



Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Course Notes

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression Course Notes was developed by Marc Huber and Danny Modlin. Additional contributions were made by Lee Bennett, Chris Daman, Tarek Elnaccash, Joanne Lo, Bob Lucas, Diane K. Michelson, Mike Patetta, and Catherine Truxillo, 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.

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression Course Notes

Copyright © 2017 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 E70972, course code LWST142/ST142, prepared date 25Apr2017. LWST142_001

ISBN 978-1-63526-090-8

Table of Contents

Chapter 1 Course Overview and Review of Concepts	1-1
1.1 Course Overview	1-3
Demonstration: Ames Home Sales Data Set Exploration	1-11
1.2 Quick Review of Statistical Concepts	1-33
1.3 One-Sample <i>t</i> -Tests.....	1-42
Demonstration: PROC TTEST for a One-Sample <i>t</i> -Test.....	1-47
Exercises.....	1-52
1.4 Two-Sample <i>t</i> -Tests	1-53
Demonstration: Two-Sample <i>t</i> -Test	1-57
Exercises.....	1-61
1.5 Solutions.....	1-63
Chapter 2 ANOVA and Regression	2-1
2.1 Graphical Analysis	2-3
Demonstration: Exploring Associations.....	2-9
2.2 One-Way ANOVA.....	2-20
Demonstration: Performing a One-Way ANOVA.....	2-36
Exercises.....	2-43
2.3 ANOVA Post Hoc Tests	2-44
Demonstration: Post Hoc Pairwise Comparisons	2-50
Exercises.....	2-56
2.4 Pearson Correlation	2-57
Demonstration: Data Exploration, Correlations, and Scatter Plots	2-66
Exercises.....	2-76

2.5	Simple Linear Regression	2-78
	Demonstration: Performing Simple Linear Regression	2-88
	Exercises.....	2-95
2.6	Solutions	2-96
Chapter 3 More Complex Linear Models		3-1
3.1	Two-Way ANOVA and Interactions	3-3
	Demonstration: Two-Way ANOVA.....	3-6
	Demonstration: Two-Way ANOVA with an Interaction	3-22
	Exercises.....	3-30
3.2	Multiple Regression	3-31
	Demonstration: Fitting a Multiple Linear Regression Model.....	3-39
	Exercises.....	3-46
3.3	Solutions.....	3-48
Chapter 4 Model Building and Effect Selection.....		4-1
4.1	Stepwise Selection Using Significance Level.....	4-3
	Demonstration: Stepwise Regression	4-10
	Exercises.....	4-20
4.2	Information Criterion and Other Selection Options	4-21
	Demonstration: Model Selection Using AIC, AICC, BIC, and SBC	4-24
	Exercises.....	4-37
4.3	All Possible Selection (Self-Study)	4-38
	Demonstration: All Possible Model Selection	4-42
	Exercises.....	4-49
4.4	Solutions	4-50

Chapter 5 Model Post-Fitting for Inference5-1

5.1 Examining Residuals	5-3
Demonstration: Residual Plots	5-10
Exercises.....	5-17
5.2 Influential Observations	5-18
Demonstration: Looking for Influential Observations	5-25
Exercises.....	5-36
5.3 Collinearity	5-37
Demonstration: Example of Collinearity	5-42
Demonstration: Collinearity Diagnostics	5-44
Exercises.....	5-50
5.4 Solutions	5-51

Chapter 6 Model Building and Scoring for Prediction6-1

6.1 Brief Introduction to Predictive Modeling	6-3
Demonstration: Predictive Model Building.....	6-9
Exercises.....	6-15
6.2 Scoring Predictive Models.....	6-16
Demonstration: Scoring Using PROC PLM and PROC GLMSELECT	6-19
Exercises.....	6-21
6.3 Solutions	6-22

Chapter 7 Categorical Data Analysis7-1

7.1 Describing Categorical Data	7-3
Demonstration: Examining Distributions	7-10
7.2 Tests of Association	7-19
Demonstration: Chi-Square Test.....	7-27

Demonstration: Detecting Ordinal Associations	7-35
Exercises.....	7-38
7.3 Introduction to Logistic Regression	7-40
Demonstration: Simple Logistic Regression Model.....	7-49
Exercises.....	7-61
7.4 Logistic Regression with Categorical Predictors.....	7-62
Demonstration: Multiple Logistic Regression with Categorical Predictors	7-69
Exercises.....	7-75
7.5 Stepwise Selection with Interactions and Predictions	7-76
Demonstration: Logistic Regression: Backward Elimination with Interactions	7-79
Demonstration: Logistic Regression: Predictions Using PROC PLM	7-88
Exercises.....	7-90
7.6 Solutions.....	7-91
Appendix A References	A-1
A.1 References	A-3
Appendix B Sampling from SAS Data Sets	B-1
B.1 Random Samples	B-3
Appendix C Additional Topics.....	C-1
C.1 Paired <i>t</i> -Tests	C-3
Demonstration: Paired <i>t</i> -Test.....	C-5
C.2 One-Sided <i>t</i> -Tests	C-11
Demonstration: One-Sided <i>t</i> -Test.....	C-13
C.3 Nonparametric ANOVA.....	C-15
Demonstration: The NPAR1WAY Procedure for Hospice Referral Data	C-21

Demonstration: The NPAR1WAY Procedure for Small Samples.....	C-31
C.4 Partial Regression Plots.....	C-34
Demonstration: Partial Regression Plots	C-36
C.5 Exact Tests for Contingency Tables	C-40
Demonstration: Fisher's Exact <i>p</i> -Values for the Pearson Chi-Square Test.....	C-45
C.6 Empirical Logit Plots	C-47
Demonstration: Fisher's Exact <i>p</i> -Values for the Pearson Chi-Square Test.....	C-51
Appendix D Percentile Definitions.....	D-1
D.1 Calculating Percentiles	D-3
Appendix E Writing and Submitting SAS® Programs in SAS® Enterprise Guide®	E-1
E.1 Writing and Submitting SAS Programs in SAS Enterprise Guide	E-3
Demonstration: Adding a SAS Program to a Project	E-11

To learn more...



For information about 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 <http://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 the 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 SAS Bookstore on the web at <http://support.sas.com/publishing/> for a complete list of books and a convenient order form.

Chapter 1 Course Overview and Review of Concepts

1.1 Course Overview.....	1-3
Demonstration: Ames Home Sales Data Set Exploration.....	1-11
1.2 Quick Review of Statistical Concepts	1-33
1.3 One-Sample <i>t</i>-Tests.....	1-42
Demonstration: PROC TTEST for a One-Sample <i>t</i> -Test	1-47
Exercises	1-52
1.4 Two-Sample <i>t</i>-Tests.....	1-53
Demonstration: Two-Sample <i>t</i> -Test.....	1-57
Exercises	1-61
1.5 Solutions.....	1-63
Solutions to Exercises.....	1-63
Solutions to Student Activities (Polls/Quizzes).....	1-70

1.1 Course Overview

Objectives

- Give overview of the models presented in the course.
- Compare inferential statistics with predictive modeling.
- Introduce the Ames Home Sales data set.
- Decide what tasks to complete before analyzing the data.
- Produce descriptive statistics for both categorical and interval level variables.

3



1.01 Multiple Choice Poll

How much training do you have in statistics?

- a. None whatsoever
- b. One introductory course
- c. Several courses
- d. An undergraduate degree in statistics or related field
- e. An advanced degree in statistics or a related field

4



Overview of Models in This Course

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Logistic Regression	Logistic Regression	Logistic Regression

This course deals with statistical modeling. The type of modeling depends on the level of measurement of two types of variables.

The first type of variable is called *Response*. These are the variables that generally are the focus of business or research. They are also known as *outcome variables* or *target variables* or (in designed experiments) *dependent variables*.

The second type of variable is referred to as *predictor variables*. These are the measures that are theoretically associated with the response variables. They can therefore be used to “predict” the value of the response variables. They are also known as *independent variables* in analysis of data from designed experiments.

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

Linear Models Terminology

- Linear Model
- Inferential Statistics (Explanatory Modeling)
- Predictive Modeling
- Explanatory (Input, Predictor) Variable
- Response (Target) Variable

6

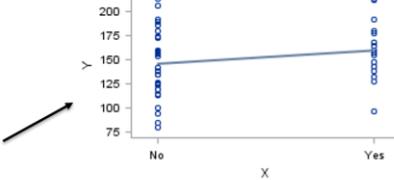
Copyright © SAS Institute Inc. All rights reserved.



Overview of Models

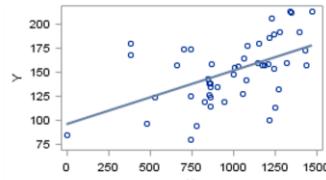
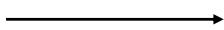
- General Linear Models

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



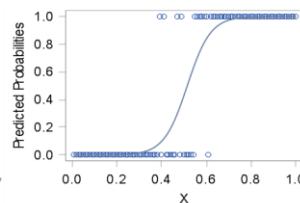
- Analysis of Variance (ANOVA)

- Regression



- Logistic Regression

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



Copyright © SAS Institute Inc. All rights reserved.



Models in this course can be most generally categorized as *Generalized Linear Models*. In every case, there is a *response* (or *target*) variable, which is the variable of interest, and the *explanatory* (or *predictor*) variable(s), which are used to model and predict the level of the response variable.

When the response variable is continuous and you can assume a normal distribution of errors, you can use a *General Linear Model* to model the relationship between predictor variables and response variables. You perform ordinary least squares regression, analysis of variance (ANOVA), or analysis of covariance (ANCOVA), depending on whether the explanatory variables are all continuous, all categorical, or a combination of continuous and categorical, respectively.

When the response variable is categorical, there can be an indirect modeling of the variable, by use of a *link function*. When the response variable is binary (can take on only 2 values) then the link function is typically the *logit* and the analysis is called *logistic regression*, regardless of the level of measurement of any of the explanatory variables. This type of modeling will be described in greater detail in a later chapter.

The defining feature of linear models is the linear function of the explanatory variables. The regression coefficients are just numbers and they are multiplied by the explanatory variable values. These products are then summed to get the individual's *predicted value*.

Explanatory versus Predictive Modeling

Explanatory Modeling

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_k X_{ki}$$

- How is X related to Y?
- Descriptions
- Small Sample
- Few Variables
- Assessed using *p*-values and confidence intervals

Predictive Modeling

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_k X_{ki}$$

- If I know X_i , can I predict Y_i ?
- Predictions
- Large Sample
- Many Variables
- Assessed using holdout sample validation

Explanatory Modeling typically refers to modeling using *inferential statistical methods*. These are the classical statistical methods that most students learn in their first statistics courses. The focus is on descriptions about the nature of relationships among variables and inference about pre-stated hypotheses about the data and the relationships among variables. Looking at the model equations, explanatory modeling focuses on the estimates of the beta coefficients. Those values say something about the explanatory variables' relationships with the response variable. Confidence limits and *p*-values are analyzed closely to determine confidence about estimates and about decisions about the existence of nonzero relationships between the predictor variables and the response variable. Distributional assumptions are vital and the goal is finding the "true relationships" among variables. In explanatory models, samples are usually small and there are few explanatory variables.

Predictive Modeling typically refers to methods for finding the most accurate predictions for future values of the response or target variable. Sample sizes are usually quite large, rendering statistical hypothesis testing virtually useless, as nearly every relationship appears *statistically significant*. The focus is not so much on the parameters of the model as it is in the predictions of observations. In the model equations shown above, these are represented on the left side of the equation. In predictive models, sample sizes are typically very large and there are many more explanatory variables (often referred to as *predictor variables* or *inputs*).

Assessment methods of the adequacy of a model differ between explanatory and predictive models. Whereas the adequacy of explanatory models is usually assessed using classic statistical metrics, such as *p*-values and confidence intervals of parameter estimates, the adequacy of predictive models is usually assessed by comparing observed to predicted values on a holdout sample of data not used to create the model.

The Ames Home Sales Data Set



Sas

The data for the instructional demonstrations in this course were collected by Dr. Dean DeCock, of Truman State University in Kirksville, Missouri, USA. The full description of the data set is provided in the [Journal of Statistics Education](#). The data set contains information about the sale of individual residential property in Ames, Iowa, from 2006 to 2010.

The data set **STAT1.AmesHousing** contains all original data from Dr. DeCock, including 2,930 observations and a large number of explanatory variables involved in assessing home values. In addition, some summary variables were calculated, as well as a variable calculated as the natural log of the sale price of the home.

The data set **STAT1.AmesHousing2** is a subset of the full data set, for homes with normal sales conditions (to avoid analyzing foreclosure or distressed sales) and gross living area of 1,500 square feet or less (to focus on homes of modest size).

The data set **STAT1.AmesHousing3** is a random sample of 300 houses, which will be used for all of the demonstrations in the course.

The FREQ Procedure

General form of the FREQ procedure:

```
PROC FREQ DATA=SAS-data-set;
  TABLES table-requests </ options>;
  RUN;
```

10

Copyright © SAS Institute Inc. All rights reserved.



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.

Note: PROC FREQ can generate large volumes of output as the number of variables or the number of variable levels (or both) increases.

The UNIVARIATE Procedure

General form of the UNIVARIATE procedure:

```
PROC UNIVARIATE DATA=SAS-data-set <options>;
  VAR variables;
  HISTOGRAM variables </options>;
  INSET keywords </options>;
  RUN;
```

11

Copyright © SAS Institute Inc. All rights reserved.



The UNIVARIATE procedure not only computes descriptive statistics, but also provides greater detail about 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.
- HISTOGRAM creates high-resolution histograms.
- 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.

SAS Studio

In class, you can use SAS Studio to develop code. SAS Studio is the new browser-based SAS programming environment.



<http://www.sas.com/gobot/SASSstudioTutorial>

12

Copyright © SAS Institute Inc. All rights reserved.

Sas

Interactive Mode

Some SAS procedures, such as PROC IMSTAT, are interactive. That means they remain active until you submit a QUIT statement, or until you submit a new PROC or DATA step.

In SAS Studio, you can use the code editor to run these procedures, as well as other SAS procedures, in interactive mode.



By default, SAS Studio does not run in interactive mode.
This icon in SAS Studio toggles interactive mode on and off.

13

Copyright © SAS Institute Inc. All rights reserved.

Considerations for Running in Interactive Mode

- Interactive mode starts a new SAS session.
- Librefs and macro variables used in the course must be defined for each new SAS session.
- SAS Studio Documentation: <http://sww.sas.com/gobot/SASSStudioDoc>

14

Copyright © SAS Institute Inc. All rights reserved.



Ames Home Sales Data Set Exploration

Example: Use the PRINT procedure to list the first 10 observations in the **STAT1.AmesHousing3** data set. Then use PROC UNIVARIATE to generate plots and descriptive statistics for continuous variables and PROC FREQ to generate plots and tables for categorical variables.

Note: Throughout the course, predefined SAS Studio tasks are used to generate SAS code for the analysis. Equivalent code is also provided.

Note: Open and submit the program **st100d05.sas** before running the programs in this course. This program calls the data creation programs, as well as the format program.

```
/*st100d05.sas*/
%let homefolder=S:\Workshop;
libname STAT1 "&homefolder";
%include "&homefolder\st100d01.sas";
%include "&homefolder\st100d02.sas";
%include "&homefolder\st100d03.sas";
%include "&homefolder\st100d04.sas";
```

Note: Change the location of **homefolder** to the folder where you have stored the program files for this course.

Note: Currently, SAS Studio does not include the option for user-specified formats. Some of the output plots produced in this course included format options that could be reproduced using the code provided.

Partial Log

```
1 /*st100d05.sas*/
3 %let homefolder=S:\Workshop;
4 %include "&homefolder\st100d01.sas";

NOTE: Missing values were generated as a result of performing an operation on missing values.
      Each place is given by: (Number of times) at (Line):(Column).
      1 at 494:19
NOTE: The data set STAT1.AMESHOUING has 2930 observations and 98 variables.
NOTE: DATA statement used (Total process time):
      real time          0.03 seconds
      cpu time          0.03 seconds

3476 %include "&homefolder\st100d02.sas";

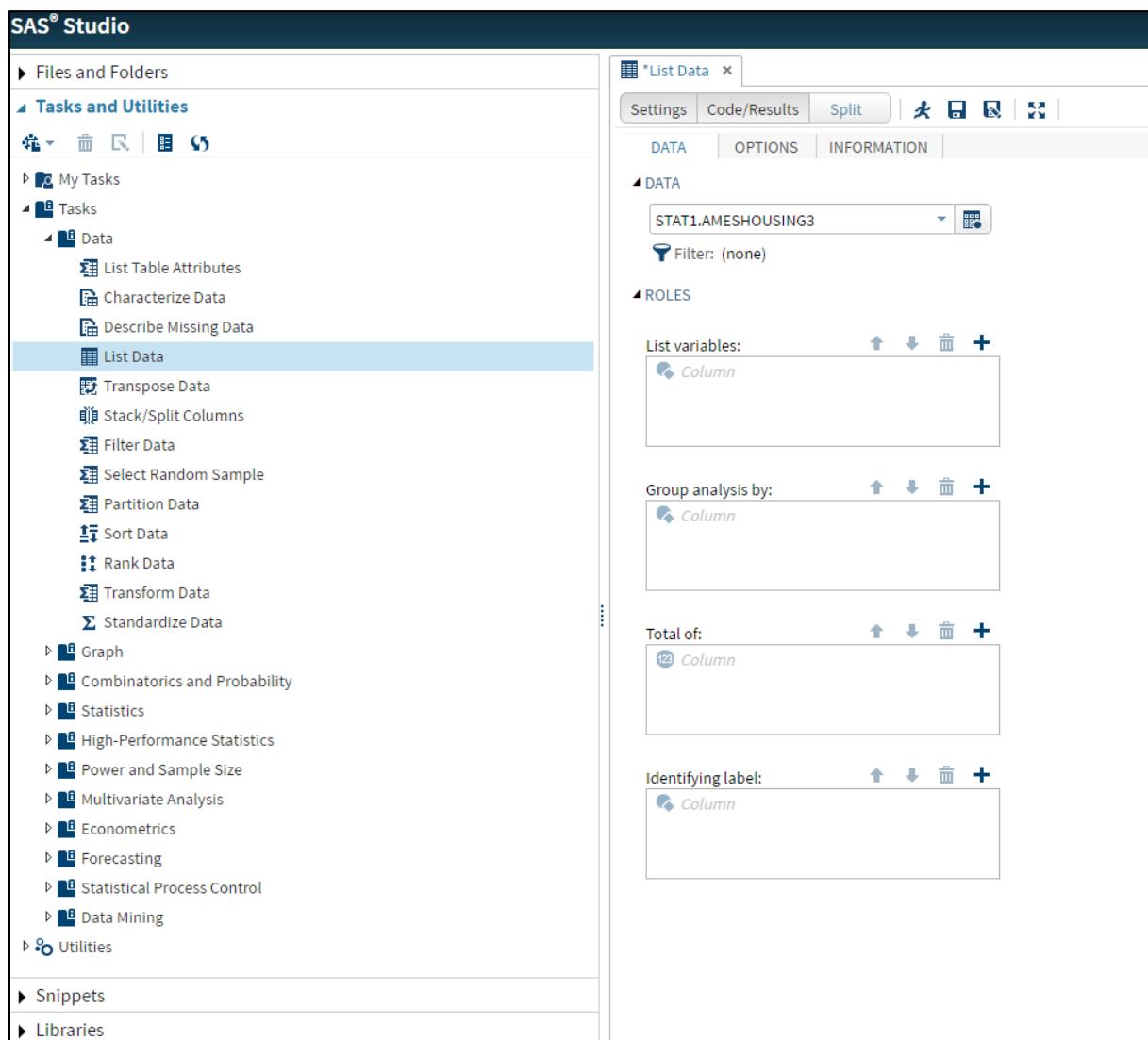
NOTE: There were 1361 observations read from the data set STAT1.AMESHOUING.
      WHERE (Sale_Condition='Normal') and (Gr_Liv_Area<=1500);
NOTE: The data set STAT1.AMESHOUING2 has 1361 observations and 30 variables.
NOTE: DATA statement used (Total process time):
      real time          0.01 seconds
      cpu time          0.01 seconds

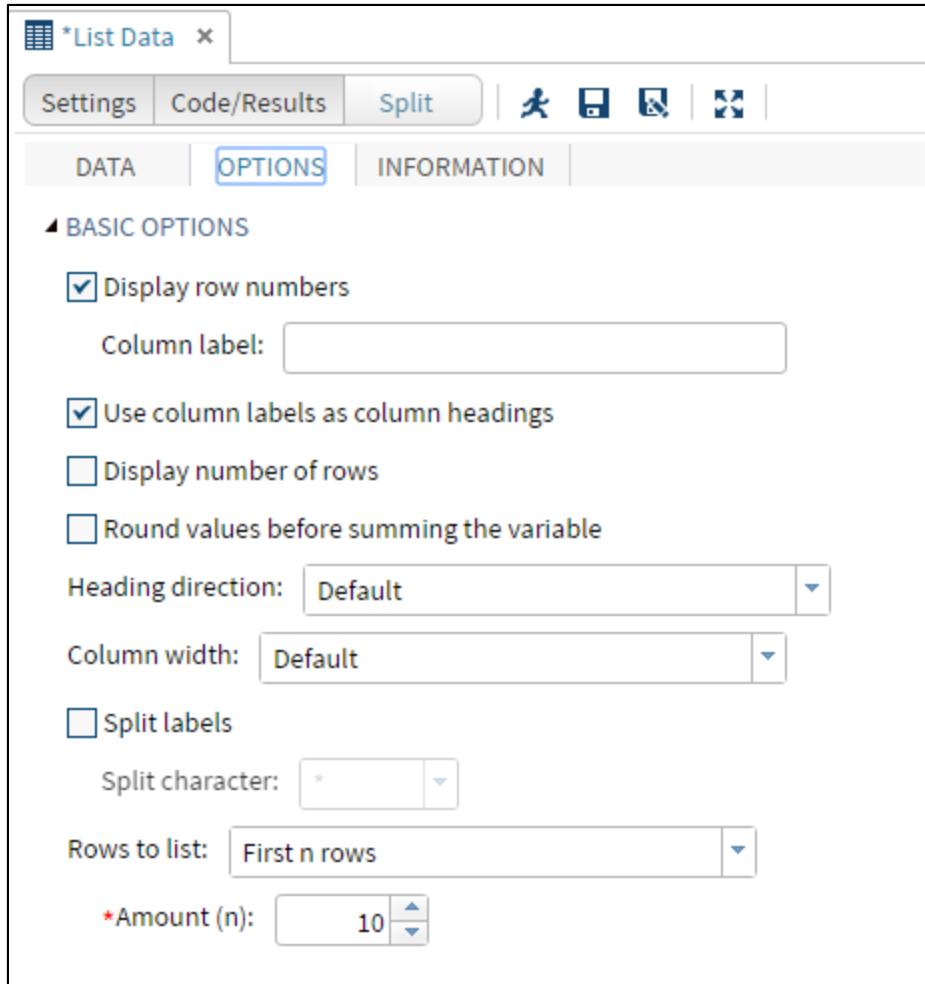
NOTE: The data set STAT1.AMESHOUING3 has 300 observations and 30 variables.
NOTE: PROCEDURE SURVEYSELECT used (Total process time):
      real time          0.00 seconds
```

Listing the First 10 Observations in the Data Set

1. On the left side of the SAS Studio interface, expand **Tasks** and then expand **Data**. Double-click **List Data** to initiate it.
2. Click the **Select a Table** button  under the Data property to navigate to the **STAT1** library. Select the **AmesHousing3** data table.
3. To specify the number of rows to print, click the **OPTIONS** tab on top. Use the drop-down menu and change **Rows to List** from **All rows** to **First n rows**. Enter the number of rows to print. The default number of rows is 10.
4. To submit the code and run the List Data task, click the **Run** toolbar button .

These steps are summarized in the display below.





A subset of the contents of the CODE window is shown below.

```
title1 "List Data for STAT1.AMESHOUSING3";

proc print data=STAT1.AMESHOUSING3
  (obs=10) label;
run;

title1;
```

Partial Output

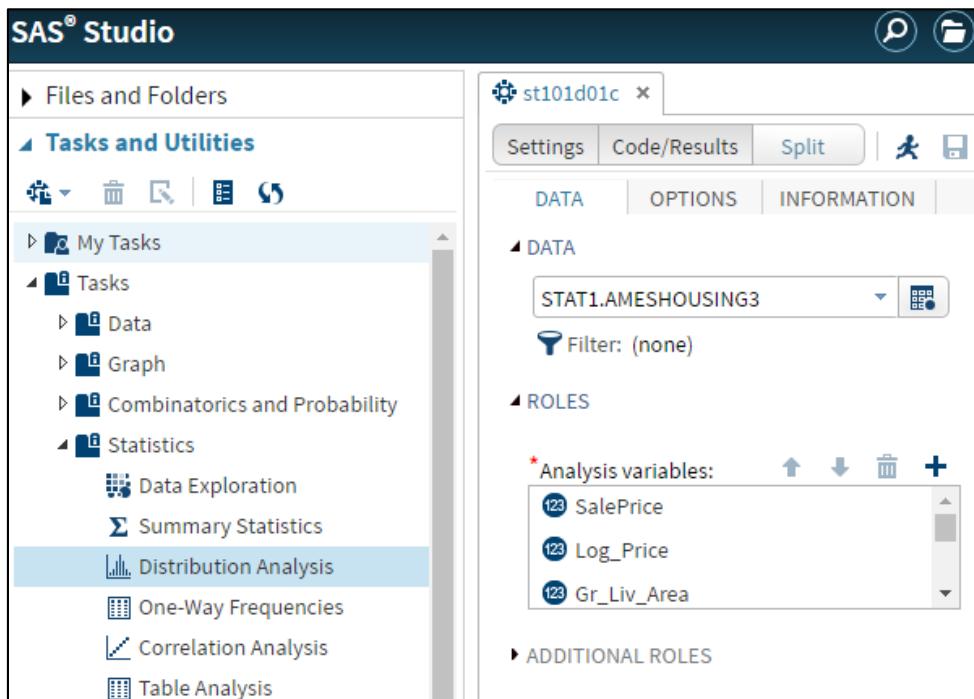
List Data for STAT1.AMESHOUSING3

Obs	PID	Sale price in dollars	Natural log of the sale price	Above grade (ground) living area square feet	Bedrooms above grade	Number of full bathrooms	Number of half bathrooms	Total number of bathrooms (half bathrooms counted 10%)	Total area of decks and porches in square feet
1	0527127150	213500	12.271392112	1338	2	3	0	3	0
2	0527145080	191500	12.162643088	1280	2	2	0	2	226
3	0527425090	115000	11.652687407	864	3	1	0	1	0
4	0528228285	160000	11.982929094	1145	2	2	0	2	216
5	0528250100	180000	12.10071213	1430	3	2	1	2.1	180
6	0531452050	125000	11.736069016	752	2	2	0	2	443
7	0533253210	206000	12.235631448	1226	1	2	0	2	301
8	0534401110	159000	11.976659481	1209	3	2	0	2	0
9	0534403410	180500	12.103486057	1152	3	2	0	2	227
10	0534430080	142125	11.864462231	1078	2	2	0	2	366

Obtaining Descriptive Statistics for Continuous Variables

5. Open the **Distribution Analysis** task under Statistics. Note that **Ameshousing3** is already selected as the data. SAS Studio automatically selects the last data set that was used.

6. Under Roles, use the plus sign  to select **Analysis Variables**. In this example, select all the continuous variables: **SalePrice**, **Log_Price**, **Gr_Liv_Area**, **Basement_Area**, **Garage_Area**, **Deck_Porch_Area**, **Lot_Area**, **Age_Sold**, **Bedroom_AbvGr**, **Full_Bathroom**, **Half_Bathroom**, and **Total_Bathroom**.



7. On the OPTIONS tab, check the boxes to include a normal curve and kernel density estimate for the histograms. To include specific summary statistics, check the option to **Add inset statistics** and expand the **Inset Statistics** properties and select **Mean** and **Standard deviation** in addition to the default of **Number of observations**.

DATA OPTIONS INFORMATION

▲ EXPLORING DATA

Histogram

Classification variables: (2 items) ↑ ↓ - +

Column

Add normal curve

Add kernel density estimate

Add inset statistics

▲ Inset Statistics

Number of observations

Mean

Median

Standard deviation

Variance

Skewness

Kurtosis

▲ CHECKING FOR NORMALITY

Histogram and goodness-of-fit tests

Normal probability plot

Normal quantile-quantile plot

► FITTING DISTRIBUTIONS

8. Submit the code by clicking the **Run** toolbar button.

The code generated by SAS Studio is as follows:

```
/** Exploring Data ***/
proc univariate data=STAT1.AMESHOUSING3;
ods select Histogram;
var SalePrice Log_Price Gr_Liv_Area Basement_Area Garage_Area
    Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
    Full_Bathroom Half_Bathroom Total_Bathroom;
histogram SalePrice Log_Price Gr_Liv_Area Basement_Area
    Garage_Area Deck_Porch_Area Lot_Area Age_Sold
    Bedroom_AbvGr Full_Bathroom Half_Bathroom
    Total_Bathroom / normal kernel;
inset n mean std / position=ne;
run;
```

Note: Alternatively, you can write the code directly in SAS. Macro variables are created to help organize the use of variables and make modification of the SAS code easier. The **%LET** statements are used to name the macro variables and set their values. The macro variables are referred to in the SAS code as **&categorical** and **&interval** to distinguish those names from those of variables. Example use of macro variable is shown below.

```
/*st101d01.sas*/ /*Part A*/
/*Exploration of all variables that are available for analysis.*/
/*%let statements define macro variables containing lists of */
/*dataset variables*/
%let categorical=House_Style Overall_Qual Overall_Cond Year_Built
    Fireplaces Mo_Sold Yr_Sold Garage_Type_2 Foundation_2
    Heating_QC Masonry_Veneer Lot_Shape_2 Central_Air;
%let interval=SalePrice Log_Price Gr_Liv_Area Basement_Area
    Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
    Full_Bathroom Half_Bathroom Total_Bathroom ;
```

```
/*st101d01.sas*/ /*Part C*/
/*PROC UNIVARIATE provides summary statistics and plots for */
/*interval variables. The ODS statement specifies that only */
/*the histogram be displayed. The INSET statement requests */
/*summary statistics without having to print out tables.*/
ods select histogram;
proc univariate data=STAT1.ameshousing3 noprint;
var &interval;
histogram &interval / normal kernel;
inset n mean std / position=ne;
title "Interval Variable Distribution Analysis";
run;
```

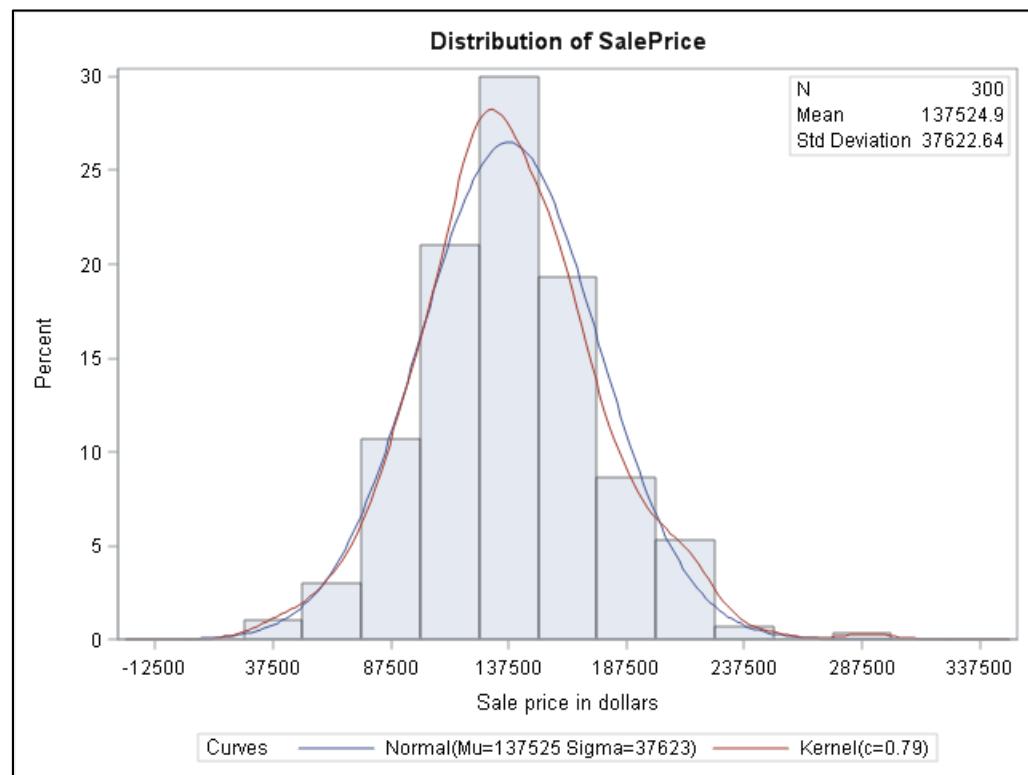
Selected option for HISTOGRAM and PROBPLOT statements:

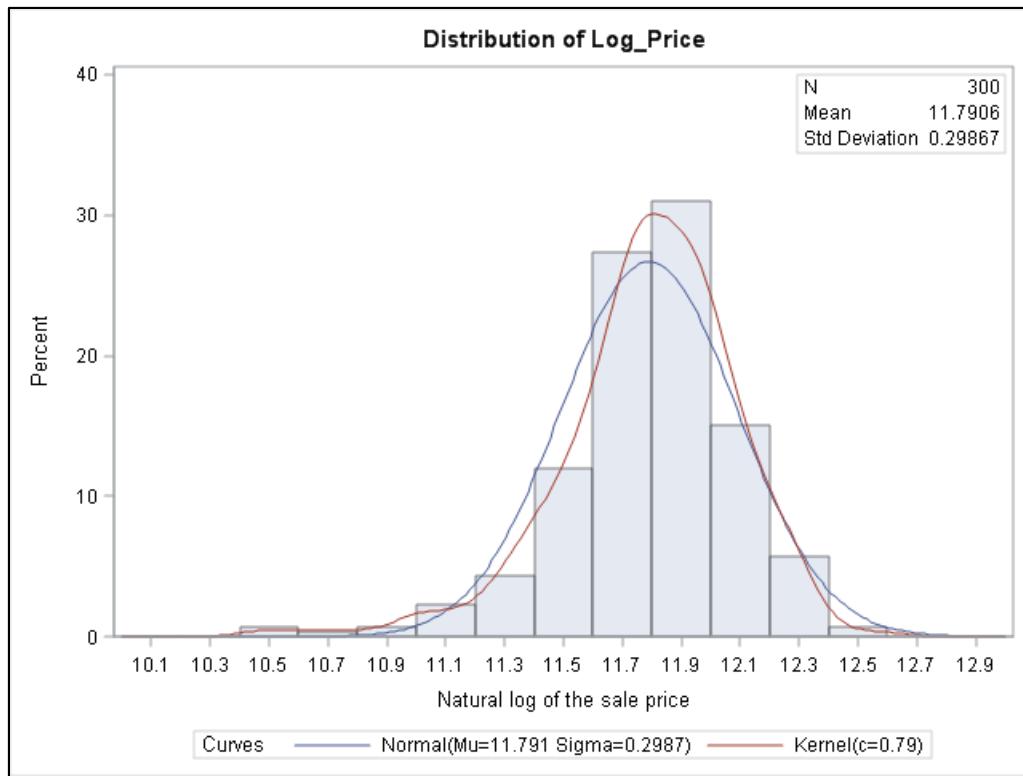
- NORMAL<(options)> creates a normal probability plot. Options (MU= SIGMA=) determine the mean and standard deviation of the normal distribution used to create reference lines (normal curve overlay in HISTOGRAM and diagonal reference line in PROBPLOT). To use sample means and standard deviations for the normal curve parameters, use MU=EST and SIGMA=EST. If these options are not specified, then MU=EST and SIGMA=EST are the default settings.
- KERNEL<(options)> Superimposes kernel density estimates on the histogram.

Selected option for INSET statement:

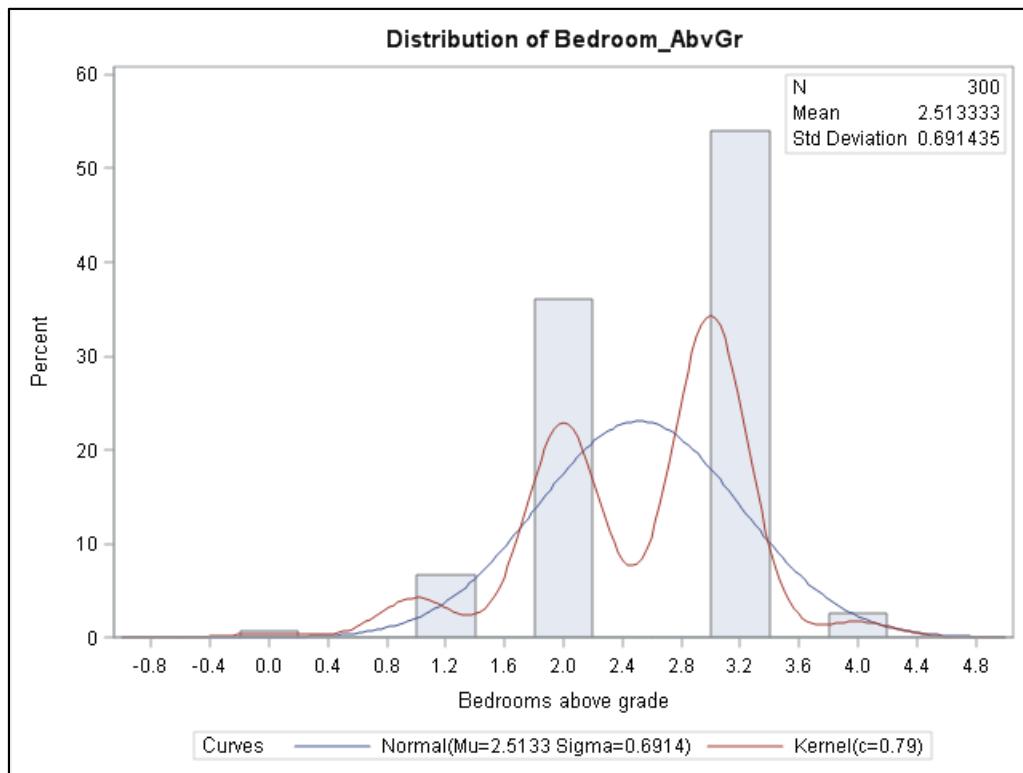
- POSITION=NE Determines the position of the inset. The position is a compass point keyword, a margin keyword, or a pair of coordinates (x,y). You can specify coordinates in axis percent units or axis data units. The default value is NW, which positions the inset in the upper left (northwest) corner of the display.

Partial Output

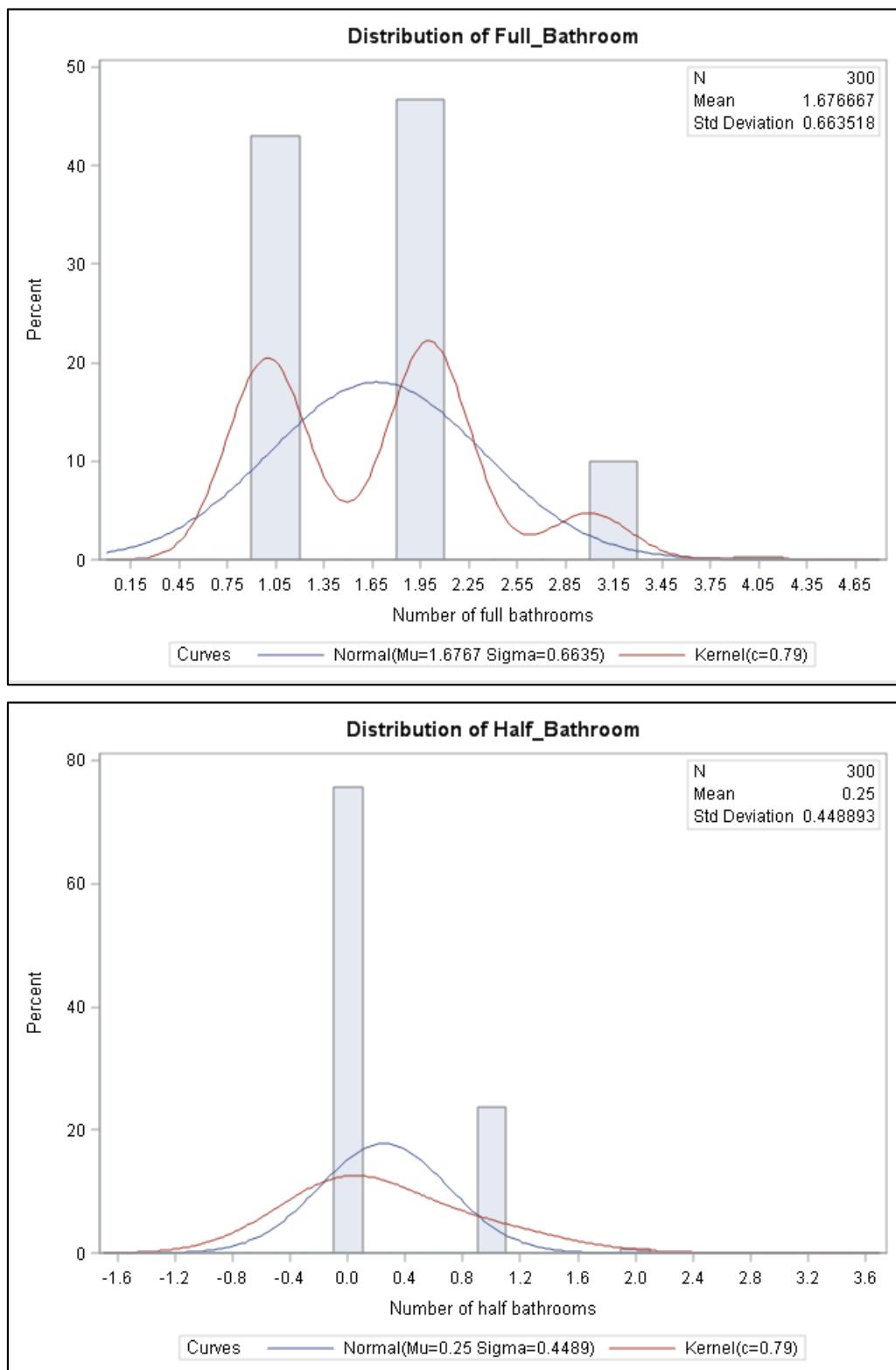


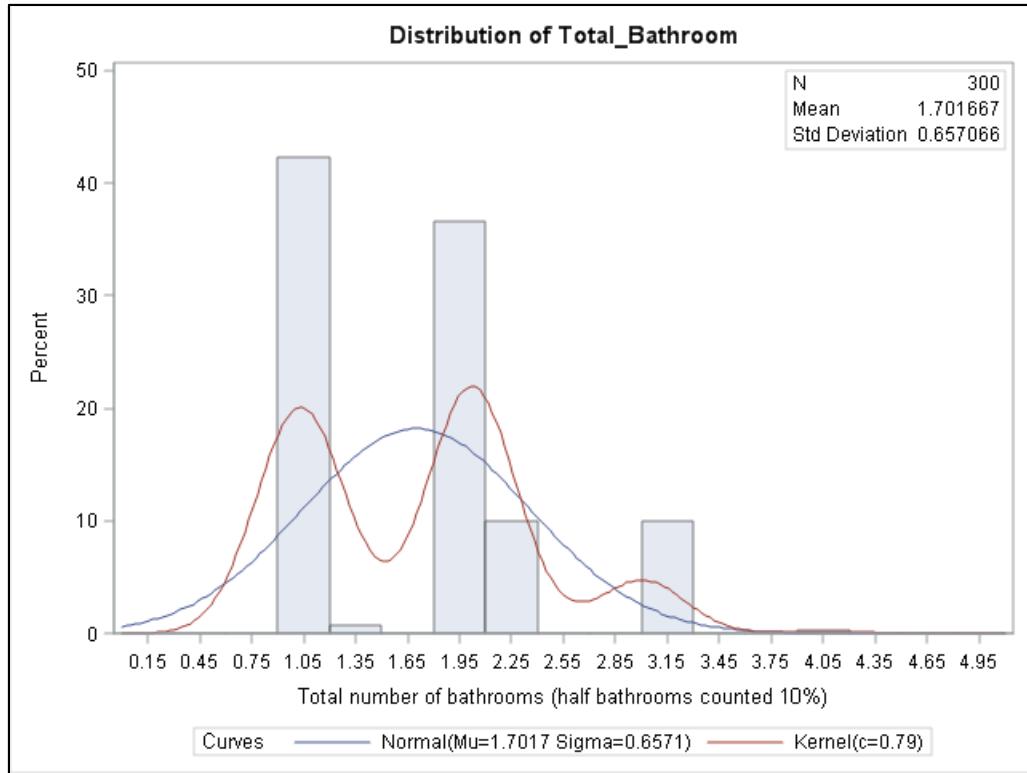


Sale price for the entire data set is skewed to the right. **Log_Price** is relatively normally distributed for the entire data set. However, for houses 1,500 square feet or less in gross living area, **Sale_Price** itself is relatively normally distributed. This can be seen by comparing the normal and kernel density curves. They are relatively similar in shape.



The number of bedrooms above grade (**Bedroom_AbvGr**) is discrete with relatively few observed values. It could be treated as a categorical (ordinal) variable in analysis.





Half-bathrooms are counted as 0.1 full bathrooms for this variable.

Generating Plots and Tables for Categorical Variables

9. Select the **Table Analysis** task under **Statistics**.
10. Under the **Roles** property, add variables of interest to the **Row variables:** **House_Style**, **Overall_Qual**, **Overall_Cond**, **Year_Built**, **Fireplaces**, **Mo_Sold**, **Yr_Sold**, **Garage_Type_2**, **Foundation_2**, **Heating_QC**, **Masonry_Veneer**, **Lot_Shape_2**, **Central_Air**.

SAS® Studio

Files and Folders

Tasks and Utilities

- My Tasks
- Tasks
 - Data
 - Graph
- Combinatorics and Probability
- Statistics
 - Data Exploration
 - Summary Statistics
 - Distribution Analysis
 - One-Way Frequencies
 - Correlation Analysis
 - Table Analysis** (selected)
 - t Tests
 - One-Way ANOVA
 - Nonparametric One-Way ANOV
 - N-Way ANOVA
 - Analysis of Covariance
 - Linear Regression
 - Binary Logistic Regression
 - Predictive Regression Models
 - Generalized Linear Models

st101d01b

DATA **OPTIONS** **INFORMATION**

DATA

STAT1.AMESHOUSING3

Filter: (none)

ROLES

Row variables:

- House_Style
- Overall_Qual
- Overall_Cond

Column variables:

Strata variables:

ADDITIONAL ROLES

11. On the **OPTIONS** tab, select the options to display **Cell** percentages and cumulative **Frequencies and percentage** and clear the box for **Chi-square statistics** under **STATISTICS**.

12. Submit the code.

DATA OPTIONS INFORMATION

► PLOTS

◀ FREQUENCY TABLE

- ▲ Frequencies
 - Observed
 - Expected
 - Deviation
- ▲ Percentages
 - Cell
 - Row
 - Column
- ▲ Cumulative
 - Column percentages
 - Frequencies and percentages
- ◀ STATISTICS
 - Chi-square statistics
 - Measures of association
 - Cochran-Mantel-Haenszel statistics

The generated SAS code is shown below.

```
ods noproctitle;
proc freq data=STAT1.AMESHOUSING3;
   tables (House_Style Overall_Qual Overall_Cond Year_Built
Fireplaces Mo_Sold
         Yr_Sold Garage_Type_2 Foundation_2 Heating_QC
Masonry_Veneer Lot_Shape_2
         Central_Air) / norow nocol plots(only)=(freqplot);
run;
```

Note: If you enter the code directly in the editor, the code below produces the necessary output.

```

/*st101d01.sas*/ /*Part B*/
/*PROC FREQ is used with categorical variables*/
ods graphics;

proc freq data=STAT1.ameshousing3;
  tables &categorical / plots=freqplot ;
  format House Style $House Style.
    Overall_Qual Overall.
    Overall Cond Overall.
    Heating_QC $Heating_QC.
    Central_Air $NoYes.
    Masonry_Veneer $NoYes.
  ;
  title "Categorical Variable Frequency Analysis";
run;

```

Selected ODS statement option:

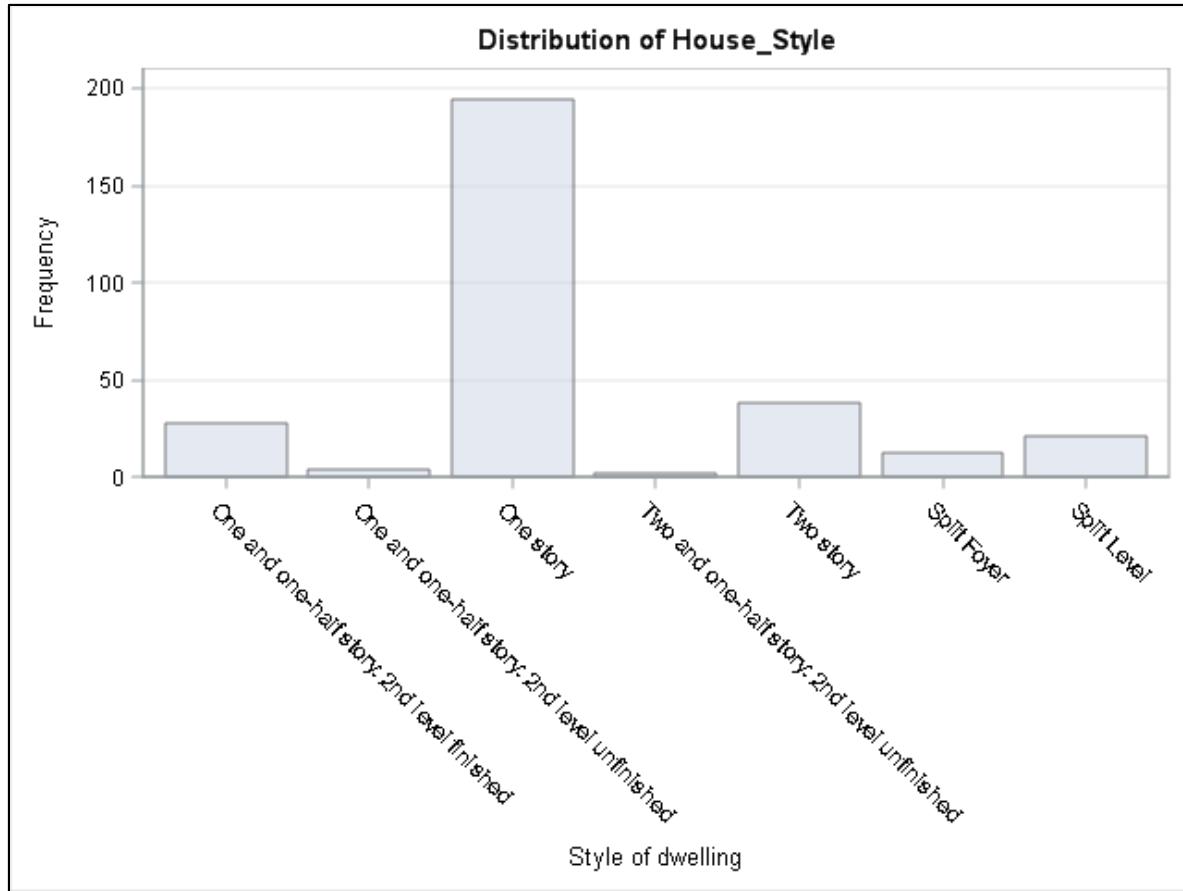
ODS GRAPHICS enables or disables ODS Statistical Graphics processing and sets graphics environment options. This statement affects ODS template-based graphics only. The ODS GRAPHICS statement does not affect device-based graphics.

Selected TABLES statement plot request:

PLOTS= FREQPLOT requests a frequency plot. Frequency plots are available for frequency and crosstabulation tables. For multiway crosstabulation tables, PROC FREQ provides a two-way frequency plot for each stratum (two-way table).

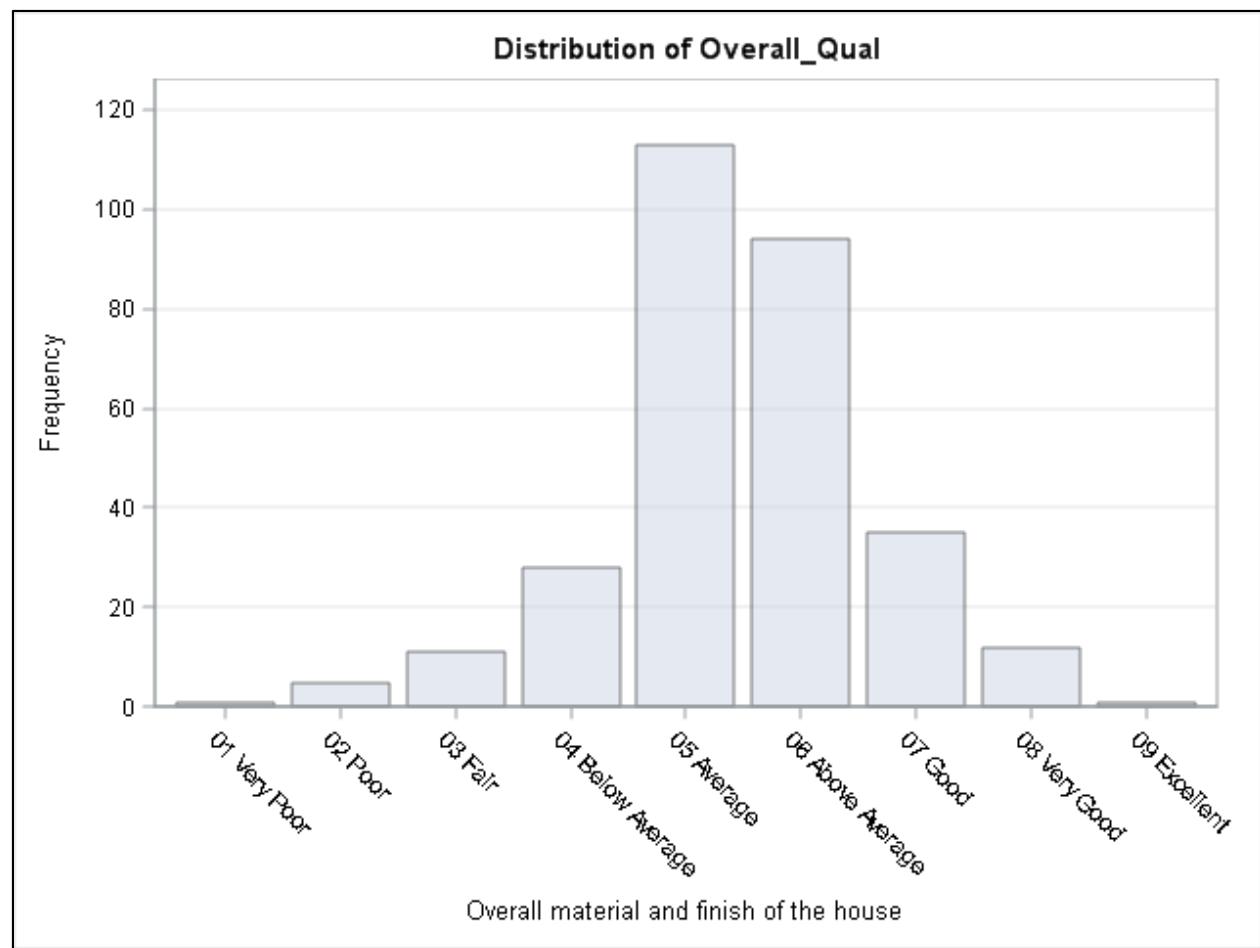
Partial Output

Style of dwelling					
House_Style	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
One and one-half story: 2nd level finished	28	9.33	28	9.33	
One and one-half story: 2nd level unfinished	4	1.33	32	10.67	
One story	194	64.67	226	75.33	
Two and one-half story: 2nd level unfinished	2	0.67	228	76.00	
Two story	38	12.67	266	88.67	
Split Foyer	13	4.33	279	93.00	
Split Level	21	7.00	300	100.00	

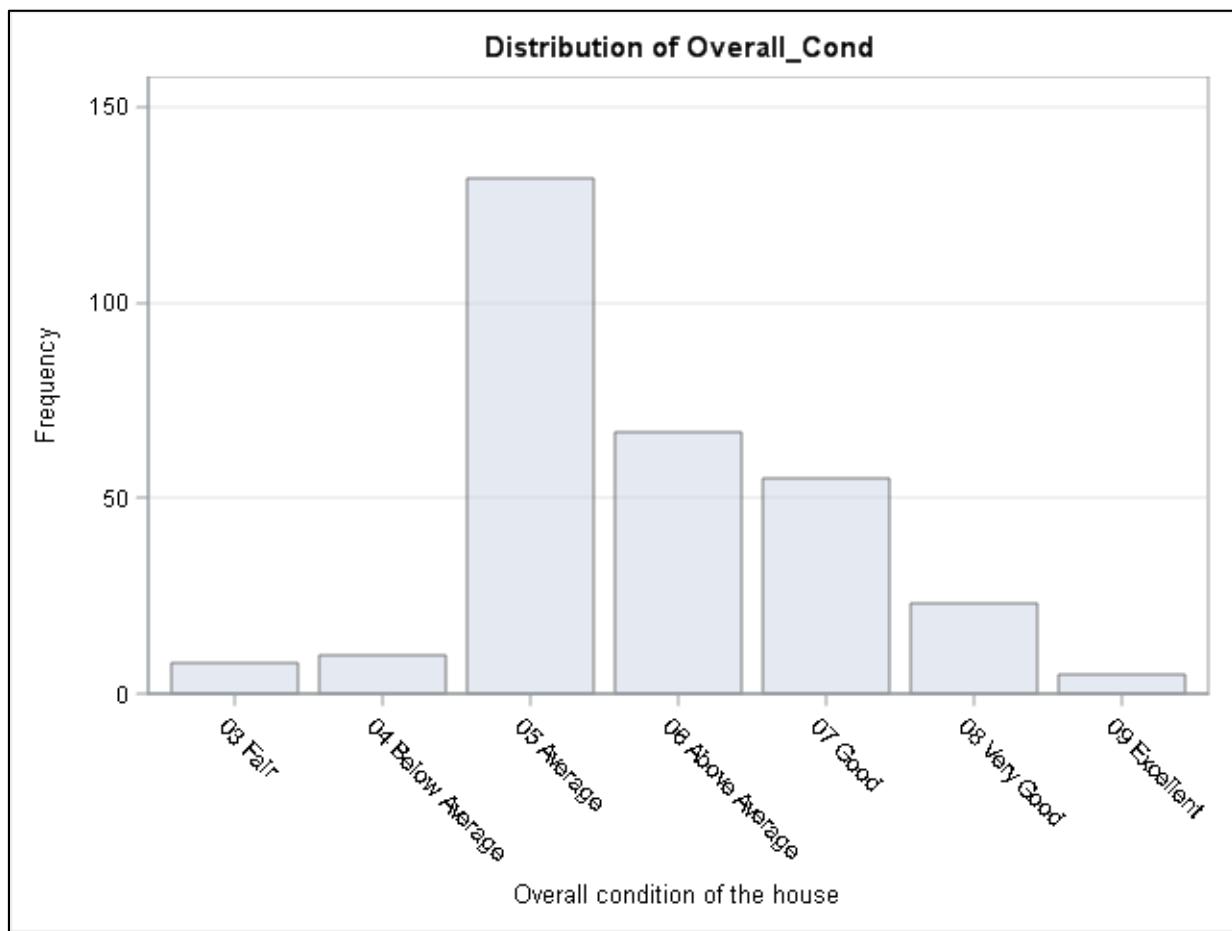


The categories with “2nd level unfinished” have too few members to analyze, so they will be merged with *One story* and *Two story* in the variable **House_Style2**.

Overall material and finish of the house					
Overall_Qual	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
01 Very Poor	1	0.33	1	0.33	
02 Poor	5	1.67	6	2.00	
03 Fair	11	3.67	17	5.67	
04 Below Average	28	9.33	45	15.00	
05 Average	113	37.67	158	52.67	
06 Above Average	94	31.33	252	84.00	
07 Good	35	11.67	287	95.67	
08 Very Good	12	4.00	299	99.67	
09 Excellent	1	0.33	300	100.00	

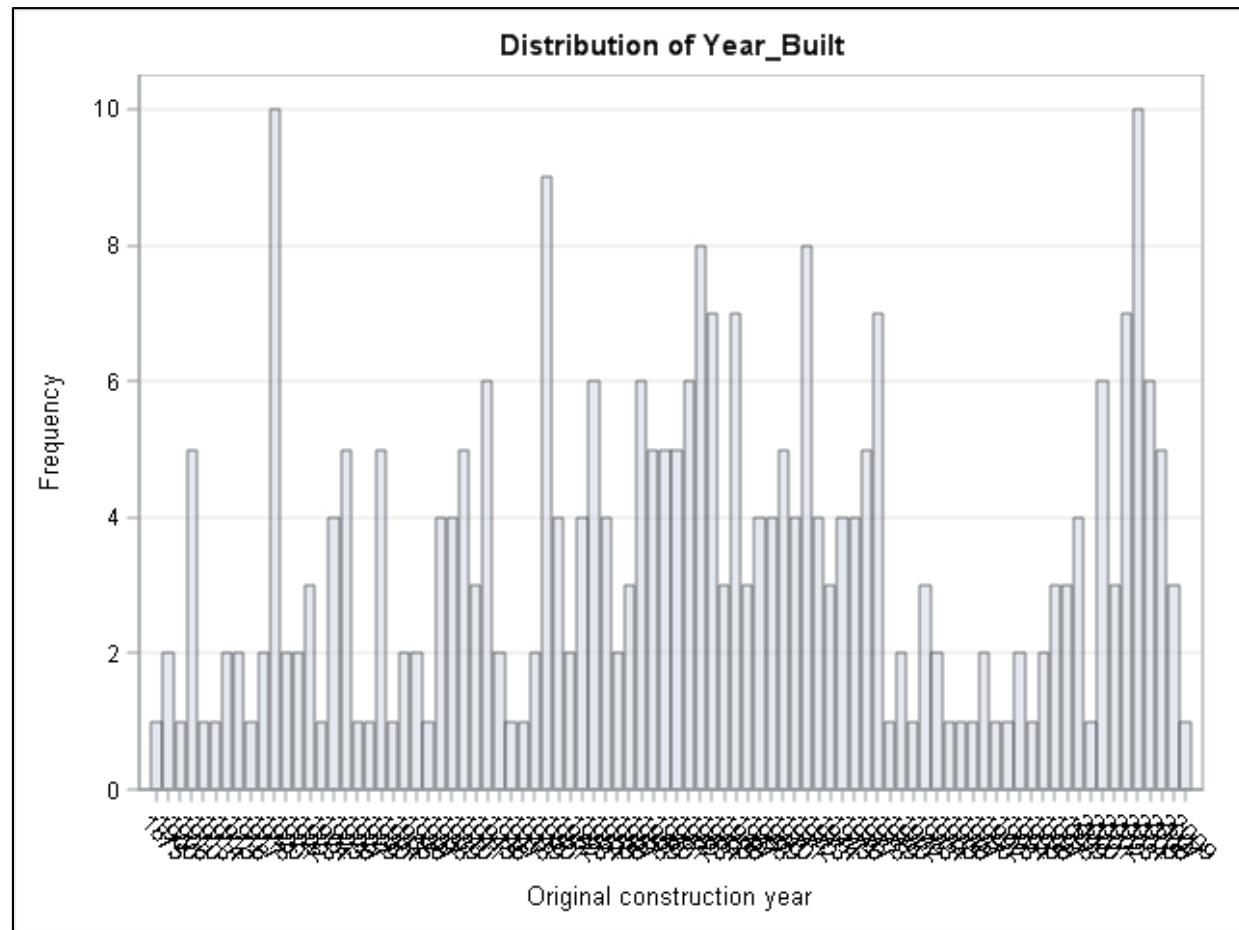


Overall condition of the house					
Overall_Cond	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
03 Fair	8	2.67	8	2.67	
04 Below Average	10	3.33	18	6.00	
05 Average	132	44.00	150	50.00	
06 Above Average	67	22.33	217	72.33	
07 Good	55	18.33	272	90.67	
08 Very Good	23	7.67	295	98.33	
09 Excellent	5	1.67	300	100.00	



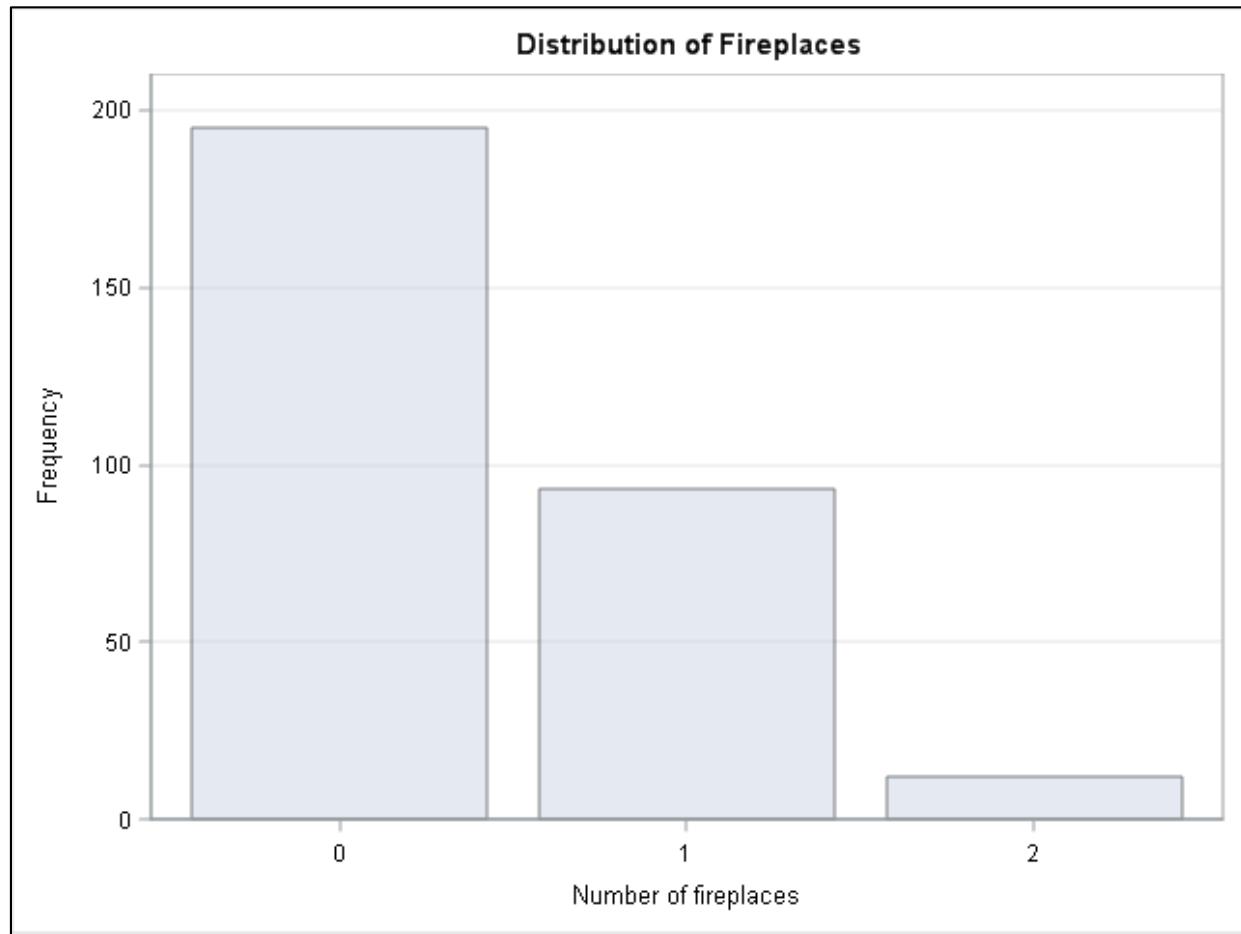
Overall_Qual and **Overall_Cond** have many levels with small frequencies. The variables will be trichotomized into **Below Average**, **Average**, and **Above Average**, in the variables **Overall_Qual2** and **Overall_Cond2**.

