines in the plot.

- 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. Regression Lines with Confidence Intervals

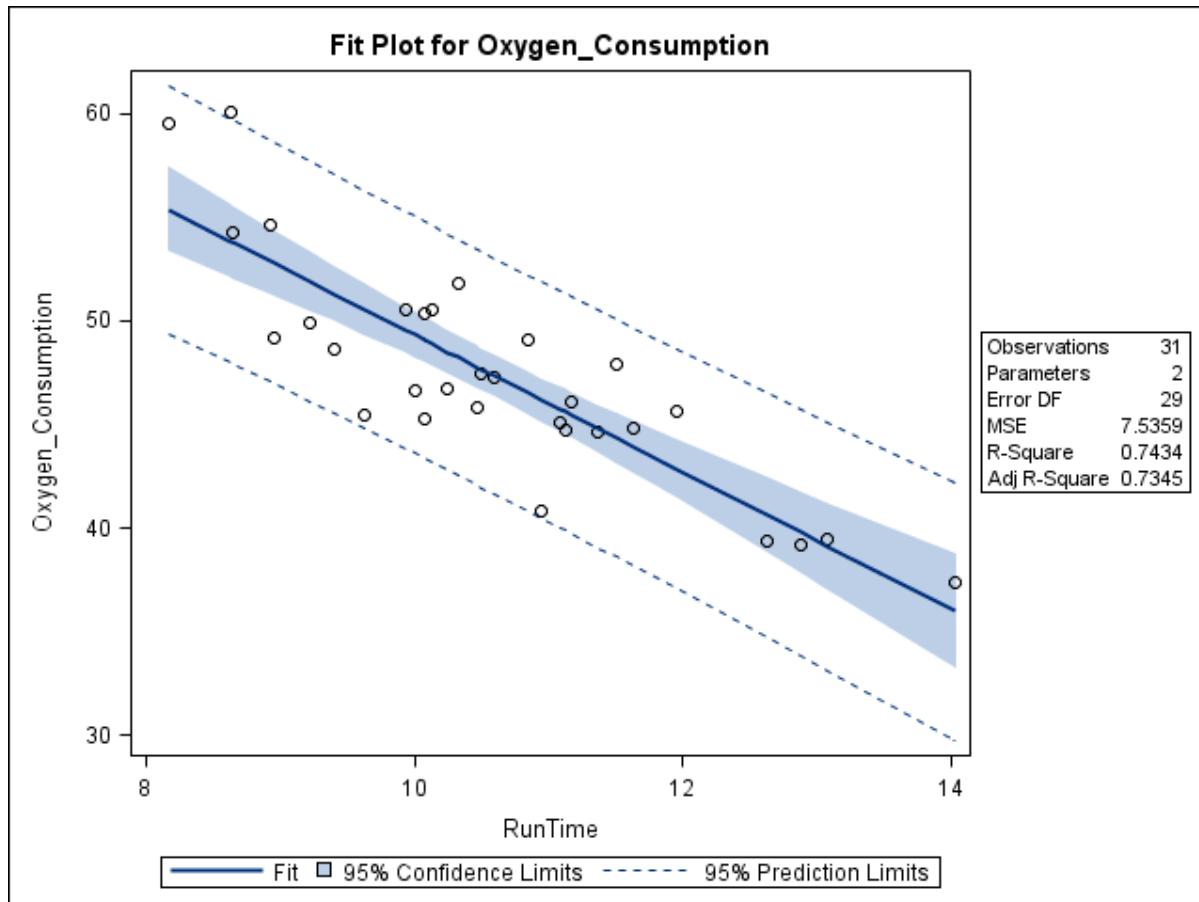


Confidence and Predicted Limits

The following is a summary of what you will accomplish in this demonstration:

- Investigate the graph containing Confidence and Prediction Limits.

Return to the output from the last demonstration and open the Fit Plot.



The Confidence Interval for the mean is represented by the shaded region. The Prediction Interval for observations is the area between the dotted lines. Model statistics are reported in the inset by default.

Printed tables for the confidence and prediction intervals at each observed data point can be obtained by adding the options CLM and CLI to the MODEL statement:

```
/*st006d02*/
ods graphics off;
proc reg data=st092.fitness;
  model Oxygen_Consumption = RunTime / clm cli;
  title 'Predicting Oxygen_Consumption from RunTime';
run;
quit;
```

Partial Output

Predicting Oxygen_Consumption from RunTime									
The REG Procedure									
Model: MODEL1									
Dependent Variable: Oxygen_Consumption									
Output Statistics									
Obs	Dependent Variable	Predicted Value	Std Error Mean	95% CL Mean	95% CL Predict	Residual			
1	59.5700	55.3753	1.0024	53.3250	57.4255	49.3982	61.3524	4.1947	
2	60.0600	53.8523	0.8616	52.0900	55.6145	47.9677	59.7368	6.2077	
3	54.3000	53.7860	0.8557	52.0359	55.5362	47.9051	59.6670	0.5140	
4	54.6300	52.8921	0.7780	51.3008	54.4834	47.0565	58.7277	1.7379	
5	49.1600	52.7928	0.7697	51.2186	54.3670	46.9618	58.6238	-3.6328	
6	49.8700	51.8989	0.6976	50.4721	53.3256	46.1059	57.6918	-2.0289	
7	48.6700	51.3029	0.6532	49.9669	52.6389	45.5317	57.0741	-2.6329	
8	45.4400	50.5414	0.6020	49.3102	51.7726	44.7935	56.2893	-5.1014	

The columns labeled 95% CL Mean are the lower and upper confidence limits for the mean, respectively.
The columns labeled 95% CL Predict are the lower and upper prediction limits.

Producing Predicted Values

What is `Oxygen_Consumption` when `RunTime` is **9, 10, 11, 12, or 13** minutes?

18

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

The following is a summary of what you will accomplish in this demonstration:

- Use PROC REG to produce predicted values.

Example: Produce predicted values of **Oxygen_Consumption** when **RunTime** is 9, 10, 11, 12, or 13.

```
/*st006d03*/
data Need_Predictions;
  input RunTime @@;
  datalines;
9 10 11 12 13
;
run;

data PredOxy;
  set Need_Predictions
    st092.fitness;
run;

ods graphics off;
proc reg data=PredOxy;
  model Oxygen_Consumption=RunTime / p;
  id RunTime;
  title 'Oxygen_Consumption=RunTime 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.

PROC REG Output

Oxygen_Consumption=RunTime 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.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			
Root MSE		2.74515	R-Square	0.7434	
Dependent Mean		47.37581	Adj R-Sq	0.7345	
Coeff Var		5.79442			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	82.42494	3.85582	21.38	<.0001
RunTime	1	-3.31085	0.36124	-9.17	<.0001

The model output is not affected by the extra 5 observations, because they were not used in any calculations, due to missing values.

Partial Output

Oxygen_Consumption=RunTime with Predicted Values					
The REG Procedure					
Model: MODEL1					
Dependent Variable: Oxygen_Consumption					
Output Statistics					
Obs	Run Time	Dependent Variable	Predicted Value	Residual	
1	9.00	.	52.6272	.	
2	10.00	.	49.3164	.	

3	11.00	.	46.0055	.
4	12.00	.	42.6947	.
5	13.00	.	39.3838	.
6	8.17	59.5700	55.3753	4.1947
7	8.63	60.0600	53.8523	6.2077
.
35	13.08	39.4400	39.1190	0.3210
36	14.03	37.3900	35.9736	1.4164

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

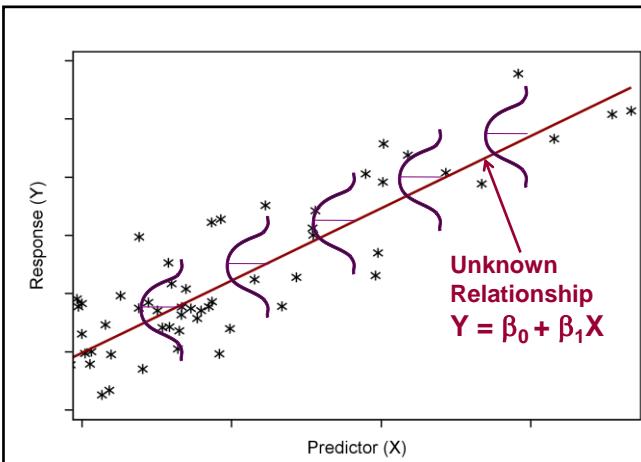
The output shows that the estimated value of **Oxygen_Consumption** is 52.6272 when **RunTime** equals 9. When **RunTime** is 13, the predicted **Oxygen_Consumption** value is 39.3838.

6.2 Examining Assumptions

Objectives

- Review the assumptions of linear regression.
- Examine the assumptions with scatter plots and residual plots.

Assumptions for Regression

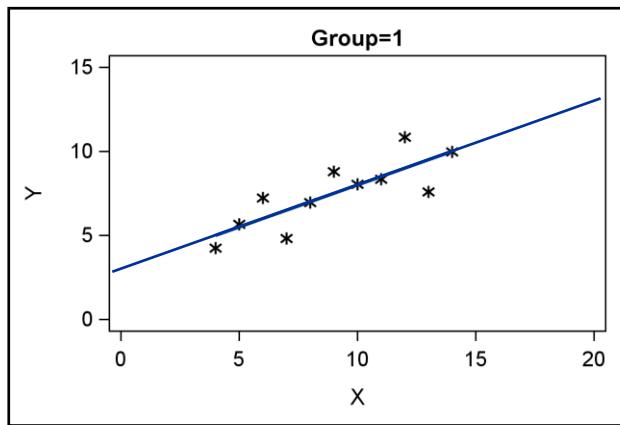


22

Recall that the model for the linear regression has the form $Y = \beta_0 + \beta_1 X + \varepsilon$. 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.

Scatter Plot of Correct Model



$$Y = 3.0 + 0.5X$$

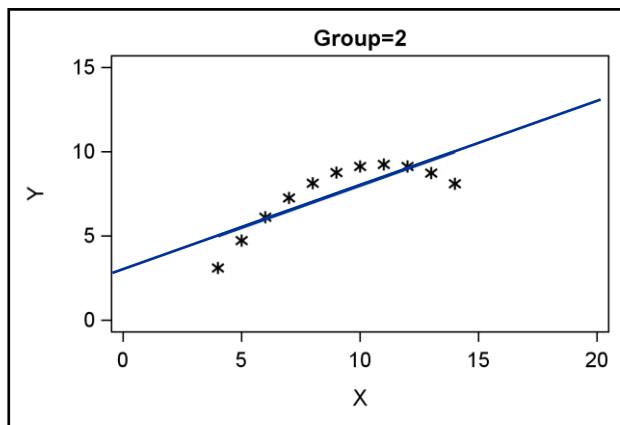
$$R^2 = 0.67$$

23

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^2 statistic are the same.

In the first plot, a regression line adequately describes the data.

Scatter Plot of Curvilinear Model



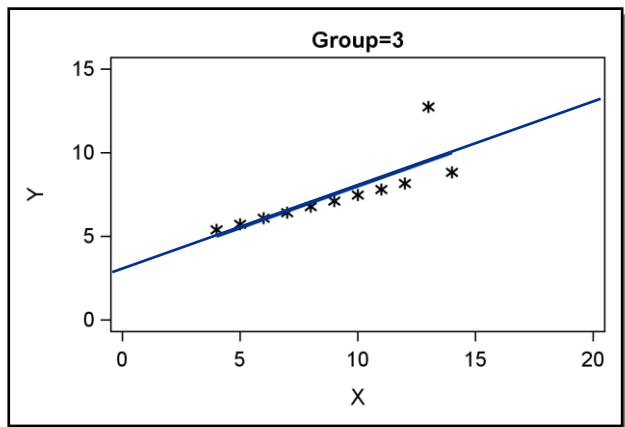
$$Y = 3.0 + 0.5X$$

$$R^2 = 0.67$$

24

In the second plot, a simple linear regression model is not appropriate because you are fitting a straight line through a curvilinear relationship.

Scatter Plot of Outlier Model



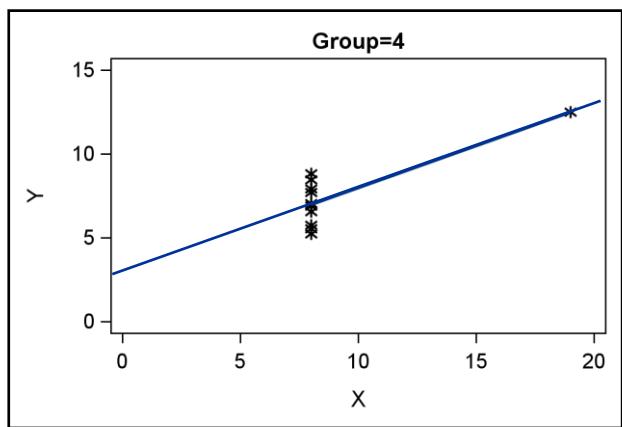
$$Y = 3.0 + 0.5X$$

$$R^2 = 0.67$$

25

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.

Scatter Plot of Influential Model



$$Y = 3.0 + 0.5X$$

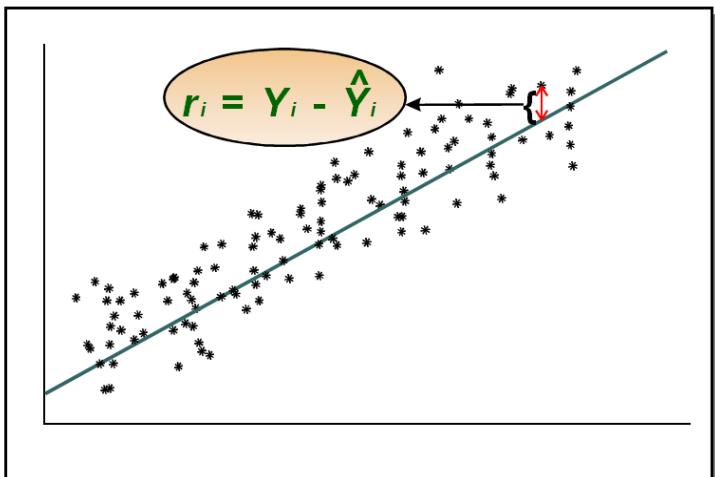
$$R^2 = 0.67$$

26

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.

Verifying Assumptions



27

To verify the assumptions for regression, you can use the residual values from the regression analysis. Residuals are defined as

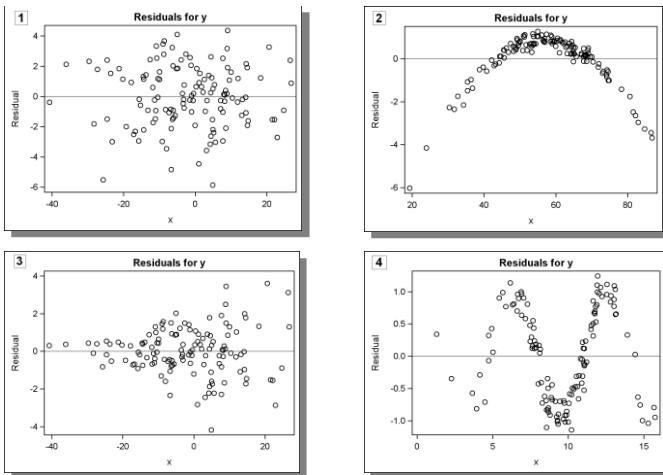
$$r_i = Y_i - \hat{Y}_i$$

where \hat{Y}_i is the predicted value for the i^{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

Examining Residual Plots

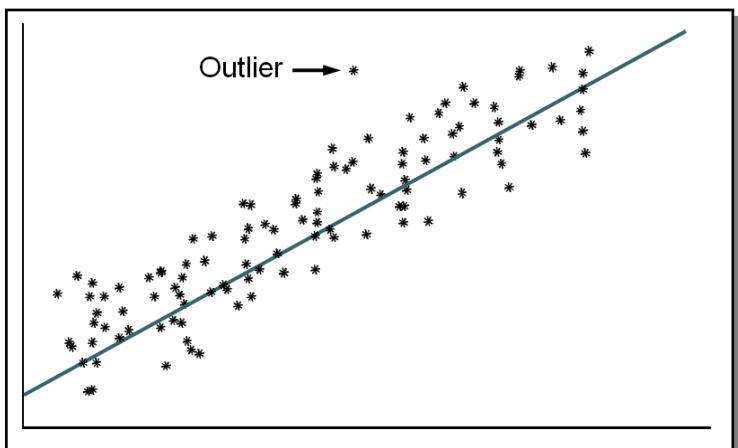


28

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



29

It is also important to check for outliers, which are observations that are far away from the bulk of the data. These observations are often data errors or reflect unusual circumstances, and it is good practice to detect these points and try to 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

30



Residual Plots

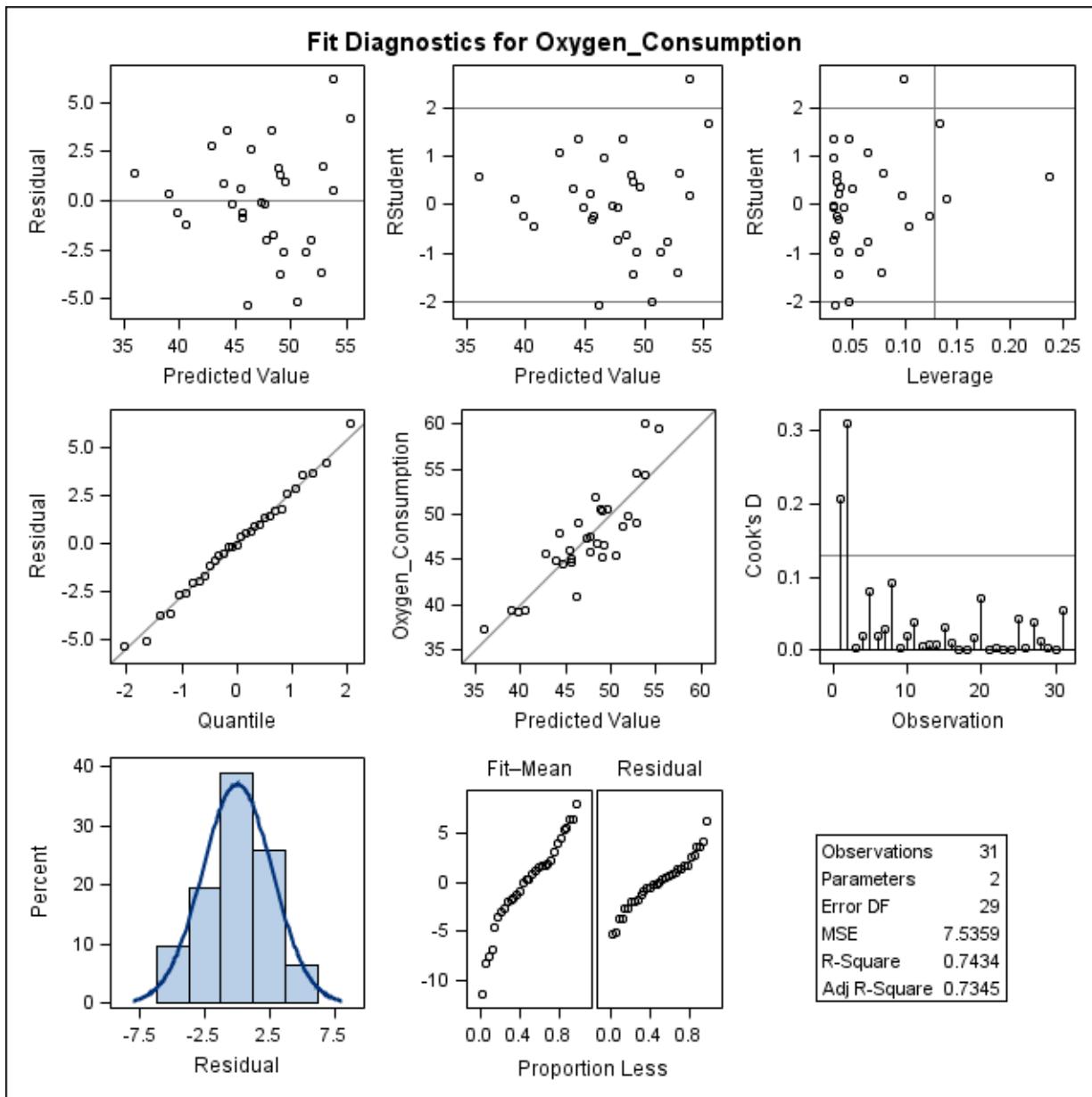
The following is a summary of what you will accomplish in this demonstration:

- Produce Residual Plots to verify the assumption of Simple Linear Regression.

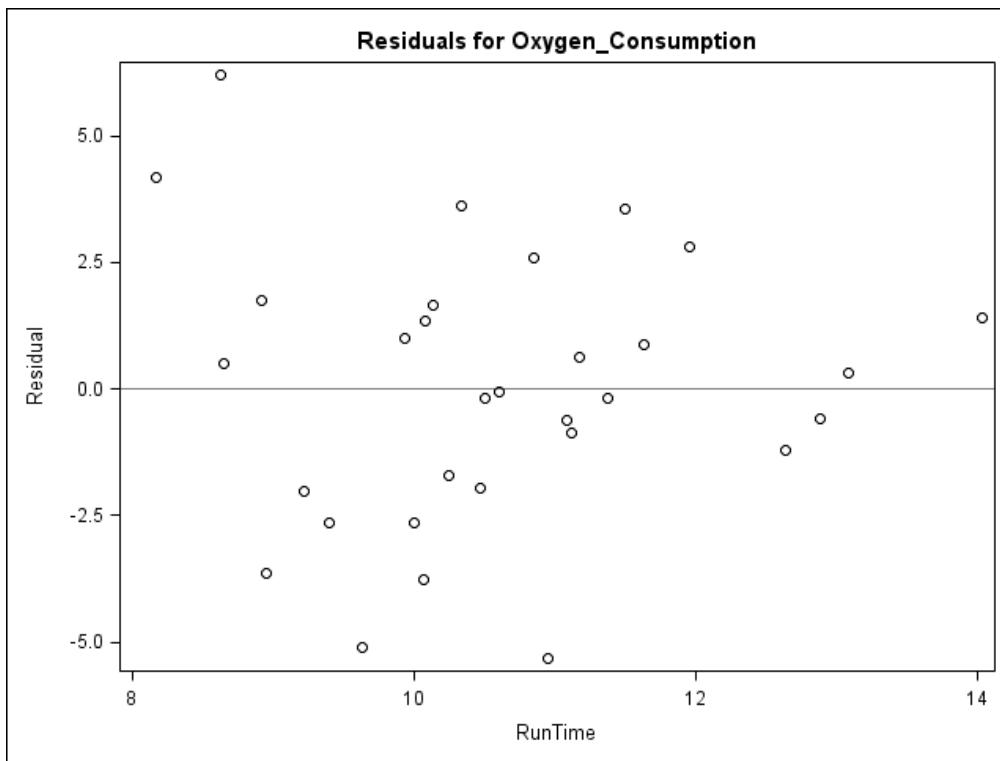
```
/*st006d04*/
ods graphics on;
ods select diagnosticsPanel ResidualPlot;
proc reg data=st092.fitness;
model Oxygen_Consumption
    = runtime;
title 'Plots of Diagnostic Statistics for runtime';
run;
quit;
ods graphics off;
```

ODS SELECT specifies that only the DIAGNOSTICS PANEL and the RESIDUAL PLOT be printed to the output destination.

The graphs are shown below.



Several residual plots and diagnostic plots are produced in the DIAGNOSTICS panel plot. The plots to verify assumptions are ; the histogram ; the first scatter plot of residuals vs Predicted Values and the QQPlot (quantile vs residuals).



The plot of the residuals versus the values of the independent variable, **Runtim**, is shown above. There are no obvious trends or patterns in the residuals. Recall that independence of residual errors (no trends) is an assumption for linear regression, as is constant variance across all levels of all predictor variables (and across all levels of the predicted values, which is seen below).

If you want to view the DIAGNOSTICS panel plots separately, specify PLOTS=DIAGNOSTICS(UNPACK) in the PROC REG statement. You could also specify each plot separately individually by name. Individual plots are produced full-sized.

```
/*st006d04*/
ods graphics on;
proc reg data=st092.fitness
plots(only)=(QQ
             RESIDUALBYPREDICTED
             RESIDUALHISTOGRAM
             RESIDUALPLOT
             RSTUDENTBYPREDICTED);
model Oxygen_Consumption
      = runtime;
title 'Plots of Diagnostic Statistics for runtime';
run;
quit;
ods graphics off;
```

Selected REG procedure PLOTS= options:

PLOTS(ONLY)= produces only the plots listed and suppresses printing of default plots.

QQ residual Quantile-Quantile plot to assess the normality of the residual error.

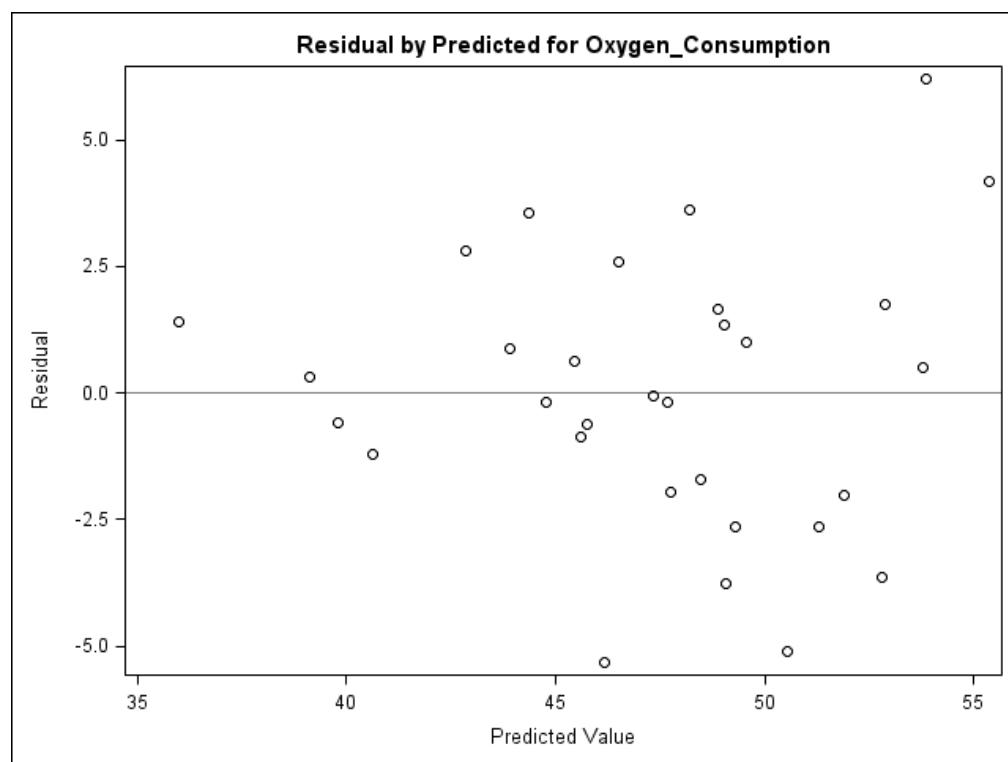
RESIDUALBYPREDICTED residuals by predicted values

RESIDUALHISTOGRAM histogram of residuals

RSTUDENTBYPREDICTED Studentised by predicted values

-  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 summary of the student residuals in tabular rather than graphic form. The R option is used in the next section.

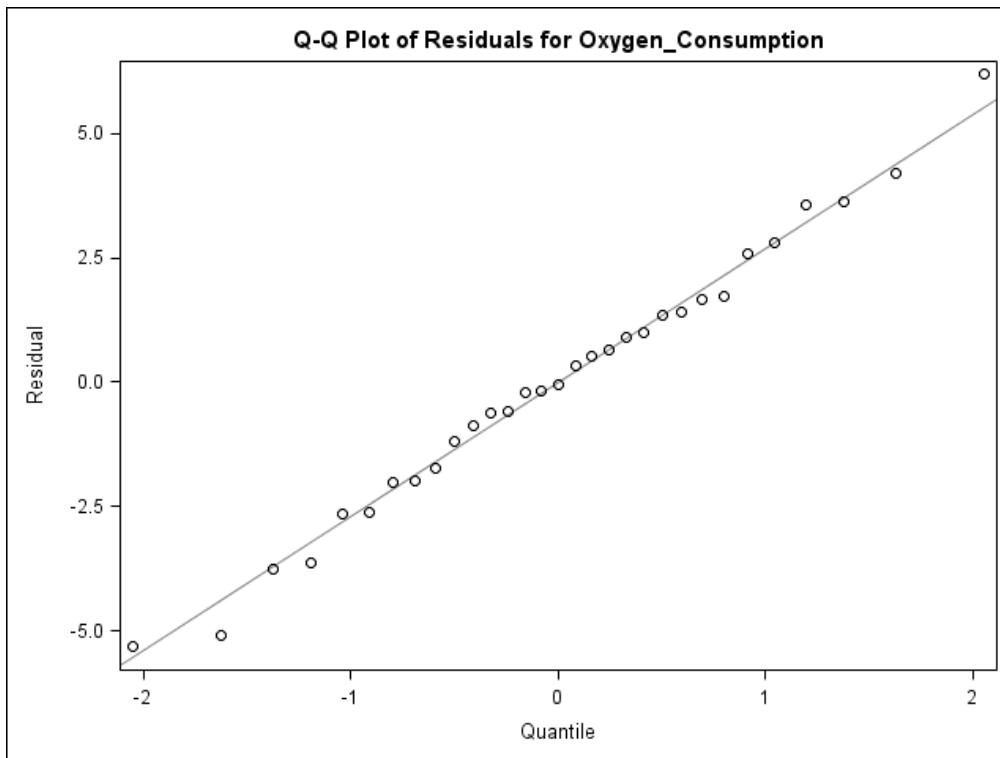
The plots of the residuals by predicted values of **Oxygen_Consumption** and runtime is shown below. The residual values appear to be randomly scattered about the reference line at 0. There are no apparent trends or patterns in the residuals.



The plot of the residuals against the normal quantiles is shown below. If the residuals are normally distributed, the plot should appear to be a straight, diagonal line. If the plot deviates substantially from the reference line, then there is evidence against normality.

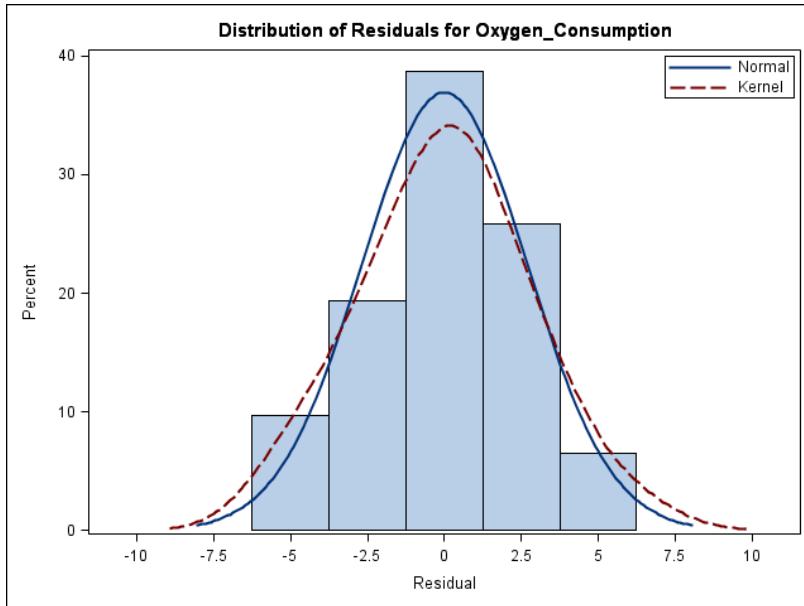
The plot below shows little deviation from the expected pattern. Thus, you can conclude that the residuals do not significantly violate the normality assumption. If the residuals did violate the normality assumption, then a transformation of the response variable or a different model might be warranted.

PROC REG Output (Continued)

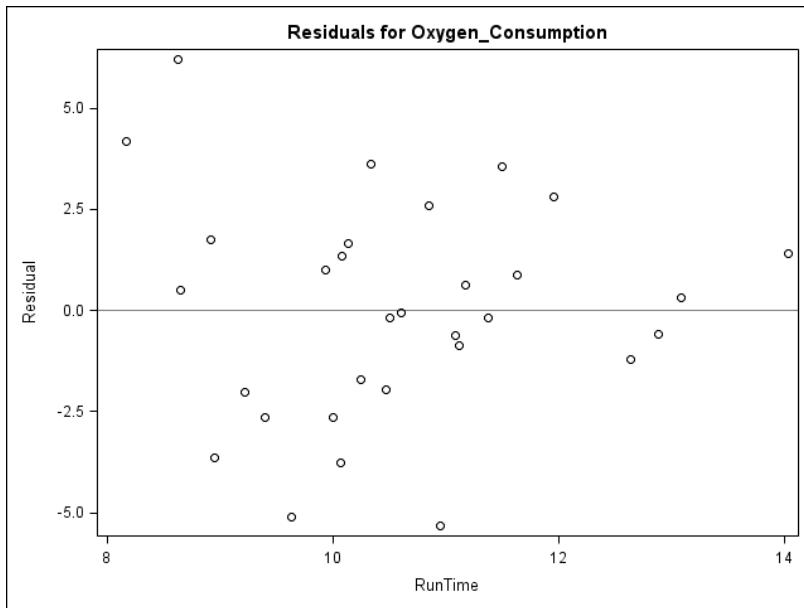


You can use the NORMAL option in the UNIVARIATE procedure to generate a hypothesis test on whether the residuals are normally distributed. This could be necessary if you feel the plot above shows a violation of the normality assumption. First you must create an output data set with the residuals in PROC 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.

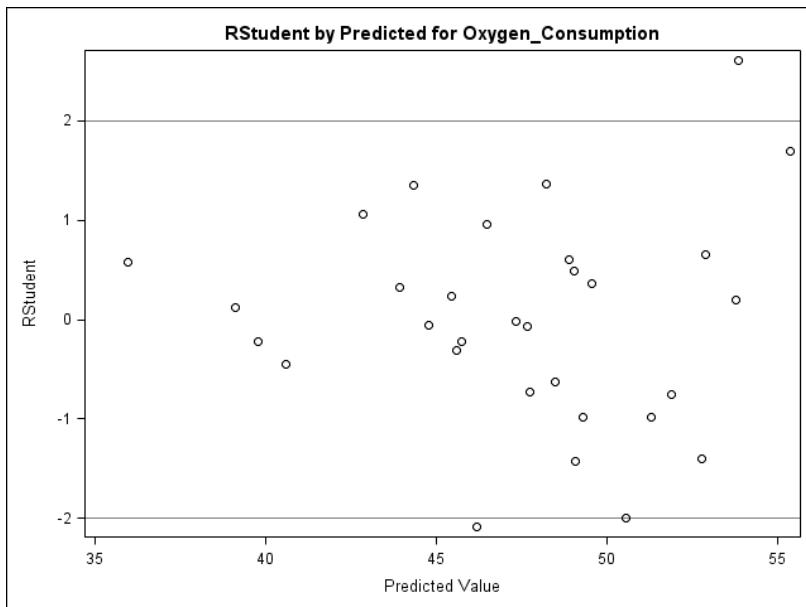
The histogram of the residuals is shown below. If the residuals are normally distributed, the histogram will follow a normal distribution. This graph shows that there is no deviation from the normal distribution therefore our normality assumption holds.



The plot for residuals vs runtime shows a random scatter around 0. Therefore no obvious patterns or trends were missed when viewing the initial scatter plot.



No clear patterns are indicated.



You may want to investigate the two observations outside the lines indicated but remember this is a suggested guideline and these points would be expected.

Chapter 7 Multiple Linear Regression

7.1 Concepts of Multiple Regression	7-3
Demonstration: Fitting a Linear Regression Model with Two Predictor Variables	7-9
Demonstration: Fitting a Multiple Linear Regression Model	7-16
7.2 Model Building and Interpretation	7-18
Demonstration: Stepwise Regression.....	7-24
Demonstration: Examining Assumptions	7-35

7.1 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.

Multiple Linear Regression with Two Variables

Consider 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.

ε is the error term.

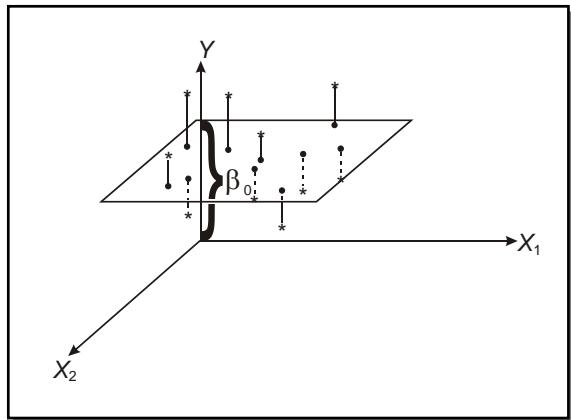
β_0 , β_1 , and β_2 are unknown parameters.

4

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).

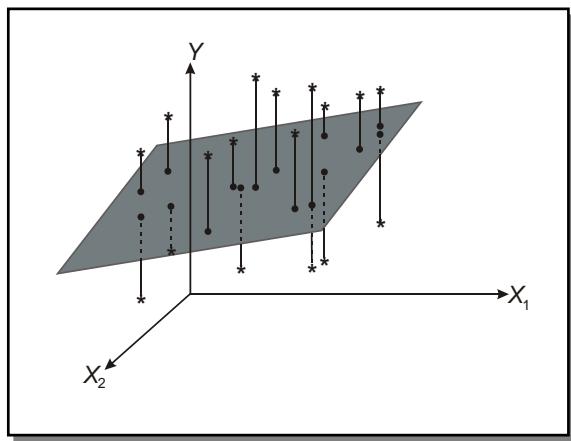
Picturing the Model: No Relationship



5

If there is no relationship among Y and X_1 and X_2 , the model is a horizontal plane passing through the point ($Y = \beta_0, X_1 = 0, X_2 = 0$).

Picturing the Model: A Relationship



6

If there is a relationship among Y and X_1 and X_2 , the model is a sloping plane passing through three points:

- ($Y = \beta_0, X_1 = 0, X_2 = 0$)
- ($Y = \beta_0 + \beta_1, X_1 = 1, X_2 = 0$)
- ($Y = \beta_0 + \beta_2, X_1 = 0, X_2 = 1$)

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

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

7

You investigate the relationship among $k + 1$ variables (k predictors + 1 response) using a k -dimensional surface for prediction.

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

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



The model has $p = k + 1$ parameters (the β s), including the intercept, β_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 β_i s equal zero.

8

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.

Adjusted R^2

The Adjusted R^2 is a modification of R^2 that adjusts for the number of explanatory/predictor variables in a model.

- The Adjusted R^2 increases only if the new term improves the model more than would be expected by chance.
- The Adjusted R^2 can be negative, and will always be less than or equal to R^2 .

9

The Adjusted R^2 is used to compare Multiple Linear Regression models that have been built on the same data. The Adjusted R^2 adjusts the R^2 depending on the number of predictor variables in the model.

$$R_{ADJ}^2 = 1 - \frac{(n - i)(1 - R^2)}{n - p}$$

$i=1$ if there is an intercept and 0 otherwise

n =the number of observations used to fit the model

p =the number of parameters in the model



Fitting a Linear Regression Model with Two Predictor Variables

The following is a summary of what you will accomplish in this demonstration:

- Use PROC REG to perform linear regression analysis of **oxygen** on **runtime** and **rest_pulse**. Interpret the output for the two-variable model.

```
/*st007d01*/
ods graphics off;
proc reg data=st092.fitness;
  model Oxygen_Consumption = runtime rest_pulse;
  title 'Plots of Diagnostic Statistics for runtime';
run;
quit;
```

The only required statement for Proc REG is the MODEL statement. The syntax for the MODEL statement is: MODEL Y = X₁ X₂;

where

Y is the target variable

X₁ and X₂ are the predictor variables.

PROC REG Partial Output

Source	DF	Analysis of Variance		F Value	Pr > F
		Sum of Squares	Mean Square		
Model	2	633.14465	316.57232	40.58	<.0001 ①
Error	28	218.40991	7.80035		
Corrected Total	30	851.55455			
Root MSE		2.79291	R-Square	0.7435 ③	
Dependent Mean		47.37581	Adj R-Sq	0.7252	
Coeff Var		5.89523			
Variable	DF	Parameter Estimates		t Value	Pr > t
		Parameter Estimate	Standard Error		
Intercept	1	82.68886	4.42340	18.69	<.0001 ②
RunTime	1	-3.28691	0.41164	-7.98	<.0001
Rest_Pulse	1	-0.00968	0.07496	-0.13	0.8982

① Examine the Analysis of Variance Table first

Model DF	is 2, the number of parameters minus one (the intercept counts as a parameter).
Error DF	is 28, the total number of observations (31) minus the number of parameters (3).
Corrected Total DF	is 30, the number of observations minus one.
Model Sum of Squares	is the total variation in the target explained by the model.
Error Sum of Squares	is the variation in the target not explained by the model.
Corrected Total Sum of Squares	is the total variation in the target variable.
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 the underlying variance.
F value	is the $\frac{MSM}{MSE}$
Pr > F	is less than 0.0001.

Step 1- Set Hypothesis

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \text{At least one } \beta \text{ is not equal to 0.}$$

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$\text{p-value} < 0.0001.$$

Step 4- Decision Rule.

The p-value < α , Reject H_0 . therefore you reject the null hypothesis and conclude at least one slope parameter is not 0.

② Review the parameter estimate table, including significance of p-values.

The parameter estimate for **runtime** is very similar to what it was in the previous model (-3.31085), but the interpretation is now different. An increase in one unit of **runtime** amounts to a -3.3 decrease in **oxygen** adjusted for **rest_pulse**. In other words, if we could hold **rest_pulse** constant, this would be the slope of the line between **oxygen** and **runtime**. The p-value is less than 0.0001, so the slope is significantly different to 0.

The parameter estimate for **rest_pulse** is -0.00968, which means an increase in one unit of **rest_pulse** amounts to a 0.01 decrease in **oxygen** adjusted for **runtime**. This parameter estimate is a very small value in comparison to its standard error. This is reflected in a small t-value and large p-value. We would conclude that the parameter estimate is not significantly different to 0.

The model we have fitted can be written as:

$$\text{oxygen} = 82.68886 - 3.28691 * (\text{runtime}) - 0.00968 * (\text{rest_pulse}).$$

We would use this model to produce predicted values of oxygen consumption based on runtime and resting pulse.

- ③ If you have another model on the same data, compare the R^2 and the adjusted R^2 .

The R^2 for this model, 0.7435, is only slightly larger than the R^2 for the model in which **runtime** is the only predictor variable, 0.7434

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

The adjusted R^2 is a similar measure to the R^2 , but it takes account of the number of terms in the model. The adjusted R^2 for this model is 0.7252, which is smaller than the adjusted R^2 of 0.7345 for the **runtime** only model. This strongly suggests that the variable **rest_pulse** does not explain any more of oxygen consumption if you know **runtime**.



Many analysts would argue that because this model has a lower adjusted R^2 , and the parameter estimate for **rest_pulse** is not significantly different to 0 that it would be “better” to use the previous model rather than this one. We will discuss some model selection strategies in the next section.

Assumptions for Linear Regression

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

11

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

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.

12

The advantage of performing multiple linear regression over a series of simple linear regression models far outweighs 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 might depend upon the purposes of the analysis, as well as subject-matter expertise.

Common Applications

Multiple linear regression is a powerful tool for:

- 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.

13

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, such as GLM, 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

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

14

Most investigators do not ignore the terms in the model (the Xs), the values of their coefficients (the β 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.

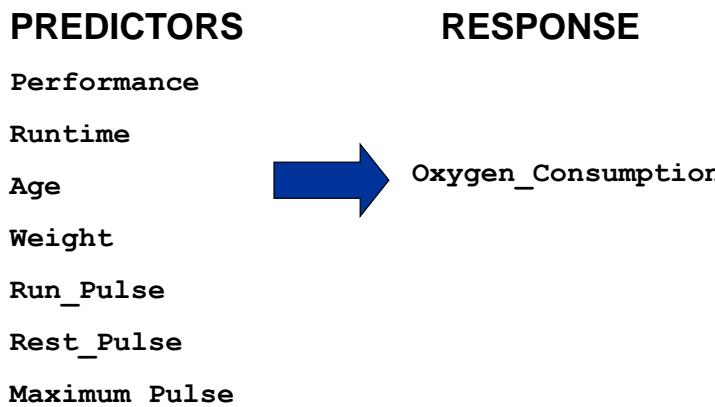
Analytical or Explanatory Analysis

- 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.

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

15

Multiple Regression Example



16

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 **RunTime** as the predictor variable.

The R^2 for this model was 0.7434, which suggests that 25.64% of the variation in **Oxygen_Consumption** is still unexplained.

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



Fitting a Multiple Linear Regression Model

The following is a summary of what you will accomplish in this demonstration:

- Use the PROC REG and perform multiple linear regression analysis of **Oxygen_Consumption** against all other continuous variables.

```
/*st007d02*/
ods graphics off;
proc reg data=st092.fitness;
  model oxygen_consumption = performance
    runtime
    rest_pulse
    run_pulse
    maximum_pulse
    age
    weight;
  title 'Regression of Oxygen Consumption';
run;
quit;
```

Partial Output of Proc REG.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.66124	103.23732	18.42	<.0001 ①
Error	23	128.89331	5.60406		
Corrected Total	30	851.55455			
Root MSE		2.36729	R-Square	0.8486	
Dependent Mean		47.37581	Adj R-Sq	0.8026	
Coeff Var		4.99683			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	131.78249	72.20754	1.83	0.0810
Performance	1	-0.12619	0.30097	-0.42	0.6789 ②
RunTime	1	-3.86019	2.93659	-1.31	0.2016
Rest_Pulse	1	-0.01512	0.06817	-0.22	0.8264
Run_Pulse	1	-0.36207	0.12324	-2.94	0.0074
Maximum_Pulse	1	0.30102	0.13981	2.15	0.0420
Age	1	-0.46082	0.58660	-0.79	0.4401
Weight	1	-0.05812	0.06892	-0.84	0.4078

1

Step 1- Set Hypothesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_K = 0$$

H_1 : At least one β is not equal to 0.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$p\text{-value} < 0.0001.$$

Step 4- Decision Rule.

The $p\text{-value} < \alpha$, Reject H_0 . therefore you reject the null hypothesis and conclude at least one slope parameter is not 0.

The analysis of variance table shows that the regression model fits the data better than the baseline model. The adjusted R^2 for this model is 0.8026, which is higher than the adjusted R^2 for the Simple Linear Regression model, 0.7345, indicating that this model explains more of the variability in oxygen consumption than the first model.

2 Many of the parameter estimates for the predictor variables have large p -values and are therefore not significantly different to 0. If the predictor does not explain much of the variability in the target variable, then it can be removed from the model. Removing unimportant predictors will leave a model that is easier to explain and interpret, but is still good at explaining the variability in the target variable.



Model selection options are discussed in the next section.

7.2 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.

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.

20

A process for selecting models might be to start with all the variables in the **fitness** data set and eliminate the least significant terms, based on p-values.

For a small data set, a final 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 would have to continue this process until only terms with *p*-values lower than some threshold value, such as 0.10 or 0.05, remain.

Model Selection Options

The Linear Regression Task Supports several possible model selection techniques.

Among those selection methods are;

- FORWARD Selection
- BACKWARD Elimination
- STEPWISE Selection

SELECTION=NONE is the default.

21

Stepwise Selection Methods



FORWARD
SELECTION



BACKWARD
ELIMINATION



STEPWISE
SELECTION

22

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:

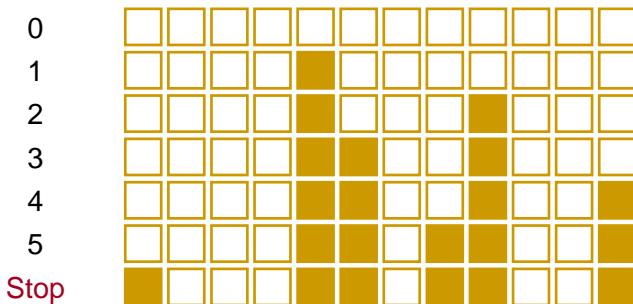
FORWARD first selects the best one-variable model. Then it selects the best two variables among those that contain the first selected variable. FORWARD continues this process, but stops when it reaches the point where no additional variables have a *p*-value level < 0.50.

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 *p*-value < 0.10.

STEPWISE works like a combination of the two. The default entry *p*-value is 0.15 and the default stay *p*-value is also 0.15.

 The SLENTRY= and SLSTAY= options can be used to change the default values.

Forward Selection

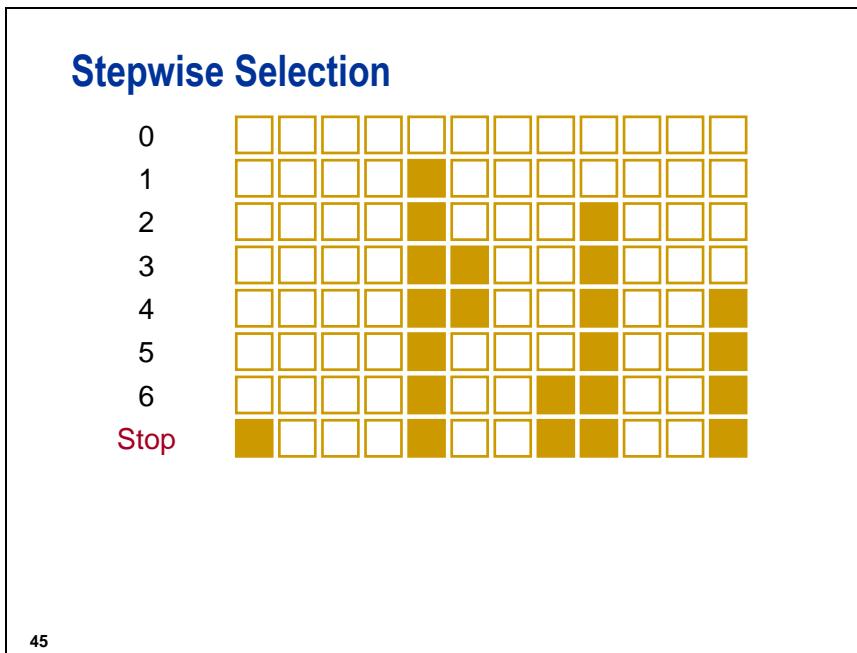


29

Forward selection starts with an empty model. The method computes an F statistic for each predictor variable not in the model and examines the largest of these statistics. If it is significant at a specified significance level (specified by the SLENTRY= option), the corresponding variable is added to the model. After a variable is entered in the model, it is never removed from the model. The process is repeated until none of the remaining variables meet the specified level for entry. By default, SLENTRY=0.50.

Backward Elimination

Backward elimination starts off with the full model. Results of the *F* test for individual parameter estimates are examined, and the least significant variable that falls above the specified significance level specified by the SLSTAY= option) is removed. After a variable is removed from the model, it remains excluded. The process is repeated until no other variable in the model meets the specified significance level for removal. By default, SLSTAY=0.10.



Stepwise selection is similar to forward selection in that it starts with an empty model and incrementally builds a model one variable at a time. However, the method differs from forward selection in that variables already in the model do not necessarily remain. The backward component of the method removes variables from the model that do not meet the significance criteria specified in the SLSTAY= option. The stepwise selection process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.

Stepwise selection (Forward, Backward, and Stepwise) has some serious shortcomings. Simulation studies (Derkzen and Keselman 1992) evaluating variable selection techniques found the following:

1. The degree of collinearity among the predictor variables affected the frequency with which authentic predictor variables found their way into the final model.
2. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.
3. The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.

One recommendation is to use the variable selection methods to create several candidate models, and then use subject-matter knowledge to select the variables that result in the best model within the scientific or business context of the problem. Therefore, you are simply using these methods as a useful tool in the model-building process (Hosmer and Lemeshow 2000).



Stepwise Regression

The following is a summary of what you will accomplish in this demonstration:

- Use the selection methods discussed, FORWARD, BACKWARD and STEPWISE methods, to select a model for predicting **Oxygen_Consumption** in the **fitness** data set.

```
/*st007d03*/
ods graphics on;
proc reg data=st092.fitness plots(only)=adjrsq;
  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 'Best Models Using Stepwise Selection';
run;
quit;
```

Selected PLOTS option;

ADJRSQ produces a plot showing the Adjusted R² against the Step number to show which model is the best according to the highest Adjusted R².

Partial PROC REG Output

Best Models Using Stepwise Selection

The REG Procedure

Model: FORWARD

Dependent Variable: Oxygen_Consumption

Number of Observations Read 31
Number of Observations Used 31

Forward Selection: Step 1

Variable RunTime Entered: R-Square = 0.7434 and C(p) = 11.9967

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			

Variable	Parameter	Standard	Type	II	SS	F Value	Pr > F
	Estimate	Error					
Intercept	82.42494	3.85582	3443.63138		456.97	<.0001	
RunTime	-3.31085	0.36124	633.01458		84.00	<.0001	

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable Age Entered: R-Square = 0.7647 and C(p) = 10.7530

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	651.19281	325.59640	45.50	<.0001
Error	28	200.36175	7.15578		
Corrected Total	30	851.55455			

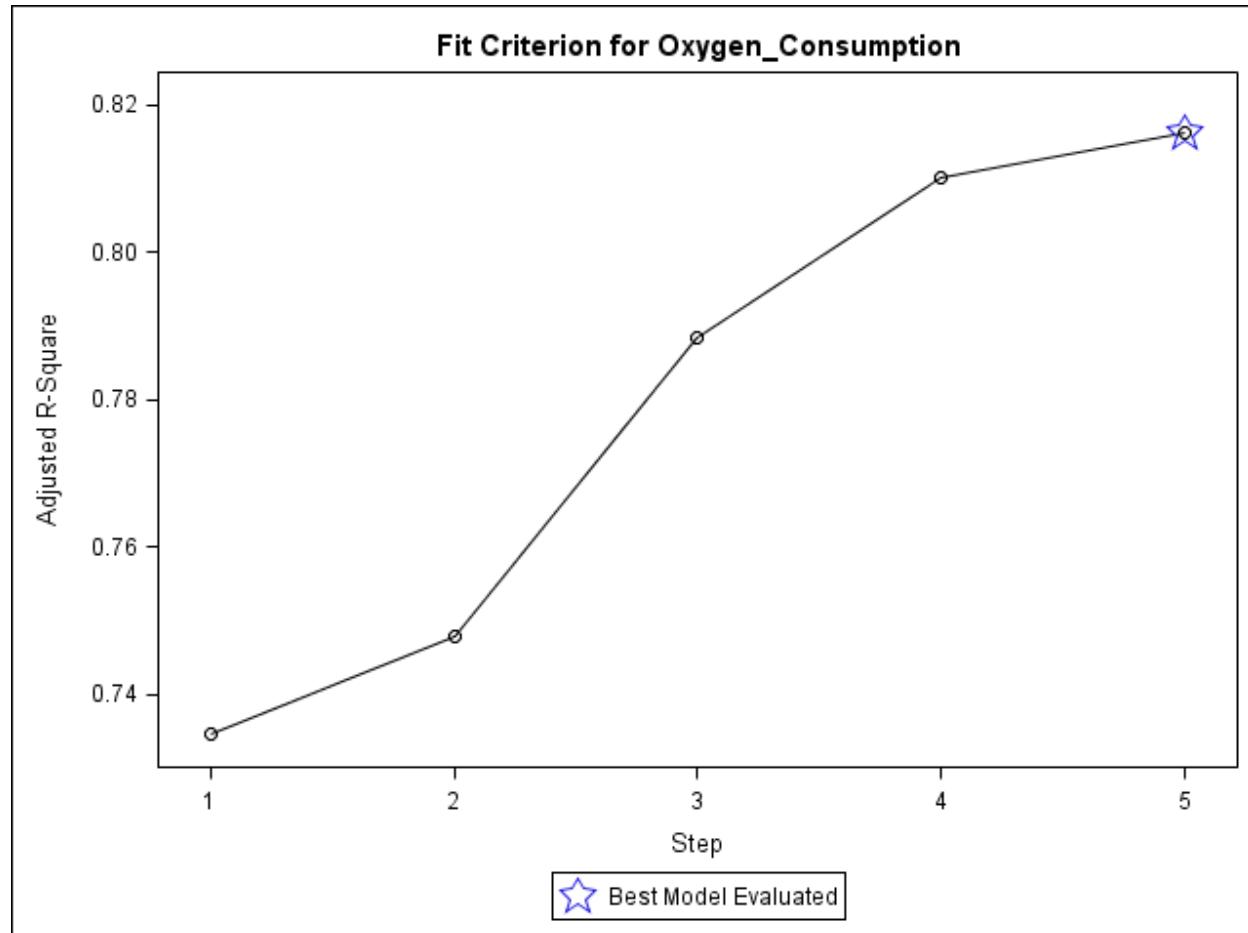
Variable	Parameter	Standard	Type	II	SS	F Value	Pr > F
	Estimate	Error					
Intercept	88.43358	5.32255	1975.38438		276.05	<.0001	
RunTime	-3.19917	0.35892	568.50196		79.45	<.0001	
Age	-0.15082	0.09463	18.17822		2.54	0.1222	

Bounds on condition number: 1.0396, 4.1585

Partial PROC REG Output (Continued)

Best Models Using Stepwise Selection							
The REG Procedure Model: FORWARD Dependent Variable: Oxygen_Consumption							
Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RunTime	1	0.7434	0.7434	11.9967	84.00	<.0001
2	Age	2	0.0213	0.7647	10.7530	2.54	0.1222
3	Run_Pulse	3	0.0449	0.8096	5.9367	6.36	0.0179
4	Maximum_Pulse	4	0.0259	0.8355	4.0004	4.09	0.0534
5	Weight	5	0.0115	0.8469	4.2598	1.87	0.1836

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^2 shows the increase in the model R^2 as each term was added.



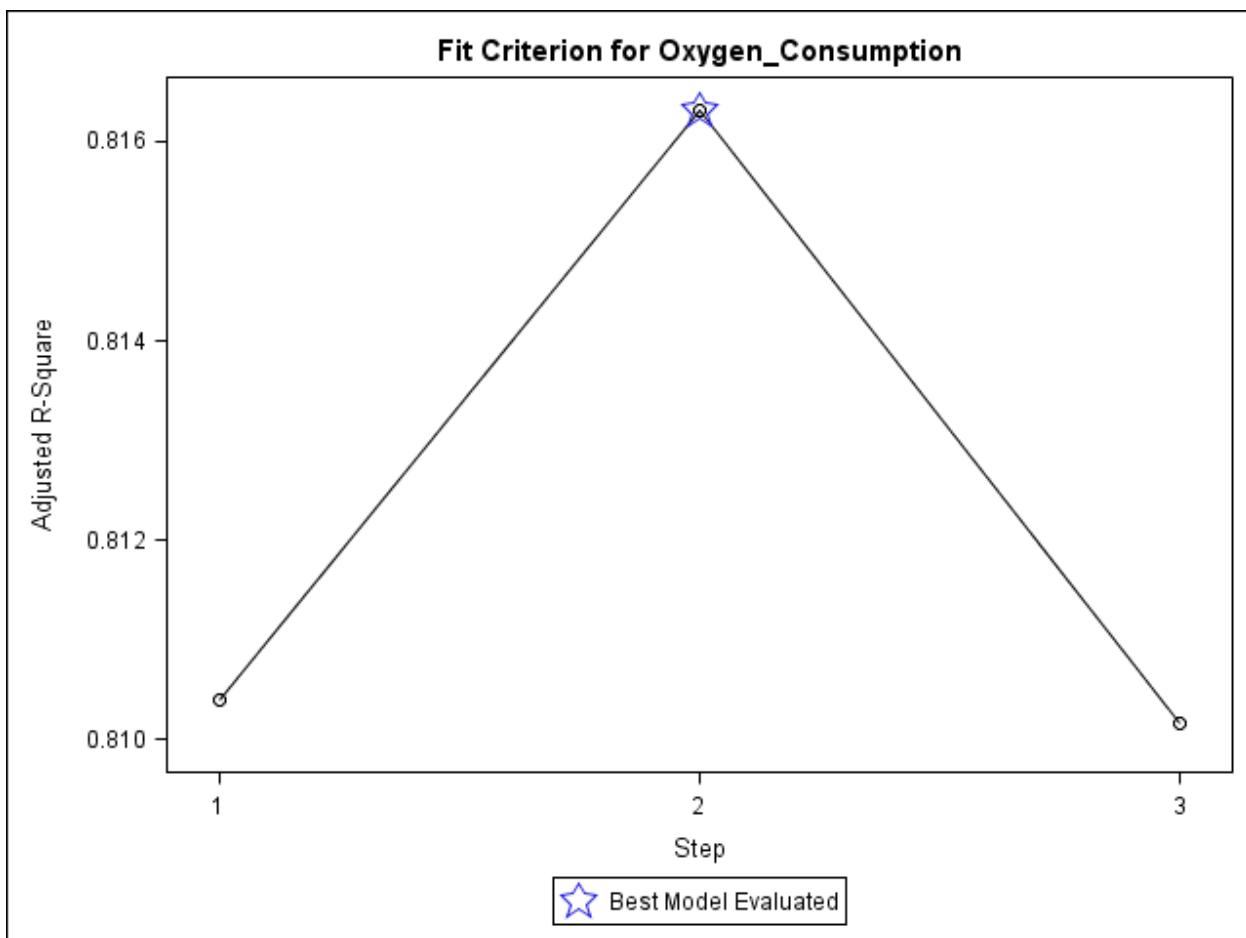
The Adjusted R-Square plot shows the progression of that statistic at each step. The star denotes the best model of the 5 tested. This is not necessarily the highest Adjusted R-Square value of all possible subsets, but is the best of the five tested in the Forward model.

Partial PROC REG Output (Continued)

Best Models Using Stepwise Selection					
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 = 0.8486 and C(p) = 8.0000					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.66124	103.23732	18.42	<.0001
Error	23	128.89331	5.60406		
Corrected Total	30	851.55455			
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	131.78249	72.20754	18.66607	3.33	0.0810
Performance	-0.12619	0.30097	0.98519	0.18	0.6789
RunTime	-3.86019	2.93659	9.68350	1.73	0.2016
Age	-0.46082	0.58660	3.45842	0.62	0.4401
Weight	-0.05812	0.06892	3.98514	0.71	0.4078
Run_Pulse	-0.36207	0.12324	48.37354	8.63	0.0074
Rest_Pulse	-0.01512	0.06817	0.27581	0.05	0.8264
Maximum_Pulse	0.30102	0.13981	25.97886	4.64	0.0420
Bounds on condition number: 162.85, 2262.9					

Partial PROC REG Output (Continued)

Backward Elimination: Step 3																																									
Variable Weight Removed: R-Square = 0.8355 and C(p) = 4.0004																																									
Analysis of Variance																																									
<table> <thead> <tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr> </thead> <tbody> <tr><td>Model</td><td>4</td><td>711.45087</td><td>177.86272</td><td>33.01</td><td><.0001</td></tr> <tr><td>Error</td><td>26</td><td>140.10368</td><td>5.38860</td><td></td><td></td></tr> <tr><td>Corrected Total</td><td>30</td><td>851.55455</td><td></td><td></td><td></td></tr> </tbody> </table>						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															
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																																							
<table> <thead> <tr><th>Variable</th><th>Parameter Estimate</th><th>Standard Error</th><th>Type III SS</th><th>F Value</th><th>Pr > F</th></tr> </thead> <tbody> <tr><td>Intercept</td><td>97.16952</td><td>11.65703</td><td>374.42127</td><td>69.48</td><td><.0001</td></tr> <tr><td>RunTime</td><td>-2.77576</td><td>0.34159</td><td>355.82682</td><td>66.03</td><td><.0001</td></tr> <tr><td>Age</td><td>-0.18903</td><td>0.09439</td><td>21.61272</td><td>4.01</td><td>0.0557</td></tr> <tr><td>Run_Pulse</td><td>-0.34568</td><td>0.11820</td><td>46.08558</td><td>8.55</td><td>0.0071</td></tr> <tr><td>Maximum_Pulse</td><td>0.27188</td><td>0.13438</td><td>22.05933</td><td>4.09</td><td>0.0534</td></tr> </tbody> </table>						Variable	Parameter Estimate	Standard Error	Type III 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
Variable	Parameter Estimate	Standard Error	Type III 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																																									
<hr/>																																									
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																																			
1	Rest_Pulse	6	0.0003	0.8483	6.0492	0.05																																			
2	Performance	5	0.0014	0.8469	4.2598	0.22																																			
3	Weight	4	0.0115	0.8355	4.0004	1.87																																			
						0.1836																																			



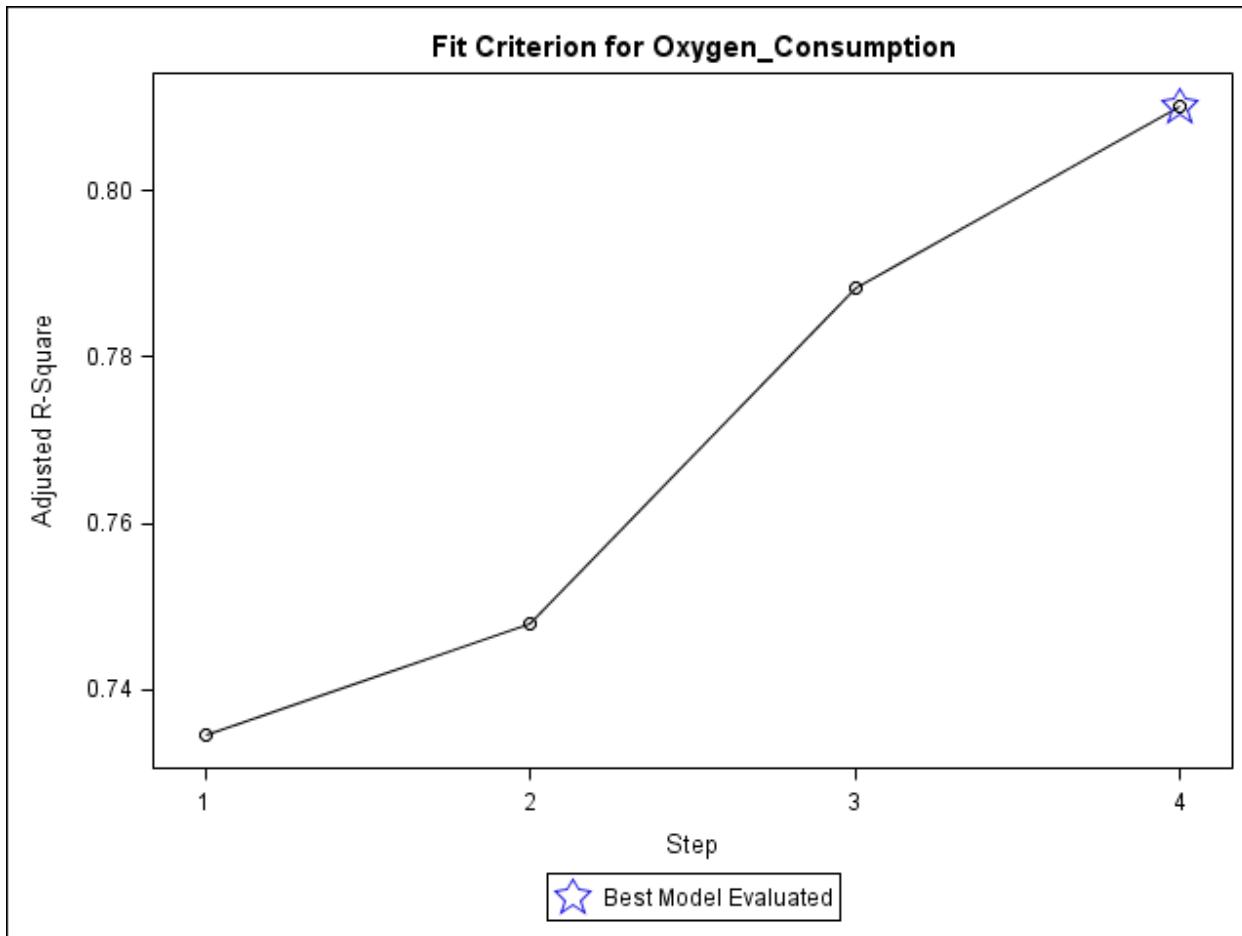
The Adjusted R-Square for the model at step 2 (before **Weight** was removed) was greatest of the three tested.

Partial PROC REG Output (Continued)

Stepwise Selection: Step 4					
Variable Maximum_Pulse Entered: R-Square = 0.8355 and C(p) = 4.0004					
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			
Parameter Standard					
Variable	Estimate	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					

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					
Variable Step Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p) F Value Pr > F
1 RunTime		1	0.7434	0.7434	11.9967 84.00 <.0001
2 Age		2	0.0213	0.7647	10.7530 2.54 0.1222
3 Run_Pulse		3	0.0449	0.8096	5.9367 6.36 0.0179
4 Maximum_Pulse		4	0.0259	0.8355	4.0004 4.09 0.0534

Using the STEPWISE option and the default p -value, the same subset resulted as that using the FORWARD option.



The SLENTRY= default criterion is $p < .50$ for the FORWARD method and $p < .15$ for the STEPWISE method. After **RunTime** was entered into the model, **Age** was entered at step 2 with a p -value of 0.1222.

If the SLENTRY= criterion were set to something less than 0.10, the final model would have been quite different. It would have included only one variable, **RunTime**. This underscores the precariousness of relying on one stepwise method for defining a “best” model.

Stepwise Regression Models

FORWARD	Runtime, Age, Weight, Run_Pulse, Maximum_Pulse
BACKWARD	Runtime, Age, Run_Pulse, Maximum_Pulse
STEPWISE	Runtime, Age, Run_Pulse, Maximum_Pulse

47

The final models obtained using the default SLENTRY= and SLSTAY= criteria are displayed. It is important to note that the choice of criterion levels can greatly affect the final models that are selected using stepwise methods.

Stepwise Models, Alternative Criteria

FORWARD (slentry=0.05)	Runtime
BACKWARD (slstay=0.05)	Runtime, Run_Pulse, Maximum_Pulse
STEPWISE (slentry=0.05, slstay=0.05)	Runtime

48

The final models using 0.05 as the forward and backward step criteria resulted in very different models than those chosen using the default criteria.

Some Problems

Some additional problems of multiple linear regression are:

- Influential observations
- Collinearity

49

Influential observations

Detecting outliers and influential points in simple linear regression could be achieved by looking at Studentized residuals. This may be more difficult in multiple linear regression because scatter plots may not reveal influential points. One solution is to use Partial Leverage Plots, which are a graphical method for visualising the test of significance for the parameter estimate in the full model.

Collinearity

Collinearity can be a problem if there are many highly related variables in a model. In this case one variable can provide nearly as much information as the other variable. This can make the model unstable and both variables may appear to be non-significant. There are collinearity statistics that can be calculated in Proc REG to enable you to detect collinearity. However, if you use a SELECTION method for choosing the variables in the model, it is unlikely to choose two variables that are highly related.

Stepwise Regression has produced a model with four predictor variables: **runtime**, **age**, **run_pulse**, and **max_pulse**. We are now going to examine some residual plots to check the assumptions.

Examining the Model

We should examine some plots to check assumptions:

- Residuals versus predicted values
- Residuals versus each predictor variable
- Normal quantile plot



Examining Assumptions

The following is a summary of what you will accomplish in this demonstration:

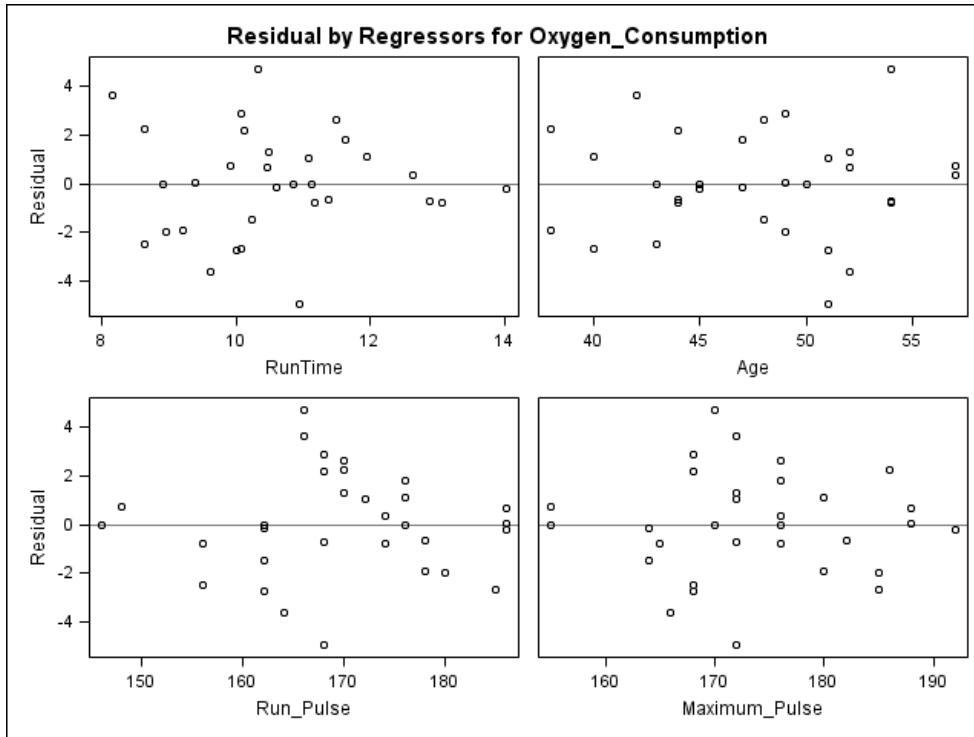
- Use the model selected in the STEPWISE selection method and examine the assumptions of this model.

```
/*st007d04*/
ods graphics on;
proc reg data=st092.fitness
plots(only)=(QQ
             RESIDUALBYPREDICTED
             RESIDUALHISTOGRAM
             RESIDUALPLOT
             RSTUDENTBYPREDICTED);
model oxygen_consumption = RunTime Age
                           Run_Pulse Maximum_Pulse;
title 'Plots of Diagnostic Statistics';
run;
quit;
```

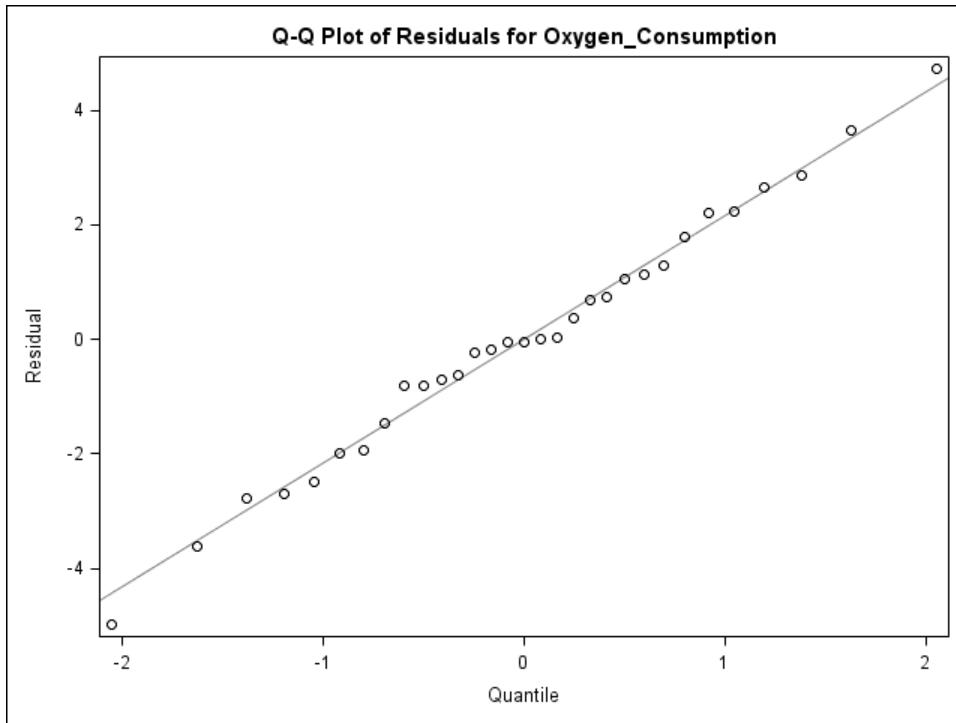
Residual Plots could have been produced when running the previous PROC REG, here we have split them out for demonstration purposes.

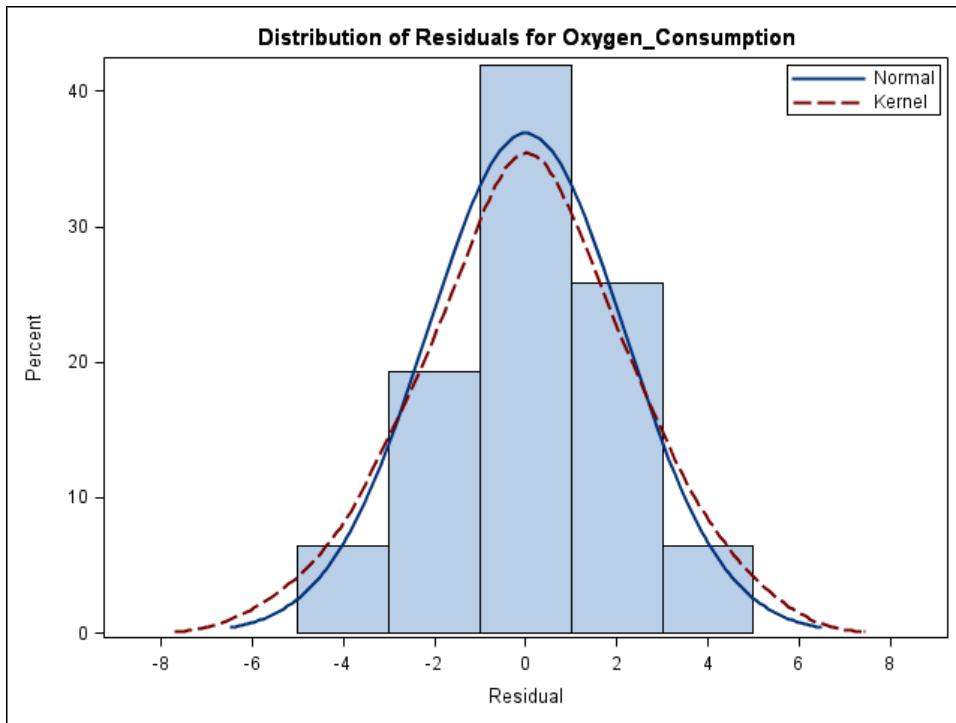
Remember Assumptions should be checked before the regression output is analyzed.

The Panel Plot shows that all graphs show a random scatter around 0. This proves that no unexpected relationships were missed at Explanatory Data phase.

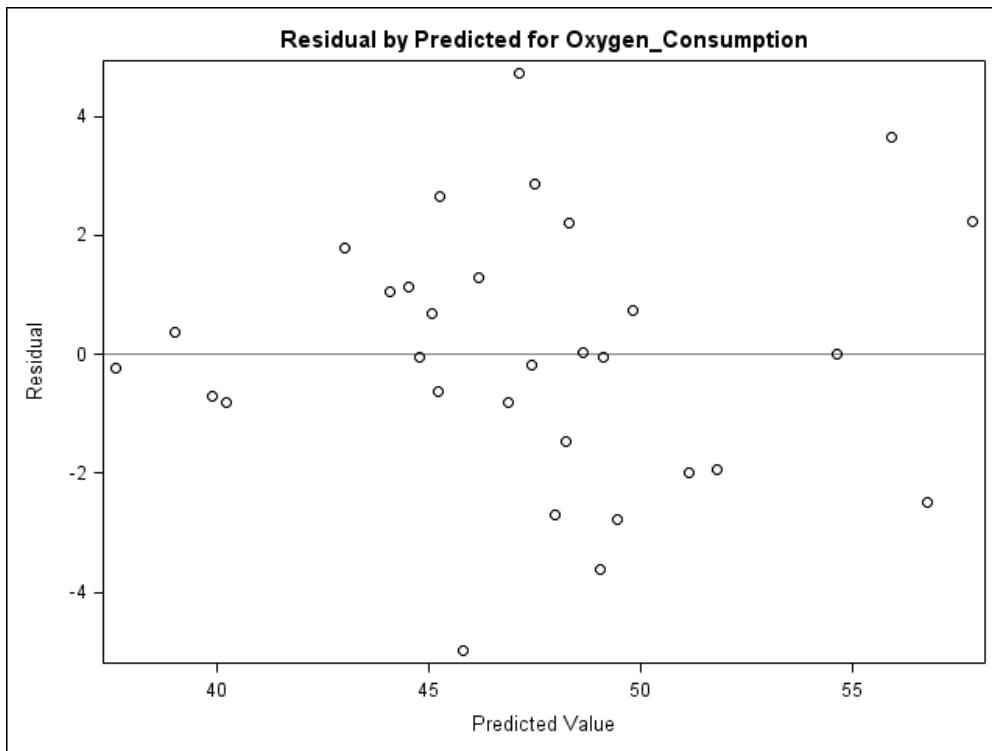


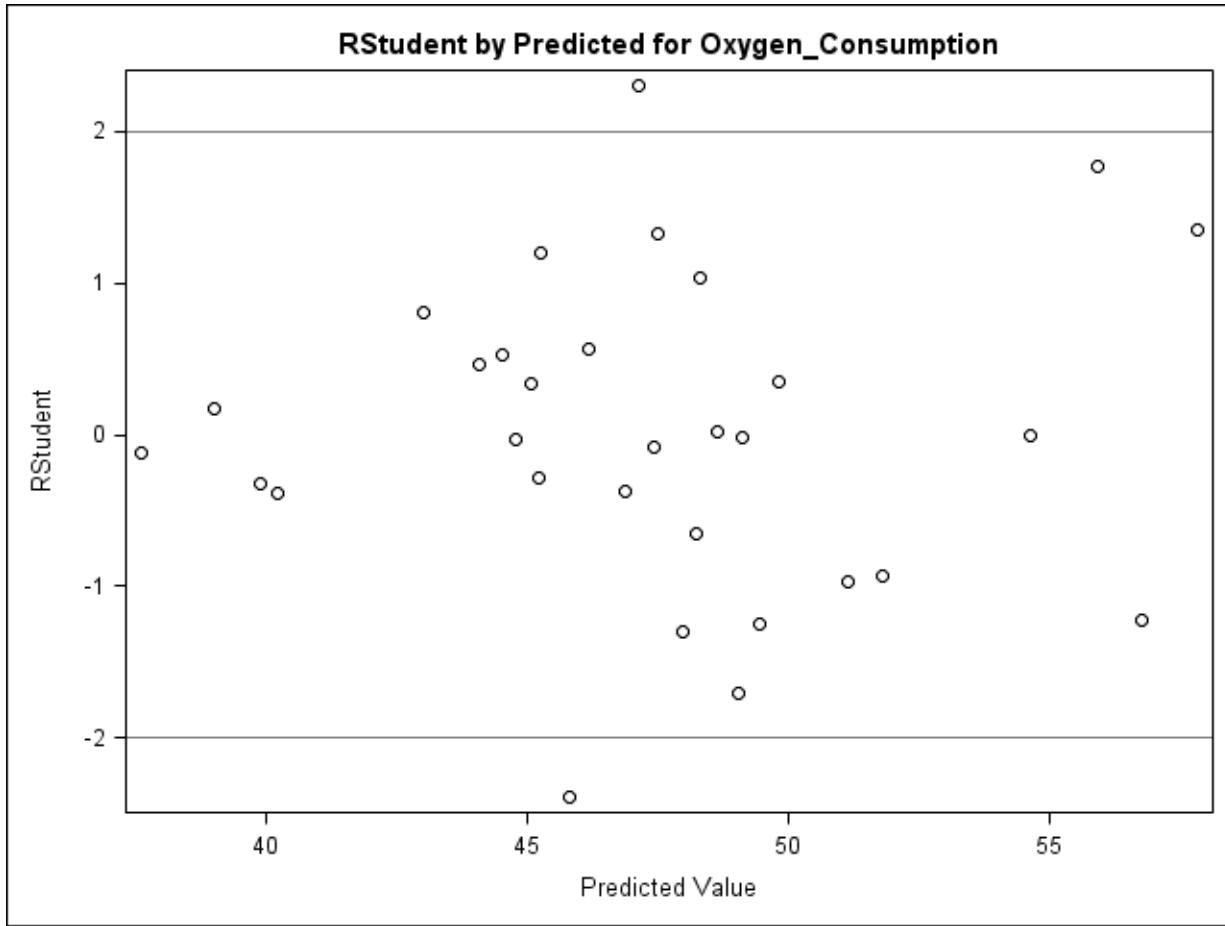
The Q-Q Plot follows a straight line and the residual histogram follows a normal distribution, therefore the normality assumption is verified.





The plot of residuals versus predicted values is shown below. The residual values appear to be randomly scattered about the reference line at 0.





There appear to be a couple of large studentised residuals, over $|2|$, which would be expected but are no cause for concern.

Chapter 8 Categorical Data Analysis

8.1 Describing Categorical Data	8-3
Demonstration: Examining Distributions	8-14
Demonstration: Ordering Values in a Frequency Table	8-23
8.2 Tests of Association	8-26
Demonstration: Chi-Square Test.....	8-33
Demonstration: Detecting Ordinal Associations.....	8-41

8.1 Describing Categorical Data

Objectives

- Recognise 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.

Overview

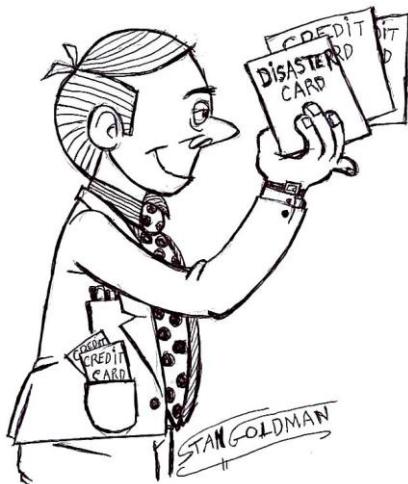
		Type of Predictors		
Type of Response	Categorical	Continuous	Categorical and Continuous	
Continuous	Analysis of Variance	Linear Regression	Analysis of Covariance (Regression with dummy variables)	
Categorical	Logistic Regression or Contingency Tables	Logistic Regression	Logistic Regression	

9

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.

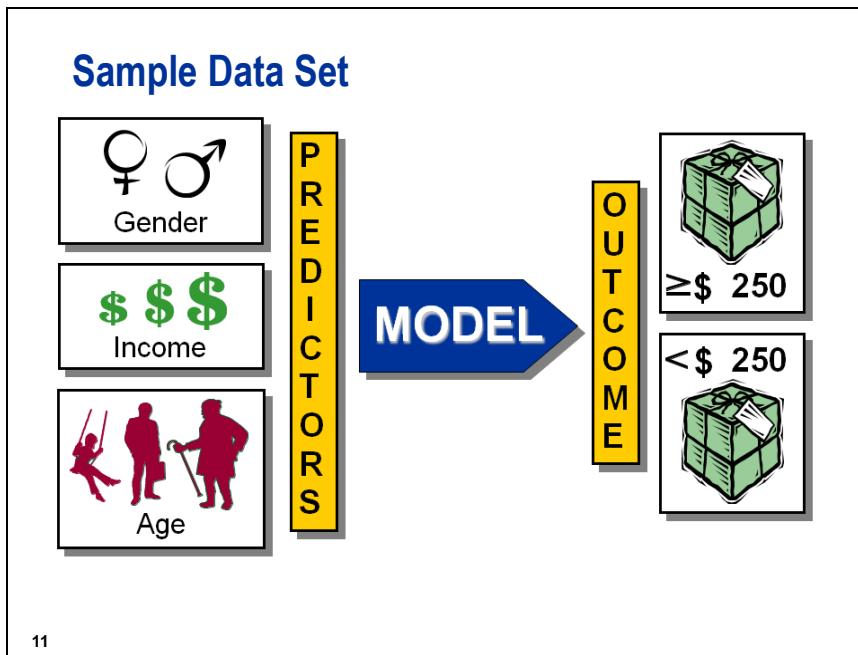
Credit Offer Example

A company wants to identify ‘prime’ customers to make credit card offers to, based on information from the past six months.



10

Example: A company that sells its products via a catalog wants to identify those customers to whom they will offer credit cards. It has been decided that customers who spend 250 dollars or more in the past six months are the target group. They will be considered “prime” customers. They would like to make the credit card offer only to this group. The data is stored in the **st092.sales** data set.



The variables in the data set are

PURCHASE purchase price (1=250 dollars or more, 0=Under 250 dollars)

AGE age of customers in years

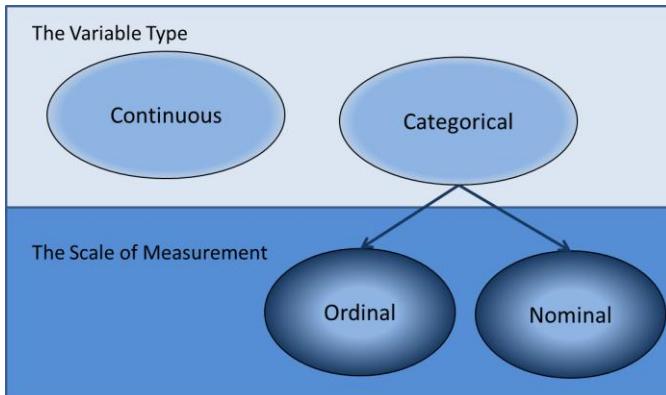
GENDER gender of customer (Male, Female)

INCOME annual income (Low, Middle, High).

 This is a hypothetical data set.

Scale of Measurement

Before analyzing, identify :



Scale of Measurement: Nominal

Nominal Variable: Type of Beverage

OR

Binary Nominal Variable: Gender

Variables with two levels are often referred to as Binary/Nominal

13

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 have values with no logical ordering. In the **st092.sales** data set, **GENDER** is a nominal variable.

Ordinal variables have values with a logical order. However, the relative distances between the values are not clear. In the **st092.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

- determine the frequency of data values
- recognize possible associations among variables.

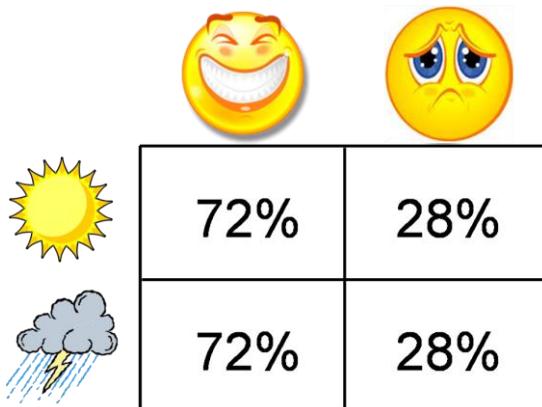
14

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.

15

No Association

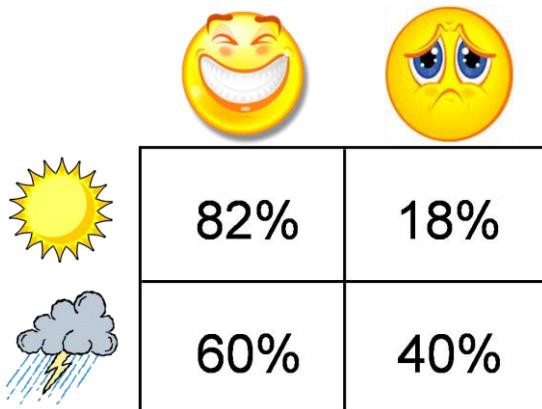


Is your manager's mood associated
with the weather?

16

There appears to be no association here because the row percentages are the **same** in each column.

Association



Is your manager's mood associated
with the weather?

17

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.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

18

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 ₁₁	cell ₁₂	...	cell _{1c}
row 2	cell ₂₁	cell ₂₂	...	cell _{2c}
...
row r	cell _{r1}	cell _{r2}	...	cell _{rc}

19

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;
```

20

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

The following is a summary of what you will accomplish in this demonstration:

- Use the PROC FREQ to create one-way frequency tables for the variables **GENDER**, **INCOME**, and **PURCHASE**
- Use the PROC FREQ to create two-way frequency tables for the variables **PURCHASE** and **GENDER**, and **PURCHASE** and **INCOME**
- Use the FORMAT procedure to format the values of **PURCHASE**.
- Use the UNIVARIATE procedure to describe the continuous variable **AGE**.

```
/*st008d01*/
title;
proc format;
  value purfmt 1 = ">=$250"
    0 = "<$250"
    ;
run;
ods graphics on;
ods listing close;
ods rtf file='freq.rtf';
proc freq data=st092.sales;
  tables Purchase Gender Income
    Gender*Purchase Income*Purchase /
      Plots(only)=(freqplot);
  format Purchase purfmt.;
run;

ods select histogram probplot;
proc univariate data=st092.sales;
  var Age;
  histogram Age / normal (mu=est sigma=est);
  probplot Age / normal (mu=est sigma=est);
run;

ods rtf close;
ods listing;
ods graphics off;
```

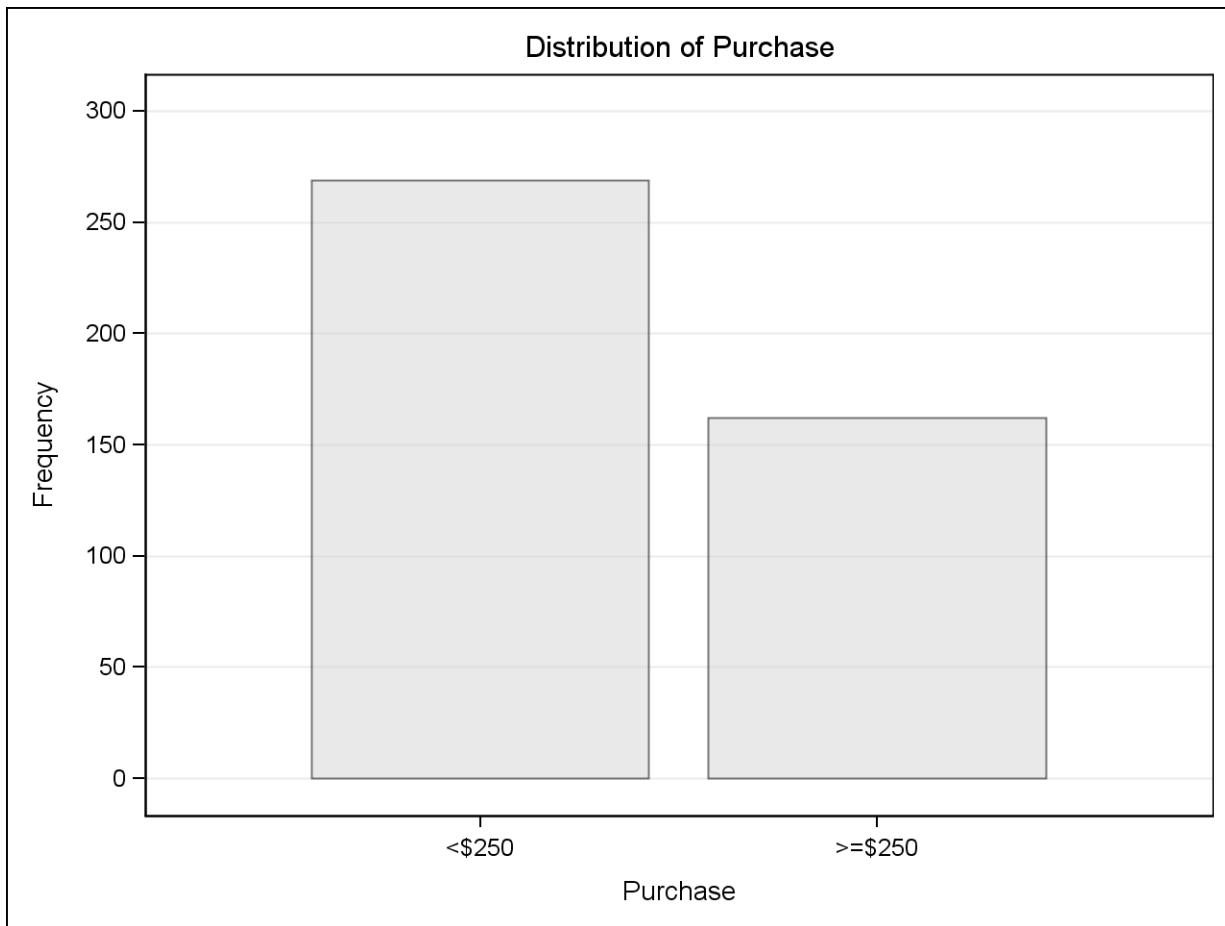
Selected TABLES statement option:

PLOTS(ONLY)=FREQPLOT

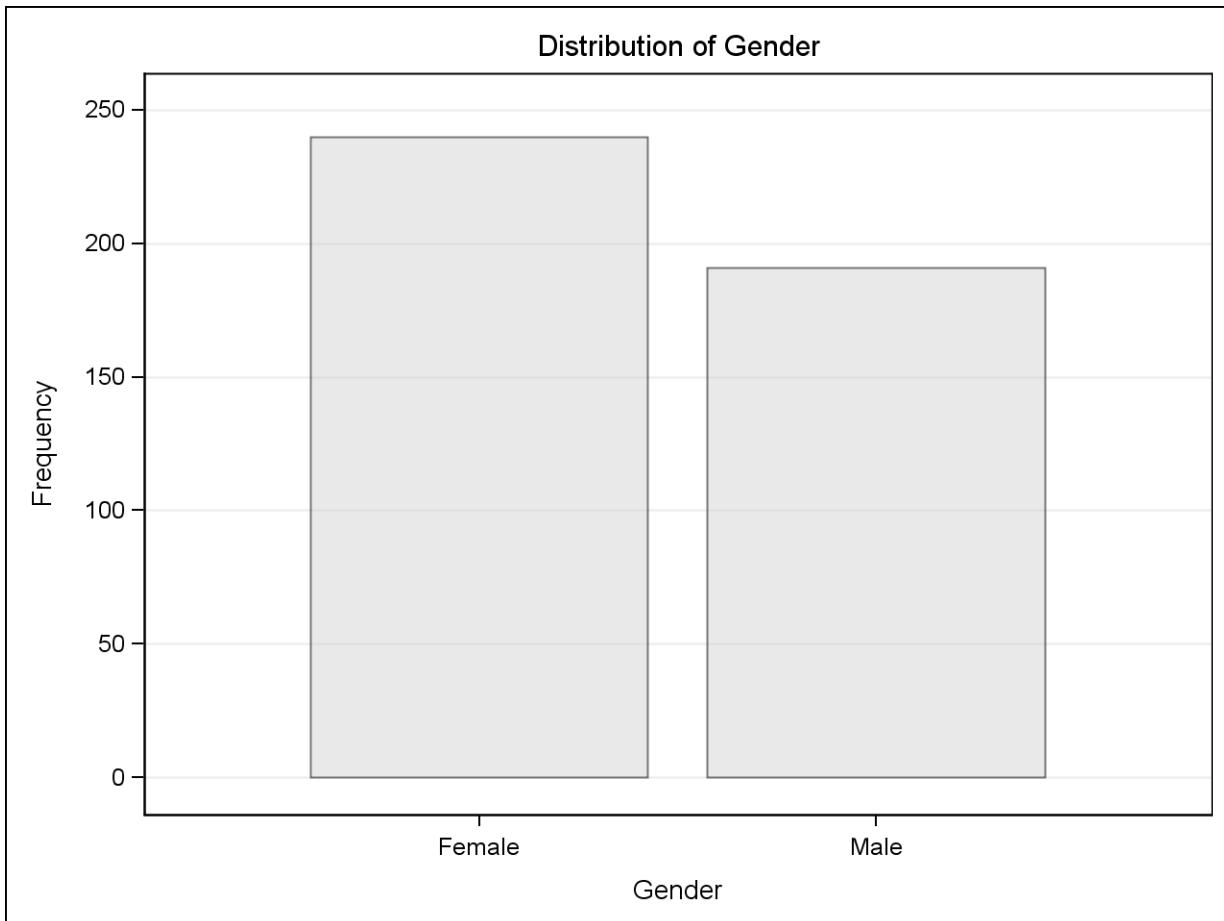
Requests only a frequency plot for one way contingency tables.
By default, a cumulative frequency table would have been otherwise also produced

PROC FREQ Output

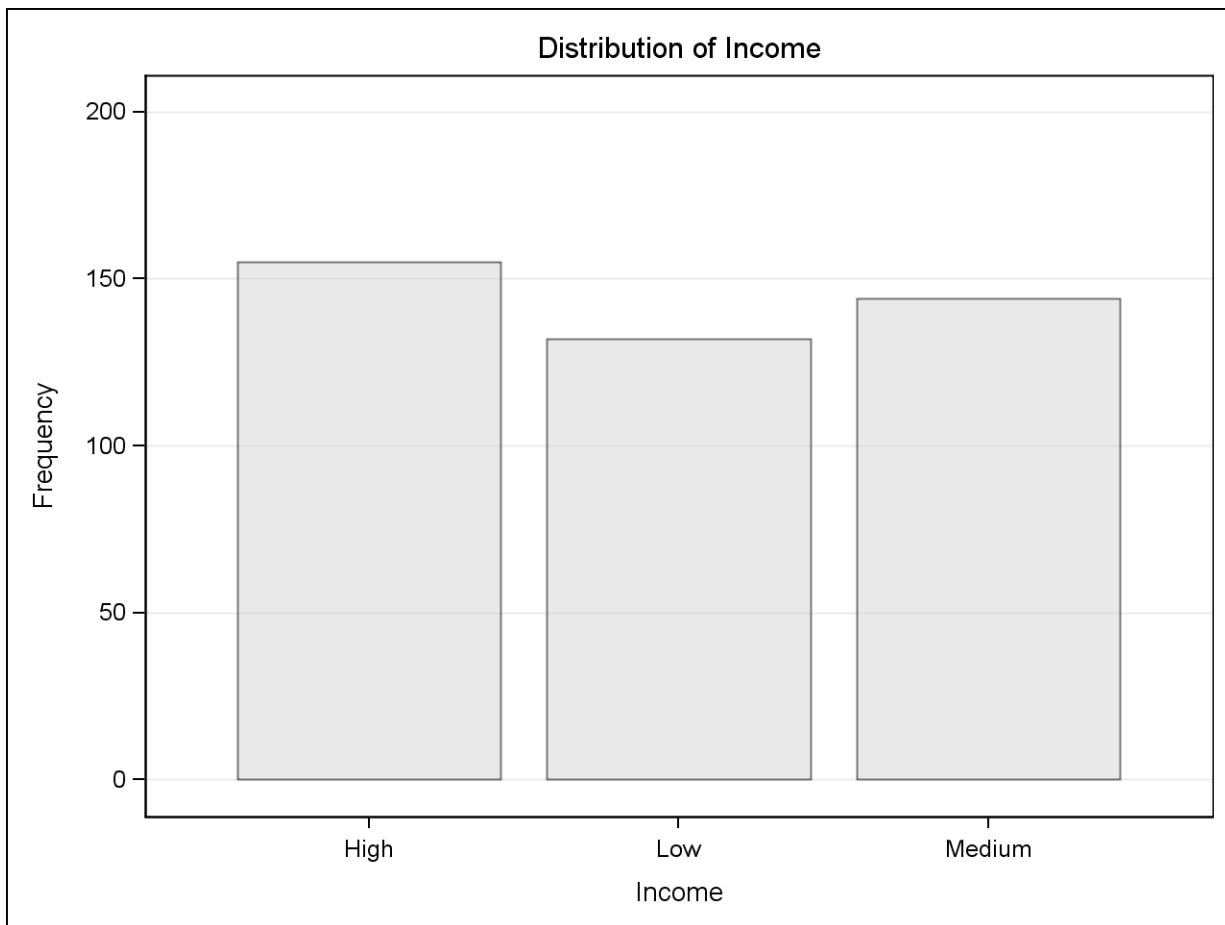
Purchase	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<\$250	269	62.41	269	62.41
>=\$250	162	37.59	431	100.00



Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	240	55.68	240	55.68
Male	191	44.32	431	100.00



<i>Income</i>	<i>Frequency</i>	<i>Percent</i>	<i>Cumulative Frequency</i>	<i>Cumulative Percent</i>
<i>High</i>	155	35.96	155	35.96
<i>Low</i>	132	30.63	287	66.59
<i>Medium</i>	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.

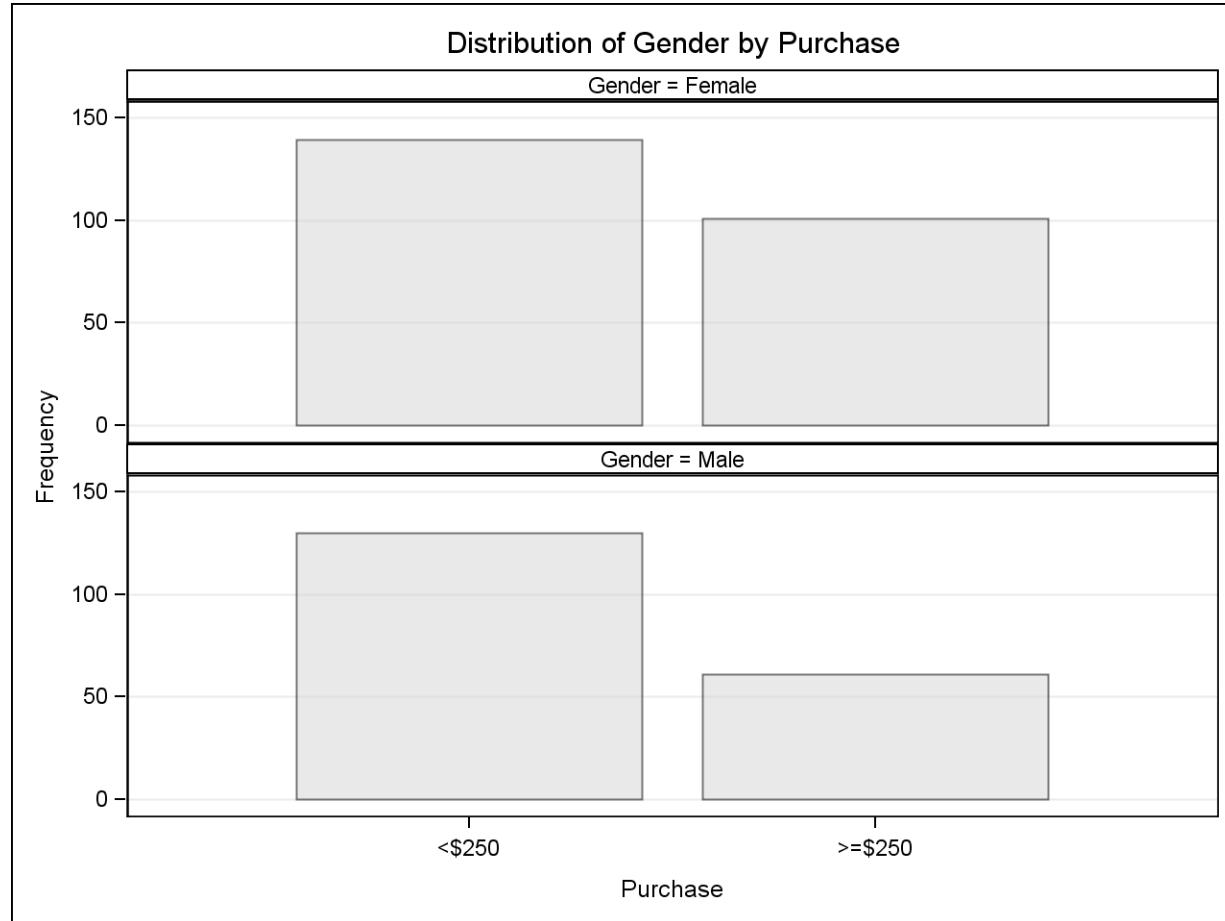
The requested two-way frequency tables are shown below. You can get a preliminary idea whether there are associations between the outcome variable, **PURCHASE**, and the predictor variables, **GENDER** and **INCOME**, by examining the distribution of **PURCHASE** for each value of the predictors.

PROC FREQ Output (Continued)

Table of Gender by Purchase

Gender	Purchase		
	Frequency	Percent	Row Pct
Col Pct	<\$250	>=\$250	Total
Female	139	32.25	240
	57.92	23.43	55.68
	51.67	42.08	
Male	130	30.16	191
	68.06	14.15	44.32
	48.33	31.94	
		37.65	
Total	269	62.41	431
	37.59		100.00

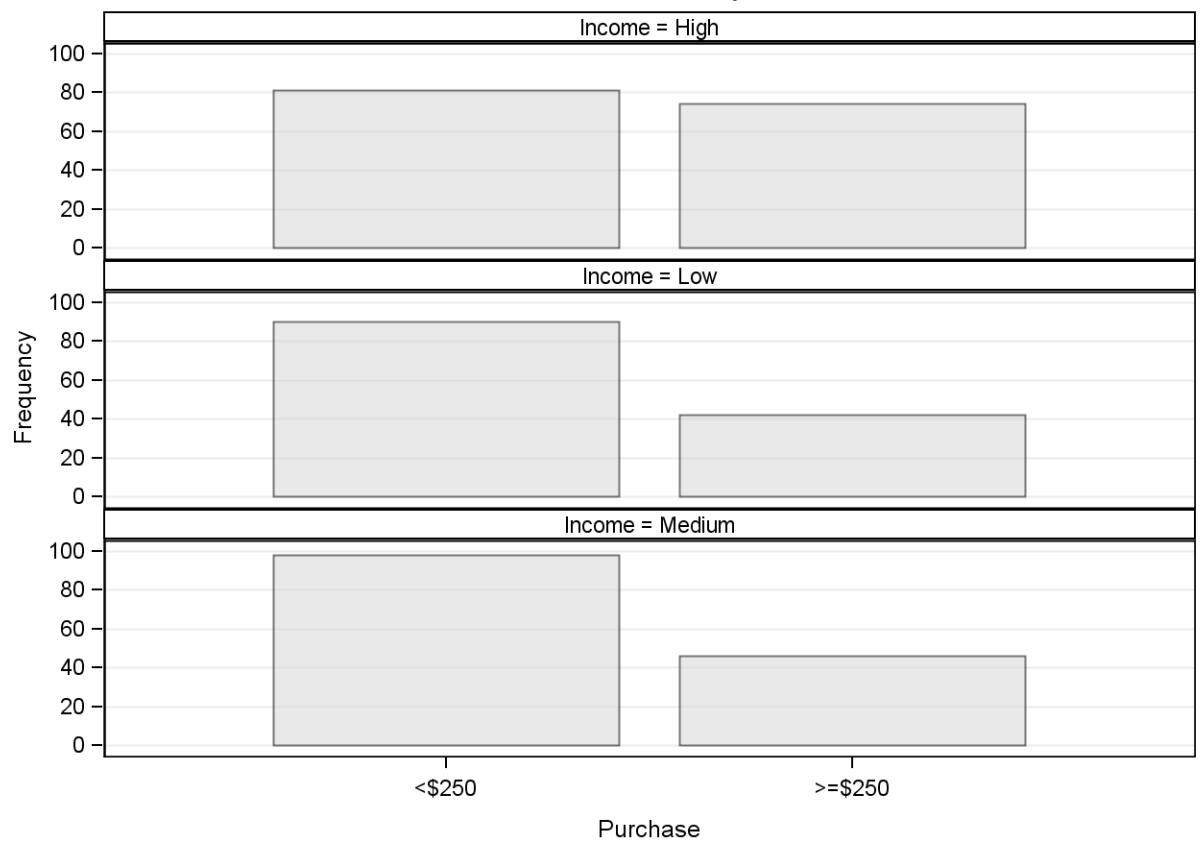
Distribution of Gender by Purchase



By examining the row percentages, you see that PURCHASE may be associated with GENDER.