Original construction year					
Year_Built	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
1875	1	0.33	1	0.33	
1900	2	0.67	3	1.00	
1906	1	0.33	4	1.33	
1910	5	1.67	9	3.00	
1913	1	0.33	10	3.33	
....	
2004	10	3.33	285	95.00	
2005	6	2.00	291	97.00	
2006	5	1.67	296	98.67	
2007	3	1.00	299	99.67	
2009	1	0.33	300	100.00	



The construction year has more values than is practical to treat as a categorical variable in a statistical model with only 300 observations.

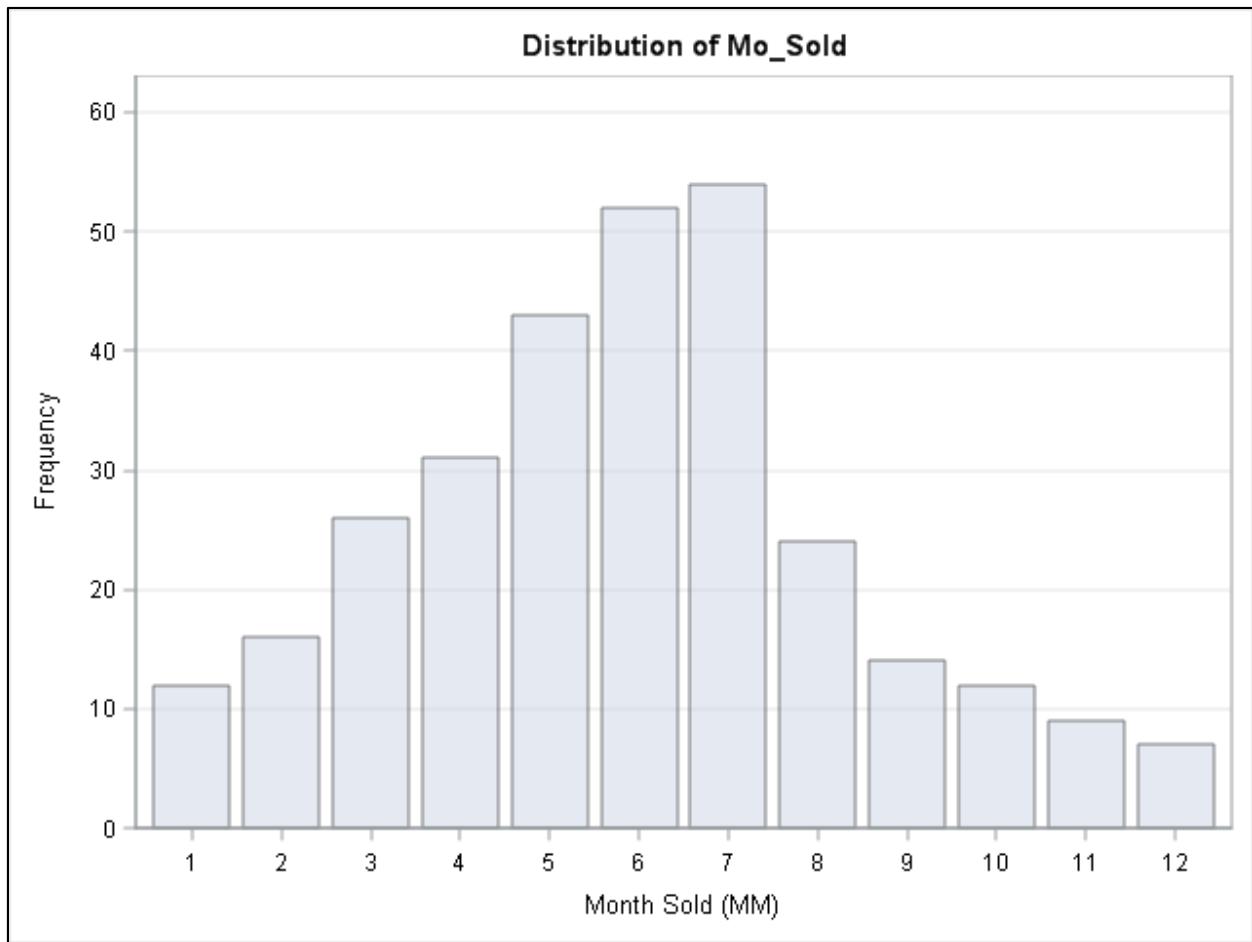
Number of fireplaces				
Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	195	65.00	195	65.00
1	93	31.00	288	96.00
2	12	4.00	300	100.00



Fireplaces can be treated as an ordinal variable in analyses.

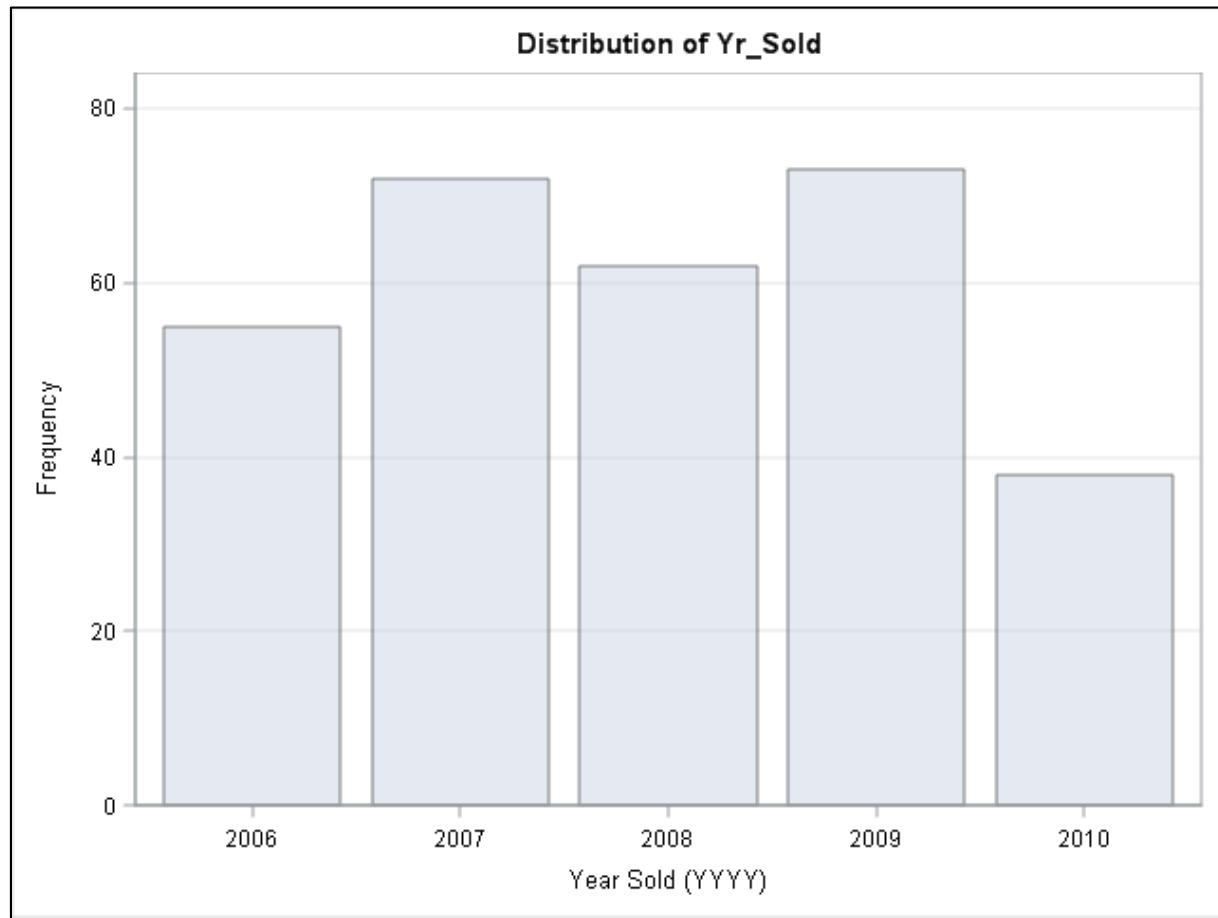
Month Sold (MM)				
Mo_Sold	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	12	4.00	12	4.00
2	16	5.33	28	9.33
3	26	8.67	54	18.00
4	31	10.33	85	28.33
5	43	14.33	128	42.67
6	52	17.33	180	60.00
7	54	18.00	234	78.00
8	24	8.00	258	86.00

Month Sold (MM)					
Mo_Sold	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
9	14	4.67	272	90.67	
10	12	4.00	284	94.67	
11	9	3.00	293	97.67	
12	7	2.33	300	100.00	



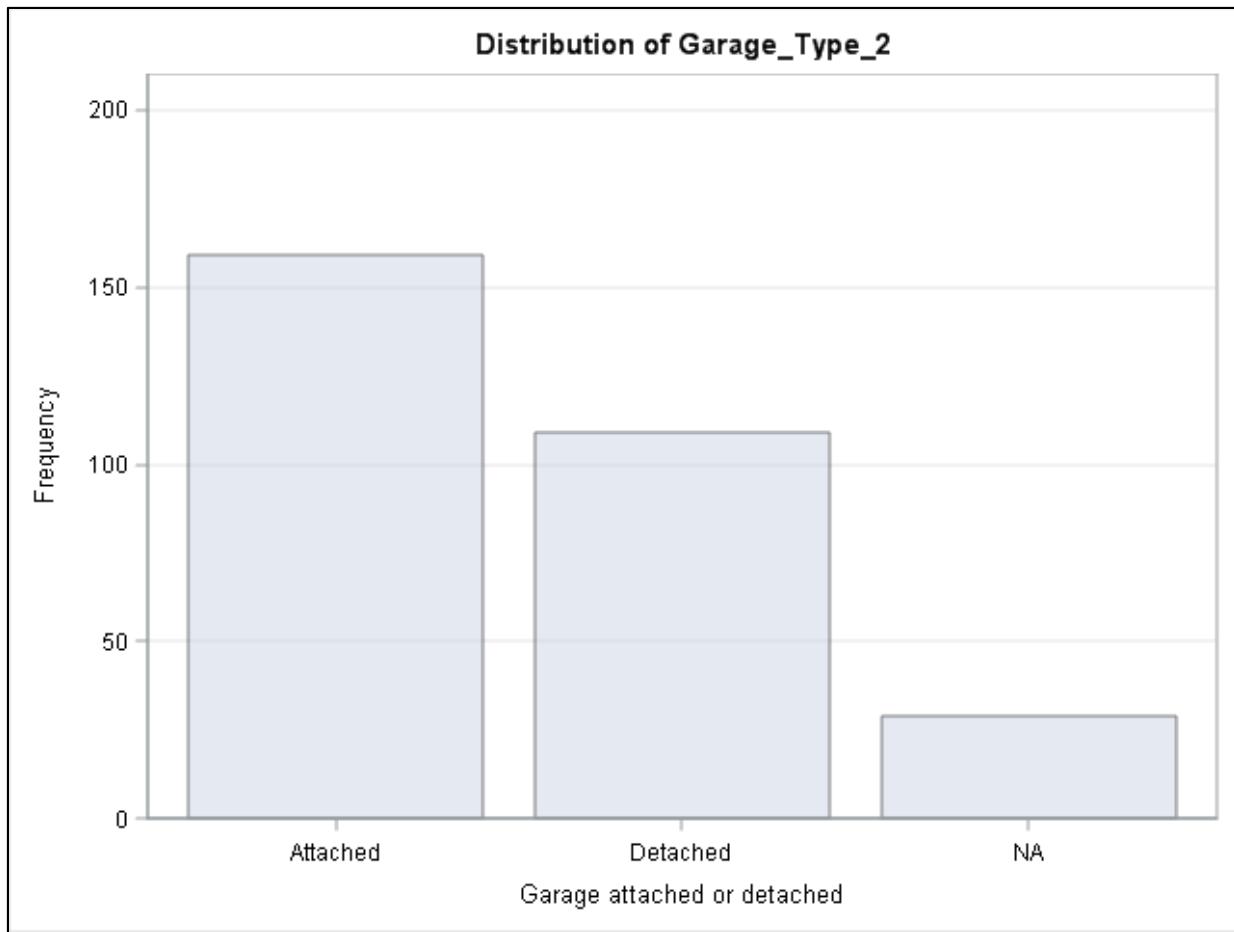
Mo_Sold shows a clear trend toward sales in July and June. Due to small numbers in some months, the variable **Season_Sold** was created and used for subsequent analyses. Season 1 is from month 12 to month 2; season 2 is from month 3 to month 5; season 3 is from month 6 to month 8; and season 4 is from month 9 to month 11.

Year Sold (YYYY)				
Yr_Sold	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2006	55	18.33	55	18.33
2007	72	24.00	127	42.33
2008	62	20.67	189	63.00
2009	73	24.33	262	87.33
2010	38	12.67	300	100.00



Garage attached or detached				
Garage_Type_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Attached	159	53.54	159	53.54
Detached	109	36.70	268	90.24
NA	29	9.76	297	100.00

Frequency Missing = 3



The value 'NA' was used for houses that had no garages. There are three missing values.

End of Demonstration

1.2 Quick Review of Statistical Concepts

Objective

- Define some common terminology related to hypothesis testing and confidence intervals.

17

Copyright © SAS Institute Inc. All rights reserved.



Population Parameters and Sample Statistics

 \bar{x}

estimates

 μ
 S

estimates

 σ

18

Copyright © SAS Institute Inc. All rights reserved.



In inferential statistics, the focus is on learning about *populations*. Examples of populations are all people with a certain disease, all drivers with a certain level of insurance, or all customers, both current and potential, at a bank.

Parameters are evaluations of characteristics of populations. They are generally unknown and must be estimated through the use of samples. A *sample* is a group of measurements from a population. In order for inferences to be valid, the sample should be representative of the population.

A *sample statistic* is a measurement from a sample. You infer information about population parameters through the use of sample statistics.

A *point estimate* is a single, best estimate of a population parameter.

Normal (Gaussian) Distribution

Proportions of the Normal Curve

Copyright © SAS Institute Inc. All rights reserved.

Sas

Because sampling involves variability, parameter estimates have variability. Often, the variability of sample statistics is approximately normal. Another name for the normal distribution is the *Gaussian* distribution. The normal distribution is bell-shaped, symmetric, and defined by two parameters, μ (the population mean) and σ (the population standard deviation). The mean locates the midpoint of the distribution. The standard deviation describes its spread.

The formula for a normal distribution of x around a mean, μ , with standard deviation, σ , is

$$f(x, \mu, \sigma) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

The standard normal curve has $\mu=0$ and $\sigma=1$. The area under the curve between any two values can be calculated. In statistics, think about probabilities related to the normal curve. Given the variability around the center (the mean, or point estimate of the parameter), you can think about the probability of sampling a value within some distance, $z\sigma$, from the mean. It is the area under the normal probability density curve in an area ranging from $-z\sigma$ to $z\sigma$. Some well-known values are shown in the slide.

Approximately 68% of the total area lies within 1 standard deviation of the mean. Approximately 95% of the total area lies within 1.96 standard deviations of the mean. Approximately 99.7% of the area lies within 3 standard deviations of the mean.

Standard Error of the Mean

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

It differs from the sample standard deviation in that:

- the sample standard deviation is a measure of the variability of data;
- the standard error of the mean is a measure of the variability of the sample mean.
- Standard error of the mean = $\frac{s}{\sqrt{n}} = S_{\bar{x}}$

20

Copyright © SAS Institute Inc. All rights reserved.



In statistics, assumptions are often made about distributions of parameters. A common one is that the sampling distribution of parameters is normal. This does not necessarily mean that the units of the population are normally distributed. It is often assumed that the parameter itself is normally distributed. Even though most statisticians only take one sample and get one point estimate for the population parameters, it is useful if they can assume normality of the parameter. That makes calculations of confidence intervals and *p*-values relatively easy. The variability of a parameter is measured by its standard error.

The standard error of the mean is computed as follows:

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

where

s is the sample standard deviation.

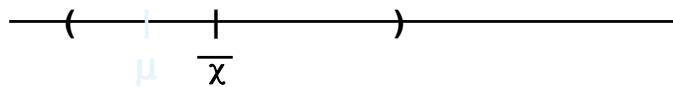
n is the sample size.

The standard error of the mean is a measure of precision of the parameter estimate. The smaller the standard error, the more precise your estimate.

Note: You can improve the precision of an estimate (reduce the standard error) by increasing the sample size.

Confidence Intervals

95% Confidence



- A 95% confidence interval represents a range of values within which you are 95% certain that the true population mean exists.
 - One interpretation is that if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

21

Copyright © SAS Institute Inc. All rights reserved.

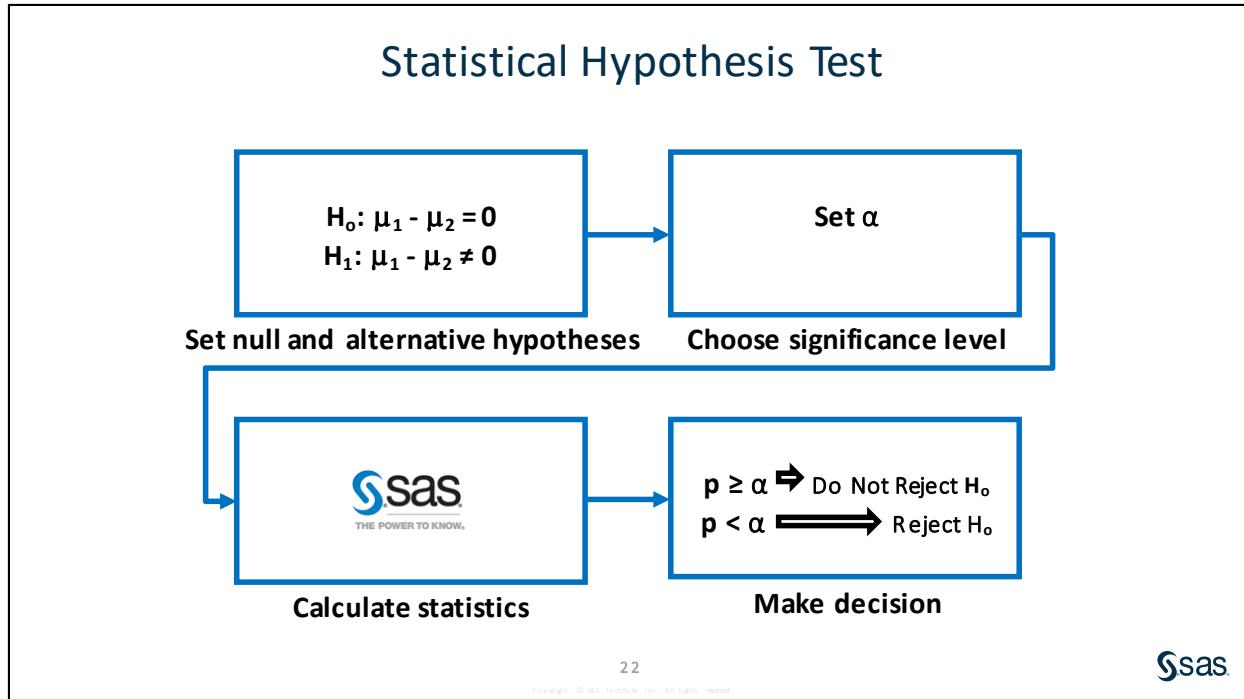
A confidence interval

- is a range of values that you believe is likely to contain the population parameter of interest
- is defined by an upper and lower bound around a parameter estimate.

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

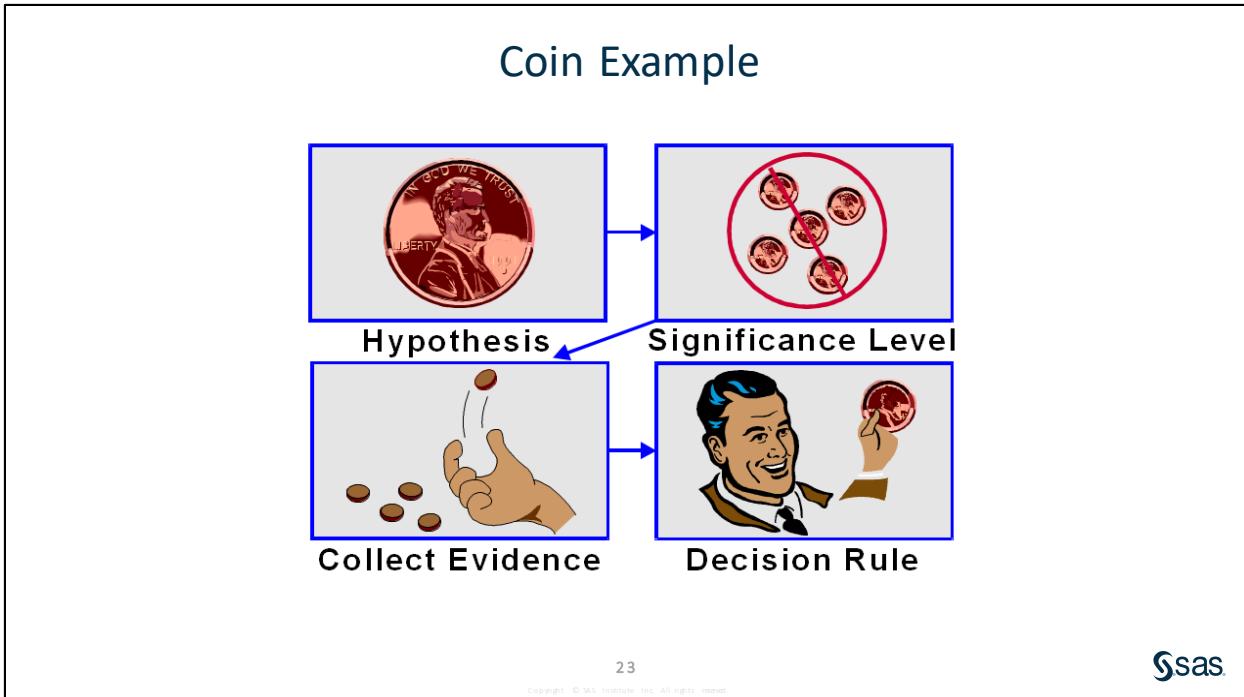
A 95% confidence interval is commonly used to assess the variability of the sample mean. In the Ames housing sales example, you interpret a 95% confidence interval by stating that you are 95% confident that the interval contains the mean sale price for your population of home sales.

Note: You want to be as confident as possible, but remember that if you increase the confidence level too much, the width of your interval increases beyond the point where it is informative. For example, a 100% confidence interval would have confidence bounds of negative and positive infinity.



In inferential statistics, you infer information about population parameters through the use of statistics. The inferences are not exact. As you have seen, there is variability of parameter estimates. You phrase questions as tests of hypotheses about population parameters. The answers are typically phrased as a probability that a specific statement about the parameter is true, given the evidence provided by the data. That statement is called the *null hypothesis*. The probability calculated from the data is called the *p-value*.

When the *p*-value is low, it provides doubt about the truth of the null hypothesis. How low does the *p*-value need to be before you reject the null hypothesis completely? That depends on you. That threshold that you choose is called the *significance level* of your test.



Test whether a coin is fair.

1. You suspect that the coin is **not** fair, but begin by assuming that the coin is fair. In other words, you assume the null hypothesis to be true.
2. You select a significance level: if you observe five heads in a row or five tails in a row, you conclude that the coin is not fair. Otherwise, you decide that there is not enough evidence to show that the coin is not fair. With a fair coin (a true null hypothesis), the probability of observing 5 heads or tails in a row in five trials is $2*(\frac{1}{2})^5=1/16$. In other words, the significance level is 1/16, or 0.0625.
3. In order to collect evidence, 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 either that there is
 - enough evidence to reject the assumption that the coin is fair (either all trials are heads or all trials are tails), or
 - not enough evidence to reject the assumption that the coin is fair (not all trials are either heads or tails).

What Is an Alpha Level?

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 $p(\text{Type II} H_1) = \beta$	Type II Error $p(\text{Type II} H_1) = \beta$
Reject Null	Type I Error $p(\text{Type I} H_0) = \alpha$	Correct $(1 - \beta) = \text{Power}$

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. This is the *significance level* of the hypothesis test.

- In the coin example, it is the probability that you conclude that the coin is not fair when it is fair.

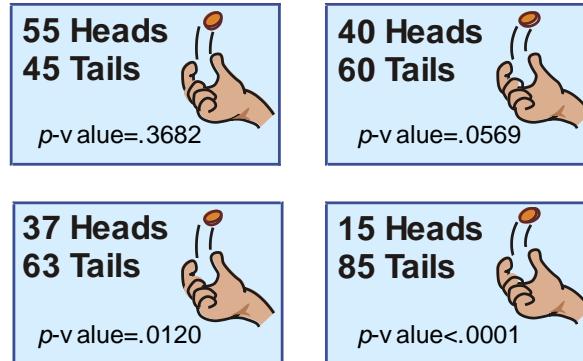
The probability of a Type II error, often denoted β , is the probability that you fail to reject the null hypothesis when it is false.

- In the coin example, it is the probability that you fail to find that the coin is not fair when it is not fair.

Note: 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, given some assumed values of the true population mean and standard deviation in the population and the sample size.

p-Value – Effect Size Influence

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



25

The *effect size* refers to the magnitude of the difference in sampled population from the null hypothesis. In this example, the null hypothesis of a fair coin suggests 50% heads and 50% tails. If the true coin flipped were actually weighted to give 55% heads, the effect size would be 5%.

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, as the difference between the number of heads and tails increases, you have more evidence that the coin is not fair.

A *p*-value measures the probability of observing a value as extreme as or more extreme than the one observed, simply by chance, given that the null hypothesis is true. 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.

A large *p*-value means that you would often see a test statistic value this large in experiments with a fair coin. A small *p*-value means that you would rarely see differences this large from a fair coin. In the latter situation, you have evidence that the coin is not fair, because if the null hypothesis were true, a random sample selected from it would not likely have the observed statistic values.

p-Value – Sample Size Influence

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



A *p*-value is not only affected by the effect size. 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 be 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, when the number of trials (k) on which the proportion is based increases. As you saw in the section about confidence intervals, the variability around a mean estimate is smaller, when the sample size is larger. For larger sample sizes, you can measure means more precisely. Therefore, 40% of the heads out of 400 flips would make you more certain that this was not a chance difference from 50% than would 40% out of 10 flips. The smaller *p*-values reflect this confidence. The *p*-value here assesses the probability that this difference from 50% occurred purely by chance.

1.02 Multiple Choice Poll

Which of the following affects alpha?

- a. The p -value of the test
- b. The sample size
- c. The number of Type I errors
- d. All of the above
- e. Answers a and b only
- f. None of the above

1.3 One-Sample t -Tests

Objective

- Perform a hypothesis test using the TTEST procedure.

Performing a *t*-Test

To test the null hypothesis $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$, SAS software calculates the value of student's *t* statistic:

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

For the Ames homes sales price example:

$$t = \frac{137,525 - 135,000}{2,172.1} = 1.16$$

The null hypothesis is rejected when the calculated value is more extreme (either positive or negative) than would be expected by chance if H_0 were true.


Copyright © SAS Institute Inc. All rights reserved.
31

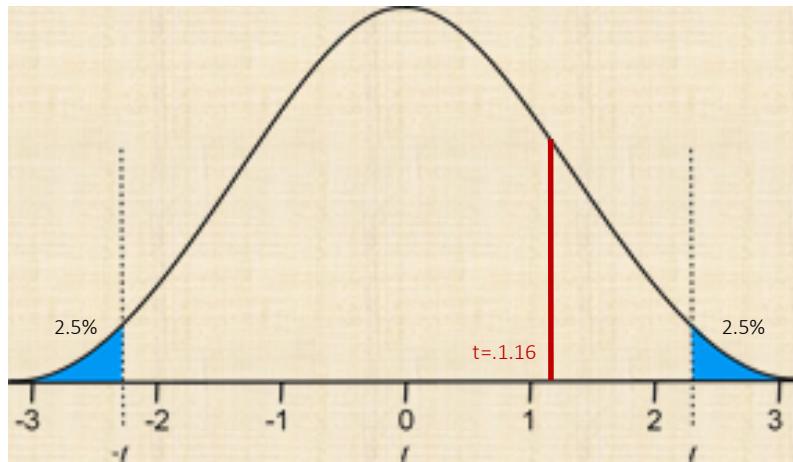
As mentioned in a previous section, when you do not know the true population standard deviation, σ , then you must estimate it from the sample. Then you must also use student's *t*-distribution, rather than the normal distribution, for calculating *p*-values and confidence limits. Student's *t*-distribution approaches the normal distribution as sample size increases.

A one-sample *t*-test compares the mean calculated from a sample to a hypothesized mean. The null hypothesis of the test is generally that the difference between the two means is zero.

For the example, suppose that you would like to know whether the mean sale price for houses in Ames, Iowa, is \$135,000. μ_0 is the hypothesized value of 135,000, \bar{x} is the sample mean of **SalePrice**, and $s_{\bar{x}}$ is the standard error of the mean.

- The student's *t* statistic measures how far \bar{x} is from the null hypothesized mean, in standard error units.
- 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 normal.

Rejection Region for Two-Sided Test



32

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.

Note: The alpha and t -distribution mentioned here are the same as those in the section about 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 ends.

The TTEST Procedure

General form of the TTEST procedure:

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

33

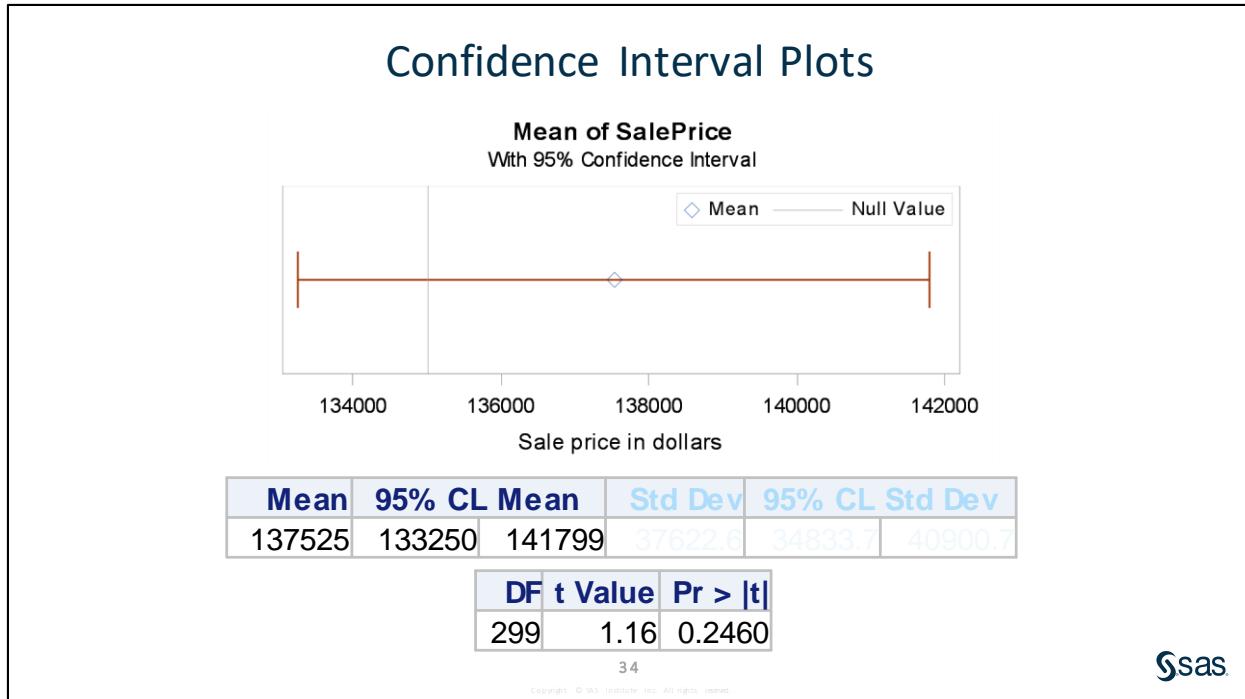
Copyright © SAS Institute Inc. All rights reserved.



The TTEST procedure performs *t*-tests and computes confidence limits for one sample, paired observations, two independent samples, and the AB/BA crossover design. With ODS Statistical Graphics, PROC TTEST can also be used to produce histograms, Quantile-Quantile plots, box plots, and confidence limit plots.

Selected TTEST procedure statements:

- | | |
|---------------------------|--|
| CLASS | specifies the two-level variable for the analysis. Only one variable is allowed in the CLASS statement. If no CLASS statement is included, a one-sample <i>t</i> -test is performed. |
| PAIRED <i>PairLists</i> ; | specifies the <i>PairLists</i> to identify the variables to be compared in paired comparisons. You can use one or more PairLists. |
| VAR | specifies <i>numeric</i> 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. |



A *confidence interval plot* is a visual display of the sample statistic value (of the mean, in this case) and the confidence interval calculated from the data. If there is a null hypothesized value for the parameter, it can be drawn on the plot as a reference line. In this way, the statistical significance of a test can be visually assessed. If the $(1.00-\alpha)\%$ confidence interval does not include the null hypothesis value, then that implies that the null hypothesis can be rejected at the α significance level. If the confidence interval includes the null hypothesis value, then that implies that the null hypothesis cannot be rejected at that significance level.



PROC TTEST for a One-Sample *t*-Test

Example: Use the TTEST procedure to test whether the mean of **SalePrice** is \$135,000 in the data set **STAT1.AmesHousing3**.

1. Open the **t Tests** task under **Statistics** to conduct a one-sample *t* test.
2. Assign **SalePrice** as the Analysis variable.

The screenshot shows the SAS Studio interface. On the left, the 'Tasks and Utilities' sidebar is open, with 't Tests' selected under the 'Statistics' section. In the main workspace, the 'st101d01b' tab is active. The 'DATA' tab is selected in the 't test' configuration pane. The 'Analysis variable' dropdown is set to 'SalePrice'. The 't test:' dropdown is set to 'One-sample test'.

3. On the OPTIONS tab, set the **Alternative hypothesis: mu^=** to 135,000.
4. Uncheck the option to conduct **Tests for normality**.
5. To select plots, expand the **PLOTS** property and use the drop-down menu to choose the option **Selected plots**. Select the option to display **Confidence interval plot** in addition to the default plots of **histogram and box plot** and **normality plot**.
6. Run the code.

The screenshot shows the SAS Test Selection dialog box. The 'OPTIONS' tab is selected. Under the 'TESTS' section, 'Tails:' is set to 'Two-tailed test'. The 'Alternative hypothesis: mu ^=' field contains '135,000'. Under the 'PLOTS' section, 'Plots:' is set to 'Selected plots'. Three checkboxes are checked: 'Histogram and box plot', 'Normality plot', and 'Confidence interval plot'.

The generated code is as follows:

```
/** t Test **/
proc ttest data=STAT1.AMESHOUSING3 sides=2 h0=135000
            plots(only showh0)=(summaryPlot intervalPlot qqplot);
    var SalePrice;
run;
```

Note: The equivalent SAS programming code is as follows.

```
/*st101d02.sas*/
ods graphics;

proc ttest data=STAT1.ameshousing3
            plots(shownull)=interval
            H0=135000;
    var SalePrice;
    title "One-Sample t-test testing whether mean SalePrice=$135,000";
run;
```

Selected PROC TTEST statement option:

PLOTS(SHOWNULL) The PLOTS option ONLY suppresses the default plots. INTERVAL requests plots of confidence interval for means. The SHOWNULL option places a vertical reference line at the null hypothesis value chosen with the H0 option.

H0=m Requests tests against a null value of m.

One-Sample t-test testing whether SalePrice=\$135,000

The TTEST Procedure

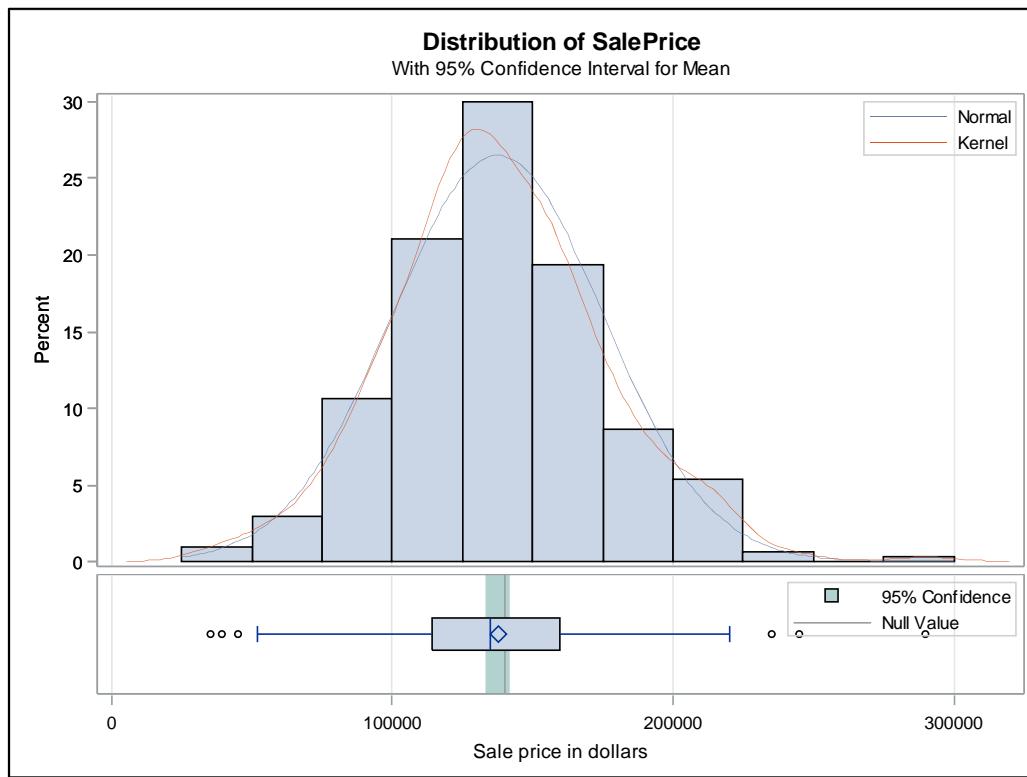
Variable: SalePrice (Sale price in dollars)

N	Mean	Std Dev	Std Err	Minimum	Maximum
300	137525	37622.6	2172.1	35000.0	290000

Mean	95% CL Mean	Std Dev	95% CL Std Dev
137525	133250	141799	37622.6

DF	t Value	Pr > t
299	1.16	0.2460

The mean value is \$137,525. The t -value associated with that is 1.16. The p -value is 0.2460. Therefore, you would reach the conclusion that the mean sale price of homes is not statistically different from \$135,000.





The confidence interval plot shows the confidence interval around the mean estimate of sale price. Its intersection with the \$135,000 reference line shows that the mean value in the sample is not statistically significantly different from \$135,000 at an alpha level of 0.05.

Note: The confidence bounds can be changed using an ALPHA= option in the PROC TTEST statement. Set alpha equal to 1-confidence. For example, for a 99% confidence interval, specify “ALPHA=0.01”.



Neither the histogram nor the q-q plot show extreme departures from normality. Therefore, the Student's *t*-test is valid.

End of Demonstration

1.03 Quiz

What is the null hypothesis for a one-sample t -test?

- a. $H_0: \mu = \mu_0$
- b. $H_0: \mu_0 = 0$
- c. $H_0: \mu - \mu_0 = 0$
- d. $H_0: \mu_0 - 0 = 0$



Exercises

1. Performing a One-Sample *t*-Test

The data in **STAT1.NormTemp** come from an article in the *Journal of Statistics Education* by Dr. Allen L. Shoemaker from the Psychology Department at Calvin College. The data are based on an article in a 1992 edition of *JAMA (Journal of the American Medical Association)*, which questions the notion that the true mean body temperature is 98.6. There are 65 males and 65 females. There is also some question about whether mean body temperatures for women are the same as for men. The variables in the data set are as follows:

ID Identification number

BodyTemp Body temperature (degrees Fahrenheit)

Gender Coded (**Male**, **Female**)

HeartRate Heart rate (beats per minute)

- a. Look at the distribution of the continuous variables in the data set using PROC UNIVARIATE, including producing histograms and insets with means, standard deviations and sample size.
- b. Perform a one-sample *t*-test to determine whether the mean of body temperatures (the variable **BodyTemp** in **STAT1.NormTemp**) is 98.6. Produce a confidence interval plot of **BodyTemp** with the value 98.6 used as a reference.
 - 1) What is the value of the *t* statistic and the corresponding *p*-value?
 - 2) Do you reject or fail to reject the null hypothesis at the 0.05 level that the average temperature is 98.6 degrees?

End of Exercises

1.4 Two-Sample t-Tests

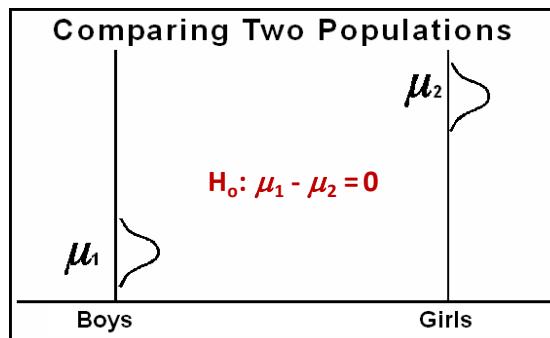
Objectives

- Use the TTEST procedure to analyze the differences between two population means.
- Verify the assumptions of a two-sample *t*-test.

41

Sas

Comparing Two Population Means



- Statistical Assumptions:
 - independent observations
 - normally distributed population means
 - equal population variances

42

Sas

In a one-sample *t*-test, the sample's mean is compared against some hypothesized mean value. For example, in the previous example, the mean sale price was compared against \$140,000. So, the null hypothesis is $\mu_1=140,000$.

If you want to compare the means of two different groups, you can specify the hypothesis in either or two ways: $\mu_1=\mu_2$; or $\mu_1-\mu_2=0$.

Before you start the analysis, examine the data to verify that the statistical assumptions are valid.

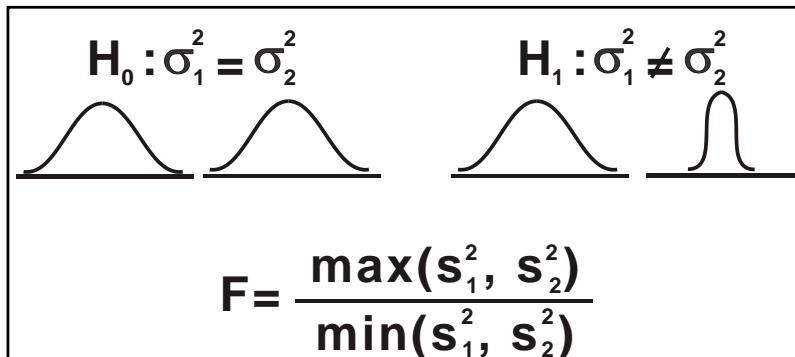
The assumption of independent observations means that no observations provide any information about any other observation that you collect. For example, measurements are not repeated on the same subject. This assumption can be verified during the design stage.

The assumption of normality can be verified if the data are approximately normally distributed or if enough data are collected. For small samples, 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.

Folded F Test for Equality of Variances



To evaluate the assumption of equal variances in each group, you can use the Folded *F* test for equality of variances. The null hypothesis for this test is that the variances are equal. The *F* value is calculated as a ratio of the greater of the two variances divided by the lesser of the two. Thus, if the null hypothesis is true, *F* tends to be close to 1.0 and the *p*-value for *F* is statistically nonsignificant ($p>0.05$).

If you reject the null hypothesis, it is recommended that you use the unequal variance *t*-test in the PROC TTEST output for testing the equality of group means.

Note: 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 use Levene's test or the Brown-Forsythe test for homogeneity of variances. These are available as options HOVTEST=LEVENE and HOVTEST=BF in the MEANS statement in the GLM procedure.

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 ②

Unequal Variance *t* Test (Satterthwaite):

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

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

Equality of Variances Test (Folded *F*):_____

$F' = 1.51$ DF = (3,3) Prob > F' = 0.7446 ①

Copyright © SAS Institute Inc. All rights reserved.



- ① Check the assumption of 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.
- ② Therefore, 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.

Note: The equality of variances *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):

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

Unequal Variance *t* Test (Satterthwaite):

$T = -2.4518 \quad DF = 11.1 \quad Prob > |T| = 0.0320 \quad \textcircled{2} \leftarrow$

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

Equality of Variances Test (Folded *F*):_____

$F' = 15.28 \quad DF = (9,4) \quad Prob > F' = 0.0185 \quad \textcircled{1}$

Copyright © SAS Institute Inc. All rights reserved.

Sas

- ① Again, check the assumption of equal variances and then 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.
- ② Therefore, 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 0.05 level.

Note: If you choose the equal variance *t*-test, you would **not** reject the null hypothesis at the 0.05 level. This shows the importance of choosing the appropriate *t*-test.



Two-Sample t-Test

Example: Use the TTEST procedure to test whether the mean of **SalePrice** is the same for homes with masonry veneer and those without.

1. Create a new **t Tests** task in SAS Studio.
2. On the DATA tab, use the drop-down menu for **t test** under **ROLES** and select **Two-sample test**.
3. Assign **SalePrice** as the dependent variable and add **Masonry_Veneer**, the group identification variable, to the **Groups variable** field.

The screenshot shows the SAS Studio interface with the 't test' configuration. The 'DATA' tab is selected. Under 'DATA', the dataset 'STAT1.AMESHOUSING3' is chosen. Under 'ROLES', the 't test' dropdown is set to 'Two-sample test'. A 'Groups variable' field contains 'Masonry_Veneer'. The 'OPTIONS' tab is not visible in this screenshot.

4. On the OPTIONS tab, uncheck the option to conduct **Tests for normality**.
5. Select the option to display a **Confidence interval plot**.
6. Submit the code.

The generated SAS syntax is shown below.

```
ods noproctitle;
ods graphics / imagemap=on;

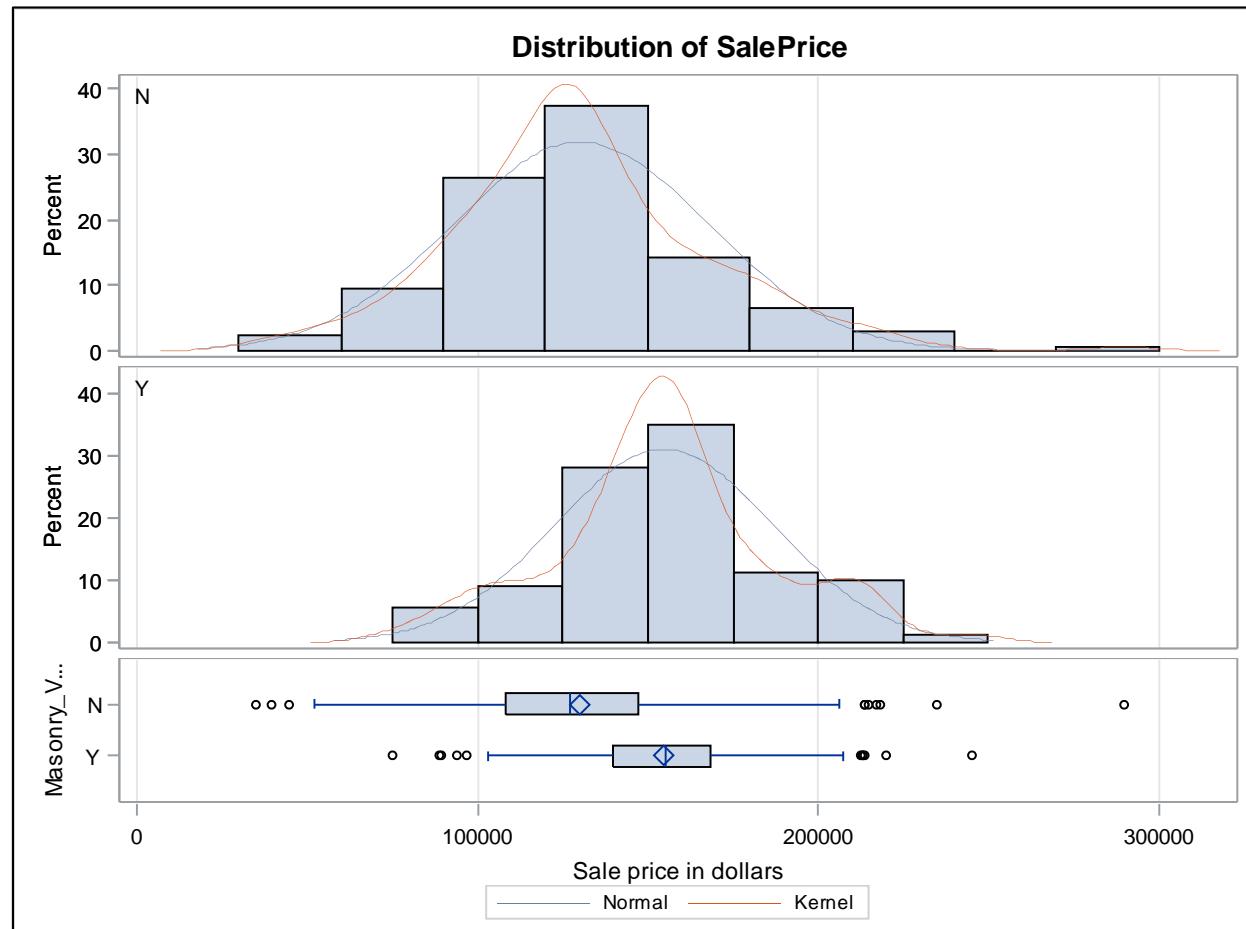
/*** t Test ***/
proc ttest data=STAT1.AMESHOUSING3 sides=2 h0=0
            plots(only showh0)=(summaryPlot intervalPlot qqplot);
    class Masonry_Veneer;
    var SalePrice;
run;
```

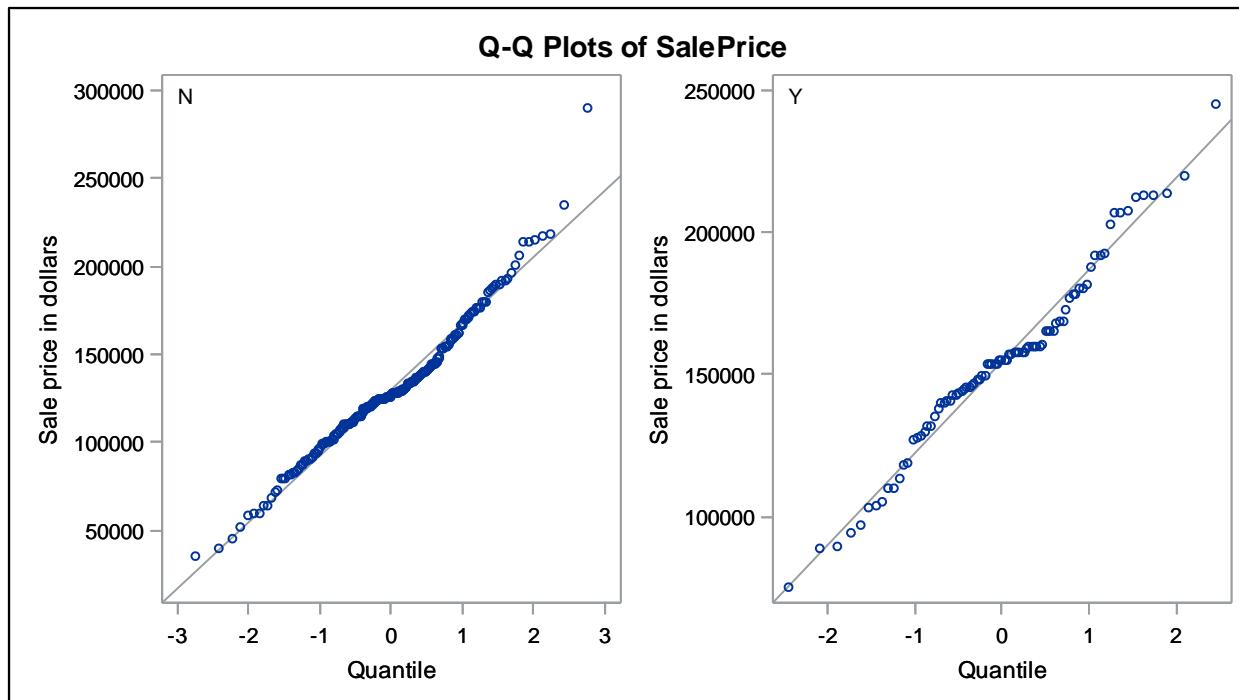
Note: If you produce the code by entering it directly in the editor, the code below produces the necessary output.

```
/*st101d03.sas*/
ods graphics;

proc ttest data=STAT1.ameshousing3 plots(shownull)=interval;
  class Masonry_Veneer;
  var SalePrice;
  format Masonry Veneer $NoYes.;
  title "Two-Sample t-test Comparing Masonry Veneer, No vs. Yes";
run;
```

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 Q-Q plot.





The Q-Q plots seem to indicate that the data from each group approximate a normal distribution. There seems to be one potential outlier in each group at the upper end of the distribution.

Note: If assumptions are not met, you can do an equivalent nonparametric test, which does not make distributional assumptions. PROC NPAR1WAY is one procedure for performing this type of test. It is described in an appendix.

The statistical tables for the TTEST procedure are displayed below.

❶		N	Mean	Std Dev	Std Err	Minimum	Maximum
	Masonry_Veneer	209	130172	37531.7	2596.1	35000.0	290000
	N	209	130172	37531.7	2596.1	35000.0	290000
	Y	89	154705	32239.8	3417.4	75000.0	245000
	Diff (1-2)		-24533.0	36039.6	4561.6		

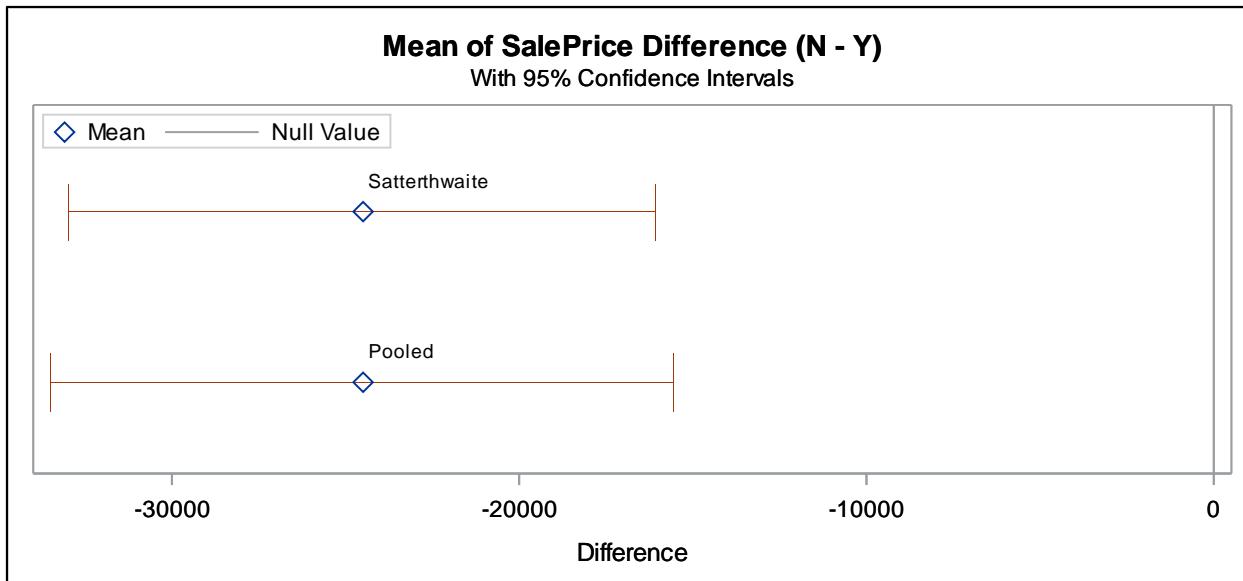
Masonry_Veneer	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
N		130172	125054	37531.7	34245.4
Y		154705	147914	32239.8	28099.7
Diff (1-2) ❸	Pooled	-24533.0	-33510.3	-15555.6	33355.6
Diff (1-2)	Satterthwaite	-24533.0	-32997.9	-16068.0	39197.1

Method	Variances	DF	t Value	Pr > t
Pooled ❸	Equal	296	-5.38	<.0001
Satterthwaite	Unequal	191.85	-5.72	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	208	88	1.36	0.1039

- ① In the Statistics table, examine the descriptive statistics for each group and their differences.
- 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 0.1039. Because this value is greater than the alpha level of 0.05, do not reject the null hypothesis of equal variances (This is equivalent to saying 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 reject the null hypothesis that the group means are equal. The mean difference between no masonry veneer and masonry veneer is -\$24,533. Because the p -value is less than 0.05 ($\text{Pr} > |t| < .0001$), you conclude that there is a statistically significant difference in the sale price between houses with the two types of veneer.

Note: The 95% confidence interval for the mean difference (-33510.3, -15555.6) does not include 0. This also implies statistical significance at the 0.05 alpha level.



Confidence intervals are shown in the output object titled Mean of SalePrice Difference (N – Y). This plot reflects the values from the confidence interval for the mean differences.

End of Demonstration



Exercises

2. Using PROC TTEST for Comparing Groups

Elli Sagerman, a Masters of Education candidate in German Education at the University of North Carolina at Chapel Hill in 2000, collected data for a study. She looked at the effectiveness of a new type of foreign language teaching technique on grammar skills. She selected 30 students to receive tutoring; 15 received the new type of training during the tutorials and 15 received standard tutoring. Two students moved away from the district before completing the study. Scores on a standardized German grammar test were recorded immediately before the 12-week tutorials and then again 12 weeks later at the end of the trial. Sagerman wanted to see the effect of the new technique on grammar skills. The data are in the **STAT1.GERMAN** data set.

Change Change in grammar test scores

Group The assigned treatment, coded **Treatment** and **Control**

Analyze the data using PROC TTEST. Assess whether the treatment group improved more than the control group.

- a. Do the two groups appear to be approximately normally distributed?
- b. Do the two groups have approximately equal variances?
- c. Does the new teaching technique seem to result in significantly different change scores compared with the standard technique?

End of Exercises

1.04 Multiple Answer Poll

How do you tell PROC TTEST that you want to do a two-sample *t*-test?

- a. SAMPLE=2 option
- b. CLASS statement
- c. GROUPS=2 option
- d. PAIRED statement

1.5 Solutions

Solutions to Exercises

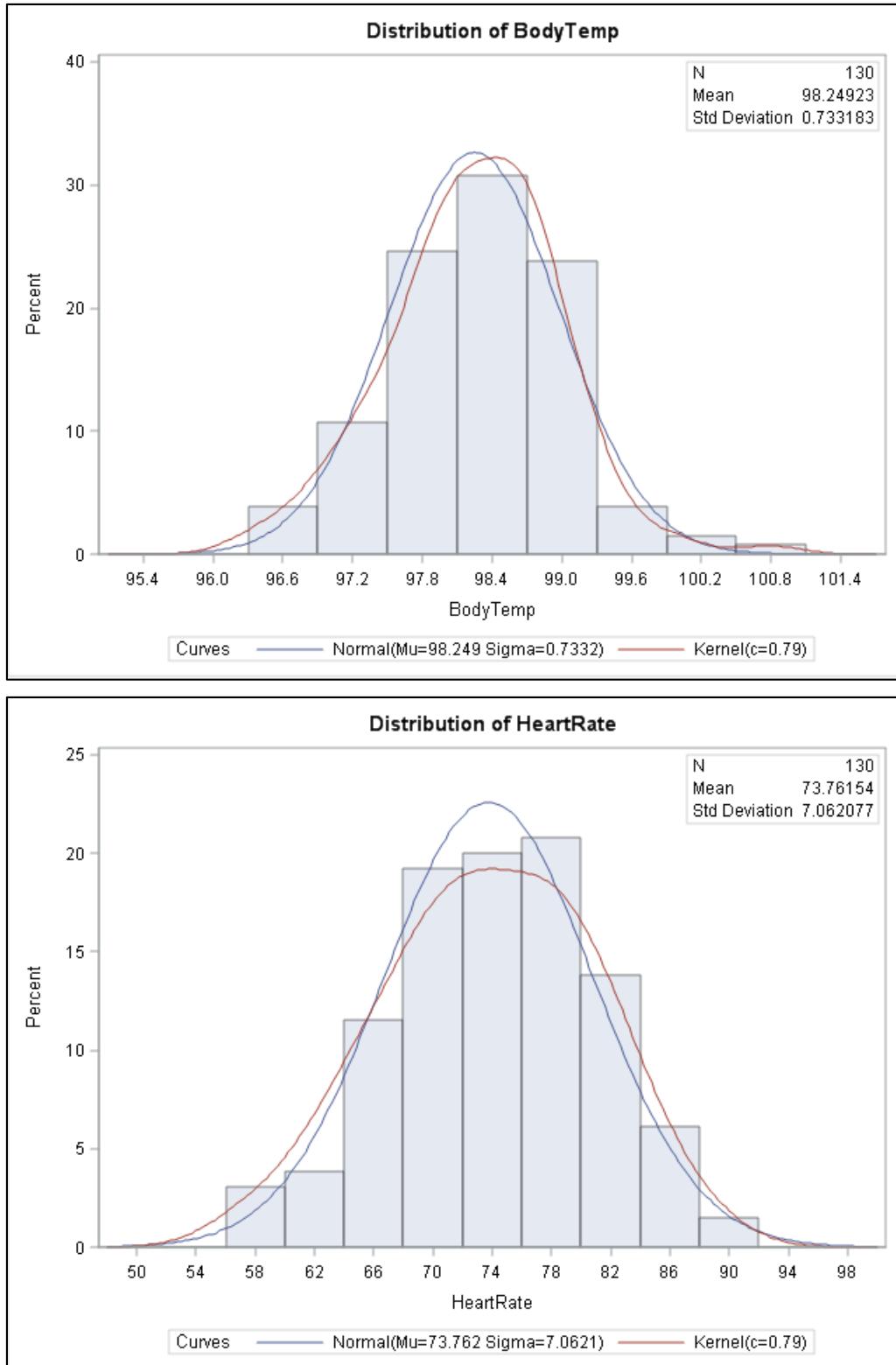
1. Performing a One-Sample *t*-Test

- a. Look at the distribution of the continuous variables in the data set using PROC UNIVARIATE, including producing histograms and insets with means, standard deviations and sample size.
 - 1) Open the Distribution Analysis task.
 - 2) On the DATA tab, do the following:
 - Select the **NORMTEMP** data set.
 - Set the continuous variables, **BodyTemp** and **HeartRate**, as the Analysis variables.
 - 3) On the OPTIONS tab, do the following:
 - Select the options to add normal curve, kernel density estimate, and inset statistics to the histogram.
 - Expand the Inset Statistics tab and select Number of observations, Mean, and Standard deviation.
 - 4) Run the code.