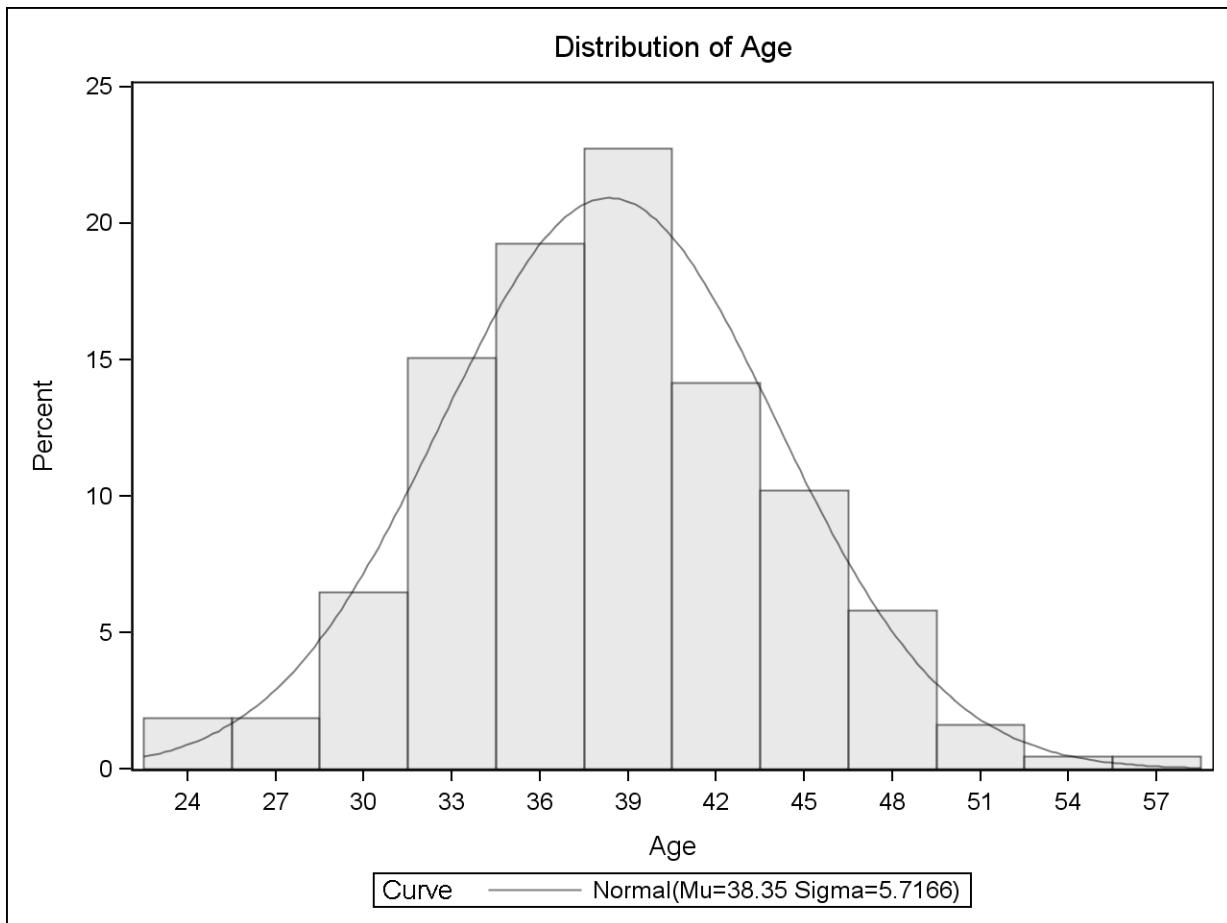
Table of Income by Purchase

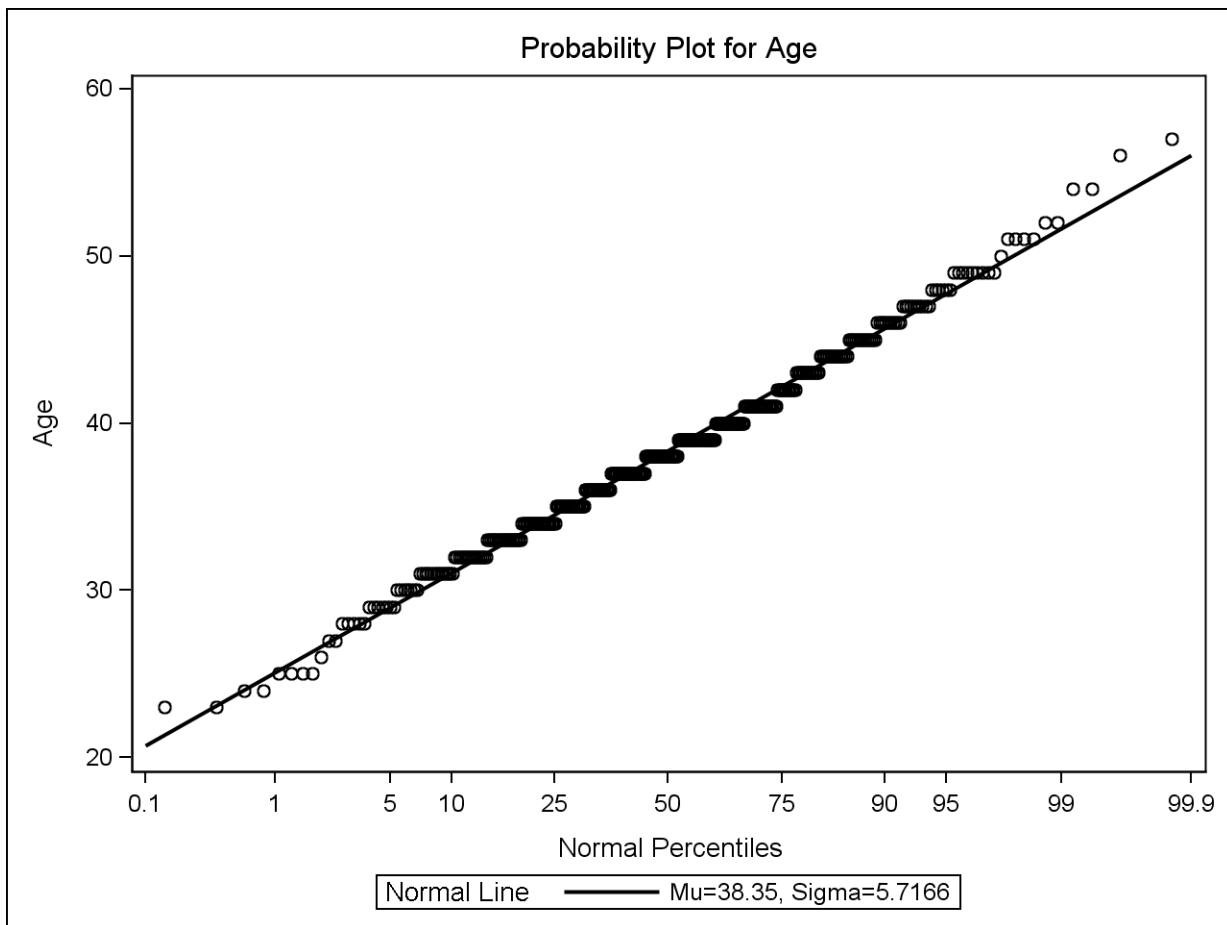
<i>Income</i>	<i>Purchase</i>		
<i>Frequency</i>			
<i>Percent</i>			
<i>Row Pct</i>			
<i>Col Pct</i>	<\$250	>=\$250	<i>Total</i>
<i>High</i>	81 18.79 52.26 30.11	74 17.17 47.74 45.68	155 35.96
<i>Low</i>	90 20.88 68.18 33.46	42 9.74 31.82 25.93	132 30.63
<i>Medium</i>	98 22.74 68.06 36.43	46 10.67 31.94 28.40	144 33.41
<i>Total</i>	269 62.41	162 37.59	431 100.00

Distribution of Income by Purchase

There also seems to be an association between **PURCHASE** and **INCOME**.

The plots show the distribution of the one continuous variable, **AGE**.





The distribution of **AGE** appears approximately normal, with no obvious outliers.

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

22

The default ordering in PROC FREQ for character variables is alphanumeric. High comes first among the three categories in alphabetical order, followed by Low and Medium. This ordering will not only affect the visual presentation of tables and graphs produced from the variable `INCOME`, but also any statistical analyses where you want to treat the variable as ordinal.

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 a Frequency Table

The following is a summary of what you will accomplish in this demonstration:

- Obtain a logical order in a frequency table for the values in the variable **INCOME**.
- Create a format for the new ordered variable.
- Produce a frequency table using PROC FREQ.

Create a new variable named **INCLEVEL** so that the sort order corresponds to its logical order.

```
/*st008d02*/
data st092.sales_inc;
  set st092.sales;
  if Income='Low' then IncLevel=1;
  else If Income='Medium' then IncLevel=2;
  else If Income='High' then IncLevel=3;
run;
```

Use PROC FORMAT to create user-defined formats.

```
proc format;
  value incfmt 1='Low Income'
            2='Medium Income'
            3='High Income';
run;
```

Use PROC FREQ with a FORMAT statement.

```
ods graphics on;
ods rtf style=statistical
  file='freq2.rtf';
proc freq data=st092.sales_inc;
  tables IncLevel*Purchase;
  format IncLevel incfmt. Purchase purfmt.;
  title1 'Create variable IncLevel to correct Income';
run;
ods rtf close;
```



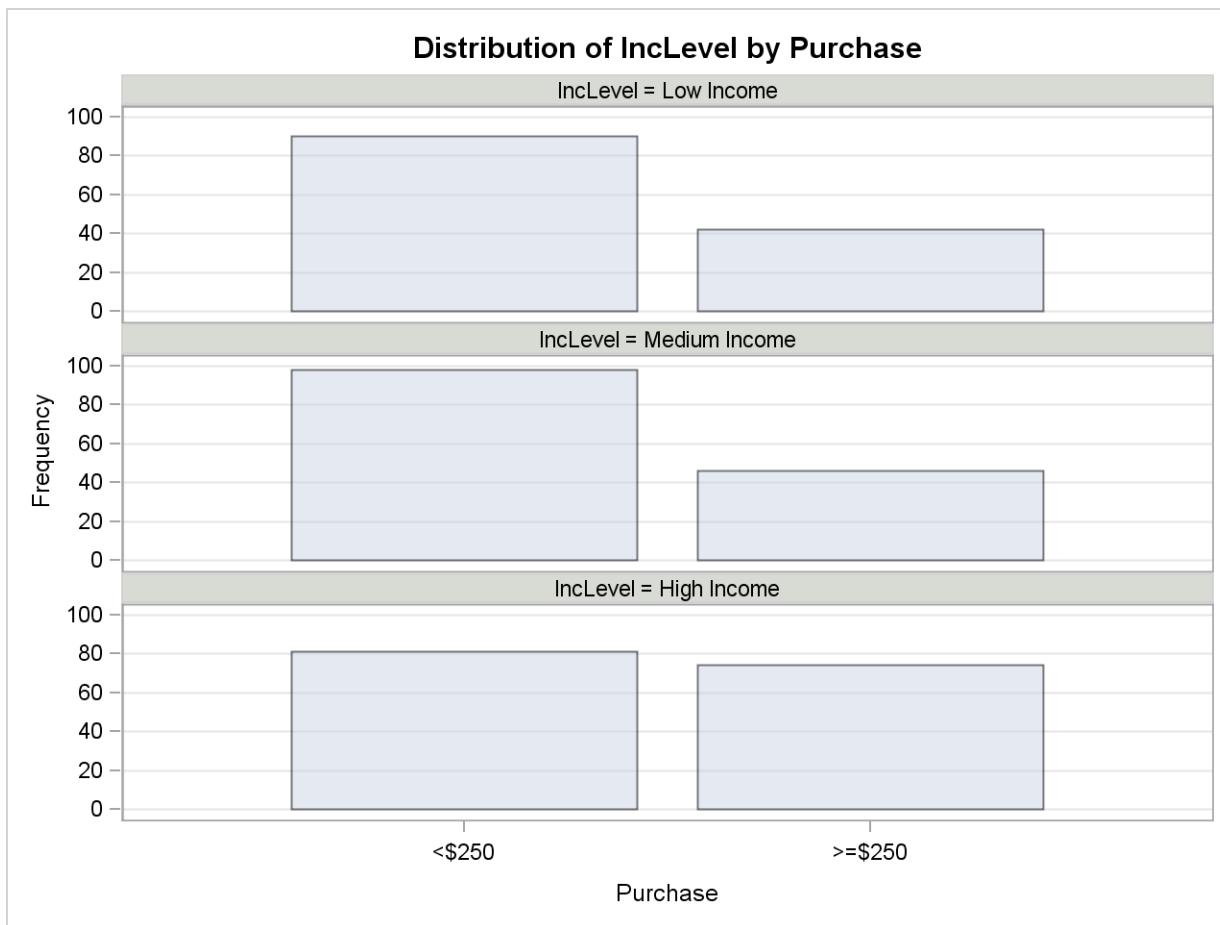
If your data is in a logical order in a data set, you can use the ORDER=DATA option in PROC FREQ.

The crosstabulation of **INCLEVEL * PURCHASE** is shown below. The values of **INCLEVEL** are now in a logical order.

Table of IncLevel by Purchase			
IncLevel	Purchase		
Frequency			
Percent			
Row Pct			
Col Pct	<\$250	>=\$250	Total
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

By examining the row percentages, you see that **PURCHASE** is associated with **INCOME**. For example, 48% of the high-income customers made purchases of 250 dollars or more compared to 32% of the low-income customers and 32% of the medium-income customers.

These percentages are shown graphically in the Frequency Plot.



8.2 Tests of Association

Objectives

- Perform a chi-square test for association.
- Examine the strength of the association.
- Perform a Mantel-Haenszel chi-square test.

25

Introduction

Table of Gender by Purchase		
Gender	Purchase	
Row Pct	< \$250	>=\$250
Female	57.92	42.08
Male	68.06	31.94

26

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 250 dollars or more (0.42) is significantly different from the probability of males purchasing items of 250 dollars or more (0.32).

Null Hypothesis

- There is no association between **Gender** and **Purchase**.
- The probability of purchasing items of 250 dollars or more is the same whether you are male or female.

Alternative Hypothesis

- There is an association between **Gender** and **Purchase**.
- The probability of purchasing items over 250 dollars is different between males and females.

Chi-Square Test

NO ASSOCIATION

observed frequencies = expected frequencies

ASSOCIATION

observed frequencies \neq expected frequencies

- ✍ The expected frequencies are calculated by the formula: $(\text{row total} * \text{column total}) / \text{sample size}$.

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.

- ✍ Under the null hypothesis of no association between Row and Column variable, the “expected” percentage in any $R*C$ cell will be equal to the percent in that cell’s row (R / T) times the percent in the cell’s column (C / T). The expected count is then just that expected percentage times the total sample size. So, the expected count = $(R / T) * (C / T) * T = (R * C) / T$.

Purchase & Gender

	< \$250	$\geq \$250$	Total
Female	139	101 (42%)	240
Male	130	61 (32%)	191
Total	269	162 (38%)	431

29

- ✍ In the entire data set 38% (162/431) of people purchased more than or equal to \$250

We need to calculate the number of males and females we would expect to purchase more than or equal to \$250 if there was no association between **GENDER** and **PURCHASE**.

We would expect 38% of females to purchase more than or equal to \$250 if there was no association and we would also expect 38% of males to purchase more than or equal to \$250 if there was no association.

38% of 240 Females is 90.2. This calculation can be generalised to:

Expected Frequencies

Expected frequency for females purchasing more than \$250 (assuming no association)

$$\frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{240 \times 162}{431} = 90.2$$

30

If 90.2 females are expected to purchase more than or equal to \$250 then 149.8 (240 - 90.2) females are expected not to purchase more than or equal to \$250. The values for males can also be worked out by subtraction because the totals are fixed.

Expected Frequencies

	< \$250	$\geq \$250$	Total
Female	149.8	90.2	240
Male	119.2	71.8	191
Total	269	162	431

31

Chi-Square Test

Calculate for all cells : $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

32

We now need to compare the observed frequencies, and the expected frequencies assuming no association, to see how different they are.

If we just calculate the observed minus expected ($O - E$) for each cell they will sum to zero because the totals are fixed. We therefore need a non-negative deviation and we obtain this by squaring ($O - E$).

We also need to include a scaling factor because the difference between 10 and 20 is the same number of units as the difference between 110 and 120, but in the first example this has been a doubling, whereas in the second example there is only a small percentage increase.

The chi-square statistic incorporates the square deviation and the scaling factor.

The p -value for the Chi-Square test only indicates how confident you can be that the null hypothesis of no association is false. It does not tell you the magnitude of an association. The value of the chi-square statistic also does not tell you the magnitude of the association. If you double the size of your sample by duplicating each observation, you double the value of the chi-square statistic, even though the strength of the association does not change.

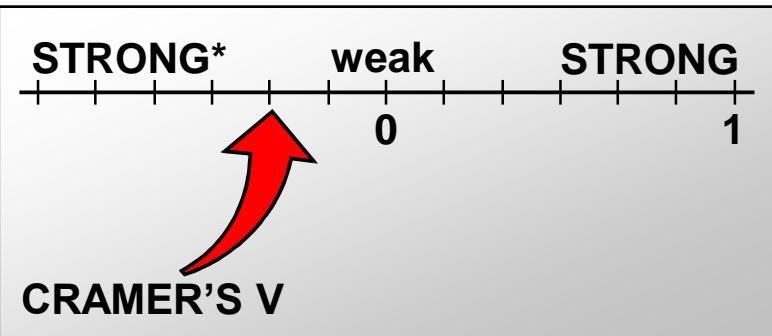
Chi-Square Tests

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.

33

Measures of Association



* Cramer's V is always nonnegative for tables larger than 2*2.

34

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

The following is a summary of what you will accomplish in this demonstration:

- 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.

```
/*st008d03*/
ods graphics off;
proc freq data=st092.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			
Gender		Purchase	
			①
Frequency			
Expected			
Cell Chi-Square			
Row Pct	<\$250	>=\$250	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

① It appears that the cell for **PURCHASE** = 1 (250 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 ②
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	

② Check the null hypothesis that there is no association between **GENDER** and **PURCHASE**.

Step 1- Set Hypothesis

H_0 : No Association exists between **GENDER** and **PURCHASE**.

H_1 : Association exists between **GENDER** and **PURCHASE**

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

p-value=0.0307.

Step 4- Decision Rule.

The p-value< α , therefore, 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.

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.

Association among Ordinal Variables

Is



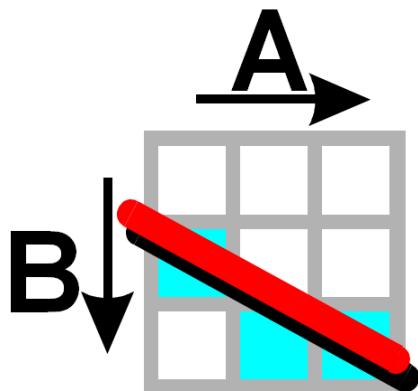
associated
with



36

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

37

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

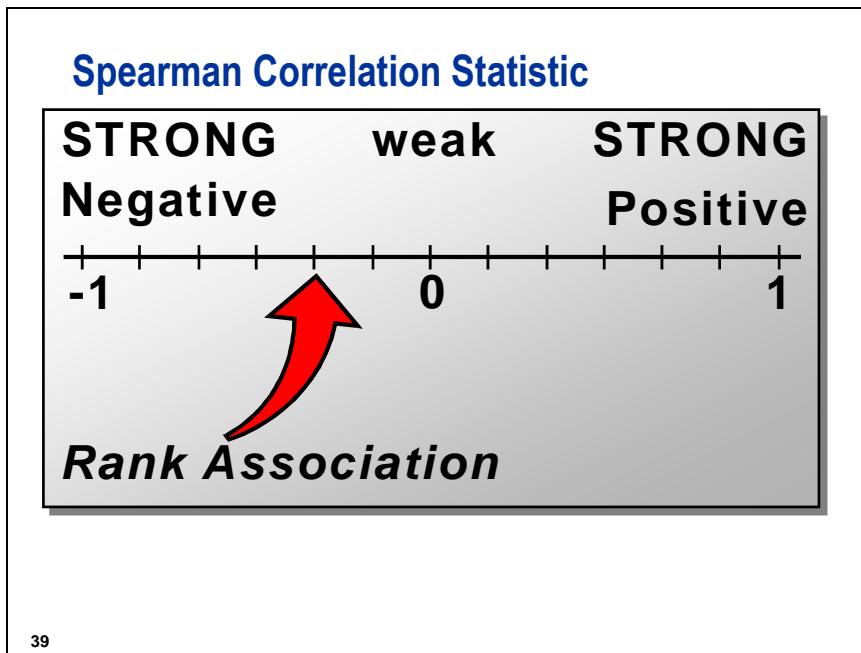
The Mantel-Haenszel chi-square test

- determines whether an ordinal association exists
- does not measure the strength of the ordinal association
- depends upon and reflects the sample size.

38

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.



39

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 ordinal scaled and the values are in a logical order.

Spearman versus Pearson

- The Spearman correlation uses ranks of the data.
- The Pearson correlation uses the observed values when the variable is numeric.

40

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

The following is a summary of what you will accomplish in this demonstration:

- 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.

```
/*st008d04*/
ods graphics off;
proc freq data=st092.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 cross tabulation is shown below.

Ordinal Association between INCLEVEL and PURCHASE?				
The FREQ Procedure				
Table of IncLevel by Purchase				
IncLevel		Purchase		
Frequency	Percent			①
Row Pct	Col Pct	<\$250	>=\$250	Total
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

- ① Confirm the frequency is correct.

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	②
Phi Coefficient		0.1571		
Contingency Coefficient		0.1552		
Cramer's V		0.1571		

② Step 1- Set Hypothesis

H_0 : No Ordinal Association exists between **PURCHASE** and **INCLEVEL**.

H_1 : An Ordinal Association exists between **PURCHASE** and **INCLEVEL**.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

p-value=0.0044

Step 4- Decision Rule.

The p-value < α , Reject H_0 , therefore, you can conclude at the 0.05 significance level that there is evidence of an ordinal association between **PURCHASE** and **INCLEVEL**.

The Spearman correlation statistic and the 95% confidence bounds are shown below.

Statistic	Value	ASE	95% Confidence Limits	
			Lower Limit	Upper Limit
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 ③
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				

③ The Spearman Correlation (0.1391) indicates that there is a relatively small positive ordinal relationship between **INCLEVEL** and **PURCHASE** (as **INCLEVEL** levels increase, **PURCHASE** tends to 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.

Chapter 9 Logistic Regression

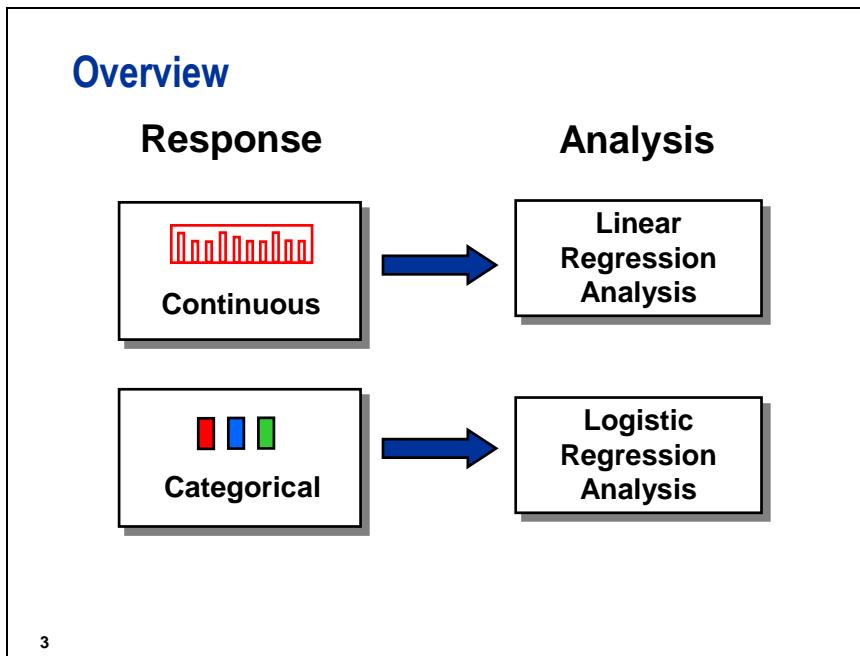
9.1 Introduction to Logistic Regression.....9-3

Demonstration: Binary Logistic Regression 9-15

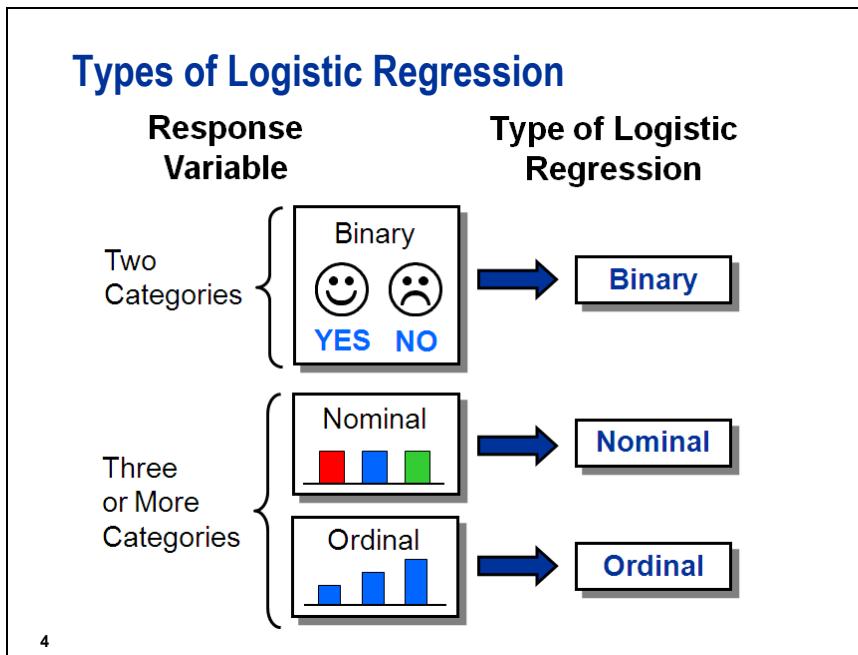
9.1 Introduction to Logistic Regression

Objectives

- Explain the concepts of logistic regression.
- Fit a binary logistic regression model using the LOGISTIC procedure.
- Explain reference cell coding for CLASS variables.
- Explain the standard output from the LOGISTIC procedure.



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.



4

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.

Why Not Ordinary Least Squares Regression?

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- If the response variable is categorical, then how do you code the response numerically?
- If the response is coded (1=Yes and 0=No) and your regression equation predicts 0.5 or 1.1 or -0.4, what does that mean practically?
- If there are only two (or a few) possible response levels, is it reasonable to assume constant variance and normality?

5

One might be tempted to analyze a regression model with a binary response variable using PROC REG. However, there are problems with that. Besides the arbitrary nature of the coding, there is the problem that the predicted values will take on values that have no intrinsic meaning, with regards to your response variable. There is also the mathematical inconvenience of not being able to assume normality and constant variance when the response variable has only two values.

What About a Linear Probability Model?

$$p_i = \beta_0 + \beta_1 X_{1i}$$

- Probabilities are bounded, but linear functions can take on any value. (Once again, how do you interpret a predicted value of -0.4 or 1.1?)
- Given the bounded nature of probabilities, can you assume a linear relationship between X and p throughout the possible range of X ?
- Can you assume a random error with constant variance?
- What is the observed probability for an observation?

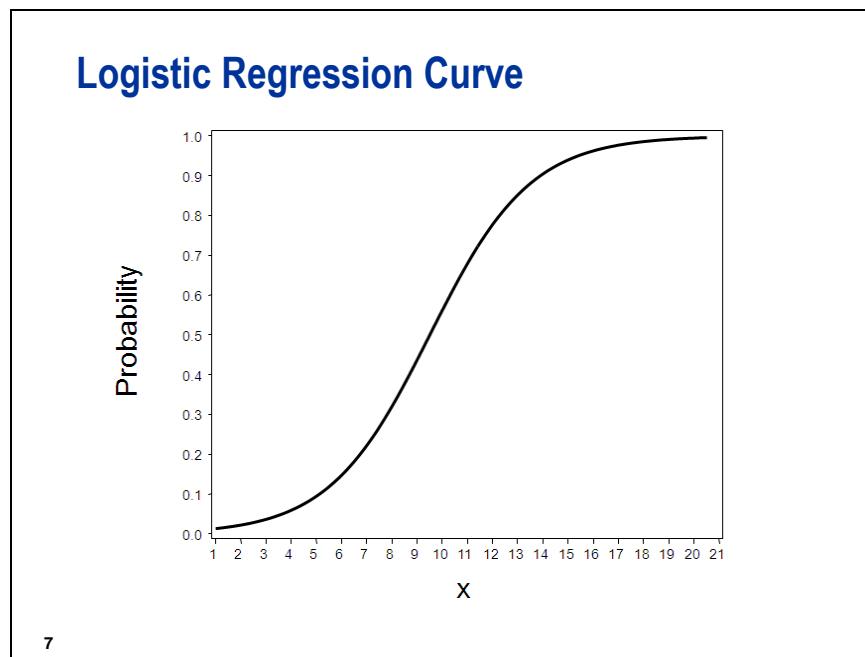
6

Instead of modeling the 0's and 1's directly, another way of thinking about modeling a binary variable is to model the probability of either the 0 or the 1. If you can model the probability of the 1 (call that p), then you have also modeled the probability of the 0, which would be $(1 - p)$. Probabilities are truly continuous and so this line of thinking might sound compelling at first.

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.

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 (a "sigmoidal" relationship).

Probabilities do not have a random normal error associated with them, but rather a binomial error of $p^*(1-p)$. That error is greatest at probabilities close to 0.5 and lowest near 0 and 1. They do not have constant error associated with them.



This plot shows a model of the relationship between a continuous predictor and the probability of an event or outcome. The linear model clearly will not fit if this is the true relationship between X and the probability. In order to model this relationship directly, you must use a nonlinear function. One such function is displayed.

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 can be represented by a straight, horizontal line that shows an equal probability of the event for everyone.

The β values cannot be computed in PROC REG. This is not a linear model.

Logit Transformation

Logistic regression models transform probabilities called logits.

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

where

- i indexes all cases (observations)
- p_i is the probability the event (a sale, for example) occurs in the i^{th} case
- \ln is the natural log (to the base e).

8

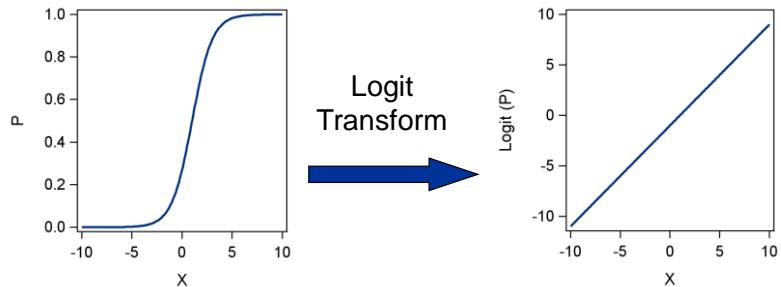
A logistic regression model applies a logit transformation to the probabilities. Two of the problems you saw with modeling the probability directly were that probabilities were bounded between 0 and 1, and that there was not likely a straight line relationship between predictors and probabilities.

First, deal with the problem of restricted range of the probability. What about the range of a logit? As p approaches its maximum value of 1, the value $\ln(p / (1 - p))$ approaches infinity. As p approaches its minimum value of 0, $p / (1 - p)$ approaches 0. The natural log of something approaching 0 is something approaching negative infinity. So, the logit has no upper or lower bounds.

If you can model the logit, then simple algebra will allow you to model the odds or the probability. The logit transformation ensures that the model generates estimated probabilities between 0 and 1.

The logit is the natural log of the odds.

Assumption



9

Assumption in logistic regression:

The logit transformation of the probabilities results in a linear relationship with the predictor variables.

If the thoughts about the nature of the direct relationship between X and p are correct, then the logit will have a straight line relationship with X . In other words, a linear function of X can be used to model the logit. In that way, you can indirectly model the probability.

To verify this assumption, it would be useful to plot the logits by the predictor variable. Logit plots are illustrated in a later section (self-study).

Logistic Regression Model

$$\text{logit } (p_i) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

where

$\text{logit } (p_i)$ = logit of the probability of the event

β_0 = intercept of the regression equation

β_k = parameter estimate of the k^{th} predictor variable

10

For a binary outcome variable, the linear logistic model with more than 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>;
  UNITS independent1=list ... </ options>;
  ODDSRATIO <'label'> variable </ options>;
  OUTPUT OUT=SAS-data-set keyword=name
        </ options>;
RUN;
```

11

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.
- UNITS enables you to obtain an odds ratio estimate for a specified change in a predictor variable. The unit of change can be a number, standard deviation (SD) or a number times the standard deviation (2*SD).
- ODDSRATIO produces odds ratios for variables even when the variables are involved in interactions with other covariates, and for classification variables that use any parameterization. You can specify several ODDSRATIO statements.

What Does a CLASS Statement Actually Do?

- PROC LOGISTIC assumes a linear relationship between predictors and the logit for the response.
 - For categorical variables, that assumption cannot be met.
- The CLASS statement creates a set of “design variables” representing the information in the categorical variables.
 - The design variables are the ones actually used in model calculations.
 - There are many different “parameterizations” of the design variables.

12

Reference Cell Coding: Two Levels

Design Variable		
<u>CLASS</u>	<u>Value</u>	<u>1</u>
Gender	Female	1
	Male	0

13

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 Female estimates the difference between Female and Male. You can choose the reference level in the CLASS statement.

Reference Cell Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Female}}$$

β_0 = the value of the logit when gender is Male

β_1 = the difference between the logits for Female and Male

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

14

The parameter estimate and *p*-value for Gender=1 reflect the difference between Gender=Female and Gender=Male (the reference level). It is important to know what type of parameterization you are using in order to interpret and report the results of this table.



Binary Logistic Regression

The following is a summary of what you will accomplish in this demonstration:

- 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 250 dollars or more and request profile likelihood confidence intervals around the estimated odds ratios.

```
/*st009d01*/
ods graphics on;
proc logistic data=st092.sales_inc plots(only)=(effect
oddsratio);
  class Gender (param=effect 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. 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. The CLODDS= option also enables production of the ODDSRATIO plot.

Selected PLOTS= options:

EFFECT If you have CLASS and continuous covariates, then a plot of the predicted probability versus the first continuous covariate at up to 10 cross-classifications of the CLASS covariate levels, while fixing all other continuous covariates at their means and all other CLASS covariates at their reference levels, is displayed.

ODDSRATIO Displays plots of odds ratios and confidence limits for the model when the CLODDS= option or ODDSRATIO statements are also specified.

 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	ST092.SALES_INC	①
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

- ❶ 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 250 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

- ❷ 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 -2Log 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

③ 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).

Step 1- Set Hypothesis

H_0 : The probability of males spending more than \$250 is equal to the probability of females spending more than \$250.

H_1 : The probability of males spending more than \$250 is not equal to the probability of females spending more than \$250. Or this model is better than the baseline model.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

p-value=0.0302,

Step 4- Decision Rule.

The *p*-value< α , Reject H_0 , therefore, this model is better than assuming the probability of purchasing is equal for the levels of gender.

Type 3 Analysis of Effects ④			
Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
Gender	1	4.6436	0.0312

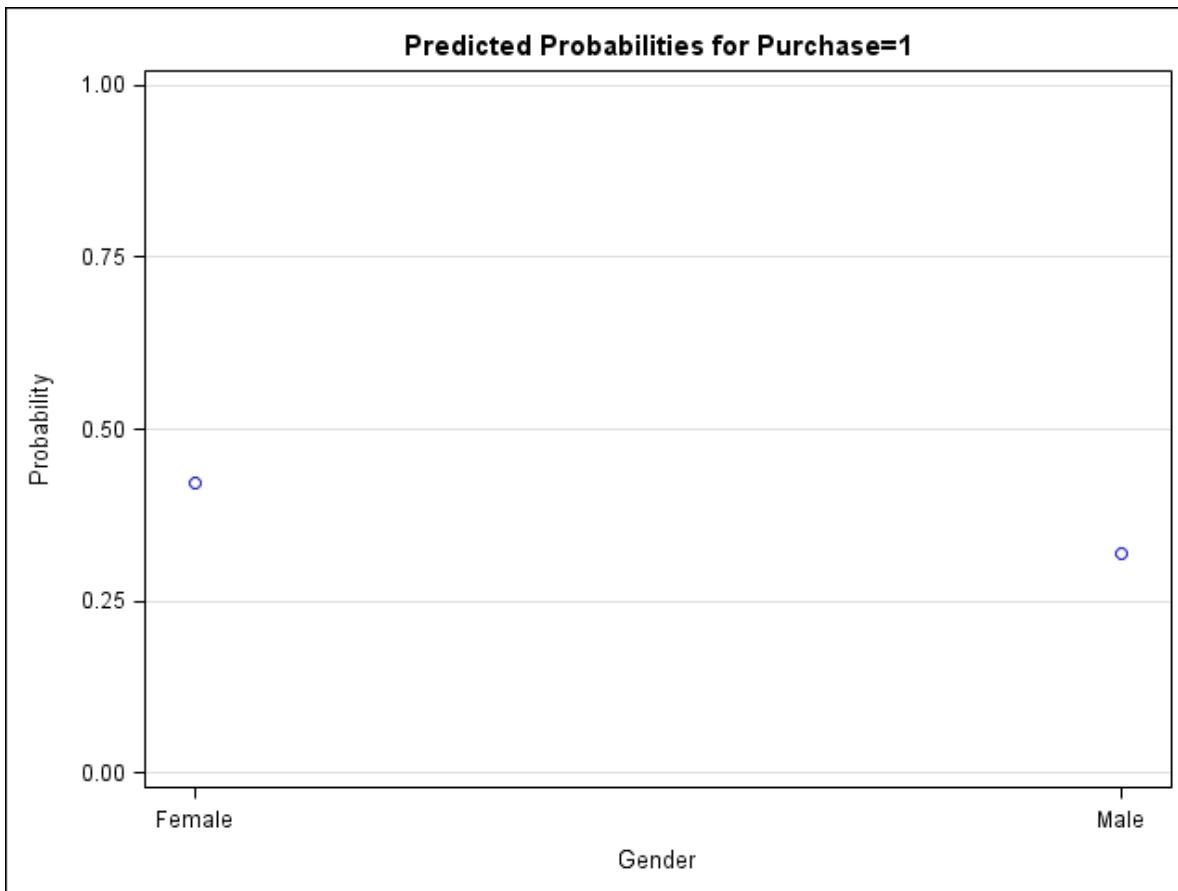
④ The Type 3 Analysis of Effects table is generated when a predictor variable is used in the CLASS statement. The listed effect (variable) is tested using the Wald Chi-Square statistic (in this example, 4.6436 with a *p*-value of 0.0312). This analysis is similar to the individual *t*-test in the REG 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 ⑤					
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

⑤ 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.



The Effect plot shows the difference between levels of the CLASS predictor variable on the probability scale.

Odds Ratios

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 do the odds of purchasing 250 dollars or more in items by males compare to those of females?

$$\text{Odds} = \frac{p_{\text{event}}}{1 - p_{\text{event}}}$$

16

The odds ratio can be used a measure the strength of association for 2 * 2 tables.

Do not mistake odds for probability. Odds are calculated from probabilities as shown in the next slides.

Probability versus Odds of an Outcome

		Outcome		Total
		No	Yes	
Group A	20	60		80
	10	90		100
Total	30	150		180

Probability of a **Yes outcome**
in Group B = 90/100 (**0.90**)

Probability of a **No outcome**
in Group B = 10/100 (**0.10**)

17

There is a 90% probability of having the outcome in group B.

What is the probability of having the outcome in group A?

Odds

Odds of Outcome in Group B

$$\frac{\text{Probability of a Yes outcome in Group B}}{\text{Probability of a No outcome in Group B}}$$

$$0.90 \div 0.10 = 9$$

18

The odds of an outcome are the ratio of the expected probability that the outcome will occur to the expected probability that the outcome will **not** occur. The odds for group B are 9 indicating that you expect 9 times as many occurrences as non-occurrences in group B.

What are the odds of having the outcome in group A?

Odds Ratio

Odds Ratio of Group B to Group A

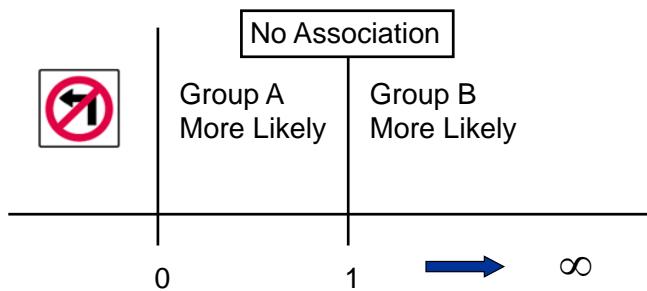
$$\frac{\text{Odds of outcome in Group B}}{\text{Odds of outcome in Group A}}$$

$$9 \div 3 = 3$$

19

The odds ratio of group B to A equals 3, indicating that the odds of getting the outcome in group B are 3 times those in group A

Properties of the Odds Ratio, B to A



20

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 B is more likely to have the outcome. If the odds ratio is less than 1, then group A 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):

$$\text{odds}_{\text{females}} = e^{\beta_0 + \beta_1}$$

$$\text{odds}_{\text{males}} = e^{\beta_0}$$

$$\text{odds ratio} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

21

Remember that in logistic regression you model the natural log of the odds and not the odds or probability directly. For interpretation, often the parameter estimates are converted into something more interpretable – an odds ratio. In order to understand this, write out the linear model predicting the natural log of the odds. In order to see that in terms of odds, the natural log is “undone” by exponentiation. Exponentiation of the right side of the equation must also be done to maintain equality. You thereby can look at the model in terms of odds and can estimate odds for females or males. The odds ratio is then the ratio of the odds of one group to the odds of another group.

The odds ratio reported by PROC LOGISTIC is for a 1-unit difference for a variable. Because you used reference cell coding for **GENDER** and used Male as the reference level, females are coded 1 and males are coded 0. Therefore, a 1-unit increase in **GENDER** corresponds to the difference between females and males.

LOGISTIC MODEL (1):Purchase=Gender

6

The LOGISTIC Procedure

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits
Gender	Female vs Male	1.549	1.040 2.305

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 Odds Ratios				
Effect	Unit	Estimate	95% Confidence Limits	
gender	1.0000	1.549	1.043	2.312

22

- ⑥ The odds ratio indicates that females have 1.55 times the odds to purchase 250 dollars or more, relative to 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.

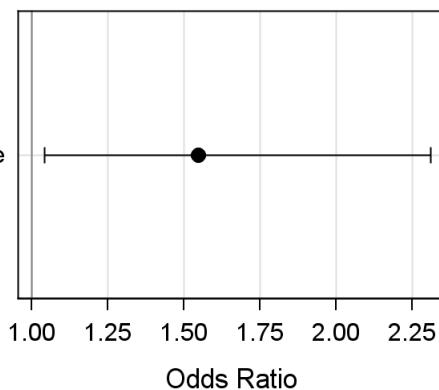
-  If you want a different significance level for the confidence intervals, you can use the ALPHA= option in the MODEL statement. The value must be between 0 and 1. The default value of .05 results in the calculation of a 95% confidence interval.

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 Plot

Odds Ratios with 95% Profile-Likelihood Confidence Limits

Gender Female vs Male



23

The odds ratio plot displays the odds ratio and confidence interval based on the method chosen in the CLODDS= option of the MODEL statement.

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.

24

Concordant versus Discordant

Customer Purchasing \$250 or More			
Customer Purchasing Less Than \$250	Predicted Outcome Probability	Females (0.42)	Males (0.32)
	Females (0.42)	Tie	Discordant Pair
	Males (0.32)	Concordant Pair	Tie

25

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.

A pair is *discordant* if the observation with the outcome has a **lower** predicted outcome probability than the observation without the outcome.

A pair is *tied* if it is neither concordant nor discordant (the probabilities are the same).

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 250 dollars or more (Female=0.42 and Male=0.32). For all pairs of observations with different outcomes (making purchases of 250 dollars or more versus making purchases of less than 250 dollars), a comparison is made of the predicted outcome probabilities. If the observation with the outcome (in this case making purchases of 250 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

Association of Predicted Probabilities and Observed Responses 7

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

7 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 250 dollars or more and 269 observations with an outcome of Under 250 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.

Chapter 10 Multiple Logistic Regression

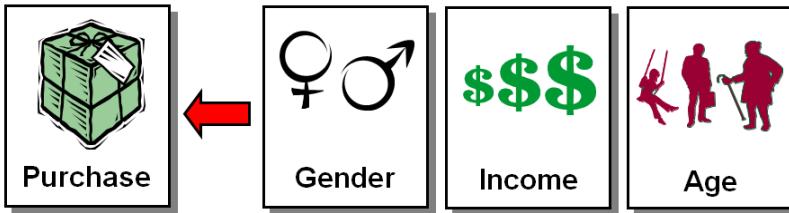
10.1 Multiple Logistic Regression	10-3
Demonstration: Multiple Logistic Regression.....	10-7
10.2 Multiple Logistic Regression with Interactions (Optional).....	10-15
Demonstration: Multiple Logistic Regression with Interactions	10-19
10.3 Logit Plots (Self-Study)	10-30
Demonstration: Plotting Estimated Logits	10-34

10.1 Multiple Logistic Regression

Objectives

- Define and explain the adjusted odds ratio.
- Fit a multiple logistic regression model using the backward elimination method.

Multiple Logistic Regression

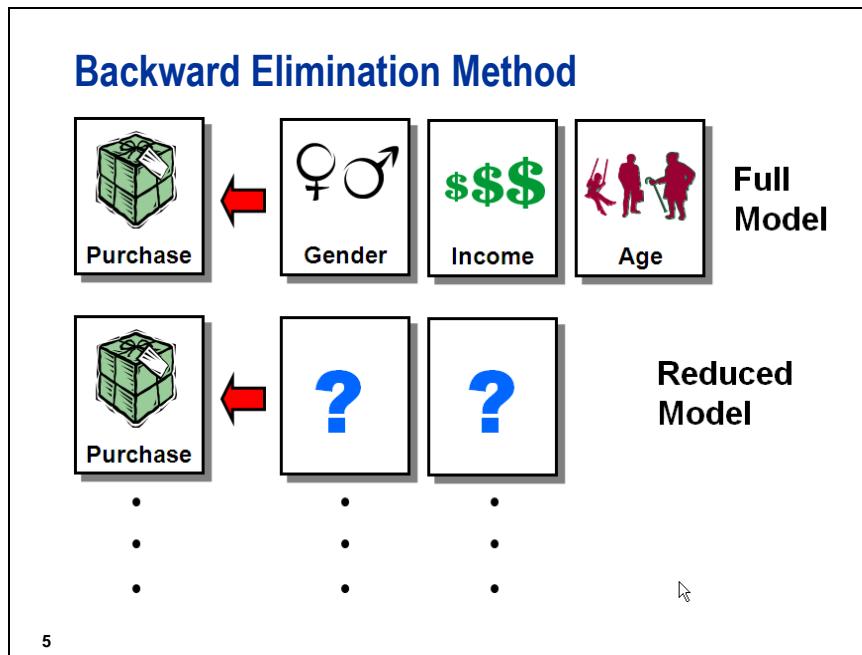


$$\text{logit}(p) = \beta_0 + \beta_1 X_{\text{Female}} + \beta_2 X_{\text{High}} + \beta_3 X_{\text{Medium}} + \beta_4 X_{\text{Age}}$$

4

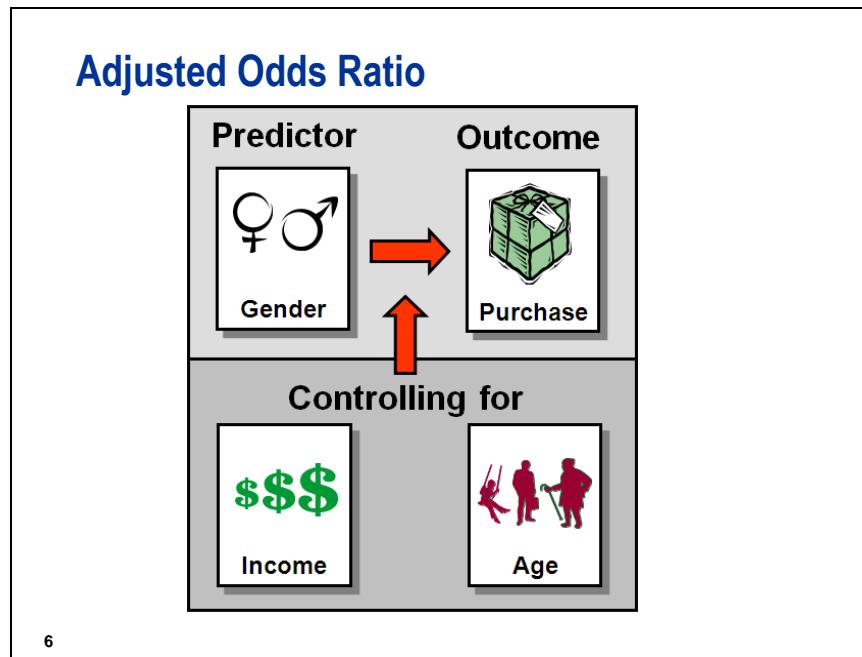
In multiple logistic regression models, several continuous or categorical predictor variables are trying to explain the variability of the response variable. The goal in multiple logistic regression is similar to that in linear multiple regression. Find the best subset of variables by eliminating unnecessary 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.



6

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

The following is a summary of what you will accomplish in this demonstration:

- Fit a multiple logistic regression model using the backward elimination method. The full model should include all the main effects.

```
/*st010d01*/
ods graphics on;
proc logistic data=st092.sales_inc plots(only)=(effect
oddsratio);
  class Gender (param=ref ref='Male')
    Income (param=ref ref='Low');
  units Age=10;
  model Purchase(event='1')=Gender Age Income /
    selection=backward clodds=pl;
  title1 'LOGISTIC MODEL (2): Purchase=Gender Age Income';
run;
ods graphics off;
```

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. 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

UNITS requests that the odds ratio for a continuous variable is calculated for a change of n units instead of the default of 1. For a categorical variable, a value of -1 inverts the odds ratio.

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 ST192.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 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 Model Fit Statistics and Testing Global Null Hypothesis tables are presented.

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	554.401
SC	576.715	574.732
-2 Log L	570.649	544.401

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	26.2480	4	<.0001
Score	25.7528	4	<.0001
Wald	24.3206	4	<.0001

NOTE: No (additional) effects met the 0.05 significance level for removal from the model.

The note that says that no additional effects met the 0.05 significance level for removal from the model indicates that all of the variables at Step 0 had *p*-values below the default SLSTAY criterion of 0.05.

The LOGISTIC Procedure

Type 3 Analysis of Effects

Effect	DF	Chi-Square	Pr > ChiSq
Gender	1	6.0563	0.0139
Age	1	9.5102	0.0020
Income	2	13.0023	0.0015

The Type 3 Analysis of Effects table for this model indicates that the coefficients for **GENDER**, **AGE** and **INCOME** are statistically different from 0 at the 0.05 level of significance. Note that **INCOME** has two degrees of freedom because it is categorical and has three levels.

Analysis of Maximum Likelihood Estimates					
Parameter		DF	Estimate	Standard Error	Wald Chi-Square
Intercept		1	-3.3071	0.7589	18.9930
Gender	Female	1	0.5204	0.2115	6.0563
Age		1	0.0560	0.0182	9.5102
Income	High	1	0.8186	0.2556	10.2523
Income	Medium	1	0.1064	0.2656	0.1605
					0.6887

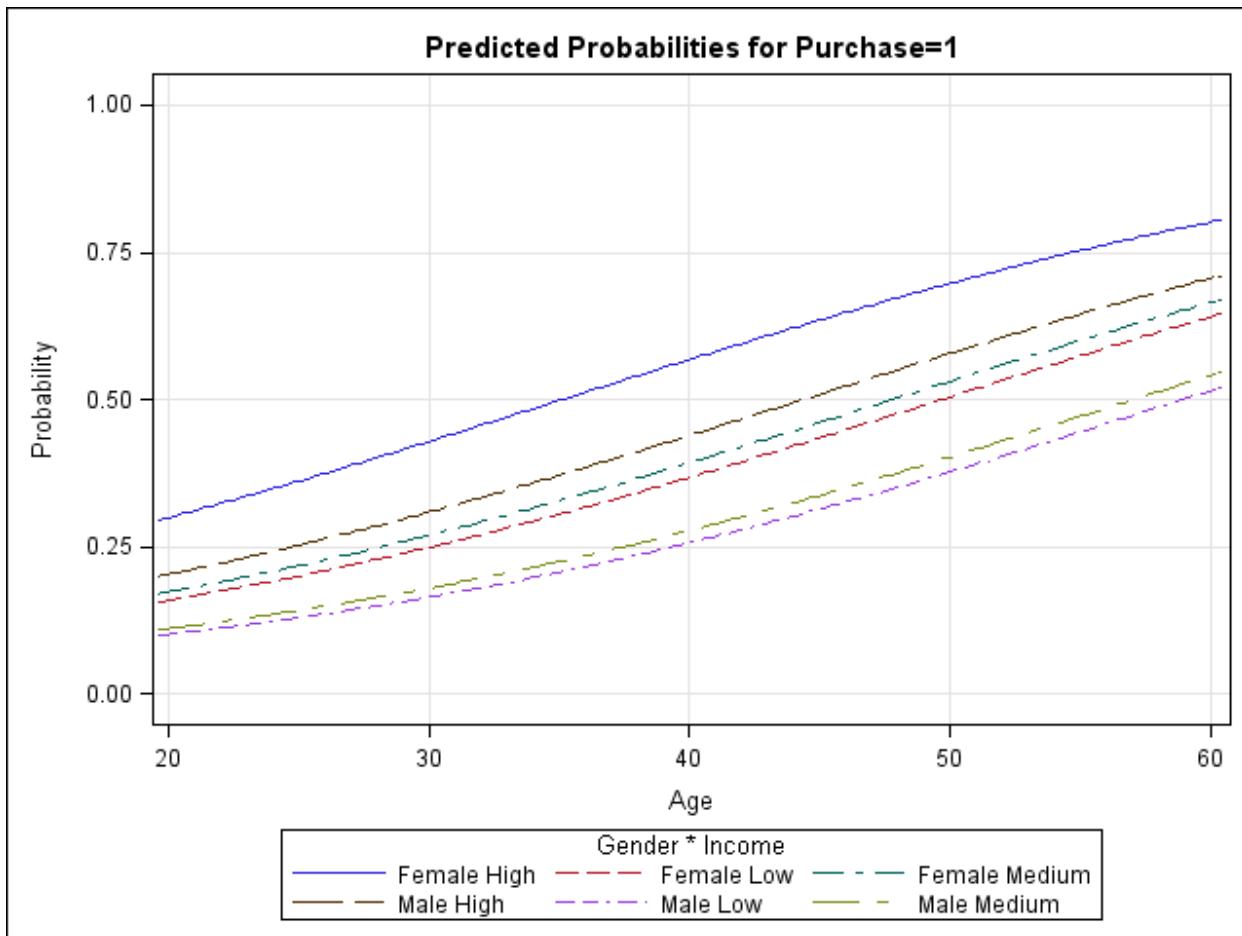
The Analysis of Maximum Likelihood Estimates table is now examined. The *p*-value for **GENDER** Female (0.0139) indicates that its coefficient is statistically different from 0 at the 0.05 level of significance. In other words, females and males are statistically different from one another in terms of purchasing 250 dollars or more of product from the catalog company.

The positive estimate (0.0560) and significant *p*-value (0.0020) for **AGE** shows that older people have a greater tendency to buy more than 250 dollars worth of product than do younger people.

The coefficient for **INCOME** =High is also statistically different from 0, based on its *p*-value (0.0014). 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 odds of purchasing 250 dollars or more. When examining **INCOME** =Medium, the *p*-value of 0.6887 indicates that this coefficient is not statistically 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 odds. This result is not surprising, given the cell values in the **INCOME** ***PURCHASE** crosstabulation..

- ✍ 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 for **HighInc**.

The EFFECT plot shows the relationship of **AGE** with the probability for **PURCHASE**, plotted at each level of **GENDER** and **INCOME**.



The positive trend for **AGE** is clearly seen here, with probability values ranging from about 0.10 to 0.30 at age 20 to about 0.50 to 0.80 at age 60.