Note: Alternatively, write the SAS programming code directly.

```
/*st101s01.sas*/ /*Part A*/
%let interval=BodyTemp HeartRate;

ods graphics;
ods select histogram;
proc univariate data=STAT1.NormTemp noprint;
  var &interval;
  histogram &interval / normal kernel;
  inset n mean std / position=ne;
  title "Interval Variable Distribution Analysis";
run;
```



- b. Perform a one-sample *t*-test to determine whether the mean of body temperatures (the variable **BodyTemp** in **STAT1.NormTemp**) is 98.6. Produce a confidence interval plot of **BodyTemp** with the value 98.6 used as a reference.

- 1) Open the t Tests task.
- 2) On the DATA tab, do the following:
 - Select **NORMTEMP** as the data.
 - Set **BodyTemp** as the **Analysis variables**.
- 3) On the OPTIONS tab, do the following:
 - Specify the H_0 value for the Alternative hypothesis as **98.6**.
 - Under **PLOTS** choose the Selected plots option, and check to display only the **Confidence interval plot**.
- 4) Run the code.

Note: Alternatively, write the SAS Programming code directly.

```
/*st101s01.sas*/ /*Part B*/
proc ttest data=STAT1.NormTemp h0=98.6
            plots(only shownull)=interval;
  var BodyTemp;
  title 'Testing Whether the Mean Body Temperature=98.6';
run;
title;
```

Partial Output

Testing Whether the Mean Body Temperature=98.6

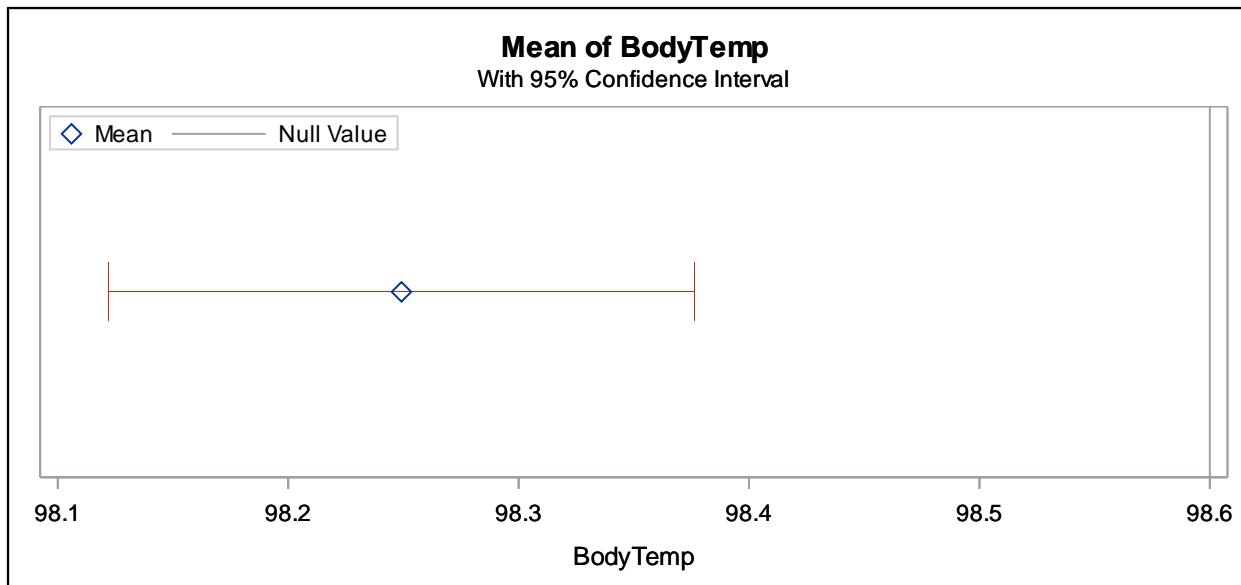
The TTEST Procedure

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

DF	t Value	Pr > t
129	-5.45	<.0001



- 5) What is the value of the t statistic and the corresponding p -value?

The t value is -5.45. The p -value is <.0001.

- 6) Do you reject or fail to reject the null hypothesis at the 0.05 level that the average temperature is 98.6 degrees?

You reject the null hypothesis at the 0.05 level.

2. Using PROC TTEST for Comparing Groups

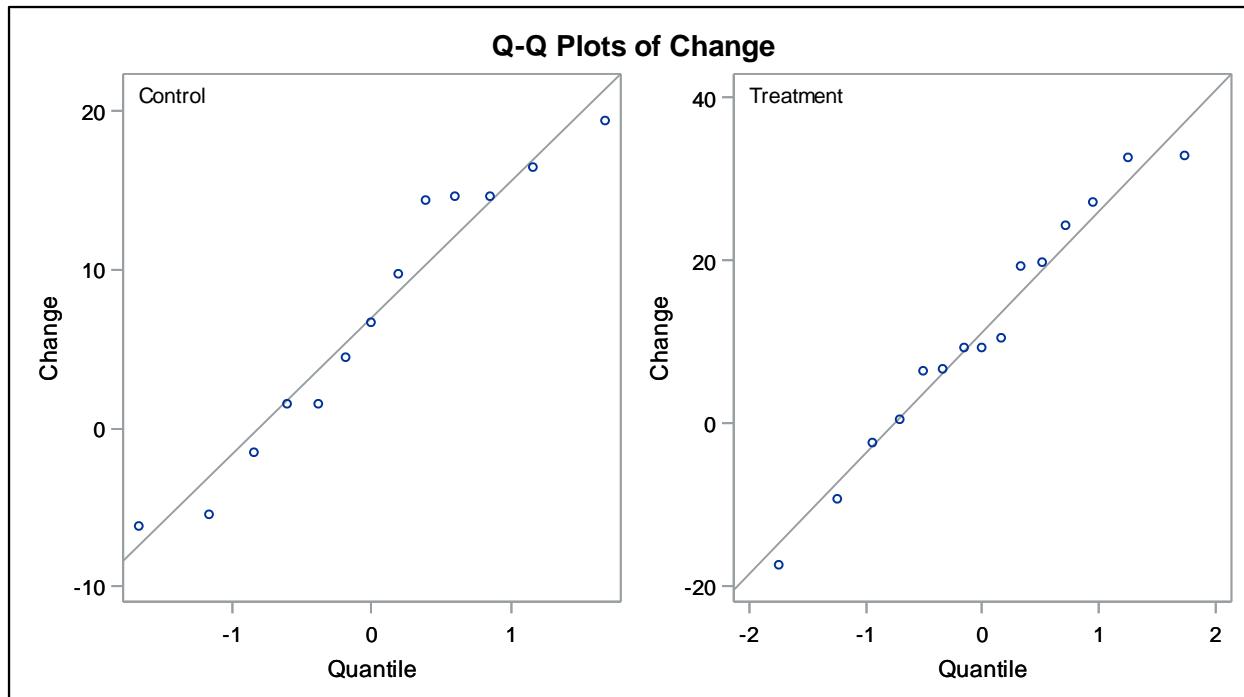
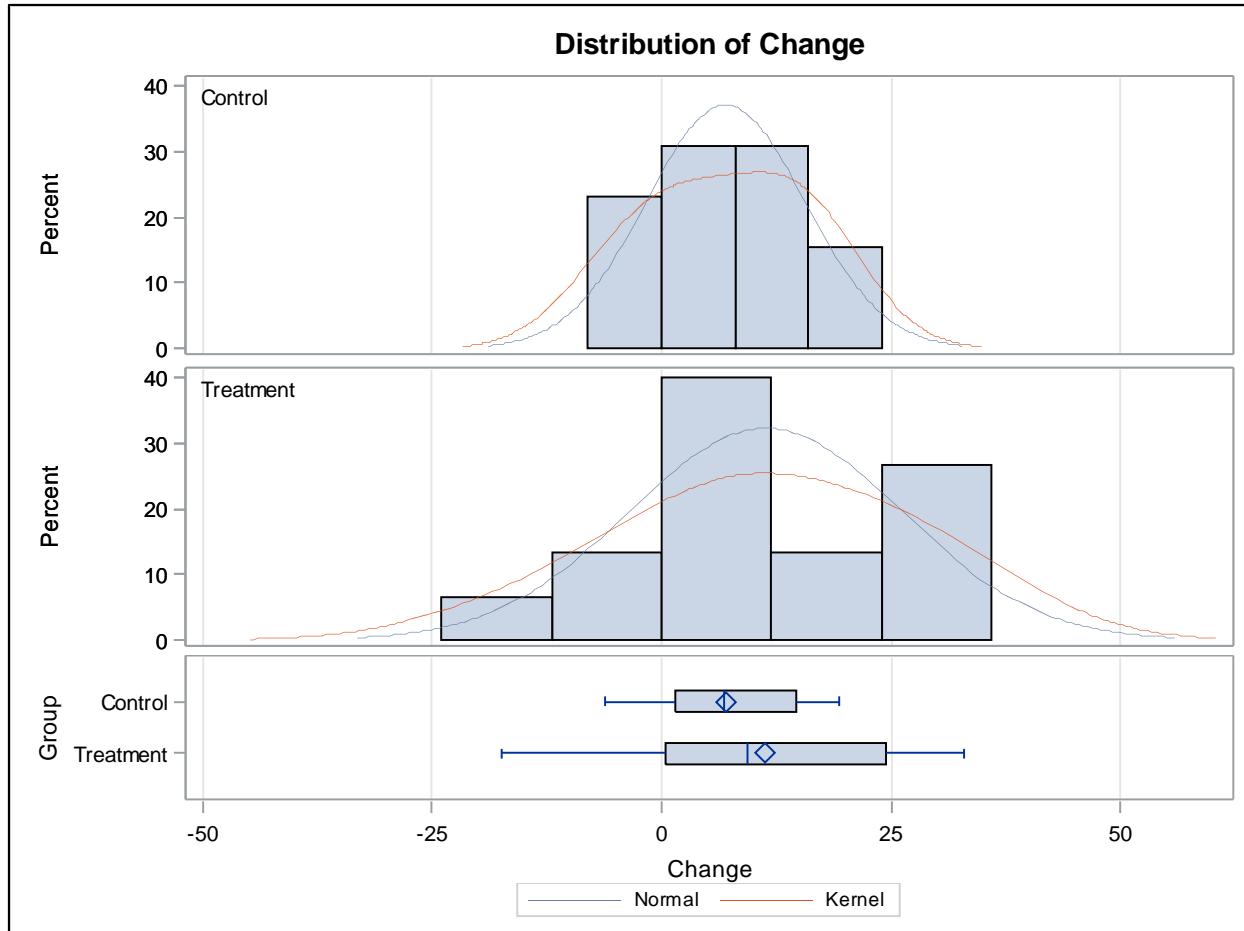
Analyze the data using PROC TTEST. Assess whether the treatment group improved more than the control group.

- Open the t Tests task.
- On the DATA tab, do the following:
 - Select the **GERMAN** data set.
 - Change to **Two-sample test** from One-sample test under **ROLES**.
 - Assign **Change** as the Analysis variable and **Group** as the Groups variable.
- On the OPTIONS tab, do the following:
 - Uncheck the option to conduct **Tests for normality**.
 - Under PLOTS choose the **Selected plots** option. Select the **Histogram and box plot**, **Normality plot**, and **Confidence interval plot** check boxes.

Note: Alternatively, write the SAS programming code directly.

```
/*st101s02.sas*/
ods graphics;
proc ttest data=STAT1.German plots(shownull)=interval;
  class Group;
  var Change;
  title "German Grammar Training, Comparing Treatment to Control";
run;
```

- d. Do the two groups appear to be approximately normally distributed?



The plots show evidence supporting approximate normality in both groups.

- e. Do the two groups have approximately equal variances?

From the bottom of the PROC TTEST output:

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	12	2.97	0.0660

Because the *p*-value for the Equality of Variances test is greater than the alpha level of 0.05, you would not reject the null hypothesis. This conclusion supports the assumption of equal variance (the null hypothesis being tested here).

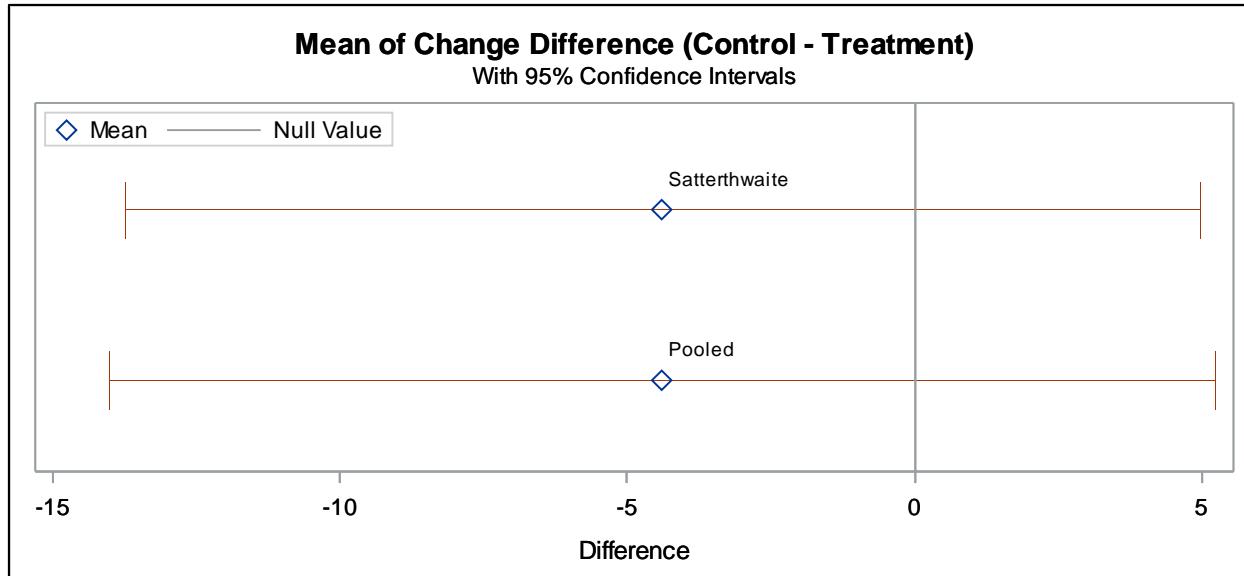
- f. Does the new teaching technique seem to result in significantly different change scores compared with the standard technique?

Group	N	Mean	Std Dev	Std Err	Minimum	Maximum
Control	13	6.9677	8.6166	2.3898	-6.2400	19.4100
Treatment	15	11.3587	14.8535	3.8352	-17.3300	32.9200
Diff (1-2)		-4.3910	12.3720	4.6882		

Group	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
Control		6.9677	1.7607 12.1747	8.6166	6.1789 14.2238
Treatment		11.3587	3.1331 19.5843	14.8535	10.8747 23.4255
Diff (1-2)	Pooled	-4.3910	-14.0276 5.2457	12.3720	9.7432 16.9550
Diff (1-2)	Satterthwaite	-4.3910	-13.7401 4.9581		

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	26	-0.94	0.3576
Satterthwaite	Unequal	22.947	-0.97	0.3413

The *p*-value for the Pooled (Equal Variance) test for the difference between the two means shows that the two groups are not statistically significantly different. Therefore, there is not strong enough evidence to say conclusively that the new teaching technique is different from the old. The Difference Interval plot displays these conclusions graphically.



The confidence interval includes the value zero, indicating a lack of statistical significance of the mean difference.

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

1.02 Multiple Choice Poll – Correct Answer

Which of the following affects alpha?

- a. The p -value of the test
- b. The sample size
- c. The number of Type I errors
- d. All of the above
- e. Answers a and b only
- f. None of the above

1.03 Quiz – Correct Answer

What is the null hypothesis for a one-sample t -test?

- a. $H_0: \mu = \mu_0$
- b. $H_0: \mu_0 = 0$
- c. $H_0: \mu - \mu_0 = 0$
- d. $H_0: \mu_0 - 0 = 0$

1.04 Multiple Answer Poll – Correct Answer

How do you tell PROC TTEST that you want to do a two-sample *t*-test?

- a. SAMPLE=2 option
- b. CLASS statement
- c. GROUPS=2 option
- d. PAIRED statement

Chapter 2 ANOVA and Regression

2.1 Graphical Analysis	2-3
Demonstration: Exploring Associations	2-9
2.2 One-Way ANOVA	2-20
Demonstration: Performing a One-Way ANOVA	2-36
Exercises.....	2-43
2.3 ANOVA Post Hoc Tests.....	2-44
Demonstration: Post Hoc Pairwise Comparisons	2-50
Exercises.....	2-56
2.4 Pearson Correlation	2-57
Demonstration: Data Exploration, Correlations, and Scatter Plots	2-66
Exercises.....	2-76
2.5 Simple Linear Regression.....	2-78
Demonstration: Performing Simple Linear Regression.....	2-88
Exercises.....	2-95
2.6 Solutions	2-96
Solutions to Exercises	2-96
Solutions to Student Activities (Polls/Quizzes)	2-116

2.1 Graphical Analysis

Objectives

- Explain what an association is.
- Graphically explore associations in the AmesHousing3 data set.

3



Associations

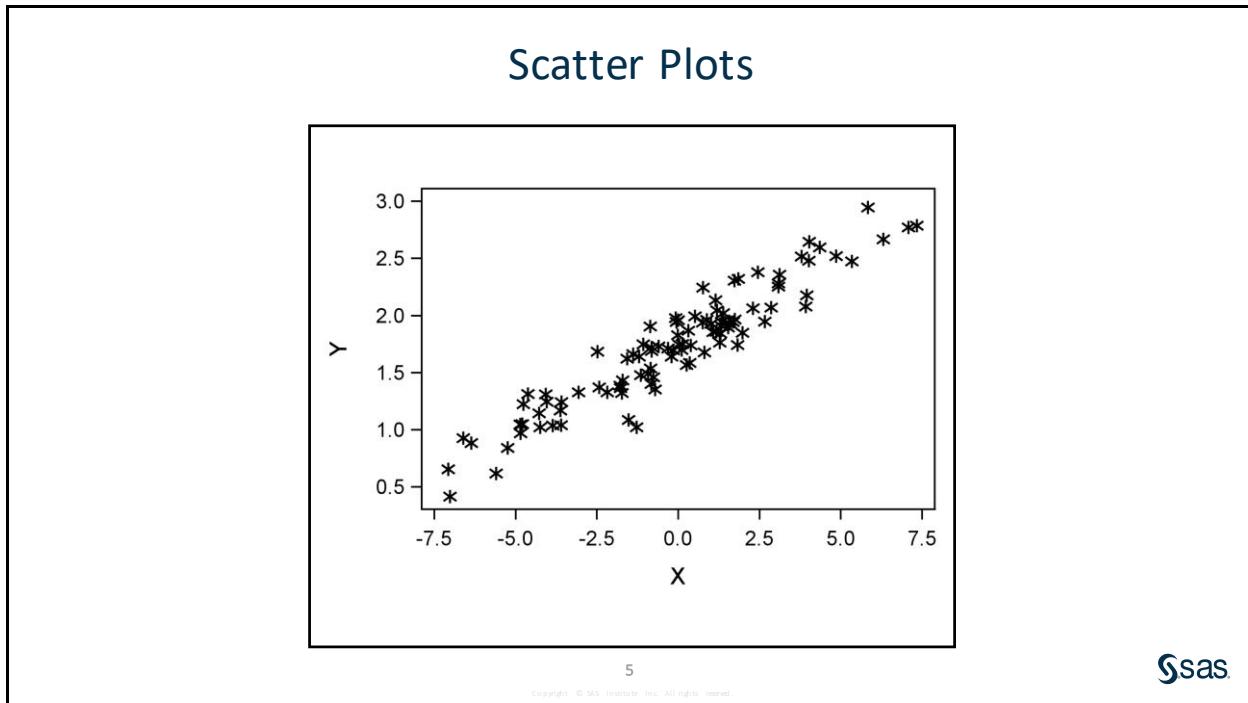
- An association exists between two variables when the expected value of one variable differs at different levels of the other variable.
- A *linear* association between two continuous variables can be inferred when the general shape of a scatter plot of the two variables is a straight line.

4



ANOVA and linear regression tests linear associations between predictor and response variables. In linear regression models, the predictor variable is continuous. In ANOVA, the predictor variable is categorical.

Typically, the categorical predictor is converted into binary dummy variables for purposes of model calculations. The following slides illustrate associations in ANOVA and regression.



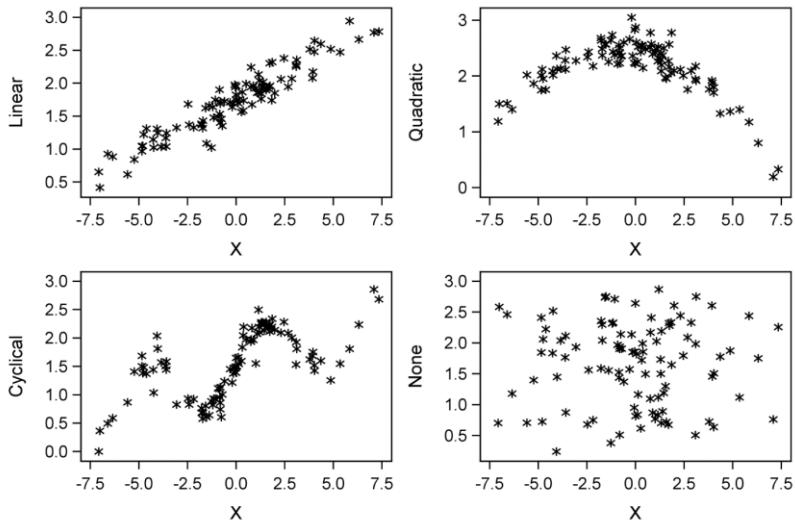
Scatter plots are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.

Scatter plots are useful to accomplish the following:

- 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

The predicted value can be thought of as the best estimate of the value of the response at a given value of the predictor variable. Scatter plots show graphically the relationship between predictor variables and response variables. Traditionally, predictor variables are plotted on the x axis and response variables are plotted on the y-axis. A preliminary analysis of associations involves discovery of the presence of associations and their nature.

Relationships between Continuous Variables

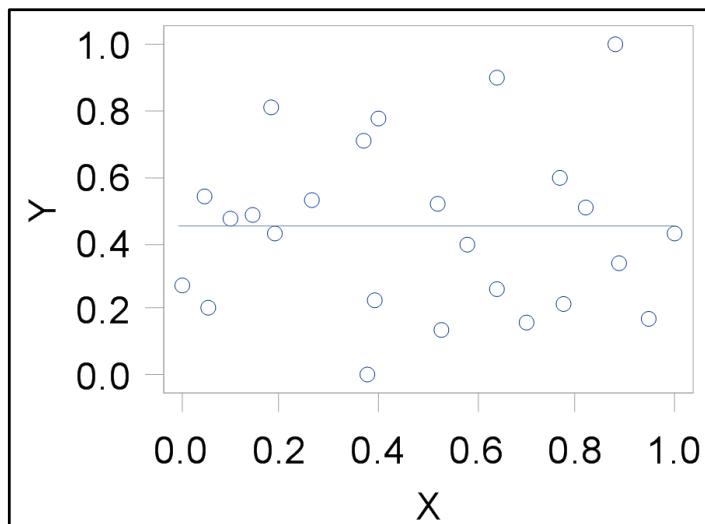


6

Describing the relationship between two continuous variables is an important first step in any statistical analysis. The scatter plot is the most important tool that 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.

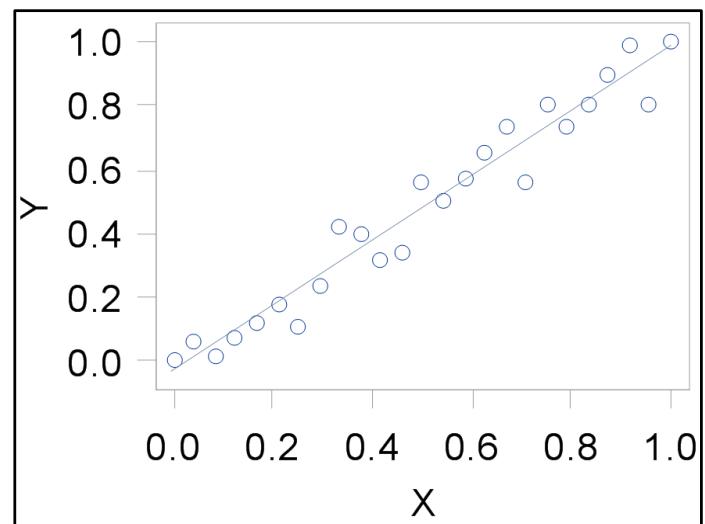
No Association – Continuous Predictor



7

If the value of x is unknown, then the best prediction of y would be the mean of y . The question becomes whether knowing x affects the best prediction of y . In regression, “best” is defined as the model (in this case, the regression line) that minimizes the sum of the squared differences between all actual y values and the corresponding predicted y values. If there is no association between x and y , then the best prediction of y will remain the mean value of y , even when x is known. The regression line will be horizontal at the value of the mean of y .

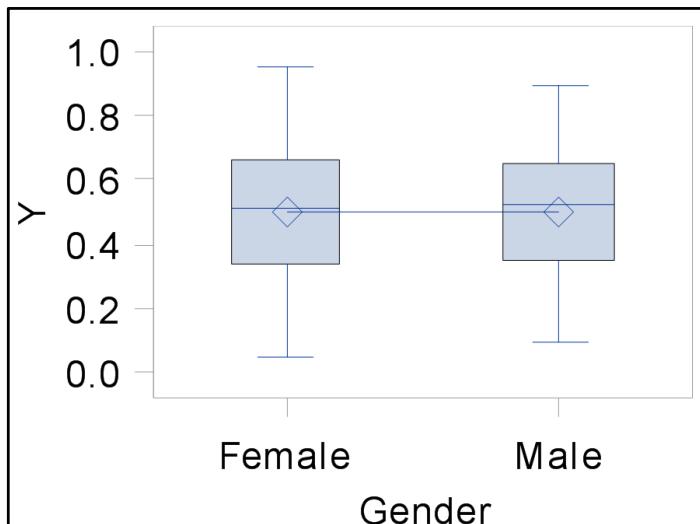
Linear Association – Continuous Predictor



8

If there is an association between x and y, then the best prediction of y would depend of the value of x. The regression line will not be horizontal.

No Association – Categorical Predictor

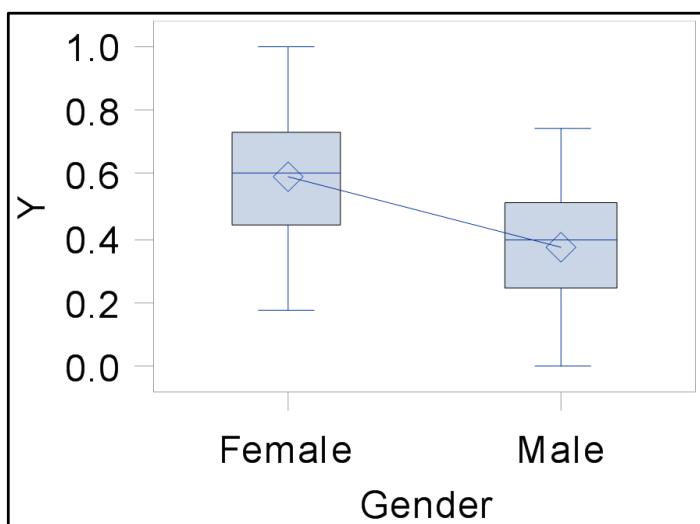


9

Sas

Similarly, when the predictor variable is categorical, lack of association with the response variable means that the best prediction of y is the mean of y, regardless of the value of x. Once again, a horizontal line through the means of all levels of the categorical predictor will indicate lack of association and therefore equal means.

Association – Categorical Predictor



10

Sas

Where there is an association between the categorical predictor and the continuous response, a plot will show a non-horizontal line connecting the category-specific means of y . The category-specific means will be better predictions than the overall mean of y . In other words, the average squared differences between the actual y and the predicted y will be smaller when the group-specific mean is the predicted value of y than when the overall mean is the predicted value of y .

2.01 Multiple Answer Poll

What is the visual cue in a scatter plot that there is no association between the response variable and the explanatory variables?

- a. All of the y values fall on a straight line.
- b. None of the y values fall on a straight line.
- c. All of the y mean values fall on a straight line.
- d. None of the y mean values fall on a straight line.
- e. All of the y values are the same.
- f. All of the y mean values are the same.



Exploring Associations

Example: Create scatter plots to show relationships between continuous predictors and **SalePrice** and comparative box plots to show relationships between categorical predictors and **SalePrice**.

1. Open the **Scatter Plot** task under Graph and select **Gr_Liv_Area** as the X variable and **SalePrice** as the Y Variable for the **AmesHousing3** data set.

The screenshot shows the SAS Studio interface with the following details:

- Left Sidebar (Tasks and Utilities):**
 - My Tasks
 - Tasks
 - Data
 - Graph
 - Bar Chart
 - Bar-Line Chart
 - Box Plot
 - Bubble Plot
 - Histogram
 - Line Chart
 - Mosaic Plot
 - Pie Chart
 - Scatter Plot (highlighted)
 - Series Plot
 - Simple HBar
 - Combinatorics and Probability
 - Statistics
 - High-Performance Statistics
 - Power and Sample Size
 - Multivariate Analysis
 - Econometrics
 - Forecasting
 - Statistical Process Control
 - Data Mining
 - Utilities
- Right Panel (DATA tab):**
 - Project: SetUpStat1.sas (active)
 - File: st102d01a
 - Buttons: Settings, Code/Results, Split
 - DATA, OPTIONS, INFORMATION tabs
 - DATA section: STAT1.AMESHOUSING3 selected
 - WHERE CLAUSE FILTER
 - ROLES
 - X variable: Gr_Liv_Area (selected)
 - Y variable: SalePrice (selected)
 - Group variable: Column
 - Marker label variable: Column
 - URL variable: Column
 - FIT PLOTS

2. Expand the **FIT PLOTS** property and select the **Regression** box to add a regression fit to the scatterplot.

DATA OPTIONS INFORMATION

▲ DATA

STAT1.AMESHOUSING3

▶ WHERE CLAUSE FILTER

▲ ROLES

* X variable: (1 item) Delete +
123 Gr_Liv_Area

* Y variable: (1 item) Delete +
123 SalePrice

Group variable: (1 item) Delete +
Column

Marker label variable: (1 item) Delete +
Column

URL variable: (1 item) Delete +
Column

▲ FIT PLOTS

Regression

Confidence limits for means

Prediction limits for individuals

*Alpha: Up Down

*Degree: Up Down

SAS Studio produces the follow code:

```
ods graphics / reset imagemap;

/*--SGPLOT proc statement--*/
proc sgplot data=STAT1.AMESHOUSING3;
    /*--Fit plot settings--*/
    reg x=Gr_Liv_Area y=SalePrice / nomarkers name='Regression';

    /*--Scatter plot settings--*/
    scatter x=Gr_Liv_Area y=SalePrice / transparency=0.0
        name='Scatter';

    /*--X Axis--*/
    xaxis grid;

    /*--Y Axis--*/
    yaxis grid;
run;

ods graphics / reset;
```

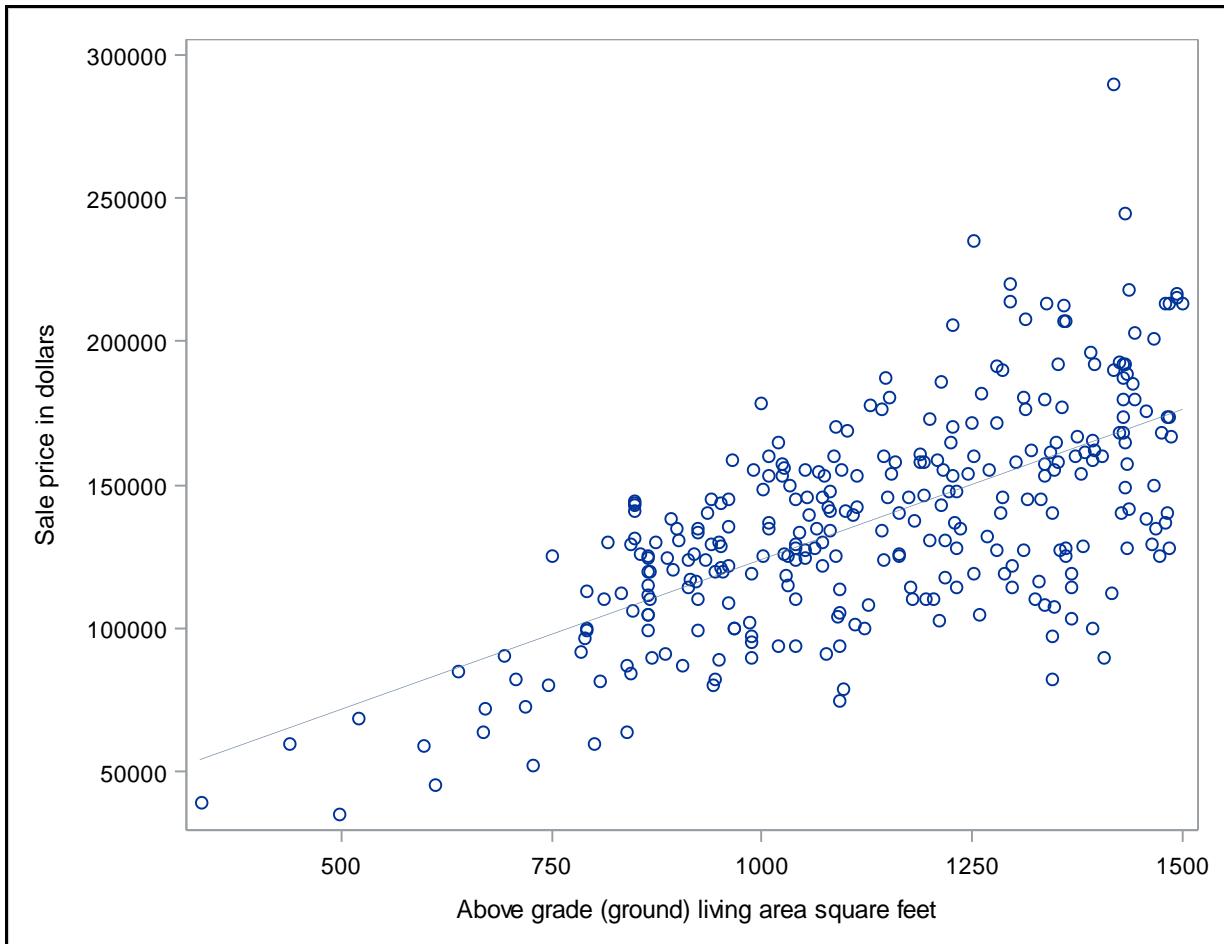
Note: Alternatively, you can write PROC SGSCATTER code:

```
/*st102d01.sas*/ /*Part A*/
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*Gr_Liv_Area / reg;
    title "Associations of Above Grade Living Area with Sale Price";
run;
```

Selected PLOT statement option:

REG Adds a regression fit to the scatter plot.

3. Submit the code.

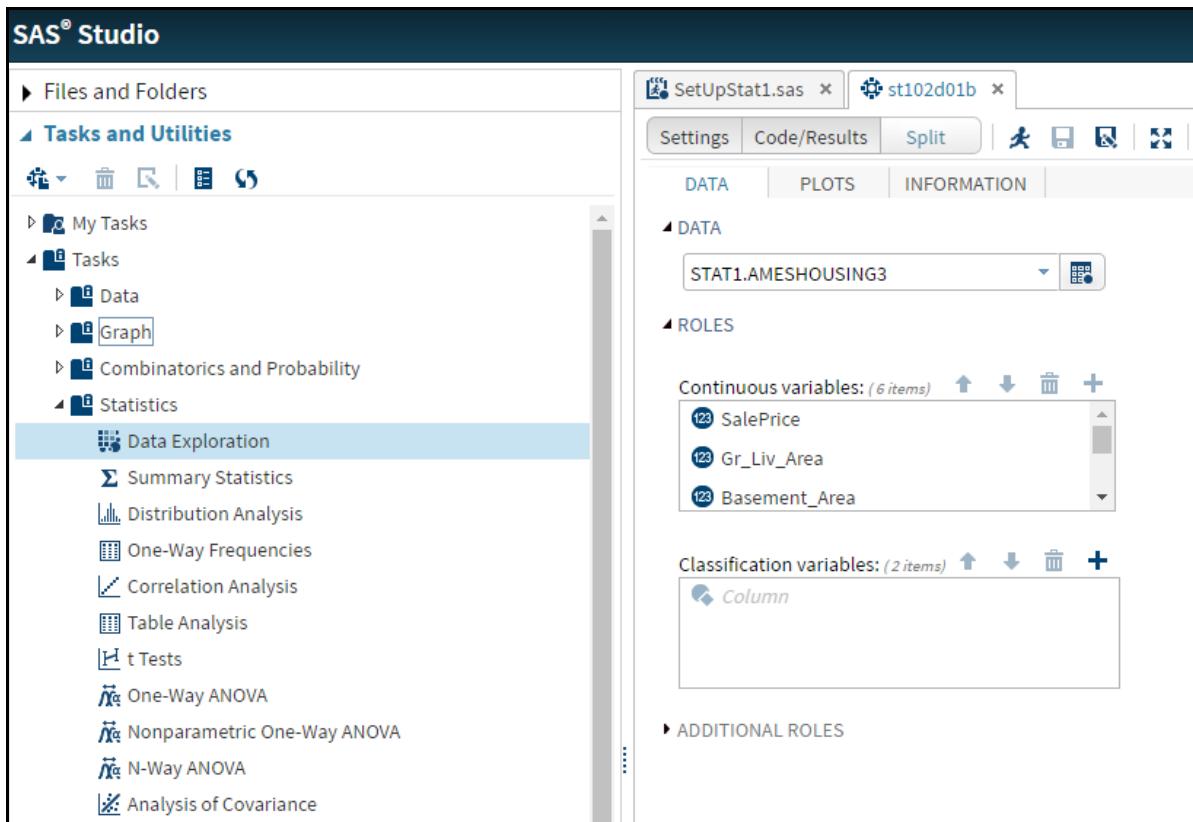


There does seem to be a nonzero association between above grade living area and sale price.

Note: There seems to be more variability in sale price at higher living area values. This is called *heteroscedasticity*. This topic is discussed in detail in the Statistics 2: ANOVA and Regression course.

Multiple Scatterplots

4. Open the **Data Exploration** task under Statistics to plot multiple correlation plots at once.
5. Assign variables of interest (**SalePrice**, **Gr_Liv_Area**, **Deck_Porch_Area**, **Lot_Area**, **Basement_Area**, and **Garage_Area**) to Continuous variables.



6. On the PLOTS tab, clear the option to output a scatter plot matrix.
7. Select the option to plot **Regression scatter plots** box and select **SalePrice** as the response variables. Select the option to include a fitted line to the scatter plot.

8. Run the code

The screenshot shows the SAS Studio Data Exploration task interface. At the top, there are three tabs: DATA, PLOTS, and INFORMATION. The PLOTS tab is selected. Below the tabs, there are three main sections: SCATTER PLOT MATRIX, PAIRWISE SCATTER PLOTS, and REGRESSION SCATTER PLOTS. Under REGRESSION SCATTER PLOTS, the 'Regression scatter plots' checkbox is checked. A dropdown menu titled 'Select response variables' contains three items: SalePrice, Gr_Liv_Area, and Basement_Area. Below this, there are three checkboxes for fitted lines: 'Add a fitted line' (checked), 'Add a loess fit' (unchecked), and 'Add a fitted, penalized B-spline curve' (unchecked). At the bottom, there is a section for HISTOGRAM AND BOX PLOT with the note: 'Available when no classification variable is specified.' and a checkbox for 'Histogram and box plot'.

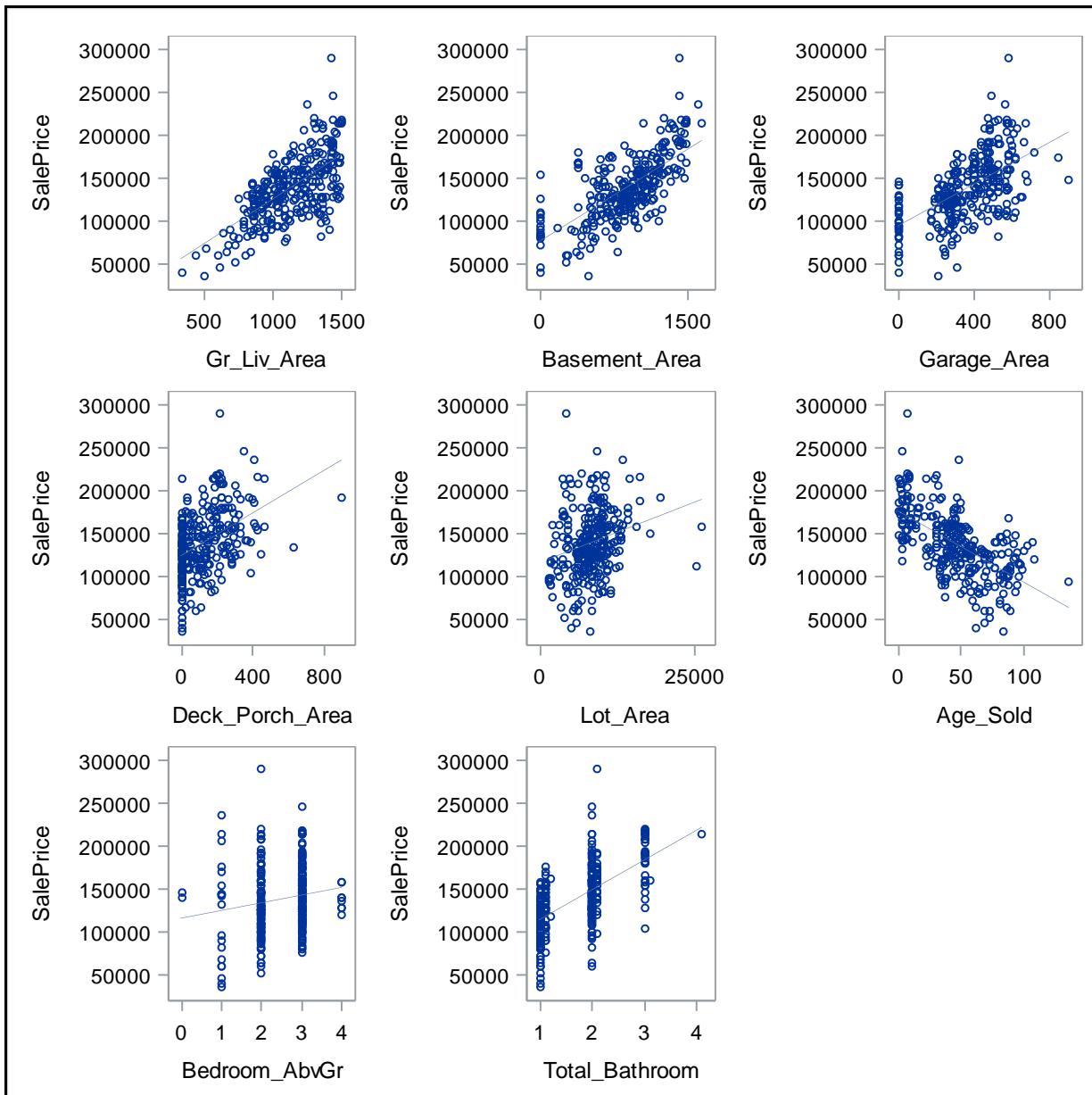
Note: The SAS Studio Data Exploration task limits the number of continuous variables to 6 and outputs individual scatter plots. To plot more than 5 variables at once in a panel plot, use PROC SGSCATTER. Example code and output are as follows:

```
/*st102d01.sas*/ /*Part B*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
    Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

options nolabel;
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*(&interval) / reg;
    title "Associations of Interval Variables with Sale Price";
run;
```

OPTIONS statement option:

NOLABEL Does not allow SAS procedures to use labels with variables.



There seems to be some association between each of the predictor variables and **SalePrice**.

Box plots for categorical predictors.

9. Open the **Box Plot** task under Graph.
10. Set the Analysis variable to **SalePrice**.

11. Set the category variable to **Central_Air**.

The screenshot shows the SAS Studio interface. On the left, the Tasks and Utilities sidebar is open, displaying various analysis and visualization tools. The main workspace is titled "DATA" and contains the following configuration:

- Analysis variable:** SalePrice
- Category variable:** Central_Air
- Group variable:** Column
- BY variable:** Column

12. On the OPTIONS tab, expand the Analysis Axis property and uncheck the Show grid option.

13. Run the code.

DATA OPTIONS INFORMATION

▲ TITLE AND FOOTNOTE

Title:

Set title font size

*Font size:

Footnote:

Set footnote font size

*Font size:

▶ BOX DETAIL

▶ GROUP LAYOUT

▲ CATEGORY AXIS

Reverse

Show values in data order

Show label

Custom label:

▲ ANALYSIS AXIS

Show grid

Show label

Custom label:

▶ LEGEND DETAILS

▶ GRAPH SIZE

Note: Equivalently, you can write the code directly in SAS.

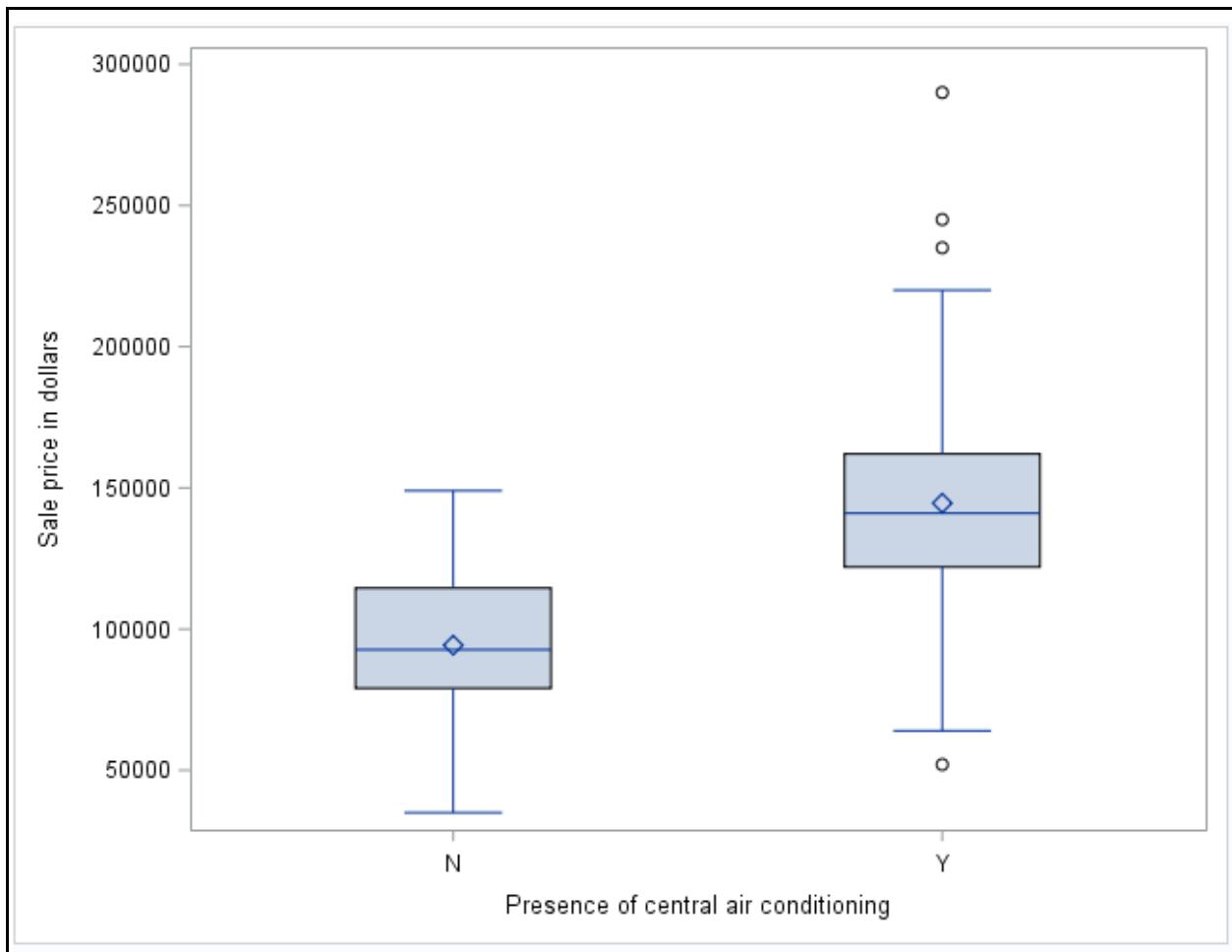
```
/*st102d01.sas*/ /*Part C*/
proc sgplot data=STAT1.ameshousing3;
  vbox SalePrice / category=Central_Air
    title "Sale Price Differences across Central Air";
run;
```

PROC SGLOT statement:

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

VBOX statement options:

CATEGORY= Specifies the category variable for the plot. A box plot is created for each distinct value of the category variable.



Houses with central air sell on average at higher prices than houses without central air. Therefore, there is a nonzero association between **Central_Air** and **SalePrice**.

Note: The Data Exploration task can be used to generate two separate box plots in one task. In order to generate multiple comparative box plots, one can use macro code. An example is shown.

```
%macro box(dsn      = ,
           response = ,
           charvar  = ) ;

%let i = 1 ;

%do %while(%scan(&charvar,&i,%str( )) ^= %str()) ;
```

```
%let var = %scan(&charvar,&i,%str( ));

proc sgplot data=&dsn;
  vbox &response / category=&var
    grouporder=ascending
    connect=mean;
  title "&response across Levels of &var";
run;

%let i = %eval(&i + 1) ;

%end;

%mend box;
```

The main part of the macro is the following code:

```
proc sgplot data=&dsn;
  vbox &response / category=&var
    grouporder=ascending
    connect=mean;
  title "&response across Levels of &var";
run;
```

This is similar to the code run in the previous example, which used data set values for category, data set name, and response variable. The **%charvar** macro variable is read by SAS as a string, with spaces separating members of the list of variables. The following statement creates a macro variable called **&var**, which is chosen to be the i^{th} member of **&charvar** string:

```
%let var = %scan(&charvar,&i,%str( ));
```

The members are defined to be separated by a space using **%str()**, which is the third argument in the **%scan** function. The **&i**, which is the second argument to the function indexes the ordered value of the member of the string. The **%do %while** loop uses the **%scan** function to check to see whether there are any more members in the macro variable **&charvar**.

The macro is called using the following code:

```
%box(dsn      = STAT1.ameshousing3,
      response = SalePrice,
      charvar  = &categorical);
```

The output is not displayed here.

End of Demonstration

2.2 One-Way ANOVA

Objectives

- Use the GLM procedure to analyze the differences between population means.
- Verify the assumptions of analysis of variance.

15

Copyright © SAS Institute Inc. All rights reserved.



Overview of Statistical Models

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

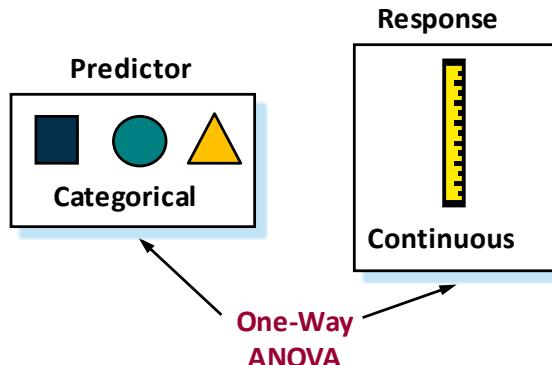
16

Copyright © SAS Institute Inc. All rights reserved.



Overview

Are there any differences among the population means?



Another way of asking: Does information about group membership help predict the level of a numeric response?

Copyright © SAS Institute Inc. All rights reserved.



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 the following:

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

Research Questions for One-Way ANOVA

Do accountants. on average, earn more than teachers? *



* Is this a case for a *t* test?

Copyright © SAS Institute Inc. All rights reserved.



If you analyze the difference between two means using ANOVA, you reach the same conclusions as you reach using a pooled, two-group *t*-test. Performing a two-group mean comparison in PROC GLM gives

you access to graphical and assessment tools different from those available in performing the same comparison with PROC TTEST.

Research Questions for One-Way ANOVA

Do people treated with one of two new drugs have higher average T-cell counts than people in the control group?



Placebo



Treatment 1



Treatment 2

Copyright © SAS Institute Inc. All rights reserved.



When there are three or more levels for the grouping variable, a simple approach is to run a series of *t*-tests between all the pairs of levels. For example, you might be interested in T-cell counts in patients taking three medications (including one placebo). You could simply run a *t*-test for each pair of medications. A more powerful approach is to analyze all the data simultaneously. The mathematical model is called a *one-way analysis of variance* (ANOVA), and the test statistic used is the *F* ratio, rather than the Student's *t* value.

Research Questions for One-Way ANOVA

Do people spend different amounts depending on which type of credit card they have?



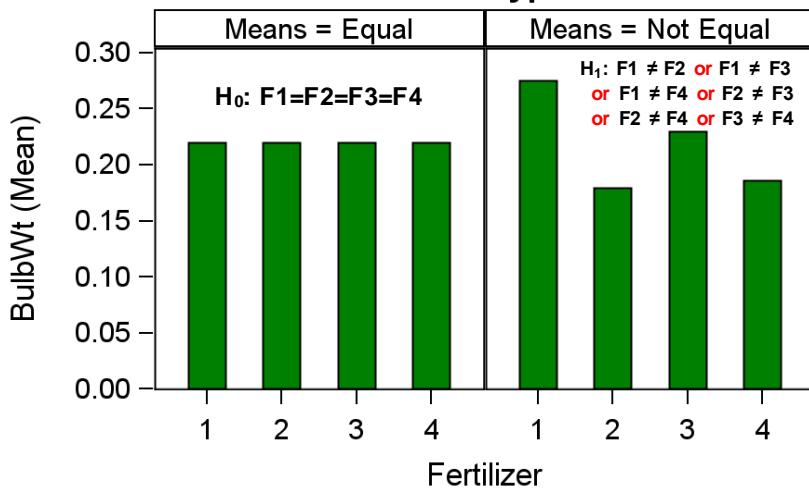
20

Copyright © SAS Institute Inc. All rights reserved.

The SAS logo, which consists of the word "SAS" in a stylized, italicized font with a registered trademark symbol.

The ANOVA Hypothesis

Null and Alternative Hypotheses



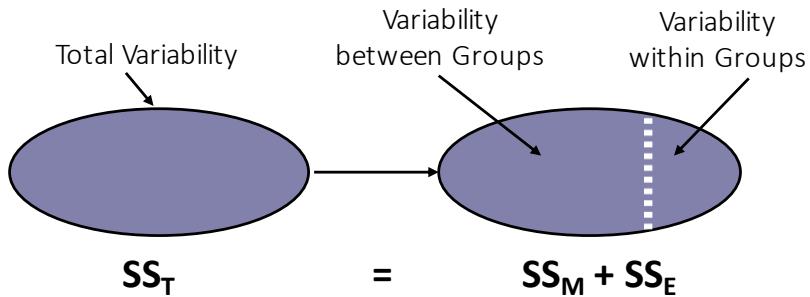
21

Copyright © SAS Institute Inc. All rights reserved.

The SAS logo, which consists of the word "SAS" in a stylized, italicized font with a registered trademark symbol.

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

Partitioning Variability in ANOVA



22

Copyright © SAS Institute Inc. All rights reserved.



In ANOVA, the Total Variation (as measured by the corrected total sum of squares) is partitioned into two components, the Between Group Variation (displayed in the ANOVA table as the Model Sum of Squares) and the Within Group Variation (displayed as the Error Sum of Squares). As its name implies, ANalysis Of VAriance analyzes, or breaks apart, the variance of the dependent variable to determine whether the between-group variation is a significant portion of the total variation. ANOVA compares the portion of variation in the response variable attributable to the grouping variable to the portion of variability that is unexplained. The test statistic, the *F* Ratio, is a ratio of the model variance to the error variance. The calculations are shown below.

Total Variation

the *overall* variability in the response variable. It is calculated as the sum of the squared differences between each observed value and the overall mean, $\sum \sum (Y_{ij} - \bar{Y})^2$. This measure is also referred to as the *Total Sum of Squares* (SS_T).

Between Group Variation

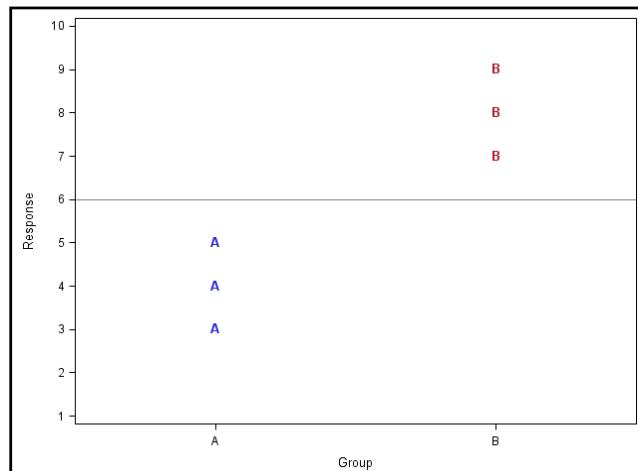
the variability explained by the independent variable and therefore represented by the between-treatment sum of squares. It is calculated as the weighted (by group size) sum of the squared differences between the mean for each group and the overall mean, $\sum n_i (\bar{Y}_i - \bar{Y})^2$. This measure is also referred to as the *Model Sum of Squares* (SS_M).

Within Group Variation

the variability not explained by the model. It is also referred to as *within treatment variability* or *residual sum of squares*. It is calculated as the sum of the squared differences between each observed value and the mean for its group, $\sum \sum (Y_{ij} - \bar{Y}_i)^2$. This measure is also referred to as the *Error Sum of Squares* (SS_E).

Note: $SS_T = SS_M + SS_E$, meaning that the model sum of squares and the error sum of squares sums to the total sum of squares.

Sums of Squares



$$\text{Overall mean} = \bar{y} = \frac{3+4+5+7+8+9}{6} = 6$$

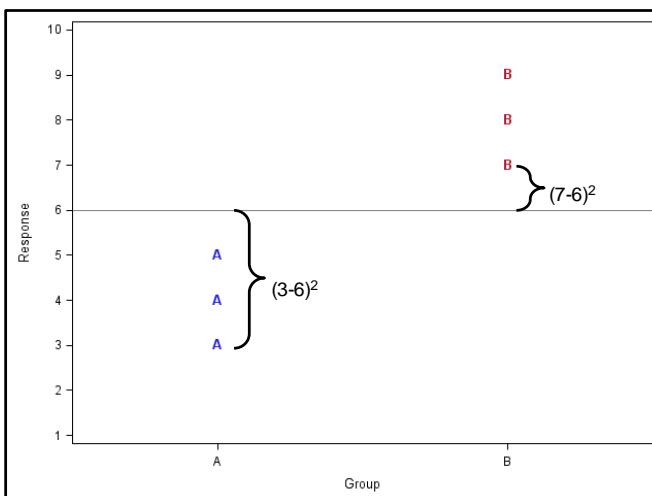
23



Copyright © SAS Institute Inc. All rights reserved.

A simple example of the various sums of squares is shown in this set of slides. First, the overall mean of all data values is calculated.

Total Sum of Squares



$$SS_T = (3-6)^2 + (4-6)^2 + (5-6)^2 + (7-6)^2 + (8-6)^2 + (9-6)^2 = 28$$

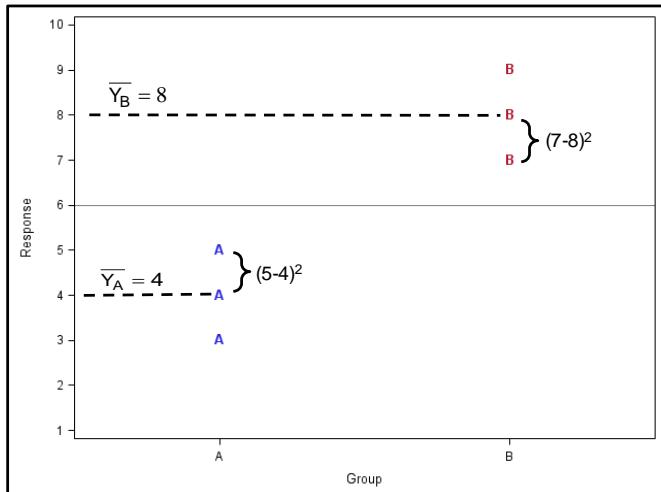
24



Copyright © SAS Institute Inc. All rights reserved.

The total sum of squares, SS_T , is a measure of the total variability in a response variable. It is calculated by summing the squared distances from each point to the overall mean. Because it is correcting for the mean, this sum is sometimes called the *corrected total sum of squares*.

Error Sum of Squares



$$SS_E = (3-4)^2 + (4-4)^2 + (5-4)^2 + (7-8)^2 + (8-8)^2 + (9-8)^2 = 4$$

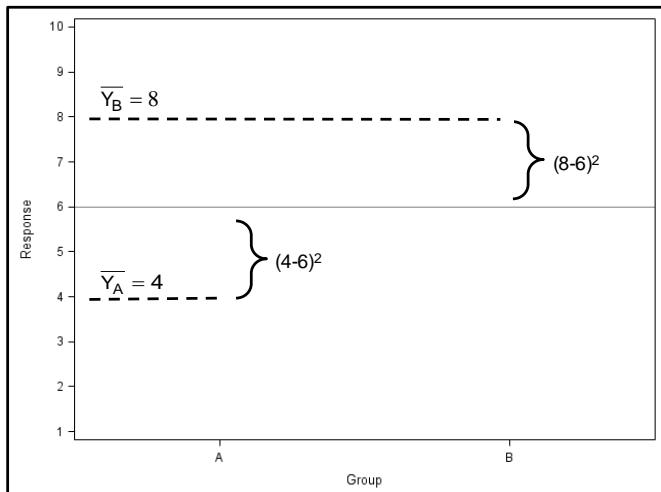
25



Copyright © SAS Institute Inc. All rights reserved.

The error sum of squares, SS_E , measures the random variability ***within*** groups; it is the sum of the squared deviations between observations in each group and that group's mean. This is often referred to as the *unexplained variation* or *within-group variation*.

Model Sum of Squares



$$SS_M = 3^*(4-6)^2 + 3^*(8-6)^2 = 24$$

26



Copyright © SAS Institute Inc. All rights reserved.

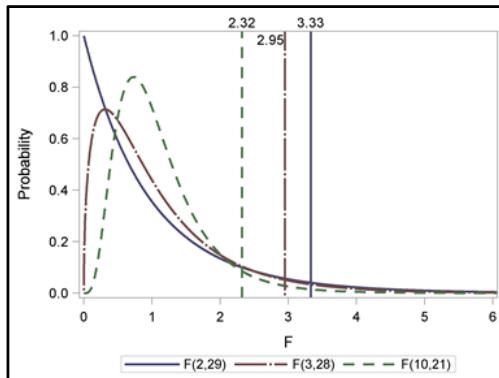
The model sum of squares, SS_M , measures the variability *between* groups; it is the sum of the squared deviations between each group mean and the overall mean, weighted by the number of observations in each group. This is often referred to as the *explained variation*. The model sum of squares can also be calculated by subtracting the error sum of squares from the total sum of squares: $SS_M = SS_T - SS_E$.

In this example, the model explains approximately 85.7%, $((SS_M / SS_T) * 100\%)$, of the variability in the response. The other 14.3% represents unexplained variability, or process variation. In other words, the variability due to differences between the groups (the explained variability) makes up a larger proportion of the total variability than the random error within the groups (the unexplained variability).

The total sum of squares (SS_T) refers to the *overall* variability in the response variable. The SS_T is computed under the null hypothesis (that the group means are all the same). The error sum of squares (SS_E) refers to the variability *within* the treatments not explained by the independent variable. The SS_E is computed under the alternative hypothesis (that the model includes nonzero effects). The model sum of squares (SS_M) refers to the variability *between* the treatments explained by the independent variable.

The basic measures of variation under the two hypotheses are transformed into a ratio of the model and the error variances, which has a known distribution (Snedecor's F distribution) under the null hypothesis that all group means are equal. The F ratio can be used to compute a p -value.

F Statistic and Critical Values at $\alpha=0.05$



$$F(\text{Model df, Error df}) = MS_M / MS_E$$

The null hypothesis for analysis of variance is tested using an F statistic. The F statistic is calculated as the ratio of the Between Group Variance to the Within Group Variance. In the output of PROC GLM, these values are shown as the Model Mean Square and the Error Mean Square. The mean square values are calculated as the sum of square value divided by the degrees of freedom.

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 sample size minus the number of treatments (or 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 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.
- $F = \frac{MSM}{MSE} = \frac{\frac{SS_M}{df_M}}{\frac{SS_E}{df_E}}$.
- The *p*-value for the test is then calculated from the F distribution with appropriate degrees of freedom.

Note: *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 sample mean (\bar{x}) is an unbiased estimate of the population mean (μ), bias=0 and MSE measures the variance.

Coefficient of Determination

$$R^2 = \frac{SS_M}{SS_T}$$

“Proportion of variance accounted for by the model”

The *coefficient of determination*, R^2 , is a measure of the proportion of variability in the response or dependent variables explained by the explanatory or independent variables in the analysis. This statistic is calculated as $R^2 = \frac{SS_M}{SS_T}$

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 R² closer to 1 are preferred, judging the magnitude of R² depends on the context of the problem.

The ANOVA Model

$$\text{SalePrice} = \text{Base Level} + \text{Central_Air} + \text{Unaccounted for Variation}$$

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

29

Copyright © SAS Institute Inc. All rights reserved.



The model, $Y_{ik}=\mu+\tau_i+\varepsilon_{ik}$, is one way of representing the relationship between the dependent and independent variables in ANOVA.

- Y_{ik} the k^{th} value of the response variable for the i^{th} treatment.
- μ the overall population mean of the response for example, sale price.
- τ_i the difference between the population mean of the i^{th} treatment and the overall mean, μ . This is referred to as the *effect* of treatment i .
- ε_{ik} the difference between the observed value of the k^{th} observation in the i^{th} group and the mean of the i^{th} group. This is called the *error term*.

Note: PROC GLM uses a parameterization of categorical variables in its CLASS statement that will not directly estimate the values of the parameters in the model shown. The correct parameter estimates can be obtained by adding the SOLUTION option in the MODEL statement in PROC GLM and then using simple algebra. Parameter estimates and standard errors can also be obtained using ESTIMATE statements. These issues are discussed in depth in the Statistics 2: ANOVA and Regression course and in the SAS documentation.

Note: In this data set, the list of categories observed in the categorical variables is exhaustive. In other words, there are no other levels imagined possible. In some applications this would be considered a *fixed effect*. If the observed levels of a categorical variable comprise just a sample of many that could have been used (for example, if you used neighborhood as an explanatory variable and only looked at houses in 10 neighborhoods, but were really interested in generalizing to all communities), the sampling variability of that variable would need to be taken into account in the model. In that case, the variable would be treated as a *random effect*. Random effects are discussed in the Statistics 2: ANOVA and Regression course.

2.02 Multiple Answer Poll

What does the R-Square value measure?

- a. The correlation between the independent and dependent variables.
- b. The proportion of total sum of squares accounted for by the model.
- c. Model sum of squares over error sum of squares.
- d. Something to do with variability.

30

Copyright © SAS Institute Inc. All rights reserved.

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;
```

32

Copyright © SAS Institute Inc. All rights reserved.

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 Least Squared Means, which are means adjusted 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.

Note: 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.

What Does a CLASS Statement Actually Do?

- The CLASS statement creates a set of “design variables” representing the information in the categorical variables.
- PROC GLM performs linear regression on the design variables, but reports the output in a manner interpretable as group mean differences.

Note: There is only one “parameterization” available in PROC GLM.

The CLASS statement creates a set of “design variables” (sometimes referred to as “dummy variables”) representing the information contained in any categorical variables. Linear regression is then performed on the design variables. ANOVA can be thought of as linear regression on dummy variables. It is only in the interpretation of the model that a distinction is made.

Even if categorical variables are represented by numbers such as 1, 2, 3, the CLASS statement tells SAS to set up design variables to represent the categories. If a numerically coded categorical variable were not included in the CLASS variable list, then PROC GLM would interpret it as a continuous variable in the regression calculations.

GLM Coding of CLASS Variables

<u>CLASS</u>	<u>Value</u>	<u>Label</u>	Design Variables		
			<u>1</u>	<u>2</u>	<u>3</u>
IncLevel	1	Low Income	1	0	0
	2	Medium Income	0	1	0
	3	High Income	0	0	1

For CLASS variable coding in PROC GLM, the number of design variables created is the number of levels of the CLASS variable. For example, because the variable **IncLevel** has three levels, three design variables are created. Each design variable is a binary indicator of membership in a particular level of the CLASS variable. So, each observation in the data set will be assigned values on all three of these new variables in PROC GLM.

In this parameterization scheme, however, a third design variable is always redundant when the other two are included. For example, if you know that **IncLevel** is not 1 and **IncLevel** is also not 2, then you do not need a third variable to tell you that **IncLevel** is 3. Because the design variables are read in order, it is the third design variable that is considered redundant.

Note: If you would like to see the regression equation estimates for the design variables, you can add the SOLUTION option to the MODEL statement in PROC GLM.

Note: Interpretation of regression parameters will be discussed in a later section.

Assumptions for ANOVA

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

35

Copyright © SAS Institute Inc. All rights reserved.



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.

- ***Independence*** implies that the ε_{ij} occurrences in the theoretical model are uncorrelated.
- The ***errors are assumed to be normally distributed*** for every group or treatment.
- Approximately ***equal error variances*** are assumed across treatments.

Assessing ANOVA Assumptions

- In many cases, good data collection designs can help ensure the independence assumption.
- Diagnostic plots from PROC GLM can be used to verify the assumption that the error is approximately normally distributed.
- PROC GLM produces a test of equal variances with the HOVTEST option in the MEANS statement. H₀ for this hypothesis test is that the variances are equal for all populations.

36

Copyright © SAS Institute Inc. All rights reserved.



Note: Additional tests and remedies for violations of these assumptions are described in the Statistics 2: ANOVA and Regression course.

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.

Observation	Heating_QC	Observed	Predicted	Residual
1	Ex	213500.0000	154919.1869	58580.8131
2	Ex	191500.0000	154919.1869	36580.8131
3	TA	115000.0000	130573.5294	-15573.5294
4	Ex	160000.0000	154919.1869	5080.8131
5	Ex	180000.0000	154919.1869	25080.8131
6	TA	125000.0000	130573.5294	-5573.5294
7	TA	206000.0000	130573.5294	75426.4706
8	Gd	159000.0000	130844.0862	28155.9138
9	TA	180500.0000	130573.5294	49926.4706
10	Gd	142125.0000	130844.0862	11280.9138

37

Copyright © SAS Institute Inc. All Rights Reserved.



The residuals from the ANOVA are calculated as the actual values minus the predicted values (the group means in ANOVA). Diagnostic plots (including normal quantile-quantile plots of the residuals) can be used to assess the normality assumption. With a reasonably sized sample and approximately equal groups (balanced design), only severe departures from normality are considered a problem. Residual values sum to 0 in ANOVA and ordinary least squares regression.

Note: 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 visually assess whether the variability is constant across groups.

2.03 Multiple Choice Poll

If you have 20 observations in your ANOVA and you calculate the residuals, to which of the following would they sum?

- a. -20
- b. 0
- c. 20
- d. 400
- e. Unable to tell from the information given

38

Copyright © SAS Institute Inc. All rights reserved.

2.04 Multiple Choice Poll

If you have 20 observations in your ANOVA and you calculate the squared residuals, to which of the following would they sum?

- a. -20
- b. 0
- c. 20
- d. 400
- e. Unable to tell from the information given

40

Copyright © SAS Institute Inc. All rights reserved.



Performing a One-Way ANOVA

Example: Run an analysis of variance with **SalePrice** as the response variable and **Heating_QC** as the categorical predictor variable. Output diagnostic plots and look at Levene's test of homogeneity of variances.

1. Open the **One-Way ANOVA** task under Statistics.
2. Select the **AmesHousing3** data set, assign **SalePrice** as the Dependent variable, and assign **Heating_QC** as the Categorical variable.
3. On the OPTIONS tab, uncheck the box for **Welch's variance-weightedANOVA**, and change the **Comparisons method** option to **None**. Under Display plots, select the **Selected plots** option, and check the **Box Plot** and **Diagnostics plot** boxes.

SAS Studio generates the following code:

```
Title;
ods noproctitle;
ods graphics / imagemap=on;

proc glm data=STAT1.AMESHOUSING3 plots(only)=(boxplot diagnostics);
  class Heating_QC;
  model SalePrice=Heating_QC;
  means Heating_QC / hovtest=levene plots=none;
  run;
quit;
```

Note: Alternatively, you can write the code directly in SAS.

```
/*st102d02.sas*/
ods graphics;

proc glm data=STAT1.ameshousing3 plots=diagnostics;
  class Heating_QC;
  model SalePrice=Heating_QC;
  means Heating_QC / hovtest=levene;
  format Heating_QC $Heating_QC.;
  title "One-Way ANOVA with Heating Quality as Predictor";
run;
quit;
```

Selected PLOTS option:

DIAGNOSTICS produces a panel display of diagnostic plots for linear models.

Note: The UNPACK option can be used in order to separate the individual plots in the panel display.

Selected MEANS statement option:

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

Partial Output:

Turn your attention to the first two tables of the output. The first table specifies the number of levels and the values of the class variable. When a FORMAT statement is used in the SAS code, the formatted values are displayed.

One-Way ANOVA with Heating Quality as Predictor

Class Level Information		
Class	Levels	Values
Heating_QC	4	Average/Typical Excellent Fair Good

Note: The output from the task shows the unformatted values.

Class Level Information		
Class	Levels	Values
Heating_QC	4	Ex Fa Gd TA

The second table shows both the number of observations read and the number of observations used. These values are the same because there are no missing values in for any variable in the model. If any row has **missing data** for a predictor or response variable, that row is **dropped** from the analysis.

Number of Observations Read	300
Number of Observations Used	300

The second part 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 effect of the independent variable in the model

Look at each of these parts separately.

Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

The *F* statistic and corresponding *p*-value are reported in the Analysis of Variance table. Because the reported *p*-value (<.0001) is less than 0.05, you reject the null hypothesis of no difference between the means.

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.157920	25.23100	34698.90	137524.9

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 measurement.

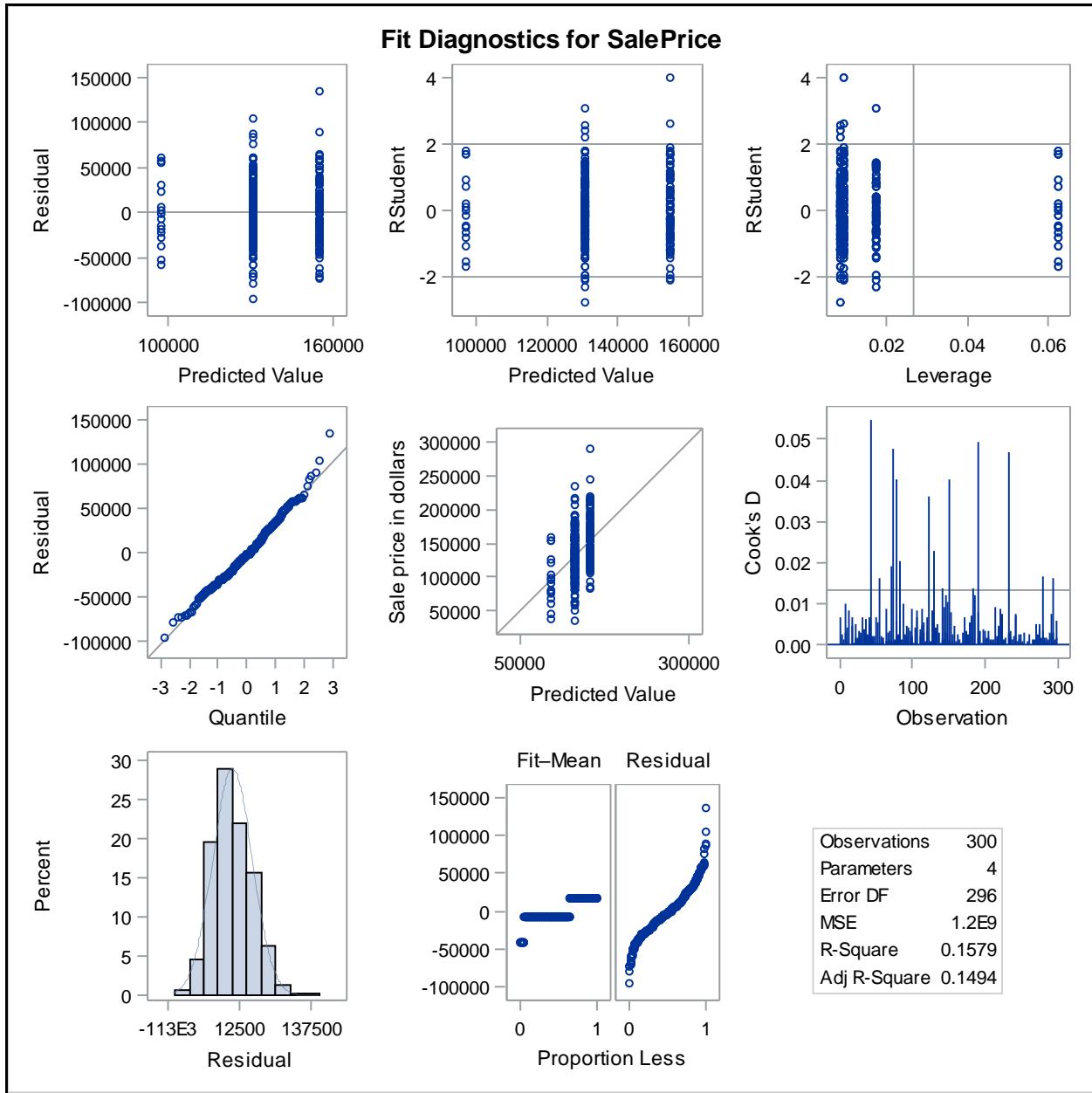
The **SalePrice Mean** is the mean of all of the data values for the variable **SalePrice**, without regard for **Heating_QC**.

As discussed previously, the R^2 value is often interpreted as the “proportion of variance accounted for by the model.” Therefore, you might say that in this model, **Heating_QC** explains about 16% of the variability of **SalePrice**.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001

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



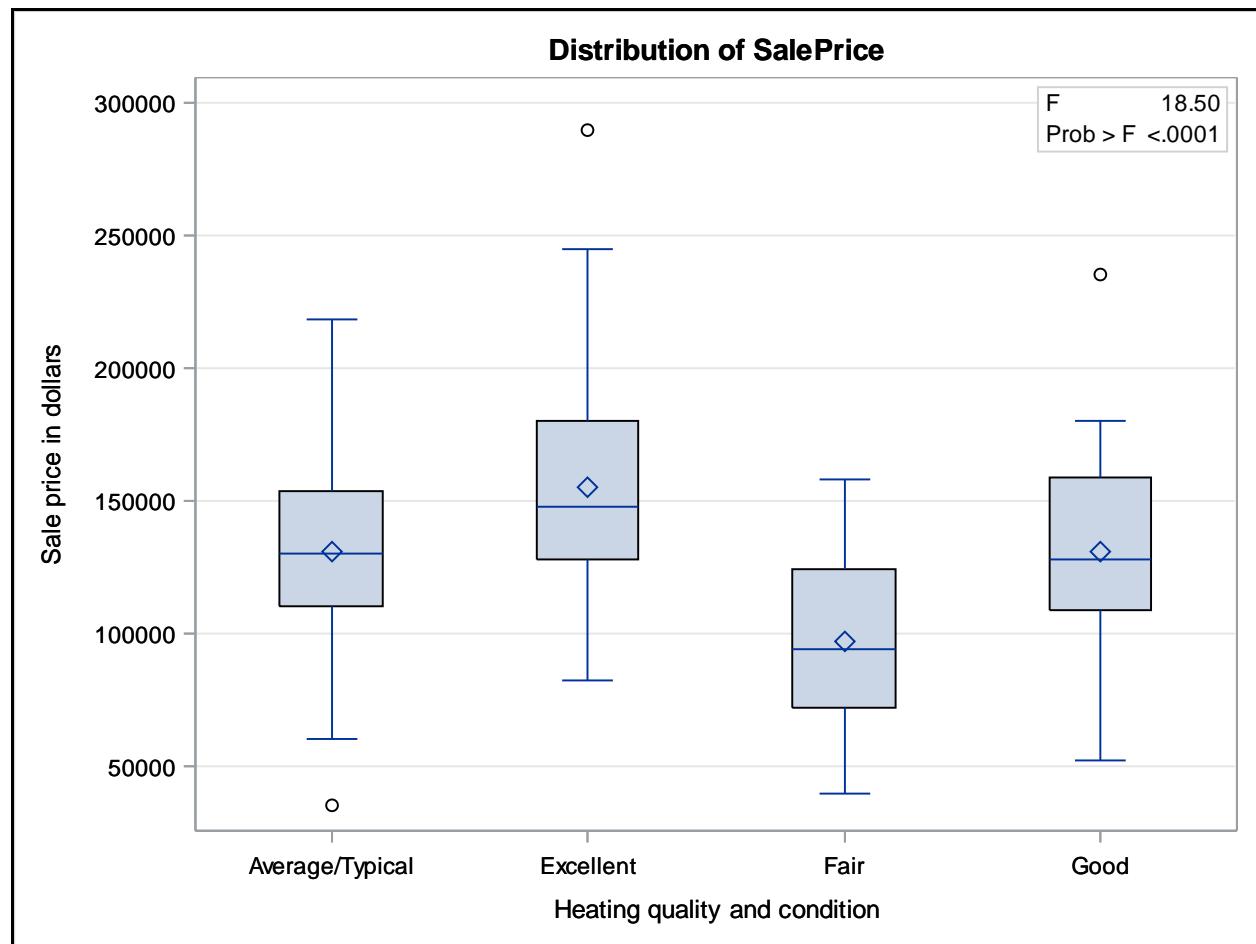
The plot in the upper left panel shows the residuals plotted against the fitted values from the ANOVA model. Essentially, you are looking for a random scatter within each group. Any patterns or trends in this plot can indicate model misspecification.

To check the normality assumption, look at the residual histogram and Q-Q plot, which are at the bottom left and middle left, respectively. The histogram is approximately symmetric. The data values in the quantile-quantile plot stay close to the diagonal reference line and give support to the assumption of normally distributed errors.

The default plot created with this code is a box plot.

Note: Notice that the ordering of the box plots uses the format label of the levels rather than the values in the data. In SAS Studio, the format line would need to be added to adjust the ordering.

```
format Heating_QC $Heating_QC.;
```



Potential outliers are evident in all groups, except for *Fair*.

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

Levene's Test for Homogeneity of SalePrice Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Heating_QC	3	5.931E18	1.977E18	0.58	0.6305
Error	296	1.014E21	3.426E18		

The output above is the result of the Hovtest option in the MEANS statement. The null hypothesis is that the variances are equal over all Heating_QC groups. The *p*-value of 0.6305 is not smaller than your alpha level of 0.05 and therefore you do not reject the null hypothesis. Another of your assumptions is met.

Note: At this point, if you determined that the variances were not equal, you could add the WELCH option to the MEANS statement. This requests Welch's (1951) variance-weighted one-way ANOVA. This alternative to the usual ANOVA is robust to the assumption of equal variances. This is similar to the unequal variance *t*-test for two populations. See the appendix for more information.

End of Demonstration

Analysis Plan for ANOVA – Summary

Null Hypothesis: All means are equal.

Alternative Hypothesis: At least one mean is different.

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

43

Sas

Garlic Example



45

Sas



Exercises

Example: Montana Gourmet Garlic is a company that grows garlic using organic methods. It specializes in hardneck varieties. Knowing a little about experimental methods, the owners design an experiment to test whether growth of the garlic is affected by the type of fertilizer used. They limit the experimentation to a Rocambole variety named Spanish Roja, and test three organic fertilizers and one chemical fertilizer (as a control). They blind themselves to the fertilizer by using containers with numbers 1 through 4. (In other words, they design the experiment in such a way that they do not know which fertilizer is in which container.) One acre of farmland is set aside for the experiment. It is divided into 32 beds. They randomly assign fertilizers to beds. At harvest, they calculate the average weight of garlic bulbs in each of the beds. The data are in the **STAT1.Garlic** data set.

These are the variables in the data set:

Fertilizer The type of fertilizer used (1 through 4)

BulbWt The average garlic bulb weight (in pounds) in the bed

BedID A bed identification number

1. Analysis of Variance with Garlic Data

Consider an experiment to study four types of fertilizer, labeled 1, 2, 3, and 4. One fertilizer is chemical and the rest are organic. You want to see whether the average of weights of garlic bulbs are significantly different for plants in beds using different fertilizers.

Test the hypothesis that the means are equal. Be sure to check that the assumptions of the analysis method that you choose are met. What conclusions can you reach at this point in your analysis?

End of Exercises

2.3 ANOVA Post Hoc Tests

Objectives

- Perform pairwise comparisons among groups after finding a significant effect of an independent variable in ANOVA.
- Demonstrate graphical features in PROC GLM for performing post hoc tests.
- Interpret a diffogram.
- Interpret a control plot.

48

Copyright © SAS Institute Inc. All rights reserved.



2.05 Multiple Choice Poll

With a fair coin, your probability of getting heads on one flip is 0.5. If you flip a coin once and got heads, what is the probability of getting heads on the second try?

- 0.50
- 0.25
- 0.00
- 1.00
- 0.75

49

Copyright © SAS Institute Inc. All rights reserved.



2.06 Multiple Choice Poll

With a fair coin, your probability of getting heads on one flip is 0.5. If you flip a coin twice, what is the probability of getting at least one head out of two?

- a. 0.50
- b. 0.25
- c. 0.00
- d. 1.00
- e. 0.75

51

Copyright © SAS Institute Inc. All rights reserved.



Multiple Comparison Methods

Number of Groups Compared	Number of Comparisons	Experimentwise Error Rate ($\alpha=0.05$)
2	1	.05
3	3	.14
4	6	.26
5	10	.40

Comparisonwise Error Rate = $\alpha = 0.05$

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

53

Copyright © SAS Institute Inc. All rights reserved.

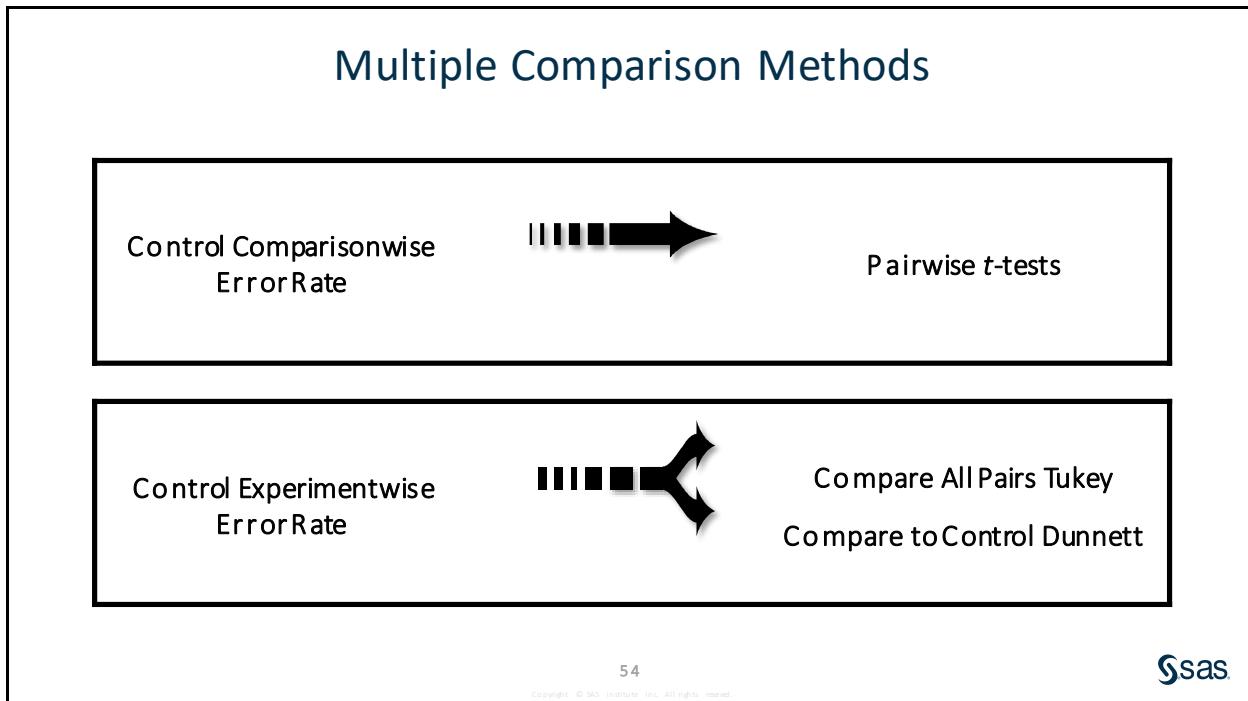


When you control the comparisonwise error rate (CER), you fix the level of alpha for a single comparison, without taking into consideration all the pairwise comparisons that you are making.

The experimentwise error rate (EER) uses an alpha that takes into consideration all the pairwise comparisons that you are making. Presuming no differences exist, the chance that you falsely conclude that **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.

Note: There is some disagreement among statisticians about whether and how to control the experimentwise error rate.



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

In order to call for the statistical hypothesis tests for group differences and ODS Statistical Graphics to support them, turn on ODS Graphics and then:

- For Comparisonwise Control LSMEANS / PDIFF=ALL ADJUST=T
- For Experimentwise Control LSMEANS / PDIFF=ALL ADJUST=TUKEY or
PDIFF=CONTROL('control level') ADJUST=DUNNETT

Note: Many other available options control the experimentwise error rate. For information about these options, see the SAS Documentation.

Note: One-tailed tests against a control level can be requested using the CONTROLL (lower tail) or CONTROLU (upper tail) options in the LSMEANS statement.

Tukey's HSD Test

HSD=Honest Significant Difference

This method is appropriate when you consider pairwise comparisons.

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.
- Also known as the Tukey-Kramer Test

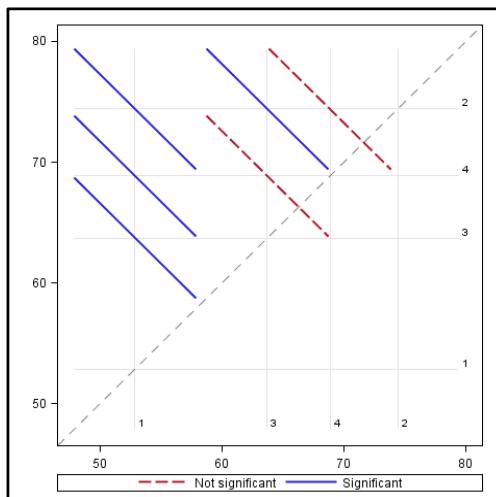
55



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 that 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



56



A *diffogram* can be used to quickly tell whether 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 vertical and horizontal lines drawn at group mean values. The downward-sloping diagonal lines show the confidence intervals for the differences. The upward-sloping line is a reference line showing where the group means would be equal. Intersection of the downward-sloping diagonal line for a pair with the upward-sloping, broken gray diagonal line implies that the confidence interval includes zero and that the mean difference between the two groups is not statistically significant. In that case, the diagonal line for the pair will be broken. If the confidence interval does not include zero, then the diagonal line for the pair will be solid. With ODS Statistical Graphics, these plots are automatically generated when you use the PDIFF=ALL option in the LSMEANS statement.

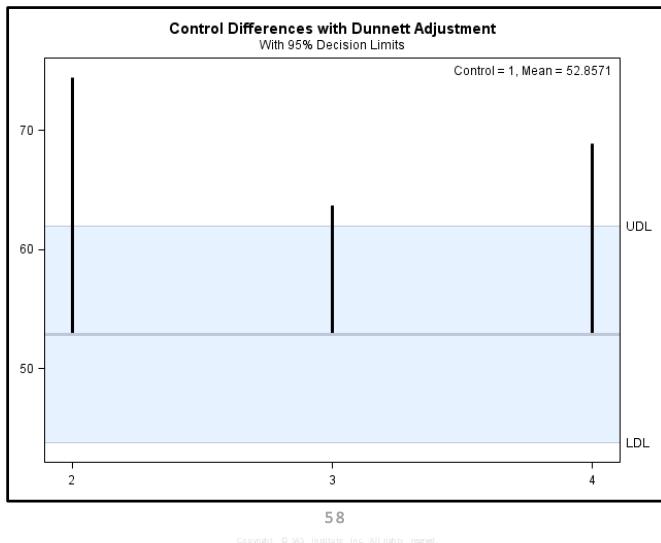
Special Case of Comparing to a Control

Comparing to a control is appropriate when there is a natural reference group, such as a placebo group in a drug trial.

- Experimentwise error rate is no greater than the stated alpha.
- Comparing to a control takes into account the correlations among tests.
- One-sided hypothesis tests against a control group can be performed.
- Control comparison computes and tests $k-1$ groupwise differences, where k is the number of levels of the CLASS variable.
- An example is the Dunnett method.

Dunnett's method is recommended when there is a true control group. When appropriate (when a natural control category exists, against which all other categories are compared) it is more powerful than methods that control for all possible comparisons. In order to do a one-sided test, use the option PDIFF=CONTROLL (for lower-tail tests when the alternative hypothesis states that a group's mean is less than the control group's mean) or PDIFF=CONTROLU (for upper-tail tests when the alternative hypothesis states that a group's mean is greater than the control group's mean).

Control Plots



LS-mean control plots are produced only when you specify PDIFF=CONTROL or ADJUST=DUNNETT in the LSMEANS statement, and in this case they are produced by default. The value of the control is shown as a horizontal line. The shaded area is bounded by the UDL and LDL (Upper Decision Limit and Lower Decision Limit). If the vertical line extends past the shaded area, that means that the group represented by that line is significantly different from the control group.



Post Hoc Pairwise Comparisons

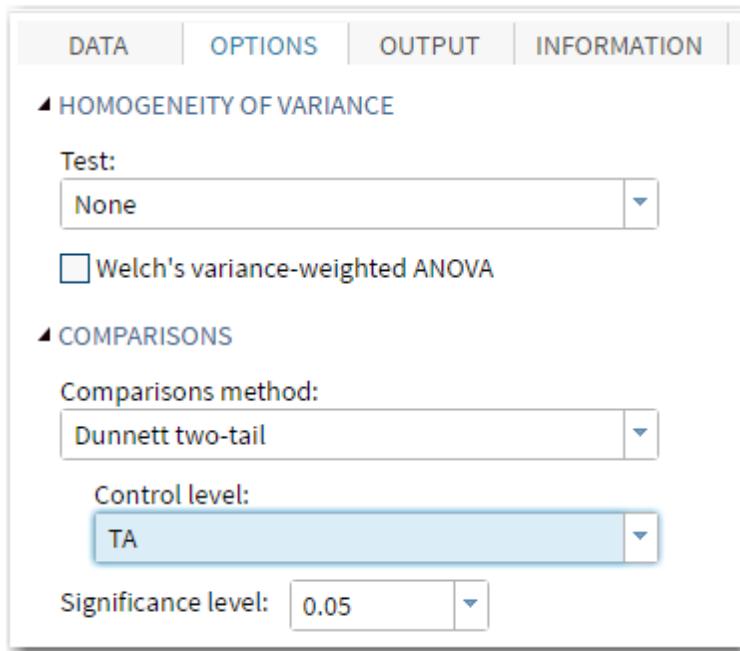
Example: Use the LSMEANS statement in PROC GLM to produce comparison information about the mean sale prices of the different heating system quality ratings.

1. Repen the **One-Way ANOVA** task from the previous demonstration.
2. Select the OPTIONS tab.
3. Set the HOMOGENEITY OF VARIANCE Test to **None**.
4. Check only the Display plot **LS-mean difference plot**.
5. Use the drop-down menu to choose **Tukey** as the comparisons method.

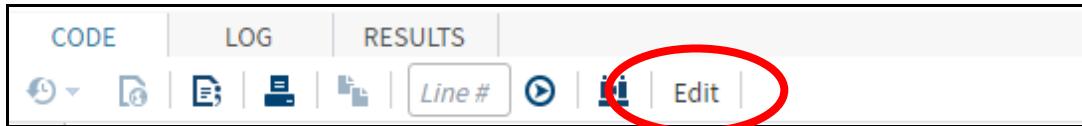
The screenshot shows the 'OPTIONS' tab selected in the SAS PROC GLM dialog box. The 'COMPARISONS' method is set to 'Tukey'. Under 'PLOTS', the 'LS-mean difference plot' is checked. Other plot options like 'Box plot' and 'Means plot' are unchecked.

Test:	None
Comparisons method:	Tukey
Significance level:	0.05
Display plots:	Selected plots
LS-mean difference plot	<input checked="" type="checkbox"/>
Box plot	<input type="checkbox"/>
Means plot	<input type="checkbox"/>
Diagnostics plot	<input type="checkbox"/>
Maximum number of plot points:	Default(5,000)

6. To use Dunnett's method, choose the **Dunnett two-tail** option and specify the **Control level** using the drop-down menu.



7. Typically, only one type of multiple comparison method would be used, and SAS Studio conducts one comparisons method at a time. To produce output for multiple comparisons methods, one can run the tasks separately or edit the generated code manually to include multiple comparisons statements. In the code window, click **Edit** to add in the code for the second comparisons method.



8. Edit the generated code by adding multiple **lsmeans** statements. The following edited code will provide comparison information using both Tukey's HSD Test and Dunnett's method.

```
Title;
ods noproctitle;
ods graphics / imagemap=on;

proc glm data=STAT1.AMESHOUSING3;
  class Heating_QC;
  model SalePrice=Heating_QC;
  means Heating_QC / hovtest=levene plots=none;
  lsmeans Heating_QC / adjust=tukey pdiff alpha=.05
plots=(diffplot);
  lsmeans Heating_QC / adjust=dunnett pdiff=control('TA')
alpha=.05 plots=(controlplot);
run;
quit;
```

Note: Alternatively, one can write the code directly altogether.

```

/*st102d03.sas*/
ods graphics;

ods select lsmeans diff diffplot controlplot;
proc glm data=STAT1.ameshousing3
    plots(only)=(diffplot(center) controlplot);
    class Heating_QC;
    model SalePrice=Heating_QC;
    lsmeans Heating_QC / pdiff=all
        adjust=tukey;
    lsmeans Heating_QC / pdiff=control('Average/Typical')
        adjust=dunnett;
    format Heating_QC $Heating_QC.;
    title "Post-Hoc Analysis of ANOVA - Heating Quality as
Predictor";
run;
quit;

```

Multiple LSMEANS statements are permitted, although typically only one type of multiple comparison method would be used for each LSMEANS effect. Two different methods are shown for illustration here. You might use Dunnett comparisons if you are interested only in knowing whether any of the heating quality levels differed from **Average/Typical**.

Note: In the code above, because **Heating_QC** uses a format, the formatted value, rather than the internal coded value, must be specified as the control level in the second LSMEANS statement,

Selected PLOTS= options:

CONTROLPLOT requests a display in which least squares means are compared against a reference level. LS-mean control plots are produced only when you specify PDIFF=CONTROL or ADJUST=DUNNETT in the LSMEANS statement, and in this case they are produced by default.

DIFFPLOT modifies the diffogram produced by an LSMEANS statement with the PDIFF=ALL option (or only PDIFF, because ALL is the default argument). The CENTER option marks the center point for each comparison. This point corresponds to the intersection of two least squares means.

Selected LSMEANS statement options:

PDIFF= requests *p*-values for the differences, which is 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. For details, see the documentation for LSMEANS under the GLM procedure.

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

Note: 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 following output is for the Tukey LSMEANS comparisons.

The GLM Procedure

Least Squares Means

Adjustment for Multiple Comparisons: Tukey-Kramer

Heating_QC	SalePrice LSMEAN	LSMEAN Number
Average/Typical	130573.529	1
Excellent	154919.187	2
Fair	97118.750	3
Good	130844.086	4

Least Squares Means for effect Heating_QC Pr > |t| for H0: LSMean(i)=LSMean(j)

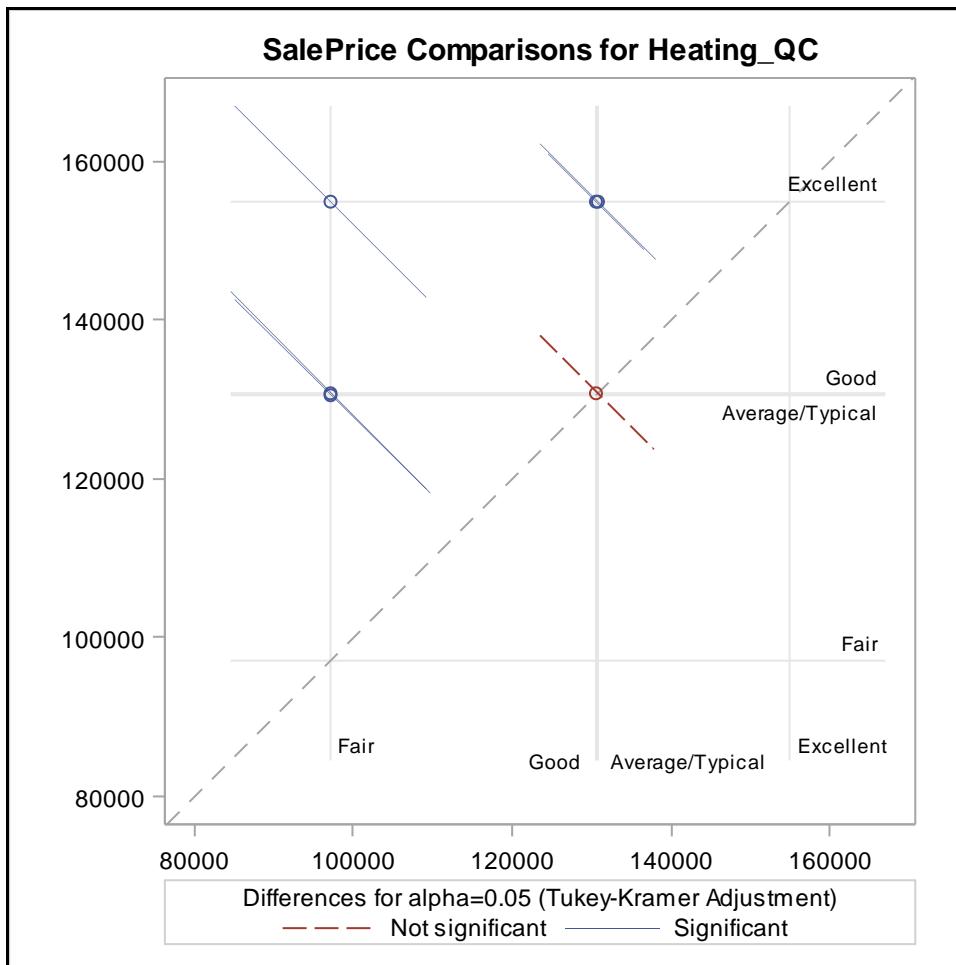
Dependent Variable: SalePrice

i/j	1	2	3	4
1		<.0001	0.0020	1.0000
2	<.0001		<.0001	0.0002
3	0.0020	<.0001		0.0037
4	1.0000	0.0002	0.0037	

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 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 on, are blank, because it does not make any sense to compare a mean to itself.

The only nonsignificant pairwise difference is between *Average/Typical* and *Good*.

The Least Square Means are shown graphically in the mean plot. The Tukey-adjusted differences among the LSMEANS are shown in the diffogram.



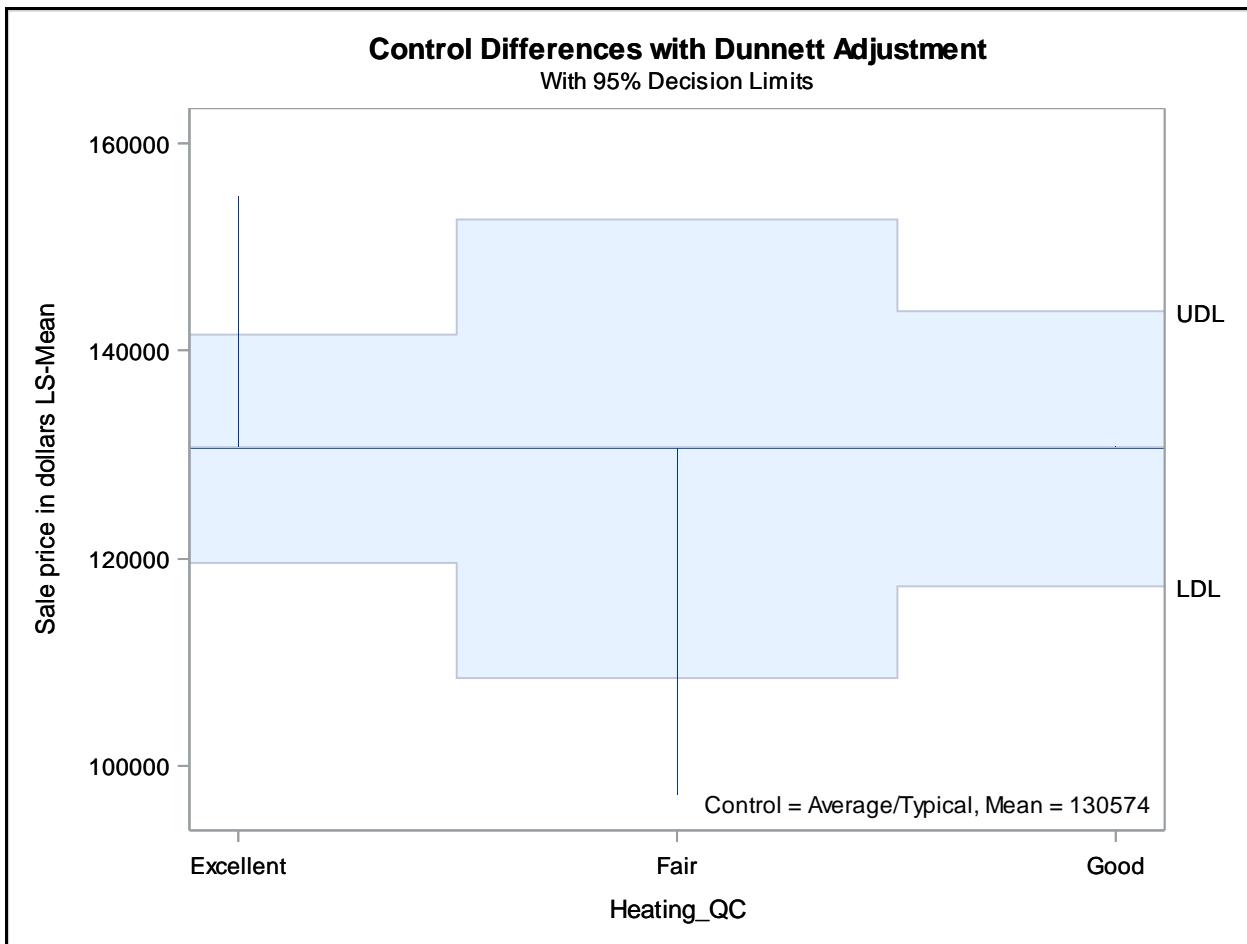
The solid line denotes significant differences between heating quality levels. (Confidence intervals for the difference do not cross the diagonal equivalence line.)

The following output is for the Dunnett LSMEANS comparisons:

Heating_QC	SalePrice LSMEAN	H0:LSMean=Control	
			Pr > t
Average/Typical	130573.529		
Excellent	154919.187	<.0001	
Fair	97118.750	0.0010	
Good	130844.086	0.9999	

In this case, all other quality levels are compared to **Average/Typical**. Once again, **Good** is the only level that is not statistically significantly different from that control level.

The Control plot is below:



This plot corresponds to the tables that were summarized. The horizontal line is drawn at the least squared mean for *Average/Typical*, which is 130574. The three other means are represented by the ends of the vertical lines extending from the horizontal control line. The mean value for *Good* is so close to *Average/Typical* that it cannot be seen here.

Note: Notice that the blue areas of non-significance vary in size. This is because different comparisons involve different sample sizes. Smaller sample sizes require larger mean differences to reach statistical significance.

End of Demonstration



Exercises

2. Post Hoc Pairwise Comparisons

Consider again the analysis of the **STAT1.Garlic** data set. There was a statistically significant difference among means for sales for the different fertilizers. Perform a post hoc test to look at the individual differences among means.

- a. Conduct pairwise comparisons with an experimentwise error rate of $\alpha=0.05$. (Use the Tukey adjustment.) Which types of fertilizer are significantly different?
- b. Use level 4 (the chemical fertilizer) as the control group and perform a Dunnett comparison with the organic fertilizers to see whether they affected the average weights of garlic bulbs differently from the control fertilizer.
- c. (Extra) Perform unadjusted tests of all pairwise comparisons to see what would have happened if the multi-test adjustments had not been made. How do the results compare to what you saw in the Tukey adjusted tests?

End of Exercises

2.4 Pearson Correlation

Objectives

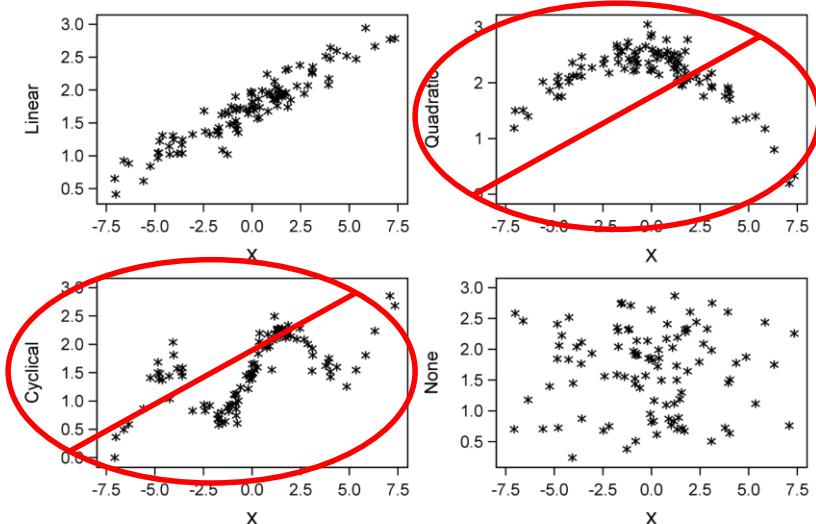
- Use a scatter plot to examine the linear relationship between two continuous variables.
- Use correlation statistics to quantify the degree of association between two continuous variables.
- Describe potential misuses of the correlation coefficient.
- Use the CORR procedure to obtain Pearson correlation coefficients.

63



Copyright © SAS Institute Inc. All rights reserved.

Pearson Correlation – Linear Relationships



64



Copyright © SAS Institute Inc. All rights reserved.

Recall that you can visualize the relationship between variables using a scatter plot. If both variables are measured on a continuous scale, you can also see whether you can detect a pattern in the relationship. Before you perform any statistical analysis, it is important to first understand the nature of the

relationship.

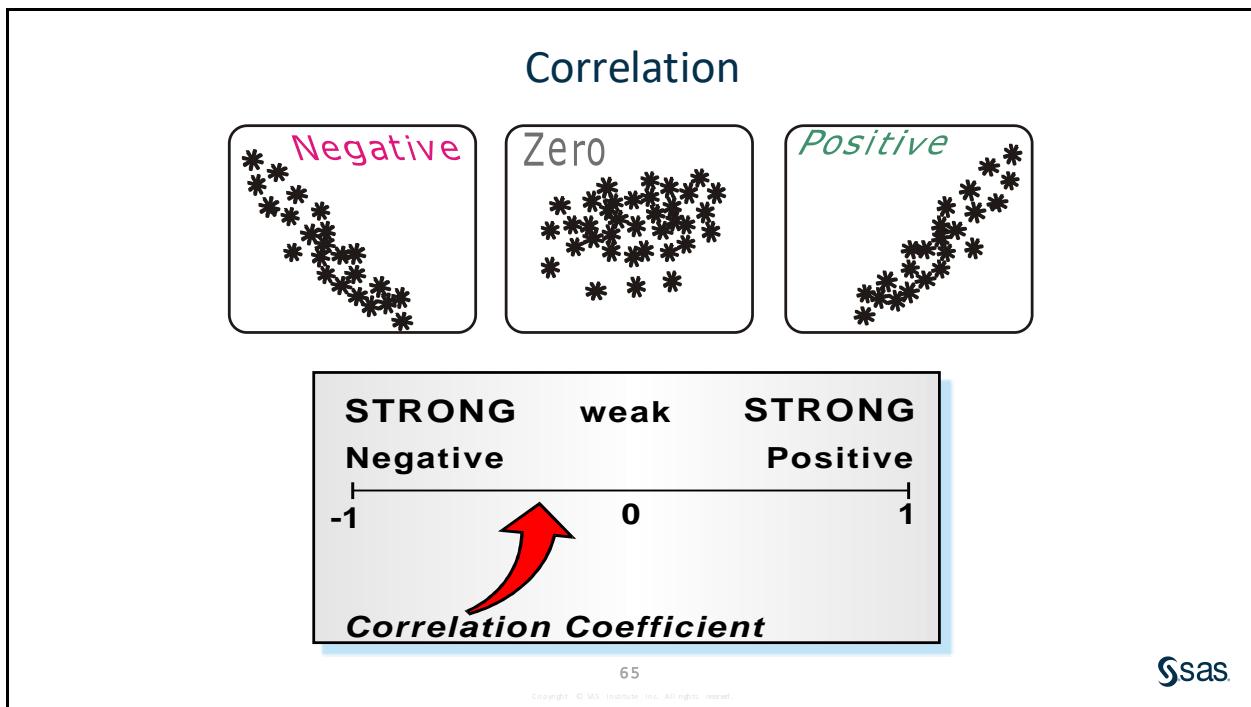
The Pearson product moment correlation is a measure of the linear relationship between two continuous variables. The formula for the population correlation is

$$\rho_{xy} = \frac{Cov(x, y)}{\sqrt{V(x)V(y)}} = \frac{E((x - E(x))(y - E(y)))}{\sqrt{E(x - E(x))^2 E(y - E(y))^2}}$$

The formula for the sample correlation is

$$r_{xy} = \frac{\sum_i ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

If a scatter plot of the relationship is distinctly non-linear, then this measure of association is invalid.



As you examine the scatter plot, you can find evidence of the nature of the correlation between the variables.

Values of the Pearson product-moment correlation 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 (upward-sloping) linear association
- less than 0 if there is a negative (downward-sloping) linear association

Hypothesis Test for a Correlation

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

66

Copyright © SAS Institute Inc. All rights reserved.



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 is the case with many statistics) because of very large sample sizes. For example, even a correlation coefficient 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 whether it is meaningfully large.

2.07 Multiple Answer Poll

Which of the following statements is/are true?

- A Pearson correlation coefficient is a measure of linear association.
- A nonsignificant p -value for a Pearson correlation means no relationship.
- A negative Pearson correlation indicates a low degree of linear association.
- A random cloud of data implies a negative correlation.

67

Copyright © SAS Institute Inc. All rights reserved.

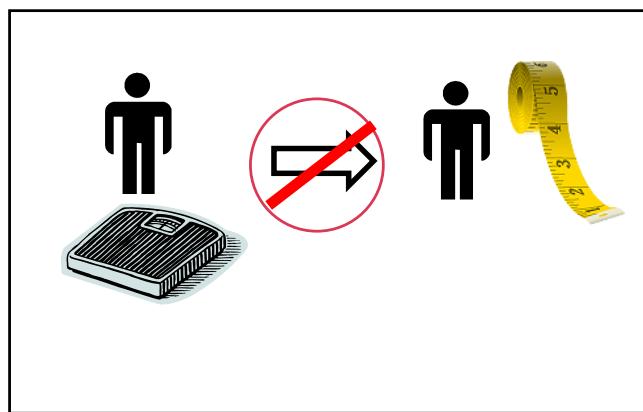


Cautions in Interpreting Correlations

- Correlation between X and Y does not imply causation because:
 - It does not imply the direction of any possible causal relationship between X and Y
 - There might be a third variable that is the cause of changes in both variables, which means the association between X and Y is only indirect.
- The Pearson Correlation only measures the linear association between X and Y.

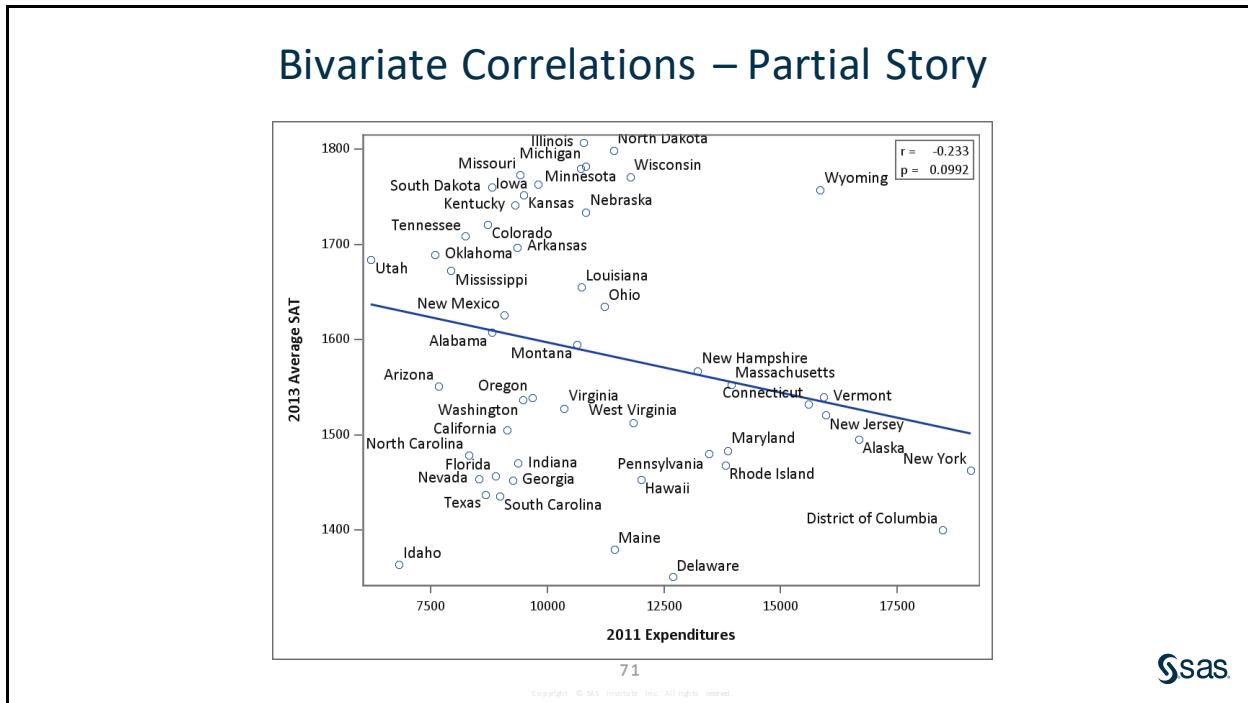
The next several slides describe in more detail some cautions in interpreting the Pearson correlation coefficient.

Correlation versus Causation



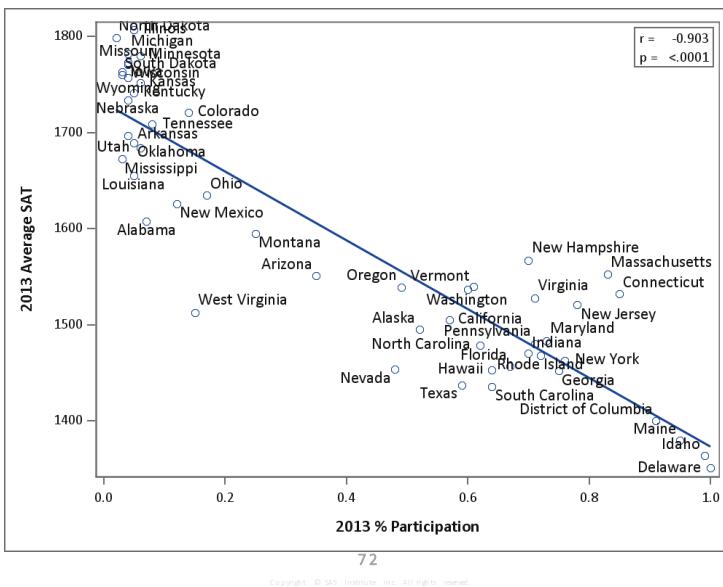
Common errors can be made when you interpret the correlation between variables. One example of this is using correlation coefficients to conclude a cause-and-effect relationship.

- 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.”



Bivariate (two-variable) correlations describe the measurable degree of linear association between the variables involved. However, often the relationship is just an artifact of both variables' relationships with some third variable. An example of reaching errant conclusions comes from U.S. Department of Education data from the Scholastic Aptitude Test (SAT) from 2013. The scatter plot above shows each state's average total SAT score versus the average state expenditure in 2011 in U.S. dollars per public school student. The correlation between the two variables is -0.233 . Looking at the plot and at this statistic, you might be led to the non-intuitive conclusion that state spending on education hurts student performance. While the calculated correlation statistic is factual, the simplicity of the relationship implied by it is not.

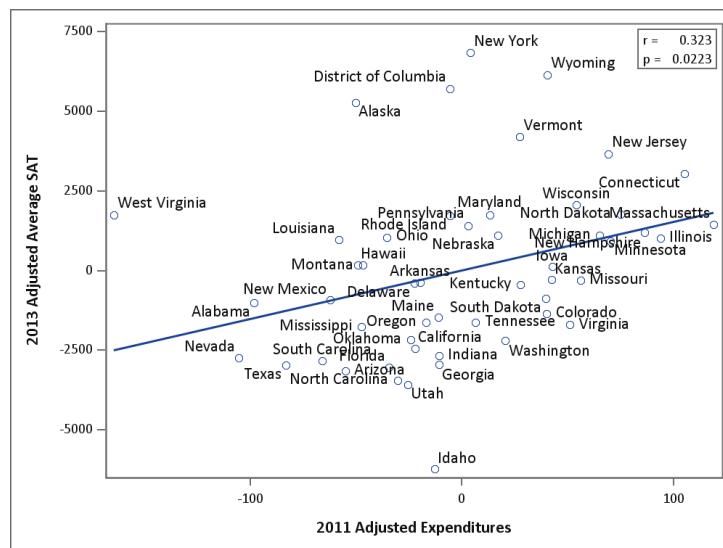
Missing Link



sas

The 2013 report did not take into account the differences among the states in the percentage of students taking the SAT. There are many reasons for the varying participation rates. Some states have lower participation because their students primarily take the rival ACT standardized test. Others have rules requiring even non-college-bound students to take the test. In low participating states, often only the highest performing students choose to take the SAT. Another reported table shows the relationship between participation rate (percent taking the SAT) and average SAT total score. The correlation is -0.903 , indicating that states with lower participation rates tend to have higher average scores.

The Truer Story



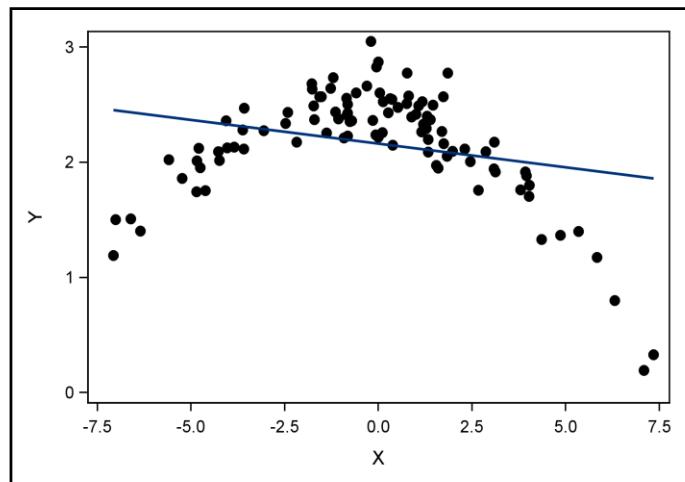
sas

If you adjust for differences in participation rates, the conclusions about the effect of expenditures might change. In this case, there seems to be a slight positive linear relationship between expenditures and average total score on the SAT when you first adjust for participation rates.

Simple correlations often do not tell the whole story.

Note: These types of adjustments are described in greater detail in the sections about multiple regression.

Missing Another Type of Relationship



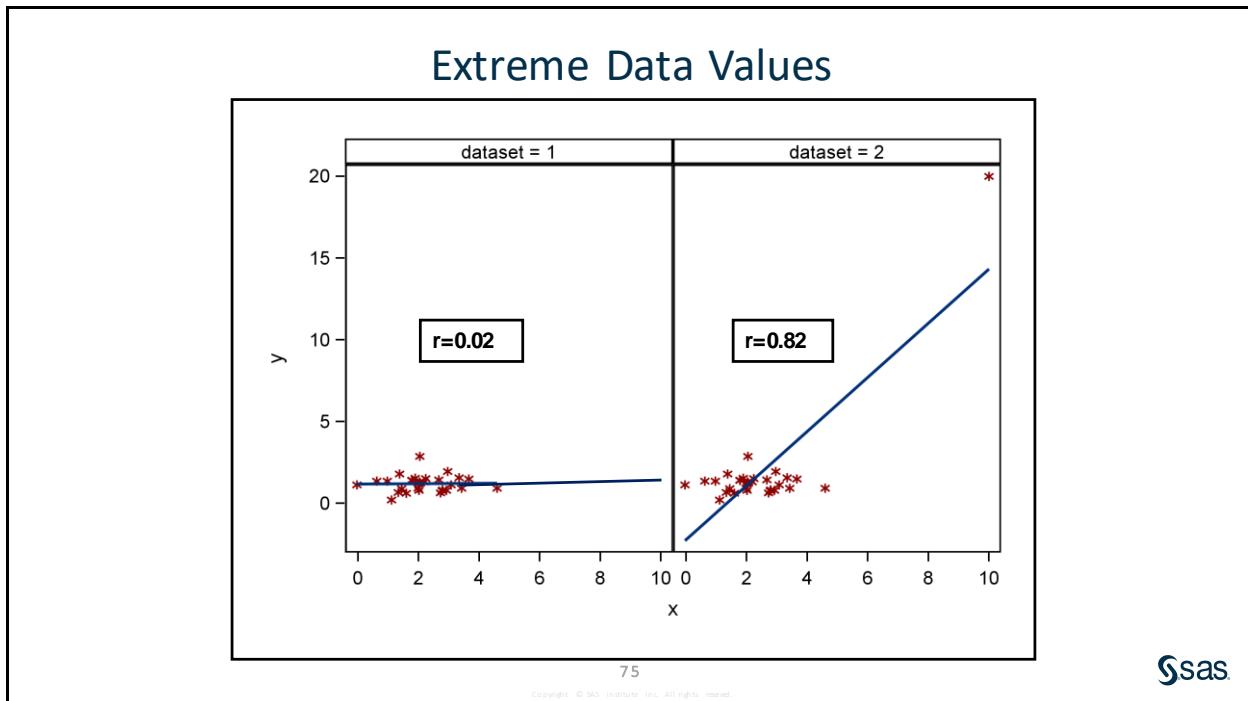
74

Copyright © SAS Institute Inc. All rights reserved.

In the scatter plot, 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 that there is no relationship of any kind between the two variables.

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



Correlation coefficients are highly affected by a few extreme values on either variable's range. The scatter plots show 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;
```

76

Copyright © SAS Institute Inc. All rights reserved.



You can use the CORR procedure to produce correlation statistics and scatter plots for your data. By default, PROC CORR produces Pearson correlation coefficients 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 specifies one or more additional tip variables to identify observations in scatter plots and scatter plot matrices.

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.



Data Exploration, Correlations, and Scatter Plots

Examine the relationships between **SalePrice** and the continuous predictor variables in the data set. Use the CORR procedure.

1. Use the **Correlation Analysis** Task under Statistics.
2. Select the **AmesHousing3** data set. Assign the continuous variables to **Analysis variables** and **SalePrice** as the variable to **Correlate with**.