The next part of the output provides the Odds Ratio Estimates table.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gender Female vs Male	1.683	1.112	2.547
Age	1.058	1.021	1.096
Income High vs Low	2.267	1.374	3.742
Income Medium vs Low	1.112	0.661	1.872

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 (0.661, 1.872) includes 1.000.

The next to last part of the output shows the table of Predicted Probabilities and Observed responses.

Association of Predicted Probabilities and Observed Responses

Percent Concordant	63.2	Somers' D	0.273
Percent Discordant	35.9	Gamma	0.276
Percent Tied	1.0	Tau-a	0.128
Pairs	43578	c	0.637

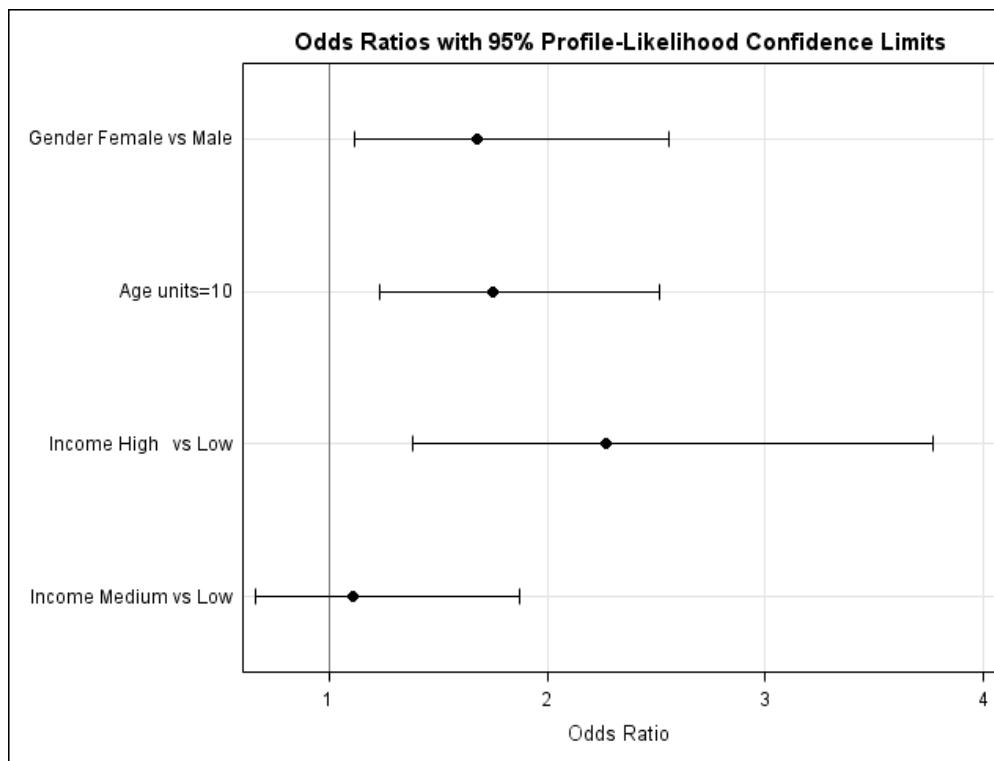
When you compare the percentages of this model with the previous model where **GENDER** was the only predictor variable, the concordant percentage increased (from 30.1 to 63.2), but the discordant percentage also increased (from 19.5 to 35.9). The tied percentage showed the most change, decreasing from 50.4 to 1.0. Tied pairs become rare when continuous variables are included in a model.

The *c* statistic increased (0.553 to 0.637) from the simple **GENDER** model – a desirable effect.

The last part of the output shows the odds ratios and profile likelihood confidence intervals. The ODDSRATIO plot provides a visual representation of those estimates and confidence intervals. Recall that the UNITS statement requested that the odds ratio for **AGE** calculate values for 10 year differences for the table and plot.

Profile Likelihood Confidence Interval for Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
Gender Female vs Male	1.0000	1.683	1.115	2.557
Age	10.0000	1.752	1.231	2.515
Income High vs Low	1.0000	2.267	1.381	3.767
Income Medium vs Low	1.0000	1.112	0.661	1.877



The only confidence interval that crosses the equality reference line of 1 is for the Income Medium vs. Income Low design variable.

Comparing Models

<i>Gender Only</i>	
AIC	569.951
SC	578.084
-2 Log L	565.951
c	0.553

<i>Gender + Age + Income</i>	
AIC	554.401
SC	574.732
-2 Log L	544.401
c	0.637

8

Adding **INCOME** to the model decreased AIC and increased value of the c statistic. The SC decreased slightly. The model with **AGE** and **INCOME** seems improved over the one with **GENDER** only.

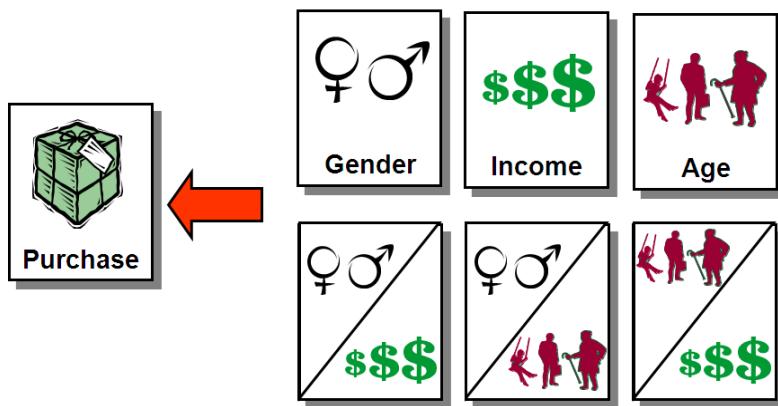
10.2 Multiple Logistic Regression with Interactions (Optional)

Objectives

- Define and explain Logistic Regression with Interactions

10

Multiple Logistic Regression

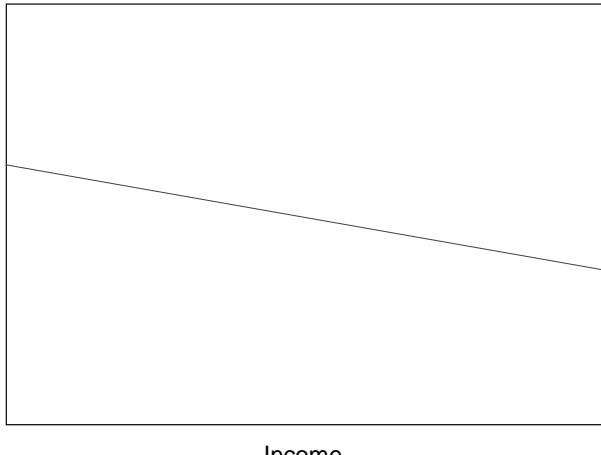


11

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 250 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.

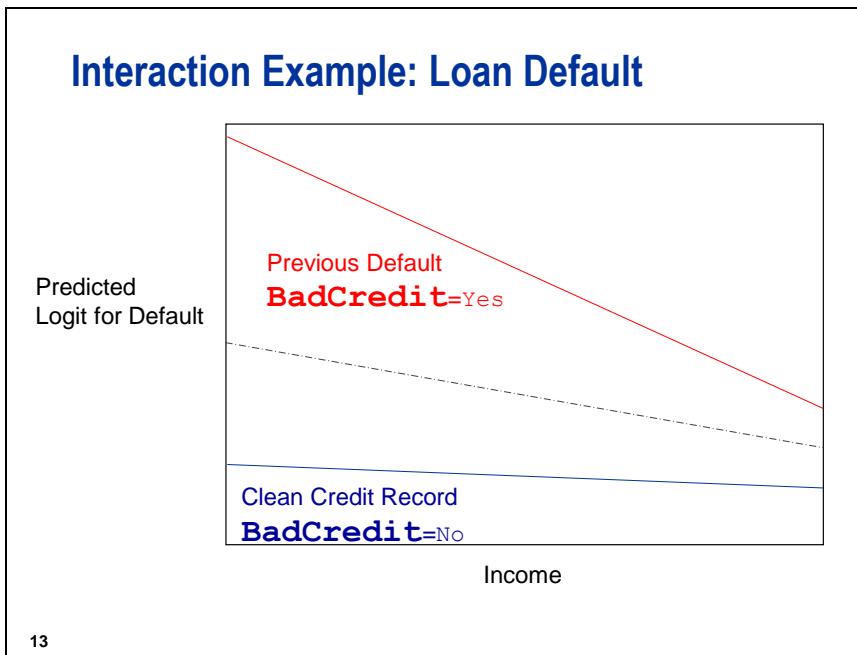
Interaction Example: Loan Default

Predicted
Logit for Default



12

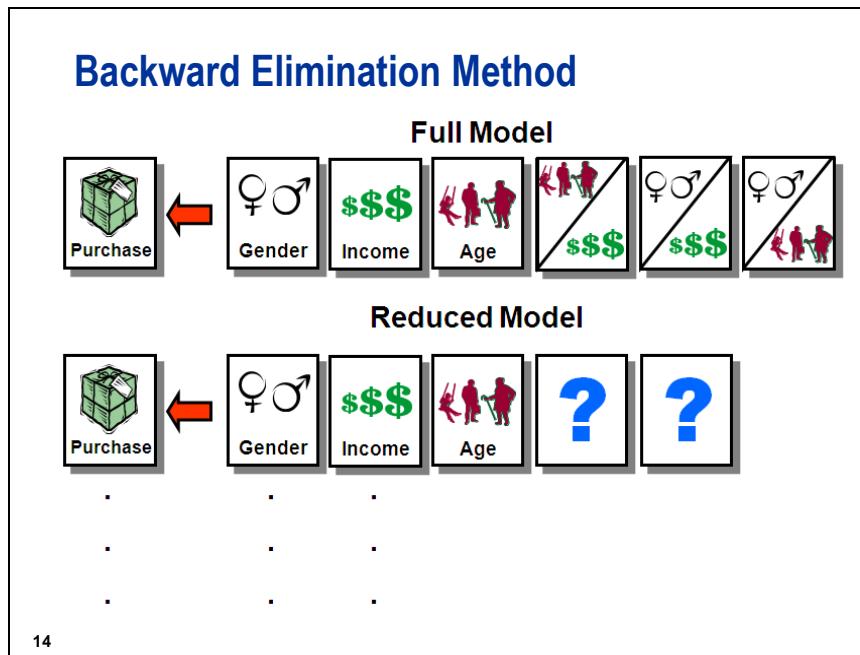
Assume that one dollar of income has the same effect for all potential customers.



However, if you consider the previous credit experience of the customer, there seems to be a great difference in the effect of income on loan default between those with good credit histories and those who have previously defaulted. 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 **INCOME** depends on the level of **BadCredit**. For **BadCredit=Yes**, as **INCOME** increases, the probability of defaulting decreases sharply. However, for **BadCredit=No**, as **INCOME** increases, the probability of defaulting does not change markedly (and remains consistently relatively low).

Therefore, there is an **INCOME** by **BadCredit** interaction in predicting the probability of default.



14

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

The following is a summary of what you will accomplish in this demonstration:

- 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.

```
/*st010d02*/
ods graphics on;
proc logistic data=st092.sales_inc plots(only)=(effect
oddsratio);
  class Gender (param=ref ref='Male')
    Income (param=ref ref='Low');
  model Purchase(event='1')=Gender|Age|Income @2/
    selection=backward clodds=pl;
  units Age=10;
  title1 'LOGISTIC MODEL (3): main effects and 2-way
interactions';
  title2 '/ sel=backward';

run;
ods graphics off;
```



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	ST092.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	0
	Low	0	0
	Medium	0	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	553.069	
SC	576.715	593.730	
-2 Log L	570.649	533.069	
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	37.5804	9	<.0001
Score	35.5320	9	<.0001
Wald	31.7454	9	0.0002

Step 1. Effect Age*Income is removed:

The LOGISTIC Procedure			
Model Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied.			
Model Fit Statistics			
Criterion	Intercept Only	Intercept and Covariates	
AIC	572.649	550.362	
SC	576.715	582.891	
-2 Log L	570.649	534.362	
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	36.2874	7	<.0001
Score	33.4793	7	<.0001
Wald	30.0042	7	<.0001
Residual Chi-Square Test			
Chi-Square	DF	Pr > ChiSq	
1.2852	2	0.5259	

The Residual Chi-Square test is performed at each step and shows the significance of the difference between the model at this step and the fullest possible model (from the model statement).

PROC LOGISTIC Output (Continued)

Step 2. Effect Age*Gender is removed:

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	572.649	550.334
SC	576.715	578.796
-2 Log L	570.649	536.334

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.3157	6	<.0001
Score	32.4897	6	<.0001
Wald	30.1705	6	<.0001

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
3.2267	3	0.3580

NOTE: No (additional) effects met the 0.05 significance level for removal from the model.

PROC LOGISTIC Output (Continued)

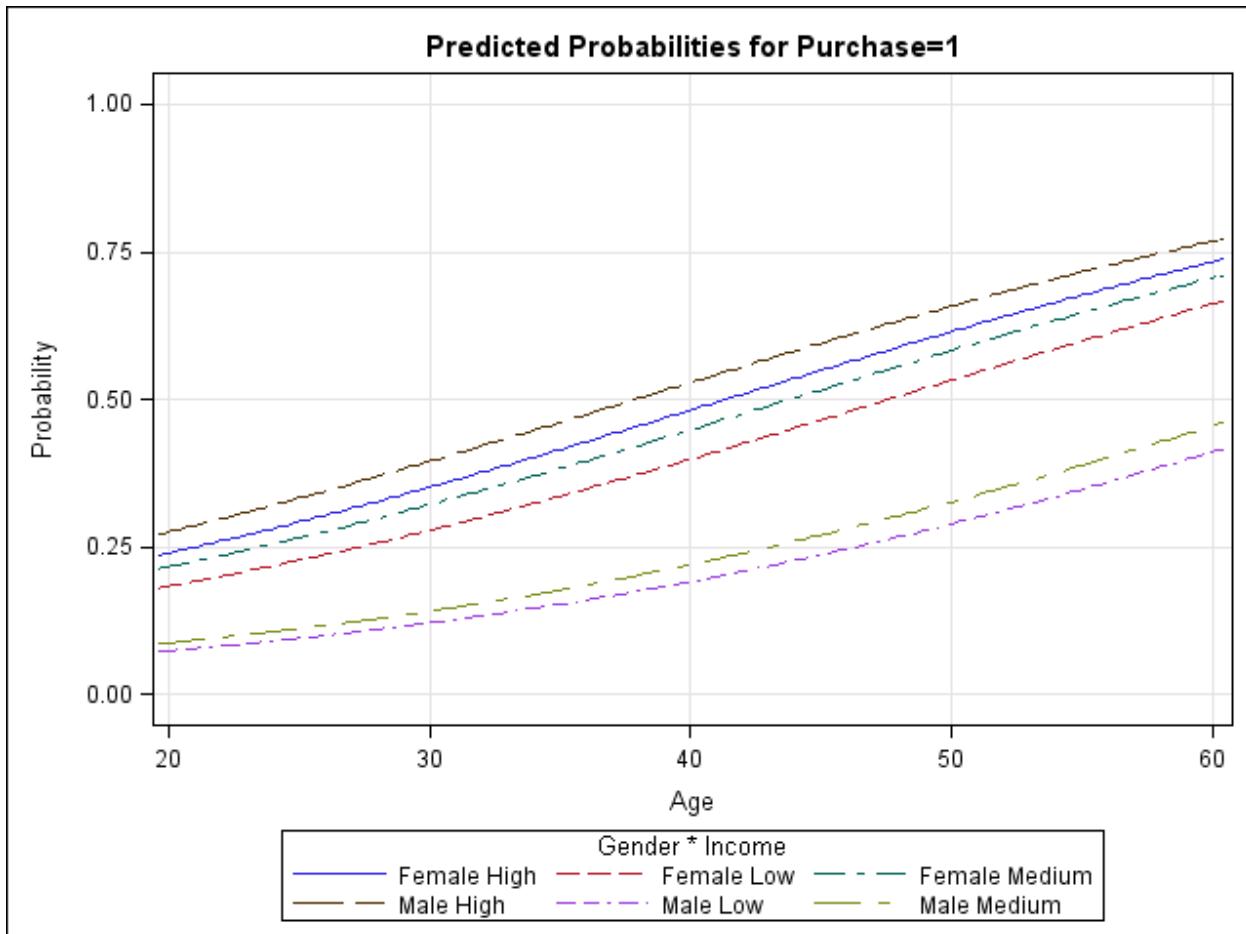
LOGISTIC MODEL (3): main effects and 2-way interactions / sel=backward					
The LOGISTIC Procedure					
Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Age*Income	2	5	1.2789	0.5276
2	Age*Gender	1	4	1.9397	0.1637

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	5.1612	0.0231
Age	1	8.6169	0.0033
Income	2	19.0440	<.0001
Gender*Income	2	7.9368	0.0189

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, it could have been dropped from the model at Step 3, but its *p*-value was smaller than 0.05, so it remained.

Analysis of Maximum Likelihood Estimates					
Parameter		DF	Estimate	Standard Error	Wald Chi-Square
Intercept		1	-3.6026	0.8331	18.6985
Gender	Female	1	1.0286	0.4528	5.1612
Age		1	0.0540	0.0184	8.6169
Income	High	1	1.5547	0.4595	11.4449
Income	Medium	1	0.1756	0.4913	0.1278
Gender*Income	Female High	1	-1.2133	0.5579	4.7298
Gender*Income	Female Medium	1	0.0295	0.5904	0.0025
					0.9602

The effect plot is shown below. It is slightly different from the one shown previously, due to the inclusion here of the interaction term between **GENDER** and **INCOME**.



Odds Ratio Estimates

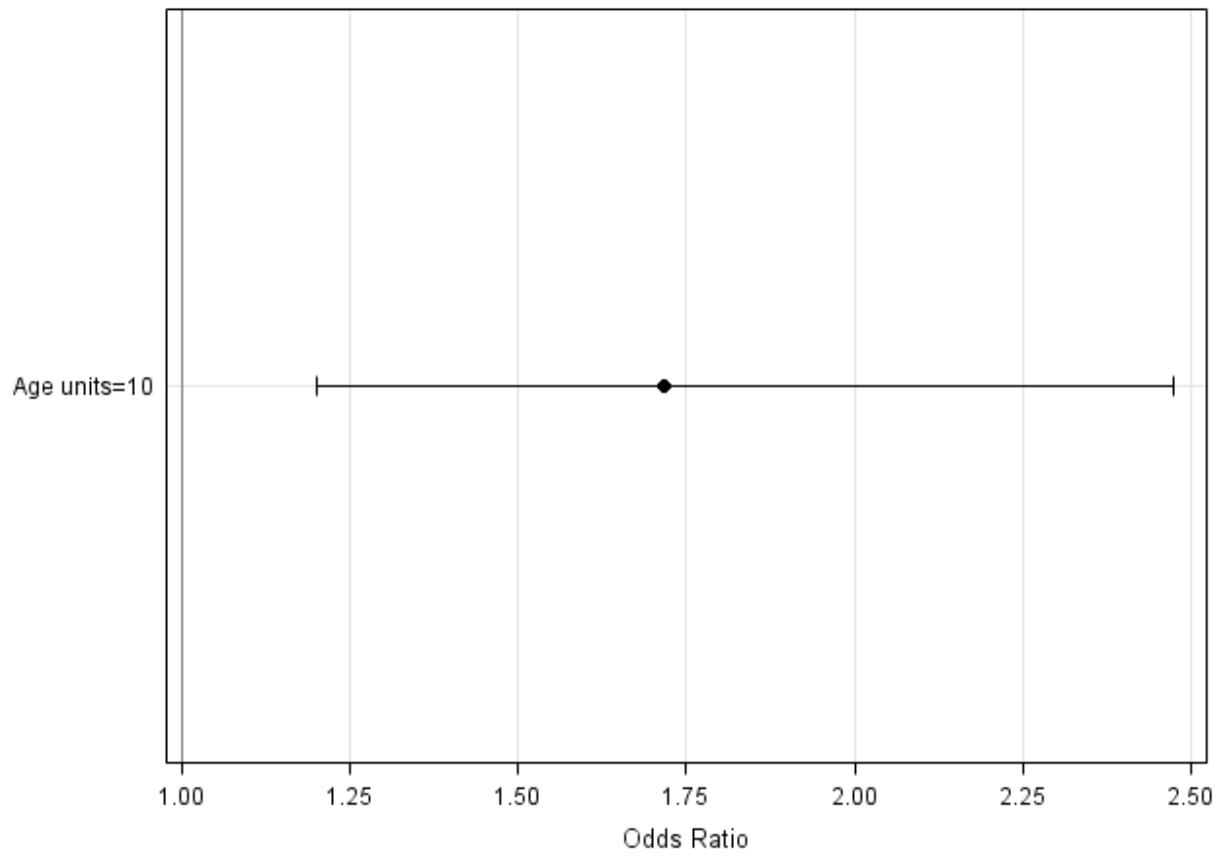
Effect	Point Estimate	95% Wald Confidence Limits
Age	1.055	1.018 1.094

Association of Predicted Probabilities and Observed Responses

Percent Concordant	64.9	Somers' D	0.308
Percent Discordant	34.0	Gamma	0.312
Percent Tied	1.1	Tau-a	0.145
Pairs	43578	c	0.654

Profile Likelihood Confidence Interval for Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits
Age	10.0000	1.716	1.201 2.475

Odds Ratios with 95% Profile-Likelihood Confidence Limits

The odds ratios are only reported for **AGE**. Odds ratios for terms that are involved in interactions are not shown or displayed in plots. The reason is that, in the presence of an interaction, the odds ratio for a main effect in that interaction would be misleading. It would only show the odds ratio for that variable, holding constant the other variable at the value 0, which may or may not even be a valid value. Remember that an interaction means that the effect of one variable differs at different levels of another variable.

One way to report and display the values of odds ratios is to calculate and plot them separately at each value of the interacting variable. This can be done using the ODDSRATIO statement.

```
/*st010d02*/
ods select OddsRatiosPL ORPlot;
proc logistic data=st092.sales_inc plots(only)=(oddsratio);
  class Gender (param=ref ref='Male')
    Income (param=ref ref='Low');
  model Purchase(event='1')=Gender|Income Age;
  units Age=10;
  oddsratio Age / cl=pl;
  oddsratio Gender / diff=ref at (Income=all) cl=pl;
  oddsratio Income / diff=ref at (Gender=all) cl=pl;
  title1 'LOGISTIC MODEL (3): main effects and 2-way
interactions';
  title2 '/ sel=backward';
run;
```

The final model from

Selected ODDSRATIO statement options:

AT(covariate=value-list | REF | ALL<...covariate=value-list | REF | ALL>)

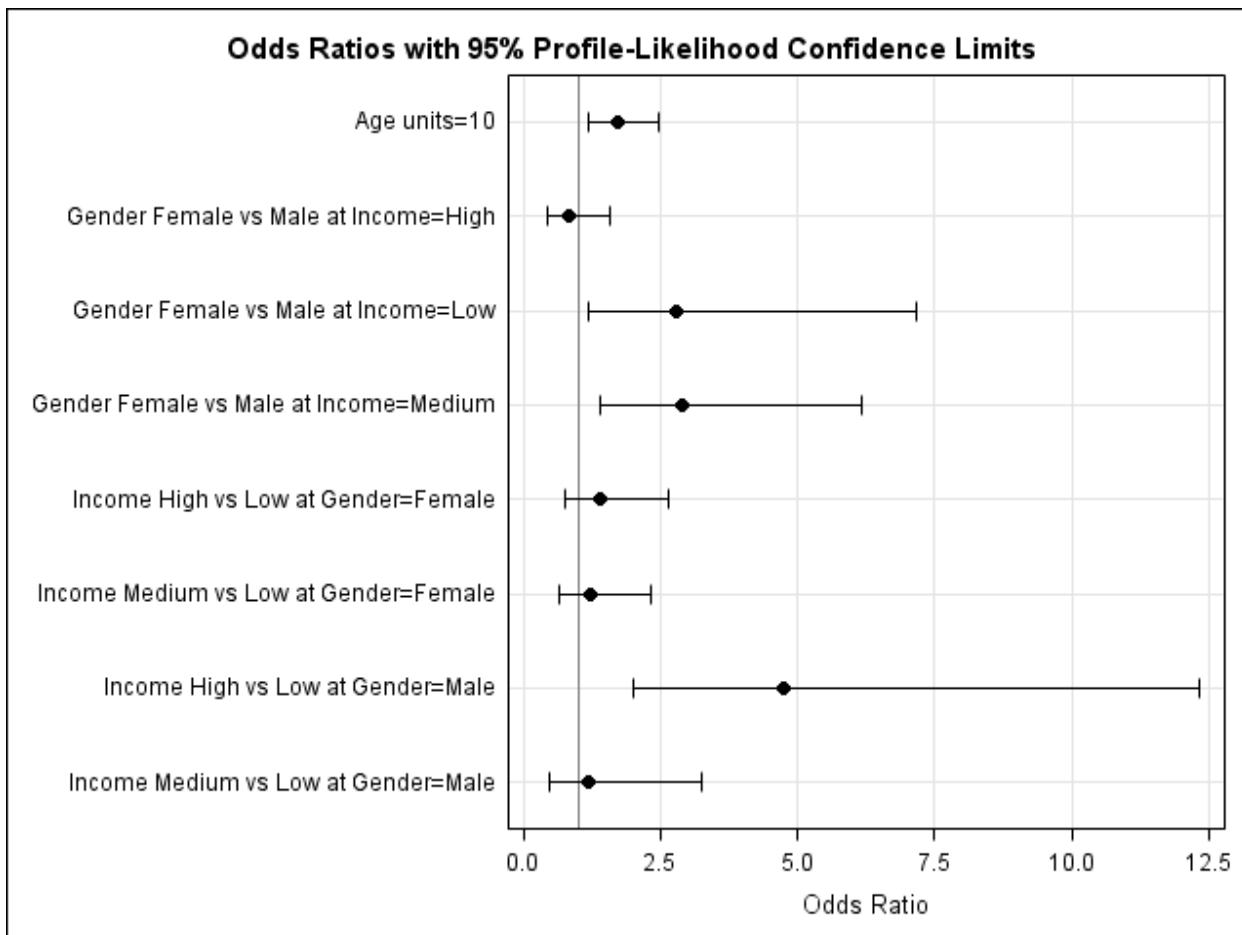
specifies fixed levels of the interacting covariates. If a specified covariate does not interact with the variable, then its AT list is ignored. If REF is chosen, the odds ratio is calculated at the reference level of the covariate. If ALL is chosen, the odds ratios are calculated at all levels of the covariate (for CLASS covariates only).

DIFF= specifies whether the odds ratios for a CLASS variable are computed against the reference level, or all pairs of variable are compared. By default, DIFF=ALL. The DIFF= option is ignored when variable is continuous.

CL= specifies how confidence limits are calculated (WALD|PL|BOTH)..

PROC LOGISTIC Output

LOGISTIC MODEL (3): main effects and 2-way interactions / sel=backward			
The LOGISTIC Procedure			
Profile Likelihood Confidence Interval for Odds Ratios			
Label	Estimate	95% Confidence Limits	
Age units=10	1.716	1.201	2.475
Gender Female vs Male at Income=High	0.831	0.437	1.577
Gender Female vs Male at Income=Low	2.797	1.195	7.181
Gender Female vs Male at Income=Medium	2.881	1.390	6.173
Income High vs Low at Gender=Female	1.407	0.755	2.635
Income Medium vs Low at Gender=Female	1.228	0.646	2.336
Income High vs Low at Gender=Male	4.733	1.998	12.314
Income Medium vs Low at Gender=Male	1.192	0.464	3.250



Females have higher odds than males at low and medium income levels, but not at high income levels. High-income males have significantly greater odds than low income males, but the income effect is not significant between any other groups for either gender. **AGE** is once again statistically significant.

Comparing Models

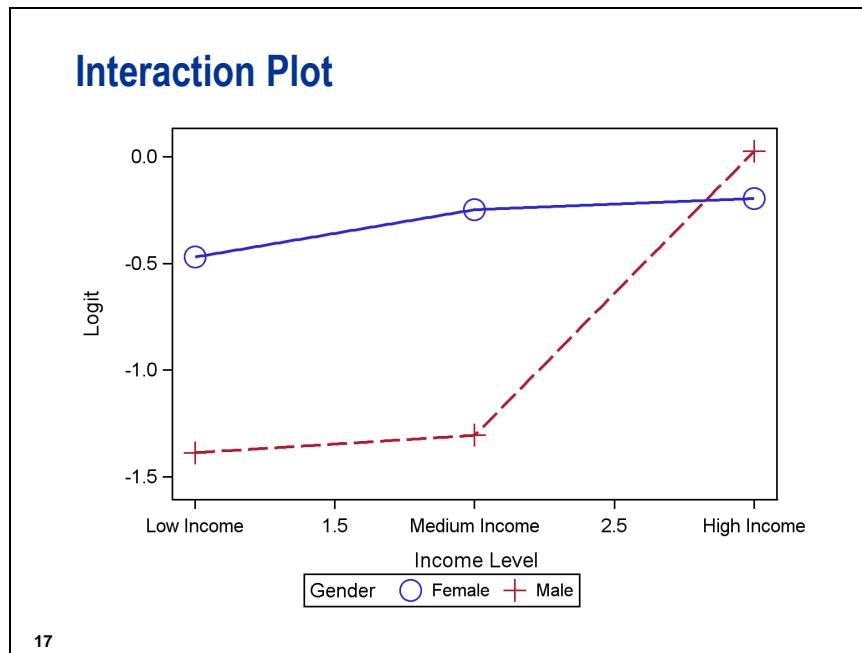
<i>Gender + Age + Income</i>	
AIC	554.401
SC	574.732
-2 Log L	544.401
c	0.637

<i>Gender, Income, Age, Gender*Income</i>	
AIC	550.334
SC	578.796
-2 Log L	536.334
c	0.654

16

AIC decreased (improved) for this model, but SC increased.

What is the purpose of your model? The “better” predictive model would include only the main effects based on the Schwarz criterion. However, using AIC, the better explanatory model would include the interaction term.



17

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 250 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 the Advanced Programs appendix.

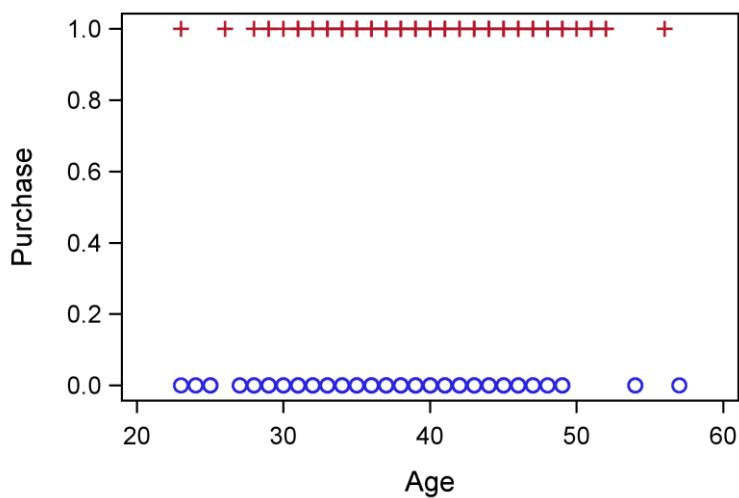
10.3 Logit Plots (Self-Study)

Objectives

- Explain the concept of logit plots.
- Plot estimated logits for continuous and ordinal variables.

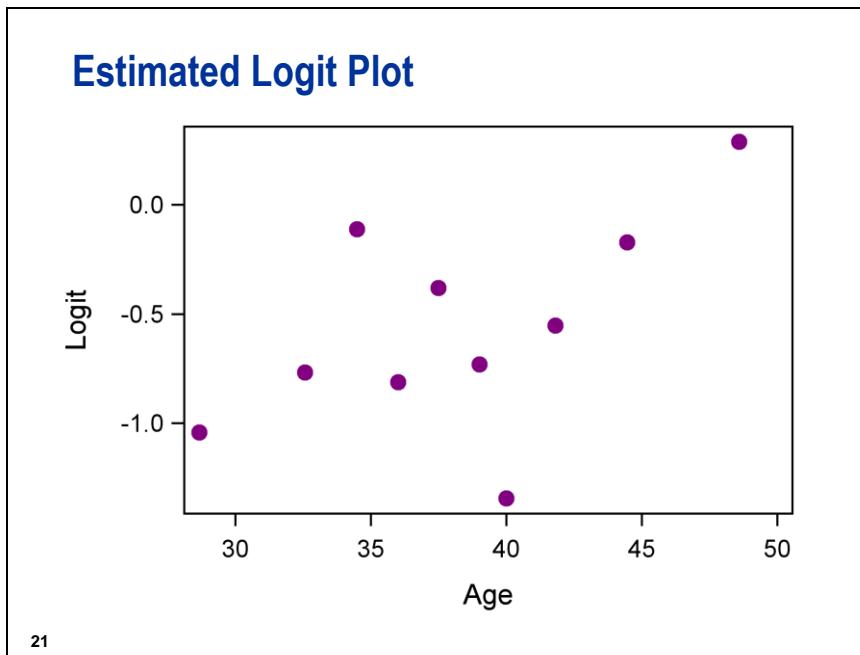
19

Scatter Plot of Binary Response Data



20

For continuous data, a recommended step before building a regression model is to analyze the bivariate relationships between the regressors and the response variables. The goal is not only to detect outliers, but also to analyze the shape of the relationships to determine if there might be some nonlinear trend that should be modeled in the analysis. For binary response variables, a scatter plot contributes little to these ends.



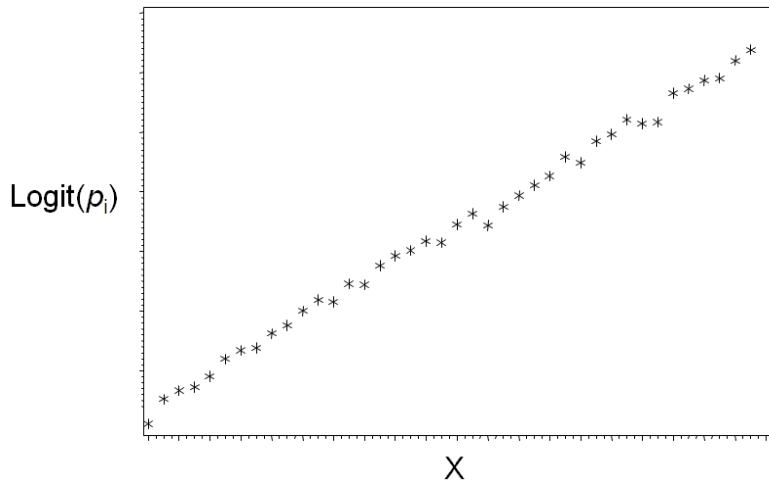
21

The logistic model asserts a linear relationship with the logit (not with the actual binary values). However, a logit for one observation will be infinite in either the positive or negative direction ($\ln(p / (1 - p)) = \ln(1 / 0)$ or $\ln(0 / 1)$). A recommendation, however, is to group the data into approximately equally sized bins, based on the values of the predictor variable. The bin size should be adequate in number of observations to reduce the sample variability of the logits. One can then assume that the average probability within each bin is approximately the value of the proportion in the bin with the event. The estimated logit is then approximately equal to $\ln(\text{proportion} / (1 - \text{proportion}))$.



If the predictor variable is a nominal variable, then there is no need to create a logit plot.

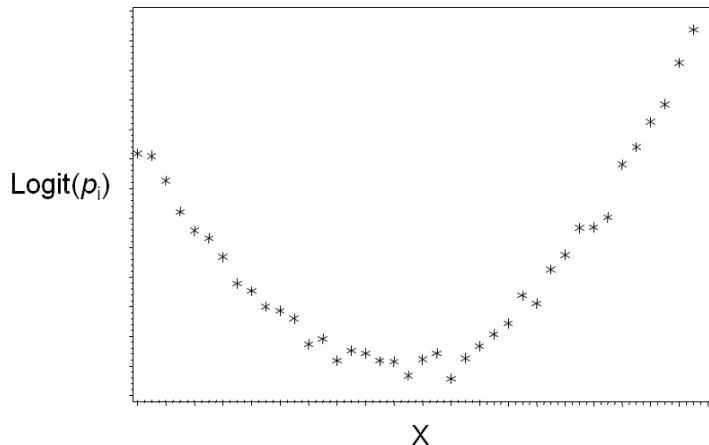
Linear Logit Plot



22

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.

Quadratic Logit Plot



23

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

24

A common approach in computing logits is to take the log of the odds. The path from the definition of a logit to the formula above is shown below. M represents the total number in the bin and m represents the total number of positive events in the bin.

$$\left(\frac{p}{(1-p)} \right) = \left(\frac{\frac{m}{M}}{\left(\frac{M}{M} - \frac{m}{M} \right)} \right) = \left(\frac{m}{(M-m)} \right)$$

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

The following is a summary of what you will accomplish in this demonstration:

- 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 250 dollars or more and the total number of customers by each level of **INCLEVEL** must be computed.
- 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.

```
/*st010d03*/
proc means data=st092.sales_inc noint nway;
  class IncLevel;
  var Purchase;
  output out=bins sum(Purchase)=Purchase n(Purchase)=BinSize;
run;

data bins;
  set bins;
  Logit=log( (Purchase+1) / (BinSize-Purchase+1) );
run;

proc sgscatter data=bins;
  plot Logit*IncLevel /
    markerattrs=(symbol=asterisk color=blue size=15);
  format IncLevel incfmt.;
  label IncLevel='Income Level';
  title 'Estimated Logit Plot of Income Level';
run;

quit;
```

Selected PROC MEANS statement option:

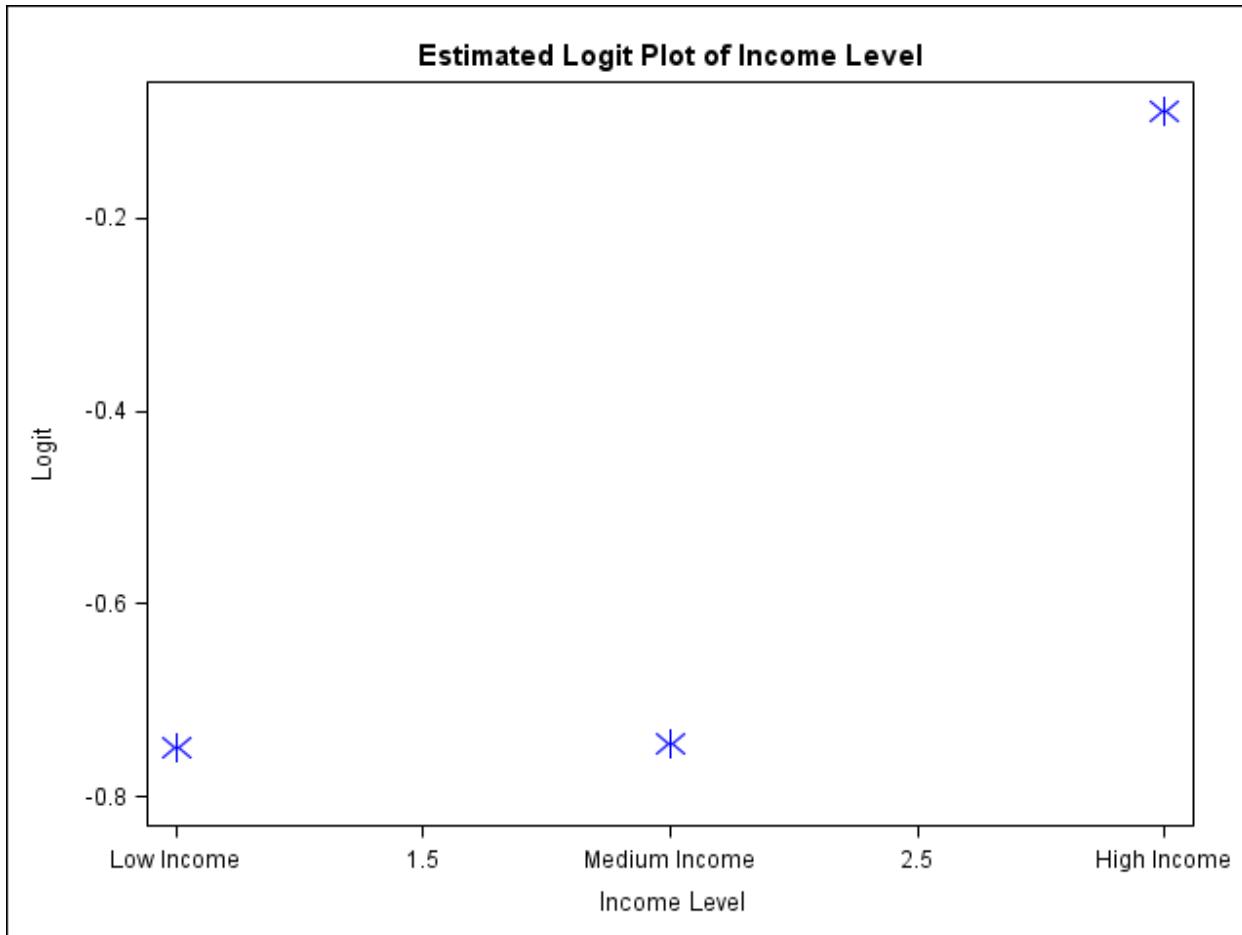
NWAY causes the output data set to have only one observation for each level of the class variable.

Selected PROC SGSCATTER PLOT statement option:

MARKERATTRS controls the display of the marker values for data points on the plot. SIZE is measured in pixels.

PROC MEANS creates a data set that contains a separate value for the requested statistics for each level of the CLASS variable. Because **PURCHASE** is coded 0/1, SUM(**Purchase**) returns the value for the count of 1's within each level of **INCLEVEL**. N(**Purchase**) returns the number of non-missing values of **Purchase**, which is the total effective sample size within each level.

The **Logit** is created in the DATA step, using the formula seen in the slide shown previously. In this case, M is represented by **BinSize** and m is represented by **Purchase**.



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 250 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 may be removed from the CLASS statement, implying that it would be a considered continuous variable. The statistical advantage of doing so would be to increase model power, due to obtaining almost the same information using fewer degrees of freedom. However, theoretical justifications should always supersede such data-driven considerations.

```
/*st010d03*/
proc rank data=st092.sales_inc groups=17 out=Ranks17;
  var Age;
  ranks Bin17;
run;

proc means data=Ranks17 noint nway;
  class Bin17;
  var Purchase Age;
  output out=Bins17 sum(Purchase)=Purchase n(Purchase)=BinSize
    mean(Age)=Age;
run;

data Bins17;
  set Bins17;
  Logit=log((Purchase+1)/(_freq_-Purchase+1));
run;

proc sgscatter data=Bins17;
  plot Logit*Age /
    reg markerattrs=(symbol=asterisk color=blue size=15);
  title "Estimated Logit Plot of Customer's Age";
run;

quit;
```

Selected PROC RANK statement option:

GROUPS=*n* bins the variables into *n* groups.

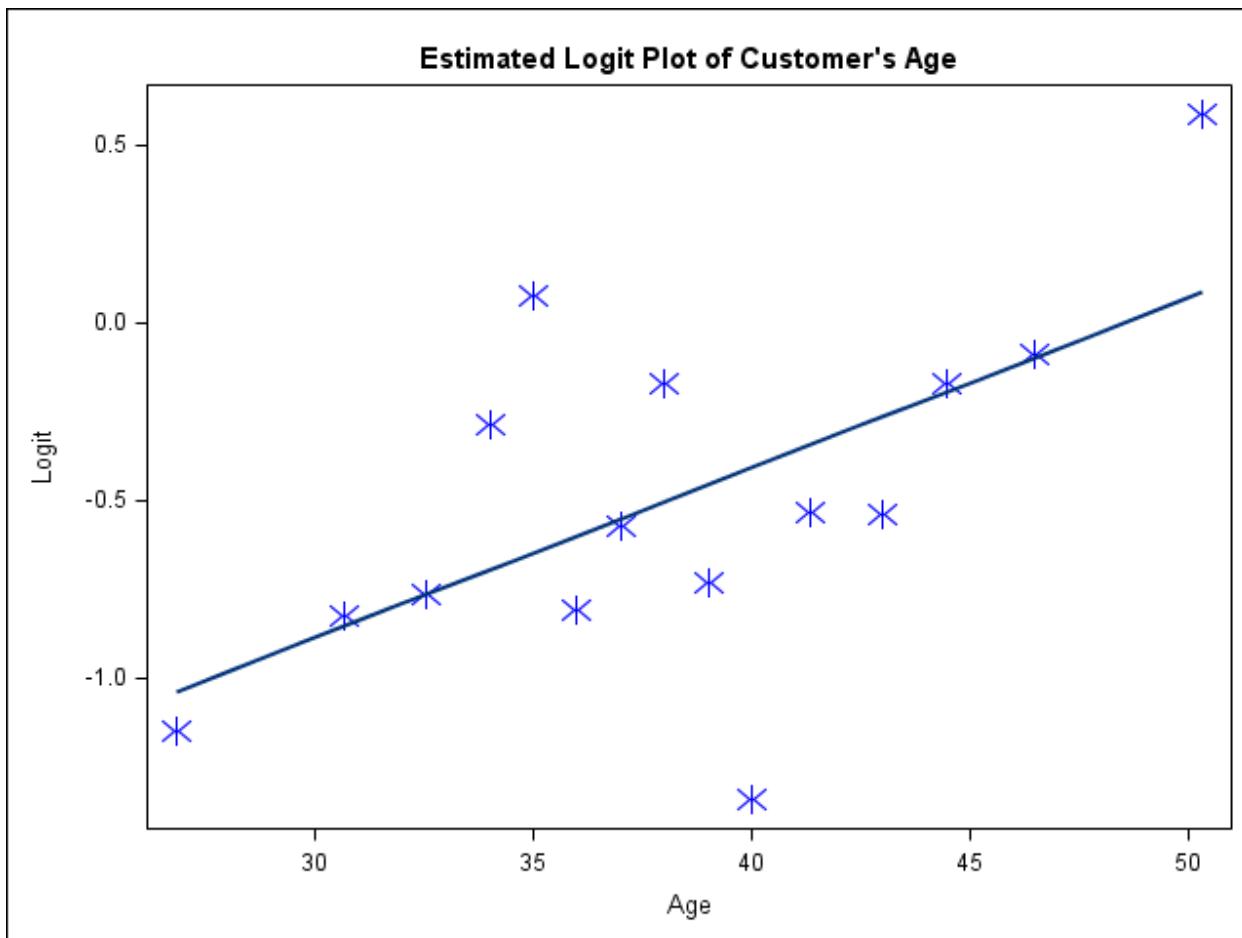
Selected RANK procedure statement:

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.

Selected PROC SGSCATTER PLOT statement option:

REG adds a regression fit to the scatter plot.

In the case of **AGE**, you do not have a made-to-order bin variable, so you must create one. You can use the RANK procedure for this purpose. You have 431 observations. It is recommended that you have about 20 to 30 observations per bin. At 25 per bin, you could create $431/25 \sim 17$ bins. That will be the option value of GROUPS=.



The regression line is used as a visual reference to aid in determining the linearity of the relationship. The estimated logit plot shows no clear deviation from linearity. Therefore, **AGE** can be entered into the model as is, without creating a categorical variable or adding a higher-level term to account for a curve. The estimated logit plot is a univariate plot and, therefore, 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). If an interaction is suspected, a model with the interaction term and main effects should be evaluated before any variable is eliminated. Estimated logit plots should never be used to eliminate variables from consideration for a multiple logistic regression model.

Appendix A Additional Topics

A.1 ODS Statistical Graphics	A-3
Demonstration: ODS Statistical Graphics Using PROC CORR	A-6
Demonstration: ODS HTML Output Using the STYLE=STATISTICAL Option	A-9
Demonstration: Using the ODS Graphics Editor	A-12
A.2 Paired <i>t</i>-Tests	A-17
Demonstration: Paired <i>t</i> -Test	A-19
A.3 Fishers Exact <i>p</i>-values	A-21
Demonstration: Exact <i>p</i> -Values for the Pearson Chi-Square Test	A-27
A.4 Nonparametric ANOVA.....	A-30
Demonstration: The NPAR1WAY Procedure	A-36
A.5 Partial Leverage Plots	A-49
Demonstration: Partial Leverage Plots	A-52

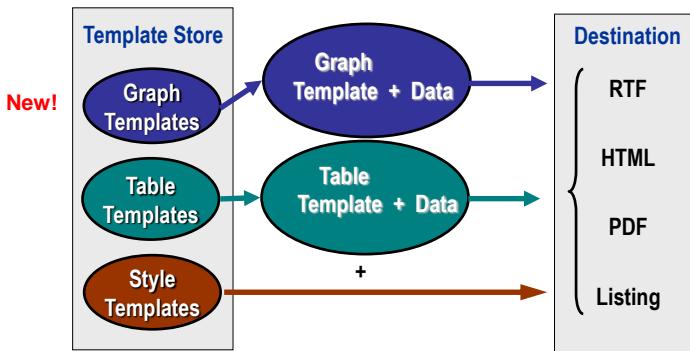
A.1 ODS Statistical Graphics

ODS Statistical Graphics

- Many statistical procedures can now create graphs as automatically as they create tables.
- New SAS/GRAF procedures streamline the process of making graphs.
- The graphics have ODS styles designed for statistical work.
- A SAS/GRAF license is required for ODS Graphics functionality.

4

ODS Process Overview



5

General Syntax of ODS Graphics

```
ODS GRAPHICS ON;  
  
    statistical procedure code;  
  
ODS GRAPHICS OFF;
```

6

In order to produce statistical graphics from SAS statistical procedures, the statement ODS GRAPHICS (or ODS GRAPHICS ON) must be submitted. This statement need only be submitted once within an interactive SAS session (or batch job) and will remain in effect until the ODS GRAPHICS OFF statement is submitted. An exception to this requirement is in the case of the new statistical graphics procedures, PROC SGLOT, PROC SGSCATTER, and PROC SGRENDER. These procedures will produce graphics whether or not the ODS GRAPHICS statement is first submitted.

ODS Graphics Output

- Some graphs are created by default.
- Procedure options (such as PLOTS=) are used to specify which graphs to create.
- You can specify where you want your graphs displayed by using ODS destination statements.
- ODS SELECT and ODS EXCLUDE statements can be used to select and exclude graphs from your output.

7

The SAS 9.2 documentation lists the graphics available within the description of the SAS procedure.

Some ODS Destinations

Destination	Viewer	Graphics File Types
HTML	Web Browser	PNG (default), GIF, JPEG, ...
RTF	Word Processor, such as Microsoft Word	Contained in RTF file
PS	PostScript viewer, such as GSview	Contained in PostScript file
PDF	PDF viewer, such as Adobe Reader	Contained in PDF file
LATEX	PostScript or PDF viewer after compiling LaTeX file	PostScript (default), EPSI, GIF, JPEG, PDF, PNG
LISTING	Default viewer in your system for file type	PNG (default), GIF, BMP, DIB, EMF, EPSI, JFIF, JPEG, ...

8

There are several destinations available for ODS Statistical Graphics. The default destination is the LISTING destination, but unlike graphs produced in traditional SAS GRAPH procedures, the graphs produced by ODS Statistical Graphics will not open by default in a Graphics Window in SAS. Instead, the graphs will open in separate windows determined by the user's default graphic viewers.



ODS Statistical Graphics Using PROC CORR

This demonstration shows how to use ODS Statistical Graphics and view the default output in SAS.

1. Turn on ODS Statistical Graphics by using the ODS GRAPHICS or ODS GRAPHICS ON statement in the SAS code. After this statement is submitted, the ODS Graphics engine will be on until turned off by the ODS GRAPHICS OFF statement. Submit the procedure and turn off the ODS Statistical Graphics.

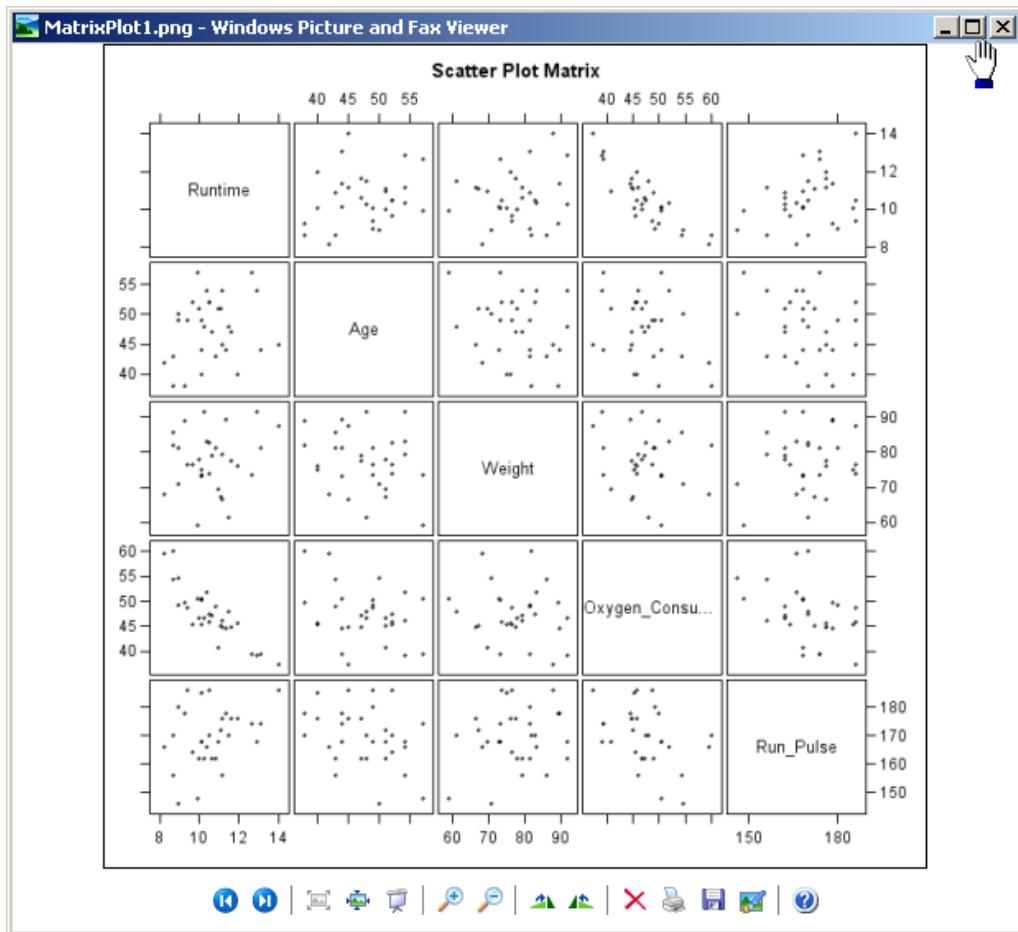
```
/*st00ad01.sas*/
ods graphics;
proc corr data=st092.fitness;
  var Runtime Age Weight Oxygen_Consumption Run_Pulse
    Rest_Pulse Maximum_Pulse Performance;
  title "Scatter Plot Matrix of Variables in Fitness Data Set";
run;

ods graphics off;
```

2. Expand the output from the Results window on the left and double-click on the icon for the desired graphic to open the picture for viewing.

	Rest_Pulse	Maximum_Pulse	Performance
Rest_Pulse	1.00000	0.30512	0.0951
Maximum_Pulse	<.0001	0.30512	1.00000
Performance	-0.31369	-0.47957	-0.22035
	0.0857	0.0063	0.2336

3. The results open in the computer's default graphics viewer.



Accessing Individual Graphs

The screenshot shows the SAS Results viewer. On the left, there is a tree view of analysis results under 'Results'. The 'Freq' node is expanded, showing 'Table prev_premr' and 'Table low'. 'Table prev_premr' is further expanded to show 'One-Way Frequencies', 'Distribution Plots', 'One-Way Chi-Square Test', and 'Deviation Plot'. 'Table low' is also expanded to show similar categories. On the right, a plot titled 'Odds Ratios and 95% Confidence Limits' is displayed. The plot has 'prev_premr by low' on the y-axis. It shows two points with horizontal error bars: one at Odds Ratio 5.4844 (2.0894, 14.396) and another at Odds Ratio 1.3889 (0.2822, 6.8355). The x-axis ranges from 0.0 to 15.0.

10

For Word documents and PowerPoint presentations use the RTF or LISTING destination. Copy and paste the graphs from the viewer. For LaTeX documents use the LATEX destination:

```
ods latex gpath="C:\mygraphs";
ods graphics on / imagefmt=ps;
```

Then, include the PostScript or PDF files into the document.

Some Recommended ODS Styles

Style	Description
DEFAULT	Color style intended for general-purpose work. This is the default for the HTML destination.
STATISTICAL	Color style recommended for output in Web pages or color print media. This is the style used in the SAS/STAT 9.2 documentation.
ANALYSIS	Color style with a somewhat different appearance from STATISTICAL.
JOURNAL and JOURNAL2	Gray-scale and pure black-and-white styles, respectively. Recommended for graphs in black-and-white publications.
RTF	Used to produce graphs to insert into a Microsoft Word document or a Microsoft PowerPoint slide.

11



ODS HTML Output Using the STYLE=STATISTICAL Option

1. Add the ODS HTML statement with the STYLE=STATISTICAL option to send the output to an HTML file. Specifying that the FILE=, GPATH=, and PATH= are optional, but are recommended. If not specified, the file will be named automatically by the procedure and the HTML output will be sent to the default location for the system. The ODS HTML CLOSE statement completes the writing of the HTML file. **&outpath** is a macro variable set by the user. For this course, the macro variable **&outpath** is set as: '%let outpath=s:\workshop;' .

```
/*st00ad02.sas*/
ods graphics;
ods html file='corr.html'
    style=statistical;

proc corr data= st092.fitness;
    var Runtime Age Weight Oxygen_Consumption Run_Pulse
        Rest_Pulse Maximum_Pulse Performance;
    title "HTML CORR output in Fitness Data Set";
run;

ods html close;
ods graphics off;
```

2. A separate Results Viewer window tab opens. Select the tab to see the HTML output.

SAS

File Edit View Go Tools Solutions Window Help

Results

Results Viewer - SAS Output

HTML CORR output in Fitness Data Set

The CORR Procedure

8 Variables: Runtime Age Weight Oxygen_Consumption Run_Pulse Rest_Pulse Maximum_Pulse Performance

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Runtime	31	10.58613	1.38741	328.17000	8.17000	14.03000
Age	31	47.67742	5.26236	1478	38.00000	57.00000

Output - (Untitled) Log - (Untitled) newfeatures.sas Results Viewer - SA...

C:\Program Files\SAS\SASI

This screenshot shows the SAS Results Viewer interface. The left pane displays a tree view of results, with 'Corr: Scale' expanded to show 'Variables', 'Simple', 'Pearls', and 'Scatter'. The main pane displays the 'HTML CORR output in Fitness Data Set' from the 'The CORR Procedure'. It lists 8 variables: Runtime, Age, Weight, Oxygen_Consumption, Run_Pulse, Rest_Pulse, Maximum_Pulse, and Performance. Below this is a 'Simple Statistics' table with columns for Variable, N, Mean, Std Dev, Sum, Minimum, and Maximum. The table shows data for Runtime and Age. The bottom of the window shows tabs for Output, Log, and newfeatures.sas, along with the Results Viewer title bar and the path C:\Program Files\SAS\SASI.

SAS

File Edit View Go Tools Solutions Window Help

Results

Results Viewer - SAS Output

Performance

	-0.98841	-0.22943	-0.10544		0.86377	-0.31369
<.0001	0.2144	0.5724		<.0001	0.0857	

Scatter Plot Matrix

40 45 50 55 40 45 50 55 60

Runtime

Output - (Untitled) Log - (Untitled) newfeatures.sas Results Viewer - SA...

C:\Program Files\SAS\SASI

Done

This screenshot shows the SAS Results Viewer interface. The left pane displays a tree view of results, with 'Corr: Scale' expanded to show 'Variables', 'Simple', 'Pearls', and 'Scatter'. The main pane displays a 'Scatter Plot Matrix' for the 'Performance' variable. The matrix shows scatter plots for pairs of variables: Runtime vs Runtime, Runtime vs Age, Runtime vs Weight, Runtime vs Oxygen_Consumption, Runtime vs Run_Pulse, Runtime vs Rest_Pulse, Runtime vs Maximum_Pulse, and Runtime vs Performance. The diagonal elements of the matrix show histograms of the individual variables. The bottom of the window shows tabs for Output, Log, and newfeatures.sas, along with the Results Viewer title bar and the path C:\Program Files\SAS\SASI. A 'Done' button is visible at the bottom left.

Modifying Your Graphs

- Use the ODS Graph Editor, a point-and-click interface
 - for data and graph-specific changes
 - to customize titles and labels, annotate data points, add text, and change the properties of graph elements.
- Make persistent changes by modifying the ODS graph template for a particular plot.

13

ODS Graphics Editor

- You can enable editing for the duration of your SAS session by first selecting the Results window and then entering SGEDIT ON from the command line.
- You can disable editing by entering SGEDIT OFF.
- To invoke the ODS Graphics Editor, submit your SAS program and then right-click in the Results window on the plot you want to edit and select Edit.

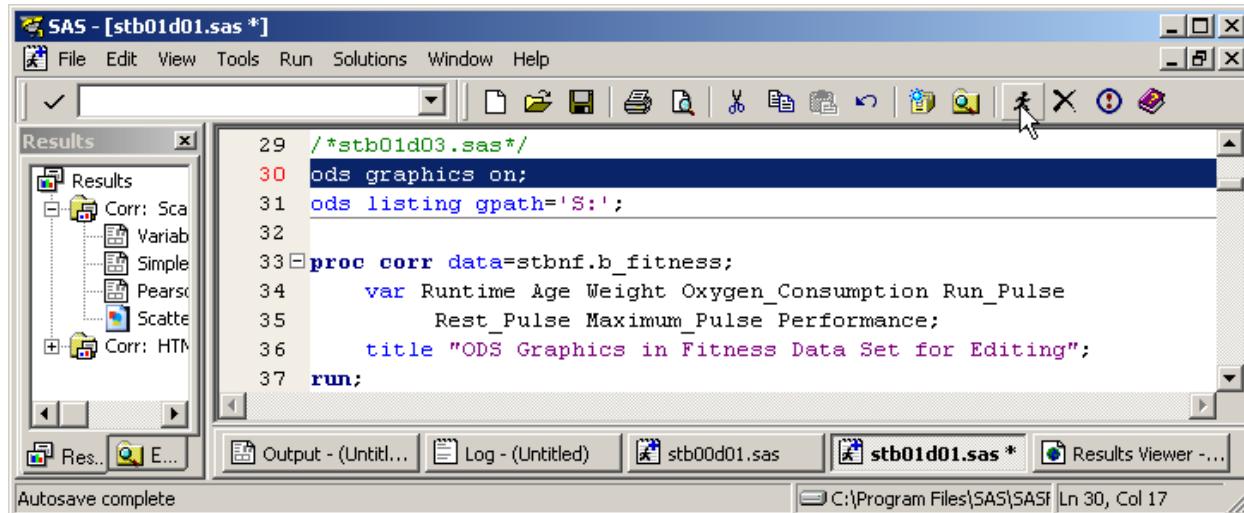
14



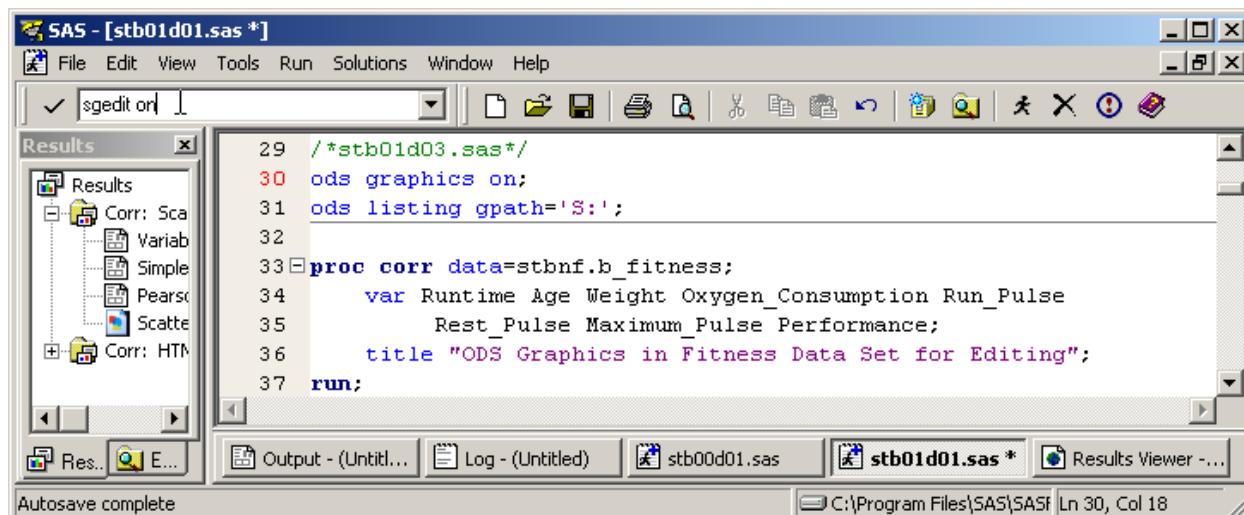
Using the ODS Graphics Editor

```
/*st00ad03.sas*/
ods graphics on;
proc corr data=st092.fitness;
  var Runtime Age Weight Oxygen_Consumption Run_Pulse
      Rest_Pulse Maximum_Pulse Performance;
  title "ODS Graphics in Fitness Data Set for Editing";
run;
```

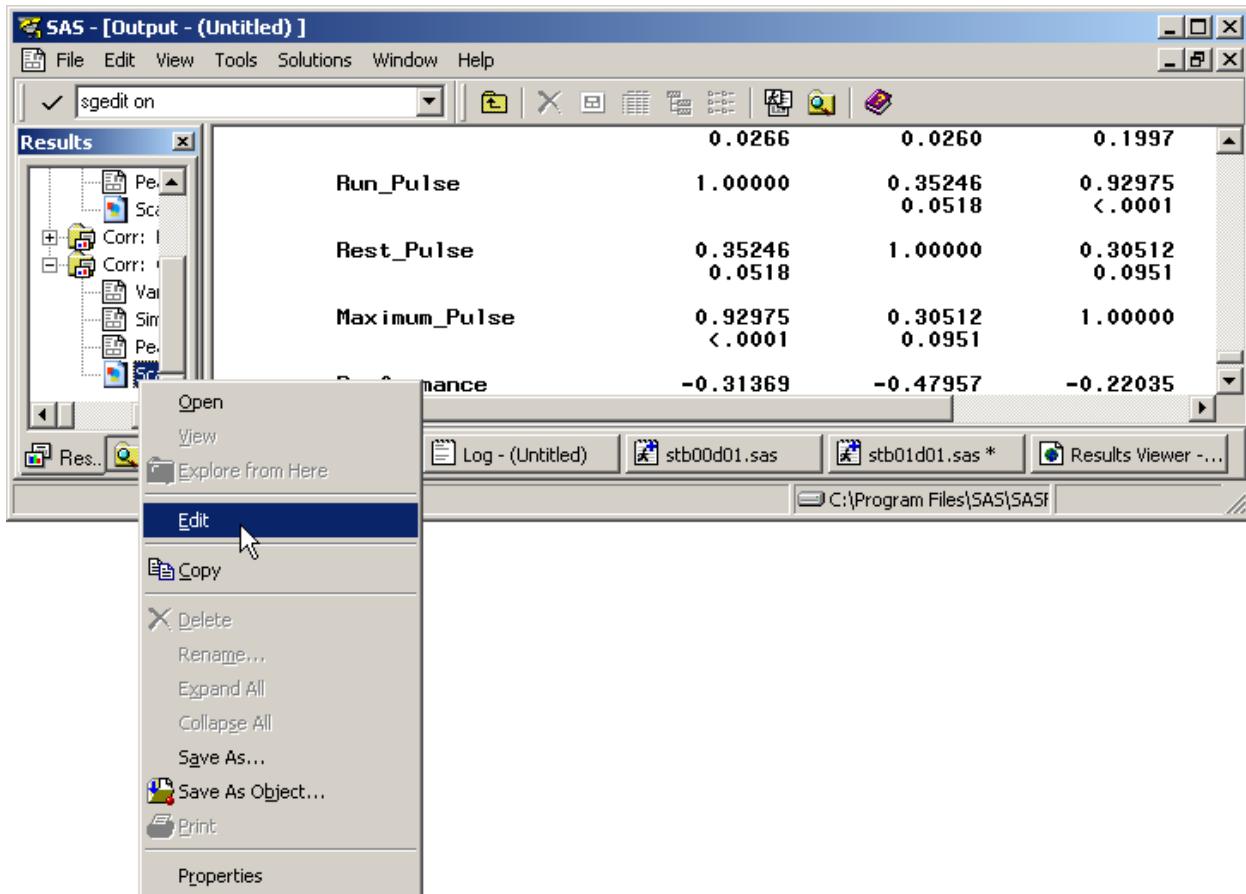
1. Turn ODS Graphics on.



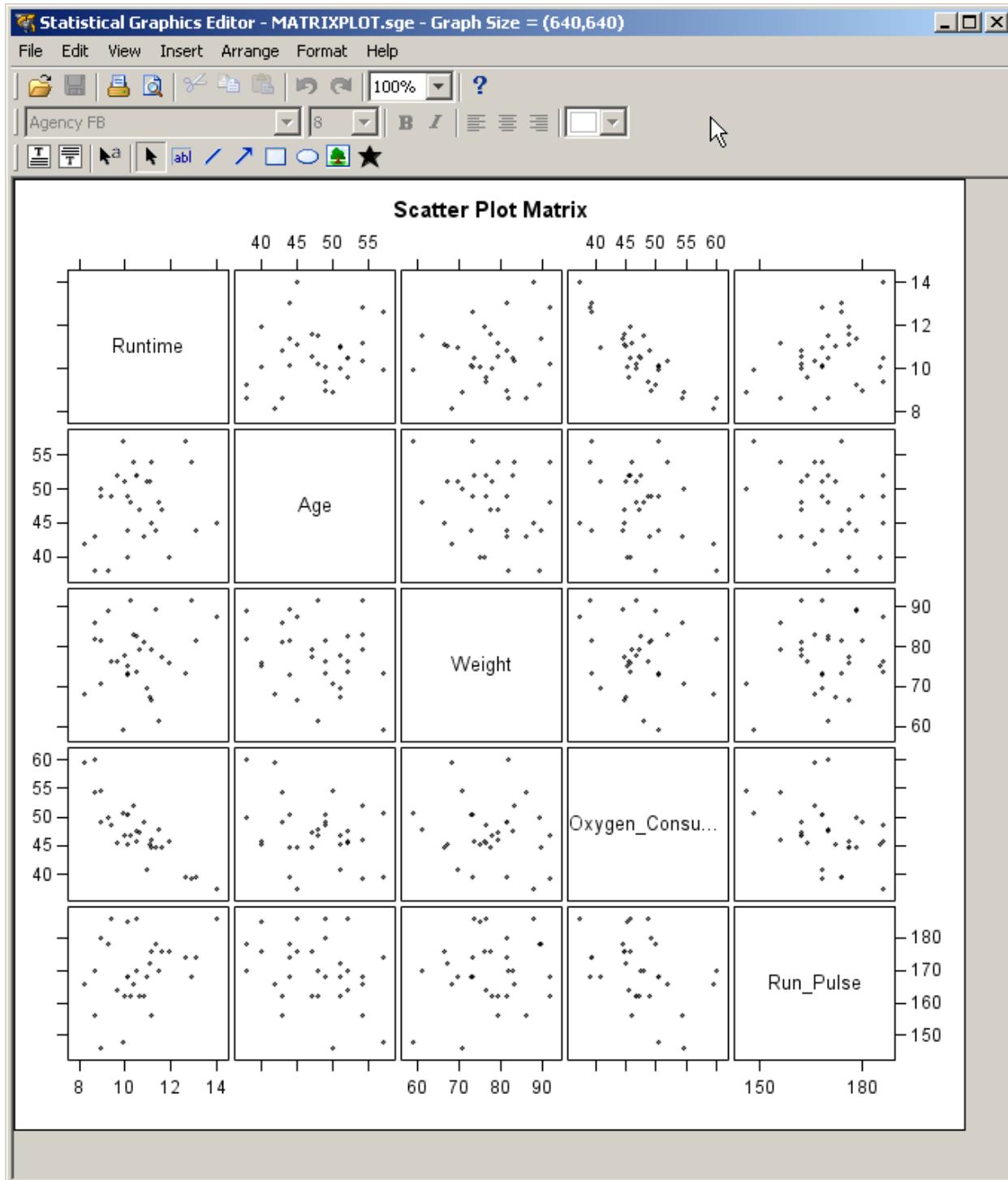
2. Turn SGEDIT on by typing **sgedit on** in the Command window and then pressing ENTER.



3. Submit the procedure, expand the output in the Results window and then right-click on the graphics icon. Select **Edit** from the drop-down menu.



- The Graphics Editor might open behind other windows and it might have to be opened from the tabs at the bottom of the computer screen.
4. The graph is now ready for editing.



Graph Templates

- Graph templates are programs, written in the Graph Template Language, that describe individual graphs.
- SAS provides a default template for every graph produced by a procedure, so you never need to write a template.
- You can modify the default templates to make persistent changes that apply every time you run your program.

16

ODS templates can be used to modify the layout and details of each graph.

Graph Templates versus ODS Graphics Editor

	Graph Template Changes	ODS Graph Editor
Appropriate for	SAS programmer familiar with the Graph Template Language	Statistical end user
Approach	Programming	Point-and-click
Type of change	Persistent	Immediate
Duration	Whenever program is re-run	Current graph only
Application	Batch processing of graphs	Papers, presentations
File saved	Modified graph template	PNG or SGE

17



Read more about editing SAS Statistical Graphics and Graph templates in the SAS 9.2 documentation.

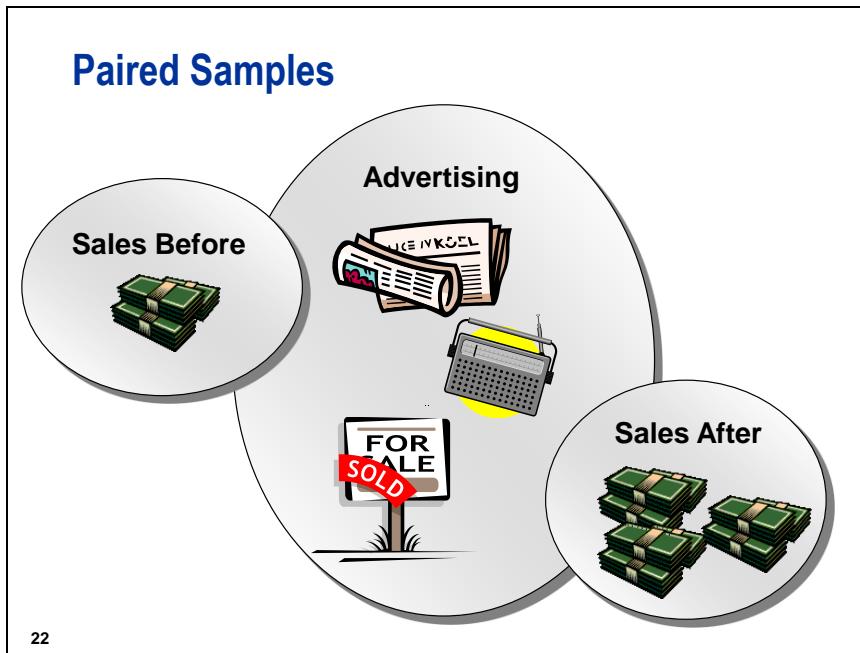
Statistical Graph Procedures in SAS/GRAPH

- PROC SGSCATTER creates single-cell and multi-cell scatter plots and scatter plot matrices with optional fits and ellipses.
- PROC SGPLOT creates single-cell plots with a variety of plot and chart types.
- PROC SGPANEL creates single-page or multi-page panels of plots and charts conditional on classification variables.
- PROC SGRENDER provides a way to create plots from graph templates that you have modified or written yourself.

18

These procedures will be described more fully in subsequent sections of this course.

A.2 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;
```

23

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.

```
/*st00ad04.sas*/
proc print data=st092.market (obs=20);
  title;
run;
```

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.

```
/*st00ad04.sas*/
proc ttest data= st092.market;
  paired post*pre;
  title 'Testing the Difference Before and After a Sales
Campaign';
run;
```

Testing the Difference Before and After a Sales Campaign					
The TTEST Procedure					
Difference: post - pre					
N	Mean	Std Dev	Std Err	Minimum	Maximum
30	0.9463	0.9271	0.1693	-0.4800	3.3400
Mean	95% CL Mean	Std Dev	95% CL Std Dev		
0.9463	0.6001	1.2925	0.9271	0.7384	1.2464
DF	t Value	Pr > t			
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.

A.3 Fishers Exact p-values

When Not to Use the Chi-square Test

When more than 20% of cells have expected counts less than five

26

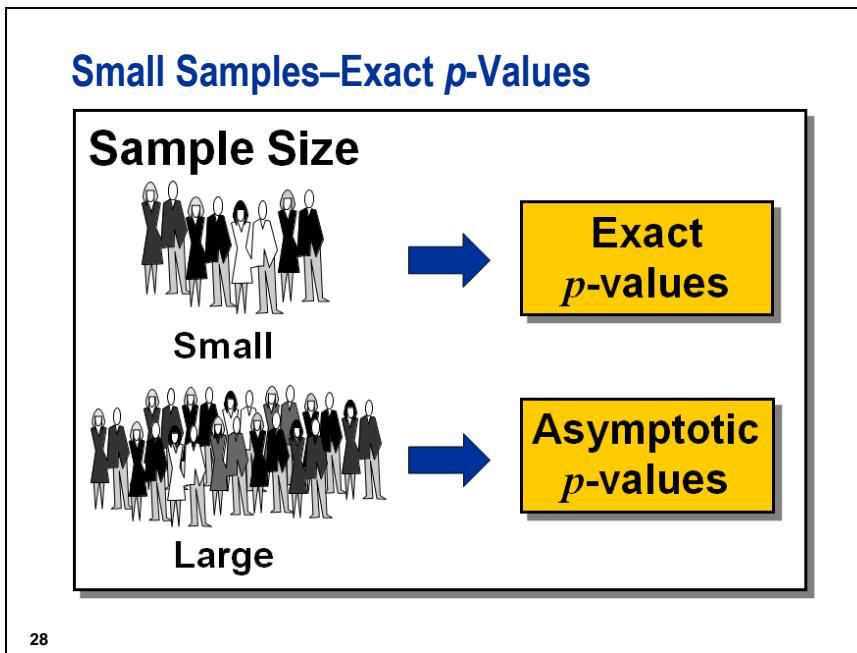
There are times when the chi-square test might not be appropriate. In fact, when more than 20% of the cells have expected cell frequencies of less than 5, the chi-square test might not be valid. This is because the *p*-values are based on the assumption that the test statistic follows a particular distribution 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

		Table of Row by Column			
		Column			
Row		1	2	3	Total
Frequency					
Expected		1	2	3	Total
1		1	5	8	14
		3.4286	4.5714	6	
2		5	6	7	18
		4.4082	5.8776	7.7143	
3		6	5	6	17
		4.1633	5.551	7.2857	
Total		12	16	21	49

27

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.



28

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

Observed Table

0	3	3
2	2	4
2	5	7

Expected Table

.86	2.14	3
1.14	2.86	4
2	5	7

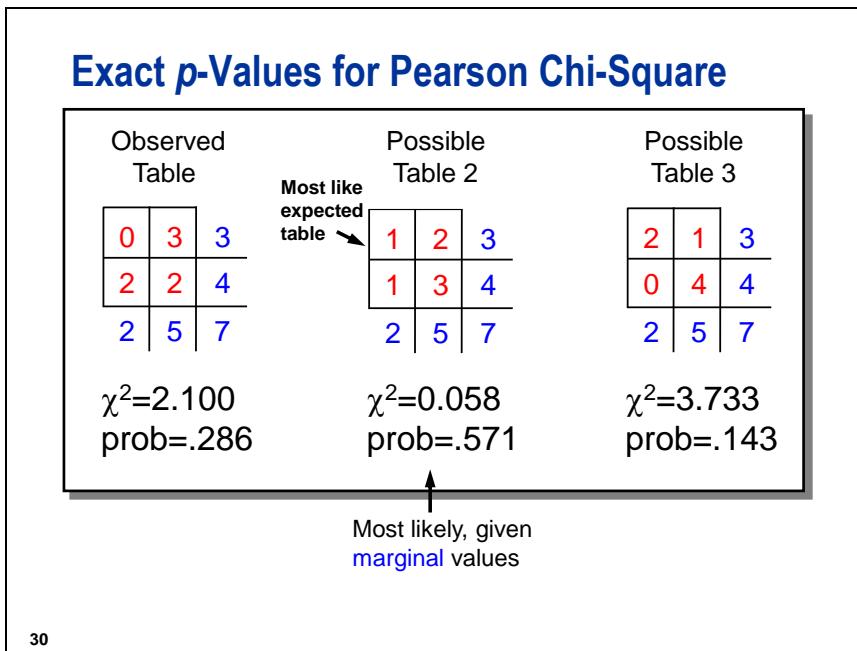
A p -value gives the probability of the value of the χ^2 value being as extreme or more extreme than the one observed, just by chance.

Could the underlined sample values have occurred just by chance?

29

Consider the table at left above. With such a small sample size, the asymptotic p -values would not be valid, because the accuracy of those p -values depends on large enough expected values in all cells.

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.



30

A key assumption behind the computation of exact p -values is that the column totals and row totals are fixed. There are only three possible tables, including the observed table, given the fixed marginal totals.

Possible Table 2 is most like the Expected Table of the previous slide. So, the probability (.571) that its cell values would occur in a table, given these row and column total values, is greatest of any possible table that could occur by chance.

Exact *p*-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	7

$$\chi^2=2.100$$

prob=.286

Possible Table 2

1	2	3
1	3	4
2	5	7

$$\chi^2=0.058$$

prob=.571

Possible Table 3

2	1	3
0	4	4
2	5	7

$$\chi^2=3.733$$

prob=.143

Exact *p*-value is the sum of probabilities of all tables with χ^2 values as great or greater than that of the Observed Table:

$$p\text{-value} = .286 + .143 = .429$$

31

To compute an exact *p*-value for this example, examine the chi-square value for each table and the probability that the table should occur by chance if the null hypothesis of no association were true (the probabilities add up to 1). Remember the definition of a *p*-value. It is the probability, if the null hypothesis is true, that you would obtain a sample statistic as great as or greater than the one you observed, just by chance. In this example, this means the probability of obtaining a table with a χ^2 value as great as or greater than the 2.100 for the Observed Table. The probability associated with every table with a χ^2 value of 2.100 or higher would be summed to compute the exact *p*-value.

The exact *p*-value would be 0.286 (Observed Table) + 0.143 (Possible Table 3) = .429. This means you have a 42.9% 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 **st092.exact** data set, which has the data from the previous example.

```
/*st00ad05.sas*/
ods graphics off;
proc freq data=st092.exact;
  tables A*B;
  exact pchi;
  title "Exact P-Values";
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.

Exact P-Values			
The FREQ Procedure			
		Table of A by B	
A	B		
Frequency		1	2
Percent		0.00	42.86
Row Pct		0.00	100.00
Col Pct		0.00	60.00
1		3	3
		42.86	42.86
2		2	2
		28.57	28.57
		50.00	50.00
		100.00	40.00
	Total	2	5
		28.57	71.43
			7
			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	

The warning message informs you that because of the small sample size, the asymptotic chi-square might not be a valid test.

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.

A.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.

34

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					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 = 19					Sum = 36				

35

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.

Median Scores

Treatment	A					B				
Response	2	5	7	8	10	6	9	11	13	15
Median Score	0	0	0	0	1	0	1	1	1	1
Median = 9.5										
Sum = 1				Sum = 4						

36

Recall that the median is the 50th 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.

Hypotheses of Interest

H_0 : all populations are identical with respect to scale, shape, and location.

H_1 : all populations are not identical with respect to scale, shape, and location.

37

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;
```

38

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?



39

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?



40

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 **st092.hosp** data set. The variables in the data set are as follows:

id	the ID number of the physician's office visited
visit	the type of visit, to the physician or to the physician's staff
code	the medical specialty of the physician
ref3p	the number of referrals three months prior to the visit
ref2p	the number of referrals two months prior to the visit
ref1p	the number of referrals one month prior to the visit
ref3a	the number of referrals three months after the visit
ref2a	the number of referrals two months after the visit
ref1a	the number of referrals one month after the visit

In addition, the following variables have been calculated:

avgprior	the average number of referrals per month for the three months prior to the visit
diff1	the difference between the number of referrals one month after the visit and the average number of referrals prior to the visit
diff2	the difference between the number of referrals two months after the visit and the average number of referrals prior to the visit
diff3	the difference between the number of referrals three months after the visit and the average number of referrals prior to the visit
diffbys1	the difference between the number of referrals one month after the visit and the number of referrals three months prior to the visit
diffbys2	the difference between the number of referrals two months after the visit and the number of referrals three months prior to the visit
diffbys3	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.

```
/*st00ad06.sas*/
proc print data= st092.hosp (obs=10);
  var visit code diffbys3;
```

```
run;
```

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**.

```
/*st00ad06.sas*/
options ps=50 ls=76;

proc sort data= st092.hosp out=sorted_hosp;
  by code;
run;
ods graphics;
ods listing gpath="&outpath";
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 sgplot data=sorted_hosp;
  vbox diffbys3 / category = code;
run;
```

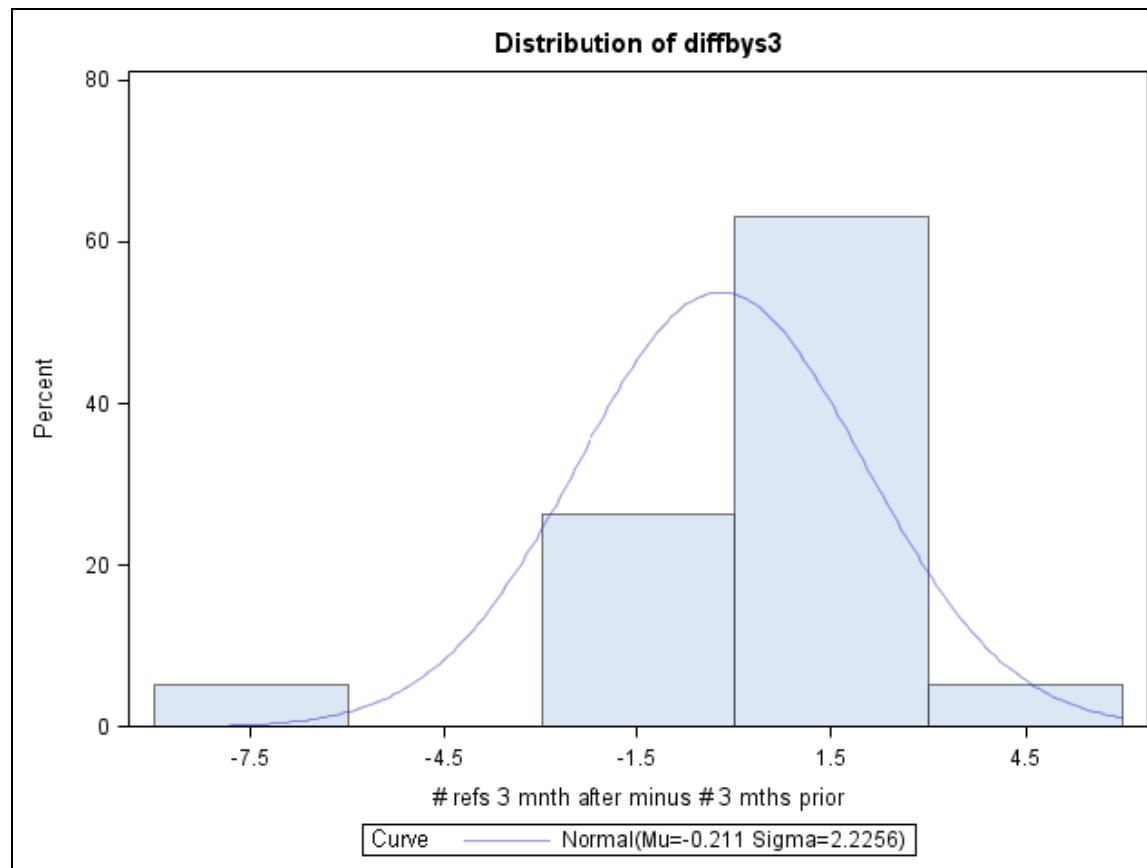
Selected PROC UNIVARIATE Output by **specialty code**

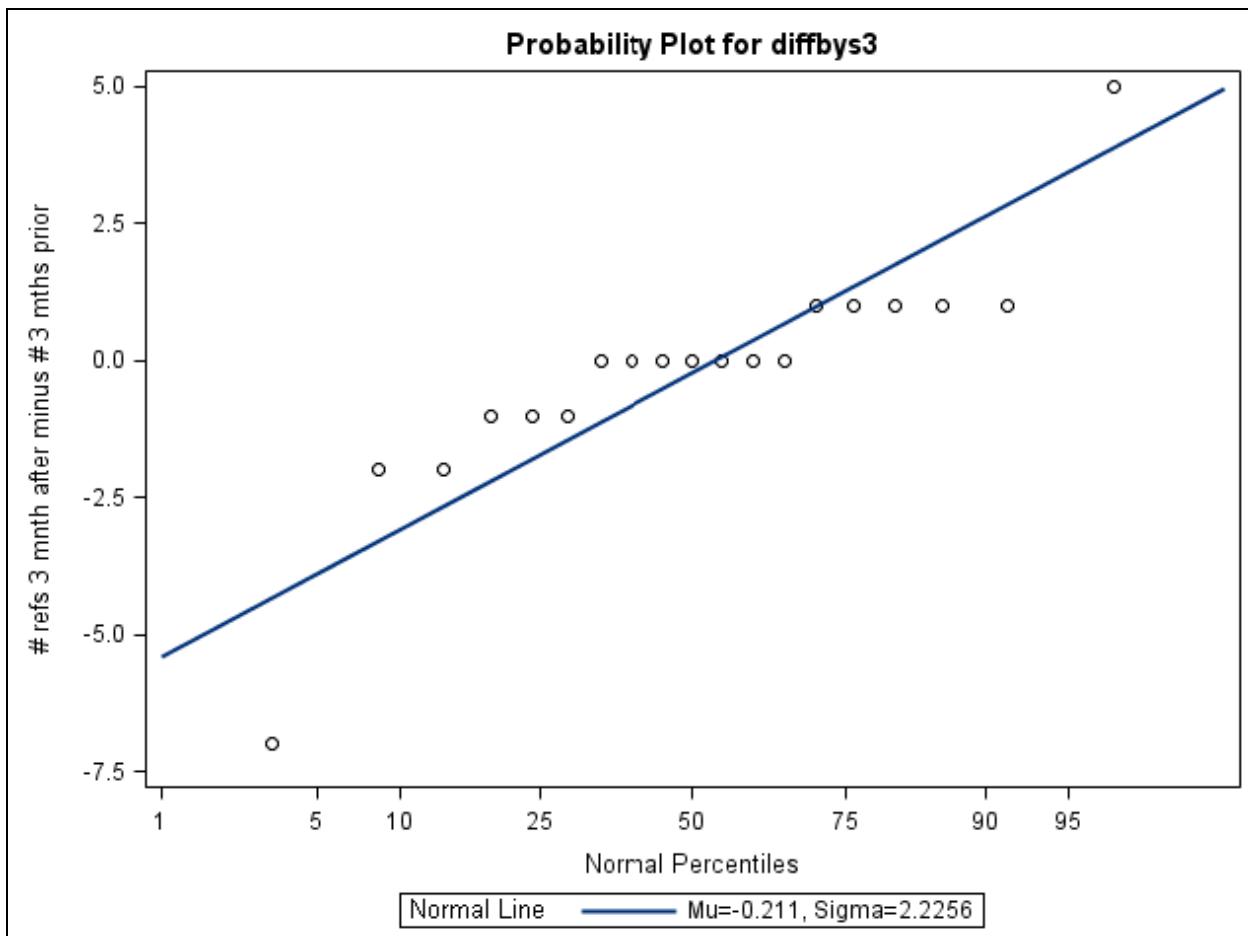
Group	Skewness	Kurtosis	Kolmogorov-Smirnov <i>p</i> -value	Cramer-von Mises <i>p</i> -value	Anderson-Darling <i>p</i> -value
oncologist	-0.988574	5.58306776	<0.010	<0.010	<0.005
internal med	0.94171457	-0.2843557	<0.010	<0.005	<0.005
family prac	-1.3336242	6.24954044	<0.010	<0.005	<0.005

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.

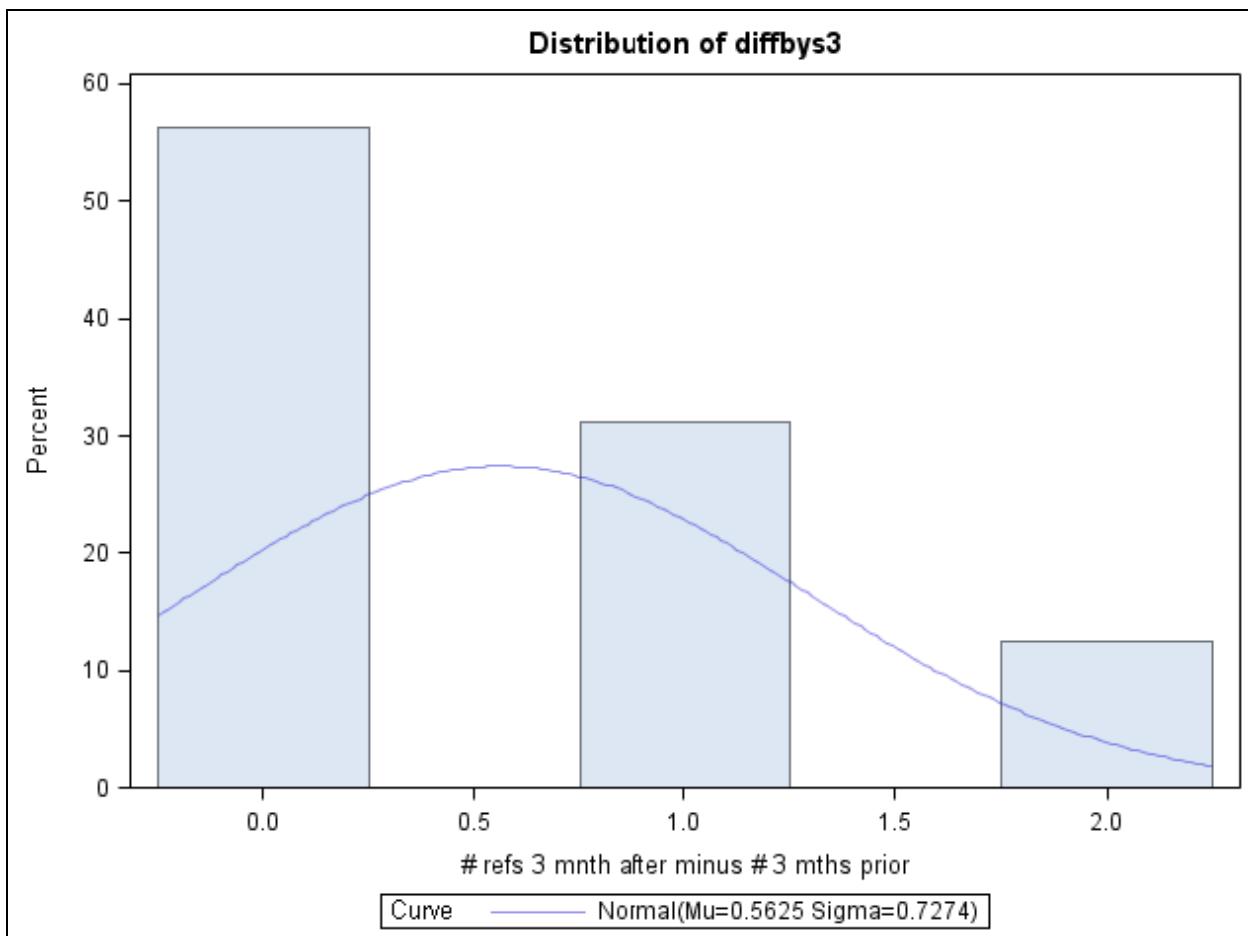
Now examine the histograms and normal probability plots for each group.

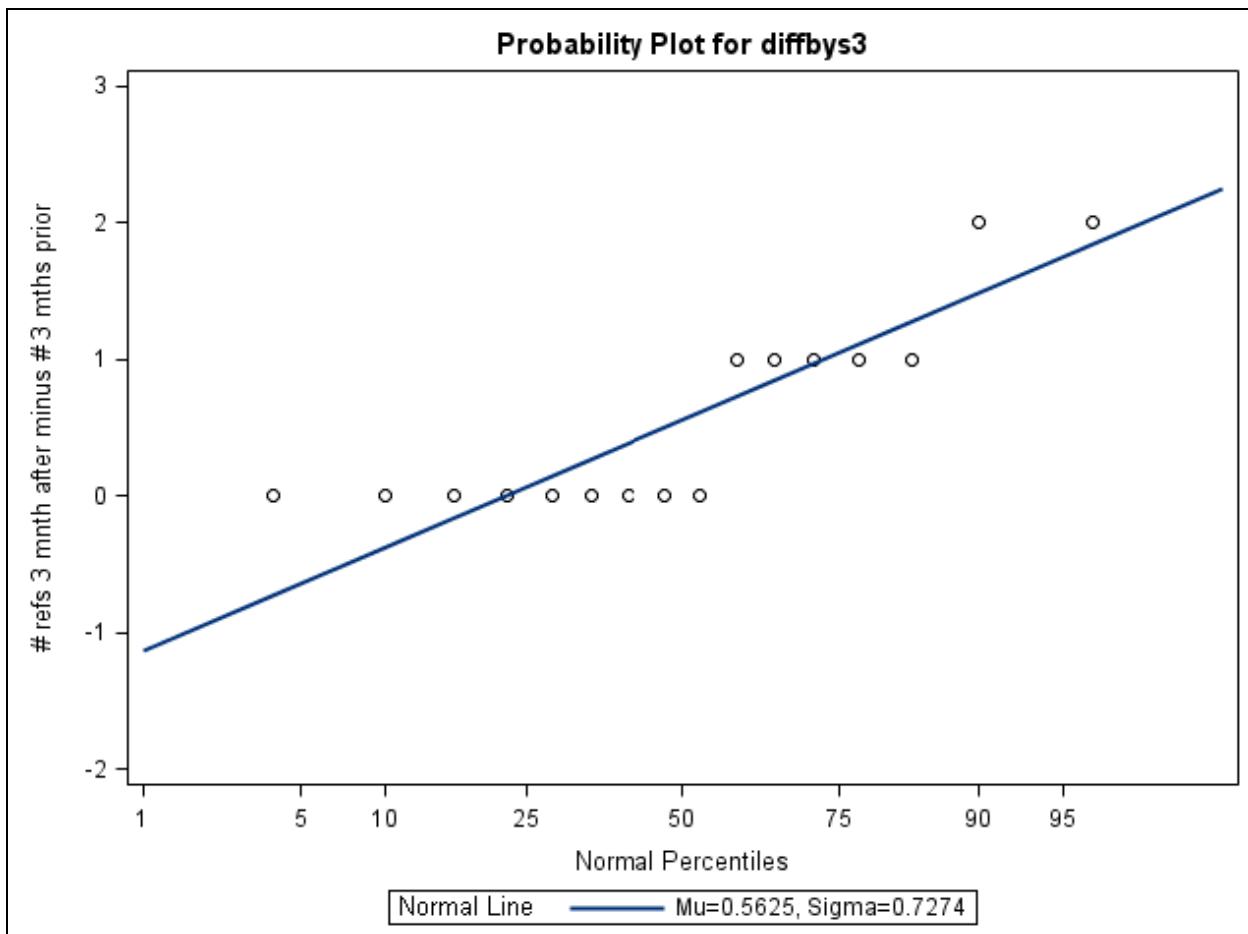
SAS/GRAFH Output (oncologists)





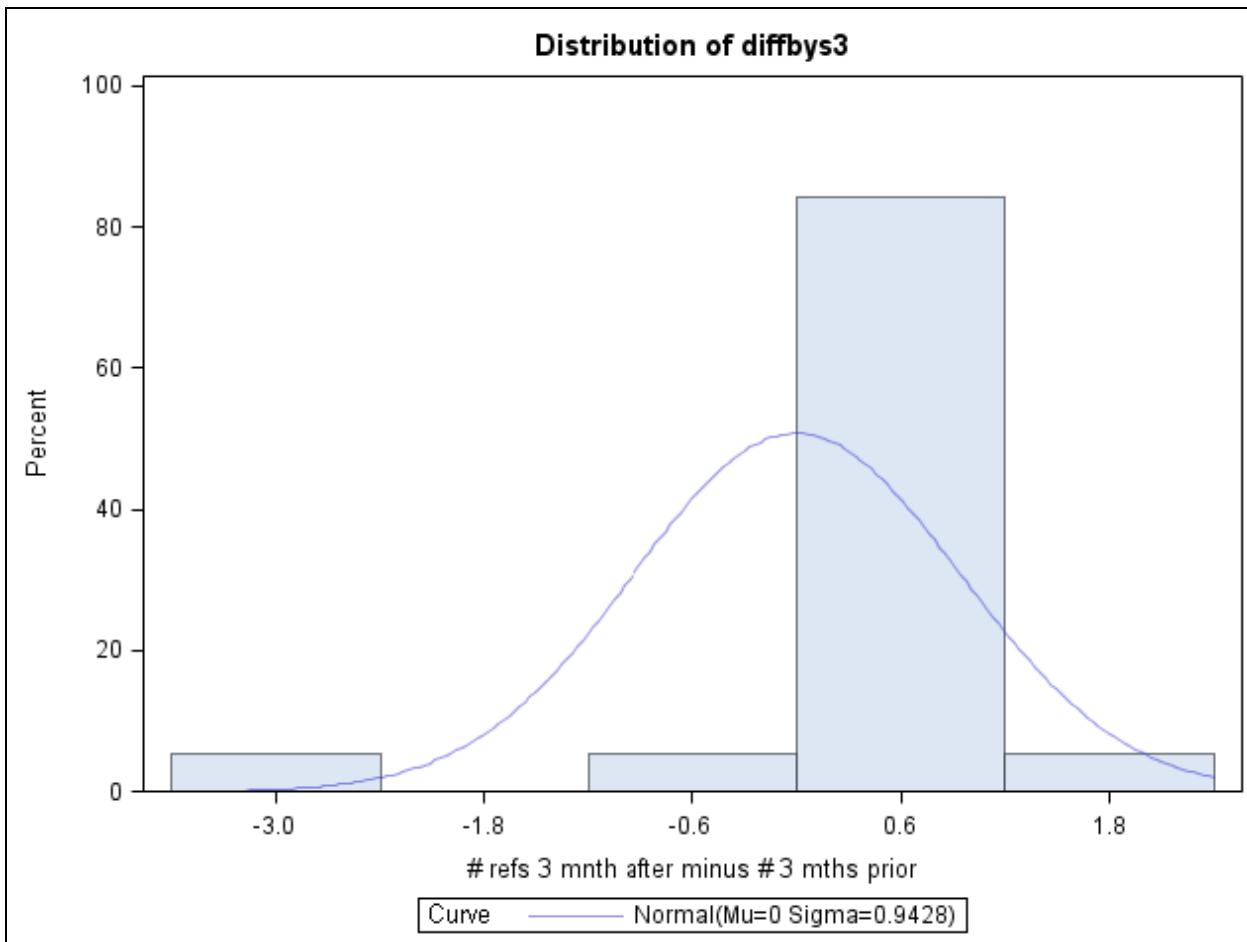
SAS/GRAPH Output (internal medicine)

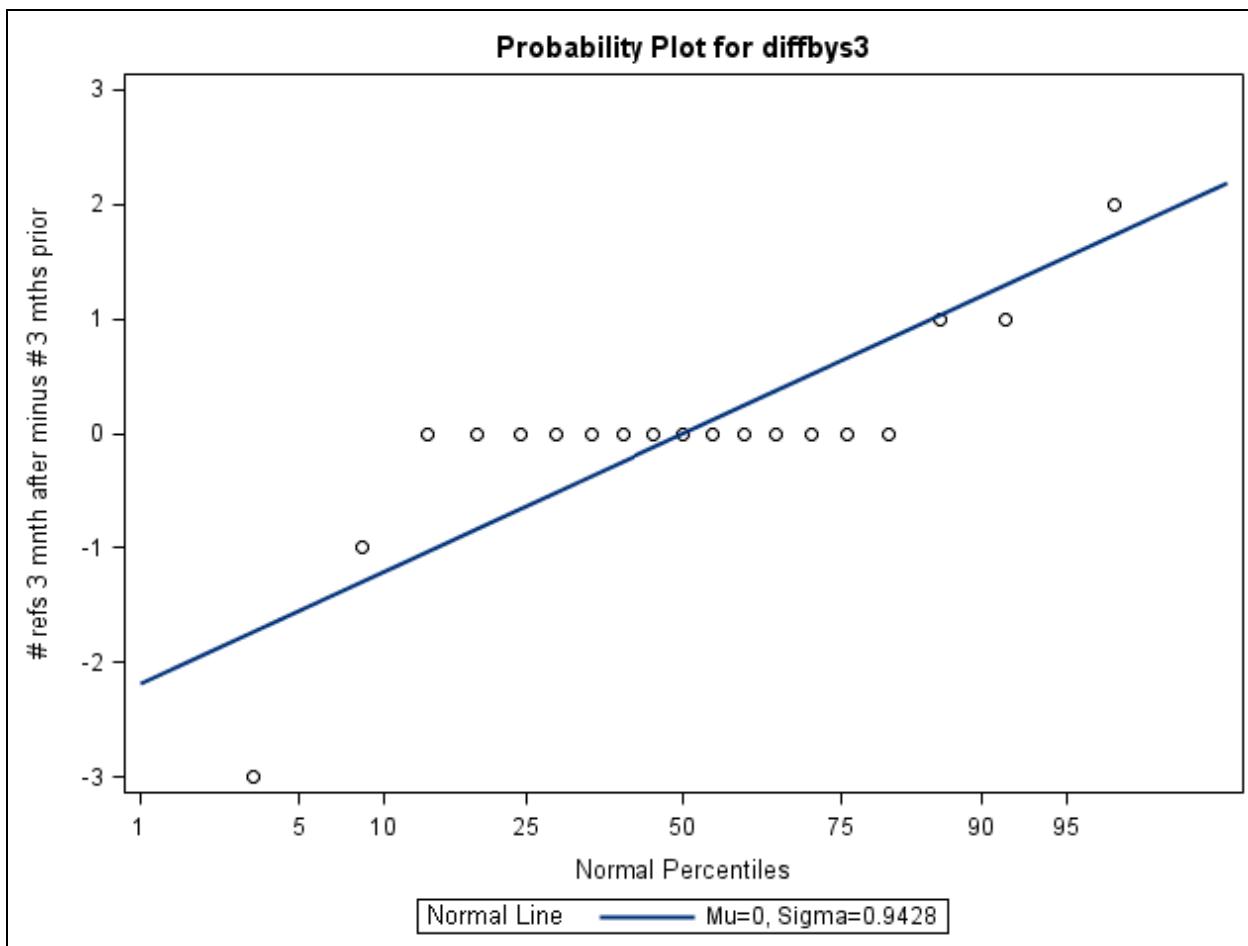




Internal medicine doctors appear to have only three values: 0, 1, and 2. The plots indicate that the data is not normal.

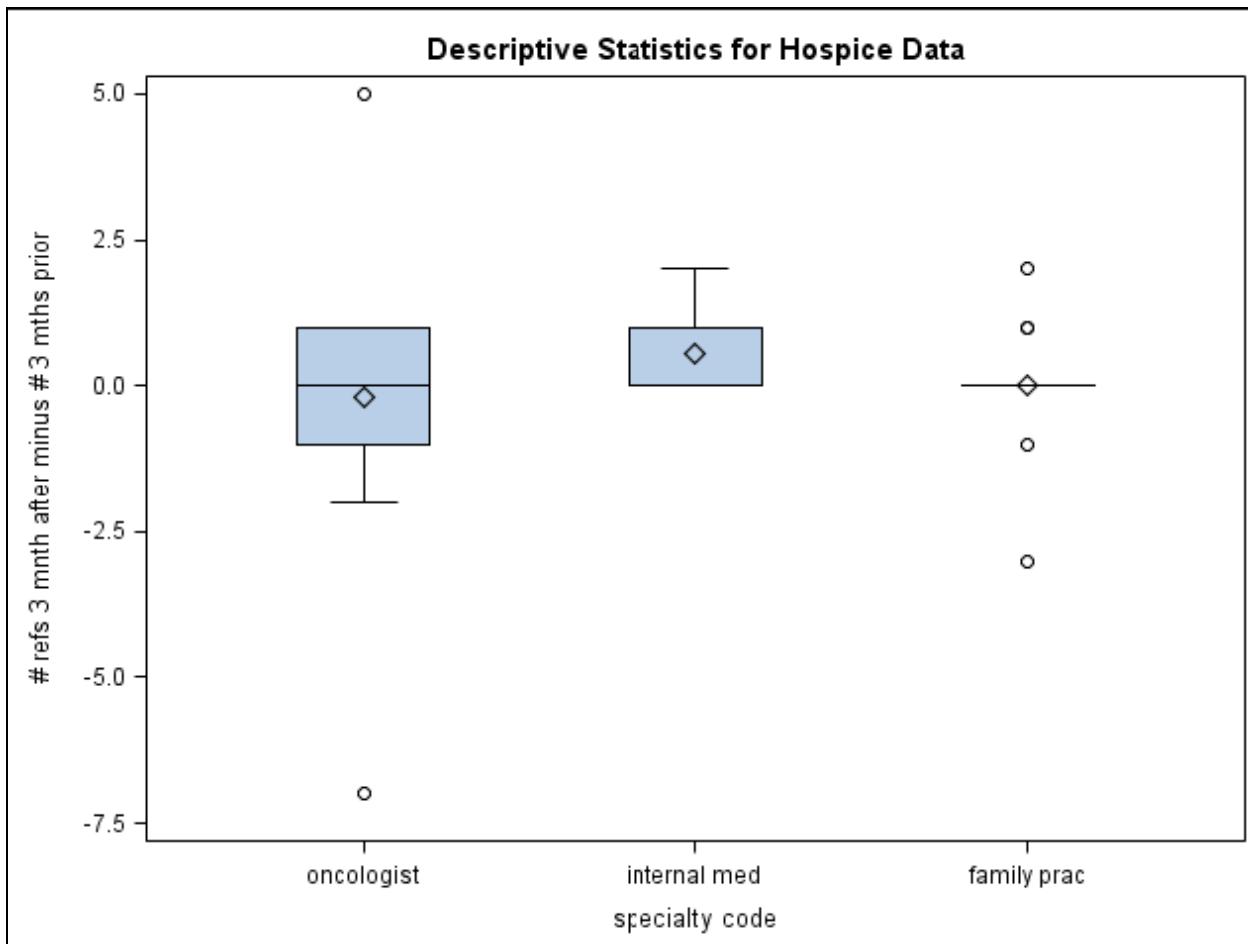
SAS/GRAFH Output (family practice)





Family practice doctors appear to have outliers in the negative direction.

Now examine the PROC SGPlot 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. .