The screenshot shows the SAS Studio interface with the following details:

- Left Sidebar (Tasks and Utilities):**
 - My Tasks**
 - Tasks** (expanded)
 - Data**
 - Graph**
 - Combinatorics and Probability**
 - Statistics** (expanded)
 - Data Exploration**
 - Summary Statistics**
 - Distribution Analysis**
 - One-Way Frequencies**
 - Correlation Analysis** (selected)
 - Table Analysis**
 - t Tests**
 - One-Way ANOVA**
 - Nonparametric One-Way ANOVA**
 - N-Way ANOVA**
 - Analysis of Covariance**
 - Linear Regression**
 - Binary Logistic Regression**
 - Predictive Regression Models**
 - Generalized Linear Models**
 - Mixed Models**
 - Partial Least Squares Regression**

3. On the OPTIONS tab, in the STATISTICS area, in the **Display statistics** drop-down menu, choose the **Selected statistics** option and check the **Correlations** and **Display p-value** boxes (they might already be checked), as well as **Descriptive statistics**.

4. Under the **PLOTS** section, under the **Type of plot:** property choose the **Individual scatter plots** option and check the **include inset statistics** box. Change **Number of variables to plot** to **8** to generate plots for all 8 variables.

5. Submit the code

SAS Studio produces the following code:

```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=STAT1.AMESHOUSING3 pearson
plots=scatter(ellipse=none nvar=8 nwith=8);
  var Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
    Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
  with SalePrice;
run;
```

Note: Alternatively, you can write the code directly in SAS Program.

```
/*st102d04.sas*/ /*Part A*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
  Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

ods graphics / reset=all imagemap;
proc corr data=STAT1.AmesHousing3 rank
  plots(only)=scatter(nvar=all ellipse=none);
  var &interval;
  with SalePrice;
  id PID;
  title "Correlations and Scatter Plots with SalePrice";
run;
```

Note: IMAGEMAP=ON in the ODS GRAPHICS statement enables tooltips to be used in HTML output. Tooltips are also functional in SAS Report output when you use SAS Enterprise Guide, starting with Version 4.3. Tooltips enable the user to identify data points by moving the cursor 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.

Selected PROC CORR statement options:

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

PLOTS creates scatter plots and scatter plot matrices using ODS GRAPHICS.

Selected PROC CORR statement:

ID when used in HTML output with IMAGEMAP, adds the listed variables to the information available with tooltips.

Suboptions for the PLOTS option:

SCATTER generates scatter plots for pairs of variables.

Suboptions for the SCATTER suboption:

NVAR=<k> specifies the maximum number of variables in the VAR list to be displayed in the matrix plot. If NVAR=ALL is specified, then all variables in the VAR list (up to a limit of 10) are displayed.

ELLIPSE=NONE suppresses the drawing of confidence ellipses on scatter plots.

The tabular output from PROC CORR is shown below. By default, the analysis generates a table of univariate statistics for the analysis variables and then a table of correlations and *p*-values.

PROC CORR Output

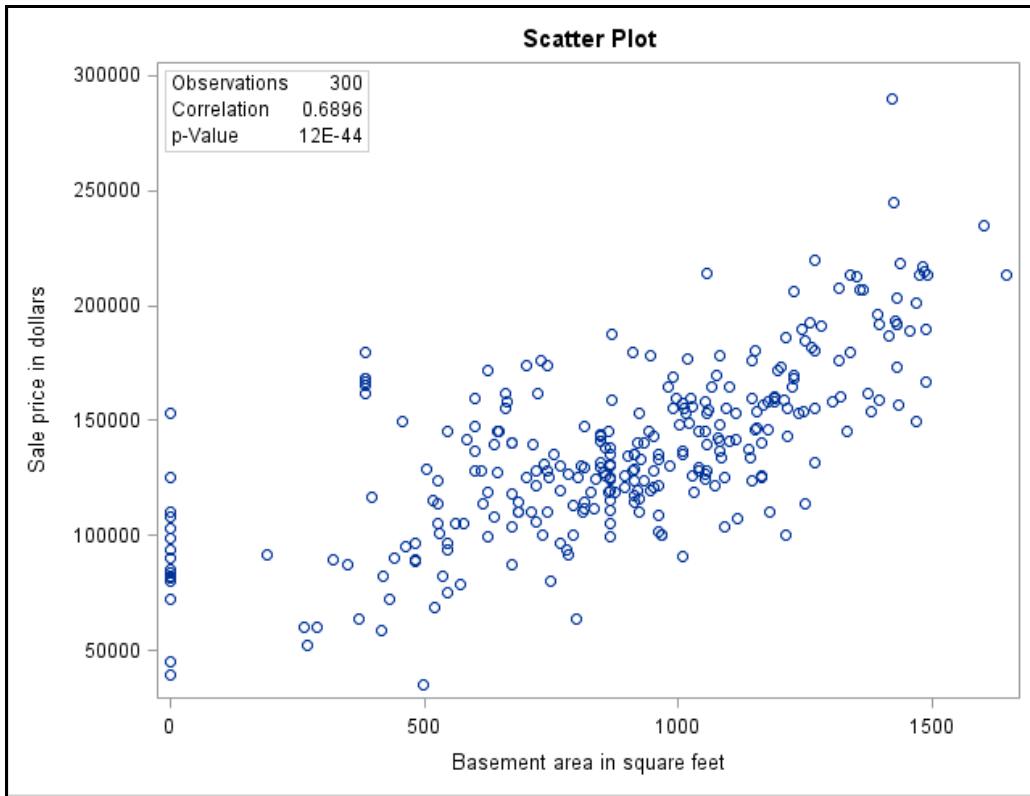
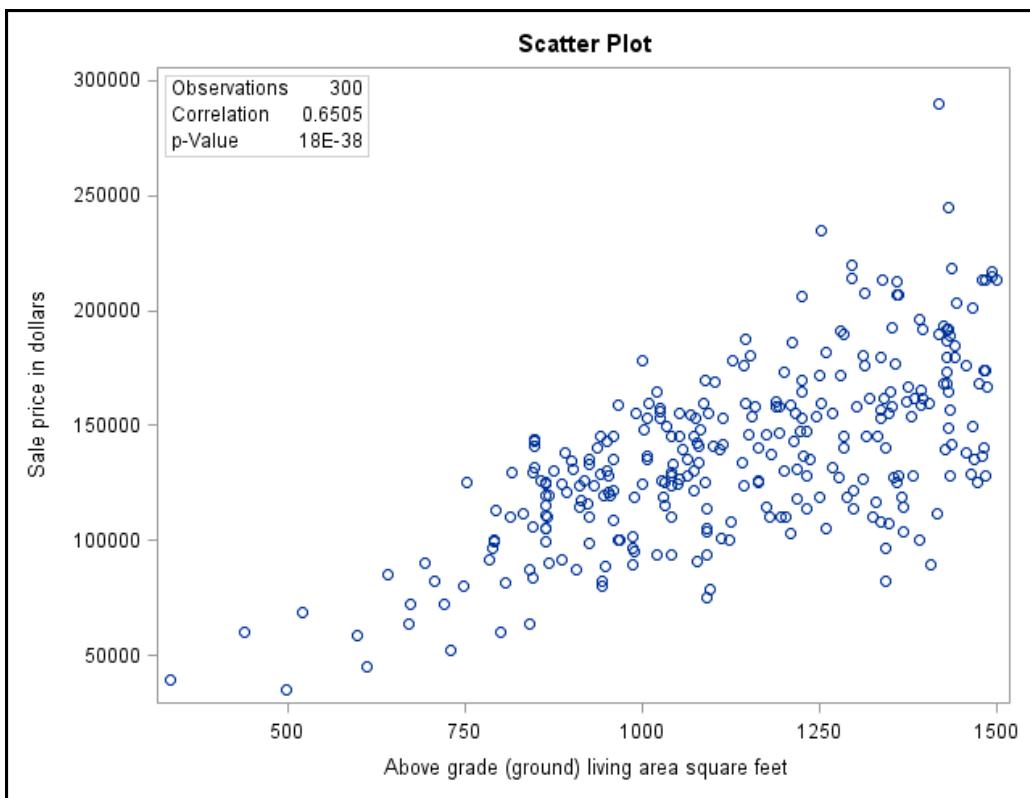
1 With Variables:	SalePrice
8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
SalePrice	300	137525	37623	41257460	35000	290000	Sale price in dollars
Gr_Liv_Area	300	1131	232.64939	339222	334.00000	1500	Above grade (ground) living area in square feet
Basement_Area	300	882.31000	359.78397	264693	0	1645	Basement area in square feet
Garage_Area	300	369.45333	176.25309	110836	0	902.00000	Size of garage in square feet
Deck_Porch_Area	300	118.26333	132.61169	35479	0	897.00000	Total area of decks and porches in square feet
Lot_Area	300	8294	3324	2488241	1495	26142	Lot size in square feet
Age_Sold	300	45.88667	27.47697	13766	1.00000	135.00000	Age of house when sold, in years
Bedroom_AbvGr	300	2.51333	0.69144	754.00000	0	4.00000	Bedrooms above grade
Total_Bathroom	300	1.70167	0.65707	510.50000	1.00000	4.10000	Total number of bathrooms (half bathrooms counted 10%)

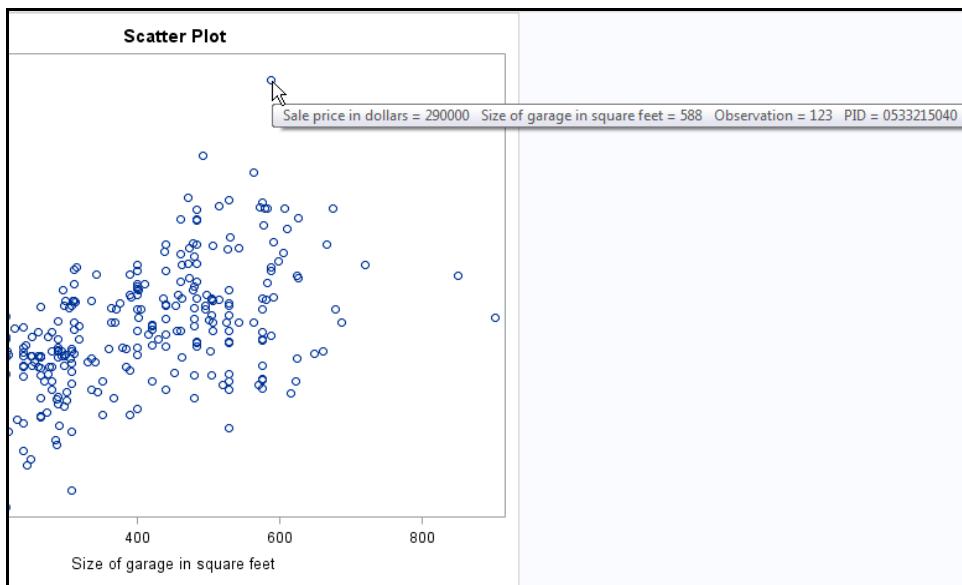
Pearson Correlation Coefficients, N = 300 Prob > r under H0: Rho=0									
SalePrice Sale price in dollars	Basement_Area 0.68956 <.0001	Gr_Liv_Area 0.65046 <.0001	Age_Sold -0.61542 <.0001	Total_Bathroom 0.60043 <.0001	Garage_Area 0.57892 <.0001	Deck_Porch_Area 0.43989 <.0001	Lot_Area 0.25335 <.0001	Bedroom_AbvGr 0.16594 0.0040	

The correlation coefficient between **SalePrice** and **Basement_Area** is 0.68956. The *p*-value is small, which indicates that the population correlation coefficient (*p*) is likely different from 0. The second largest correlation coefficient, in absolute value, is **Gr_Liv_Area**, at 0.65046.

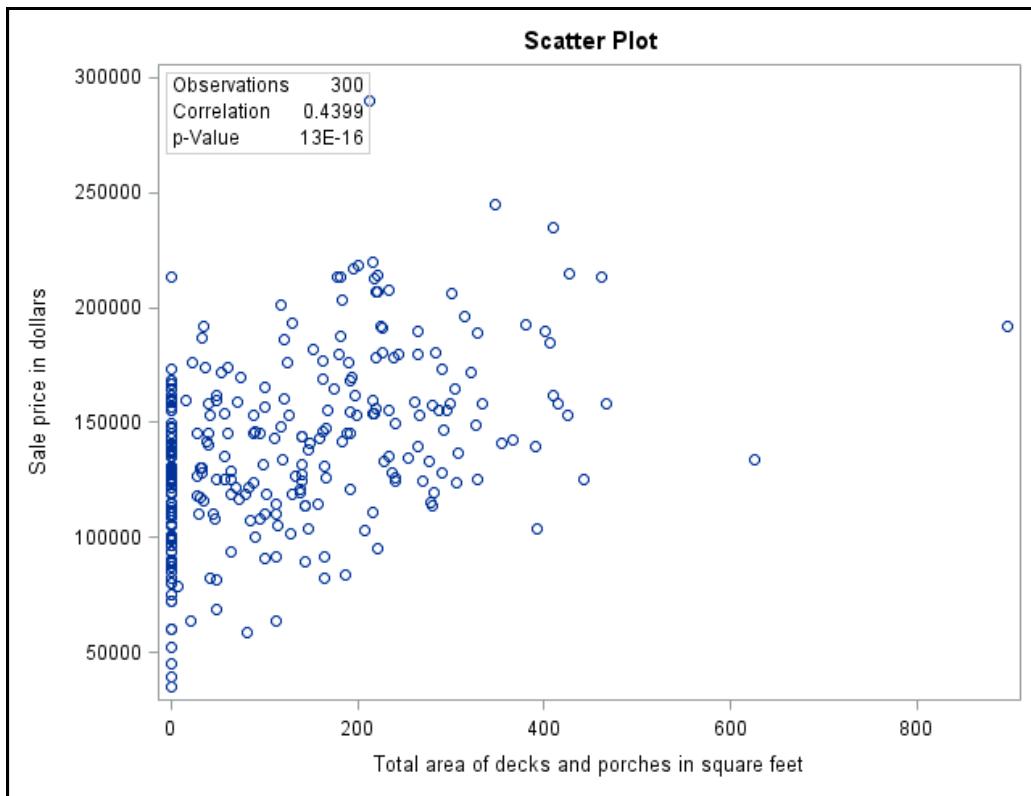
Scatter plots associated with these correlations are shown below.

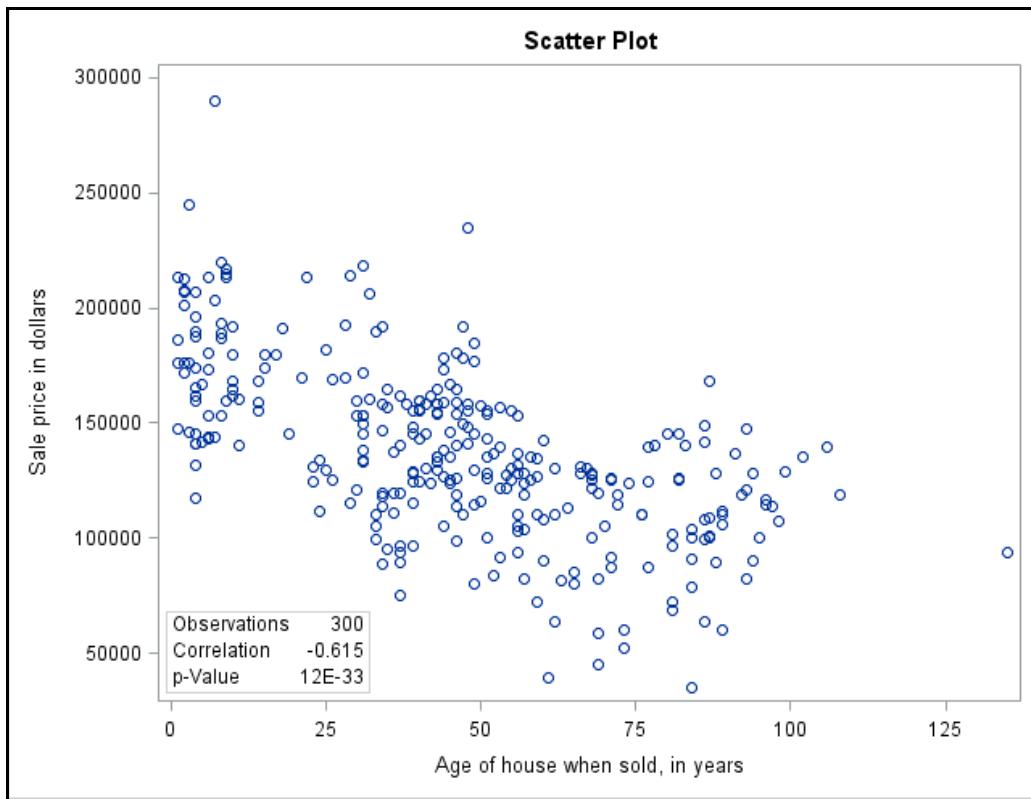
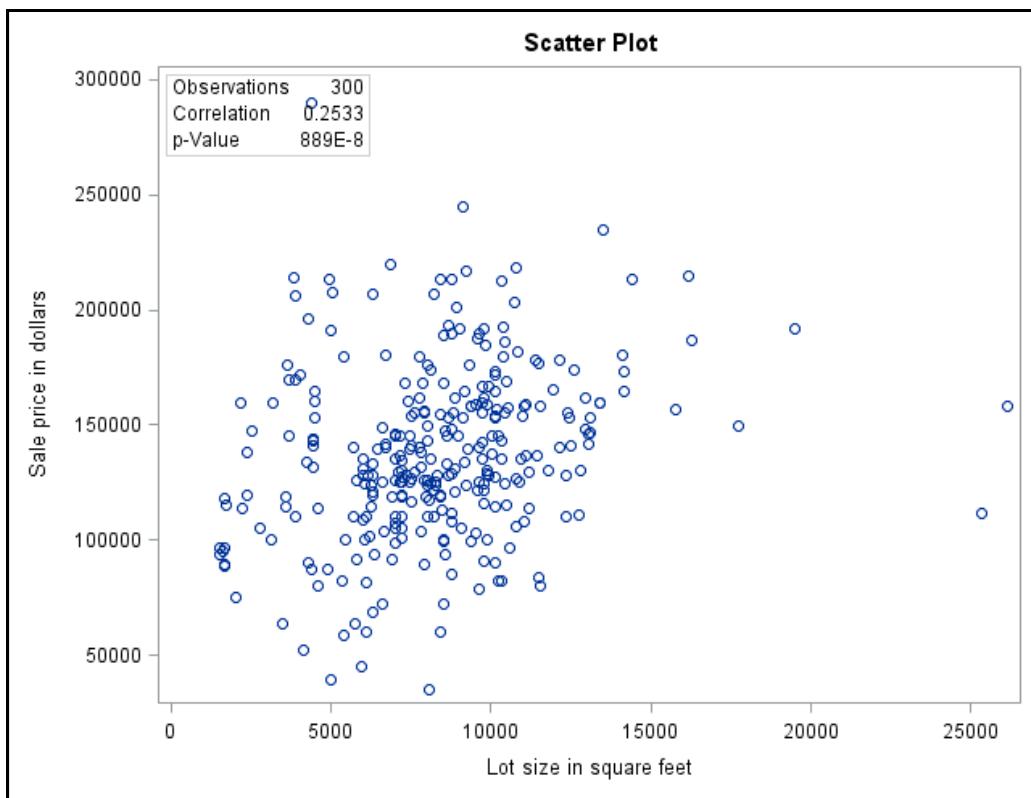


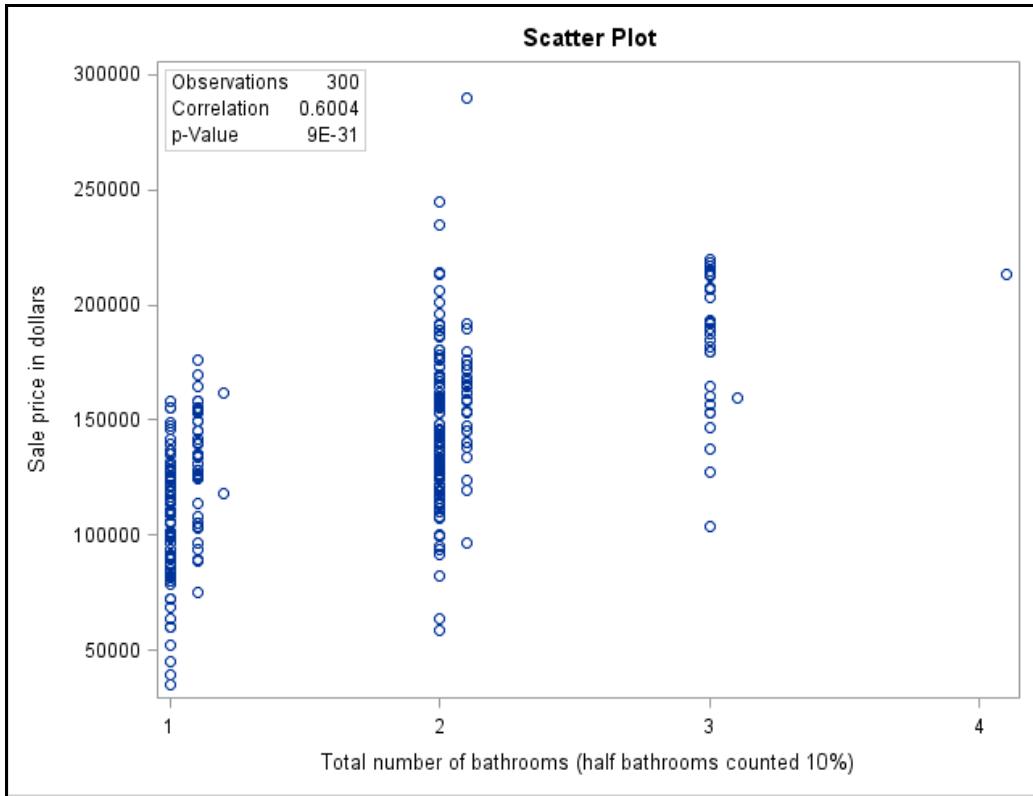
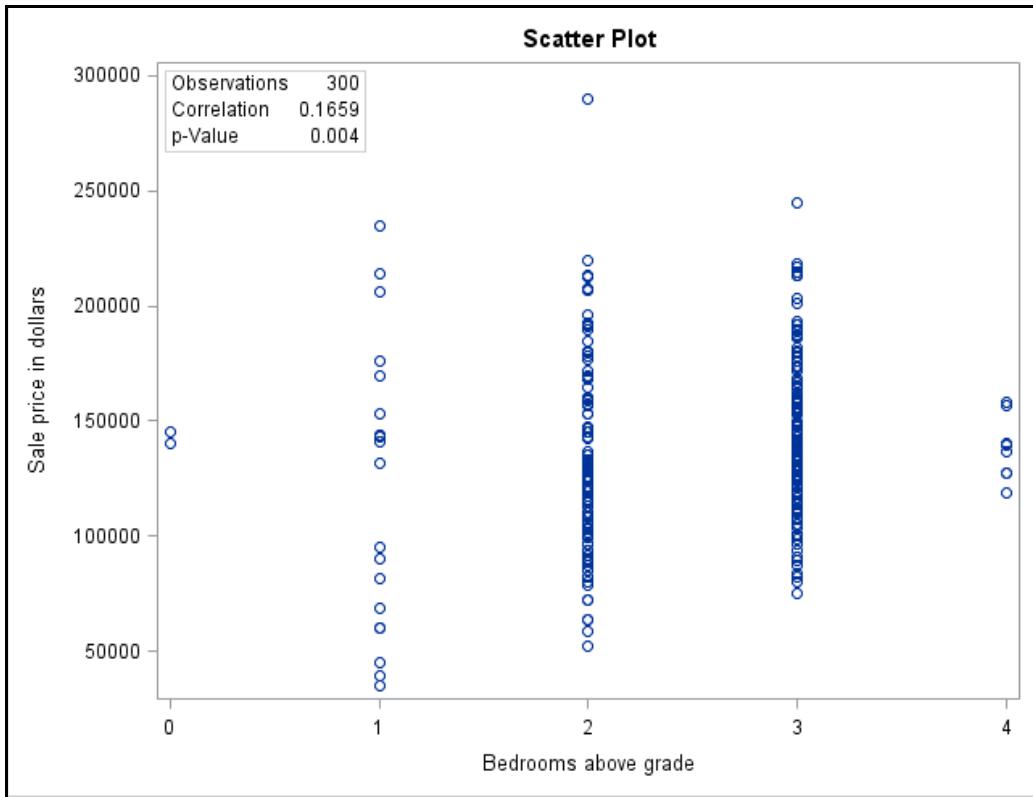
Notice that there are several houses with basements sized 0 square feet. These are houses without basements. This mixture of data can affect the correlation coefficient. You will need to take this into account later when you build a model with basement area as a predictor variable.



If you want to explore an observation further, you can move the cursor over the observation and information is displayed in a floating box. You can only do this in an HTML file with IMAGEMAP turned on. The coordinate values, observation number, and ID variable values are displayed.







The correlation and scatter plot analyses indicate that several variables might be good predictors for **SalePrice**.

When you prepare to conduct a regression analysis, it is always good practice to examine the correlations among the potential predictor variables. When you do not specify a WITH statement, you get a matrix of correlations of all VAR variables. That matrix can be very big and difficult to interpret. To limit the displayed output to only the strongest correlations, you can use a BEST= option.

- The BEST= option is not supported by SAS Studio. One can either manually edit the code produced by SAS Studio or write the code directly as follow.

```
/*st102d04.sas*/ /*Part B*/
ods graphics off;
proc corr data=STAT1.AmesHousing3
    nosimple
    best=3;
var &interval;
title "Correlations and Scatter Plot Matrix of Predictors";
run;
```

Selected PROC CORR statement option:

NOSIMPLE	suppresses printing simple descriptive statistics for each variable.
BEST=	Prints the n highest correlation coefficients for each variable, $n \geq 1$.
PLOTS(MAXPOINTS=n)	The global plot option MAXPOINTS= specifies that plots with elements that require processing more than n points be suppressed. The default is MAXPOINTS=5000. This limit is ignored if you specify MAXPOINTS=NONE.

PROC CORR Output

Variables:	Gr_Liv_Area	Basement_Area	Garage_Area	Deck_Porch_Area	Lot_Area	Age_Sold	Bedroom_AbvGr
	Total_Bathroom						

Pearson Correlation Coefficients, N = 300 Prob > r under H0: Rho=0			
Gr_Liv_Area Above grade (ground) living area square feet	Gr_Liv_Area 1.00000	Bedroom_AbvGr 0.48431 <.0001	Basement_Area 0.43985 <.0001
Basement_Area Basement area in square feet	Basement_Area 1.00000	Total_Bathroom 0.48500 <.0001	Gr_Liv_Area 0.43985 <.0001
Garage_Area Size of garage in square feet	Garage_Area 1.00000	Age_Sold -0.41346 <.0001	Total_Bathroom 0.36876 <.0001
Deck_Porch_Area Total area of decks and porches in square feet	Deck_Porch_Area 1.00000	Basement_Area 0.33689 <.0001	Gr_Liv_Area 0.28058 <.0001
Lot_Area Lot size in square feet	Lot_Area 1.00000	Bedroom_AbvGr 0.29801 <.0001	Basement_Area 0.27198 <.0001
Age_Sold Age of house when sold, in years	Age_Sold 1.00000	Total_Bathroom -0.52889 <.0001	Garage_Area -0.41346 <.0001

Pearson Correlation Coefficients, N = 300 Prob > r under H0: Rho=0			
Bedroom_AbvGr Bedrooms above grade	Bedroom_AbvGr 1.00000	Gr_Liv_Area 0.48431 <.0001	Lot_Area 0.29801 <.0001
Total_Bathroom Total number of bathrooms (half bathrooms counted 10%)	Total_Bathroom 1.00000	Age_Sold -0.52889 <.0001	Basement_Area 0.48500 <.0001

There are moderately strong correlations between **Total_Bathroom** and **Age_Sold** (-0.52889), between **Total_Bathroom** and **Basement_Area** (0.48500), and between **Bedroom_AbvGr** and **Gr_Liv_Area** (0.48431).

End of Demonstration

Body Fat Example



sas.



Exercises

3. Describing the Relationship between Continuous Variables

Percentage of body fat, age, weight, height, and 10 body circumference measurements (for example, abdomen) were recorded for 252 men by Dr. Roger W. Johnson of Calvin College in Minnesota. The data are in the **STAT1.BodyFat2** data set. Body fat, one measure of health, was accurately estimated by a water displacement measurement technique. The following variables are in the data set:

Case	Case Number
PctBodyFat2	Percent body fat using Siri's equation, $495/\text{Density} - 450$
Age	Age (years)
Weight	Weight (lbs)
Height	Height (inches)
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)
Knee	Knee circumference (cm)
Ankle	Ankle circumference (cm)
Biceps	Extended biceps circumference (cm)
Forearm	Forearm circumference (cm)
Wrist	Wrist circumference (cm) "distal to the styloid processes"

- a. Generate scatter plots and correlations for the VAR variables **Age**, **Weight**, **Height**, and the circumference measures versus the WITH variable, **PctBodyFat2**.



Important! ODS Graphics in PROC CORR limits you to 10 VAR variables at a time, so for this exercise, look at the relationships with **Age**, **Weight**, and **Height** separately from the circumference variables (**Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist**).

Note: This limitation exists only on the graphics obtained from ODS. The correlation table will display all variables in the VAR statement by default.

- 1) Can straight lines adequately describe the relationships?
- 2) Are there any outliers that you should investigate?

- 3) What variable has the highest correlation with **PctBodyFat2**?
 - a) What is the *p*-value for the coefficient?
 - b) Is the correlation statistically significant at the 0.05 level?
- b. Generate correlations among all of the variables in the previously mentioned variables minus **PctBodyFat2**. Are there any notable relationships?
- c. (Advanced) Output the correlation table into a data set. Print out only the correlations whose absolute values are 0.70 and above or note them with an asterisk in the full correlation table.

End of Exercises

2.08 Multiple Choice Poll

The correlation between tuition and rate of graduation at U.S. colleges is 0.55. What does this mean?

- a. The way to increase graduation rates at your college is to raise tuition.
- b. Increasing graduation rates is expensive, causing tuition to rise.
- c. Students who are richer tend to graduate more often than poorer students.
- d. None of the above.

2.5 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 of Statistical Models

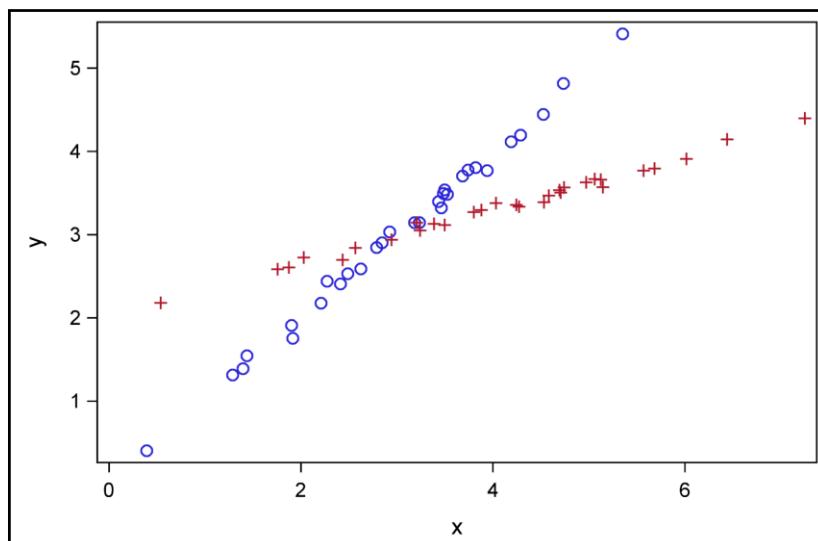
Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

85



Copyright © SAS Institute Inc. All rights reserved.

Overview



86



Copyright © SAS Institute Inc. All rights reserved.

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.

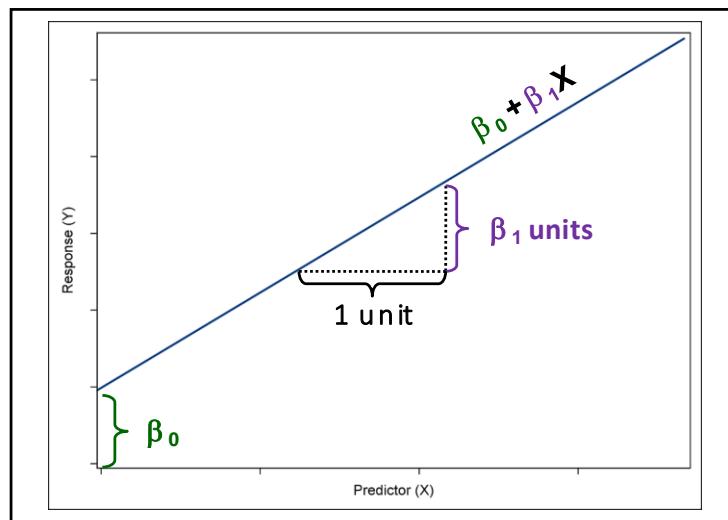
Simple Linear Regression Analysis

The objectives of simple linear regression are as follows:

- 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

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

Simple Linear Regression Model

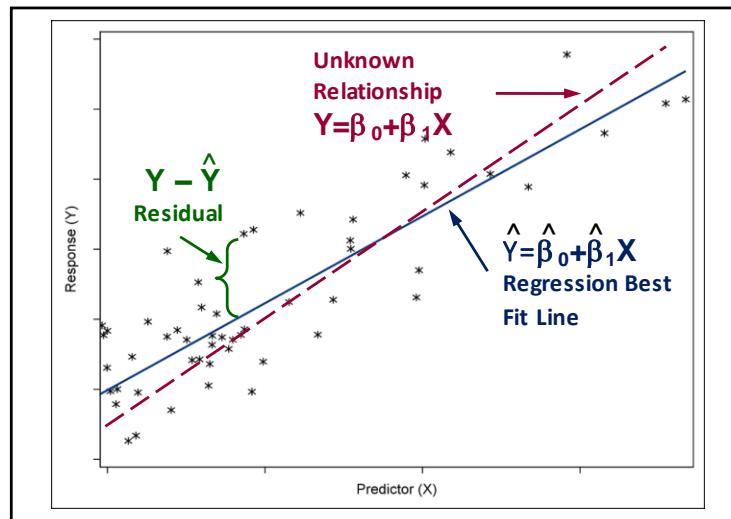


The relationship between the response variable and the predictor variable can be characterized by the equation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i=1, \dots, n$

where

- y_i is the response variable.
- x_i is the predictor variable.
- β_0 is the intercept parameter, which corresponds to the value of the response variable when the predictor is 0.
- β_1 is the slope parameter, which corresponds to the magnitude of change in the response variable given a one unit change in the predictor variable.
- ε_i is the error term representing deviations of y_i about $\beta_0 + \beta_1 x_i$.

Ordinary Least Squares (OLS) Regression



89

Copyright © SAS Institute Inc. All rights reserved.

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.

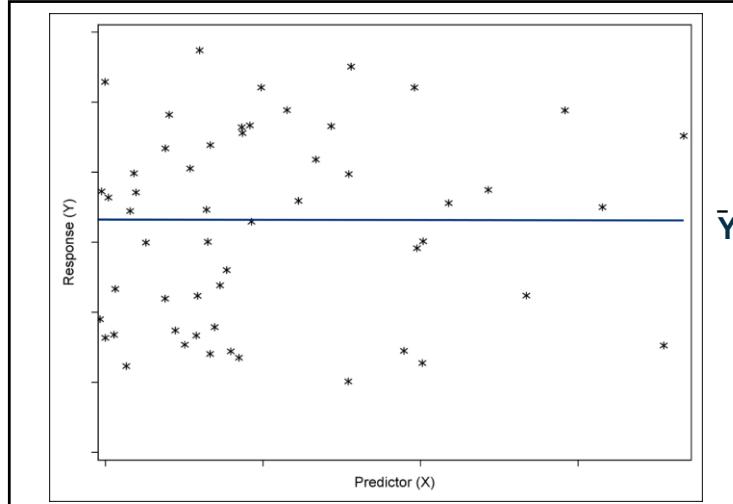
Estimates of the unknown population parameters β_0 and β_1 are obtained by the method of *ordinary least squares*. This method provides the estimates by determining the line that minimizes the sum of the squared vertical distances between the observations and the fitted line. In other words, the fitted or regression line is as close as possible to all the data points.

Ordinary 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, ordinary least squares is used by many data analysts to investigate the relationship between continuous predictor and response variables.

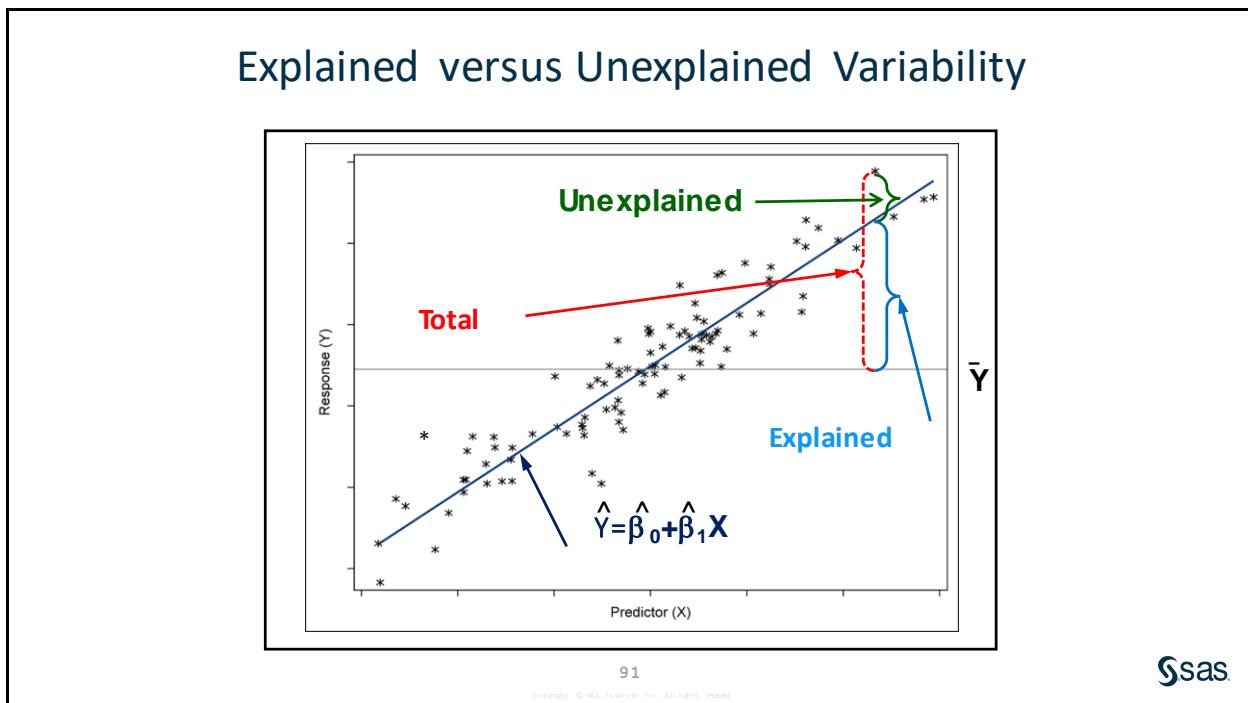
With a large and representative sample, the fitted regression line should be a good approximation of the relationship between the response and predictor variables in the population. The estimated parameters obtained using the method of ordinary least squares should be good approximations of the true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

The Baseline Model (Null Hypothesis)



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 unconditional mean (the mean calculated disregarding the predictor variables) of the response variable.



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 (SS_M) 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 (SS_E) 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 (SS_T) is the sum of the explained and unexplained variability. The corrected total sum of squares is equal to $\sum(Y_i - \bar{Y})^2$. |

Note: Remember that the relationship of the following: total=unexplained+explained applies for sums of squares over all observations and not necessarily for any individual observation.

Model Hypothesis Test

Null Hypothesis:

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

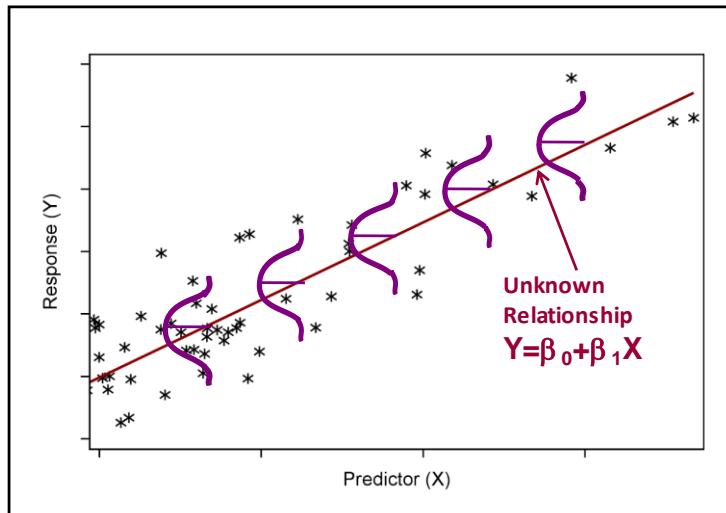
Alternative Hypothesis:

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

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 differs from zero.

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 differs from zero and that the predictor variable explains a significant amount of variability in the response variable.

Assumptions of Simple Linear Regression



93

Copyright © SAS Institute Inc. All rights reserved.



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, that is, the error is normally distributed and has constant variance across the range of the predictor variable, and observations are independent.

Note: 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;
```

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

Selected REG procedure statement:

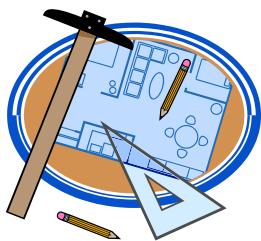
MODEL specifies the response and predictor variables. The variables must be numeric.

Note: 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.

Sale Price Regression Example

PREDICTOR

Lot_Area



RESPONSE

SalePrice



95

Copyright © SAS Institute Inc. All rights reserved.

sas

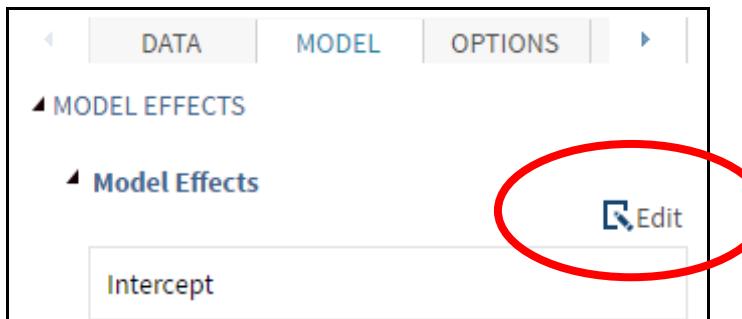
Note: If you were only performing a simple linear regression, the most correlated predictor with **SalePrice** is **Basement_Area** (0.68956). We saw that all of our predictor variables were significantly correlated with **SalePrice** and ultimately we will be performing a multiple linear regression analysis. For the following demonstration, we will select one of our predictors to illustrate a simple linear regression.



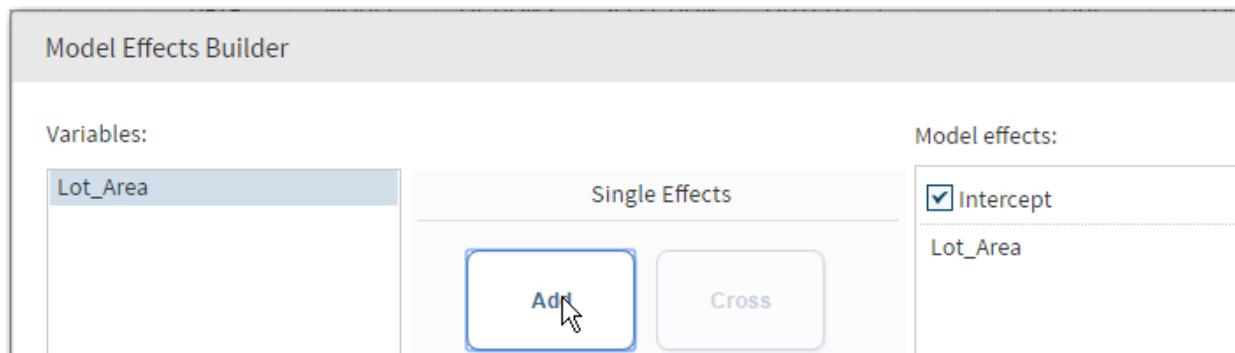
Performing Simple Linear Regression

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

1. Open the **Linear Regression** task under Statistics.
2. From the AmesHousing3 data set select **SalePrice** as the Dependent variable and **Lot_Area** as the Continuous variable.
3. On the MODEL tab, click **Edit**, to specify the **Model Effects**.



4. Select **Lot_Area** and click **Add** under **Single Effects**. Click **OK** to close the model builder window.



5. On the OPTIONS tab, in the PLOTS area, expand the Scatter Plots property and uncheck the option for plotting a scatter plot of observed values by predicted values.
6. Run the code.

Note: Alternatively, you can write the SAS code directly.

```
/*st102d05.sas*/
ods graphics;

proc reg data=STAT1.ameshousing3;
  model SalePrice=Lot_Area;
  title "Simple Regression with Lot_Area as Regressor";
run;
quit;
```

PROC REG Output

Number of Observations Read	300
Number of Observations Used	300

The Number of Observations Read and the Number of Observations Used are the same, which indicates that no missing values were detected for either **SalePrice** or **Lot_Area**.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27164711173	27164711173	20.44	<.0001
Error	298	3.960588E11	1329056404		
Corrected Total	299	4.232235E11			

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

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

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

Each of the column measurements are applied to the following sources of variation:

- 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).

The *F* value tests whether the slope of the predictor variable is equal to 0. The *p*-value is small (less than 0.05), so you have enough evidence at the 0.05 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, **Lot_Area** explains a significant amount of variability in **SalePrice**.

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

Root MSE	36456	R-Square	0.0642
Dependent Mean	137525	Adj R-Sq	0.0610
Coeff Var	26.50882		

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 is \bar{Y} .
Coeff Var	The coefficient of variation is the size of the standard deviation relative to the mean. The coefficient of variation is <ul style="list-style-type: none"> • calculated as $\left(\frac{\text{Root MSE}}{\bar{Y}} \right) * 100$ • a unitless measure, so it can be used to compare data that has different units of measurement or different magnitudes of measurement.
R Square	The coefficient of determination is also referred to as the R-square value. This value is <ul style="list-style-type: none"> • between 0 and 1. • the proportion of variability observed in the data explained by the regression line. In this example, the value is 0.0642, which means that the regression line explains 6% of the total variation in the response values. • the square of the multiple correlation between Y and the Xs. <p>Note: The R square is the squared value of the correlation that you saw earlier between Lot_Area and SalePrice (0.25335). This is no coincidence. For simple regression, the R-square value is the square of the value of the bivariate Pearson correlation coefficient.</p>
Adj R Sq	The adjusted R square is adjusted for the number of parameters in the model. This statistic is useful in multiple regression and is discussed in a later section.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	113740	5666.48352	20.07	<.0001
Lot_Area	Lot size in square feet	1	2.86770	0.63431	4.52	<.0001

The Parameter Estimates table defines the model for your data.

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 <i>t</i> statistic, which is calculated by dividing the parameter estimates by their corresponding standard error estimates.
Pr > t	is the <i>p</i> -value associated with the <i>t</i> 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.

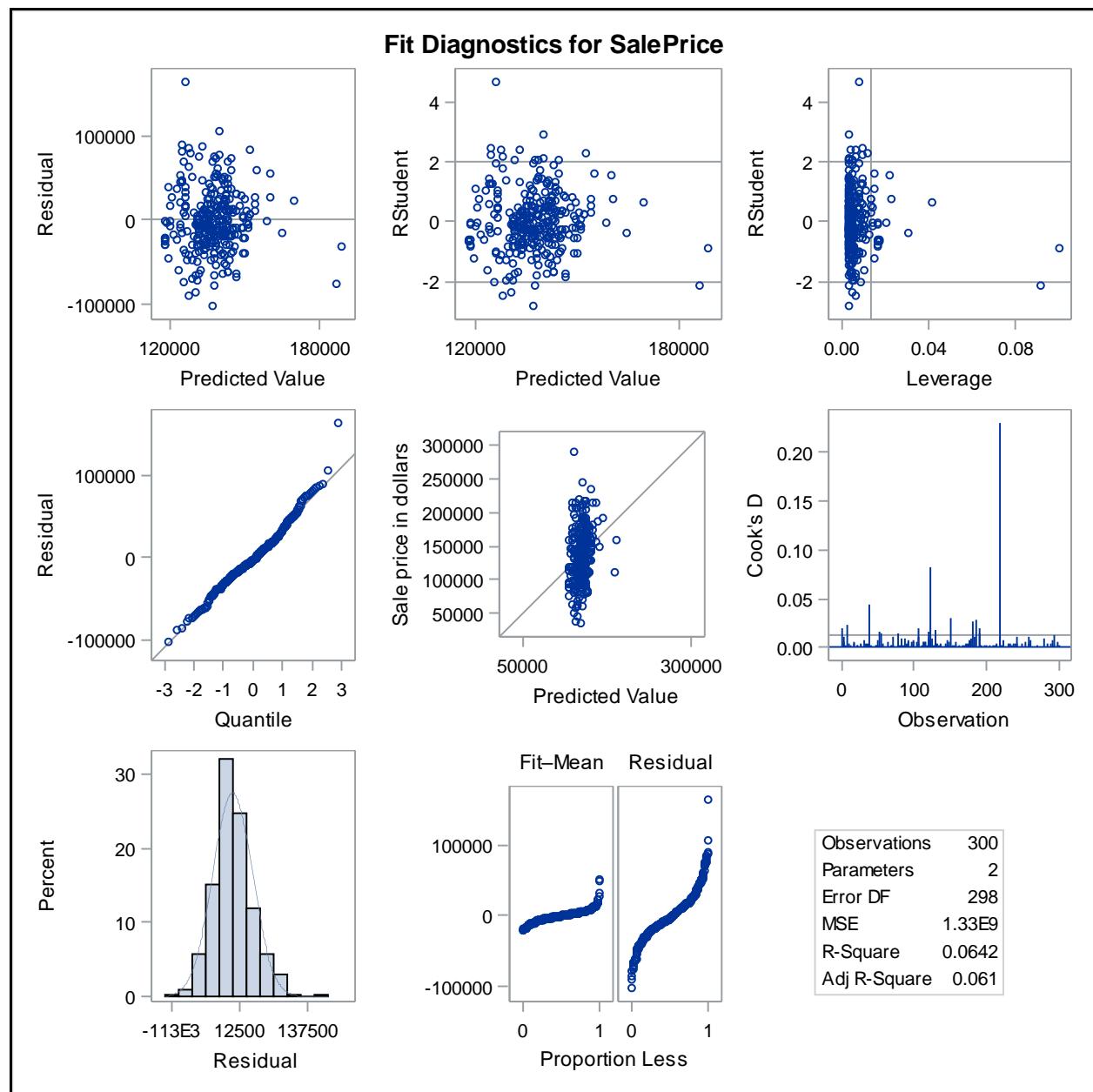
Because the estimate of $\beta_0=113740$ and $\beta_1=2.86770$, the estimated regression equation is given by **SalePrice**=\$113,740+\$2.86770*(**Lot_Area**).

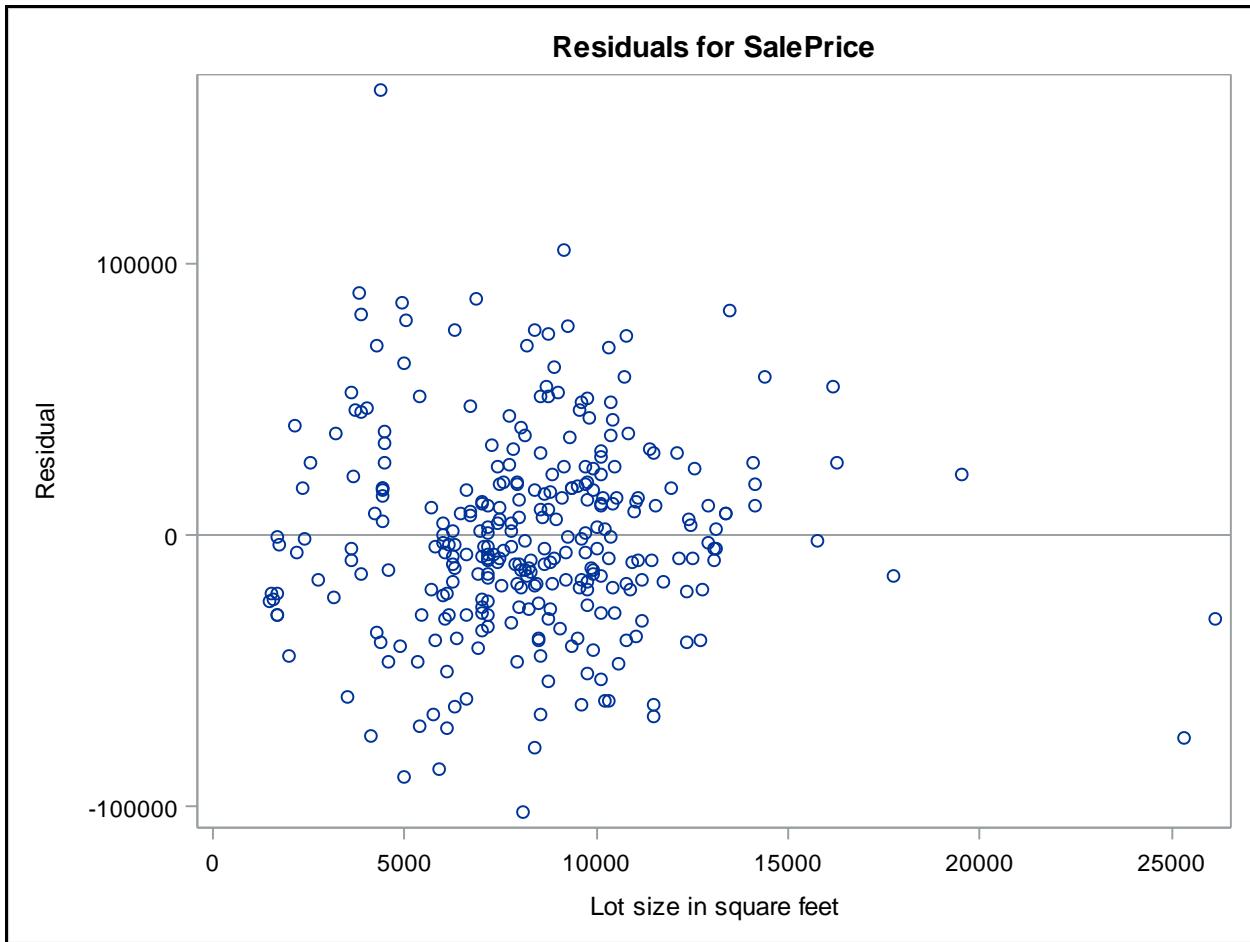
The model indicates that each additional square foot of lot area is associated with an approximately \$2.87 higher sale price.

Note: *Extrapolation of the model beyond the range of your predictor variables is inappropriate.*

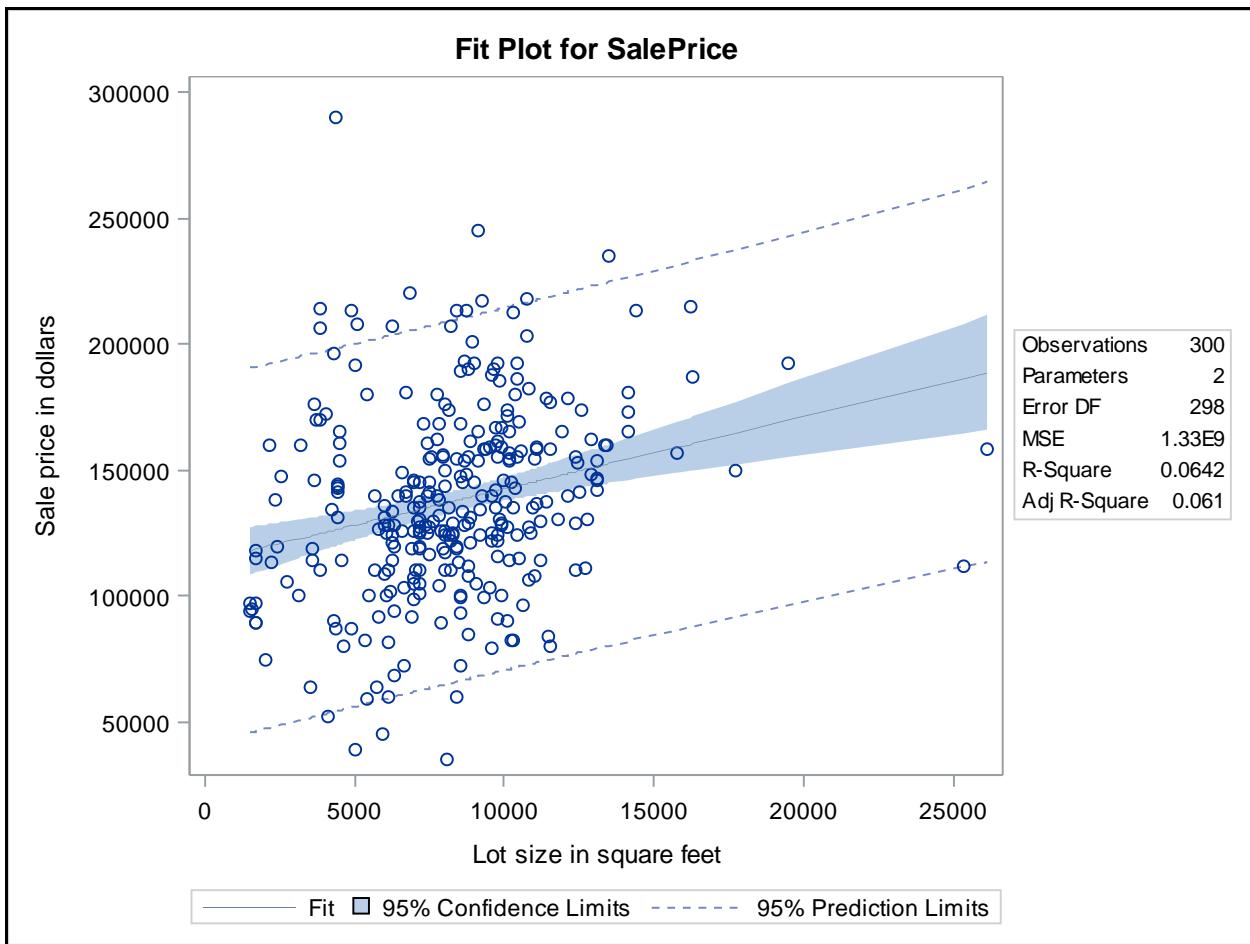
You cannot assume that the relationship maintains in areas that were not sampled from.

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 **SalePrice**=0 (giving away a house for free) is not within the range of observed values.





The diagnostics table and the residuals by **Lot_Area** table show a variety of plots designed to help with an assessment of the data's fulfillment of statistical assumptions and influential outliers. These plots are explored in detail in a later chapter.



The Fit plot produced by ODS Graphics shows the predicted regression line superimposed over a scatter plot of the data.

To assess the level of precision around the mean estimates of **SalePrice**, 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 that your interval contains the population mean of Y for a particular X.
- Confidence intervals become wider as you move away from the mean of the independent variable. This reflects the fact that your estimates become more variable as you move away from the means of X and Y.

Suppose that the mean **SalePrice** at a fixed value of **Lot_Area** is not the focus. If you are interested in making a prediction for a future single observation, you need a **prediction interval**. This is represented by the area between the broken lines in the plot.

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

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

End of Demonstration

2.09 Multiple Choice Poll

Run PROC REG with this MODEL statement: model y=x1;. If the parameter estimate (slope) of x1 is 0, then the best guess (predicted value) of y when x1=13 is which of the following?

- a. 13
- b. the mean of y
- c. a random number
- d. the mean of x1
- e. 0



Exercises

4. Fitting a Simple Linear Regression Model

Use the **STAT1.BodyFat2** data set for this exercise.

Perform a simple linear regression model with **PctBodyFat2** as the response variable and **Weight** as the predictor.

- a. What is the value of the F statistic and the associated p -value? How would you interpret this with regard to the null hypothesis?
- b. Write the predicted regression equation.
- c. What is the value of R-square? How would you interpret this?

End of Exercises

2.6 Solutions

Solutions to Exercises

1. Analysis of Variance with Garlic Data

Consider an experiment to study four types of fertilizer, labeled 1, 2, 3, and 4. One fertilizer is chemical and the rest are organic. You want to see whether the average of weights of garlic bulbs are significantly different for plants in beds using different fertilizers.

Test the hypothesis that the means are equal. Be sure to check that the assumptions of the analysis method that you choose are met. What conclusions can you reach at this point in your analysis?

a. Checking Assumptions

- 1) Open the **Summary Statistics** task under Statistics and choose the **GARLIC** data set.
- 2) Select **BulbWt** as the analysis variables and **Fertilizer** as the classification variables.
- 3) Run the code to produce summary statistics for the two groups.
- 4) Open the **Box Plot** task under Graph.
- 5) Again set the analysis variable to **BulbWt** and category variable to **Fertilizer**.
- 6) Run the code to produce box plots of bulb weight for the four groups.

b. ANOVA

- 1) Open the **One-Way ANOVA** task under Statistics.
- 2) Under ROLES, set the Dependent variable to **BulbWt** and the Categorical variable to **Fertilizer**
- 3) On the OPTIONS tab do the following:
 - Uncheck the box for **Welch's variance-weighted ANOVA**.
 - Change the **Comparisons method** option to **None**.
 - Drop down the Display plots menu, choose the **Selected plots** option, and check only the option to display the **Diagnostics plot** (uncheck all other plot boxes).
- 4) Run the code to conduct a One-Way ANOVA.

Note: The code below produces the necessary output for the three tasks.