```
/*st00ad06.sas*/
proc npar1way data=sorted_hosp wilcoxon median;
  class code;
  var diffbys3;
run;
```

Selected PROC NPAR1WAY statement options:

WILCOXON requests an analysis of the rank scores. The output includes the Wilcoxon two-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 two-sample test and the median one-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

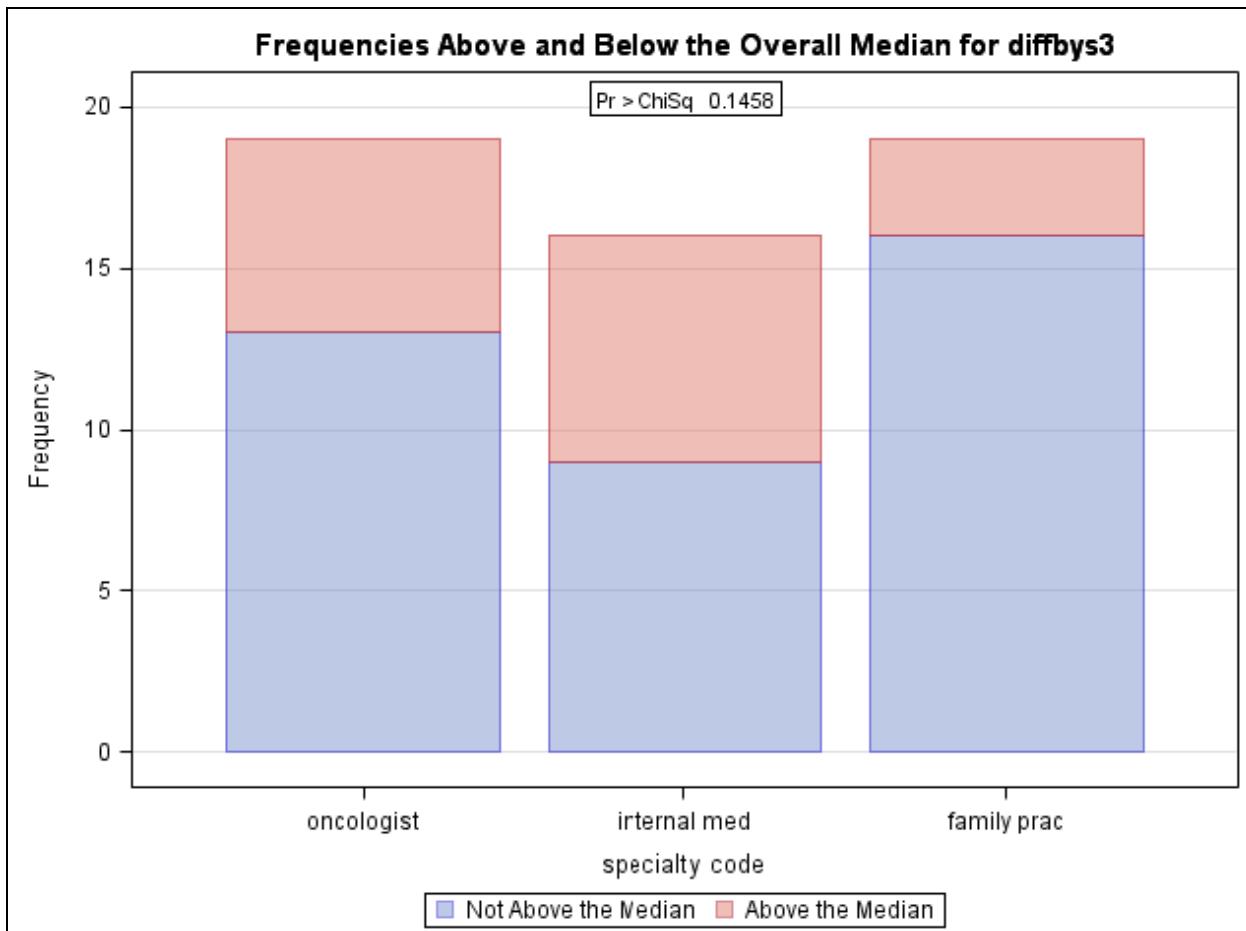
Average scores were used for ties.

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.

PROC NPAR1WAY produces a box plot similar to the one you created for exploratory data analysis. In addition, when you specify the MEDIAN option, a mosaic plot is generated showing the number of observations above and below median for each group:



Example: For an experiment to compare the durability of three brands of synthetic wood veneer, perform nonparametric one-way ANOVA. The data is stored in the **st092.ven** data set.

```
/*st00ad06.sas*/
proc print data= st092.ven;
  title 'Wood Veneer Wear Data';
run;
```

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 four 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.

```
/*st00ad06.sas*/
ods graphics;
proc npar1way data= st092.ven wilcoxon;
  class brand;
  var wear;
  exact;
run;
```

Wood Veneer Wear Data

The NPAR1WAY Procedure

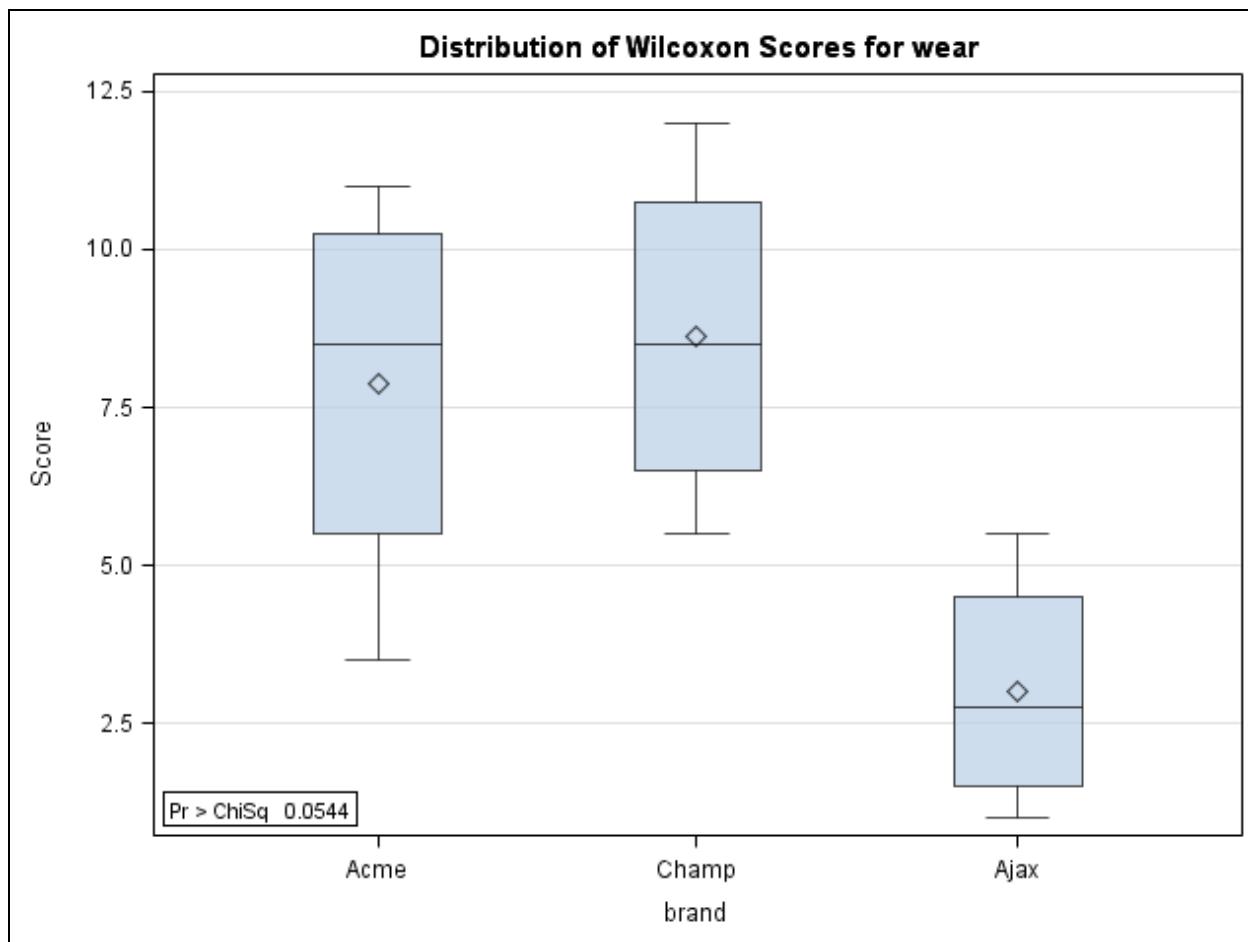
Wilcoxon Scores (Rank Sums) for Variable wear
Classified by Variable brand

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.

The following box plot is produced through ODS Graphics:



A.5 Partial Leverage Plots

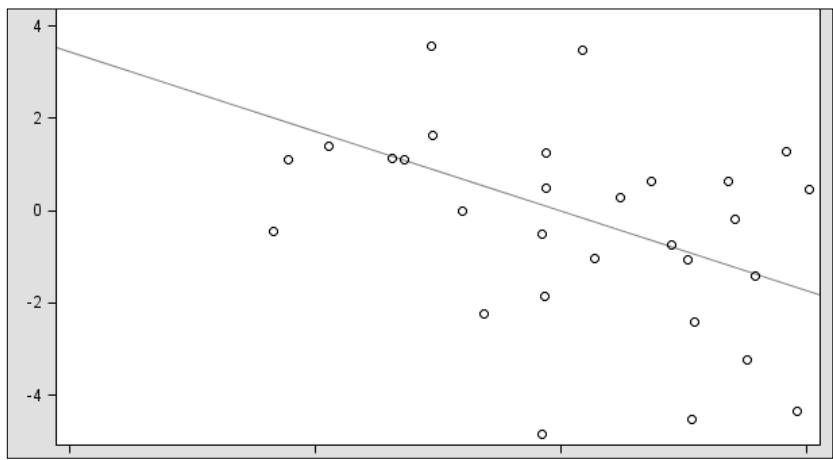
Partial Leverage Plots

- Producing scatter plots of the response (Y) versus each of the possible predictor variables (the Xs) is recommended.
- However, in the multiple regression situation, these plots can be somewhat misleading because Y might depend upon the other Xs not accounted for in the plot.
- Partial leverage plots compensate for this limitation of the scatter plots.

43

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.

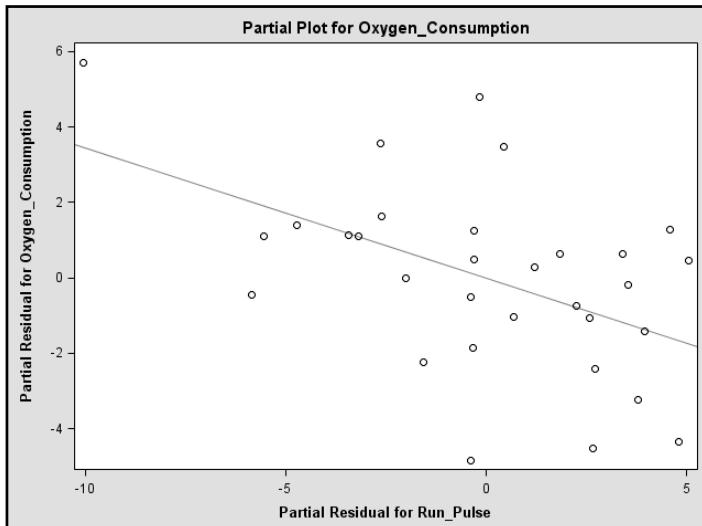
A Scatter Plot



44

In this scatter plot, there are no obvious influential observations.

Example of a Partial Leverage Plot



45

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 are produced automatically with ODS Statistical Graphics when you specify the PLOTS = PARTIAL option in the PROC REG statement.

Partial Leverage Plots

Presume that you are performing a multiple linear regression with Y as the dependent variable and X1, X2, and X3 as the independent variables.

To create a partial leverage plot for X2:

- regress Y on X1 and X3. These residuals are the vertical axis of the partial leverage plot.
- regress X2 on X1 and X3. These residuals are the horizontal axis of the partial leverage plot.

46

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 PREDICT variable model.

```
/*st00ad07.sas*/
ods graphics on / imagemap;
ods html file = 'partial.html';
proc reg data=st092.fitness plots(only) = partial(unpack);
  PREDICT: model Oxygen_Consumption
            = RunTime Age Run_Pulse Maximum_pulse/partial;
  id Name;
  title 'Producing Partial Leverage Plots';
run;
quit;
ods html close;
```

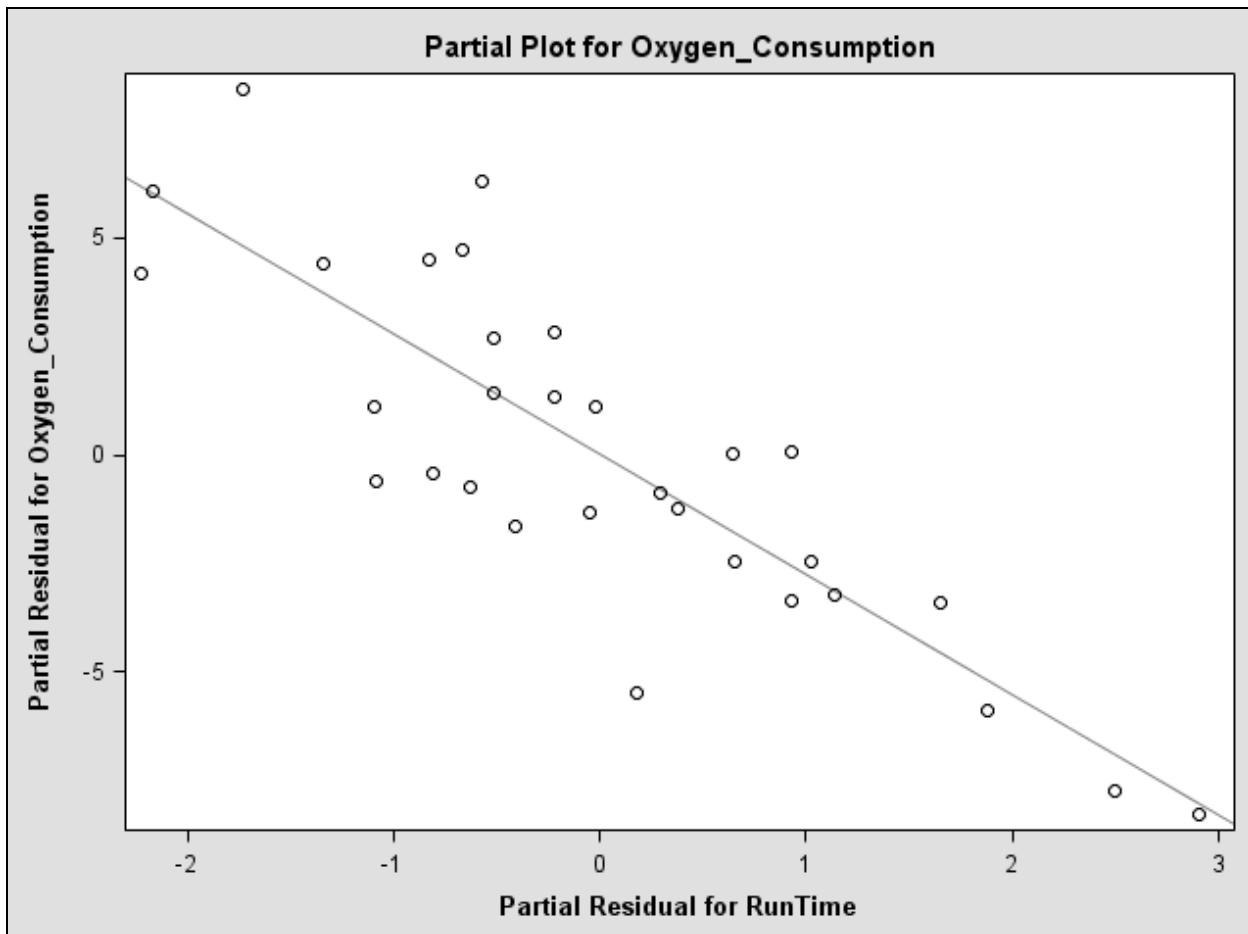
Selected MODEL statement option:

PARTIAL generates partial leverage plots for all predictor variables in the model. If you also specify PLOTS= PARTIAL in the PROC REG statement, ODS Graphics are produced.

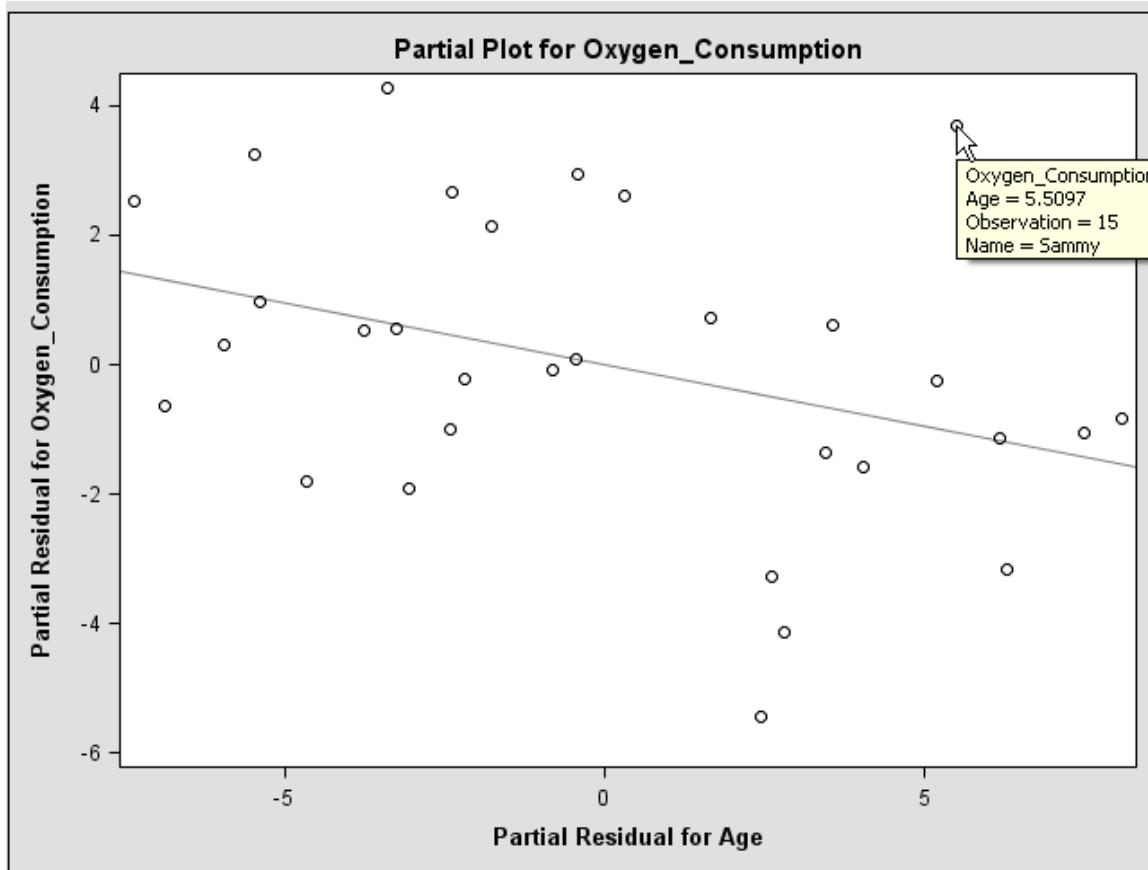
You usually do not look at the INTERCEPT plot.

Partial PROC REG Output

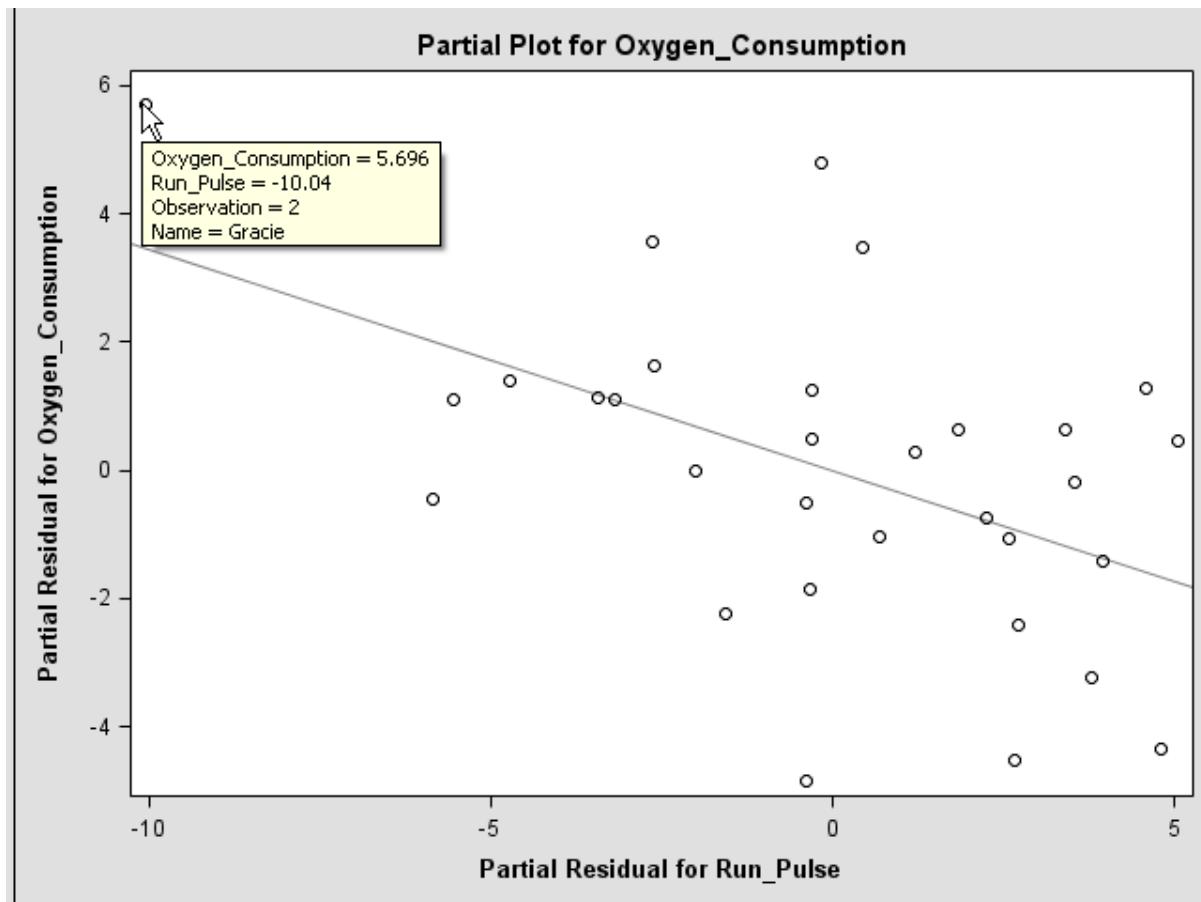
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	97.16952	11.65703	8.34	<.0001
Runtime	1	-2.77576	0.34159	-8.13	<.0001
Age	1	-0.18903	0.09439	-2.00	0.0557
Run_Pulse	1	-0.34568	0.11820	-2.92	0.0071
Maximum_Pulse	1	0.27188	0.13438	2.02	0.0534



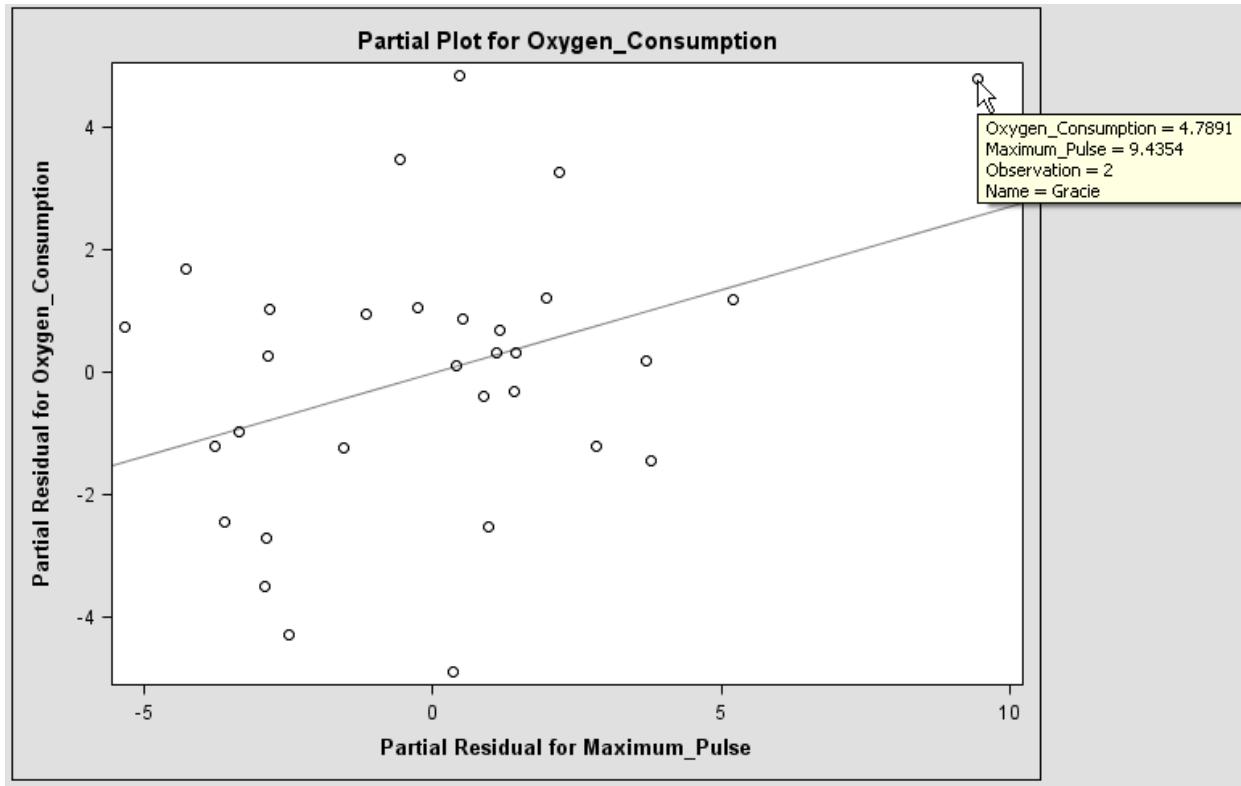
The slope of **runtime** is -2.77576. There do not appear to be any observations that stand out in the **runtime** plot.



Because you used the IMAGEMAP option and an ID statement, you can place your mouse over a data point to see the value of Name displayed as a tag on the plot.



The slope of `run_pulse` is -0.34568. Gracie appears to be influential in the slope of `run_pulse`.



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`.

48

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 B Advanced Programs

B.1 Interaction Plot.....	B-3
---------------------------	-----

B.1 Interaction Plot

```
/*st00bd01.sas*/
proc means data=st092.sales_inc noprint nway;
  class IncLevel Gender;
  var Purchase;
  output out=bins sum(Purchase)=Purchase n(Purchase)=BinSize;
run;
```

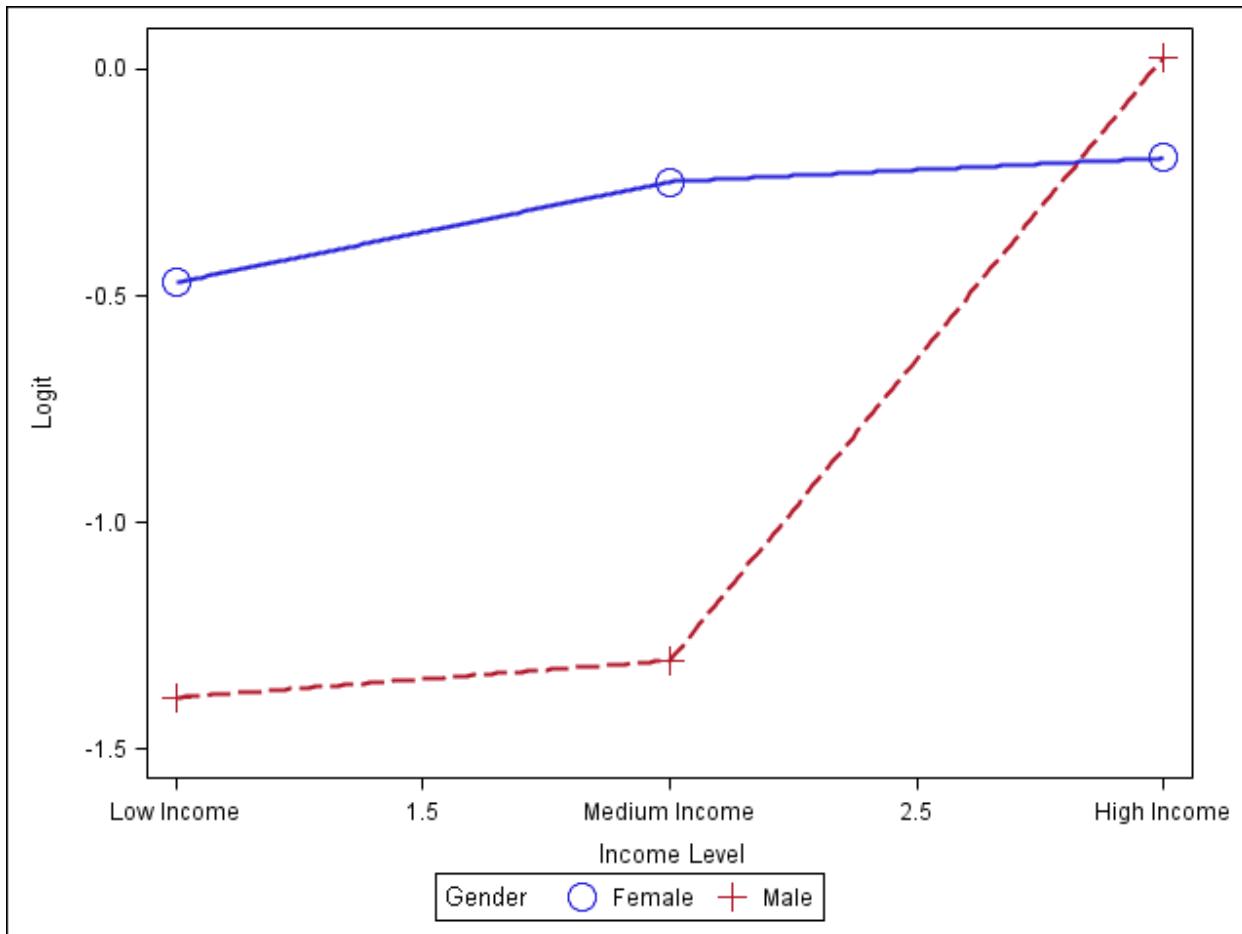
Create summary values for **Purchase** over all values of **IncLevel** and **Gender**, using a CLASS statement. The output data set will contain an observation for each **IncLevel*Gender** combination. The variable **Purchase** will contain the sum of '1' values in the input data set. The variable **BinSize** will contain the number of observations with non-missing values for **Purchase** in each bin.

```
data bins;
  set bins;
  Logit=log( (Purchase+1) / (BinSize-Purchase+1) );
run;
```

Calculate the logit as $\ln((m + 1) / (M - m + 1))$, from the formula in the logistic regression section of the course.

```
ods graphics on;
proc sgscatter data=bins;
  plot Logit*IncLevel /group=Gender markerattrs=(size=15)
    join;
  format IncLevel incfmt.;
  label IncLevel='Income Level';
  title;
run;
quit;
```

Create the plots in PROC SGSCATTER. The GROUP option creates different plotting values for observations based on the level of the **Gender** variable. In order to make the observations easier to see, the option MARKERATTRS=(SIZE=15) was used.



Appendix C Additional Resources

C.1 References	C-3
----------------------	-----

C.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.
- Chatfield, C. (1995), "Model Uncertainty, Data Mining and Statistical Inference," *Journal of the Royal Statistical Society*, 158: 419-466.
- Findley, D.F. and E. Parzen. 1995. "A Conversation with Hirotugu Akaike." *Statistical Science* Vol. 10, No. 1:104-117.
- Freedman, D.A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37: 152-155.
- Hocking, R. R. 1976. "The Analysis and Selection of Variables in Linear Regression." *Biometrics* 32:1-49
- Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression 2nd edition*, New York: John Wiley & Sons.
- Johnson, R. W. 1996. "Fitting percentage of body fat to simple body measurements" *Journal of Statistics Education*, Vol. 4, No. 1.
- 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.
- Raftery, A.E. (1995), "Bayesian Model Selection in Social Research," *Sociological Methodology*.
- 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.
- Shoemaker, A. L. 1996. "What's Normal? -- Temperature, Gender, and Heart Rate." *Journal of Statistics Education*, Vol. 4, No. 2.
- Welch, B. L. 1951. "On the Comparison of Several Mean Values: An Alternative Approach." *Biometrika* 38:330-336.

Appendix D Exercises

Chapter 2	D-3
Chapter 3	D-6
Chapter 4	D-7
Chapter 5	D-11
Chapter 6	D-15
Chapter 7	D-18
Chapter 8	D-21
Chapter 9	D-24
Chapter 10	D-25

Chapter 2

The data in **st092.NORMTEMP** come from an article in the Journal of Statistics Education by Dr. Allen L. Shoemaker at the Psychology Department at Calvin College. They were based on an article in a 1992 JAMA (Journal of the American Medical Association) article questioning the notion that the true mean body temperature was 98.6. There are 65 males and 65 females. There is also some question about whether women's body temperatures truly are the same as men's. The variables in the data set are

ID	Identification number
BODYTEMP	Body temperature (degrees Fahrenheit)
GENDER	Coded (Male, Female)
HEARTRATE	Heart rate (beats per minute)

1. Calculating Basic Statistics in PROC MEANS

The following is a summary of what you will accomplish in this exercise:

- Reinforcing understanding and use of the SAS Help.
- Use the MEANS Procedure to produce simple descriptive statistics.

a) Use PROC MEANS with **BODYTEMP** in the VAR statement to answer these questions:

- i. What is the overall mean and standard deviation of body temperature in the sample?
-

- ii. Interpret the above results- making reference to the data.
-
-

b) Use the SAS Help to explore options on the PROC MEANS STATEMENT.

- i. Which option will “Display the analysis for all requested combinations of class variables”?
-

c) Add a CLASS statement with the variable **GENDER** and use the PRINTALLTYPES option on the PROC MEANS STATEMENT to answer the following.

- i. What is the mean value for Males and the mean value for Females?
-

- iii. Interpret the mean for Males and Females- making reference to the overall mean.
-
-

2. Producing Descriptive Statistics

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of descriptive statistics
 - Use the UNIVARIATE procedure to produce descriptive statistics, histograms and normal probability plots.
 - Use the SGLOT procedure to produce a boxplot.

Use the **st092.NORMTEMP** data set to answer the following:

- a) What are the minimum, the maximum, the mean, and the standard deviation for each of the following variables in the data set? Do the variables appear to be normally distributed?

	BODYTEMP	HEARTRATE
Minimum		
Maximum		
Mean		
Standard Deviation		
Skewness		
Kurtosis		
Distribution: Normal	Yes/No	Yes/No

- b) Create box-and-whisker plots for the **BODYTEMP** and **HEARTRATE** variables. For **BODYTEMP**, display a reference line at 98.6 degrees.
- i. Does the average body temperature seem to be 98.6 degrees?
-

3. Producing Confidence Intervals

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of Confidence Intervals.
 - Verify the assumption for a confidence interval.
 - Use the MEANS Procedure to generate a 95% confidence interval.
 - Interpret the confidence interval.

Use the **st092.NORMTEMP** data set to generate the 95% confidence interval for the mean of **BODYTEMP**.

a) Is the assumption of normality met to produce a confidence interval for **BODYTEMP**?

b) What is the confidence interval for **BODYTEMP**?

c) How do you interpret this interval with regards to the true population mean for body temperature?

Chapter 3

1. Performing a One-Sample *t*-test

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of hypothesis testing and a one-sample *t*-test.
 - Create a hypothesis test.
 - Perform a one-sample *t*-test.
 - Interpret the output.

Using the data, **st092.NORMTEMP**, perform a one-sample *t*-test to determine whether the true mean body temperature is 98.6.

a) What are the Hypotheses?

H_0 : _____

H_1 : _____

b) Set Significance level

α =_____

c) Are the assumptions of the one-sample *t*-test validated in this example?

d) What is the value of the *t*-statistic and the corresponding p-value?

e) How would you interpret these values, i.e. what is your conclusion based on the Decision Rule.

Chapter 4

1. Performing a Two-Sample *t*-test

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of a two-sample *t*-test.
 - Perform a two-sample *t*-test.
 - Interpret the output.

Consider an experiment to study advertising by different media, paper and radio. Perform a two-sample *t*-test to see whether the average sales are significantly different depending on advertising. The **st092.ADS** data set contains data for these variables:

AD type of advertising

SALES level of sales in thousands of dollars

- a) Do the two groups appear to be approximately normally distributed?

- b) Do the two groups have approximately equal variances?

- c) Does the media of paper have a significant effect on sales over the media of radio?

2. One-Way ANOVA – comparing two groups.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of a One-Way ANOVA for comparing two groups.
 - Perform a Hypothesis test.
 - Use the GLM procedure to test the hypothesis.
 - Use the graphs produced by the GLM procedure to verify assumptions.

Re-visit the advertising data, **st092.ADS**, to compare the average sales depending on advertising

- a) 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?

3. One-Way ANOVA – comparing more than two groups.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of a One-Way ANOVA for comparing more than two groups.
 - Perform a Hypothesis test.
 - Use the GLM procedure to test the hypothesis.
 - Use the graphs produced by the GLM procedure to verify assumptions.

The **st092.ALL_ADS** data set contains data for four types of advertising: local newspaper ads, local radio ads, in-store salespeople, and in-store displays.

- a) 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?

4. Post Hoc Pairwise Comparison- Tukey.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of Tukey's post-hoc pairwise comparison.
 - Use the GLM procedure to perform the post hoc pairwise comparison test.
 - Interpret the output.

Adapt the code from the previous exercise to perform Tukey's multiple comparison method to test which mean(s) are significantly different from other (s).

- a) Conduct a pairwise comparison with an experimentwise (use the Tukey method) error rate of $\alpha=0.05$. Which types of advertising are significantly different?

Chapter 5

Percentage of body fat, age, weight, height, and ten body circumference measurements (for example, abdomen) were recorded for 252 men by Dr. Roger W. Johnson of Calvin College in Minnesota*. The data is in the **st092.BODYFAT** data set. Body fat, one measure of health, has been accurately estimated by an underwater weighing technique. There are two measures of percentage body fat in this data set. The following variables are in the data set:

CASE	Case Number
PCTBODYFAT1	Percent body fat using Brozek's equation, $457/\text{Density} - 414.2$
PCTBODYFAT2	Percent body fat using Siri's equation, $495/\text{Density} - 450$
DENSITY	Density (gm/cm ³)
AGE	Age (yrs)
WEIGHT	Weight (lbs)
HEIGHT	Height (inches)
ADIOPOSITY	Adiposity index = Weight/Height ² (kg/m ²)
FATFREEWT	Fat Free Weight = (1-fraction of body fat)*Weight, using Brozek's formula (lbs)
NECK	Neck circumference (cm)
CHEST	Chest circumference (cm)
ABDOMEN	Abdomen circumference (cm) "at the umbilicus and level with the iliac crest"
HIP	Hip circumference (cm)
THIGH	Thigh circumference (cm)

*Due to time constraints some variables have been removed.

1. Describing the Relationship between Continuous Variables

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of descriptive statistics between continuous variables.
 - Use the UNIVARIATE procedure to familiarise yourself with the new data.

Examine the distribution of the variables **PCTBODYFAT2**, **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH**.

- a) What conclusions can you draw about the distribution of these variables?

- b) Do there appear to be any unusual observations?

2. Scatter Plots and Correlation Statistics

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of scatter plots and correlation statistics between continuous variables.
 - Use the CORR procedure to obtain scatter plots and correlation statistics to measure the strength of the linear relationship between the target variable vs. predictor variables.

Generate scatter plots and correlations for the VAR variables, **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH** versus the WITH variable, **PCTBODYFAT2**.

 Please note! ODS Graphics in PROC CORR limits you to 10 VAR variables at a time. Correlation tables can be created using more than 10 VAR variables at a time.

a) Can straight lines adequately describe the relationships?

b) Are there any outliers you should investigate?

c) What variable has the highest correlation with **PCTBODYFAT2**?

d) What is the p-value for the coefficient of the variable that has the highest correlation with **PCTBODYFAT2**?

- e) Is it statistically significant at the 0.05 level?

3. More Scatter Plots and Correlation.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of scatter plots and correlation statistics between continuous variables.
 - Use the CORR procedure to obtain scatter plots and correlation statistics to measure the strength of the linear relationship between the predictor variables.

Generate correlations among all of the VAR variables **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH**.

- a) Are there any notable relationships?

Chapter 6

1. Fitting a Simple Linear Regression Model

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of Simple Linear Regression by using the REG procedure to;
 - Produce a simple linear regression model.
 - Analyse the output.

Use the **st092.BODYFAT** data set to perform a simple linear regression model with **PCTBODYFAT2** as the response variable and **ABDOMEN** as the predictor.

- a) What is the value of the F statistic and the associated p-value? How would you interpret this with regards to the null hypothesis?

- b) Write out the predicted regression equation.

- c) What is the value of the R^2 statistic? How would you interpret this?

2. Confidence and Prediction intervals.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of Confidence and Prediction Intervals by using the REG procedure to produce Confidence and Prediction Intervals.

- a) What is the Confidence Interval when **ABDOMEN** is 83 (see observation number 2) and how would you interpret this?

- b) What is the Prediction Interval when **ABDOMEN** is 83 (see observation number 2) and how would you interpret this?

3. Predicted values

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of producing predicted values by using the REG procedure to produce Predicted Values for **PCTBODYFAT2**.

Produce predicted values for **PCTBODYFAT2** when **ABDOMEN** is 80, 100 and 120

 The **ABDOMENPRED** data set in **st092** contains the 3 observations needed.

- a) What are the predicted values when **ABDOMEN** is 80, 100 and 120?

- b) Is it appropriate to predict **PCTBODYFAT2** when **ABDOMEN** is 200?

4. Examining Residuals

The following is a summary of what you will accomplish in this exercise:

- Use the graphs created in the REG procedure to verify the assumptions for the Simple Linear Regression model built in the previous exercise.

Assess the model obtained from using **ABDOMEN** as a predictor variable for **PCTBODYFAT2**. Create plots of the residuals by **ABDOMEN**, and by the predicted values, and a normal Quantile-Quantile plot.

- a) Do the residual plots indicate any problems with the constant variance assumption?

- b) Are there any outliers indicated in the residual plots?

- c) Does the quantile-quantile plot indicate any problems with the normality assumption?

Chapter 7

1. Performing a Regression Using the REG Procedure

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of Multiple Linear Regression by using the REG procedure to:
 - Create a Multiple Linear Regression model.
 - Re-run and evaluate the model created to eliminate unnecessary variables efficiently.

Using the **st092.BODYFAT** data set, run a regression of **PCTBODYFAT2** on the variables **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH**.

a) Compare the output with the output from the model with only **ABDOMEN** -in the previous exercise.

- i. What is different in the ANOVA tables?

- ii. How do the R^2 and the adjusted R^2 compare with these statistics for the **ABDOMEN** regression demonstration?

- iii. Did the estimate for the intercept change? Did the estimate for the coefficient of **ABDOMEN** change?

b) Simplifying the Model

- i. Rerun the model in a), but eliminate the variable with the highest p -value. Compare the output with the Exercise a) model.

ii. Did the p-value for the model change?

iii. Did the R^2 and adjusted R^2 change?

iv. Did the parameter estimates and their p-values change?

c) More Simplifying of the Model

i. Rerun the model in Exercise b), but drop the variable with the highest p -value.

ii. How did the output change from the previous model?

iii. Did the number of parameters with a p-value less than 0.05 change?

2. Using Stepwise Selection

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of stepwise selection methods by using the REG procedure to;
 - Select a candidate model using the FORWARD stepwise regression method.
 - Select a candidate model using the BACKWARD stepwise regression method.
 - Select a candidate model using the STEPWISE stepwise regression method.

Use the **st092.BODYFAT** data set to identify a set of “best” models.

- a) Use a stepwise regression method to select a candidate model; try FORWARD, STEPWISE, and BACKWARD.
- b) How many variables would have resulted from a model using FORWARD selection and a significance level for entry criterion of 0.05, instead of the default SLENTRY of 0.50?

- c) Verify the assumptions of the model found in b)

Chapter 8

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 **st092.SAFETY** data set.

The variables in the data set are as follows:

UNSAFE	dichotomized safety score (1=Below Average, 0=Average or Above)
TYPE	type of car (Large, Medium, Small, Sport/Utility, Sports)
REGION	manufacturing region (Asia, N America)
WEIGHT	weight in 1000's of pounds
SIZE	trichotomized version of TYPE (1=Small or Sports, 2=Medium, 3=Large or Sport/Utility).

1. Describing Categorical Data

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of describing categorical data by using the FREQ procedure to create one-way frequencies.

- a) What is the measurement scale of each variable?

<u>Variable</u>	<u>Measurement Scale</u>
UNSAFE	_____
TYPE	_____
REGION	_____
WEIGHT	_____
SIZE	_____

- b) Create one-way frequency tables for the variables **UNSAFE**, **TYPE**, and **REGION**.

- i. What is the proportion of cars made in North America?

- ii. For the variables **UNSAFE**, **TYPE**, and **REGION**, are there any unusual data values that warrant further investigation?

2. Performing Tests and Nominal Measures of Association.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of cross tabulations and the chi-squared test.
 - Use the FORMAT procedure to creating a temporary format.
 - Use the FREQ procedure to generate expected frequencies and a chi-square test of association.

- a) Generate a temporary format called **safefmt** to clearly identify the values of **UNSAFE**. Use the following information: 0='Average or Above Safety', 1='Below Average Safety'.
- b) Use the **st092.SAFETY** data set to calculate what percentage of cars made in Asia had a below-average safety score? Ensure the **safefmt** format is assigned.

- c) Use the FREQ procedure to perform an appropriate measure of association test between **REGION** and **UNSAFE**. Do you see a statistically significant association (at the 0.05 level)?

3. Performing Tests and Ordinal Measures of Association.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of an ordinal chi-squared test.
 - Use the FREQ procedure to generate an ordinal measure of association

Use the **st092.SAFETY** data set to examine the ordinal association between **SIZE** and **UNSAFE**.

- a) What statistic should you use to detect an ordinal association between **SIZE** and **UNSAFE**?

- b) Do you reject or fail to reject the null hypothesis at the 0.05 level?

- c) What is the strength of the ordinal association between **SIZE** and **UNSAFE**?

Chapter 9

1. Performing a Logistic Regression Model.

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of logistic regression by using the LOGISTIC procedure to;
 - Build and interpret a logistic regression model.
 - Use Reference Cell Coding.
 - Model the correct probability.

Fit a simple logistic regression model using **st092.SAFETY** with **UNSAFE** as the outcome variable and **REGION** as the predictor variable. Request reference cell coding with **Asia** as the reference level. Model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

- a) Do you reject or fail to reject the null hypothesis that all regression coefficients of the model are 0?

- b) Write out the logistic regression equation.

- c) Interpret the odds ratio for **REGION**.

Chapter 10

1. Fitting a Multiple Logistic Regression Model

The following is a summary of what you will accomplish in this exercise:

- Reinforce understanding of logistic regression by using the LOGISTIC procedure to build and interpret a multiple logistic regression model.
- Use the FORMAT procedure to generate a format.

Use **UNSAFE** as the outcome variable and **WEIGHT**, **SIZE**, and **REGION** as the predictor variables. Ensure to model the probability of below-average safety scores.

- a) Create a format called **sizefmt** to indicate 1= Small, 2= Medium, 3=Large
 - b) Use the LOGISTIC procedure and specify **REGION** and **SIZE** as classification variables using reference cell coding. Specify Asia as the reference level for **REGION**. For the **SIZE** variable, apply the **sizefmt** and specify Small as the reference level. Request any relevant plots.
-  The variable **SIZE** is coded (1, 2, 3), but the applied format requires that the formatted value be used in the CLASS statement for the REF= category.

Use **Unsafe** as the outcome variable and **Weight**, **Size**, and **Region** as the predictor variables. Use the EVENT= option to model the probability of below-average safety scores.

- i. Interpret the odds ratio estimates for **WEIGHT** and **SIZE**, in the Profile Likelihood Table.

- ii. Do you think this is a better model than the one fit with just **REGION**?

Appendix E Solutions

Chapter 2	E-3
Chapter 3	E-11
Chapter 4	E-13
Chapter 5	E-24
Chapter 6	E-31
Chapter 7	E-36
Chapter 8	E-47
Chapter 9	E-52
Chapter 10	E-55

Chapter 2

1. Calculating Basic Statistics in PROC MEANS

- a) Use PROC MEANS with **BODYTEMP** in the VAR statement to answer these questions:

```
/*st002s01.sas a)*/
proc means data=st092.normtemp;
  var BodyTemp;
  title 'Descriptive Statistics for Body Temp';
run;
```

PROC MEANS Output

Descriptive Statistics for Body Temp				
The MEANS Procedure				
Analysis Variable : BodyTemp				
N	Mean	Std Dev	Minimum	Maximum
130	98.2492308	0.7331832	96.3000000	100.8000000

- i. What is the overall mean and standard deviation of body temperature in the sample?

The overall mean is 98.25. The overall standard deviation is 0.73

- ii. Interpret the above results- making reference to the data and the notion of the true mean.

The average body temperature for our sample of 130 people is 98.25, which is 0.35 units away from the true mean stated at 98.6. The standard deviation suggests that, on average, the data points lie 0.73 degrees Fahrenheit away from the sample mean.

- b) Use the SAS Help to explore options on the PROC MEANS STATEMENT.

Help for the PROC MEANS STATEMENT can be found in the SAS Help.

Help>SAS Help and Documentation.

On the contents Tab navigate to:

SAS Products>Base SAS>Base SAS 9.2 Procedures Guide>Procedures>The MEANS Procedure> Syntax: MEANS Procedure>PROC MEANS Statement.

- i. Which option will “Display the analysis for all requested combinations of class variables”?

PrintAllTypes.

- c) Add a CLASS statement with the variable **GENDER** and use the PRINTALLTYPES option on the PROC MEANS STATEMENT to answer the following.

```
/*st002s01.sas c*/
proc means data=st092.normtemp printalltypes;
  var BodyTemp;
  class Gender;
  title 'Selected Descriptive Statistics for Body Temp';
run;
```

PROC MEANS Output

Selected Descriptive Statistics for Body Temp					
The MEANS Procedure					
Analysis Variable : BodyTemp					
Obs	N	Mean	Std Dev	Minimum	Maximum
130	130	98.2492308	0.7331832	96.3000000	100.8000000

Analysis Variable : BodyTemp						
Gender	Obs	N	Mean	Std Dev	Minimum	Maximum
Female	65	65	98.3938462	0.7434878	96.4000000	100.8000000
Male	65	65	98.1046154	0.6987558	96.3000000	99.5000000

- i. What is the mean value for Males and the mean value for Females?

The mean values of 98.39 for females and 98.10 for males were found.

- iii. Interpret the mean for Males and Females- making reference to the overall mean.

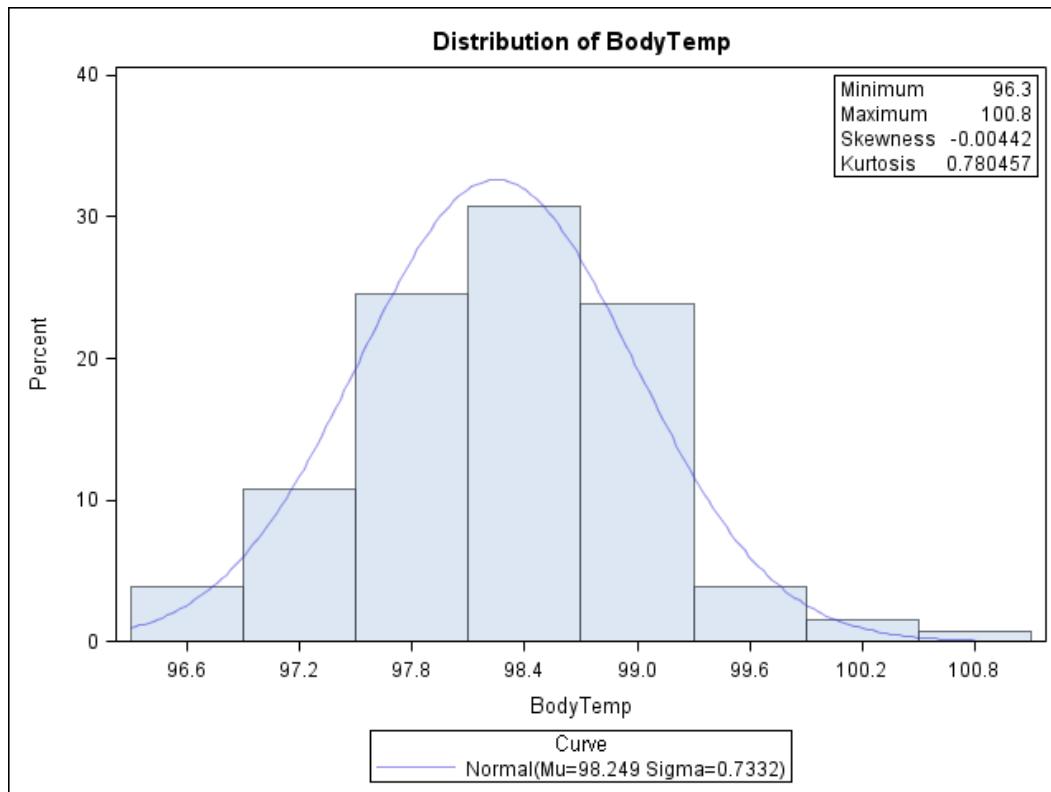
According to our sample, Females have a slightly higher average body temperature than Males. Both groups have a smaller average body temperature than that stated as the true mean, 98.6.

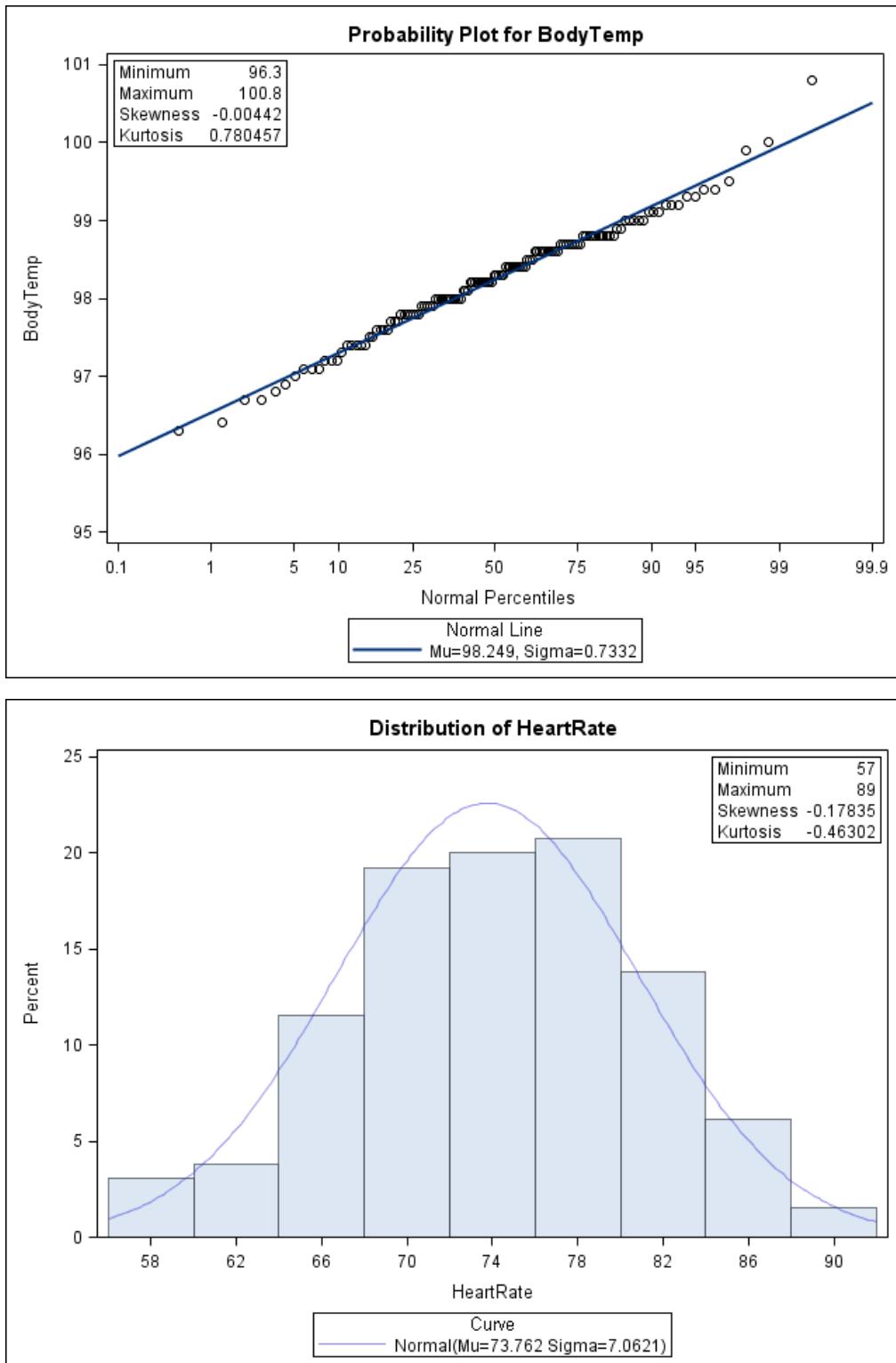
2. Producing Descriptive Statistics

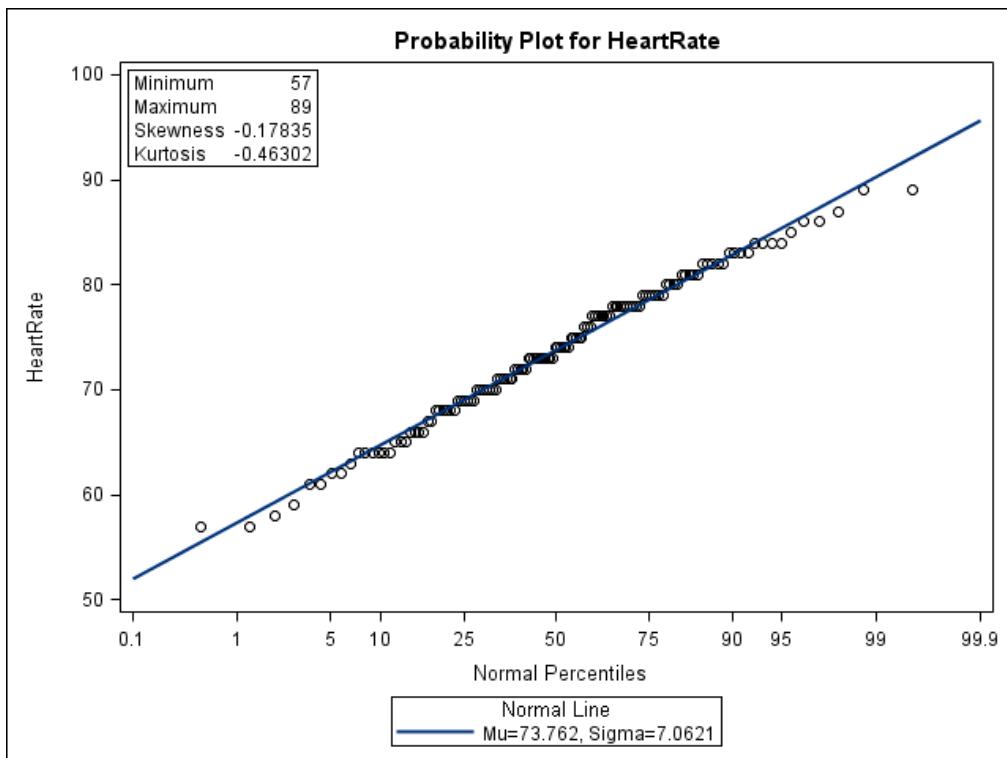
- a) What are the minimum, the maximum, the mean, and the standard deviation for each of the following variables in the data set? Do the variables appear to be normally distributed?

```
ods graphics;
proc univariate data=st092.NormTemp noint;
var BodyTemp HeartRate;
histogram BodyTemp HeartRate / normal(mu=est sigma=est
                                         noint);
inset min max skewness kurtosis / position=ne;
probplot BodyTemp HeartRate / normal(mu=est sigma=est);
inset min max skewness kurtosis;
title 'Descriptive Statistics Using PROC UNIVARIATE';
run;
```

 The NOPRINT option in both the PROC UNIVARIATE and HISTOGRAM statements suppress the printing of the tabular output. Because the statistics are being reported in the insets of the plots, they are not needed in the output tables.





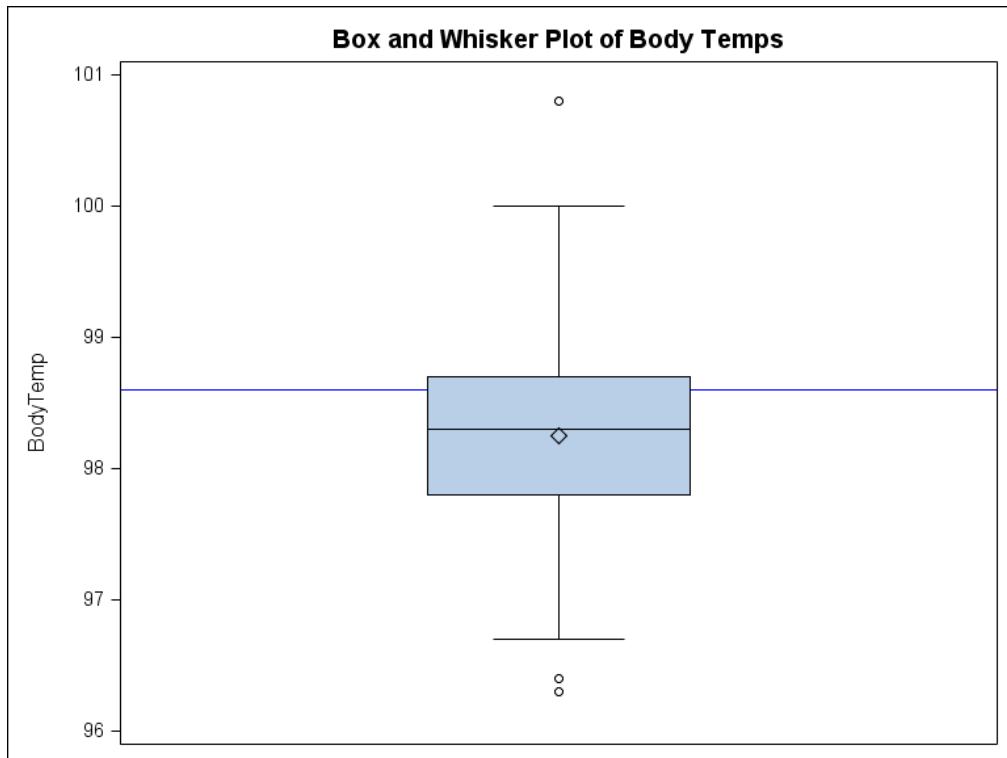


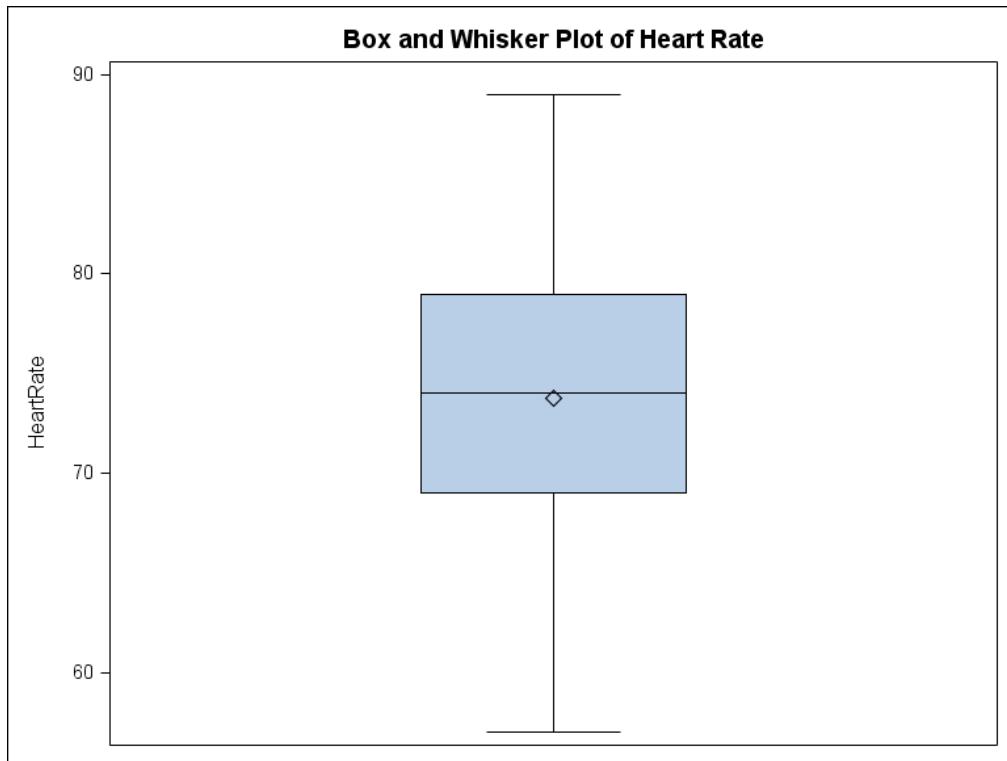
The distributions for both variables look approximately normal.

	BODYTEMP	HEARTRATE
Minimum	96.3	57
Maximum	100.8	89
Mean	98.249	73.762
Standard Deviation	0.7332	7.0621
Skewness	-0.00442	-0.17835
Kurtosis	0.780457	-0.46302
Distribution: Normal	Yes	Yes

- b) Create box-and-whisker plots for the **BODYTEMP** and **HEARTRATE** variables. For **BODYTEMP**, display a reference line at 98.6 degrees.

```
ods graphics on;
proc sgplot data=st092.NormTemp;
refline 98.6 / axis=y lineattrs=(color=blue);
vbox BodyTemp;
title "Box and Whisker Plot of Body Temps";
run;
proc sgplot data=st092.NormTemp;
vbox HeartRate;
title "Box and Whisker Plot of Heart Rate";
run;
```





- i. Does the average body temperature seem to be 98.6 degrees?

The average body temperature seems to be somewhat less than 98.6 degrees.

3. Producing Confidence Intervals

Use the **st092.NORMTEMP** data set to generate the 95% confidence interval for the mean of **BODYTEMP**.

```
/*st002s03.sas*/
proc means data=st092.NormTemp maxdec=2
            n mean stderr clm;
  var BodyTemp;
  title '95% Confidence Interval for Body Temp';
run;
```

95% Confidence Interval for Body Temp

The MEANS Procedure

Analysis Variable : BodyTemp

N	Mean	Std Error	Lower 95%	Upper 95%
CL for Mean	CL for Mean			
130	98.25	0.06	98.12	98.38

- a) Is the assumption of normality met to produce a confidence interval for these data?

Yes. Because the sample size is large enough and because the data values seemed to be normally distributed, the normality assumption seems to hold (see Chapter 2 Exercise 2).

- b) What is the confidence interval for **BODYTEMP**?

The 95% confidence interval is 98.12 to 98.38 degrees Fahrenheit.

- c) How do you interpret this interval with regards to the true population mean for body temperature?

We are 95% confident that the true mean body temperature for the population of all people in the world is somewhere between 98.12 and 98.38 degrees.

Chapter 3

1. Performing a One-Sample *t*-test

Perform a one-sample *t*-test to determine whether the true mean body temperature is 98.6.

```
/*st003s01.sas*/
ods graphics on;
proc ttest data=st092.normtemp h0=98.6;
  var bodytemp;
  title "Testing the true mean of body-temp is 98.6";
run;
ods graphics off;
```

- a) What are the Hypotheses?

$$H_0: \mu = 98.6$$

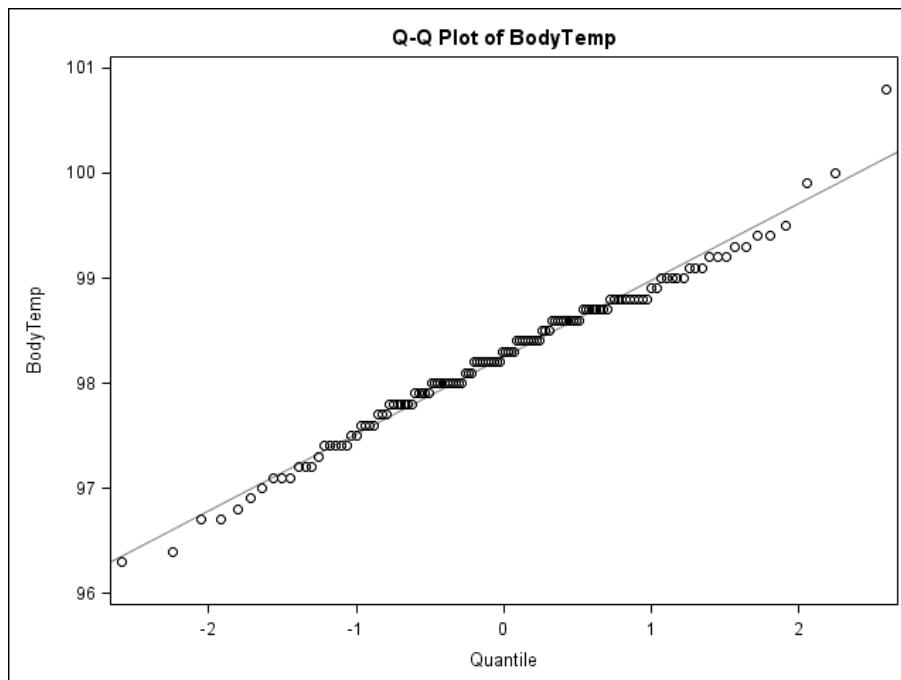
$$H_1: \mu \neq 98.6$$

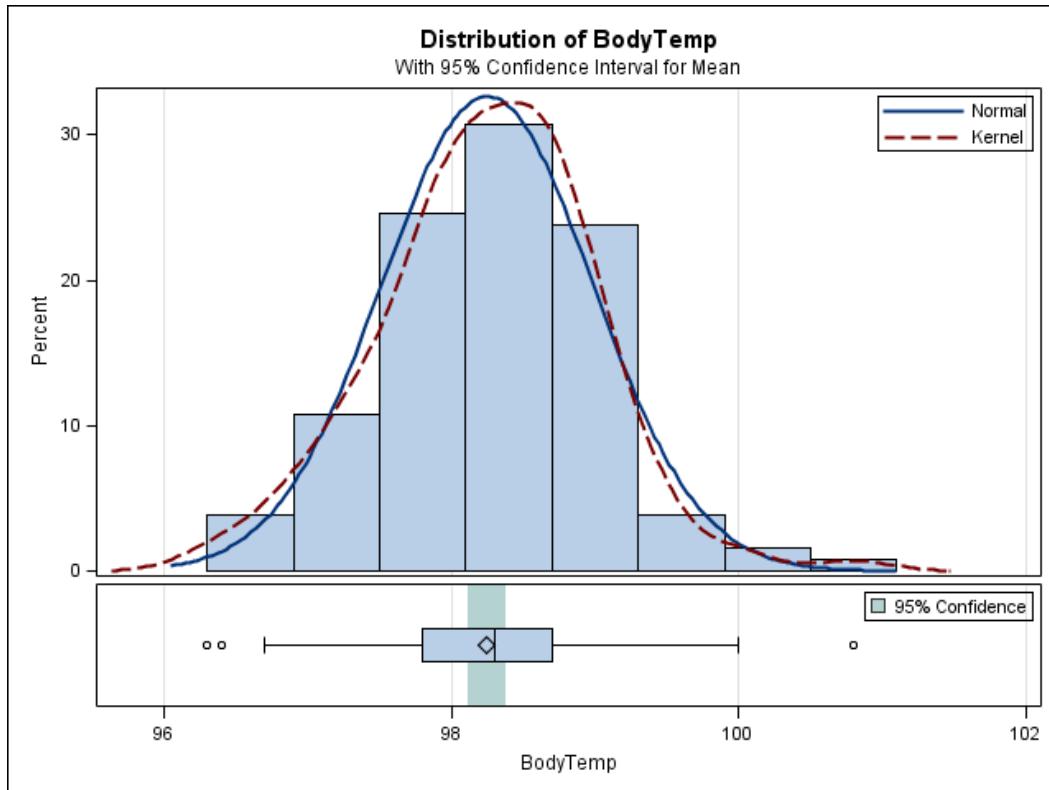
- b) Set Significance level

$$\alpha = 0.05$$

- c) Are the assumptions of the one-sample *t*-test validated in this example?

The normality assumption can be verified by invoking the central limit theorem (130 observations) or by viewing the Q-Q plot and Summary Plot. Assume the independent observations assumption is verified throughout the course.





- d) What is the value of the t -statistic and the corresponding p-value?

Testing the true mean of body-temp is 98.6							
The TTEST Procedure							
Variable: BodyTemp							
N	Mean	Std Dev	Std Err	Minimum	Maximum		
130	98.2492	0.7332	0.0643	96.3000	100.8		
Mean	95% CL Mean	Std Dev	95% CL Std Dev				
98.2492	98.1220 98.3765	0.7332	0.6536 0.8350				
DF	t Value	Pr > t					
129	-5.45	<.0001					

The t -statistic is -5.45 and the p-value is <.0001.

- e) How would you interpret these values, i.e. what is your conclusion based on the Decision Rule.

As $p < \alpha$, Reject H_0 , therefore there is no evidence to support that the true mean of Body Temperature is 98.6. The t -statistic suggests that the difference is -5.44 standard errors away from zero. As the t -statistic is negative, the hypothesised value is higher than the sample mean.

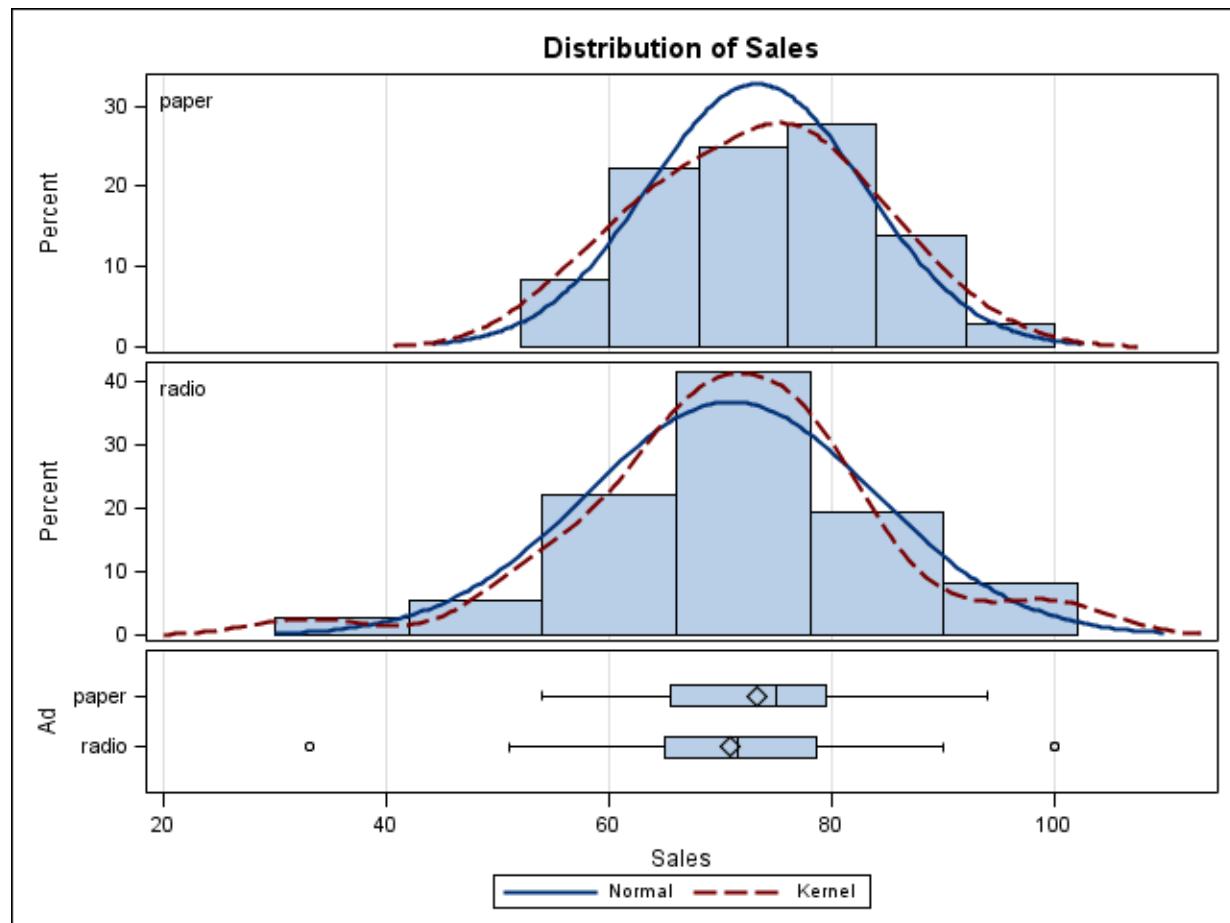
Chapter 4

1. Performing a Two-Sample *t*-test

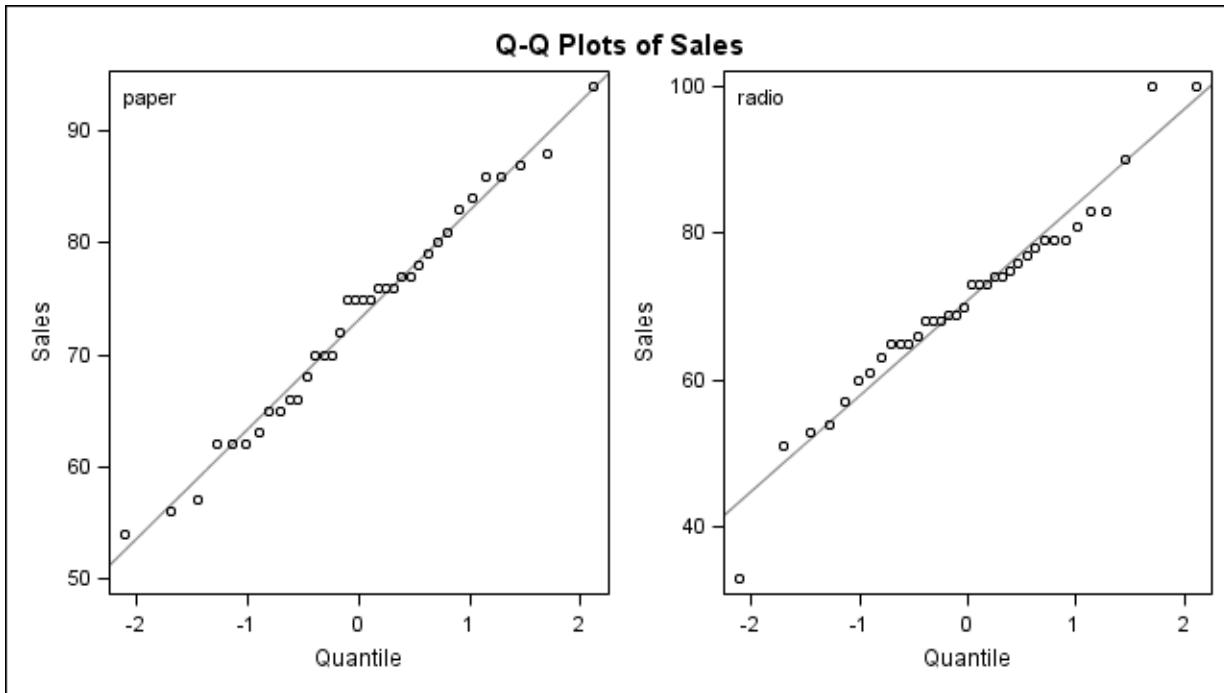
Perform a two-sample *t*-test to see whether the average sales are significantly different depending on advertising.

```
/*st004s01.sas*/
ods graphics on;
proc ttest data=st092.ads plots(shownull)=interval;
  class ad;
  var sales;
  title "Two-Sample t-test Comparing Media Types on Sales";
run;
ods graphics off;
```

- a) Do the two groups appear to be approximately normally distributed?



The SummaryPanel plot shows that the data is relatively normally distributed. Verify by looking at the Q-QPlot.



The data follows the line, the normality assumption is satisfied.

b) Do the two groups have approximately equal variances?

The variability assumption needs to be verified from the output of the *t*-test.

Equality of Variances					
Method	Num DF	Den DF	F Value	Pr > F	
Folded F	35	35	1.77	0.0942	

Step 1- Set Hypothesis

$$H_0: \sigma_{\text{paper}}^2 = \sigma_{\text{radio}}^2$$

$$H_1: \sigma_{\text{paper}}^2 \neq \sigma_{\text{radio}}^2$$

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$\text{p-value}=0.0942$$

Step 4- Decision Rule.

The p-value > α , Fail to reject H_0 , therefore there is no evidence to say the variances are not equal.

c) Does the media of paper have a significant effect on sales over the media of radio?

The TTEST Procedure						
Variable: Sales						
Ad	N	Mean	Std Dev	Std Err	Minimum	Maximum
paper	36	73.2222	9.7339	1.6223	54.0000	94.0000
radio	36	70.8889	12.9676	2.1613	33.0000	100.0
Diff (1-2)		2.3333	11.4653	2.7024		
Ad	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev	
paper		73.2222	69.9287 76.5157	9.7339	7.8950	12.6973
radio		70.8889	66.5013 75.2765	12.9676	10.5178	16.9154
Diff (1-2)	Pooled	2.3333	-3.0564 7.7231	11.4653	9.8406	13.7377
Diff (1-2)	Satterthwaite	2.3333	-3.0638 7.7305			
Method	Variances	DF	t Value	Pr > t		
Pooled	Equal	70	0.86	0.3909		
Satterthwaite	Unequal	64.937	0.86	0.3911		

Step 1- Set Hypothesis

$$H_0: \mu_{\text{paper}} = \mu_{\text{radio}}$$

$$H_1: \mu_{\text{paper}} \neq \mu_{\text{radio}}$$

Step 2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$\text{p-value}=0.3909.$$

Step 4- Decision Rule.

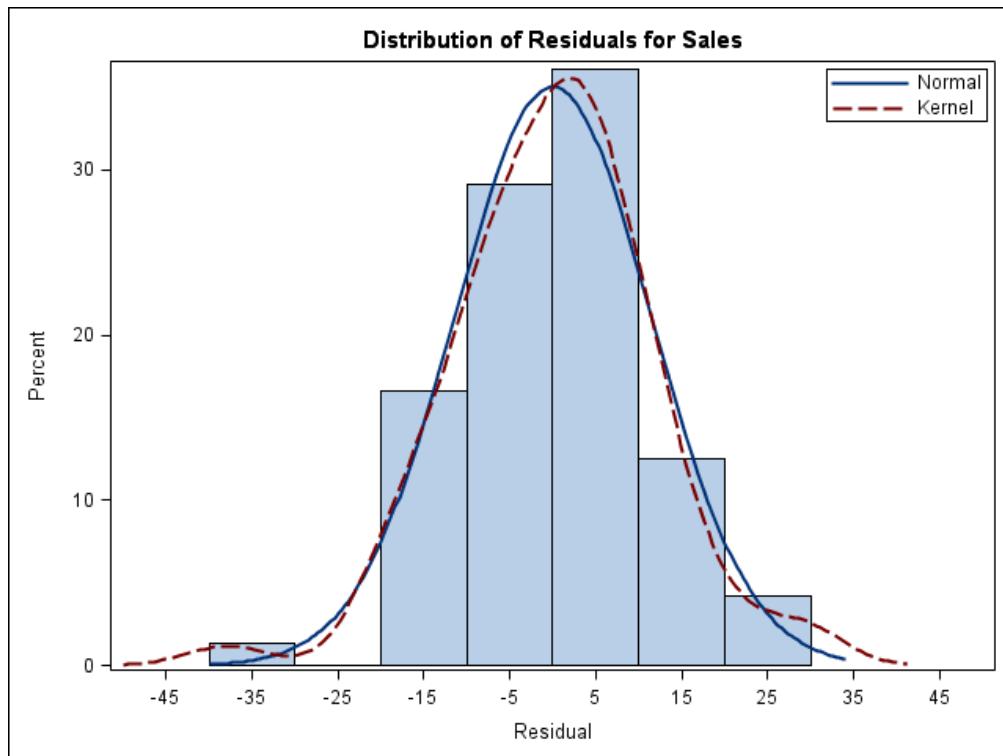
The p-value> α , Fail to reject H_0 , therefore, there is no statistical significant difference between sales when advertising via paper or radio.

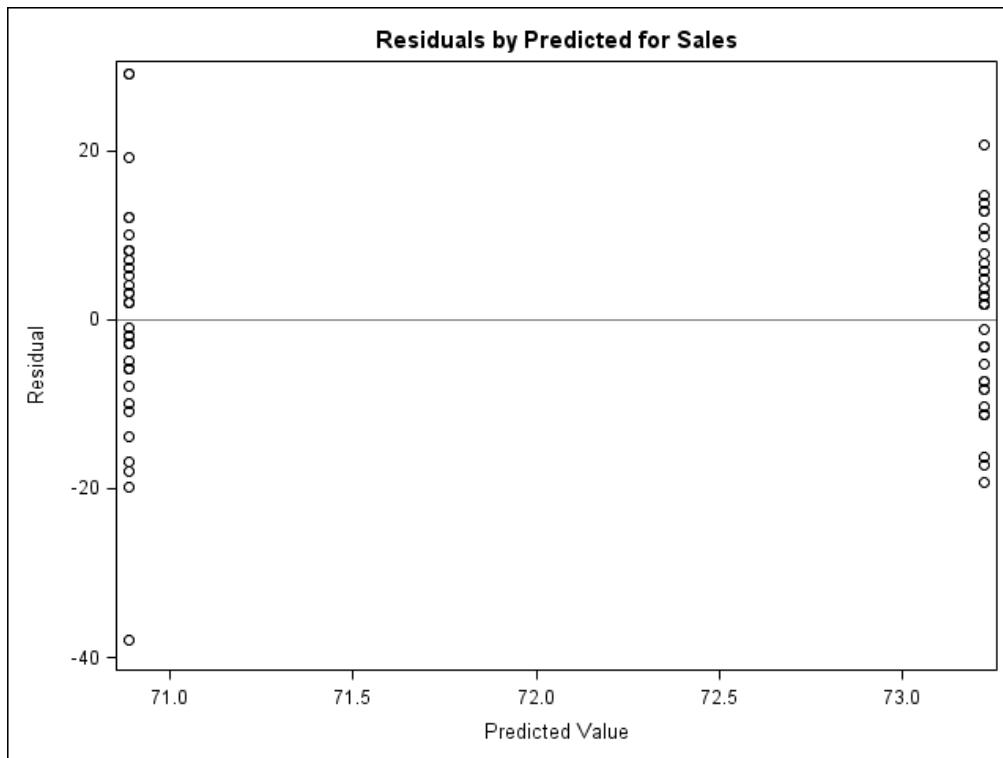
2. One-Way ANOVA – comparing two groups.

- a) 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?

```
/*st004s02.sas*/
ods graphics on;
proc glm data=st092.ads PLOTS(only)=diagnostics(unpack);
  class ad;
  model sales=ad;
  means ad / hovtest;
  title 'Testing for Equality of Means with PROC GLM';
run;
quit;
ods graphics off;
```

Firstly, check that the assumptions of the One-Way ANOVA are verified.





The graphs do not show strong evidence against normality.

Use the Levene's Test for Homogeneity

Testing for Equality of Means with PROC GLM
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	1	91687.1	91687.1	1.80	0.1841

Step 1- Set Hypothesis

$$H_0: \sigma_{\text{paper}}^2 = \sigma_{\text{radio}}^2$$

$$H_1: \sigma_{\text{paper}}^2 \neq \sigma_{\text{radio}}^2$$

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$\text{p-value}=0.1841$$

Step 4- Decision Rule.

The p-value > α , Fail to reject H_0 , therefore there is no evidence to say the variances are not equal.

Testing for Equality of Means with PROC GLM						
The GLM Procedure						
Class Level Information						
Class		Levels	Values			
Ad		2	paper radio			
Number of Observations Read			72			
Number of Observations Used			72			
Testing for Equality of Means with PROC GLM						
The GLM Procedure						
Dependent Variable: Sales						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	98.000000	98.000000	0.75	0.3909	
Error	70	9201.777778	131.453968			
Corrected Total	71	9299.777778				
R-Square Coeff Var Root MSE Sales Mean						
	0.010538	15.91180	11.46534	72.05556		

Step 1- Set Hypothesis

$$H_0: \mu_{\text{paper}} = \mu_{\text{radio}}$$

$$H_1: \mu_{\text{paper}} \neq \mu_{\text{radio}}$$

Step2-Set Significance level $\alpha=0.05$ **Step 3 -Collect evidence**

$$\text{p-value}=0.3909.$$

Step 4- Decision Rule.

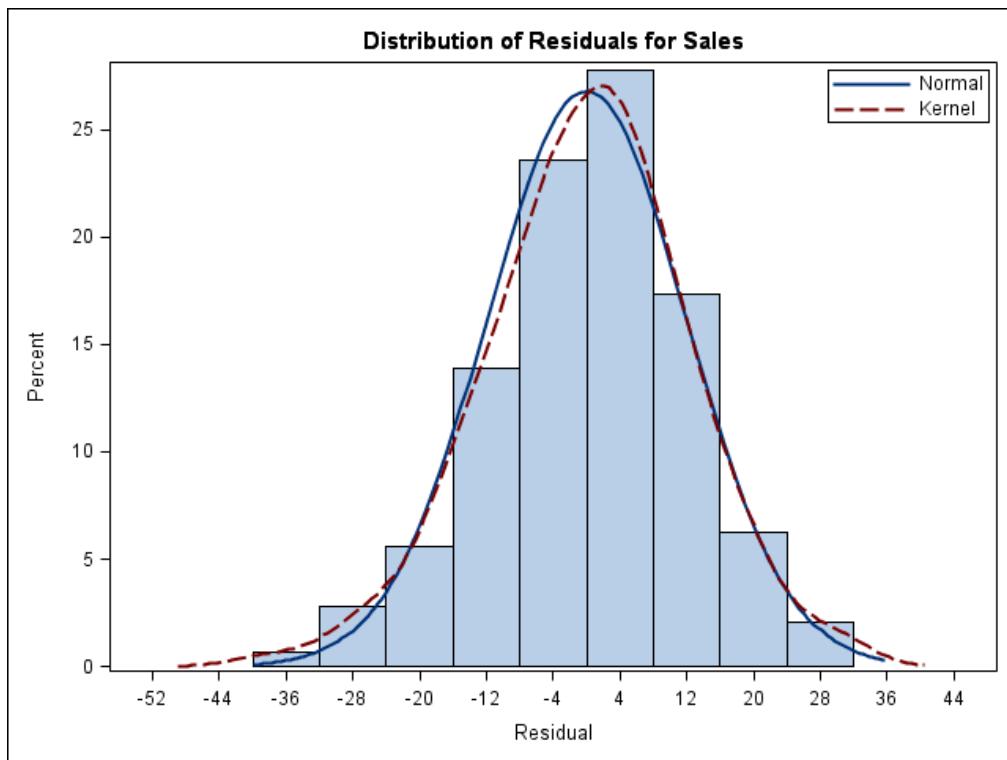
The p-value> α , Fail to reject H_0 , therefore, there is no significant difference between sales when advertising via paper or radio.

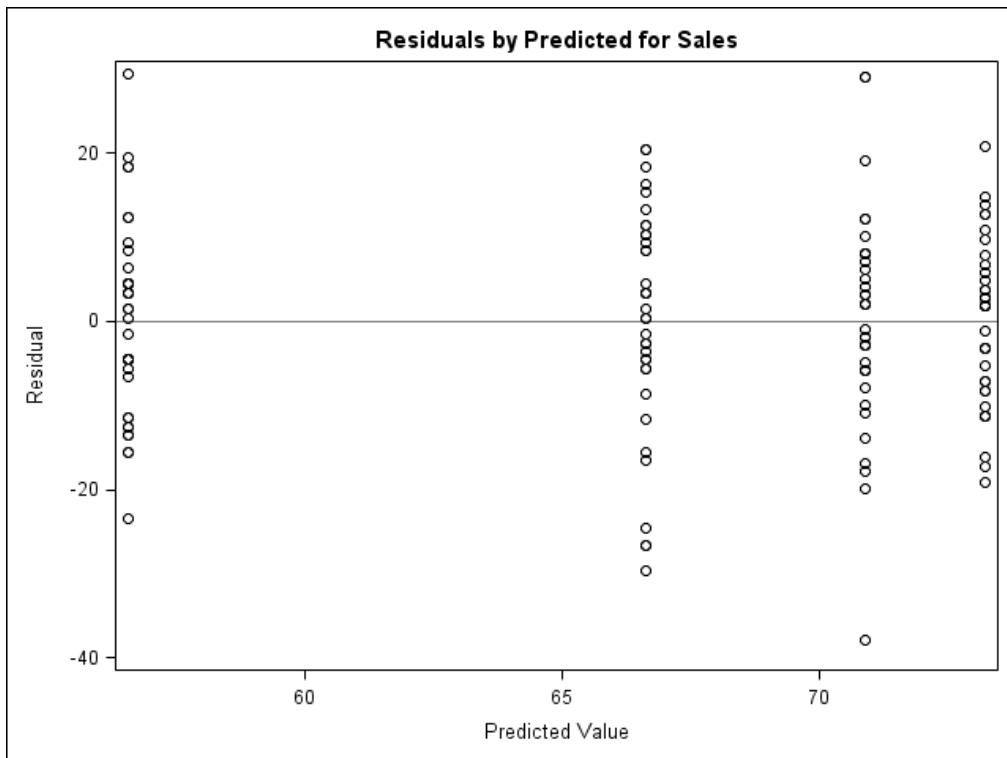
3. One-Way ANOVA – comparing more than two groups.

- a) 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?

```
/*st004s03.sas*/
ods graphics on;
proc glm data=st092.all_ads PLOTS(only)=diagnostics(unpack);
  class ad;
  model sales=ad;
  means ad / hovtest;
  title 'Testing for Equality of Means with PROC GLM';
run;
quit;
ods graphics off;
```

Test the normality assumption.





The graphs do not show strong evidence against normality.

Test the homogeneity of variance assumption using Levene's.

Testing for Equality of Means with PROC GLM					
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		

Step 1- Set Hypothesis

$$H_0: \sigma^2_{\text{paper}} = \sigma^2_{\text{radio}} = \sigma^2_{\text{people}} = \sigma^2_{\text{display}}$$

$H_1:$ At least one variance differs from another\others.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$\text{p-value}=0.3532$$

Step 4- Decision Rule.

The p-value > α , Fail to reject H_0 , therefore there is no evidence to say the variances are not equal.

The assumptions are verified; now test the hypothesis that all the means are equal.

Testing for Equality of Means with PROC					
The GLM Procedure					
Class Level Information					
Class	Levels	Values			
Ad	4	display paper people radio			
		Number of Observations Read	144		
		Number of Observations Used	144		
Testing for Equality of Means with PROC GLM					
The GLM Procedure					
Dependent Variable: Sales					
Sum of					
Source	DF	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		
Testing for Equality of Means with PROC GLM					
The GLM Procedure					
Testing for Equality of Means with PROC GLM					
The GLM Procedure					
Level of					
	-----Sales-----				
Ad	N	Mean	Std Dev		
display	36	56.5555556	11.6188134		
paper	36	73.2222222	9.7339204		
people	36	66.6111111	13.4976776		
radio	36	70.8888889	12.9676031		

Step 1- Set Hypothesis

$$H_0: \mu_{\text{paper}} = \mu_{\text{radio}} = \mu_{\text{people}} = \mu_{\text{display}}$$

$H_1:$ At least one mean is different from an/ other.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

$$\text{p-value}=<.0001$$

Step 4- Decision Rule.

The p-value< α , Reject H_0 , therefore at least one mean is different from one other mean.

4. Post Hoc Pairwise Comparison- Tukey.

- a) Conduct a pairwise comparison with an experimentwise (use the Tukey method) error rate of $\alpha=0.05$. Which types of advertising are significantly different?

```
/*st004s04.sas*/
ods graphics on;
ods select LSMeans Diff MeanPlot DiffPlot;
proc glm data=st092.all_ads;
  class ad;
  model sales=ad;
  lsmeans ad / pdiff=all adjust=tukey;
  title 'Testing for Equality of Means with PROC GLM';
run;
quit;
ods graphics off;
```

Output for PROC GLM

Testing for Equality of Means with PROC GLM
 The GLM Procedure
 Least Squares Means
 Adjustment for Multiple Comparisons: Tukey

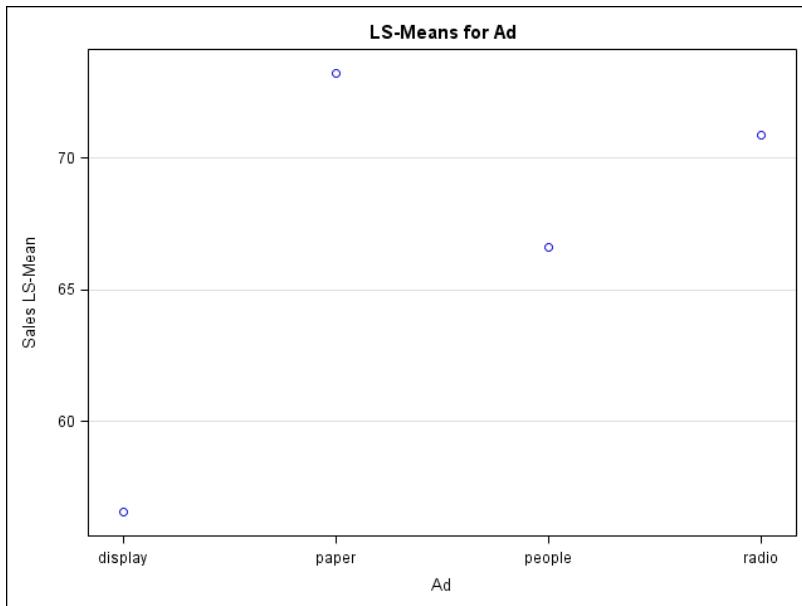
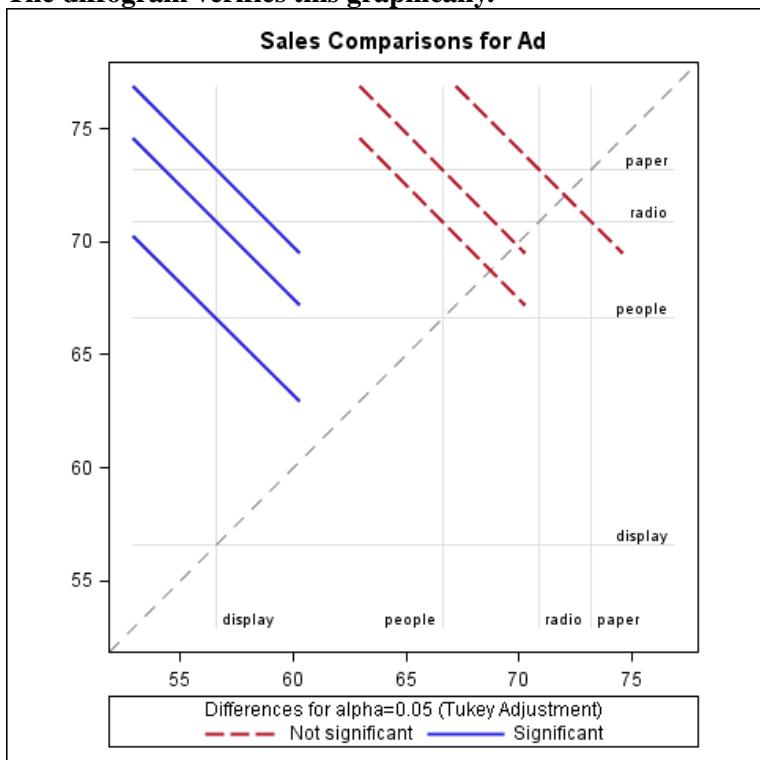
Ad	Sales	LSMEAN	Number
display	56.5555556	1	
paper	73.2222222	2	
people	66.6111111	3	
radio	70.8888889	4	

Least Squares Means for effect Ad
 Pr > |t| for H0: LSMean(i)=LSMean(j)

i/j	Dependent Variable: Sales			
	1	2	3	4
1		<.0001	0.0030	<.0001
2	<.0001		0.0964	0.8440
3	0.0030	0.0964		0.4360
4	<.0001	0.8440	0.4360	

The Tukey comparisons show significant differences between display and all other types of advertising.

The diffogram verifies this graphically.



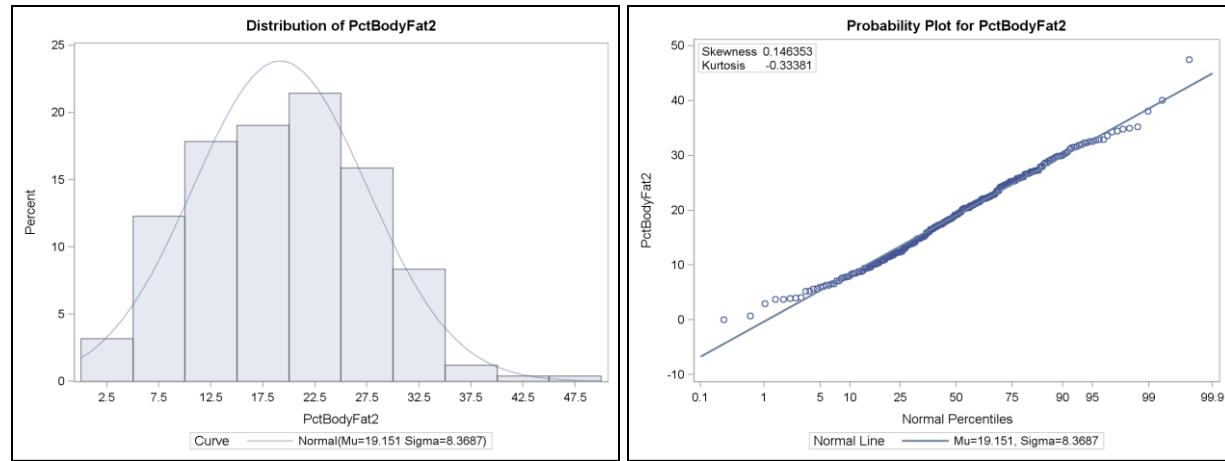
Chapter 5

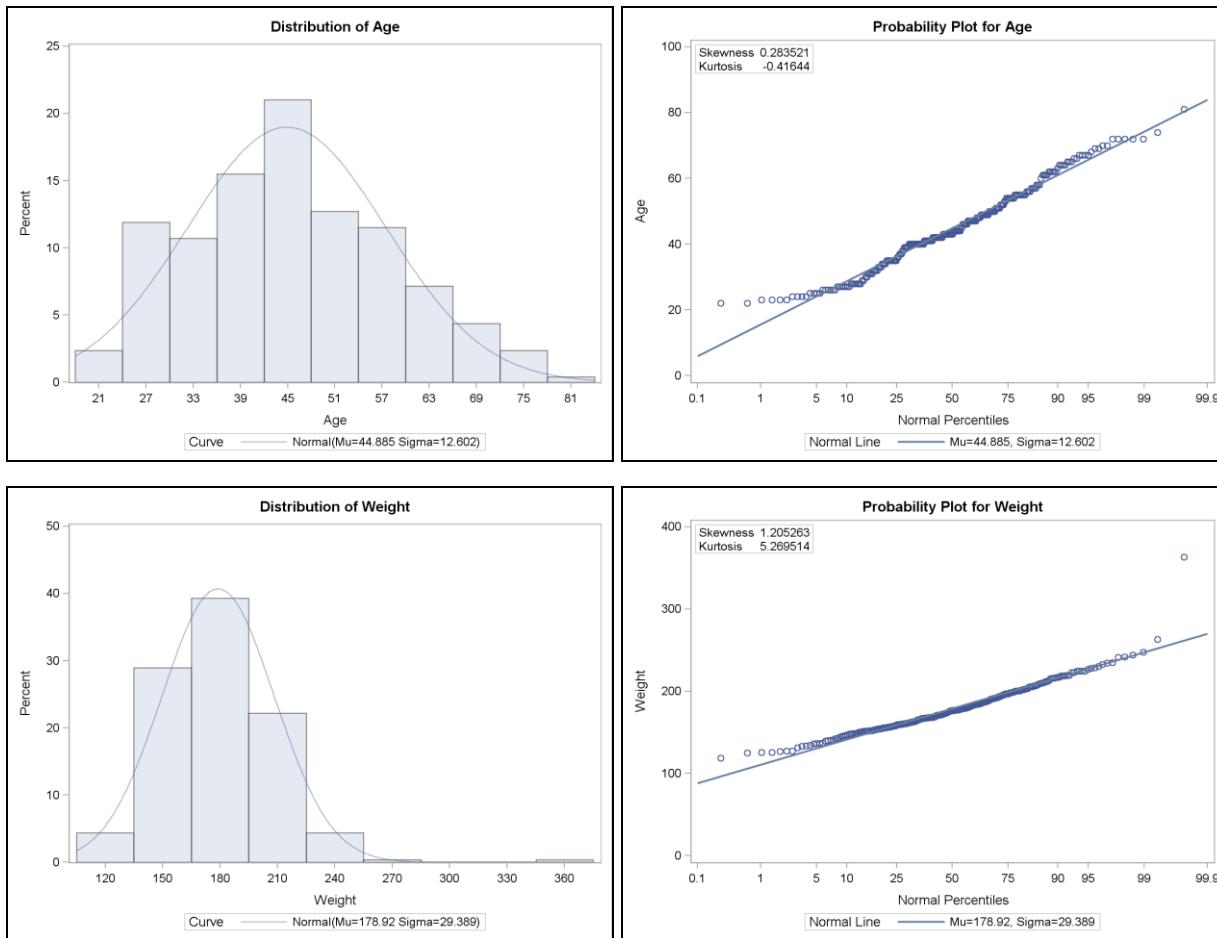
1. Describing the Relationship between Continuous Variables

Use the UNIVARIATE procedure to examine the distribution of the variables **PCTBODYFAT2**, **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH**.

```
/*st005s01.sas*/
ods graphics on;
ods listing close;
ods rtf file='bodyfat.rtf' style=statistical;
ods select histogram probplot;
proc univariate data=st092.BodyFat;
var PctBodyFat2 Age Weight Height
    Neck Chest Abdomen Hip Thigh;
histogram / normal (mu=est sigma=est);
probplot / normal (mu=est sigma=est);
inset skewness kurtosis;
title "Predictors of % Body Fat";
run;
ods rtf close;
ods listing;
ods graphics off;
```

Partial output shown below.





- a) What conclusions can you draw about the distribution of these variables?

WEIGHT, NECK, ABDOMEN, HIP, and THIGH and other measures seem to show high skewness and kurtosis. This might be due to a large outlier.

- b) Do there appear to be any unusual observations?

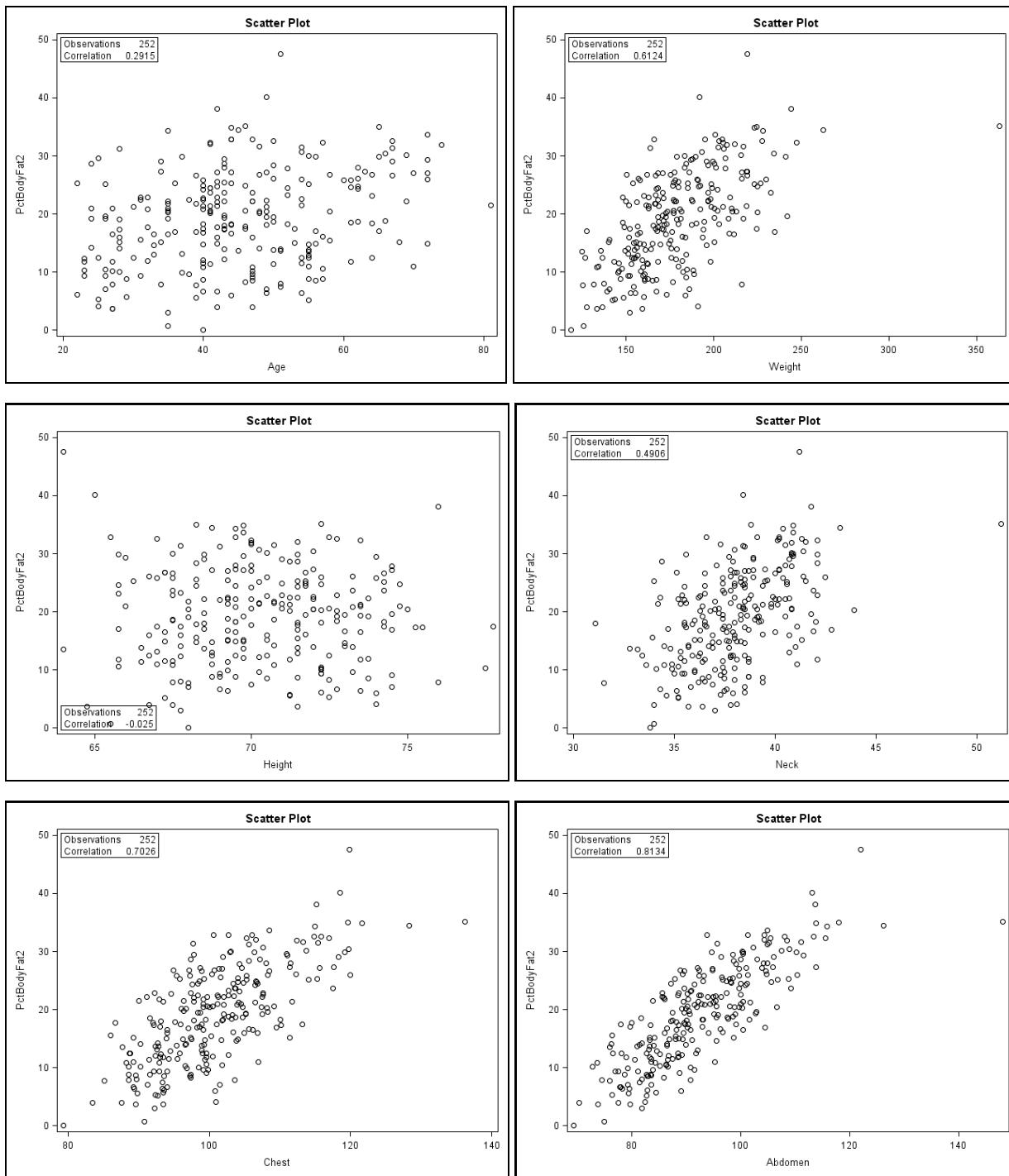
One participant seems to be different than the rest, CASE 39.

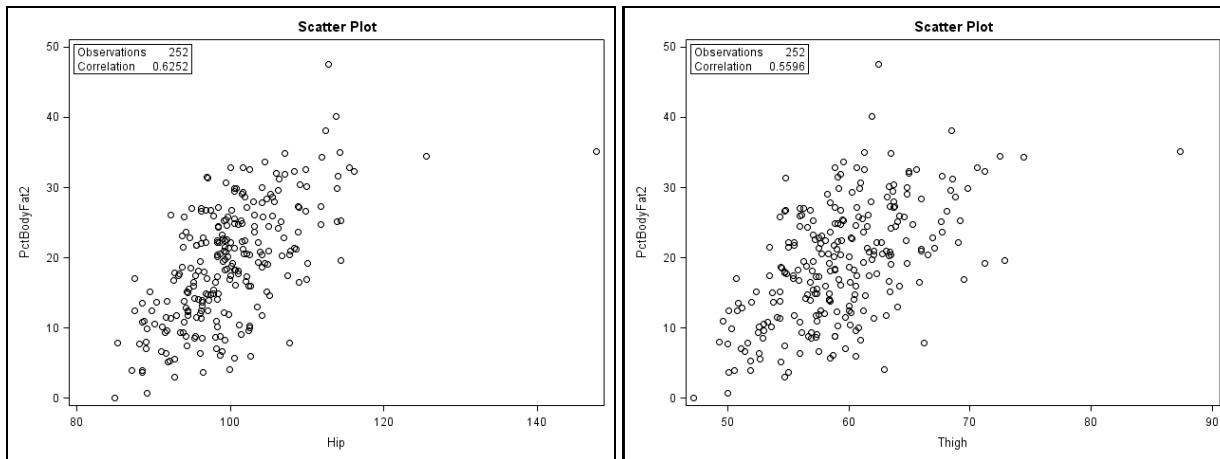
Variable_	Value
Weight	363.15
Neck	51.20
Chest	136.20
Hip	147.70
Thigh	87.30

2. Scatter Plots and Correlation Statistics

Generate scatter plots and correlations for the VAR variables, **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH** versus the WITH variable, **PCTBODYFAT2**.