```
/*st102s01.sas*/ /*Part A*/
proc means data=STAT1.Garlic;
  var BulbWt;
  class Fertilizer;
  title 'Descriptive Statistics of BulbWt by Fertilizer';
run;

proc sgplot data=STAT1.Garlic;
```

```

vbox BulbWt / category=Fertilizer connect=mean;
  title "Bulb Weight Differences across Fertilizers";
run;

/*st102s01.sas*/ /*Part B*/
ods graphics;

proc glm data=STAT1.Garlic plots(only)=diagnostics;
  class Fertilizer;
  model BulbWt=Fertilizer;
  means Fertilizer / hovtest=levene;
  title "One-Way ANOVA with Fertilizer as Predictor";
run;
quit;

```

PROC GLM Output

Class Level Information		
Class	Levels	Values
Fertilizer	4	1 2 3 4

Number of Observations Read	32
Number of Observations Used	32

Dependent Variable: BulbWt

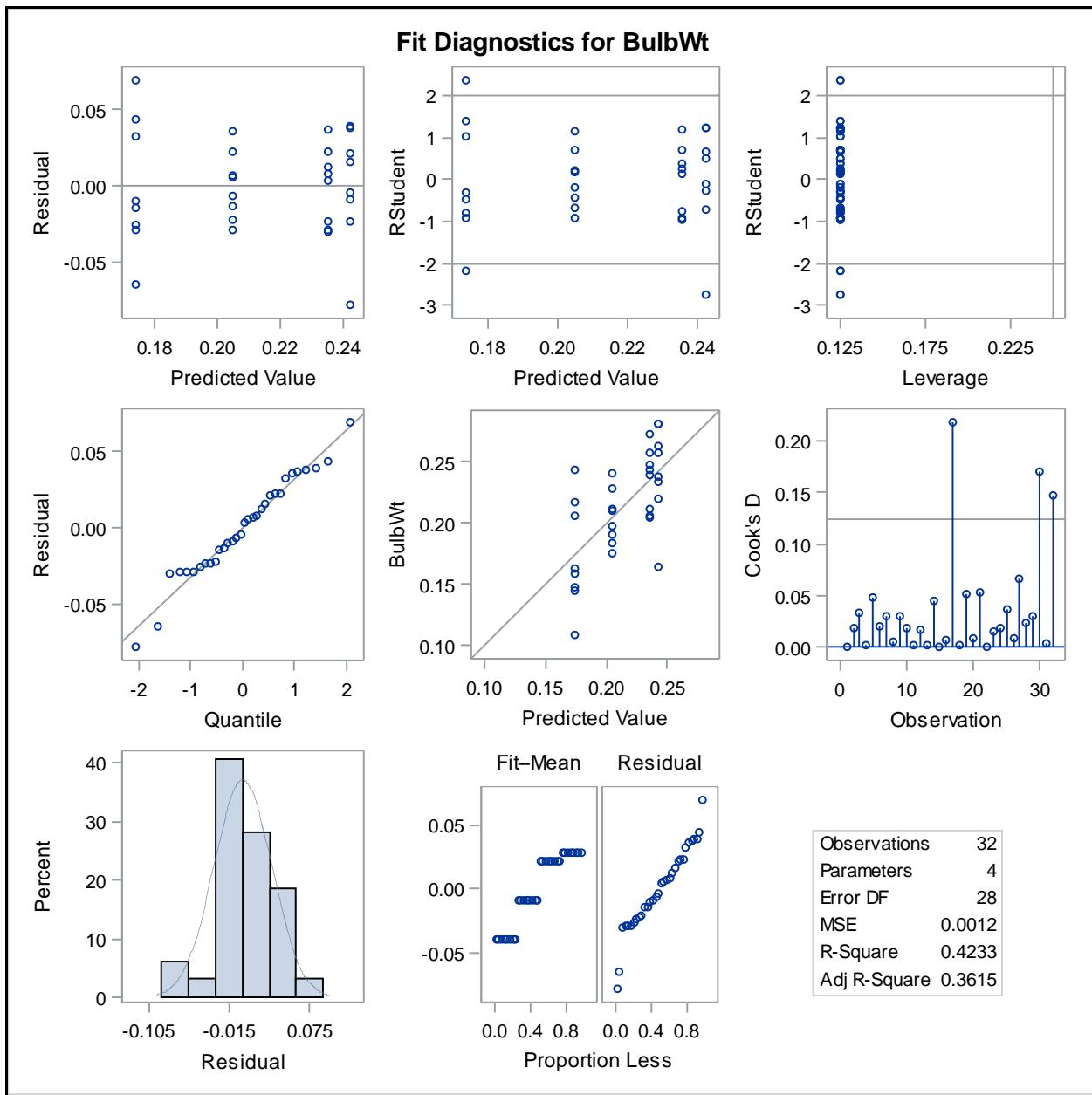
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.02370114	0.00790038	6.85	0.0013
Error	28	0.03229141	0.00115326		
Corrected Total	31	0.05599255			

The overall *F* value from the analysis of variance table is associated with a *p*-value=0.0013. Presuming that all assumptions of the model are valid, you know that at least one treatment mean is different from one other treatment mean. At this point, you do not know which means are significantly different from one another.

R-Square	Coeff Var	Root MSE	BulbWt Mean
0.423291	15.85633	0.033960	0.214172

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Fertilizer	3	0.02370114	0.00790038	6.85	0.0013

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fertilizer	3	0.02370114	0.00790038	6.85	0.0013



Both the histogram and Q-Q plot show that the residuals seem relatively normally distributed (one assumption for ANOVA).

Levene's Test for Homogeneity of BulbWt Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Fertilizer	3	9.13E-6	3.043E-6	1.54	0.2257
Error	28	0.000055	1.974E-6		

The Levene's Test for Homogeneity of Variance shows a *p*-value greater than alpha. Therefore, do not reject the hypothesis of homogeneity of variances (equal variances across Ad types). This assumption for ANOVA is met.

2. Post Hoc Pairwise Comparisons

Consider again the analysis of the **STAT1.Garlic** data set. There was a statistically significant difference among means for sales for the different fertilizers. Perform a post hoc test to look at the individual differences among means.

- a. Conduct pairwise comparisons with an experimentwise error rate of $\alpha=0.05$. (Use the Tukey adjustment.) Which types of fertilizer are significantly different?
- b. Use level 4 (the chemical fertilizer) as the control group and perform a Dunnett comparison with the organic fertilizers to see whether they affected the average weights of garlic bulbs differently from the control fertilizer.
- c. (Extra) Perform unadjusted tests of all pairwise comparisons to see what would have happened if the multi-test adjustments had not been made. How do the results compare to what you saw in the Tukey adjusted tests?
 - 1) Open the **One-Way ANOVA** task tab that you created in the previous exercise.
 - 2) Use the drop-down menu for Comparisons method, under COMPARISONS to choose **Tukey** for Tukey's HSD.
 - 3) Under the PLOTS property, use the drop-down menu and choose the selected plots option.
 - 4) Select the LS-mean Difference plot option and uncheck the Diagnostics plot option.
 - 5) Run the code.
 - 6) To compare the output for the three different comparisons methods, rerun the task with different comparisons methods.
 - 7) After specifying **Dunnett** two-tail as comparisons method, use the drop-down menu to choose **4** as the **Control level**.
 - 8) To include the control plot in the output, select the option to display default plots.
 - 9) For **unadjusted** tests, choose **Least significant difference (LSD)** as the Comparisons method.

Note: Alternatively, use the editing option and add the rest of the comparisons method in manually. The code below produces all the necessary output shown below.

```
/*st102s02.sas*/
ods graphics;

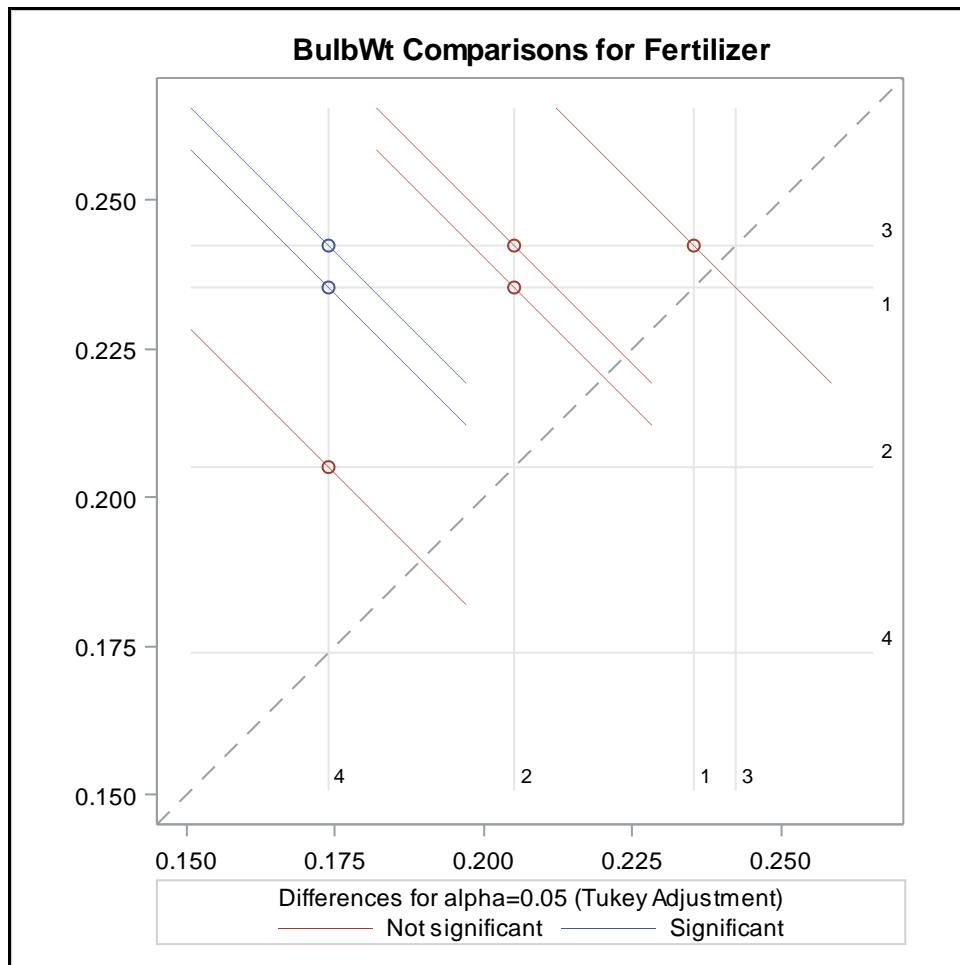
ods select lsmeans diff diffplot controlplot;
proc glm data=STAT1.Garlic
plots(only)=(diffplot(center) controlplot);
class Fertilizer;
model BulbWt=Fertilizer;
Tukey: lsmeans Fertilizer / pdiff=all adjust=tukey;
```

```
Dunnett: lsmeans Fertilizer / pdiff=control('4') adjust=dunnett;
No_Adjust: lsmeans Fertilizer / pdiff=all adjust=t;
title "Post-Hoc Analysis of ANOVA - Fertilizer as Predictor";
run;
quit;
```

PROC GLM Output

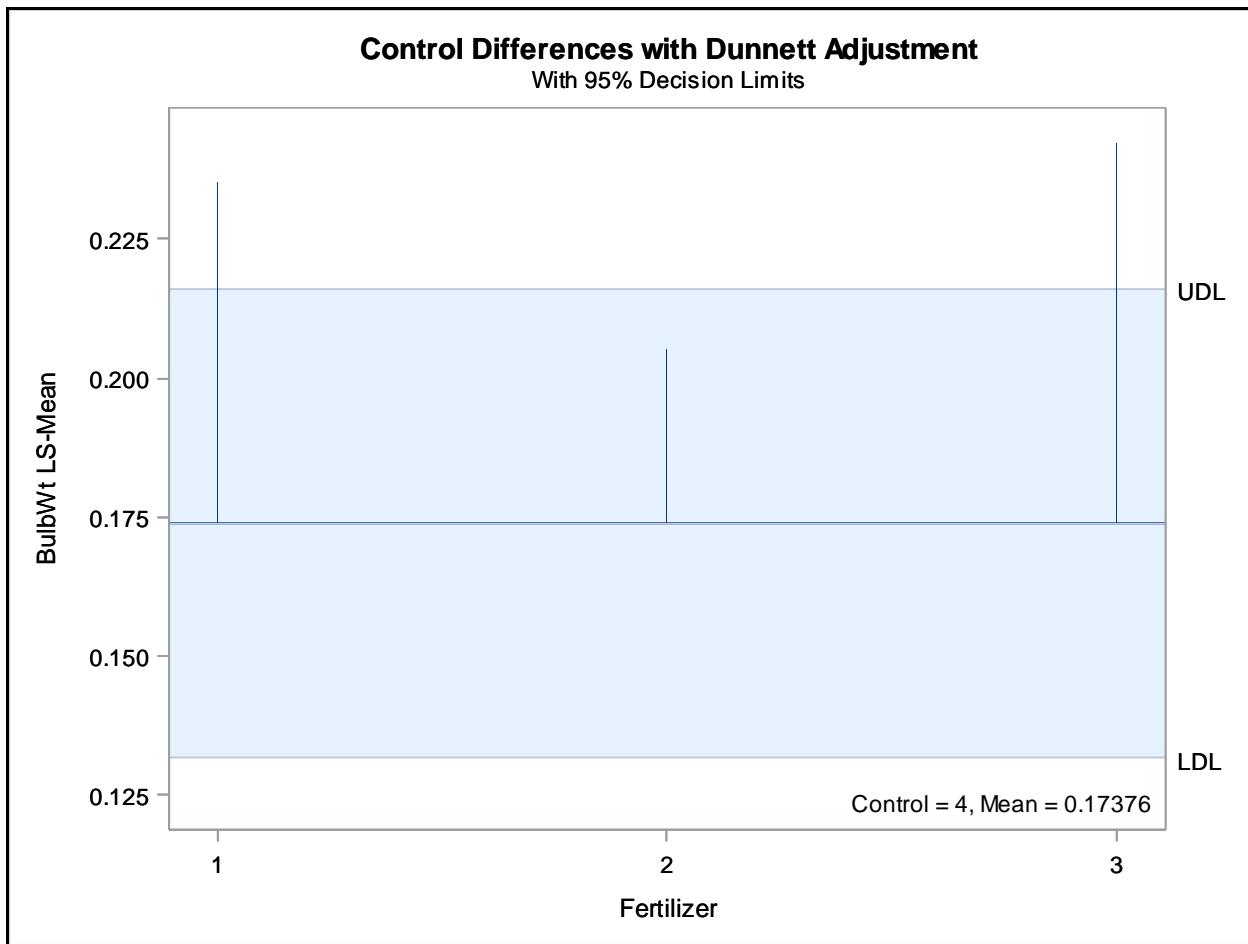
Fertilizer	BulbWt LSMEAN	LSMEAN Number
1	0.23539981	1
2	0.20511406	2
3	0.24240747	3
4	0.17376488	4

Least Squares Means for effect Fertilizer Pr > t for H0: LSMean(i)=LSMean(j)					
Dependent Variable: BulbWt					
i\j	1	2	3	4	
1		0.3021	0.9758	0.0058	
2	0.3021		0.1490	0.2738	
3	0.9758	0.1490		0.0020	
4	0.0058	0.2738	0.0020		



The Tukey comparisons show significant differences between fertilizers 3 and 4 ($p=0.0020$) and 1 and 4 ($p=0.0058$).

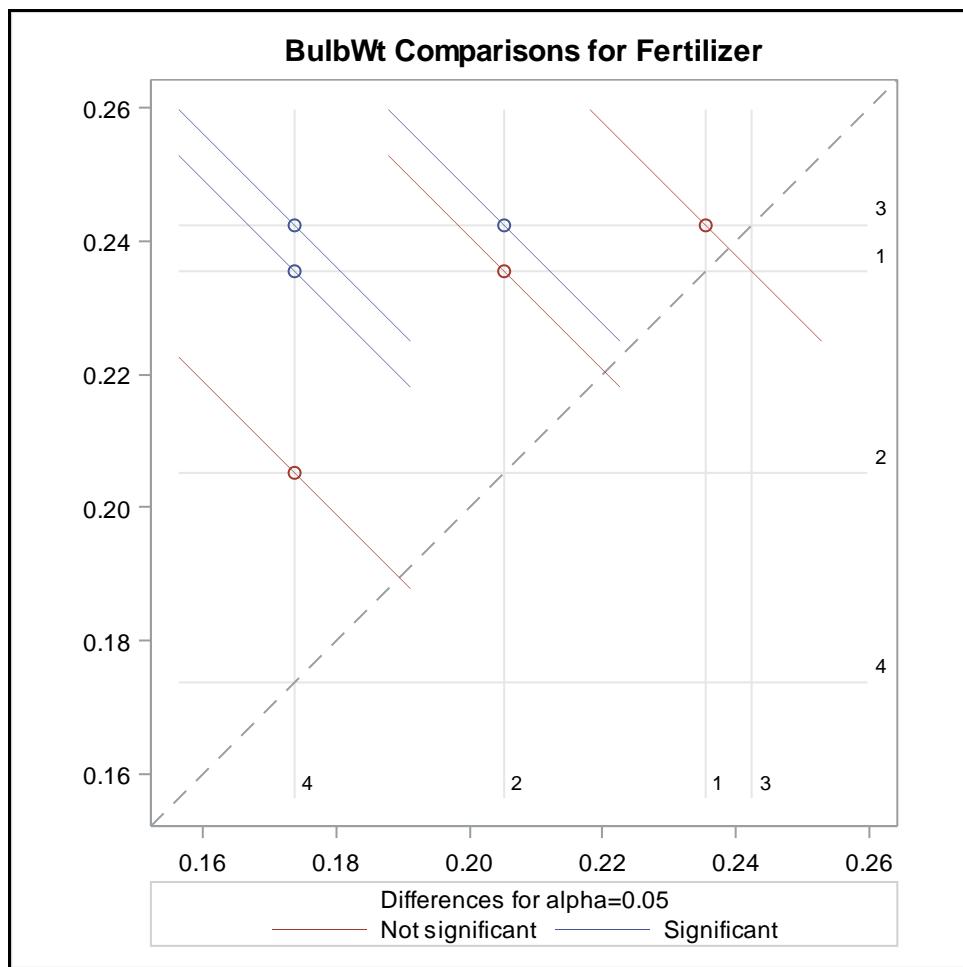
Fertilizer	BulbWt LSMEAN	H0:LSMean=Control	
			Pr > t
1	0.23539981		0.0031
2	0.20511406		0.1801
3	0.24240747		0.0011
4	0.17376488		



The Dunnett comparisons show the same pairs as significantly different, but with smaller *p*-values than with the Tukey comparisons (3 versus 4 *p*=0.0011, 1 versus 4 *p*=0.0031). This is due to the fact that the Tukey adjustment is for more pairwise comparisons than the Dunnett adjustment.

Fertilizer	BulbWt LSMEAN	LSMEAN Number
1	0.23539981	1
2	0.20511406	2
3	0.24240747	3
4	0.17376488	4

Least Squares Means for effect Fertilizer Pr > t for H0: LSMean(i)=LSMean(j)				
Dependent Variable: BulbWt				
i/j	1	2	3	4
1		0.0853	0.6830	0.0011
2	0.0853		0.0365	0.0755
3	0.6830	0.0365		0.0004
4	0.0011	0.0755	0.0004	



The unadjusted (*t*-test) comparisons all have smaller *p*-values than they had with Tukey adjustments. One additional comparison has a *p*-value below 0.05 (2 versus 3).

3. Describing the Relationship between Continuous Variables

Percentage of body fat, age, weight, height, and 10 body circumference measurements (for example, abdomen) were recorded for 252 men by Dr. Roger W. Johnson of Calvin College in Minnesota. The data are in the **STAT1.BodyFat2** data set. Body fat, one measure of health, was accurately estimated by an underwater weighing technique. There are two measures of percentage body fat in this data set.

- a. Generate scatter plots and correlations for the VAR variables **Age**, **Weight**, **Height**, and the circumference measures versus the WITH variable, **PctBodyFat2**.



Important! ODS Graphics in PROC CORR limits you to 10 VAR variables at a time, so for this exercise, look at the relationships with **Age**, **Weight**, and **Height** separately from the circumference variables (**Neck** **Chest** **Abdomen** **Hip** **Thigh** **Knee** **Ankle** **Biceps** **Forearm** **Wrist**).

- 1) Open the **Correlation Analysis** task under **Statistics** and select the **BODYFAT2** data set.
- 2) Select all the variables of interest for **Analysis variables** and select **PctBodyFat2** for the **Correlate with** variable.
- 3) On the OPTIONS tab do the following:
 - Choose the Selected statistics option under the Display statistics property and check the options to display Correlations, p-values, and Descriptive statistics.
 - Choose the Individual scatter plots option for Type of plot and check the box to include inset statistics.
 - Change the number of variables to plot option from 5 to 10.
- 4) Run the code and repeat for the rest of the variables.

Note: Alternatively, you can write the code directly in SAS program.

```
/*st102s03.sas*/ /*Part A*/
%let interval=Age Weight Height Neck Chest Abdomen Hip
            Thigh Knee Ankle Biceps Forearm Wrist;

ods graphics / reset=all imagemap;
proc corr data=STAT1.BodyFat2
           plots(only)=scatter(nvar=all ellipse=none);
  var &interval;
  with PctBodyFat2;
  id Case;
  title "Correlations and Scatter Plots";
run;

%let interval=Biceps Forearm Wrist;

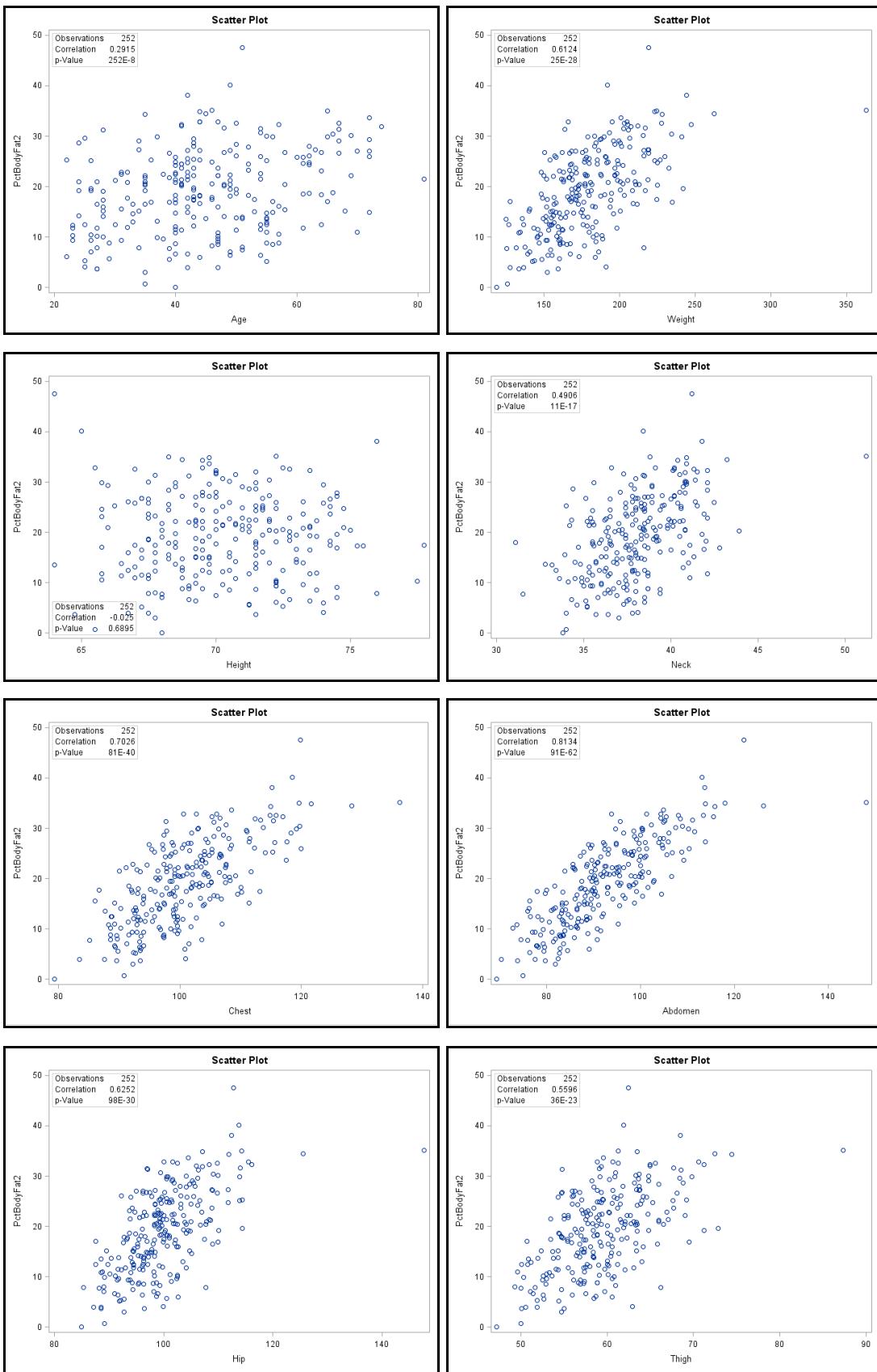
ods graphics / reset=all imagemap;
ods select scatterplot;
proc corr data=STAT1.BodyFat2
           plots(only)=scatter(nvar=all ellipse=none);
  var &interval;
  with PctBodyFat2;
  id Case;
  title "Correlations and Scatter Plots";
run;
```

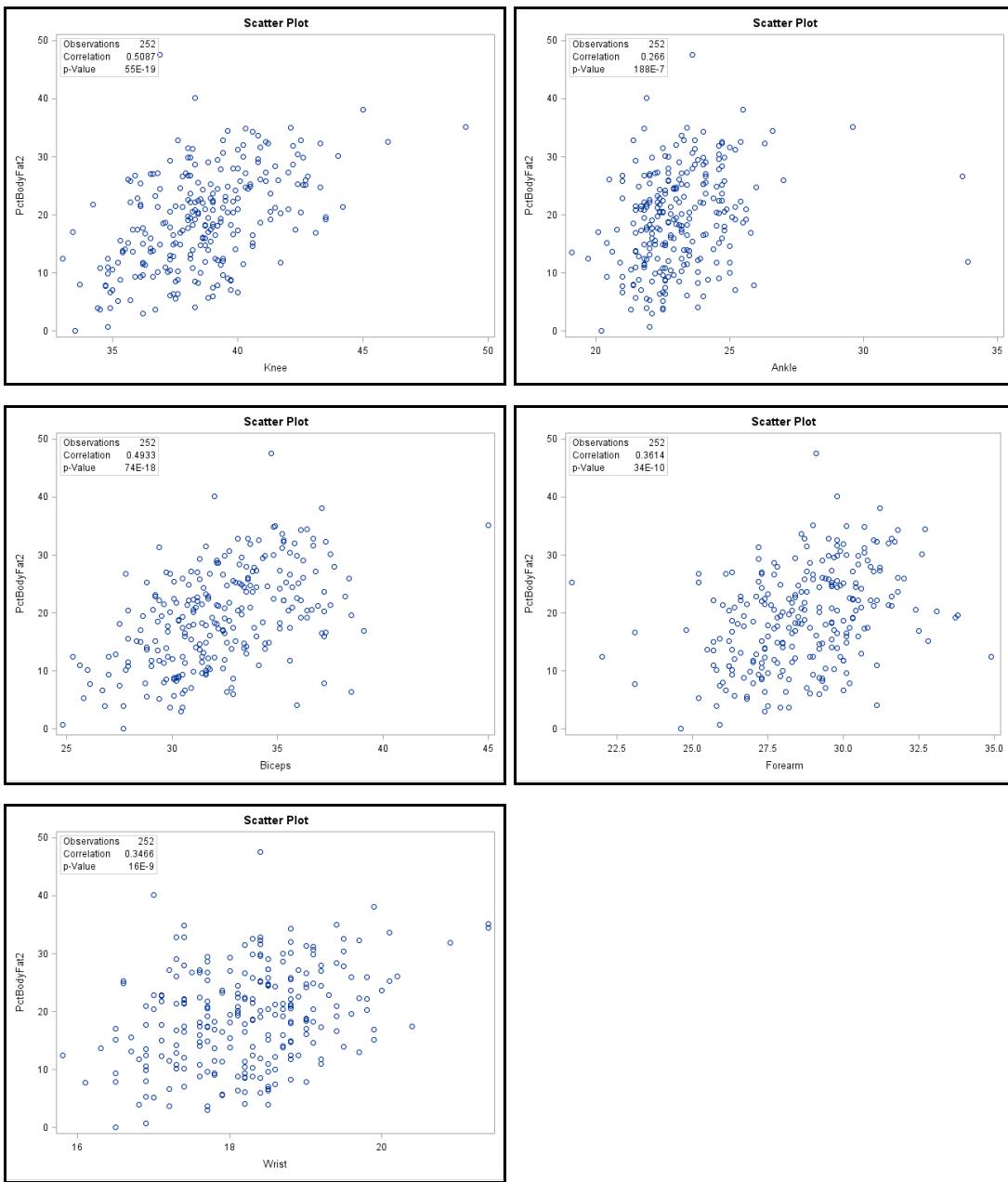
1 With Variables:	PctBodyFat2									
13 Variables:	Age Ankle Weight Biceps Height Forearm Neck Wrist Chest Abdomen Hip Thigh Knee									

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PctBodyFat2	252	19.15079	8.36874	4826	0	47.50000
Age	252	44.88492	12.60204	11311	22.00000	81.00000
Weight	252	178.92440	29.38916	45089	118.50000	363.15000
Height	252	70.30754	2.60958	17718	64.00000	77.75000
Neck	252	37.99206	2.43091	9574	31.10000	51.20000
Chest	252	100.82421	8.43048	25408	79.30000	136.20000
Abdomen	252	92.55595	10.78308	23324	69.40000	148.10000
Hip	252	99.90476	7.16406	25176	85.00000	147.70000
Thigh	252	59.40595	5.24995	14970	47.20000	87.30000
Knee	252	38.59048	2.41180	9725	33.00000	49.10000
Ankle	252	23.10238	1.69489	5822	19.10000	33.90000
Biceps	252	32.27341	3.02127	8133	24.80000	45.00000
Forearm	252	28.66389	2.02069	7223	21.00000	34.90000
Wrist	252	18.22976	0.93358	4594	15.80000	21.40000

Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0									
PctBodyFat2	Age 0.29146 <.0001	Weight 0.61241 <.0001	Height -0.02529 0.6895	Neck 0.49059 <.0001	Chest 0.70262 <.0001	Abdomen 0.81343 <.0001	Hip 0.62520 <.0001	Thigh 0.55961 <.0001	

Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0					
PctBodyFat2	Knee 0.50867 <.0001	Ankle 0.26597 <.0001	Biceps 0.49327 <.0001	Forearm 0.36139 <.0001	Wrist 0.34657 <.0001





5) Can straight lines adequately describe the relationships?

Yes

6) Are there any outliers that you should investigate?

There is one person with outlying values for several measurements. In addition, there are one or two more values that seem to be outliers for Ankle.

7) What variable has the highest correlation with PctBodyFat2?

Abdomen ($r=0.81343$).

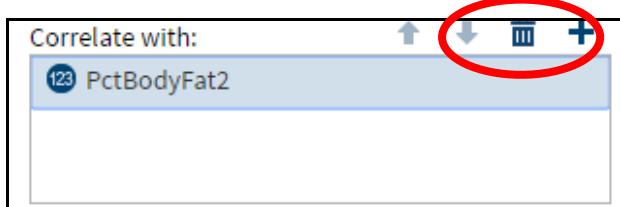
- a) What is the *p*-value for the coefficient?

The *p*-value is <.0001.

- b) Is the correlation statistically significant at the 0.05 level?

Yes

- c) Generate correlations among all of the variables in the previously mentioned variables minus **PctBodyFat2**. Select only the highest five per variable. Are there any notable relationships?
- 8) On the same **Correlation Analysis** task, in the **Correlate with** box select **PctBodyFat2** and remove it by clicking the trash can button.



- 9) On the OPTIONS tab, uncheck the **Descriptive statistics** box and select the **None** option for the **Type of plot**.
- 10) To print the highest five correlated variables for each variable, **edit** the generated code by including **best=5**.



```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=STAT1.BODYFAT2 best=5 nosimple pearson plots=none;
  var Age Weight Height Adioposity Neck Chest Abdomen Hip Thigh
  Knee
  Ankle Biceps Forearm Wrist;

run;
```

Note: Alternatively, you can write the code directly in SAS.

```
/*st102s03.sas*/ /*Part B*/
ods graphics off;
%let interval=Age Weight Height Neck Chest Abdomen Hip Thigh
            Knee Ankle Biceps Forearm Wrist;

proc corr data=STAT1.BodyFat2
    nosimple
    best=5
    out=pearson;
var &interval;
title "Correlations of Predictors";
run;
```

13 Variables: Age Weight Height Neck Chest Abdomen Hip Thigh Knee Ankle Biceps
Forearm Wrist

Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0						
Age	Age	Height	Abdomen	Wrist	Thigh	
	1.00000	-0.24521 <.0001	0.23041 0.0002	0.21353 0.0006	-0.20010 0.0014	
Weight	Weight	Hip	Chest	Abdomen	Thigh	
	1.00000	0.94088 <.0001	0.89419 <.0001	0.88799 <.0001	0.86869 <.0001	
Height	Height	Knee	Weight	Wrist	Ankle	
	1.00000	0.50050 <.0001	0.48689 <.0001	0.39778 <.0001	0.39313 <.0001	
Neck	Neck	Weight	Chest	Abdomen	Wrist	
	1.00000	0.83072 <.0001	0.78484 <.0001	0.75408 <.0001	0.74483 <.0001	
Chest	Chest	Abdomen	Weight	Hip	Neck	
	1.00000	0.91583 <.0001	0.89419 <.0001	0.82942 <.0001	0.78484 <.0001	
Abdomen	Abdomen	Chest	Weight	Hip	Thigh	
	1.00000	0.91583 <.0001	0.88799 <.0001	0.87407 <.0001	0.76662 <.0001	
Hip	Hip	Weight	Thigh	Abdomen	Chest	
	1.00000	0.94088 <.0001	0.89641 <.0001	0.87407 <.0001	0.82942 <.0001	
Thigh	Thigh	Hip	Weight	Knee	Abdomen	
	1.00000	0.89641 <.0001	0.86869 <.0001	0.79917 <.0001	0.76662 <.0001	
Knee	Knee	Weight	Hip	Thigh	Abdomen	
	1.00000	0.85317 <.0001	0.82347 <.0001	0.79917 <.0001	0.73718 <.0001	
Ankle	Ankle	Weight	Knee	Wrist	Hip	
	1.00000	0.61369 <.0001	0.61161 <.0001	0.56619 <.0001	0.55839 <.0001	
Biceps	Biceps	Weight	Thigh	Hip	Neck	
	1.00000	0.80042 <.0001	0.76148 <.0001	0.73927 <.0001	0.73115 <.0001	
Forearm	Forearm	Biceps	Weight	Neck	Wrist	
	1.00000	0.67826 <.0001	0.63030 <.0001	0.62366 <.0001	0.58559 <.0001	
Wrist	Wrist	Neck	Weight	Knee	Chest	
	1.00000	0.74483 <.0001	0.72977 <.0001	0.66451 <.0001	0.66016 <.0001	

Weight has a high correlation with nearly every other variable. Hip also is correlated with most variables.

(Advanced) Output the correlation table into a data set. Print out only the correlations whose absolute values are 0.70 and above or note them with an asterisk in the full correlation table.

Potential solution to printing out the correlation matrix with asterisks for correlations with absolute values at 0.7 and above:

```
%let big=0.7;
proc format;
  picture correlations &big -< 1 = '009.99' (prefix="*")
    -1 <- -&big = '009.99' (prefix="*")
    -&big <-< &big = '009.99';
run;

proc print data=pearson;
  var _NAME_ &interval;
  where _type_="CORR";
  format &interval correlations .;
run;
```

Obs	_NAME_	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
4	Age	1	0.01	0.24	0.11	0.17	0.23	0.05	0.20	0.01	0.10	0.04	0.08	0.21
5	Weight	0.01	1	0.48	*0.83	*0.89	*0.88	*0.94	*0.86	*0.85	0.61	*0.80	0.63	*0.72
6	Height	0.24	0.48	1	0.32	0.22	0.18	0.37	0.33	0.50	0.39	0.31	0.32	0.39
7	Neck	0.11	*0.83	0.32	1	*0.78	*0.75	*0.73	0.69	0.67	0.47	*0.73	0.62	*0.74
8	Chest	0.17	*0.89	0.22	*0.78	1	*0.91	*0.82	*0.72	*0.71	0.48	*0.72	0.58	0.66
9	Abdomen	0.23	*0.88	0.18	*0.75	*0.91	1	*0.87	*0.76	*0.73	0.45	0.68	0.50	0.61
10	Hip	0.05	*0.94	0.37	*0.73	*0.82	*0.87	1	*0.89	*0.82	0.55	*0.73	0.54	0.63
11	Thigh	0.20	*0.86	0.33	0.69	*0.72	*0.76	*0.89	1	*0.79	0.53	*0.76	0.56	0.55
12	Knee	0.01	*0.85	0.50	0.67	*0.71	*0.73	*0.82	*0.79	1	0.61	0.67	0.55	0.66
13	Ankle	0.10	0.61	0.39	0.47	0.48	0.45	0.55	0.53	0.61	1	0.48	0.41	0.56
14	Biceps	0.04	*0.80	0.31	*0.73	*0.72	0.68	*0.73	*0.76	0.67	0.48	1	0.67	0.63
15	Forearm	0.08	0.63	0.32	0.62	0.58	0.50	0.54	0.56	0.55	0.41	0.67	1	0.58
16	Wrist	0.21	*0.72	0.39	*0.74	0.66	0.61	0.63	0.55	0.66	0.56	0.63	0.58	1

Potential solution to printing out only the correlations whose absolute values are 0.7 and above:

```
%let big=0.7;
data bigcorr;
  set pearson;
  retain row 1;
  array vars{*} &interval;
  do i=1 to dim(vars);
    if i>row or abs(vars{i})<&big then vars{i}=.;
  end;
  if _type_="CORR";
  drop i _type_ row;
  row=row+1;
  output;
run;

proc print data=bigcorr;
  format &interval 5.2;
run;
```

Obs	_NAME_	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1	Age	1.00
2	Weight	.	1.00	.	0.83	0.89	0.89	0.94	0.87	0.85	.	0.80	.	0.73
3	Height	.	.	1.00
4	Neck	.	0.83	.	1.00	0.78	0.75	0.73	.	.	.	0.73	.	0.74
5	Chest	.	0.89	.	0.78	1.00	0.92	0.83	0.73	0.72	.	0.73	.	.
6	Abdomen	.	0.89	.	0.75	0.92	1.00	0.87	0.77	0.74
7	Hip	.	0.94	.	0.73	0.83	0.87	1.00	0.90	0.82	.	0.74	.	.
8	Thigh	.	0.87	.	.	0.73	0.77	0.90	1.00	0.80	.	0.76	.	.
9	Knee	.	0.85	.	.	0.72	0.74	0.82	0.80	1.00
10	Ankle	1.00	.	.	.
11	Biceps	.	0.80	.	0.73	0.73	.	0.74	0.76	.	.	1.00	.	.
12	Forearm	1.00	.
13	Wrist	.	0.73	.	0.74	1.00

4. Fitting a Simple Linear Regression Model

Use the **STAT1.BodyFat2** data set for this exercise.

Perform a simple linear regression model with **PctBodyFat2** as the response variable and **Weight** as the predictor.

Open the **Linear Regression** task under **Statistics** and select the **BodyFat2** data set.

Select **PctBodyFat2** as the Dependent variable and **Weight** as Continuous variables.

On the MODEL tab, add **Weight** as an effect to the model.

On the OPTIONS tab, expand the Scatter Plots property and clear the box for plotting observed values by predicted values.

Run the code.

Note: Alternatively, the code below produces the necessary output.

```
/*st102s04.sas*/
ods graphics on;

proc reg data=STAT1.BodyFat2;
  model PctBodyFat2=Weight;
  title "Regression of % Body Fat on Weight";
run;
quit;
```

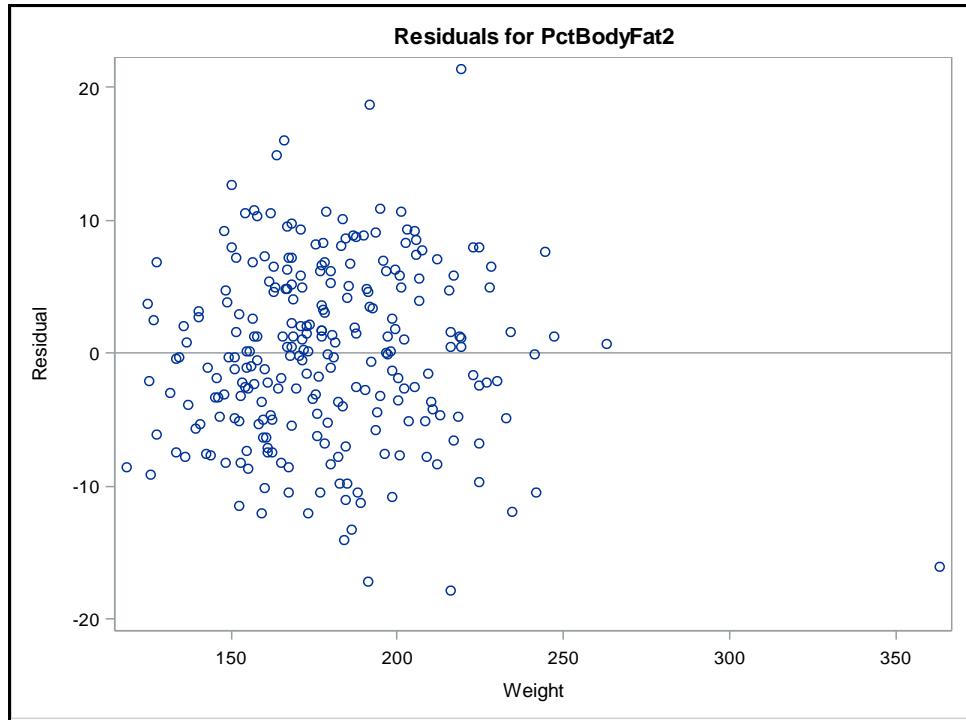
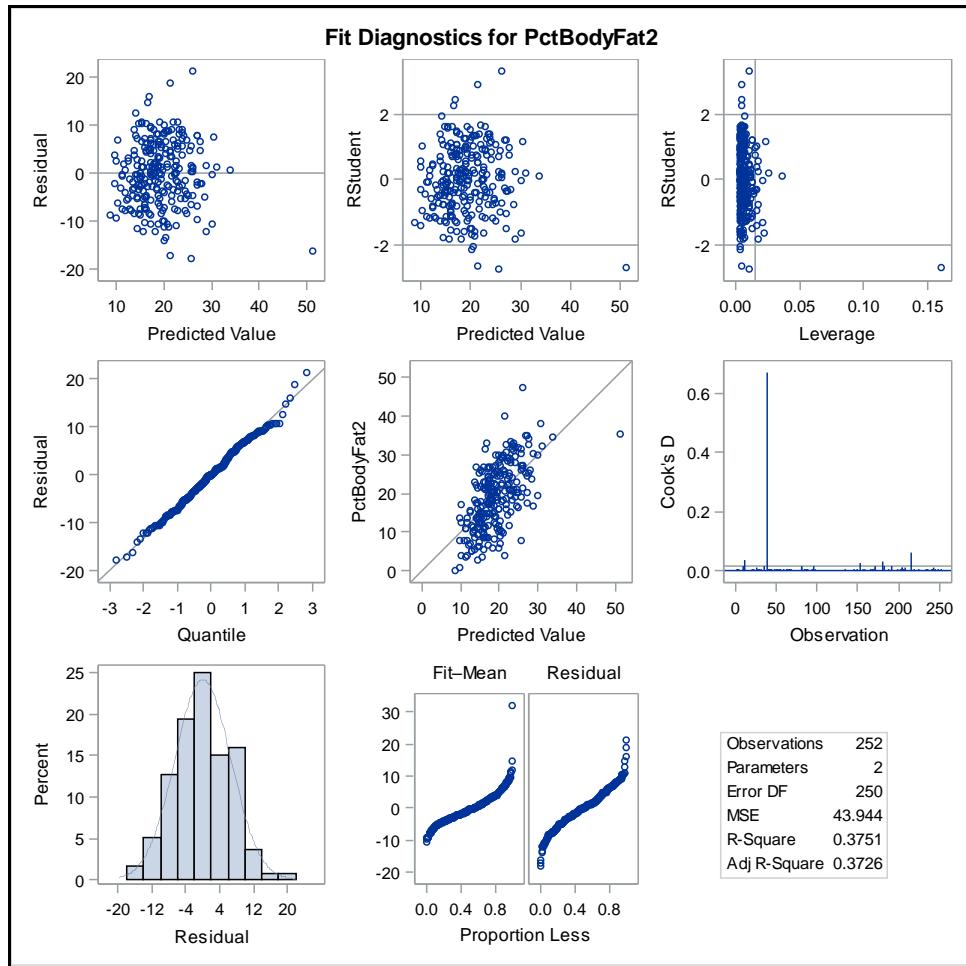
The REG Procedure**Model: MODEL1****Dependent Variable: PctBodyFat2**

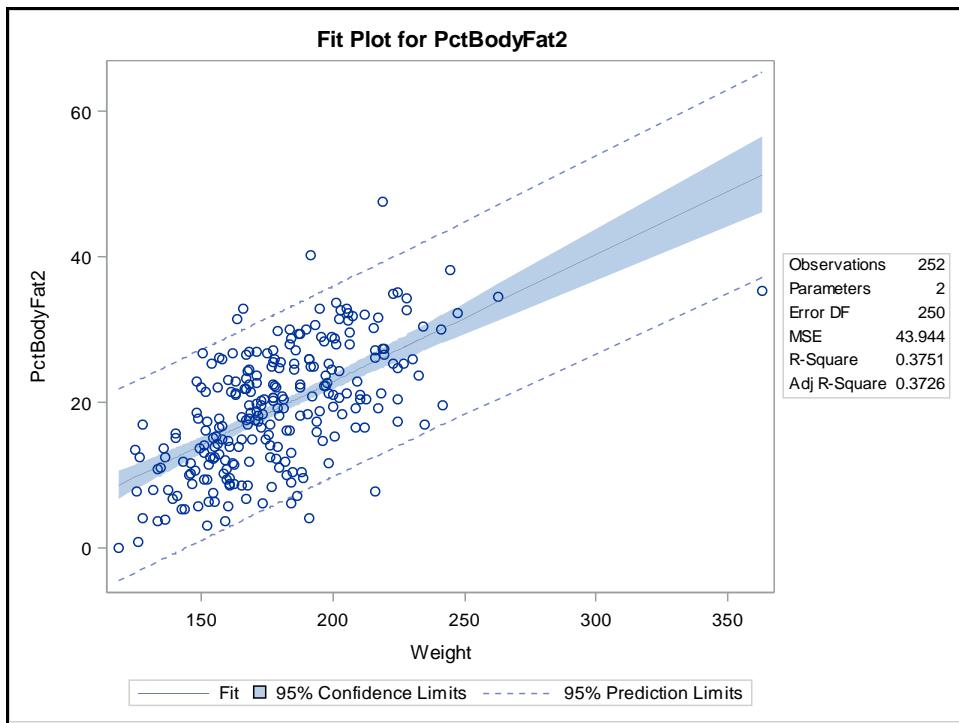
Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6593.01614	6593.01614	150.03	<.0001
Error	250	10986	43.94389		
Corrected Total	251	17579			

Root MSE	6.62902	R-Square	0.3751
Dependent Mean	19.15079	Adj R-Sq	0.3726
Coeff Var	34.61485		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-12.05158	2.58139	-4.67	<.0001
Weight	1	0.17439	0.01424	12.25	<.0001





- a. What is the value of the F statistic and the associated p -value? How would you interpret this with regard to the null hypothesis?

The value of the F statistics is 150.03 ($p < .0001$). Therefore, the null hypothesis of a zero slope for Weight (no linear association) is rejected.

- b. Write the predicted regression equation.

The prediction equation is

$$\text{PctBodyFat2} = -12.05158 + 0.17439 * \text{Weight}$$

- c. What is the value of R-square? How would you interpret this?

The R-square value is 0.3751. Weight explains approximately 37.51% of the variability in PctBodyFat2.

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

2.01 Multiple Answer Poll – Correct Answers

What is the visual cue in a scatter plot that there is no association between the response variable and the explanatory variables?

- a. All of the y values fall on a straight line.
- b. None of the y values fall on a straight line.
- c. All of the y mean values fall on a straight line.
- d. None of the y mean values fall on a straight line.
- e. All of the y values are the same.
- f. All of the y mean values are the same.

However, if the y values are all the same, then there might be problems in your data collection or measurement.

2.02 Multiple Answer Poll – Correct Answers

What does the R-Square value measure?

- a. The correlation between the independent and dependent variables.
- b. The proportion of total sum of squares accounted for by the model.
- c. Model sum of squares over error sum of squares.
- d. Something to do with variability.

2.03 Multiple Choice Poll – Correct Answer

If you have 20 observations in your ANOVA and you calculate the residuals, to which of the following would they sum?

- a. -20
- b. 0
- c. 20
- d. 400
- e. Unable to tell from the information given

39

Copyright © SAS Institute Inc. All rights reserved.

2.04 Multiple Choice Poll – Correct Answer

If you have 20 observations in your ANOVA and you calculate the squared residuals, to which of the following would they sum?

- a. -20
- b. 0
- c. 20
- d. 400
- e. Unable to tell from the information given

41

Copyright © SAS Institute Inc. All rights reserved.

2.05 Multiple Choice Poll – Correct Answer

With a fair coin, your probability of getting heads on one flip is 0.5. If you flip a coin once and got heads, what is the probability of getting heads on the second try?

- a. 0.50
- b. 0.25
- c. 0.00
- d. 1.00
- e. 0.75



2.06 Multiple Choice Poll – Correct Answer

With a fair coin, your probability of getting heads on one flip is 0.5. If you flip a coin twice, what is the probability of getting at least one head out of two?

- a. 0.50
- b. 0.25
- c. 0.00
- d. 1.00
- e. 0.75



2.07 Multiple Answer Poll – Correct Answer

Which of the following statements is/are true?

- a. A Pearson correlation coefficient is a measure of linear association.
- b. A nonsignificant p-value for a Pearson correlation means no relationship.
- c. A negative Pearson correlation indicates a low degree of linear association.
- d. A random cloud of data implies a negative correlation.

2.08 Multiple Choice Poll – Correct Answer

The correlation between tuition and rate of graduation at U.S. colleges is 0.55. What does this mean?

- a. The way to increase graduation rates at your college is to raise tuition.
- b. Increasing graduation rates is expensive, causing tuition to rise.
- c. Students who are richer tend to graduate more often than poorer students.
- d. None of the above.

2.09 Multiple Choice Poll – Correct Answer

Run PROC REG with this MODEL statement: model y=x1;. If the parameter estimate (slope) of x1 is 0, then the best guess (predicted value) of y when x1=13 is which of the following?

- a. 13
- b.** the mean of y
- c. a random number
- d. the mean of x1
- e. 0

Chapter 3 More Complex Linear Models

3.1 Two-Way ANOVA and Interactions	3-3
Demonstration: Two-Way ANOVA	3-6
Demonstration: Two-Way ANOVA with an Interaction	3-22
Exercises.....	3-30
3.2 Multiple Regression	3-31
Demonstration: Fitting a Multiple Linear Regression Model	3-39
Exercises.....	3-46
3.3 Solutions	3-48
Solutions to Exercises	3-48
Solutions to Student Activities (Polls/Quizzes)	3-58

3.1 Two-Way ANOVA and Interactions

Objectives

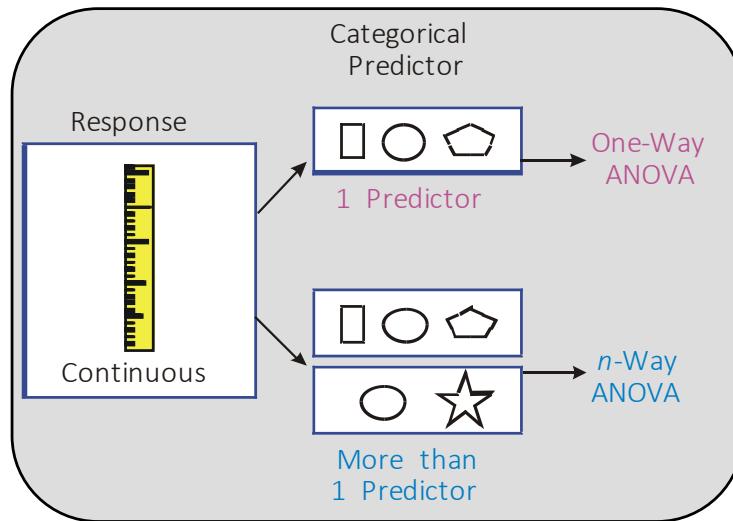
- Fit a two-way ANOVA model.
- Detect interactions between factors.
- Analyze the treatments when there is a significant interaction.

3



Copyright © SAS Institute Inc. All rights reserved.

n-Way ANOVA



4



Copyright © SAS Institute Inc. All rights reserved.

In the previous chapter, you considered the case where you had one categorical predictor variable. In this section, consider a case with two categorical predictors. In general, anytime you have more than one categorical predictor (or explanatory) variable and a continuous response variable, the analysis is called *n*-way ANOVA. The *n* can be replaced with the number of categorical predictor variables.

Additional Linear Models Terminology

- Model – a mathematical relationship between explanatory variables and response variables
- Effect – the expected change in the response that occurs due to the change in the value of an explanatory variable
 - Main Effect – the effect of a single explanatory variable (for example, x_1, x_2, x_3)
 - Interaction Effect – the effect of a simultaneous change of two or more explanatory variables
(for example, $x_1*x_2, x_1*x_2*x_3$)

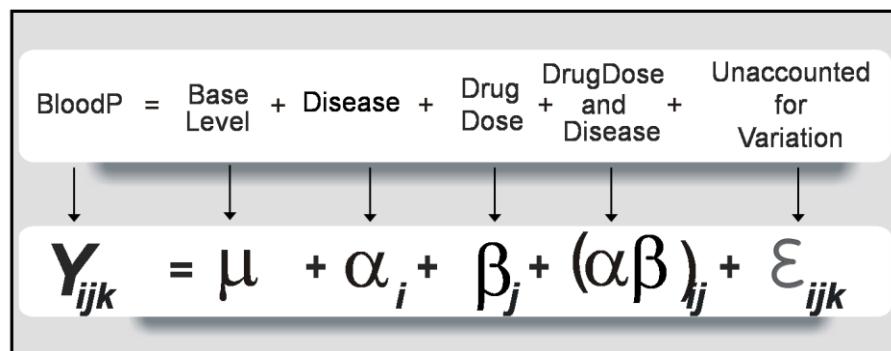
ANOVA and regression are used to estimate parameters in statistical *models*, which are simply the mathematical relationships relating explanatory variables with response variables. The same model can be expressed in a variety of ways, depending on the desired way of communicating the results.

Much of ANOVA terminology comes from the world of design of experiments (DOE) because ANOVA is often used to analyze data obtained from designed experiments.

In this course, you will encounter the term *effect* to mean the magnitude of the expected change in the response variable presumably caused by the change in value of an explanatory term in the model. The terms themselves are often referred to as effects in models. *Main effects* are effects of a single variable, averaged across the levels of all other explanatory variables.

Sometimes there is evidence that explanatory variables do not seem to relate to the response variable in an additive fashion, but rather two or more act jointly on the response variable above and beyond the individual effects of either (any) of the explanatory variables. These effects are called *interaction effects* and are explained in more detail in the next slides.

The Model



6

With each additional predictor variable, a new parameter is introduced to the model. The example in the slide represents a model predicting the change in blood pressure after the administration of a drug.

Notice that the interaction term involves both of the main effects of **DrugDose** and **Disease**. This is also known as a *crossed effect* in experimental design.

Y_{ijk} the observed change in **BloodP** for each subject from before to after treatment

μ the overall base level of the response, **BloodP**

α_i the effect of the i^{th} **Disease**

β_j the effect of the j^{th} **DrugDose**

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

ε_{ijk} error term, or residual

In the model, the following is assumed:

- Observations are independent.
- Error terms are normally distributed for each treatment.
- Variances are equal across treatments.

Note: Verifying ANOVA assumptions with more than one variable is discussed in the Statistics 2: ANOVA and Regression class.



Two-Way ANOVA

Example: Perform a two-way ANOVA of **SalePrice** with **Heating_QC** and **Season_Sold** as predictor variables.

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

1. Open the **Summary Statistics** task under **Statistics** and select the **AmesHousing3** data set.
2. Select **SalePrice** as the **Analysis variables** and **Season_Sold** and **Heating_QC** as the **Classification variables**.
3. On the OPTIONS tab, expand the **Basic Statistics** property and select the option to display **Mean** and **Standard deviation**.
4. Expand the **Additional Statistics** property and select **Variance**.
5. Run the code.

Note: Equivalent SAS code is shown here.

```
/*st103d01.sas*/ /*Part A*/
ods graphics off;
proc means data=STAT1.ameshousing3
    mean std var nway;
    class Season_Sold Heating_QC;
    var SalePrice;
    format Season_Sold Season. ;
    title 'Selected Descriptive Statistics';
run;
```

Selected PROC MEANS statement option:

NWAY When you include CLASS variables, NWAY specifies that the output data set contains only statistics for the observations with the highest _TYPE_ and _WAY_ values. NWAY corresponds to the combination of all class variables.

PROC MEANS Output (Using **FORMAT** statement in SAS code)

Analysis Variable : SalePrice Sale price in dollars						
Season when house sold	Heating quality and condition	N Obs	Mean	Std Dev	Variance	
Winter	Ex	6	145583.33	39738.42	1579141667	
	Fa	3	58100.00	17925.68	321330000	
	Gd	10	124330.00	30580.86	935189000	
	TA	16	121312.50	40979.21	1679295833	
Spring	Ex	41	153765.24	33611.64	1129742652	
	Fa	7	98657.14	21272.19	452506190	
	Gd	18	149619.83	32905.66	1082782633	
	TA	34	129404.41	27701.46	767370965	
Summer	Ex	45	154279.42	35282.20	1244833504	
	Fa	5	128800.00	36507.88	1332825000	
	Gd	22	113727.27	33988.01	1155184935	
	TA	58	134046.55	33743.78	1138642444	
Fall	Ex	15	163726.93	49360.41	2436449681	
	Fa	1	45000.00	.	.	
	Gd	8	143812.50	23398.62	547495536	
	TA	11	129345.45	21507.23	462560727	

The mean sale price is always lowest for houses with fair heating systems. Note that there is only one house in the data set with a fair heating system sold in the fall.

To further explore the numerous treatments, examine the means graphically.

6. Open the **Line Chart** task under **Graph**.
7. Select **Season Sold** as the **Category variable**, **SalePrice** as the **Response variable**, and **Heating_QC** as the **Group variable**.
8. Expand the **Statistics** property and check to plot the **Mean**.
9. On the **OPTIONS** tab, expand the **LINE DETAILS** property and modify the line thickness to **1**.

10. Run the code.

Note: To add markers to the chart for point value (as shown in the following output), use the editor to edit the generated code and specify **markers** options or directly enter the code as shown below.

```
/*st103d01.sas*/ /*Part B*/
proc sgplot data=STAT1.ameshousing3;
  vline Season_Sold / group=Heating_QC
    stat=mean
    response=SalePrice
    markers;
  format Season_Sold season. ;
run;
```

VLINE Creates a vertical line chart (the line is horizontal).

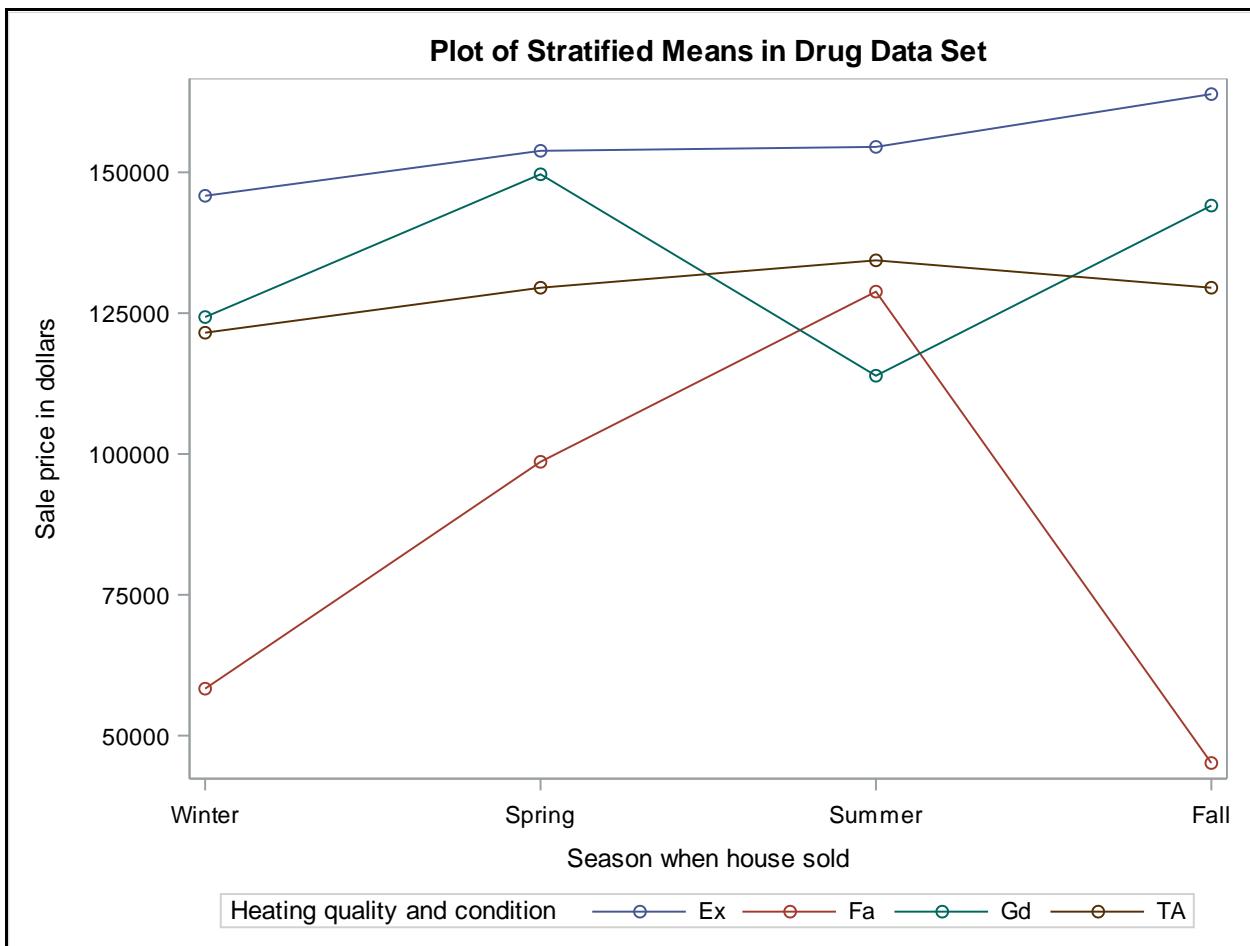
Selected VLINE statement options:

GROUP= Specifies a category variable that is used to group the data. A separate plot is created for each unique value of the grouping variable. The plot elements for each group value are automatically distinguished by different visual attributes.

STAT= specifies the statistic for the vertical axis.

RESPONSE= Specifies a numeric response variable for the plot. The summarized values of the response variable are displayed on the vertical axis.

MARKERS adds data point markers to the series plot data points.



The relationship might be clearer in the graph. The season a home is sold does not seem to affect the sale price very much, except where the heating system is fair. For those homes, the mean sale price seems markedly lower in the colder seasons. This plot is exploratory, and helps you plan your analysis. Later you see similar plotting output directly from PROC GLM.

You can use the GLM procedure to discover the effects of both **Season_Sold** and **Heating_QC**.

11. Open the **N-Way ANOVA** task under Statistics.
12. Select **SalePrice** as the **Dependent variable** and select both **Season_Sold** and **Heating_QC** as the **Factors**.

Note: Order matters in selecting the Factors here. Selection order determines the generated code for the CLASS statement. If you add **Heating_QC** first and **Season_Sold** second, the graphs produced will differ from the presented output.

The screenshot shows the SAS Studio interface. On the left, there's a sidebar with 'Files and Folders' and a tree view of 'Tasks and Utilities' under 'Statistics'. The main workspace has two tabs at the top: 'DATA' and 'MODEL'. The 'DATA' tab is active, showing a dataset named 'STAT1.AMESHOUSING3' and a filter option '(none)'. Below that, under 'ROLES', there's a section for 'Dependent variable' containing 'SalePrice' and a section for 'Factors' containing 'Season_Sold' and 'Heating_QC'. The 'MODEL' tab is also visible.

13. On the MODEL tab, specify the model by clicking on the Edit button to open the Model Effects Builder. Add **Heating_QC** and **Season_Sold** to the Model Effects.
14. Run the code.

This screenshot shows the 'Model Effects' dialog box. At the top, there's a toolbar with buttons for 'DATA', 'MODEL' (which is selected), 'OPTIONS', and other tabs. Below that, there's a section titled 'MODEL EFFECTS' with a 'Model' entry. Under 'Model Effects', there's a list box containing 'Intercept', 'Heating_QC', and 'Season_Sold'. An 'Edit' button is located above the list box.

Note: Order matters in selecting the Factors here. If you add **Season_Sold** first and **Heating_QC** second, the statistics reported will differ from the presented output.

Note: Alternatively, you can write the code directly.