```
/*st005s02.sas*/
ods graphics on;
proc corr data=st092.BodyFat rank
            plots(only)=scatter(nvar=all ellipse=none);
  var Age Weight Height Neck Chest Abdomen Hip Thigh;
  with PctBodyFat2;
  title "Correlations and Scatter Plots with Body Fat %";
run;
ods graphics off;
```





- a) Can straight lines adequately describe the relationships?

HEIGHT seems to be the only variable that shows no real linear relationship.

- b) Are there any outliers you should investigate?

The WEIGHT outlier is present again, as well as NECK, ABDOMEN and Hip .

- c) What variable has the highest correlation with PCTBODYFAT2?

ABDOMEN, with 0.81343 is the variable with the highest correlation with PCTBODYFAT2.

- d) What is the p-value for the coefficient of the variable that has the highest correlation with PCTBODYFAT2?

p-value=<.0001

- e) Is it statistically significant at the 0.05 level?

Step 1- Set Hypothesis

$H_0: \text{Rho}=0.$

$H_1: : \text{Rho} \neq 0.$

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

p-value=<.0001

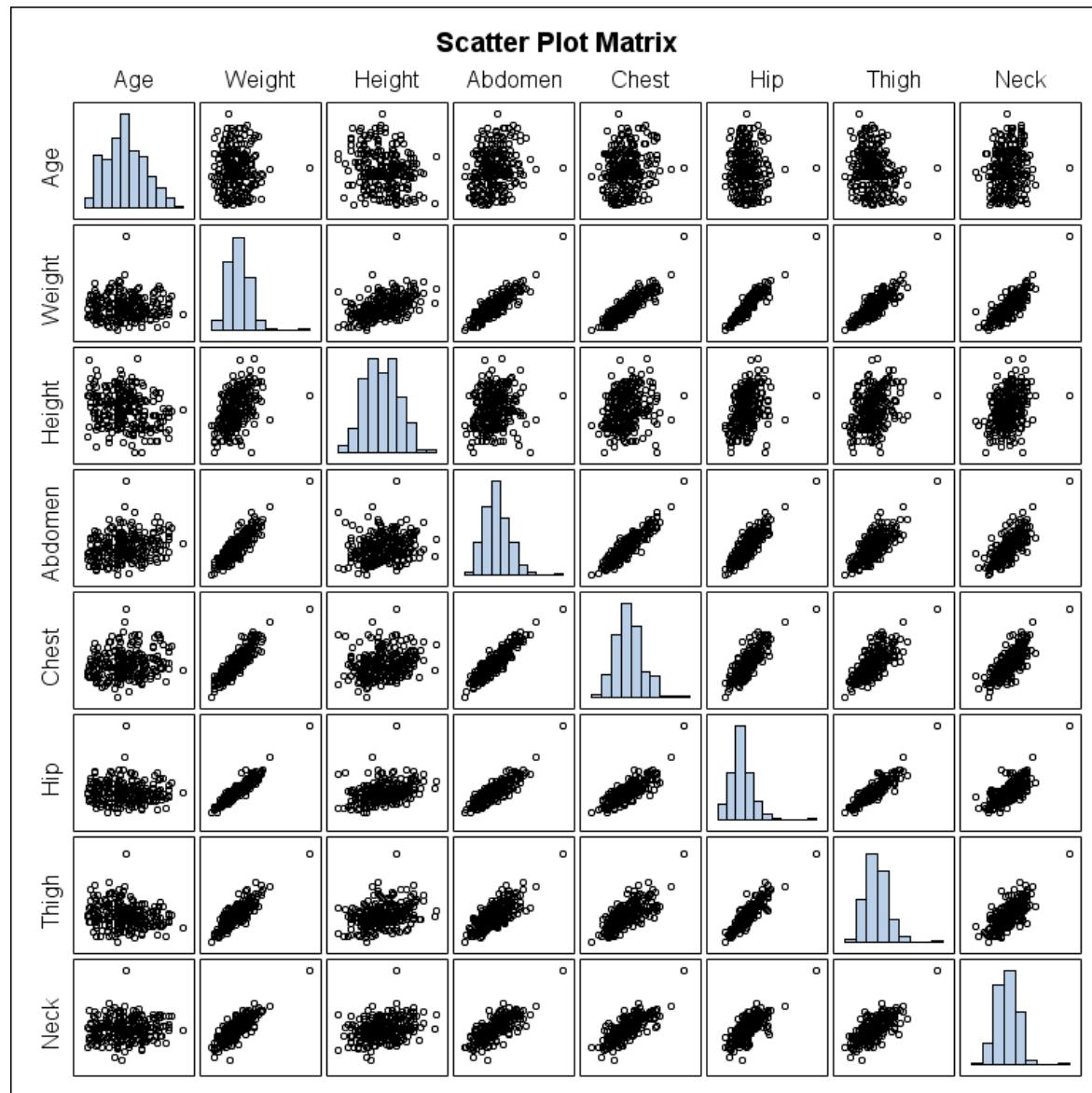
The p-value $<\alpha$, reject H_0 , therefore the coefficient of ABDOMEN is statistically significantly different to 0 at the 0.05 significance level.

3. Scatter Plots and Correlation Statistics

Generate correlations among all of the VAR variables **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH**.

```
/*st005s03.sas*/
ods graphics on;
proc corr data=st092.BodyFat nosimple
    plots=matrix(nvar=all histogram);
    var Age Weight Height Abdomen Chest Hip Thigh Neck;
    title "Correlations and Scatter Plot Matrix of Basic "
        "Measures";
run;
ods graphics off;
```

Partial Output



Correlations and Scatter Plot Matrix of Basic Measures								
The CORR Procedure								
8 Variables:	Age	Weight	Height	Abdomen	Chest	Hip	Thigh	Neck
Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0								
	Age	Weight	Height	Abdomen	Chest	Hip	Thigh	Neck
Age	1.00000	-0.01275 0.8404	-0.24521 <.0001	0.23041 0.0002	0.17645 0.0050	-0.05033 0.4263	-0.20010 0.0014	0.11351 0.0721
Weight	-0.01275 0.8404	1.00000	0.48689 <.0001	0.88799 <.0001	0.89419 <.0001	0.94088 <.0001	0.86869 <.0001	0.83072 <.0001
Height	-0.24521 <.0001	0.48689 <.0001	1.00000	0.18977 0.0025	0.22683 0.0003	0.37211 <.0001	0.33856 <.0001	0.32114 <.0001
Abdomen	0.23041 0.0002	0.88799 <.0001	0.18977 0.0025	1.00000	0.91583 <.0001	0.87407 <.0001	0.76662 <.0001	0.75408 <.0001
Chest	0.17645 0.0050	0.89419 <.0001	0.22683 0.0003	0.91583 <.0001	1.00000	0.82942 <.0001	0.72986 <.0001	0.78484 <.0001
Hip	-0.05033 0.4263	0.94088 <.0001	0.37211 <.0001	0.87407 <.0001	0.82942 <.0001	1.00000	0.89641 <.0001	0.73496 <.0001
Thigh	-0.20010 0.0014	0.86869 <.0001	0.33856 <.0001	0.76662 <.0001	0.72986 <.0001	0.89641 <.0001	1.00000	0.69570 <.0001
Neck	0.11351 0.0721	0.83072 <.0001	0.32114 <.0001	0.75408 <.0001	0.78484 <.0001	0.73496 <.0001	0.69570 <.0001	1.00000

a) Are there any notable relationships?

WEIGHT seems to correlate highly with ABDOMEN, CHEST, HIP, THIGH and NECK.

Thigh seems to correlate highly with all other predictor variables as does ABDOMEN and CHEST

Chapter 6

1. Fitting a Simple Linear Regression Model

Perform a simple linear regression model with **PCTBODYFAT2** as the response variable and **ABDOMEN** as the predictor.

```
/*st006s01.sas*/
ods graphics off;
proc reg data=st092.BodyFat;
  model PctBodyFat2=Abdomen;
  title "Regression of % Body Fat on Abdomen";
run;
quit;
```

Regression of % Body Fat on Abdomen

The REG Procedure

Model: MODEL1

Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of		Mean	
		Squares	Square	F Value	Pr > F
Model	1	11632	11632	488.93	<.0001
Error	250	5947.46303	23.78985		
Corrected Total	251	17579			

Root MSE	4.87748	R-Square	0.6617
Dependent Mean	19.15079	Adj R-Sq	0.6603
Coeff Var	25.46884		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-39.28018	2.66034	-14.77	<.0001
Abdomen	1	0.63130	0.02855	22.11	<.0001

- a) What is the value of the F statistic and the associated p-value? How would you interpret this with regards to the null hypothesis?

The F value is 488.93 and the p-value is <.0001. You would reject the null hypothesis of no relationship.

- b) Write out the predicted regression equation.

**From the parameter estimates table, the predicted value of
 $PctBodyFat2 = -39.28018 + 0.63130 * Abdomen$.**

- c) What is the value of the R² statistic value? How would you interpret this?

The R² value of 0.6617 can be interpreted to mean that 66.17% of the variability in PctBodyFat2 can be explained by ABDOMEN.

2. Confidence and Prediction intervals.

Use the REG procedure to produce Confidence and Prediction Intervals around ABDOMEN .

```
/*st006s02.sas*/
ods graphics on;
proc reg data=st092.BodyFat;
model PctBodyFat2=abdomen / clm cli;
title "Regression of % Body Fat on Abdomen";
run;
quit;
ods graphics off;
```

Partial OUTPUT

Regression of % Body Fat on Abdomen

The REG Procedure
Model: MODEL1
Dependent Variable: PctBodyFat2

Output Statistics

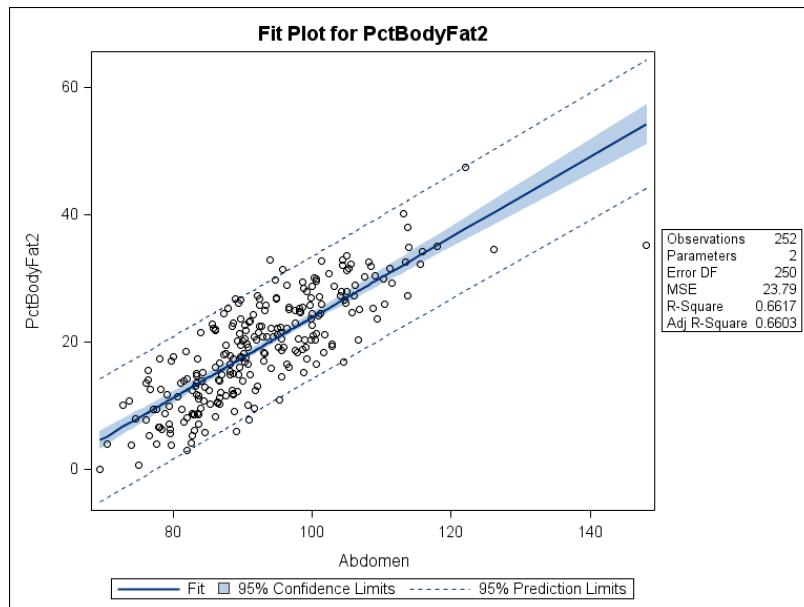
Obs	Abdomen	Dependent Variable	Predicted Value	Std Error Mean	95% CL Mean	95% CL Predict
1	85.2	12.3000	14.5069	0.3722	13.7740	15.2399
2	83.0	6.1000	13.1181	0.4109	12.3088	13.9273
3	87.9	25.3000	16.2115	0.3348	15.5521	16.8708

- a) What is the Confidence Interval when ABDOMEN is 83 (see observation number 2) and how would you interpret this?

The 95% Confidence Interval when ABDOMEN is 83 (12.3, 13.9). We are 95% confident that the population mean of PCTBODYFAT2, when ABDOMEN is 83, would lie between 12.3 and 13.9.

- b) What is the Prediction Interval when **ABDOMEN** is 83 (see observation number 2) and how would you interpret this?

The 95% Prediction Interval when ABDOMEN is 83 (3.5, 22.8). We are 95% confident that the actual value of PCTBODYFAT2, when ABDOMEN is 83, would lie between 3.5 and 22.8.



3. Predicted values.

Produce predicted values for **PCTBODYFAT2** when **ABDOMEN** is 80, 100 and 120

```
/*st006s03*/
ods graphics off;
data st092.BodyFatPRED;
  set st092.ABDOMENPRED
      st092.BodyFat;
run;

proc reg data=st092.BodyFatPRED ;
  model PctBodyFat2=Abdomen /p;
  id abdomen;
  title "Regression of % Body Fat on Abdomen";
run;
quit;
```

Partial OUTPUT.

Regression of % Body Fat on Abdomen

The REG Procedure
Model: MODEL1
Dependent Variable: PctBodyFat2

Output Statistics				
Obs	Abdomen	Dependent Variable	Predicted Value	Residual
1	80.0	.	11.2242	.
2	100.0	.	23.8503	.
3	120.0	.	36.4763	.
4	85.2	12.3000	14.5069	-2.2069
5	83.0	6.1000	13.1181	-7.0181
.....				

- a) What are the predicted values when **ABDOMEN** is 80, 100 and 120?

11.22, 23.85, 36.48 respectively

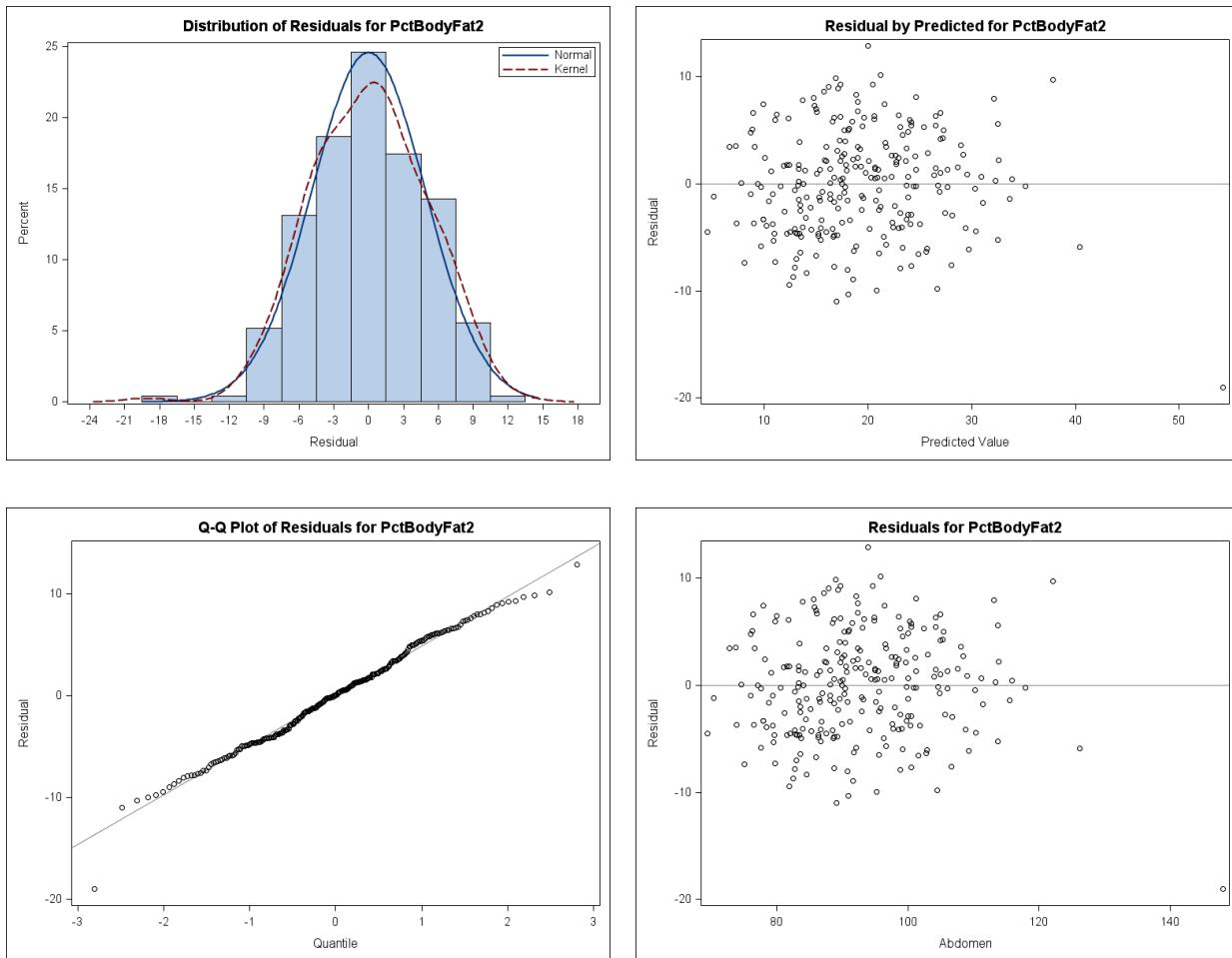
- b) Is it appropriate to predict **PCTBODYFAT2** when **ABDOMEN** is 200?

No, because there is no data in the model data set with **Abdomen** greater than 148.1. You should not predict beyond the range of your data.

4. Examining Residuals

Assess the model obtained from using **ABDOMEN** as a predictor variable for **PCTBODYFAT2**. Create plots of the residuals by **ABDOMEN**, and by the predicted values, and a normal Quantile-Quantile plot.

```
/*st006s04*/
ods graphics on;
proc reg data=st092.bodyfat
plots(only)=(QQ
             RESIDUALBYPREDICTED
             RESIDUALHISTOGRAM
             RESIDUALPLOT);
model PctBodyFat2
      = Abdomen;
title 'Plots of Diagnostic Statistics for Abdomen';
run;
quit;
ods graphics off;
```



- a) Do the residual plots indicate any problems with the constant variance assumption?

It does not appear that the data violates the assumption of constant variance.

- b) Are there any outliers indicated in the residual plots?

There is a clear outlier in **ABDOMEN** that might be worth investigating.

- c) Does the quantile-quantile plot indicate any problems with the normality assumption?

The normality assumption seems to be met.

Chapter 7

1. Performing a Regression Using the REG Procedure

Using the **st092.BODYFAT** data set, run a regression of **PCTBODYFAT2** on the variables **AGE**, **WEIGHT**, **HEIGHT**, **NECK**, **CHEST**, **ABDOMEN**, **HIP**, and **THIGH**.

```
/*st007s01 a)*/  
ods graphics off;  
proc reg data=st092.BodyFat;  
    model PctBodyFat2 = Age Weight Height  
        Neck Chest Abdomen Hip Thigh;  
    title 'Regression of PctBodyFat2 on All Predictors';  
run;  
quit;
```

Regression of PctBodyFat2 on All Predictors

The REG Procedure

Model: MODEL1

Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance

Source	DF	Sum of		Mean	
		Squares	Square	F Value	Pr > F
Model	8	12877	1609.58696	83.18	<.0001
Error	243	4702.29417	19.35100		
Corrected Total	251	17579			
Root MSE		4.39898	R-Square	0.7325	
Dependent Mean		19.15079	Adj R-Sq	0.7237	
Coeff Var		22.97021			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-23.59253	21.14877	-1.12	0.2657
Age	1	0.02789	0.02945	0.95	0.3445
Weight	1	-0.09819	0.05783	-1.70	0.0908
Height	1	-0.09358	0.17539	-0.53	0.5941
Neck	1	-0.56942	0.22293	-2.55	0.0113
Chest	1	0.02599	0.10400	0.25	0.8028
Abdomen	1	0.95753	0.08809	10.87	<.0001
Hip	1	-0.24167	0.14551	-1.66	0.0980
Thigh	1	0.33955	0.13675	2.48	0.0137

a) Compare the output with the output from the model with only **ABDOMEN** -in the previous exercise.

- i. What is different in the ANOVA tables?

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, 8 versus 1.

The Mean Square Model and the F ratio are much smaller.

- ii. How do the R^2 and the adjusted R^2 compare with these statistics for the **ABDOMEN** regression demonstration?

Both the R^2 and adjusted R^2 for the full models are larger than the simple linear regression. The multiple regression model explains almost 72 percent of the variation in the **PctBodyFat2 variable versus only about 66 percent explained by the simple linear regression model.**

- iii. Did the estimate for the intercept change? Did the estimate for the coefficient of **ABDOMEN** change?

Yes, including the other variables in the model changed both the estimate of the intercept and the slope for **ABDOMEN.**

b) Simplifying the Model

- i. Rerun the model in a), but eliminate the variable with the highest p -value. Compare the output with the Exercise a) model.

This program reruns the regression with **CHEST** removed because it has the largest p -value (0.8028).

```
/*st007s01 b*/
proc reg data=st092.BodyFat;
  model PctBodyFat2 = Age Weight Height
    Neck Abdomen Hip Thigh;
  title 'Remove Chest';
run;
quit;
```

Remove Chest

The REG Procedure

Model: MODEL1

Dependent Variable: PctBodyFat2

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	12875	1839.35525	95.42	<.0001
Error	244	4703.50311	19.27665		
Corrected Total	251	17579			
Root MSE		4.39052	R-Square	0.7324	
Dependent Mean		19.15079	Adj R-Sq	0.7248	
Coeff Var		22.92604			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-20.74865	17.79306	-1.17	0.2447
Age	1	0.02799	0.02939	0.95	0.3418
Weight	1	-0.09055	0.04900	-1.85	0.0658
Height	1	-0.10987	0.16251	-0.68	0.4996
Neck	1	-0.56879	0.22249	-2.56	0.0112
Abdomen	1	0.96548	0.08199	11.78	<.0001
Hip	1	-0.24981	0.14155	-1.76	0.0788
Thigh	1	0.33289	0.13388	2.49	0.0136

- ii. Did the p-value for the model change?

No, the *p*-value for the model did not change, up to four decimal places.

- iv. Did the R^2 and adjusted R^2 change?

The R^2 showed essentially no change. The adjusted R^2 increased from 0.7237 to 0.7248
When an adjusted R^2 increases by removing a variable from the models, it strongly implies that the removed variable was not necessary.

- v. Did the parameter estimates and their *p*-values change?

Some of the parameter estimates and their *p*-values changed slightly, none to any large degree.

c) More Simplifying of the Model

- i. Rerun the model in Exercise b), but drop the variable with the highest *p*-value.

This program reruns the regression with **HEIGHT** removed because it has the largest *p*-value (0.4996).

```
/*st007s01 c*/
proc reg data=st092.BodyFat;
model PctBodyFat2 = Age Weight
    Neck Abdomen Hip Thigh;
title 'Remove Chest and Height';
run;
quit;
```

Remove Chest and Height					
The REG Procedure					
Model: MODEL1					
Dependent Variable: PctBodyFat2					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	12867	2144.44589	111.49	<.0001
Error	245	4712.31447	19.23394		
Corrected Total	251	17579			
Root MSE		4.38565	R-Square	0.7319	
Dependent Mean		19.15079	Adj R-Sq	0.7254	
Coeff Var		22.90062			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-30.68564	10.01738	-3.06	0.0024
Age	1	0.02845	0.02935	0.97	0.3333
Weight	1	-0.11207	0.03722	-3.01	0.0029
Neck	1	-0.53624	0.21698	-2.47	0.0141
Abdomen	1	0.99354	0.07063	14.07	<.0001
Hip	1	-0.23789	0.14029	-1.70	0.0912
Thigh	1	0.35000	0.13132	2.67	0.0082

- ii. How did the output change from the previous model?

The ANOVA table did not change significantly. The R^2 remained essentially unchanged. The adjusted R^2 increased again, confirming that the variable HEIGHT did not contribute to explaining the variation in PctBodyFat2 when the other variables are in the model.

- iii. Did the number of parameters with p-values less than 0.05 change?

The p-values and parameter estimates for other variables didn't change much.

2. Using Stepwise Selection

- a) Use a stepwise regression method to select a candidate model; try FORWARD, STEPWISE and BACKWARD.

```
/*st007s02 a)*/
ods graphics on;
proc reg data=st092.BodyFat plots(only)=adjrsq;
  FORWARD: model PctBodyFat2 = Age Weight Height
             Neck Chest Abdomen Hip Thigh
             / selection=forward;
  BACKWARD: model PctBodyFat2 = Age Weight Height
             Neck Chest Abdomen Hip Thigh
             / selection=backward;
  STEPWISE: model PctBodyFat2 = Age Weight Height
             Neck Chest Abdomen Hip Thigh
             / selection=stepwise;
  title "Using Stepwise Methods for Model Selection";
run;
quit;
```

Partial Output of the final suggested model and the summary of the steps taken for the FORWARD selection.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	12875	1839.35525	95.42	<.0001
Error	244	4703.50311	19.27665		
Corrected Total	251	17579			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-20.74865	17.79306	26.21258	1.36	0.2447
Age	0.02799	0.02939	17.48463	0.91	0.3418
Weight	-0.09055	0.04900	65.82475	3.41	0.0658
Height	-0.10987	0.16251	8.81136	0.46	0.4996
Neck	-0.56879	0.22249	125.98133	6.54	0.0112
Abdomen	0.96548	0.08199	2673.17514	138.67	<.0001
Hip	-0.24981	0.14155	60.03887	3.11	0.0788
Thigh	0.33289	0.13388	119.18516	6.18	0.0136

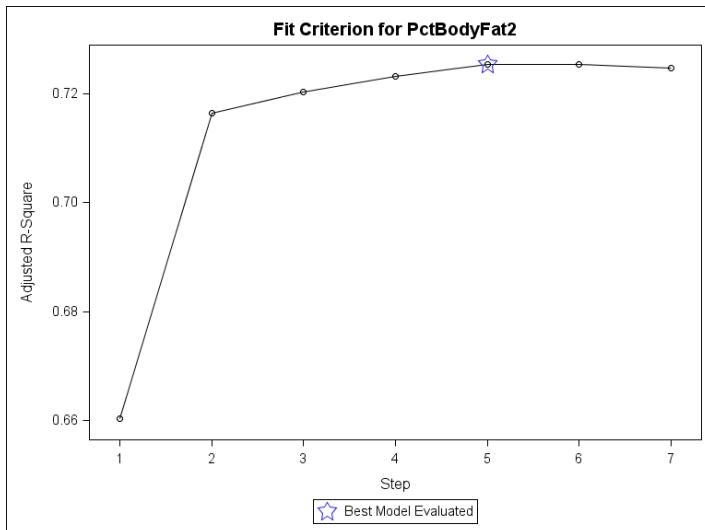
Bounds on condition number: 27.007, 454.6

No other variable met the 0.5000 significance level for entry into the model.

Using Stepwise Methods for Model Selection

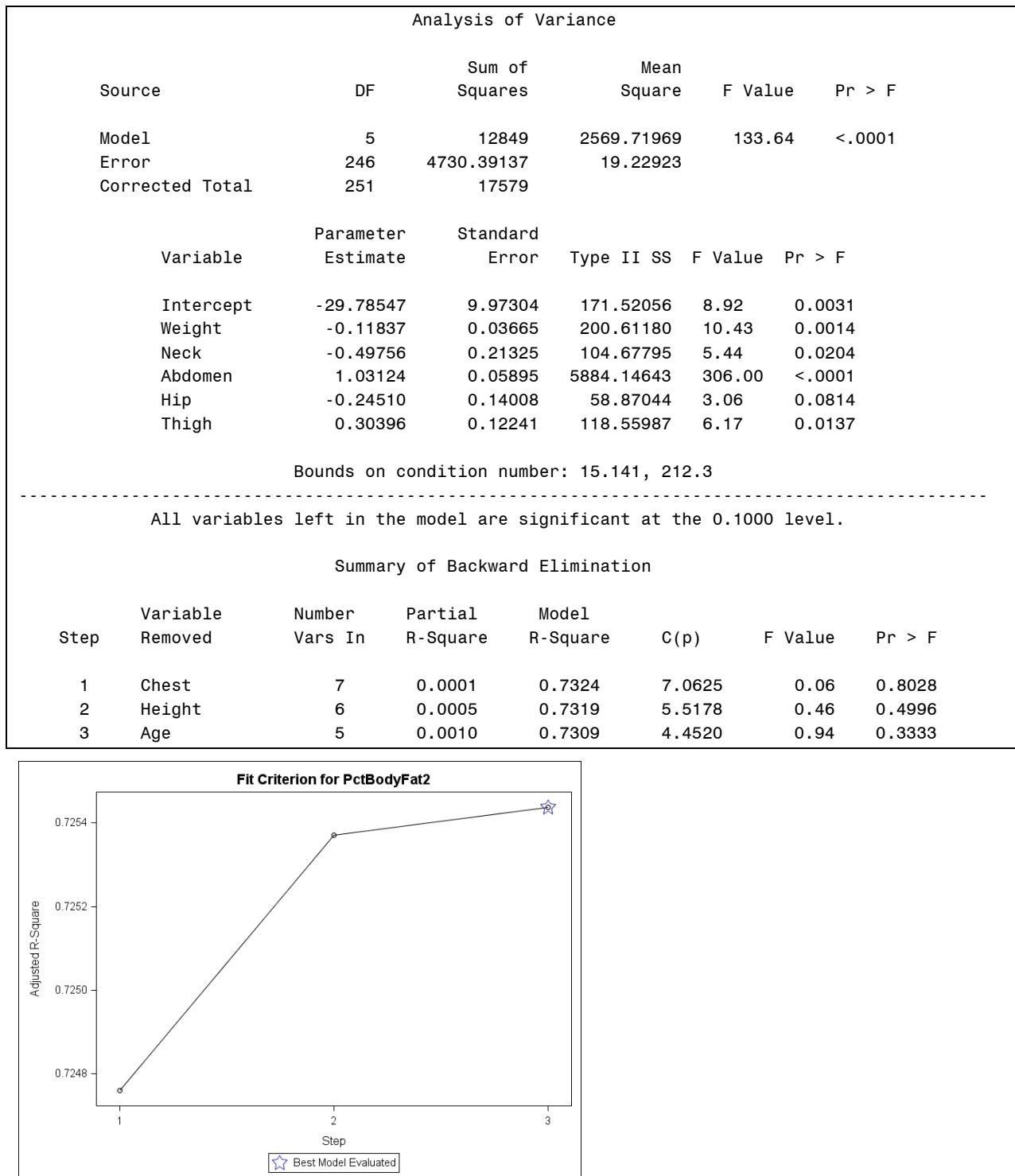
The REG Procedure

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Abdomen	1	0.6617	0.6617	59.3465	488.93	<.0001
2	Weight	2	0.0571	0.7188	9.4516	50.58	<.0001
3	Neck	3	0.0049	0.7237	6.9594	4.44	0.0361
4	Thigh	4	0.0038	0.7276	5.4942	3.46	0.0641
5	Hip	5	0.0033	0.7309	4.4520	3.06	0.0814
6	Age	6	0.0010	0.7319	5.5178	0.94	0.3333
7	Height	7	0.0005	0.7324	7.0625	0.46	0.4996



The Criterion Plot, for the FORWARD final model, shows that the “best” model according to the Adjusted R-Square is the model in Step 5. The largest increase is from step1 to step 2 and the increase is rather modest after that.

Partial Output of the final suggested model and the summary of the steps taken for the BACKWARD selection.



The Criterion Plot, for the BACKWARD final model, shows that the “best” model according to the Adjusted R-Square is the final model in Step 3. Although the increase from step 1 to step 2 looks large, it’s actually quite small.

Partial Output of the final suggested model and the summary of the steps taken for the BACKWARD selection.

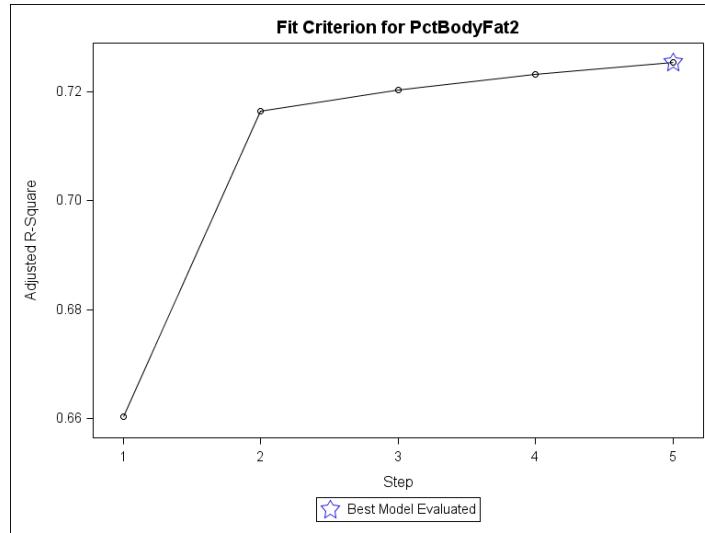
Stepwise Selection: Step 5						
Variable Hip Entered: R-Square = 0.7309 and C(p) = 4.4520						
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	12849	2569.71969	133.64	<.0001	
Error	246	4730.39137	19.22923			
Corrected Total	251	17579				

Using Stepwise Methods for Model Selection						
The REG Procedure						
Model: STEPWISE						
Dependent Variable: PctBodyFat2						
Stepwise Selection: Step 5						
Parameter	Standard					
Variable	Estimate	Error	Type II SS	F Value	Pr > F	
Intercept	-29.78547	9.97304	171.52056	8.92	0.0031	
Weight	-0.11837	0.03665	200.61180	10.43	0.0014	
Neck	-0.49756	0.21325	104.67795	5.44	0.0204	
Abdomen	1.03124	0.05895	5884.14643	306.00	<.0001	
Hip	-0.24510	0.14008	58.87044	3.06	0.0814	
Thigh	0.30396	0.12241	118.55987	6.17	0.0137	

Bounds on condition number: 15.141, 212.3

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	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Abdomen		1	0.6617	0.6617	59.3465	488.93	<.0001
2	Weight		2	0.0571	0.7188	9.4516	50.58	<.0001
3	Neck		3	0.0049	0.7237	6.9594	4.44	0.0361
4	Thigh		4	0.0038	0.7276	5.4942	3.46	0.0641
5	Hip		5	0.0033	0.7309	4.4520	3.06	0.0814



The Criterion Plot, for the STEPWISE final model, shows that the “best” model according to the Adjusted R-Square is the final model in Step 5.

- b) How many variables would have resulted from a model using FORWARD selection and a significance level for entry criterion of 0.05, instead of the default SLENTRY of 0.50?

```
/*st007s02 b) */
ods graphics on;
proc reg data=st092.BodyFat plots(only)=adjrsq;
  FORWARD05: model PctBodyFat2 = Age Weight Height
    Neck Chest Abdomen Hip Thigh
    / selection=forward slentry=0.05;
  title "Using Forward Stepwise with SLENTRY=0.05";
run;
quit;
```

Partial Output

Forward Selection: Step 3
Variable Neck Entered: R-Square = 0.7237 and C(p) = 6.9594

Analysis of Variance

Source	DF	Sum of Squares		Mean Square	F Value	Pr > F
		Model	Error			
Model	3	12723	4240.89123	216.57	<.0001	
Error	248	4856.31615	19.58192			
Corrected Total	251	17579				

Variable	Parameter Estimate	Standard Error	Type II SS			F Value	Pr > F
			F	Value	Pr > F		
Intercept	-35.01532	5.79995	713.71246	36.45	<.0001		
Weight	-0.12054	0.02443	476.62624	24.34	<.0001		
Neck	-0.43576	0.20682	86.92911	4.44	0.0361		
Abdomen	0.99713	0.05644	6110.97099	312.07	<.0001		

```

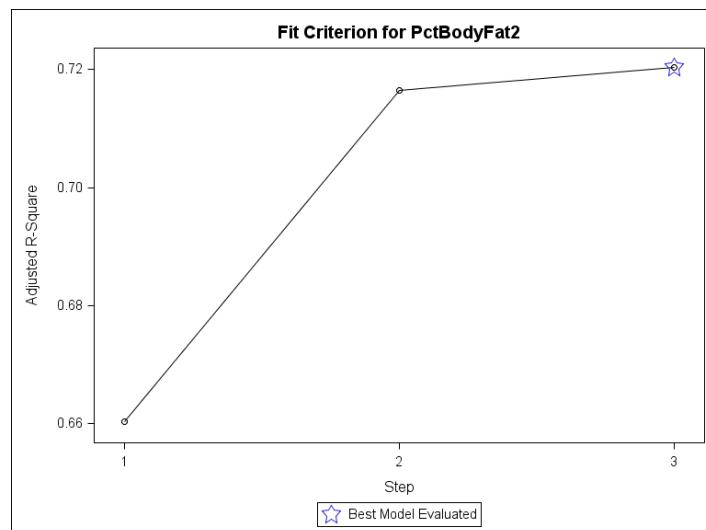
Bounds on condition number: 6.6094, 43.793
-----
No other variable met the 0.0500 significance level for entry into the model.

Using Forward Stepwise with SLENTRY=0.05
The REG Procedure
Model: FORWARD05
Dependent Variable: PctBodyFat2

Summary of Forward Selection
Step   Variable    Number Vars In Partial R-Square Model
      Entered          1       0.6617  0.6617  59.3465
      1     Abdomen      2       0.0571  0.7188  9.4516
      2     Weight       3       0.0049  0.7237  6.9594
      3     Neck

```

The model using SLENTRY=0.05 has substantially fewer (3) variables than the FORWARD default SLENTRY=0.50 final model.



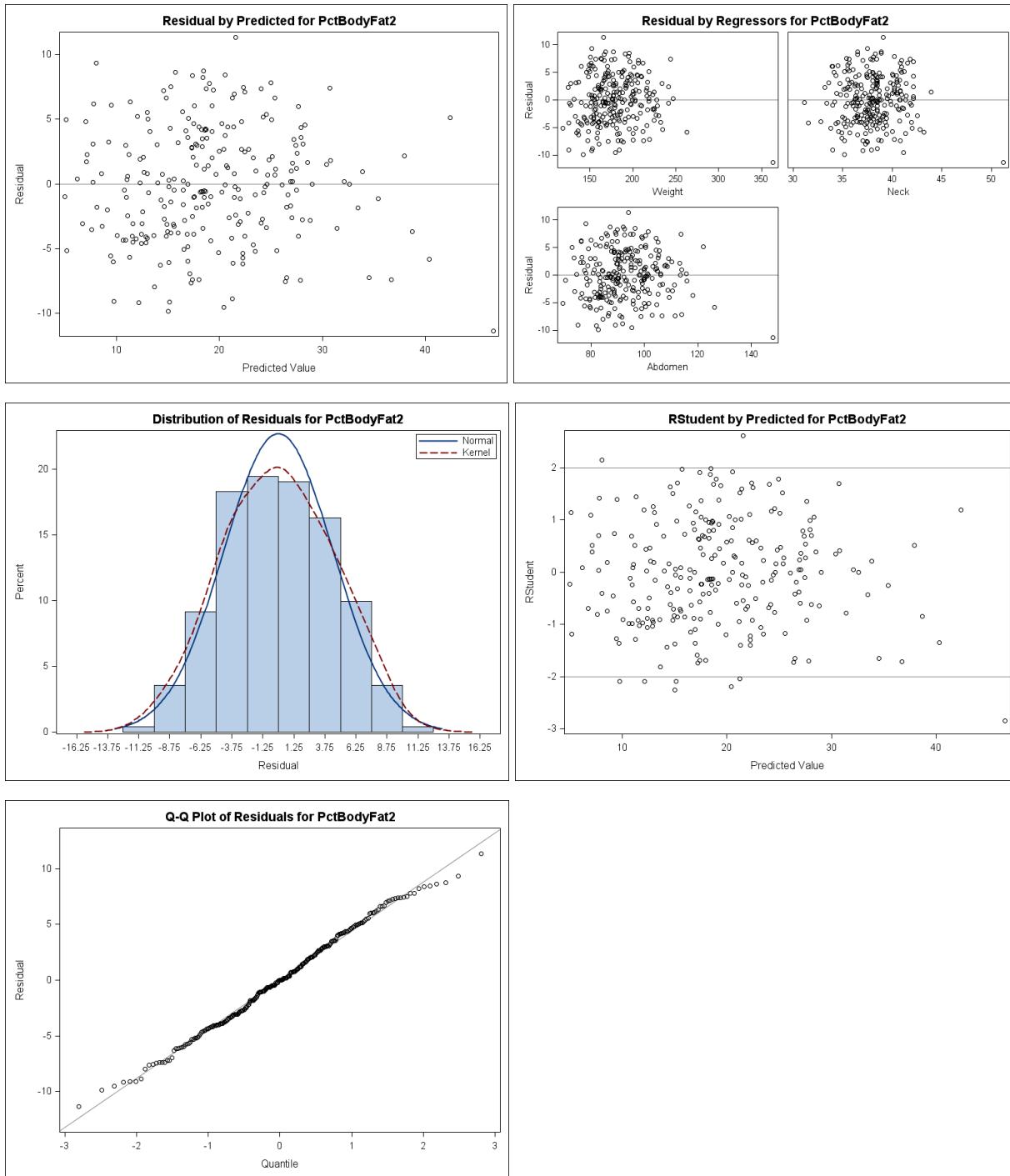
The Criterion Plot, showing adjusted R-square at each step is also produced.

c) Verify the assumptions of the model found in b).

```

/*st007s02 c)*/
ods graphics on;
proc reg data=st092.BodyFat plots(only)=(QQ
                                         RESIDUALBYPREDICTED
                                         RESIDUALHISTOGRAM
                                         RESIDUALPLOT
                                         RSTUDENTBYPREDICTED);
FORWARD05:model PctBodyFat2 = Weight
              Neck Abdomen
              / selection=forward slentry=0.05;
title "Using Forward Stepwise with SLENTRY=0.05";
run;
quit;
ods graphics off;

```



The plots do not show any violation of the assumptions but there is an outlier that may need investigating.

Chapter 8

1. Describing Categorical Data

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 **st092.SAFETY** data set.

- a) What is the measurement scale of each variable?

<u>Variable</u>	<u>Measurement Scale</u>
UNSAFE	Nominal, Binary
TYPE	Nominal
REGION	Nominal, Binary
WEIGHT	Continuous
SIZE	Ordinal

- b) Create one-way frequency tables for the variables **UNSAFE**, **TYPE**, and **REGION**.

```
/*st008s01*/
ods graphics off;
proc freq data=st092.safety;
  tables Unsafe Type Region;
  title "Safety Data Frequencies";
run;
```

Safety Data Frequencies				
The FREQ Procedure				
Unsafe	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

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	35	36.46	35	36.46
N America	61	63.54	96	100.00

- i. What is the proportion of cars made in North America?

63.54 %

- ii. For the variables **UNSAFE**, **TYPE**, and **REGION**, are there any unusual data values that warrant further investigation?

No, there are no unusual data values for UNSAFE, TYPE, and REGION.

2. Performing Tests and Nominal Measures of Association.

- a) Generate a temporary format called **safefmt** to clearly identify the values of **UNSAFE**. Use the following information: 0='Average or Above Safety', 1='Below Average Safety'.

```
/*st008s02 a)*/
proc format;
  value safefmt 0='Average or Above'
            1='Below Average';
run;
```

- b) Use the FREQ procedure and the **st092.SAFETY** data set to calculate what percentage of cars made in Asia had a below-average safety score? Ensure the **safefmt** format is assigned.

```
/*st008s02 b)*/
proc freq data=st092.safety;
  tables Region*Unsafe/ format=15.;
  format Unsafe safefmt.;
  title "Association between Unsafe and Region";
run;
```

Table of Region by Unsafe				
Region	Frequency	Unsafe		Total
		Average or Above	Below Average	
Asia	20		15	35
	20.83		15.63	36.46
	57.14		42.86	
	30.30		50.00	
N America	46		15	61
	47.92		15.63	63.54
	75.41		24.59	
	69.70		50.00	

Total	66 68.75	30 31.25	96 100.00
-------	-------------	-------------	--------------

Region is a Row variable, so look at the Row Pct value in the Below Average cell of the Asia row. The percentage of cars made in Asia with a below-average safety score is 42.86.

- c) Use the FREQ procedure to perform an appropriate measure of association test between REGION and UNSAFE. Do you see a statistically significant association (at the 0.05 level)?

```
/*st008s02 c)*/
proc freq data=st092.safety;
  tables Region*Unsafe / chisq format=15.;
  format Unsafe safefmt.;
  title "Association between Unsafe and Region";
run;
```

Partial Output

Statistics for Table of Region by Unsafe			
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	

Step 1- Set Hypothesis

H₀: No Association exists between REGION and UNSAFE.

H₁: Association exists between REGION and UNSAFE.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

p-value=0.0631

Step 4- Decision Rule.

The p-value > α , therefore, you fail to reject the null hypothesis at the 0.05 level and conclude there is no evidence of an association between Region and Unsafe.

3. Performing Tests and Ordinal Measures of Association.

Use the **st092.SAFETY** data set to examine the ordinal association between **SIZE** and **UNSAFE**.

```
/*st008s03*/
proc freq data=st092.safety;
  tables Size*Unsafe / chisq measures cl format=15.;
  format Unsafe safefmt.;
  title "Association between Unsafe and Size";
run;
```

		The FREQ Procedure		
		Table of Size by Unsafe		
Size	Unsafe			Total
		Average or Above	Below Average	
1		12 12.50 34.29 18.18	23 23.96 65.71 76.67	35 36.46
2		24 25.00 82.76 36.36	5 5.21 17.24 16.67	29 30.21
3		30 31.25 93.75 45.45	2 2.08 6.25 6.67	32 33.33
	Total	66 68.75	30 31.25	96 100.00

Statistics for Table of Size by Unsafe				
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		

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				

- a) What statistic should you use to detect an ordinal association between **SIZE** and **UNSAFE**?

The Mantel-Haenszel Chi-Square.

- b) Do you reject or fail to reject the null hypothesis at the 0.05 level?

Step 1- Set Hypothesis

H₀: No Ordinal Association exists between SIZE and UNSAFE

H₁: An Ordinal Association exists between SIZE and UNSAFE.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

p-value=<.0001

Step 4- Decision Rule.

The p-value < α , therefore, you reject the null hypothesis at the 0.05 level and conclude there is evidence of an ordinal association between **SIZE** and **UNSAFE**.

- c) What is the strength of the ordinal association between **SIZE** and **UNSAFE**?

The Spearman Correlation is -0.5425, which concludes that there is a relatively strong negative ordinal association.

Chapter 9

1. Performing a Logistic Regression Model.

Fit a simple logistic regression model using **st092.SAFETY** with **UNSAFE** as the outcome variable and **REGION** as the predictor variable. Request reference cell coding with Asia as the reference level. Model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

```
/*st09s01*/
ods graphics on;
proc logistic data=st092.safety plots(only)=(effect oddsratio);
  class Region (param=ref ref='Asia');
  model Unsafe(event='1')=Region / clodds=pl;
  title1 'LOGISTIC MODEL (1):Unsafe=Region';
run;
quit;
ods graphics off;
```

Partial PROC LOGISTIC Output

```
LOGISTIC MODEL (1):Unsafe=Region
The LOGISTIC Procedure
Model Information
Data Set                      ST092.SAFETY
Response Variable              Unsafe
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          Total
Value      Unsafe  Frequency
1             0       66
2             1       30
Probability modeled is Unsafe=1.

Class Level Information
Design
Class      Value      Variables
Region    Asia       0
          N America  1

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

Model Fit Statistics
Intercept
Intercept and
```

Criterion	Only	Covariates			
AIC	121.249	119.854			
SC	123.813	124.982			
-2 Log L	119.249	115.854			
Testing Global Null Hypothesis: BETA=0					
Test	Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio	3.3949	1	0.0654		
Score	3.4541	1	0.0631		
Wald	3.3828	1	0.0659		
Type 3 Analysis of Effects					
Effect	DF	Chi-Square	Pr > ChiSq		
Region	1	3.3828	0.0659		
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.2876	0.3416	0.7089	0.3998
Region N America	1	-0.8329	0.4528	3.3828	0.0659
Odds Ratio Estimates					
Effect	Point Estimate	95% Wald Confidence Limits			
Region N America vs Asia	0.435	0.179 1.056			
Association of Predicted Probabilities and Observed Responses					
Percent Concordant	34.8	Somers' D	0.197		
Percent Discordant	15.2	Gamma	0.394		
Percent Tied	50.0	Tau-a	0.086		
Pairs	1980	c	0.598		
Profile Likelihood Confidence Interval for Odds Ratios					
Effect	Unit	Estimate	95% Confidence Limits		
Region N America vs Asia	1.0000	0.435	0.177 1.055		

- a) Do you reject or fail to reject the null hypothesis that all regression coefficients of the model are 0?

Step 1- Set Hypothesis

H₀: The baseline model is better (assuming the probability of average or above is equal to the probability of below average safety)

H₁: This model is better than the baseline model.

Step2-Set Significance level $\alpha=0.05$

Step 3 -Collect evidence

p-value=0.0654

Step 4- Decision Rule.

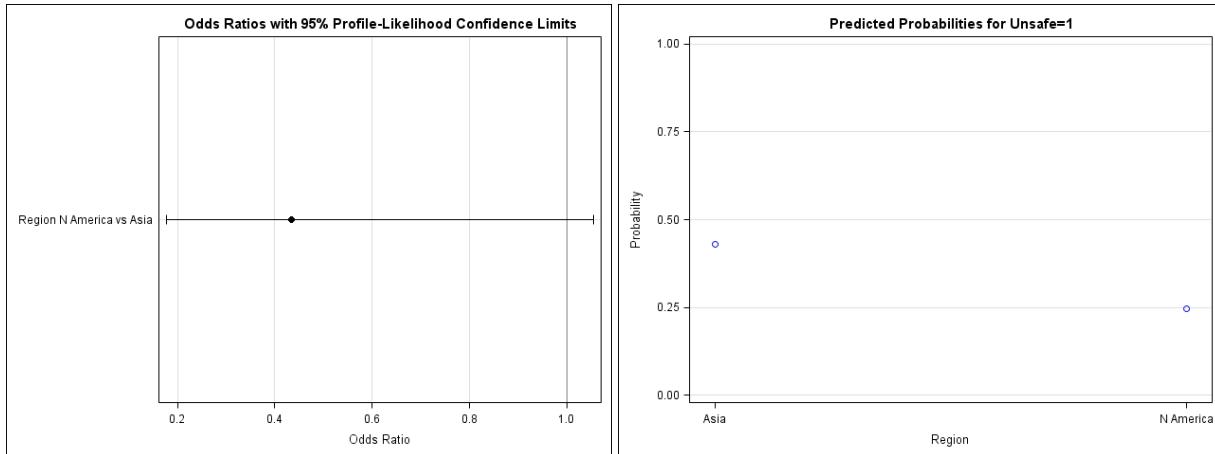
The p-value > α , Fail to Reject H_0 , therefore, baseline model is better than our model. We assume that the probability of average or above is equal to the probability of below average safety, i.e. the region does not make any difference.

- b) Write out the logistic regression equation.

The equation is: $\text{Logit(unsafe)} = -0.2876 - 0.8329 * (\text{Region} = 'N\ America')$.

- c) Interpret the odds ratio for REGION.

The odds ratio for region is 0.435, as this is less than 1, this indicates that Asia is more likely to be unsafe. To translate this, calculate $1/0.435 = 2.30$. Therefore Asia is 2.3 times more likely, with respect to odds, to be unsafe than North America.



The difference between Asian and North American cars is shown on the probability scale on the Effect Plot.

Chapter 10

1. Fitting a Multiple Logistic Regression Model

Use **UNSAFE** as the outcome variable and **WEIGHT**, **SIZE**, and **REGION** as the predictor variables. Ensure to model the probability of below-average safety scores.

- a) Create a format called **sizefmt** to indicate 1= Small, 2= Medium, 3=Large

```
/*st010s01 a)*/
ods graphics on;
proc format;
  value sizefmt 1='Small'
                2='Medium'
                3='Large';
run;
```

- b) Use the LOGISTIC procedure and specify **REGION** and **SIZE** as classification variables using reference cell coding. Specify Asia as the reference level for **REGION**. For the **SIZE** variable, apply the **sizefmt** and specify Small as the reference level. Request any relevant plots.

 The variable **SIZE** is coded (1, 2, 3), but the applied format requires that the formatted value be used in the CLASS statement for the REF= category.

Use **Unsafe** as the outcome variable and **Weight**, **Size**, and **Region** as the predictor variables. Use the EVENT= option to model the probability of below-average safety scores.

```
/*st010s01 b)*/
proc logistic data=st092.safety plots(only)=(effect oddsratio);
  class Region (param=ref ref='Asia')
        Size (param=ref ref='Large');
  model Unsafe(event='1')=Weight Region Size / clodds=pl;
  format Size sizefmt.;
  title1 'LOGISTIC MODEL (2):Unsafe=Weight Region Size';
run;
```

```
LOGISTIC MODEL (2):Unsafe=Weight Region Size
The LOGISTIC Procedure
Model Information
Data Set                      ST092.SAFETY
Response Variable              Unsafe
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      Total
          Unsafe    Frequency
```

1	0	66
2	1	30

Probability modeled is Unsafe=1.

Class Level Information

Class	Value	Design Variables	
		Region	N America
Asia	0		
N America	1		
Size	Large	0	0
	Medium	1	0
	Small	0	1

Model Convergence Status

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

The LOGISTIC Procedure

Model Fit Statistics

Criterion	Intercept		Intercept and Covariates
	Only	Covariates	
AIC	121.249	94.004	
SC	123.813	106.826	
-2 Log L	119.249	84.004	

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	35.2441	4	<.0001
Score	32.8219	4	<.0001
Wald	23.9864	4	<.0001

Type 3 Analysis of Effects

Effect	DF	Chi-Square	Pr > ChiSq
Weight	1	2.1176	0.1456
Region	1	0.4506	0.5020
Size	2	15.3370	0.0005

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard		Wald	Pr > ChiSq
		Estimate	Error	Chi-Square	
Intercept	1	0.0500	1.8008	0.0008	0.9778
Weight	1	-0.6678	0.4589	2.1176	0.1456
Region N America	1	-0.3775	0.5624	0.4506	0.5020
Size Medium	1	0.6582	0.9231	0.5085	0.4758
Size Small	1	2.6783	0.8810	9.2422	0.0024

Odds Ratio Estimates				
Effect	Point Estimate	95% Wald Confidence Limits		
Weight	0.513	0.209	1.261	
Region N America vs Asia	0.686	0.228	2.064	
Size Medium vs Large	1.931	0.316	11.793	
Size Small vs Large	14.560	2.590	81.857	

Association of Predicted Probabilities and Observed Responses				
Percent Concordant	81.9	Somers' D	0.696	
Percent Discordant	12.3	Gamma	0.739	
Percent Tied	5.8	Tau-a	0.302	
Pairs	1980	c	0.848	

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	Unit	Estimate	95% Confidence Limits	
Weight	1.0000	0.513	0.201	1.260
Region N America vs Asia	1.0000	0.686	0.225	2.081
Size Medium vs Large	1.0000	1.931	0.343	15.182
Size Small vs Large	1.0000	14.560	3.018	110.732

- i. Interpret the odds ratio estimates for **WEIGHT** and **SIZE**, in the Profile Likelihood Table.

The odds ratio for **WEIGHT** is 0.513 which indicates that odds ratio nearly doubles for each 1000 pounds lighter a car is.

The odds ratio for **SIZE Medium vs. Large** is 1.931 indicates that Medium cars have nearly twice the odds of having a below average safety rating compared with Large cars.

The odds ratio for **SIZE Small vs. Large** is 14.560 indicates that Small cars have nearly 15 times the odds of having a below average safety rating compared with Large cars.

- ii. Do you think this is a better model than the one fit with just **REGION**?

The AIC (94.004) value for this model is lower than for the previous model (119.854). The same is true for the SC (106.826 versus 124.982). This indicates that the present model fits better than the previous model. Of course, this model is statistically significant at the 0.05 level, whereas the previous model was not.