```
/*st103d01.sas*/ /*Part C*/
ods graphics on;
proc glm data=STAT1.ameshousing3 order=internal;
  class Season_Sold Heating_QC;
  model SalePrice=Heating_QC Season_Sold;
  lsmeans Season_Sold / diff adjust=tukey;
  format Season_Sold season. ;
  title "Model with Heating Quality and Season as Predictors";
run;
```

Selected PROC GLM option:

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sorting order for the levels of all classification variables. The ordering is important for the plot in this case.

PROC GLM Output

Class Level Information		
Class	Levels	Values
Season_Sold	4	1 2 3 4
Heating_QC	4	Ex Fa Gd TA

Number of Observations Read	300
Number of Observations Used	300

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	72774816066	12129136011	10.14	<.0001
Error	293	350448703445	1196070660.2		
Corrected Total	299	423223519511			

The global test is of the null hypothesis that all means are equal for all explanatory variables. Notice that the number of degrees of freedom for this test is 6. Both **Season_Sold** and **Heating_QC** account for three degrees of freedom (number of categories minus 1). The statistically significant *p*-value indicates that not all means are equal for all explanatory variables. It does not indicate which mean values differ.

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.171954	25.14764	34584.25	137524.9

The R-Square value of 0.171954 shows that about 17% of the variability in **SalePrice** is explained by the two categorical predictors.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.63	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

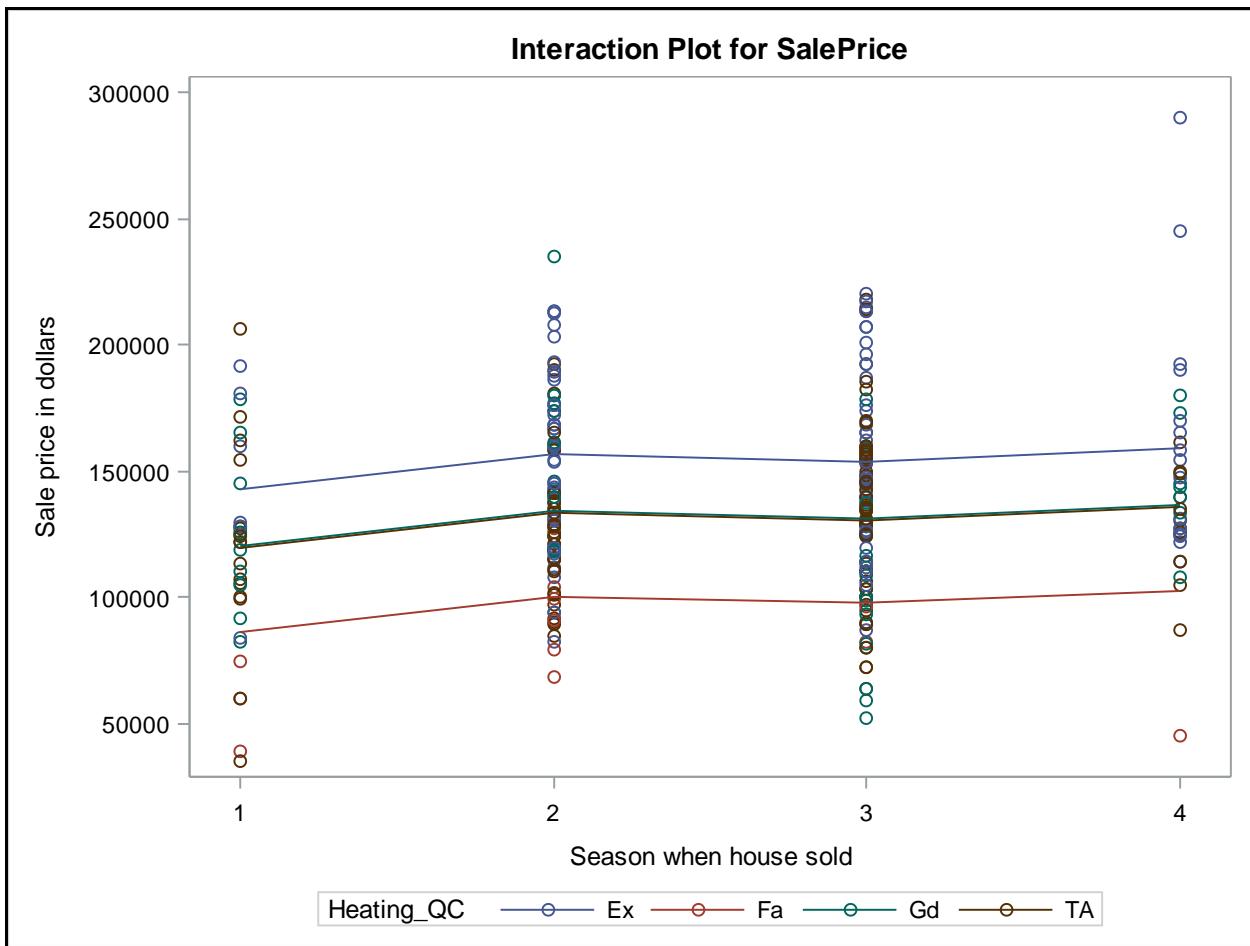
In this case, the test of **Heating_QC** is an unadjusted test, because no other terms are above it, whereas the **Season_Sold** test adjusts for **Heating_QC**, which appears before it. The model specification determines the ordering in this table.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	60050783038	20016927679	16.74	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

The Type III sums of squares are commonly called *partial sums of squares*. The Type III sum of squares for a particular variable is the increase in the model sum of squares due to adding the variable to a model that already contains all the other variables in the model. Type III sums of squares, therefore, do not depend on the order in which the explanatory variables are specified in the model. The Type III SS values are not generally additive (except in a completely balanced design, where all categories of all inputs contain exactly the same sample size). The values do not necessarily sum to the Model SS.

In this example, the Type I effects and Type III effects differ only slightly. There seems to be no significant differences across all levels of the **Season_Sold** variable, whereas there are differences across the **Heating_QC** variable.

You will generally interpret and report results based on the Type III SS.



This plot differs from the exploratory plot because it imposes a main effects model on the data. In other words, the effect of each variable is not permitted to differ at different levels of the other variable. That constraint can be relaxed by adding an interaction term.

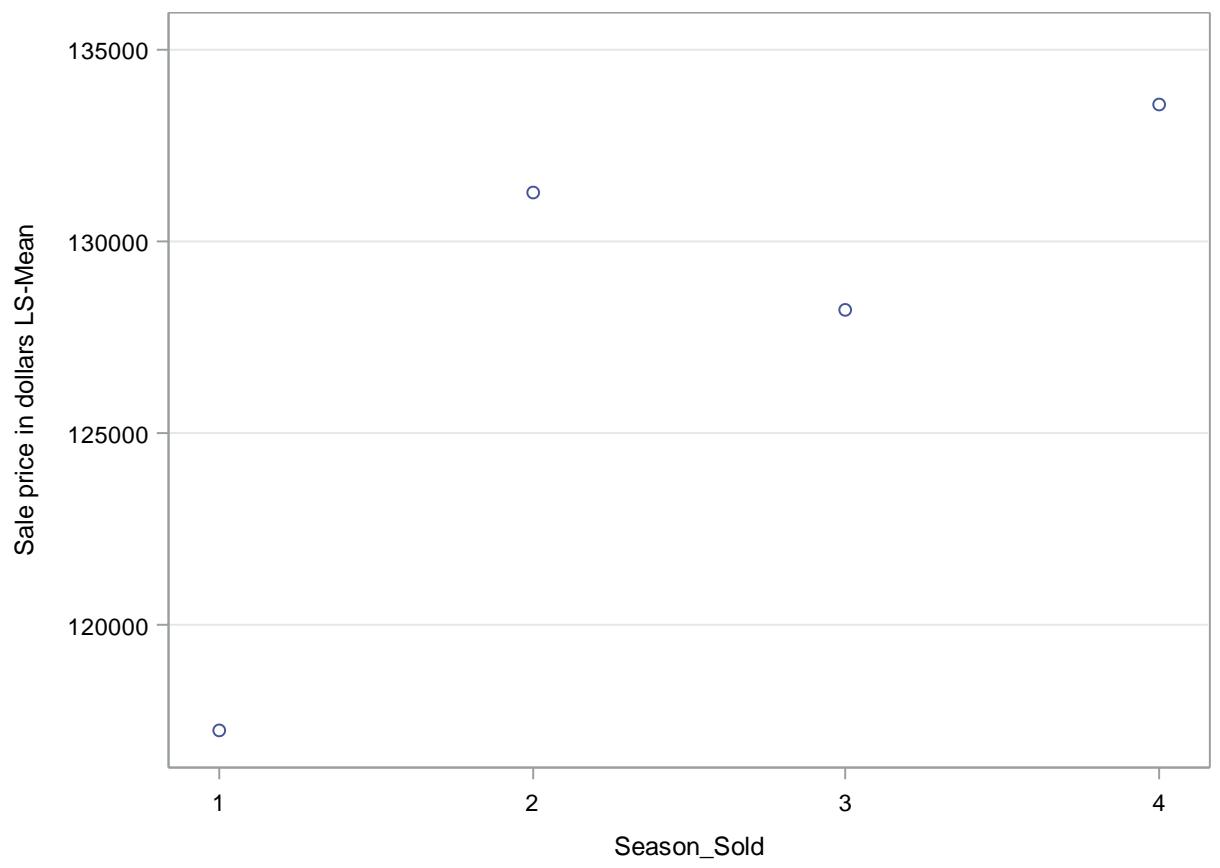
Season_Sold	SalePrice LSMEAN	LSMEAN Number
1	117255.605	1
2	131263.281	2
3	128216.231	3
4	133543.394	4

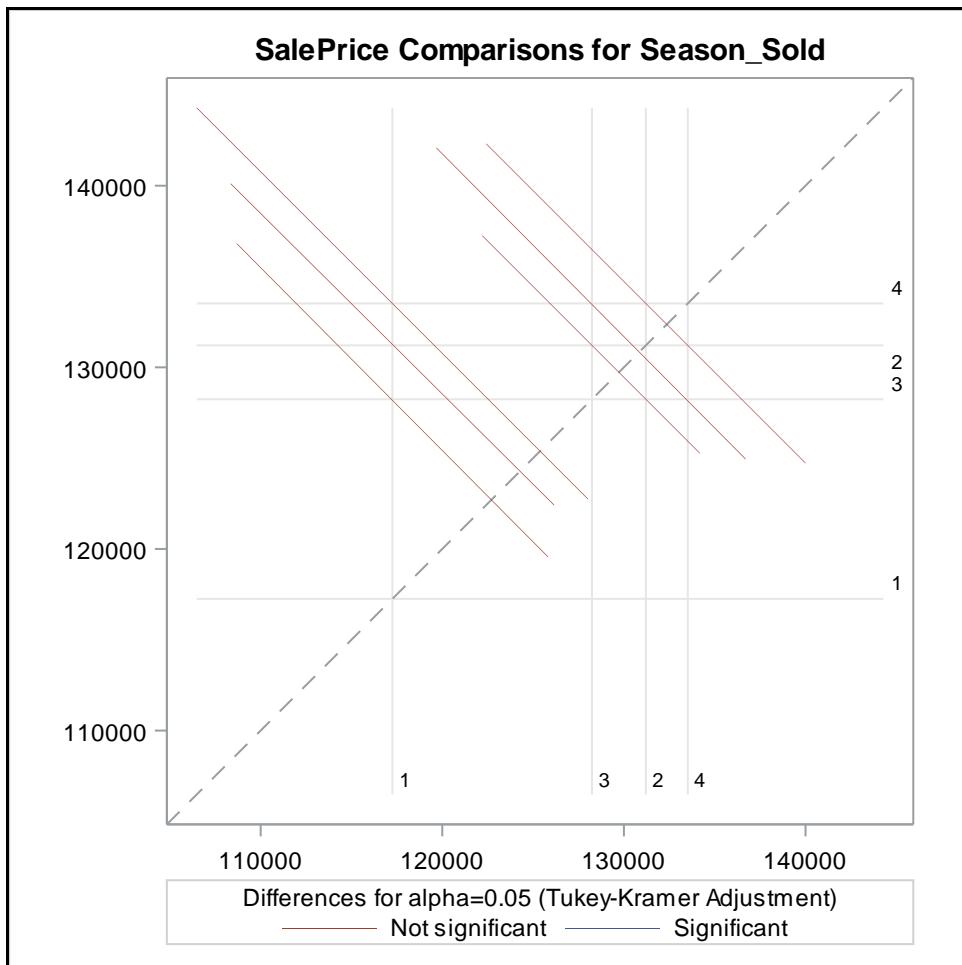
Least Squares Means for effect Season_Sold
 $Pr > |t| \text{ for } H_0: LS\text{Mean}(i) = LS\text{Mean}(j)$

Dependent Variable: SalePrice

i\j	1	2	3	4
1		0.1760	0.3529	0.2089
2	0.1760		0.9124	0.9870
3	0.3529	0.9124		0.8517
4	0.2089	0.9870	0.8517	

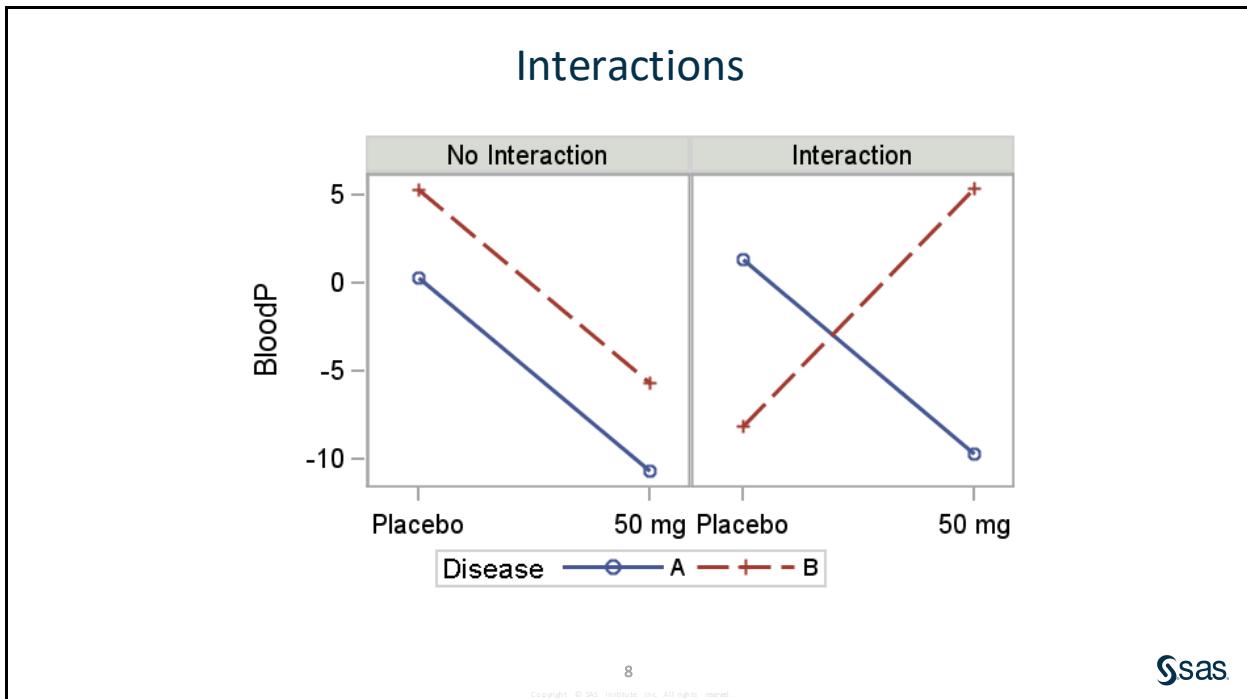
LS-Means for Season_Sold





The LSMEANS statement requests a Tukey-adjusted analysis of the difference across all seasons. There are no significant differences.

End of Demonstration



An *interaction* occurs when the differences between group means on one variable change at different levels of another variable.

The average blood pressure change over different doses was plotted in mean plots and then connected for disease A and B.

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

In the right plot, however, as the dose increases, average blood pressure **decreases** for those with disease A, but **increases** for those with disease B. This indicates an interaction between the variables **DrugDose** and **Disease**.

When you analyze an *n*-way ANOVA with interactions, you should first look at any tests for interaction among factors.

If there is no interaction between the factors, the tests for the individual factor effects (main effects) can be interpreted as true effects of that factor.

If an interaction exists between any factors, the tests for the individual factor effects might be misleading. These are known as tests of *marginal effects* and only tell part of the story about the overall effect of that variable.

What to Do with a Nonsignificant Interaction

Analyze the main effects with the interaction in the model.

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

...or...

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

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

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

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

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

- There are fewer than five degrees of freedom for the error.
- The F value for the interaction term is < 2 .

Note: When you analyze data from an observational study, it is more common to delete the non-significant interaction from the model and then analyze the main effects.

PROC GLM Interaction Specification

```
PROC GLM ...;
  MODEL A B A*B;
  or
  MODEL A|B;
RUN;
QUIT;
```

10

Copyright © SAS Institute Inc. All rights reserved.



An interaction term can be specified in PROC GLM using the * operator between two listed variables. An alternate way of specifying a full factorial model is through the use of the bar operator (|). You can shorten the specification of a large factorial model by using the bar operator. For example, two ways of writing the model for a full three-way factorial model follow:

model Y=A B C A*B A*C B*C A*B*C;

model Y=A|B|C;

When the bar (|) is used, the right and left sides become effects, and the cross of them becomes an effect. Multiple bars are permitted.

You can also specify the maximum number of variables involved in any effect that results from bar evaluation by specifying that maximum number, preceded by an @ sign, at the end of the bar effect. For example, the specification A | B | C @2 would result in only those effects that contain 2 or fewer variables: in this case, A B A*B C A*C and B*C.

STORE Statement in PROC GLM

```
STORE <OUT=>item-store-name
      </ LABEL='label'>;
```

- The STORE statement requests that the procedure save the context and results of the statistical analysis.
- The resulting item store has a binary file format that cannot be modified.
- The contents of the item store can be processed with the PLM procedure.

11

Copyright © SAS Institute Inc. All rights reserved.



The STORE statement applies to the following SAS/STAT procedures: GENMOD, GLIMMIX, GLM, GLMSELECT, LOGISTIC, MIXED, ORTHOREG, PHREG, PROBIT, SURVEYLOGISTIC, SURVEYPHREG, and SURVEYREG. This statement requests that the procedure save the context and results of the statistical analysis into an item store. An *item store* is a binary file format that cannot be modified by the user. The contents of the item store can be processed with the PLM procedure.

One example of item-store use is to perform a time-consuming analysis and to store its results by using the STORE statement. At a later time, you can then perform specific statistical analysis tasks based on the saved results of the previous analysis, without having to fit the model again.

In the STORE statement:

item-store-name is a usual one- or two-level SAS name, similar to the names that are used for SAS data sets. If you specify a one-level name, then the item store resides in the Work library and is deleted at the end of the SAS session. Because item stores usually are used to perform postprocessing tasks, typical usage specifies a two-level name of the form *libname.membername*.

label identifies the estimate on the output. A label is optional but must be enclosed in quotation marks.

The PLM Procedure

General form of the PLM procedure:

```
PROC PLM RESTORE=item-store-specification<options>;
  EFFECTPLOT <plot-type <(plot-definition options)>>
    </ options> ;
  LSMEANS <model-effects> </ options>;
  LSMESTIMATE model-effect <'label'> values
    <divisor=n>, ...<'label'> values
    <divisor=n> </ options>;
  SHOW options;
  SLICE model-effect </ options>;
  WHERE expression;
RUN;
```

12

Copyright © SAS Institute Inc. All rights reserved.



The PLM procedure performs post-fitting statistical analyses and plotting for the contents of a SAS item store that were previously created with the STORE statement in some other SAS/STAT procedure.

The statements that are available in the PLM procedure are designed to reveal the contents of the source item store via the Output Delivery System (ODS) and to perform post-fitting tasks.

The use of item stores and PROC PLM enables you to separate common post-processing tasks, such as testing for treatment differences and predicting new observations under a fitted model, from the process of model building and fitting. A numerically expensive model fitting technique can be applied once to produce a source item store. The PLM procedure can then be called multiple times and the results of the fitted model are analyzed without incurring the model fitting expenditure again.

Selected PROC PLM option:

RESTORE specifies the source item store for processing.

Selected PROC PLM procedure statements:

EFFECTPLOT The EFFECTPLOT statement produces a display of the fitted model and provides options for changing and enhancing the displays.

LSMEANS computes and compares least squares means (LS-means) of fixed effects.

LSMESTIMATE provides custom hypothesis tests among least squares means.

SHOW uses the Output Delivery System to display contents of the item store. This statement is useful for verifying that the contents of the item store apply to the analysis and for generating ODS tables.

SLICE The SLICE statement provides a general mechanism for performing a partitioned analysis of the LS-means for an interaction. This analysis is also known as an analysis of simple effects. The SLICE statement uses the same options as the LSMEANS statement.

WHERE is used in the PLM procedure when the item store contains BY-variable information and you want to apply the PROC PLM statements to only a subset of the BY groups.



Two-Way ANOVA with an Interaction

Example: Perform a two-way ANOVA of **SalePrice** with **Heating_QC** and **Season_Sold** as predictor variables. Include the interaction between the two explanatory variables.

1. Open the **N-Way ANOVA** task under **Statistics**.
2. On the **DATA** tab, select **AmesHousing3** data set, assign **SalePrice** as the Dependent variable and **Heating_QC** and **Season_Sold** (in that order) as the Factors.
3. On the **MODEL** tab, the interaction term **Heating_QC*Season_Sold** can be specified by using the **Cross** button. Alternatively, select both variables and click the **Full Factorial** button to include all three terms, **Season_Sold**, **Heating_QC**, and **Season_Sold*Heating_QC**, in the model.

Model Effects Builder

Variables:

- Heating_QC
- Season_Sold

Model effects:

- Intercept
- Heating_QC
- Season_Sold
- Heating_QC*Season_Sold

Single Effects

Add Cross (circled in red)

Nest

Standard Models

Full Factorial (circled in red) N-way Factorial

OK Cancel

4. On the **OPTIONS** tab, select the **Default and additional statistics** option and clear the **Perform multiple comparisons** option.

5. Run the code.

The code generated is:

```
ods graphics / imagemap=on;

proc glm data=ST141.AMESHOUSING3;
  class Heating_QC Season_Sold;
  model SalePrice=Heating_QC Season_Sold Heating_QC*Season_Sold /
ss1 ss3;
quit;
```

Partial PROC GLM Output

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	97609874155	6507324943.7	5.68	<.0001
Error	284	325613645356	1146526920.3		
Corrected Total	299	423223519511			

The number of degrees of freedom for the model is now 15. This includes 3 degrees of freedom for each of the main effects and $3 \times 3 = 9$ degrees of freedom for the interaction term. The model is statistically significant.

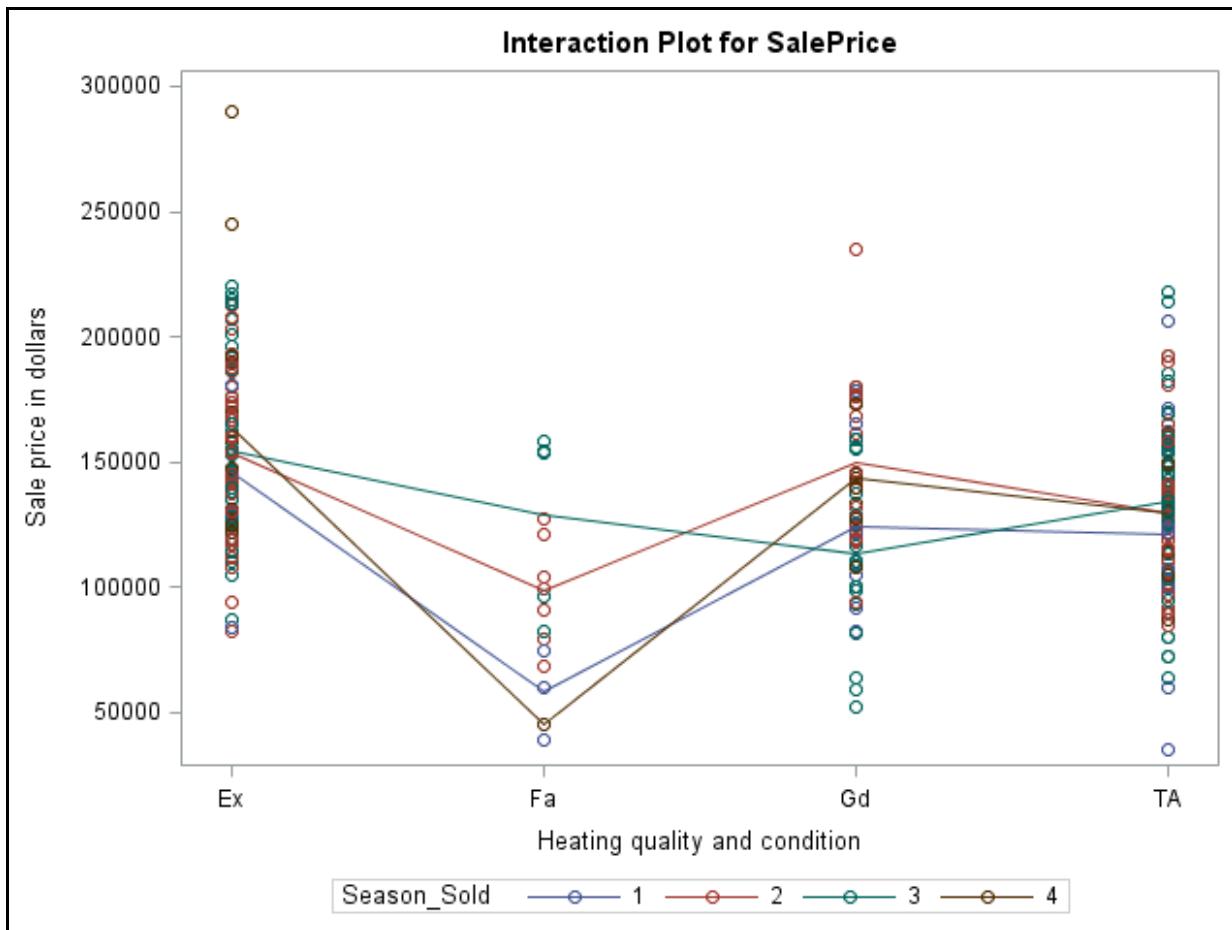
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.230634	24.62130	33860.40	137524.9

The R-Square for this model is 0.230634, which means that about 23% of the variability in **SalePrice** is explained.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	19.43	<.0001
Season_Sold	3	5939259845	1979753282	1.73	0.1617
Heating_Q*Season_Sol	9	24835058089	2759450899	2.41	0.0121

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	51116493768	17038831256	14.86	<.0001
Season_Sold	3	9318181844	3106060615	2.71	0.0455
Heating_Q*Season_Sol	9	24835058089	2759450899	2.41	0.0121

The interaction effect is statistically significant at the 0.05 alpha level. This means that the effect of **Season_Sold** differs at different levels of **Heating_QC** and that the effect of **Heating_QC** differs at different levels of **Season_Sold**. In the main effect model it appeared that **Season_Sold** was not related to **SalePrice**. Now it can be seen that it is related, but in a more complex way.



The interaction model is able to reflect the data more accurately than the main effects model.

Note: Depending on the order of the predictor variables in the CLASS statement, the interaction plot can be switched where each line is a heating quality and condition and the season sold is on the x-axis.

The SLICE= option in the LSMEANS statement enables you to look at the effect of **Season_Sold** at all levels of **Heating_QC**.

Note: Currently the SLICE= option needs to be specified manually in SAS Studio using the edit function and modify generated code. Alternatively, you can write the code directly in SAS.

```
/*st103d02.sas*/ /*Part A*/
ods graphics on;

proc glm data=STAT1.ameshousing3
    order=internal
    plots(only)=intplot;
class Season_Sold Heating_QC;
model SalePrice=Heating_QC Season_Sold Heating_QC*Season_Sold;
lsmeans Heating_QC*Season_Sold / diff slice=Heating_QC;
format Season_Sold Season. ;
store out=interact;
title "Model with Heating Quality and Season as Interacting "
```

```
"Predictors";
run;
```

Selected LSMEANS statement option:

SLICE= Specifies effects by which to partition interaction LSMEANS effects.

Season_So*Heating_QC Effect Sliced by Heating_QC for SalePrice					
Heating_QC	DF	Sum of Squares	Mean Square	F Value	Pr > F
Ex	3	1759608339	586536113	0.51	0.6746
Fa	3	12318827232	4106275744	3.58	0.0143
Gd	3	14560964166	4853654722	4.23	0.0060
TA	3	2134918196	711639399	0.62	0.6021

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

The tests displayed are of **Season_Sold** within each slice of **Heating_QC**. There is a larger seasonal effect for houses with fair and good heating systems. The note reminds you that these p-values are not adjusted for multiple tests. Adjustment is possible using PROC PLM.

Note: Currently the N-Way ANOVA task does not include the SLICE= option for multiple comparisons tests. The code below uses the output from the ANOVA analysis to adjust for multiple tests.

```
/*st103d02.sas*/ /*Part B*/
proc plm restore=interact plots=all;
  slice Heating_QC*Season_Sold / sliceby=Heating_QC adjust=tukey;
  effectplot interaction(sliceby=Heating_QC) / clm;
run;
```

Partial PROC PLM Output

Store Information	
Item Store	WORK.INTERACT
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLM
Date Created	02JUL14:10:35:36
Response Variable	SalePrice
Class Variables	Season_Sold Heating_QC
Model Effects	Intercept Heating_QC Season_Sold Season_So*Heating_QC

Class Level Information		
Class	Levels	Values
Season_Sold	4	Winter Spring Summer Fall
Heating_QC	4	Ex Fa Gd TA

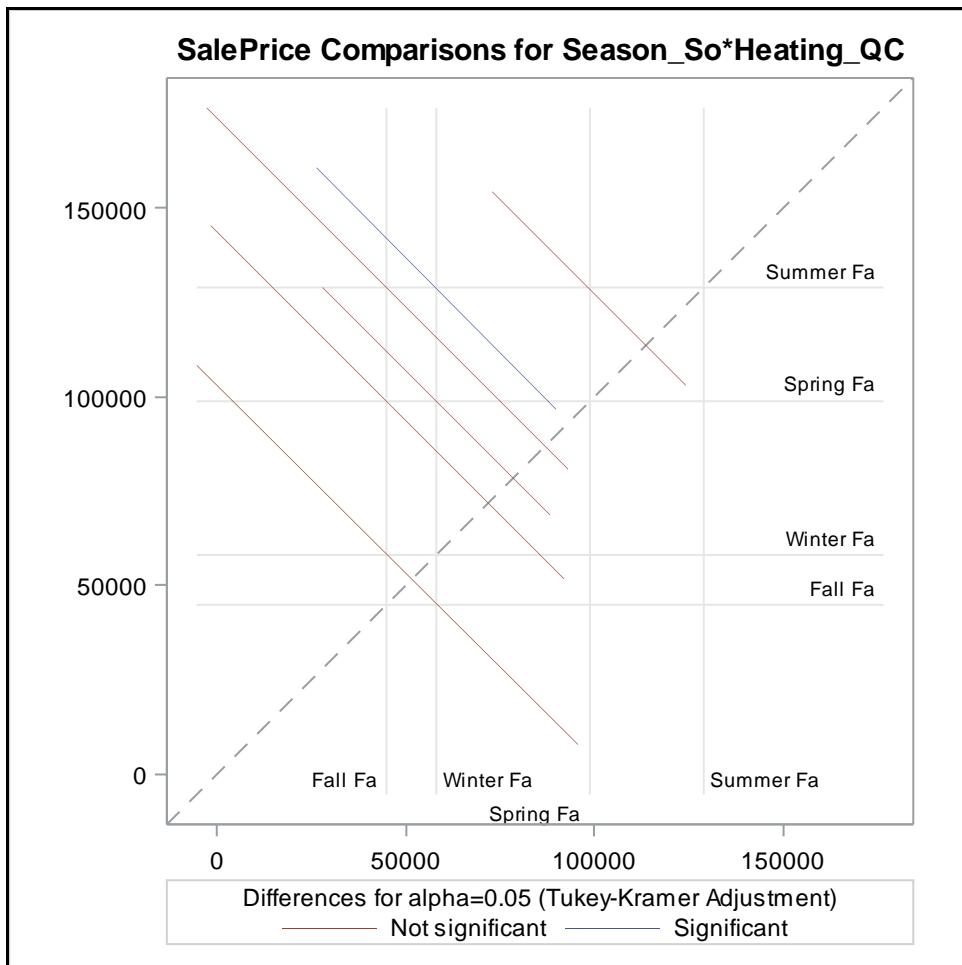
F Test for Season_So*Heating_QC Least Squares Means Slice					
Slice	Num DF	Den DF	F Value	Pr > F	
Heating_QC Ex	3	284	0.51	0.6746	

The slice of excellent heating systems shows that there is no significant effect of season. So, skip to the analysis fair heating systems:

F Test for Season_So*Heating_QC Least Squares Means Slice					
Slice	Num DF	Den DF	F Value	Pr > F	
Heating_QC Fa	3	284	3.58	0.0143	

Note that the *F* test is not adjusted. However, the pairwise tests are adjusted.

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer									
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > t	Adj P	
Heating_QC Fa	Winter	Spring	-40557	23366	284	-1.74	0.0837	0.3071	
Heating_QC Fa	Winter	Summer	-70700	24728	284	-2.86	0.0046	0.0235	
Heating_QC Fa	Winter	Fall	13100	39099	284	0.34	0.7378	0.9870	
Heating_QC Fa	Spring	Summer	-30143	19827	284	-1.52	0.1295	0.4267	
Heating_QC Fa	Spring	Fall	53657	36198	284	1.48	0.1394	0.4495	
Heating_QC Fa	Summer	Fall	83800	37092	284	2.26	0.0246	0.1102	

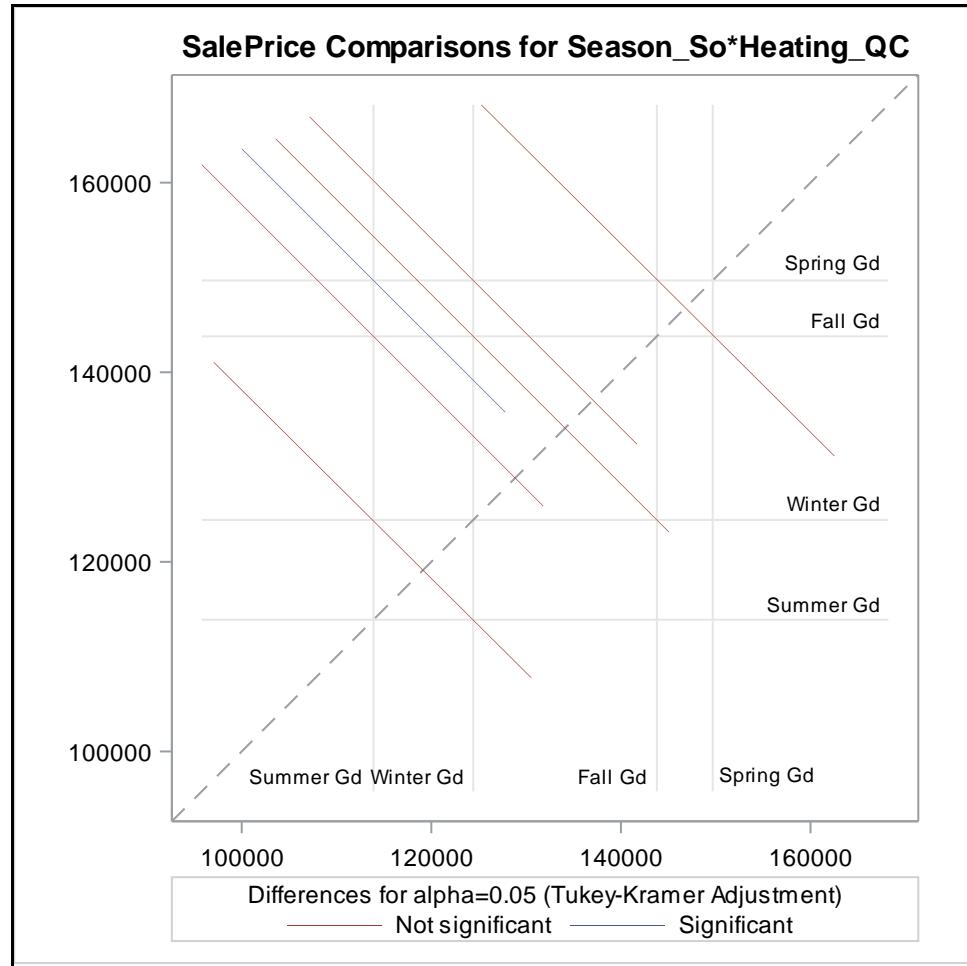


For fair systems, the only statistically significant pairwise comparison is between summer and winter.

F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Gd	3	284	4.23	0.0060

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer									
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > t	Adj P	
Heating_QC Gd	Winter	Spring	-25290	13355	284	-1.89	0.0593	0.2330	
Heating_QC Gd	Winter	Summer	10603	12914	284	0.82	0.4123	0.8445	
Heating_QC Gd	Winter	Fall	-19483	16061	284	-1.21	0.2261	0.6191	
Heating_QC Gd	Spring	Summer	35893	10762	284	3.34	0.0010	0.0053	

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > t	Adj P
Heating_QC Gd	Spring	Fall	5807.33	14388	284	0.40	0.6868	0.9777
Heating_QC Gd	Summer	Fall	-30085	13980	284	-2.15	0.0322	0.1394



There was a significant mean sale price difference for houses with good heating systems between the spring months and the winter months.

F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC TA	3	284	0.62	0.6021

End of Demonstration

3.01 Multiple Answer Poll

How can you recognize an interaction?

- a. The effect of one variable differs at different levels of another.
- b. The joint effect of two variables is greater than the sum of their individual effects.
- c. A stratified plot of the effect of one variable against the dependent variable shows different patterns across different levels of the other.
- d. Two variables are talking at a party.
- e. One variable shows statistical significance without the other in the model, but no significance with the other variable in the model.

14

Copyright © SAS Institute Inc. All rights reserved.



Exercise: Drug Example

The purpose of the study is to look at the effect of a new prescription drug on blood pressure.



17

Copyright © SAS Institute Inc. All rights reserved.





Exercises

1. Performing Two-Way ANOVA

Data were collected in an effort to determine whether different dose levels of a given drug have an effect on blood pressure for people with one of three types of heart disease. The data are in the **STAT1.Drug** data set.

The data set contains the following variables:

DrugDose dosage level of drug (1, 2, 3, 4), corresponding to (Placebo, 50 mg, 100 mg, 200 mg)

Disease heart disease category

BloodP change in diastolic blood pressure after 2 weeks treatment

- a. Examine the data with a vertical line plot. Put **BloodP** on the Y axis, **DrugDose** on the X axis, and then stratify by **Disease**. What information can you obtain from looking at the data?
- b. Test the hypothesis that the means are equal, making sure to include an interaction term if the graphical analyses that you performed indicate that would be advisable. What conclusions can you reach at this point?

End of Exercises

3.2 Multiple Regression

Objectives

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

20

Copyright © SAS Institute Inc. All rights reserved.



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.

21

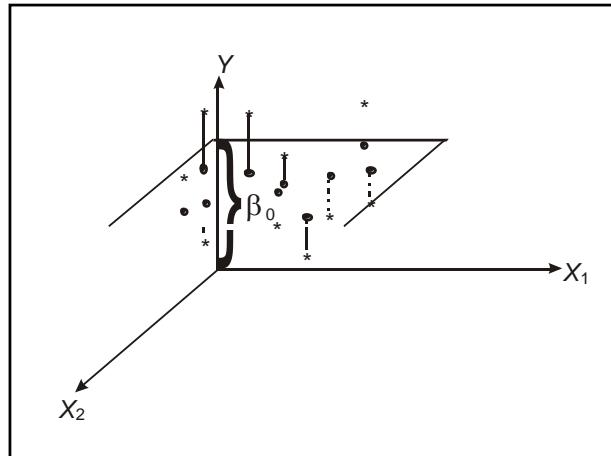
Copyright © SAS Institute Inc. All rights reserved.



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

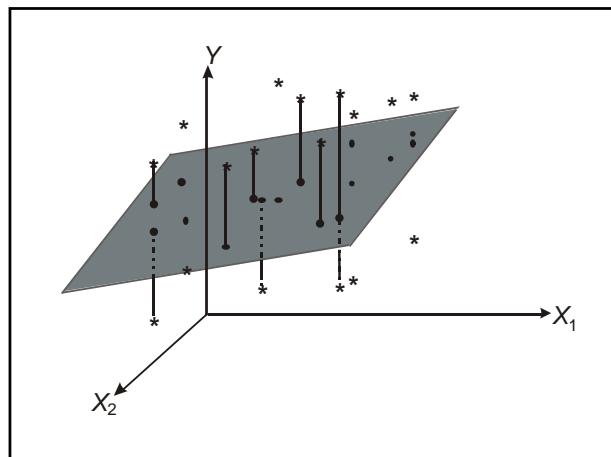


22

Copyright © SAS Institute Inc. All rights reserved.

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



23

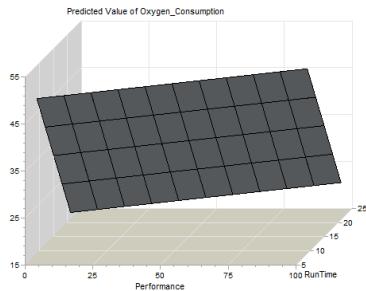
Copyright © SAS Institute Inc. All rights reserved.

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

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

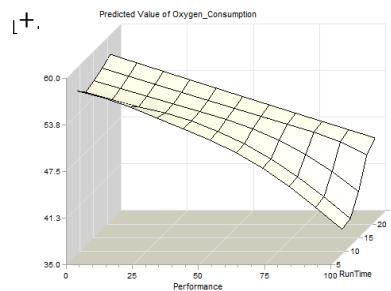
The Multiple Linear Regression Model

In general, you model the dependent variable, Y , as a linear function of k independent variables, X_1 through X_k :



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

Linear Model with
only Linear Effects



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \varepsilon$$

Linear Model with
Nonlinear Effects

24

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 Xs or cross products of one X with another, surfaces more complex than planes can be modeled. It should be noted, though, that these are still linear models because the response variable is related to the predictor terms by a linear function.

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

Note: 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.

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 differ from zero. The predictor variables do not 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 differs from zero. At least one predictor variable explains a significant amount of variability in the response variable.

Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a linear function of the Xs.
- The random error term, 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.

26

Copyright © SAS Institute Inc. All rights reserved.



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 do the following:

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

27

Copyright © SAS Institute Inc. All rights reserved.



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 on the purposes of the analysis, as well as subject-matter expertise.

Common Applications of Multiple Regression

Multiple linear regression is a powerful tool for the following tasks:

- 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

Even though multiple linear regression enables you to analyze many experimental designs, ranging from simple to complex, you 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 is probably a good analytic model. Conversely, a model developed for an analytic study is probably a good prediction model.

Myers (1999) refers to four applications of regression:

- prediction
- variable screening
- model specifications
- 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 this formula:

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

29

Copyright © SAS Institute Inc. All rights reserved.



Most investigators whose main goal is prediction 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.

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

30

Copyright © SAS Institute Inc. All rights reserved.



3.02 Multiple Answer Poll

Based on the definitions outlined in the previous slides, which type of modeler would you primarily consider yourself?

- a. Predictive.
- b. Explanatory.
- c. Both.
- d. I do not know yet.
- e. It's a secret.

31

Copyright © SAS Institute Inc. All rights reserved.



Adjusted R Square

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

$i=1$ 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

32

Copyright © SAS Institute Inc. All rights reserved.



The R-square always increases or stays the same as you include more terms in the model. Therefore, choosing the “best” model is not as simple as just making the R-square as large as possible.

The adjusted R-square is a measure similar to R-square, but it takes into account the number of terms in the model. It can be thought of as a penalized version of R-square with the penalty increasing with each parameter added to the model.



Fitting a Multiple Linear Regression Model

Example: Perform a regression model of **SalePrice** with **Lot_Area** and **Basement_Area** as predictor variables.

1. Open the **Linear Regression** task under Statistics and select the **AmesHousing3** data set.
2. Assign **SalePrice** as the Dependent variable and **Basement_Area** and **Lot_Area** as the Continuous variables.
3. On the MODEL tab, click the Edit button and specify the model by adding the two factors.
4. On the OPTIONS tab, expand the Scatter Plots area and uncheck the option to display observed values by predicted values scatter plot.
5. Run the code.

Note: Alternatively, you can write the code directly in SAS.

```
/*st103d03.sas*/ /*Part A*/
ods graphics on;

proc reg data=STAT1.ameshousing3;
  model SalePrice=Basement_Area Lot_Area;
  title "Model with Basement Area and Lot Area";
run;
```

PROC REG Output

Number of Observations Read	300
Number of Observations Used	300

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.032206E11	1.016103E11	137.17	<.0001
Error	297	2.200029E11	740750509		
Corrected Total	299	4.232235E11			

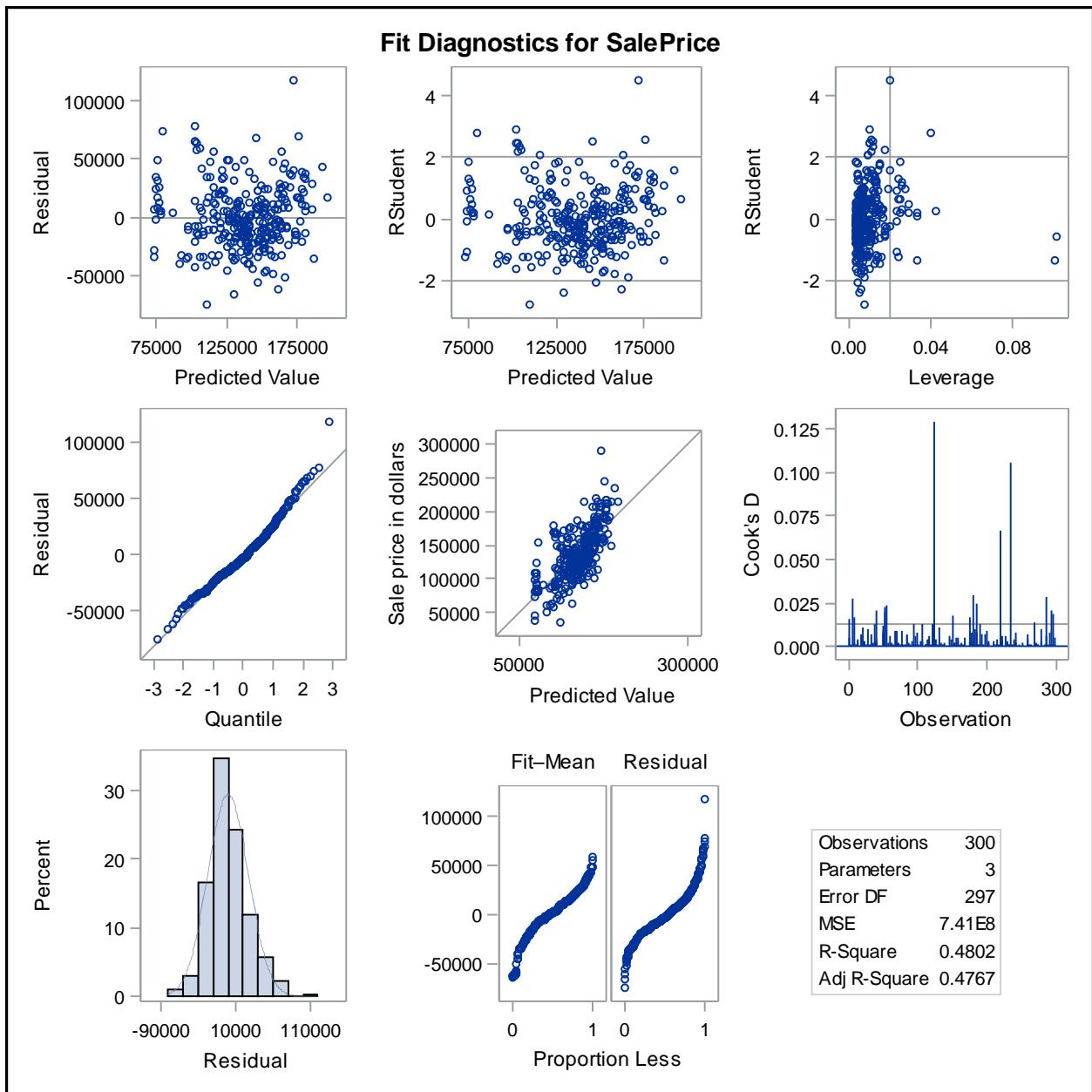
The model is statistically significant at the 0.05 alpha level.

Root MSE	27217	R-Square	0.4802
Dependent Mean	137525	Adj R-Sq	0.4767
Coeff Var	19.79041		

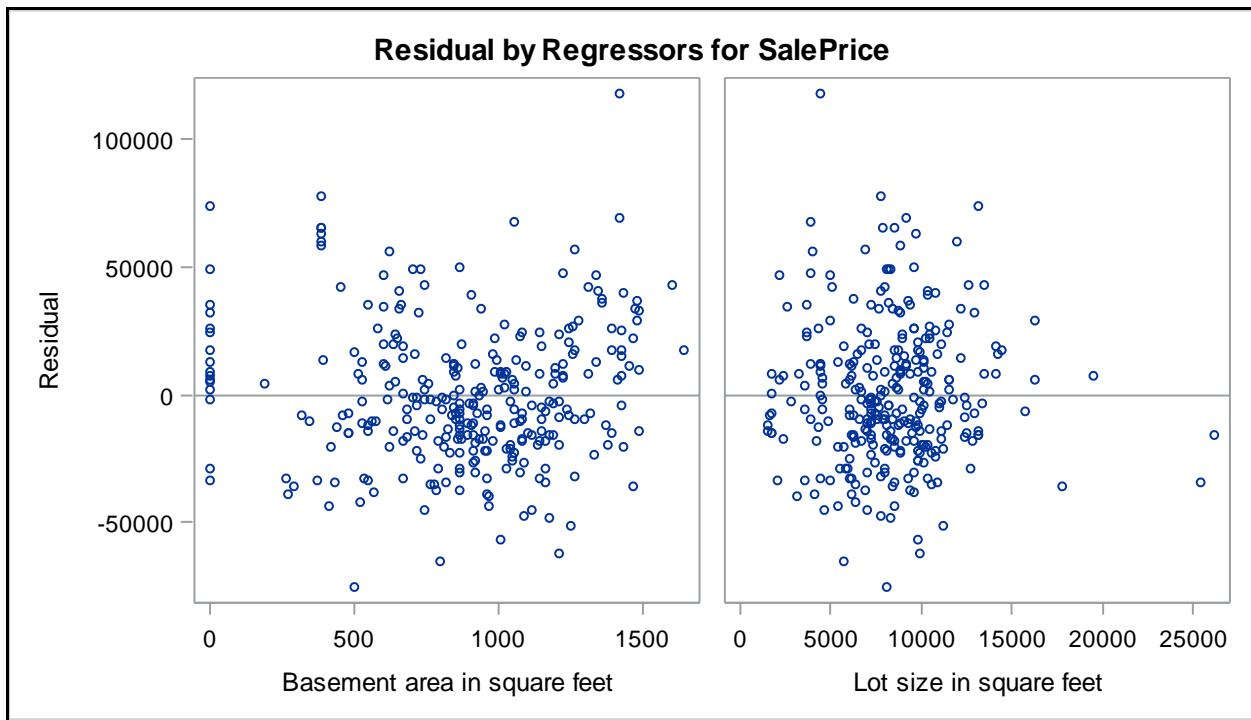
The R-Square value is much greater than for the model that included only **Lot_Area**. It is now 0.4802. The adjusted R-Square is also greater than in the simple model.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	69016	5129.52179	13.45	<.0001
Basement_Area	Basement area in square feet	1	70.08680	4.54618	15.42	<.0001
Lot_Area	Lot size in square feet	1	0.80430	0.49210	1.63	0.1032

In this model, **Lot_Area** is no longer significant. The parameter estimate for each of the explanatory variables adjust for the other variable in the model. The **Lot_Area** estimate is notably different than it was in the simple regression model (2.87 in the simple regression model and 0.80 in this model) and its p-value no longer shows statistical significance.



The quantile-quantile plot of residuals do not indicate problems with an assumption of normally distributed error.



The residuals show no pattern, although lot size does show some outliers.

Note: This same model can be run in PROC GLM. Additional plots can be obtained using PROC GLM and PROC PLM. You can write the code in SAS.

```
/*st103d03.sas*/ /*Part B*/
proc glm data=STAT1.ameshousing3
    plots(only)=(contourfit);
model SalePrice=Basement_Area Lot_Area;
store out=multiple;
title "Model with Basement Area and Gross Living Area";
run;
```

Selected PLOTS option:

CONTOURFIT modifies the contour fit plot produced by default when you have a model involving only two continuous predictors. The plot displays a contour plot of the predicted surface overlaid with a scatter plot of the observed data.

PROC GLM Output

Number of Observations Read	300
Number of Observations Used	300

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	203220618262	101610309131	137.17	<.0001
Error	297	220002901249	740750509.26		
Corrected Total	299	423223519511			

Notice that the ANOVA table contains the same information as in PROC REG

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.480173	19.79041	27216.73	137524.9

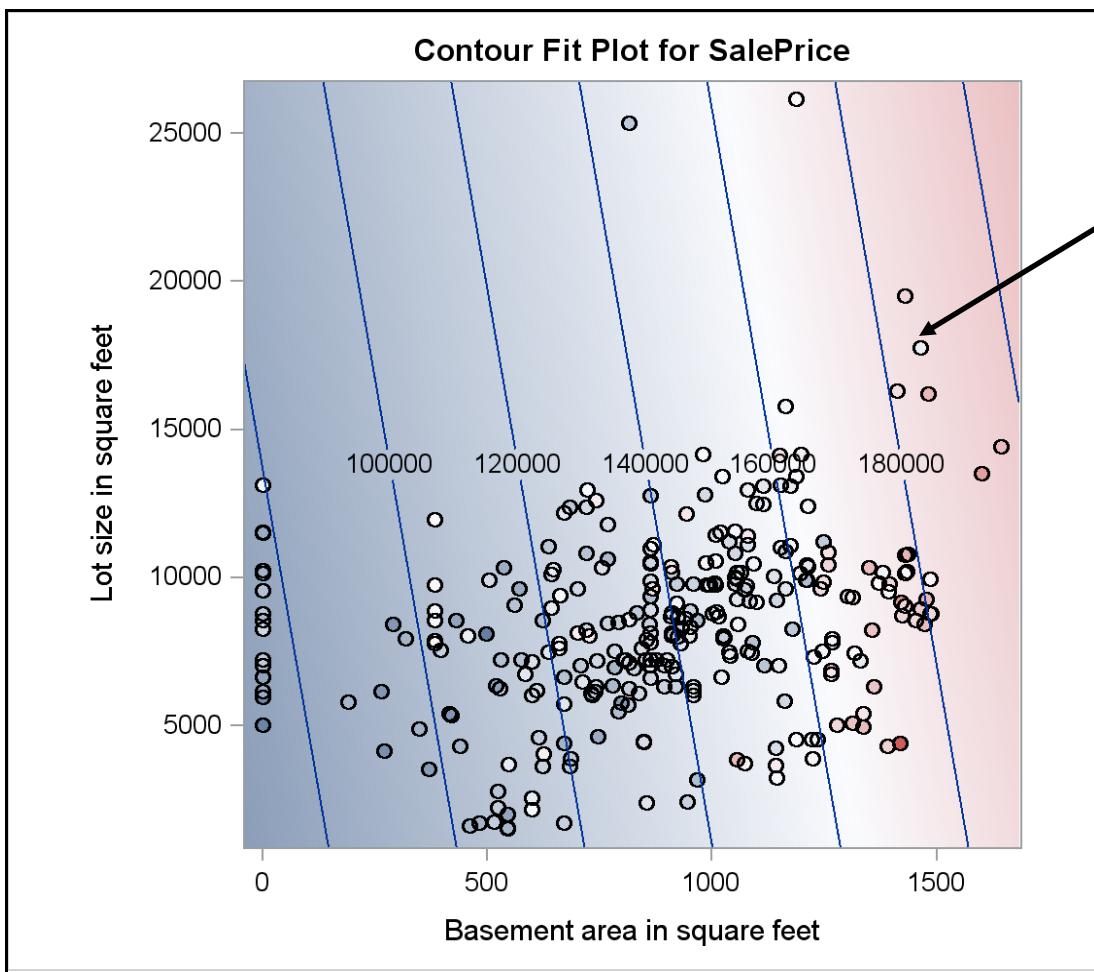
PROC GLM does not give an adjusted R-Square value.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Basement_Area	1	201241844480	201241844480	271.67	<.0001
Lot_Area	1	1978773781.7	1978773781.7	2.67	0.1032

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Basement_Area	1	176055907089	176055907089	237.67	<.0001
Lot_Area	1	1978773781.7	1978773781.7	2.67	0.1032

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	69015.61360	5129.521790	13.45	<.0001
Basement_Area	70.08680	4.546183	15.42	<.0001
Lot_Area	0.80430	0.492102	1.63	0.1032

The estimates table gives the same results (within rounding error) as the parameter estimates table in PROC REG.



The contour plot shows predicted values of **SalePrice** as gradations of color from blue (low values) to red (high values). The dots for the actual data are similarly colored. Observations that are perfectly fit would show the same color within the circle as outside the circle. The blue lines help you read the actual predictions at even intervals. For example, the circle that is being pointed at in the plot has a basement area of about 1,500 square feet, a lot size of about 17,000 square feet and a predicted value of over \$180,000 for sale price. Its color shows that its observed sale price is actually closer to about \$160,000.

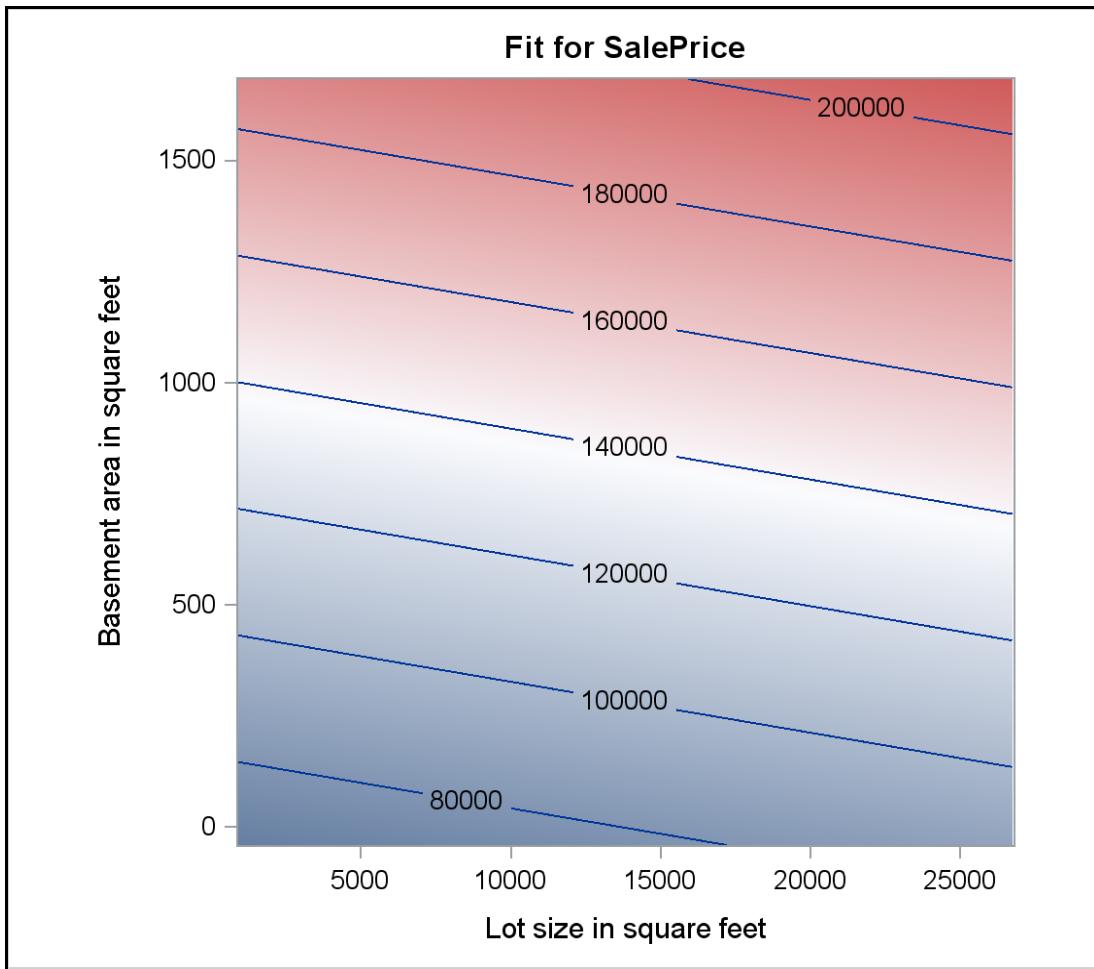
PROC PLM can use the item store for further analysis.

```
/*st103d03.sas*/ /*Part C*/
proc plm restore=multiple plots=all;
  effectplot contour (y=Basement_Area x=Lot_Area);
  effectplot slicefit(x=Lot_Area
                      sliceby=Basement_Area=250 to 1000 by 250);
run;
```

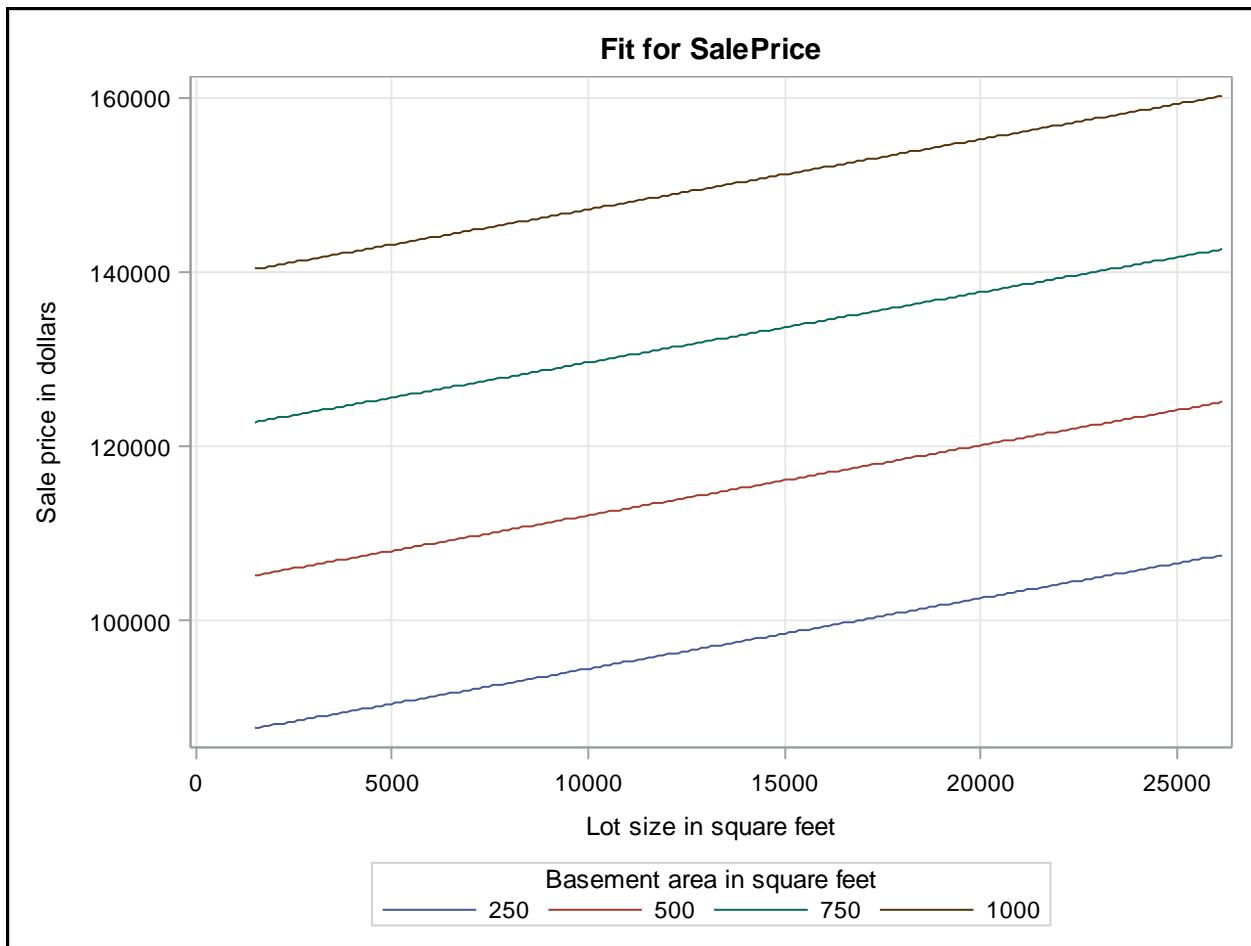
Selected EFFECTPLOT options:

- | | |
|----------|--|
| CONTOUR | Displays a contour plot of predicted values against two continuous covariates. |
| SLICEFIT | Displays a curve of predicted values versus a continuous variable grouped by the levels of another effect. |

Store Information	
Item Store	WORK.MULTIPLE
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLM
Date Created	03JUL14:11:40:15
Response Variable	SalePrice
Model Effects	Intercept Basement_Area Lot_Area



PROC PLM does not display the observed values on the contour plot.



Another way of displaying the results of a 2-predictor regression model is through the use of a slice plot. The regression lines represent the slices of **Basement_Area** that were specified in the SAS code.

End of Demonstration



Exercises

2. Performing Multiple Regression Using the REG Procedure

- a. Using the **STAT1.BodyFat2** data set, run a regression of **PctBodyFat2** on the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.
 - 1) Compare the ANOVA table with that from the model with only **Weight** in the previous exercise. What is different?
 - 2) How do the R-square and the adjusted R-square compare with these statistics for the **Weight** regression demonstration?
 - 3) Did the estimate for the intercept change? Did the estimate for the coefficient of **Weight** change?

3. Simplifying the Model

Rerun the preceding model, but eliminate the variable with the highest *p*-value. Compare the output with the preceding model

- a. Did the *p*-value for the model change notably?
- b. Did the R-square and adjusted R-square change notably?
- c. Did the parameter estimates and their *p*-values change notably?

4. More Simplifying of the Model

Again, rerun the preceding model, but eliminate the variable with the highest *p*-value. Compare the output with the preceding model.

- a. How did the output change from the previous model?
- b. Did the number of parameters with a *p*-value less than 0.05 change?

End of Exercises

3.03 Multiple Answer Poll

Which statistic(s) is/are used to test the null hypothesis that all regression slopes are zero, against the alternative hypothesis that they are not all zero?

- a. F test in the ANOVA table.
- b. F test in the Regression table.
- c. The Global t-test in the parameter estimates table.
- d. R square
- e. Adjusted R square

3.3 Solutions

Solutions to Exercises

1. Performing Two-Way ANOVA

Data were collected in an effort to determine whether different dose levels of a given drug have an effect on blood pressure for people with one of three types of heart disease. The data are in the **STAT1.Drug** data set.

The data set contains the following variables:

DrugDose dosage level of drug (1, 2, 3, 4), corresponding to (Placebo, 50 mg, 100 mg, 200 mg)

Disease heart disease category

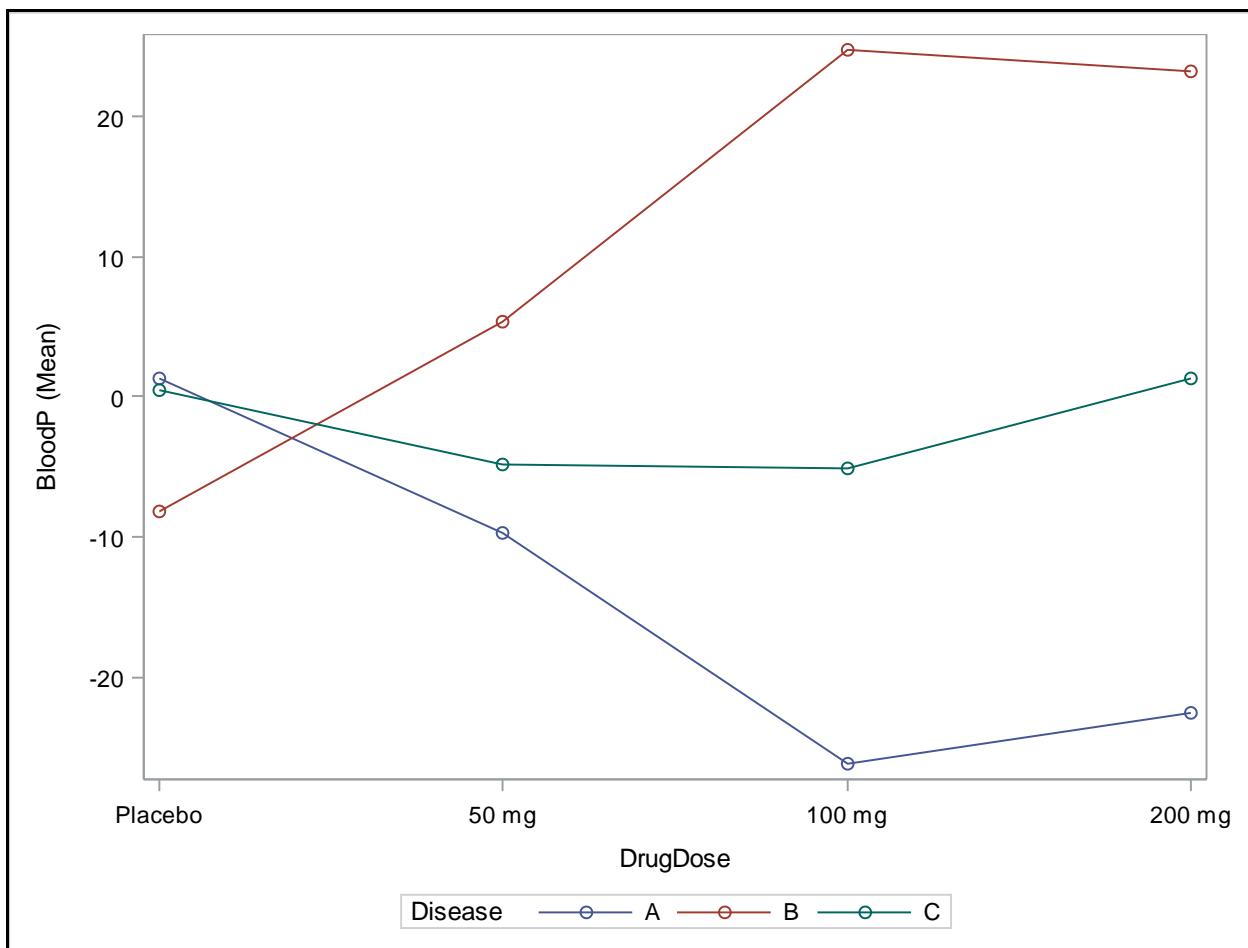
BloodP change in diastolic blood pressure after 2 weeks treatment

- a. Examine the data with a vertical line plot. Put **BloodP** on the Y axis, **DrugDose** on the X axis, and then stratify by **Disease**. What information can you obtain from looking at the data?
 - 1) Open the **Line Chart** task under Graph and select the **Drug** data set.
 - 2) Assign **DrugDose** as the Category variable, **BloodP** as the Response variable, and **Disease** as the Group variable.
 - 3) On the OPTIONS tab, expand the LINE DETAILS tab and change the line thickness to 1.
 - 4) Submit the code.

Note: Alternatively, you can enter the SAS code directly and include additional FORMAT statement.

```
/*st103s01.sas*/ /*Part A*/
proc sgplot data=STAT1.drug;
  vline DrugDose / group=Disease
    stat=mean
    response=BloodP
    markers;
  format DrugDose dosefmt.;
run;
```

PROC SGLOT Output



It appears that drug dose affects change in blood pressure. However, that effect is not consistent across diseases. Higher doses result in increased blood pressure for patients with disease B, decreased blood pressure for patients with disease A, and little change in blood pressure for patients with disease C.

- b. Test the hypothesis that the means are equal, making sure to include an interaction term if your graphical analyses indicate that would be advisable. What conclusions can you reach at this point?
- 1) Open the **N-Way ANOVA** task under **Statistics**.
 - 2) Assign **BloodP** as the dependent variable and **DrugDose** and **Disease** as the factors.
 - 3) On the **MODEL** tab, specify the model by clicking the **Edit** button and then selecting the effects, including the interaction term, and then click **OK**.
 - 4) On the **OPTIONS** tab, in the drop-down list of **Select statistics to display**, select the option to display default and additional statistics, and clear the **Perform multiple comparisons** option.
 - 5) Run the code.

PROC GLM Output

Class Level Information		
Class	Levels	Values
DrugDose	4	1 2 3 4
Disease	3	A B C

Number of Observations Read	170
Number of Observations Used	170

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	36476.8353	3316.0759	7.66	<.0001
Error	158	68366.4589	432.6991		
Corrected Total	169	104843.2941			

The global *F* test indicates a significant difference among the different groups. Because the interaction is in the model, this is a test of all combinations of DrugDose*Disease against all other combinations.

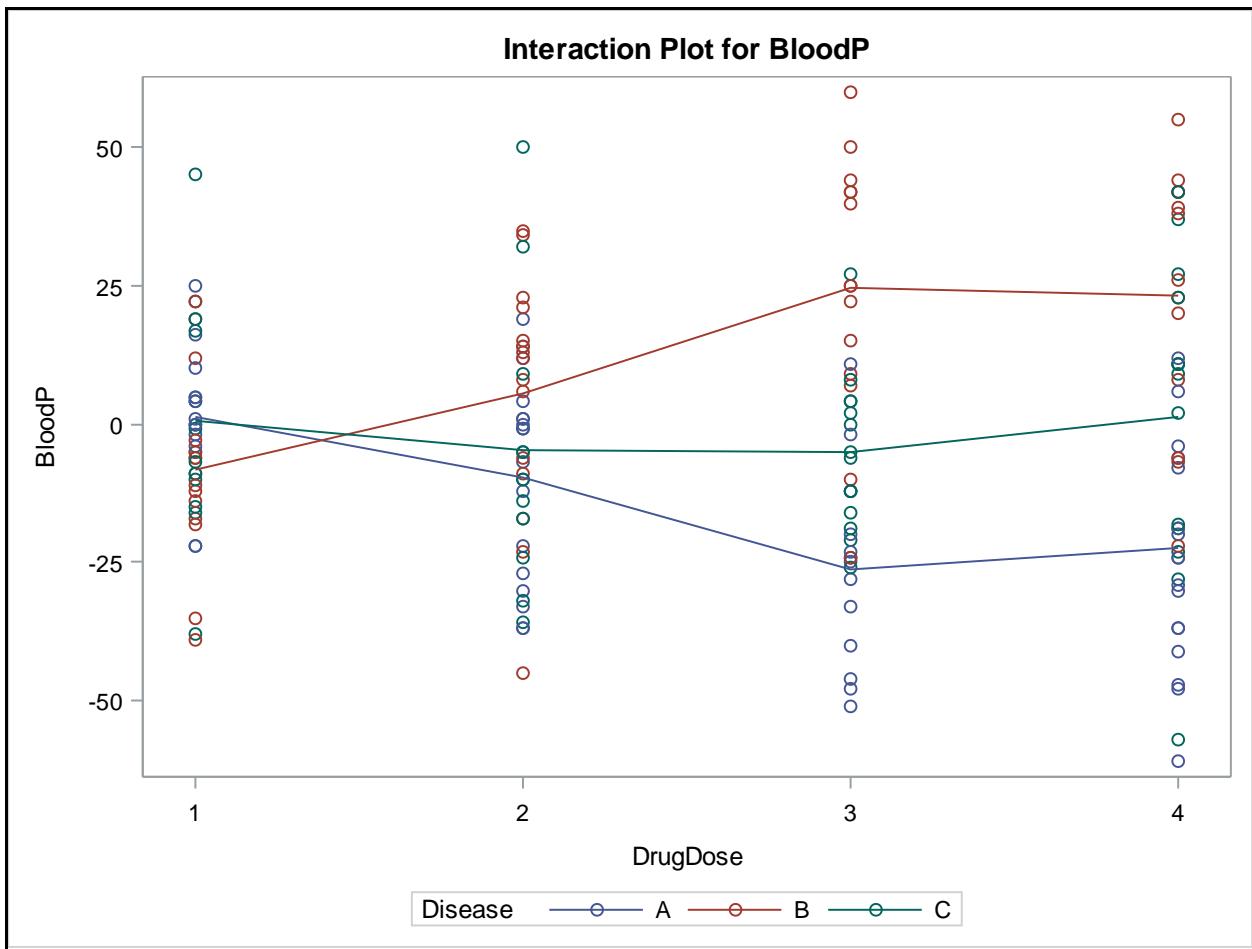
R-Square	Coeff Var	Root MSE	BloodP Mean
0.347918	-906.7286	20.80142	-2.294118

The R-Square value implies that about 35% of the variation in BloodP can be explained by variation in the explanatory variables.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DrugDose	3	54.03137	18.01046	0.04	0.9886
Disease	2	19276.48690	9638.24345	22.27	<.0001
DrugDose*Disease	6	17146.31698	2857.71950	6.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DrugDose	3	335.73526	111.91175	0.26	0.8551
Disease	2	18742.62386	9371.31193	21.66	<.0001
DrugDose*Disease	6	17146.31698	2857.71950	6.60	<.0001

The interaction term is statistically significant, as predicted by the plot of means.



- c. To investigate the interaction effect between the two factors, include the slice option by manually editing the generated code or you can write the code directly.

```
/*st103s01.sas*/ /*Part B*/
ods graphics on;

proc glm data=STAT1.drug plots(only)=intplot;
  class DrugDose Disease;
  model BloodP=DrugDose|Disease;
  lsmeans DrugDose*Disease / slice=Disease;
run;
quit;
```

DrugDose	Disease	BloodP LSMEAN
1	A	1.3333333
1	B	-8.1333333
1	C	0.4285714
2	A	-9.6875000
2	B	5.4000000
2	C	-4.8461538
3	A	-26.2307692
3	B	24.7857143
3	C	-5.1428571
4	A	-22.5555556
4	B	23.2307692
4	C	1.3076923

DrugDose*Disease Effect Sliced by Disease for BloodP					
Disease	DF	Sum of Squares	Mean Square	F Value	Pr > F
A	3	6320.126747	2106.708916	4.87	0.0029
B	3	10561	3520.222833	8.14	<.0001
C	3	468.099308	156.033103	0.36	0.7815

The sliced table shows the effects of DrugDose at each level of Disease. The effect is significant for all but disease C.

2. Performing Multiple Regression Using the REG Procedure

- a. Using the STAT1.BodyFat2 data set, run a regression of PctBodyFat2 on the variables Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, and Wrist.

Note: For this exercise, turn off ODS Graphics.

- 1) Open the **Linear Regression** task under Statistics.
- 2) On the DATA tab, select the **BodyFat2** data set and assign the variables.
- 3) On the MODEL tab, specify the model by using the model editor to add in the effects.
- 4) On the OPTIONS tab, clear the options for all the plots.
- 5) Run the code.

Note: Alternatively, you can write the code directly in SAS.

```
/*st103s02.sas*/ /*Part A*/
Ods graphics off;
proc reg data=STAT1.BodyFat2;
    model PctBodyFat2=Age Weight Height
        Neck Chest Abdomen Hip Thigh
        Knee Ankle Biceps Forearm Wrist;
    title 'Regression of PctBodyFat2 on All Predictors';
run;
quit;
```

PROC REG Output

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	13	13159	1012.22506	54.50	<.0001
Error	238	4420.06401	18.57170		
Corrected Total	251	17579			

Root MSE	4.30949	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7348
Coeff Var	22.50293		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-21.35323	22.18616	-0.96	0.3368
Age	1	0.06457	0.03219	2.01	0.0460
Weight	1	-0.09638	0.06185	-1.56	0.1205
Height	1	-0.04394	0.17870	-0.25	0.8060
Neck	1	-0.47547	0.23557	-2.02	0.0447
Chest	1	-0.01718	0.10322	-0.17	0.8679
Abdomen	1	0.95500	0.09016	10.59	<.0001
Hip	1	-0.18859	0.14479	-1.30	0.1940
Thigh	1	0.24835	0.14617	1.70	0.0906
Knee	1	0.01395	0.24775	0.06	0.9552
Ankle	1	0.17788	0.22262	0.80	0.4251
Biceps	1	0.18230	0.17250	1.06	0.2917
Forearm	1	0.45574	0.19930	2.29	0.0231
Wrist	1	-1.65450	0.53316	-3.10	0.0021

- 1) Compare the ANOVA table with that from the model with only **Weight** in the previous exercise. What is different?

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, 13 versus 1.
- The Mean Square model and the *F* ratio are much smaller.

- 2) How do the R-square and the adjusted R-square compare with these statistics for the **Weight** regression demonstration?

Both the R-square and adjusted R-square for the full models are larger than the simple linear regression. The multiple regression model explains almost 75% of the variation in the PctBodyFat2 variable versus only about 37.5% explained by the simple linear regression model.

- 3) Did the estimate for the intercept change? Did the estimate for the coefficient of **Weight** change?

Yes, including the other variables in the model changed the estimates both of the intercept and the slope for Weight. Also, the *p*-values for both changed dramatically. The slope of Weight is now not significantly different from zero.

3. Simplifying the Model

Rerun the preceding model, but eliminate the variable with the highest *p*-value. Compare the output with the preceding model.

- 1) On the **MODEL** tab, edit the model by using the trash can button to remove **Knee** from the model.

Note: Alternatively, you can simply delete the variable name from the model statement in the code.

```
/*st103s02.sas*/ /*Part B*/
proc reg data=STAT1.BodyFat2;
  model PctBodyFat2=Age Weight Height
    Neck Chest Abdomen Hip Thigh
    Ankle Biceps Forearm Wrist;
  title 'Regression of PctBodyFat2 on All '
    'Predictors, Minus Knee';
run;
quit;
```

PROC REG Output

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	12	13159	1096.57225	59.29	<.0001
Error	239	4420.12286	18.49424		
Corrected Total	251	17579			

Root MSE	4.30049	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7359
Coeff Var	22.45595		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-21.30204	22.12123	-0.96	0.3365
Age	1	0.06503	0.03108	2.09	0.0374
Weight	1	-0.09602	0.06138	-1.56	0.1191
Height	1	-0.04166	0.17369	-0.24	0.8107
Neck	1	-0.47695	0.23361	-2.04	0.0423
Chest	1	-0.01732	0.10298	-0.17	0.8666
Abdomen	1	0.95497	0.08998	10.61	<.0001
Hip	1	-0.18801	0.14413	-1.30	0.1933
Thigh	1	0.25089	0.13876	1.81	0.0719
Ankle	1	0.18018	0.21841	0.82	0.4102
Biceps	1	0.18182	0.17193	1.06	0.2913
Forearm	1	0.45667	0.19820	2.30	0.0221
Wrist	1	-1.65227	0.53057	-3.11	0.0021

- a. Did the *p*-value for the model change notably?

The *p*-value for the model did not change out to four decimal places.

- b. Did the R-square and adjusted R-square change notably?

The R-square showed essentially no change. The adjusted R-square increased from 0.7348 to 0.7359. When an adjusted R-square increases by removing a variable from the model, it implies that the removed variable was not necessary.

- c. Did the parameter estimates and their *p*-values change notably?

Some of the parameter estimates and their *p*-values changed slightly, none to any large degree.

4. More Simplifying of the Model

- 1) Again, rerun the preceding task, but drop the variable with the highest p -value, **Chest**.

This program reruns the regression with **Chest** removed, because it is the variable with the highest p -value in the previous model.

```
/*st103s02.sas*/ /*Part C*/
proc reg data=STAT1.BodyFat2;
  model PctBodyFat2=Age Weight Height
    Neck Abdomen Hip Thigh
    Ankle Biceps Forearm Wrist;
  title 'Regression of PctBodyFat2 on All '
    'Predictors, Minus Knee, Chest';
run;
quit;
```

PROC REG Output

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	11	13158	1196.21310	64.94	<.0001
Error	240	4420.64572	18.41936		
Corrected Total	251	17579			

Root MSE	4.29178	R-Square	0.7485
Dependent Mean	19.15079	Adj R-Sq	0.7370
Coeff Var	22.41044		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-23.13736	19.20171	-1.20	0.2294
Age	1	0.06488	0.03100	2.09	0.0374
Weight	1	-0.10095	0.05380	-1.88	0.0618
Height	1	-0.03120	0.16185	-0.19	0.8473
Neck	1	-0.47631	0.23311	-2.04	0.0421
Abdomen	1	0.94965	0.08406	11.30	<.0001
Hip	1	-0.18316	0.14092	-1.30	0.1950
Thigh	1	0.25583	0.13534	1.89	0.0599
Ankle	1	0.18215	0.21765	0.84	0.4035

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Biceps	1	0.18055	0.17141	1.05	0.2933
Forearm	1	0.45262	0.19634	2.31	0.0220
Wrist	1	-1.64984	0.52930	-3.12	0.0020

- a. How did the output change from the previous model?

The ANOVA table did not change greatly. The R-square remained essentially unchanged. The adjusted R-square increased again, which confirms that the variable Chest did not contribute to explaining the variation in PctBodyFat2 when the other variables are in the model.

- b. Did the number of parameters with a *p*-value less than 0.05 change?

The *p*-value for Weight changed more than any other and is now just above 0.05. The *p*-values and parameter estimates for other variables changed much less. There are no more variables in this model with *p*-values below 0.05, compared with the previous one.

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

3.01 Multiple Answer Poll – Correct Answers

How can you recognize an interaction?

- a. The effect of one variable differs at different levels of another.
- b. The joint effect of two variables is greater than the sum of their individual effects.
- c. A stratified plot of the effect of one variable against the dependent variable shows different patterns across different levels of the other.
- d. Two variables are talking at a party.
- e. One variable shows statistical significance without the other in the model, but no significance with the other variable in the model.

15

Copyright © SAS Institute Inc. All rights reserved.



3.03 Multiple Answer Poll – Correct Answer

Which statistic(s) is/are used to test the null hypothesis that all regression slopes are zero, against the alternative hypothesis that they are not all zero?

- a. F test in the ANOVA table.
- b. F test in the Regression table.
- c. The Global t-test in the parameter estimates table.
- d. R square
- e. Adjusted R square

37

Copyright © SAS Institute Inc. All rights reserved.



Chapter 4 Model Building and Effect Selection

4.1 Stepwise Selection Using Significance Level.....	4-3
Demonstration: Stepwise Regression	4-10
Exercises.....	4-20
4.2 Information Criterion and Other Selection Options.....	4-21
Demonstration: Model Selection Using AIC, AICC, BIC, and SBC	4-24
Exercises.....	4-37
4.3 All Possible Selection (Self-Study).....	4-38
Demonstration: All Possible Model Selection	4-42
Exercises.....	4-49
4.4 Solutions	4-50
Solutions to Exercises	4-50
Solutions to Student Activities (Polls/Quizzes)	4-64

4.1 Stepwise Selection Using Significance Level

Objectives

- Describe forward, backward, and stepwise model selection methodology.
- Explain the GLMSELECT procedure options for model selection using significance level.

3

Copyright © SAS Institute Inc. All rights reserved.



Model Selection

Eliminating one variable at a time manually for small data sets is a reasonable approach.

However, eliminating one variable at a time manually for large data sets can take an extreme amount of time.

4

Copyright © SAS Institute Inc. All rights reserved.



A process for selecting models might be to start with all the variables in the **STAT1.ameshousing3** 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 0.05 or 0.10, remain.

PROC GLMSELECT

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

Options within PROC GLMSELECT can assist with model selection.

Model Selection Options

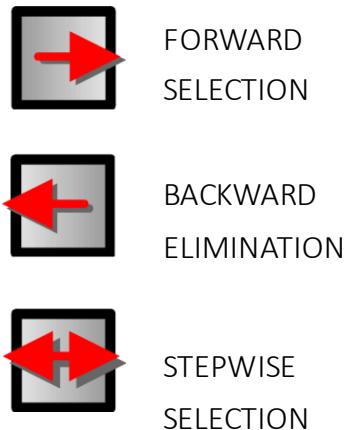
Options in the MODEL statement of PROC GLMSELECT support many model selection techniques and criteria.

- SELECTION=<option>
- CHOOSE=<option>
- SELECT=<option>
- STOP=<option>

PROC GLMSELECT offers many choices for selection techniques and criteria through the usage of different options.

- SELECTION** specifies the method used to select the model. Possible methods to choose from include NONE (specifies no model selection), FORWARD, BACKWARD, STEPWISE, LAR, LASSO, and ELASTICNET. The default is STEPWISE.
- CHOOSE** specifies the criterion for choosing the model. The specified criterion is evaluated at each step of the selection process, and the model that yields the best value of the criterion is chosen. If CHOOSE= is omitted, the model at the final step in the selection process is selected.
- SELECT** specifies the criterion that determines the order in which effects enter or leave at each step of the specified selection method. The default value is SELECT=SBC. The effect that is selected to enter or leave at a step of the selection process is the effect whose addition to or removal from the current model gives the maximum improvement in the specified criterion.
- STOP** specifies when to stop the selection process. If you do not specify the STOP= option but do specify the SELECT= option, this criterion will also be used as the STOP= criterion. Default is STOP=SBC when neither STOP= nor SELECT= is specified. If you specify STOP=n, then selection will stop at the first step for which the selected model has n effects.

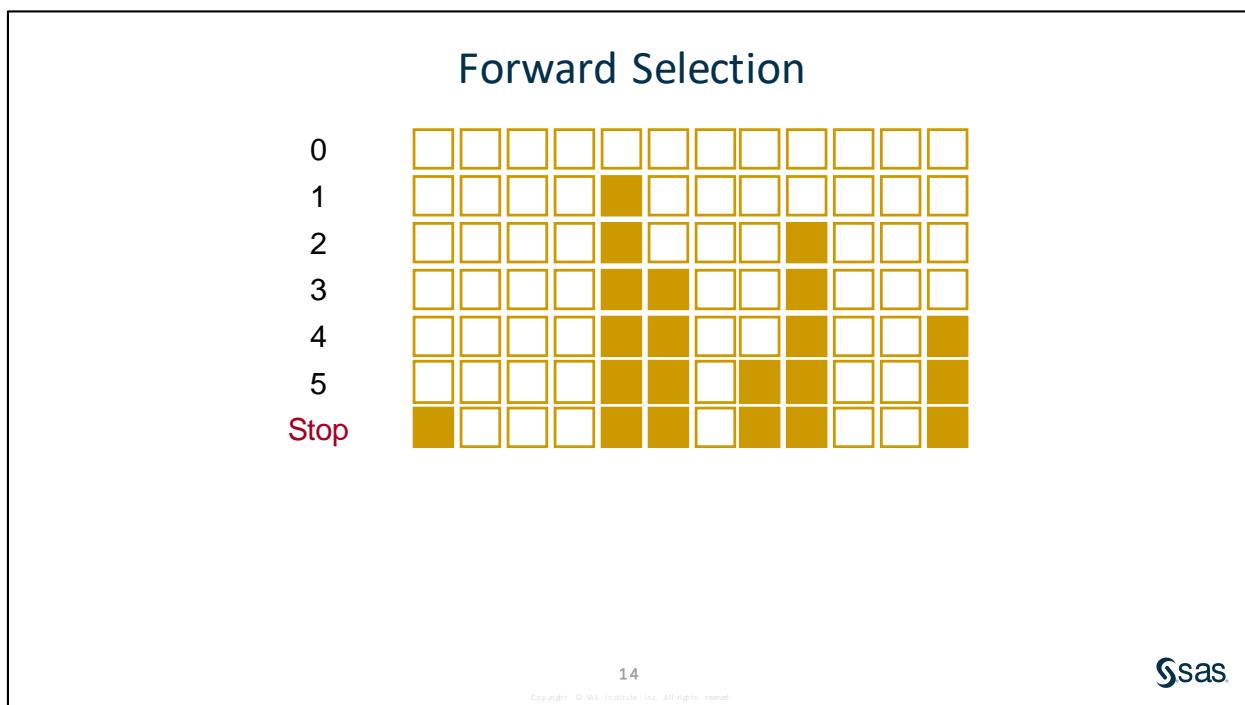
Stepwise Selection Methods



PROC GLMSELECT also offers the following stepwise SELECTION= options:

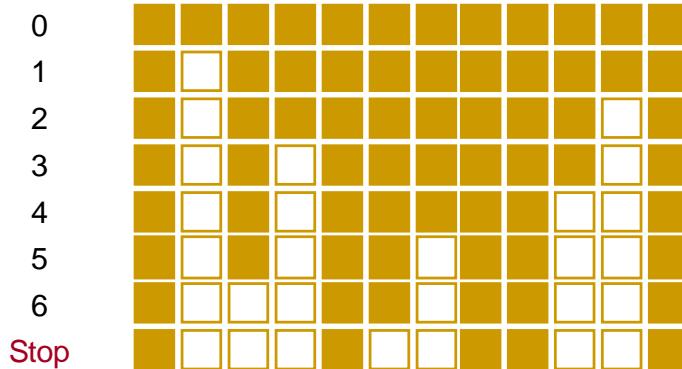
- FORWARD** first selects the best one-variable model. Then it selects the best two variables among those that contain the first selected variable. FORWARD continues this process, but stops when it reaches the point where no additional variables have p -value levels less than some stopping criterion (0.50, by default).
- 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 p -values less than a stopping criterion value (0.10, by default).
- STEPWISE** works like a combination of the FORWARD and BACKWARD method. The default entry p -value is 0.15 and the default stay p -value is also 0.15.

Note: The SLENTRY= (for forward step stopping criterion) and SLSTAY= (for backward step stopping criterion) options can be used to change the default stopping values.



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 meets the specified level for entry. By default, SLENTRY=0.50.

Backward Elimination



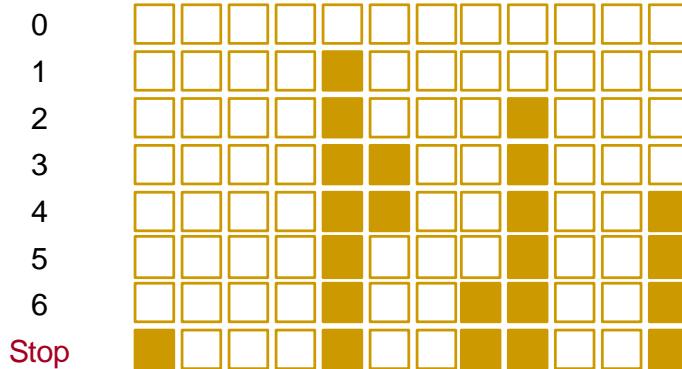
22



Copyright © SAS Institute Inc. All rights reserved.

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



30



Copyright © SAS Institute Inc. All rights reserved.

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 criterion specified in the SLSTAY= option. The stepwise selection process terminates if no further variables can be added to the model or if the variable entered into the model is the only variable removed in the subsequent backward elimination. By default, SLENTRY=0.15 and SLSTAY=0.15.

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).

Are *p*-values and Parameter Estimates Correct?

Automated model selection results in the following:

- biases in parameter estimates, predictions, and standard errors
- incorrect calculation of degrees of freedom
- *p*-values that tend to err on the side of overestimating significance
(increasing Type I Error probability)

Statisticians give warnings and cautions about the appropriate interpretation of p -values from models chosen using any automated variable selection technique. Refitting many submodels in terms of an optimum fit to the data distorts the significance levels of conventional statistical tests. However, many researchers and users of statistical software neglect to report that the models that they ended up with were chosen using automated methods. They report statistical quantities such as standard errors, confidence limits, p -values, and R-square as if the resulting model were entirely prespecified. These inferences are inaccurate, tending to err on the side of overstating the significance of predictors and making predictions with overly optimistic confidence. This problem is very evident when there are many iterative stages in model building. When there are many variables and you use stepwise selection to find a small subset of variables, inferences become less accurate (Chatfield 1995, Raftery 1994, Freedman 1983).

One solution to this problem is to split your data. One part can be used for finding the regression model and the other part can be used for inference. Another solution is to use bootstrapping methods to obtain the correct standard errors and p -values. *Bootstrapping* is a resampling method that tries to approximate the distribution of the parameter estimates to estimate the standard error.



Stepwise Regression

Example: Select a model for predicting **SalePrice** in the **STAT1.ameshousing3** data set by using the STEPWISE selection method. Use 0.05 as the significance level for entry into and staying in the model.

1. Open the **Linear Regression** task under **Statistics**.
2. On the DATA tab, select the **AmesHousing3** data set.
3. Assign **SalePrice** as the dependent variable and the interval variables as the continuous variables.
4. On the MODEL tab, specify the appropriate model by using the model effect editor.
5. On the OPTIONS tab, suppress all diagnostic plots, scatter plots, and residual plots.
6. On the SELECT tab, use the drop-down menu and choose the **Stepwise selection** option.
7. Specify using **Significance level** as the criterion to add/remove effects.
8. Under the DETAILS property, modify to show details for each step.
9. To obtain detailed graphical output, open the **editor** for the generated code and modify the plotting option to **plots=all**.
10. Run the code.

The screenshot shows the SAS GLMSELECT procedure interface. The top navigation bar includes MODEL, OPTIONS, SELECTION (which is selected), OUTPUT, and INFORM tabs. The SELECTION tab is expanded, showing the following configuration:

- Selection method:** Stepwise selection
- Add/remove effects with:** Significance level
- Stop adding/removing effects with:** Significance level
- Select best model by:** Default criterion
- *Significance level to add an effect to the model:** 0.05
- *Significance level to remove an effect from the model:** 0.05

Below the SELECTION tab, there are collapsed sections for SELECTION STATISTICS and SELECTION PLOTS. The DETAILS section is expanded, showing:

- Selection process details:** Details for each step
- Additional information for each step:**
 - ANOVA table
 - Fit statistics
 - Parameter estimates

Note: Alternatively, you can write the code directly in SAS.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
      Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d01.sas*/
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
  STEPWISE: model SalePrice=&interval / selection=stepwise
             details=steps select=SL slstay=0.05 slentry=0.05;
  title "Stepwise Model Selection for SalePrice - SL 0.05";
run;
```

Partial PROC GLMSELECT Output

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	299	4.232235E11	1415463276	
Corrected Total	299	4.232235E11		

Root MSE	37623
Dependent Mean	137525
R-Square	0.0000
Adj R-Sq	0.0000
AIC	6624.21515
AICC	6624.25555
SBC	6325.91893

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	137525	2172.144314	63.31

Recall that STEPWISE selection begins like FORWARD selection with just the intercept. Then, subject to the criterion specified, an effect enters the model, if possible.

Effect Entered: Basement_Area

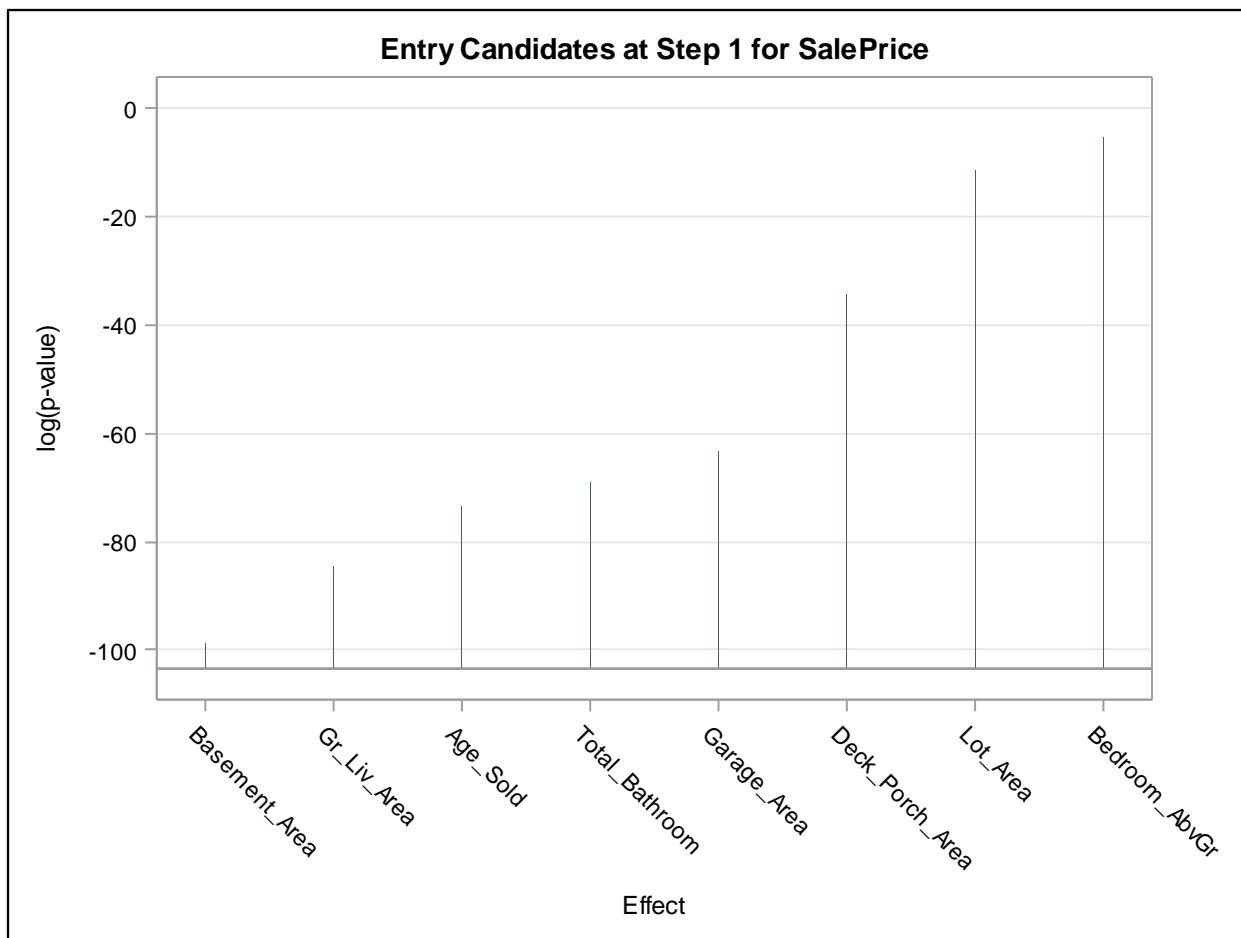
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	27293
Dependent Mean	137525
R-Square	0.4755
Adj R-Sq	0.4737
AIC	6432.62346
AICC	6432.70454
SBC	6138.03102

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

Note: The *p*-values can be displayed on the Parameter Estimates table by including the SHOWPVALUES option in the MODEL statement.

Entry Candidates				
Rank	Effect	Log pValue	Pr > F	
1	Basement_Area	-98.8577	<.0001	
2	Gr_Liv_Area	-84.6132	<.0001	
3	Age_Sold	-73.5219	<.0001	
4	Total_Bathroom	-69.1880	<.0001	
5	Garage_Area	-63.3558	<.0001	
6	Deck_Porch_Area	-34.3105	<.0001	
7	Lot_Area	-11.6303	<.0001	
8	Bedroom_AbvGr	-5.5339	0.0040	



During each step of the selection process, SAS displays both a table and graph of the entry candidates for that individual step. In step one, there are several entry candidates whose significance level is displayed as $<.0001$. To distinguish between these candidates, SAS also displays the log of the p -value for each effect. From both the table and the graph, you see that **Basement_Area** will be first to enter the model.

Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	Basement_Area		2	270.16	<.0001
2	Gr_Liv_Area		3	118.32	<.0001
3	Age_Sold		4	162.37	<.0001
4	Garage_Area		5	42.30	<.0001
5	Deck_Porch_Area		6	19.99	<.0001
6	Bedroom_AbvGr		7	6.41	0.0119
7	Lot_Area		8	4.97	0.0265

Upon completion of the selection process, SAS generates a summary table detailing the steps taken in the development of the model. The F values and p -values shown in this summary table are not the F and p -values for the selected model. These are statistics from each individual step. Final p -values and parameter estimates can be found in the table preceding this summary or at the conclusion of the output.

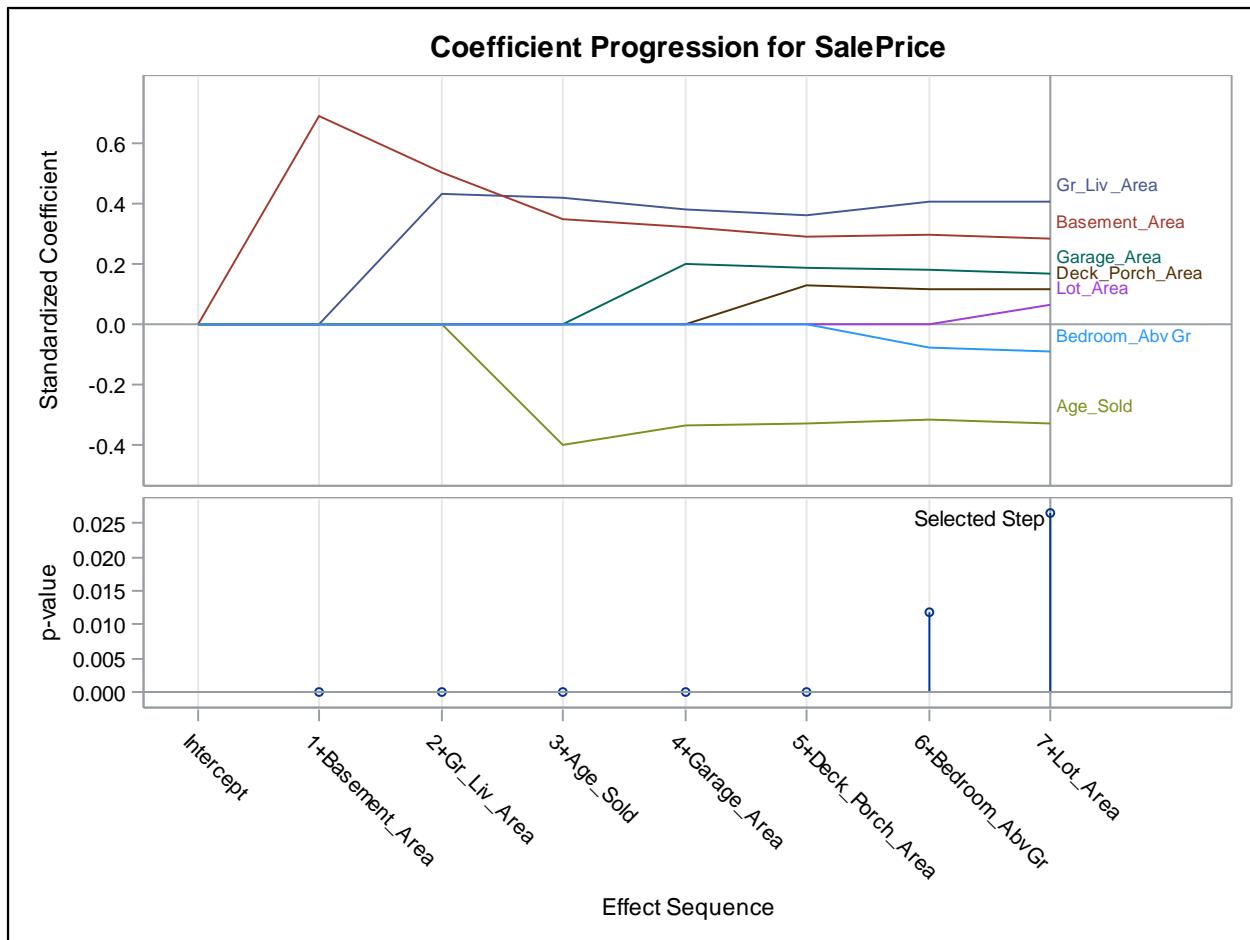
Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
-----------------	---

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

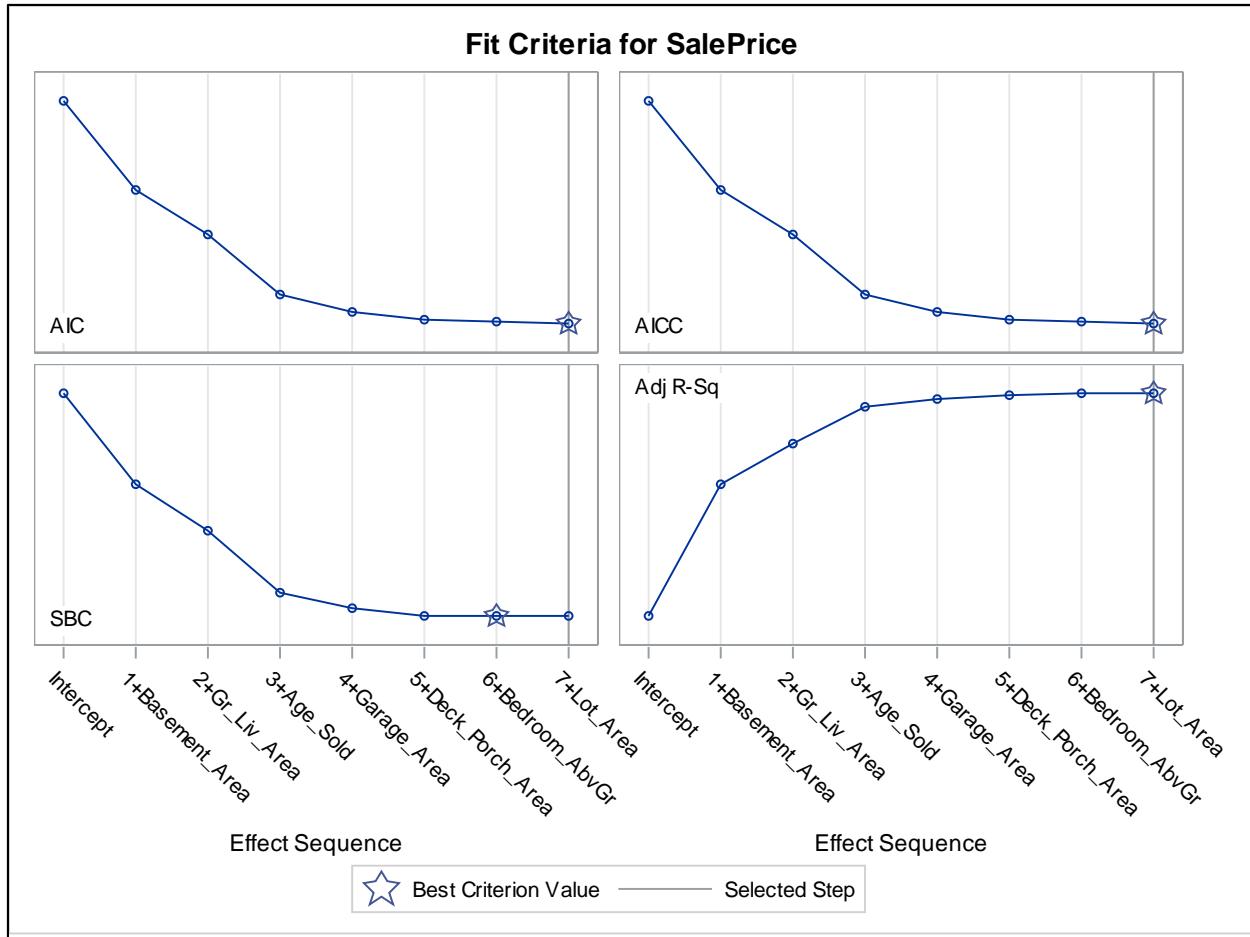
Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

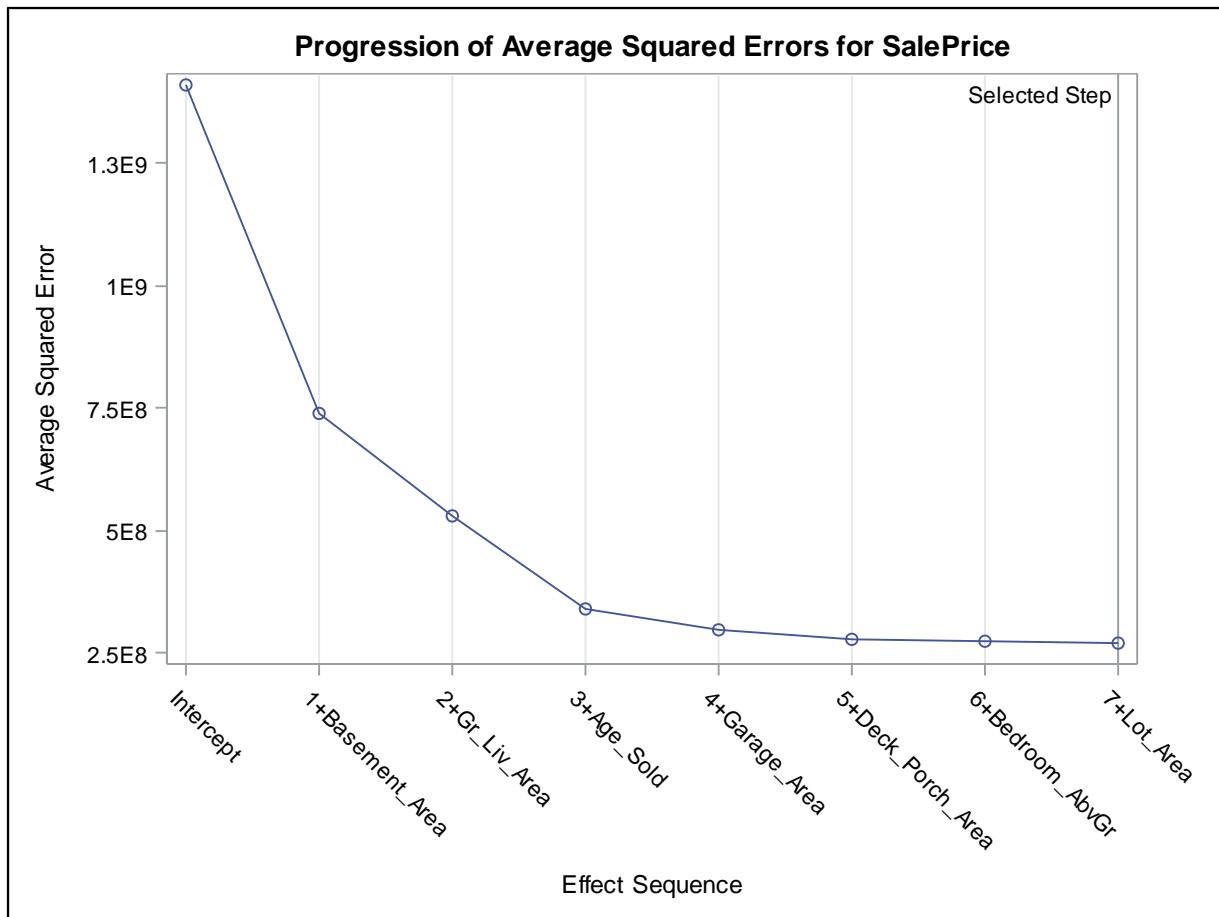
Note: Adding the SHOWPVALUES option to the MODEL statement will add p -values to the output in the Parameter Estimates table.



In the Coefficient Panel, PROC GLMSELECT displays a panel two plots showing how the standardized coefficients and the criterion used to choose the final model evolved as the selection progressed. In this image you can monitor the change in the standardized coefficients as each effect is added to or deleted from the model.



The Criteria Panel displays the progression of the adjusted R-square, AIC, AICC, SBC, as well as any other criteria that are named in the CHOOSE=, SELECT=, STOP=, or STATS= option. The star denotes the best model of the eight that were tested, in this example.



The Average Square Error Plot shows the progression of the average square error (ASE) evaluated on the training data. As more effects are added to the model, the ASE decreases for the training data. When a test or validation data set are provided, this plot will also contain information about the ASE in those data sets. This plot is best used with a hold-out data set to detect over fitting.

- To perform a different selection method, use the drop-down menu to choose the method and rerun the task.

Note: Additional code has been included that performs FORWARD and BACKWARD selection with the selection criterion set as significance level. In both cases, the SLENTRY and SLSTAY criteria have been changed to 0.05.

```
proc glmselect data=STAT1.ameshousing3 plots=all;
  FORWARD: model SalePrice=&interval / selection=forward
  details=steps select=SL slentry=0.05;
  title "Forward Model Selection for SalePrice - SL 0.05";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
  BACKWARD: model SalePrice=&interval / selection=backward
  details=steps select=SL slstay=0.05;
  title "Backward Model Selection for SalePrice - SL 0.05";
run;
```

End of Demonstration

Stepwise Models

```

FORWARD          Basement_Area,
(slentry=0.05)   Gr_Liv_Area, Age_Sold,
                  &
                  Garage_Area,
STEPWISE         Deck_Porch_Area,
(slentry=0.05,   Bedroom_AbvGr,
slstay=0.05)    Lot_Area
                  &
BACKWARD        (slstay=0.05)

```

33

Copyright © SAS Institute Inc. All rights reserved.



The final models obtained using the SLENTRY=0.05 and SLSTAY=0.05 criteria are displayed for FORWARD, BACKWARD, and STEPWISE. In this instance, all the selected models matched. It is important to note that the choice of SLENTRY and SLSTAY levels can greatly affect the final models that are selected using stepwise methods. Some analysts use larger boundaries to get models to a manageable size then do manual reduction instead of using low values for SLENTRY and SLSTAY.



Exercises

1. Using Significance Level Model Selection Techniques

Use the **STAT1.BodyFat2** data set to identify a set of “best” models.

- a. With the SELECTION=STEPWISE option, use SELECT=SL to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**. Use the default values for SLENTRY and SLSTAY.
- b. Try FORWARD.
- c. How many variables would result from a model using FORWARD selection and a significance level for entry criterion of 0.05, instead of the default SLENTRY of 0.50?

End of Exercises

4.01 Poll

The STEPWISE, BACKWARD, and FORWARD strategies result in the same final model if the same significance levels are used in all three.

- True
- False

36

Copyright © SAS Institute Inc. All rights reserved.

4.2 Information Criterion and Other Selection Options

Objectives

- Describe different criteria available within PROC GLMSELECT to perform model selection.
- Compare models provided from PROC GLMSELECT using different selection criteria.

39

Copyright © SAS Institute Inc. All rights reserved.

Information Criteria

- Akaike's information criterion (AIC)
- Corrected Akaike's information criterion (AICC)
- Sawa Bayesian information criterion (BIC)
- Schwarz Bayesian information criterion (SBC)

Smaller is better.

40

Copyright © SAS Institute Inc. All rights reserved.



Beyond significance level, there are several statistics, referred to as information criteria, that can be used both to evaluate competing models as well as direct the selection process within PROC GLMSELECT. These criteria each search for a model that will minimize the unexplained variability using as few effects within the model as possible (most parsimonious model).

Each information criterion begins $n \log \left(\frac{SSE}{n} \right)$. It then invokes a penalty representing the complexity of the model. A table of these penalties is shown below where n is the number of observations, p is the number of parameters including the intercept, and $\hat{\sigma}^2$ is the estimate of pure error variance from fitting the full model. For each of these information criteria, smaller is better.

Information Criteria	Penalty Component
AIC	$2p + n + 2$
AICC	$\frac{n(n + p)}{n - p - 2}$
BIC	$2(p + 2)q - 2q^2$
SBC	$p \log(n)$

In the BIC penalty, $q = \frac{n\hat{\sigma}^2}{SSE}$.

Adjusted R Square / Mallows' Cp

- Adjusted R Square allows proper comparison between models with different parameter counts.

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

- Mallows' Cp is a simple indicator of effective variable selection within a model.

41

Copyright © SAS Institute Inc. All rights reserved.



Other choices of selection criteria within PROC GLMSELECT include adjusted R-square and Mallows' Cp.

The R-square always increases or stays the same as you include more terms in the model. Therefore, choosing the “best” model is not as simple as just making the R-square as large as possible.

The adjusted R-square is a measure similar to R-square, but it takes into account the number of terms in the model. It can be thought of as a penalized version of R-square with the penalty increasing with each parameter added to the model. In the equation, i=1 if there is an intercept and 0 otherwise. The number of observations used to fit the model is n and the number of parameters in the model is p.

Note: More discussion about Mallow's Cp can be found in the self-study section of this chapter.



Model Selection Using AIC, AICC, BIC, and SBC

Example: Invoke PROC GLMSELECT four times on the **SalePrice** variable, regressing on the interval variables (you can use the macro %interval) in **STAT1.ameshousing3**. For each run, request STEPWISE selection with the SELECTION= option and include DETAILS=STEPS to obtain step by step information and a selection summary table. Once in PROC GLMSELECT use SELECT=AIC, SELECT=BIC, SELECT=AICC, and SELECT=SBC and compare the selected models from the output.

1. Open the **Linear Regression** task.
2. Select the **AmesHousing3** data set and assign the appropriate variables (**Gr_Liv_Area** **Bedroom_AbvGr** **Total_Bathroom** **Deck_Porch_Area** **Age_Sold** **Lot_Area** **Basement_Area** **Garage_Area**).
3. Specify the model by adding in the model effects.
4. On the OPTIONS task, expand PLOTS and suppress all plots.
5. On the SELECTION tab, select **Stepwise selection** as the selection method and then expand SELECTION PLOTS and check the option to display **Coefficient plots** in addition to the default-checked box for Criteria plots.
6. For AIC, select **Akaike's information criterion** as the criterion to add/remove effects.
7. Under the DETAILS property, select Details for each step.
8. Run the code.
9. Rerun the task and modify the information criterion. Choose Sawa Bayesian information criterion for **BIC**, Akaike's information criterion corrected for small-sample for **AICC**, and Schwarz Bayesian information criterion for **SBC**.
10. Alternatively, you can open the editor and modify the code manually.

Note: The code below produces the necessary outputs.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
      Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d02.sas*/
ods graphics on;

proc glmselect data=STAT1.ameshousing3 plots=all;
  STEPWISEAIC: model SalePrice=&interval / selection=stepwise
               details=steps select=AIC;
  title "Stepwise Model Selection for SalePrice - AIC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
  STEPWISEBIC: model SalePrice=&interval / selection=stepwise
               details=steps select=BIC;
  title "Stepwise Model Selection for SalePrice - BIC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
```

```

STEPWISEAICC: model SalePrice=&interval / selection=stepwise
               details=steps select=AICC;
   title "Stepwise Model Selection for SalePrice - AICC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
  STEPWISESBC: model SalePrice=&interval / selection=stepwise
               details=steps select=SBC;
   title "Stepwise Model Selection for SalePrice - SBC";
run;

```

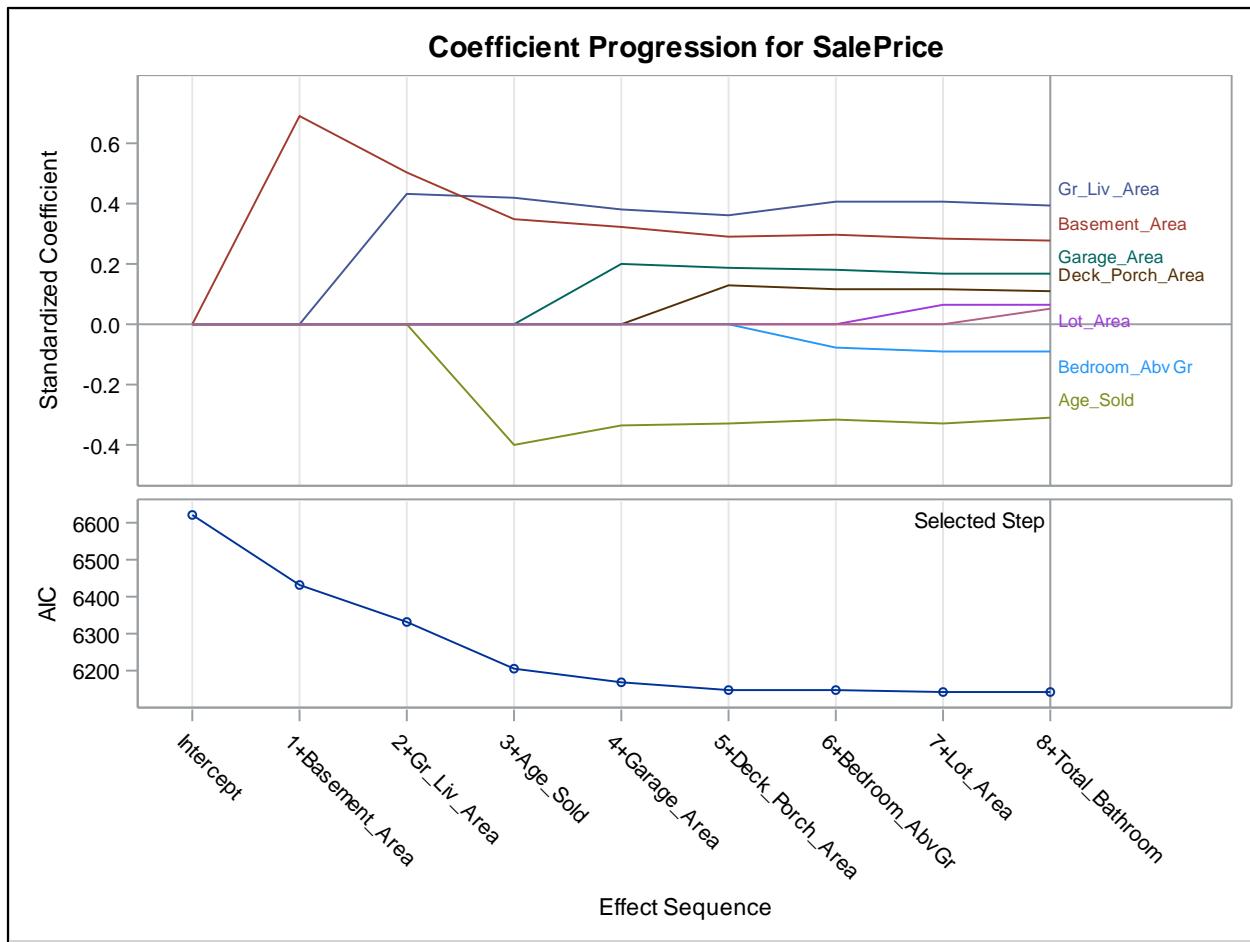
Partial PROC GLMSELECT (SELECT=AIC) Output

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	AIC
0	Intercept		1	6624.2151
1	Basement_Area		2	6432.6235
2	Gr_Liv_Area		3	6334.0262
3	Age_Sold		4	6204.8293
4	Garage_Area		5	6166.6273
5	Deck_Porch_Area		6	6148.8927
6	Bedroom_AbvGr		7	6144.4040
7	Lot_Area		8	6141.3368
8	Total_Bathroom		9	6140.7956*

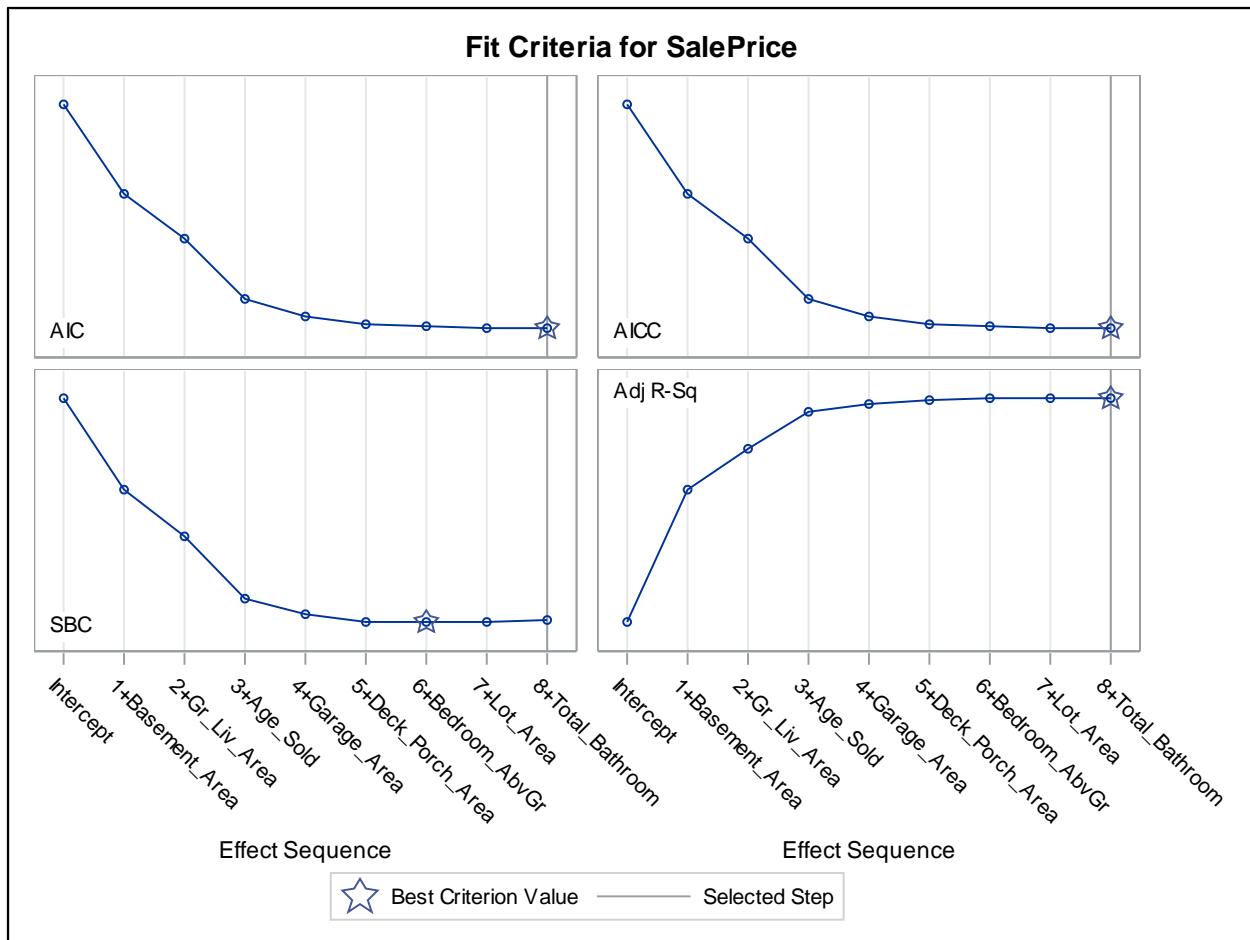
* Optimal Value Of Criterion

Selection stopped because all effects are in the final model.

Using STEPWISE selection with the SELECT criterion of AIC, all eight effects are allowed into the model before the process stops. The table above displays the order of entry for the effects.



The AIC component of the Coefficient Panel shows larger improvements to the AIC across steps one through three and moderate improvements to the AIC across steps four and five. After step five the AIC only minimally improves compared to the rest. It is also at step five that the standardized coefficients have stabilized and do not appear to vary as new effects are added to the model. One could entertain stopping after five effects and include that as an additional model that could be validated using a hold-out data set.



The Criteria Panel displays several of the other fit statistics for the model at each step. The AIC and AICC are minimized and the adjusted R-square is maximized at step eight. The SBC is shown to minimize at step six.

Effects: Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	3.431321E11	42891512314	155.84
Error	291	80091420996	275228251	
Corrected Total	299	4.232235E11		

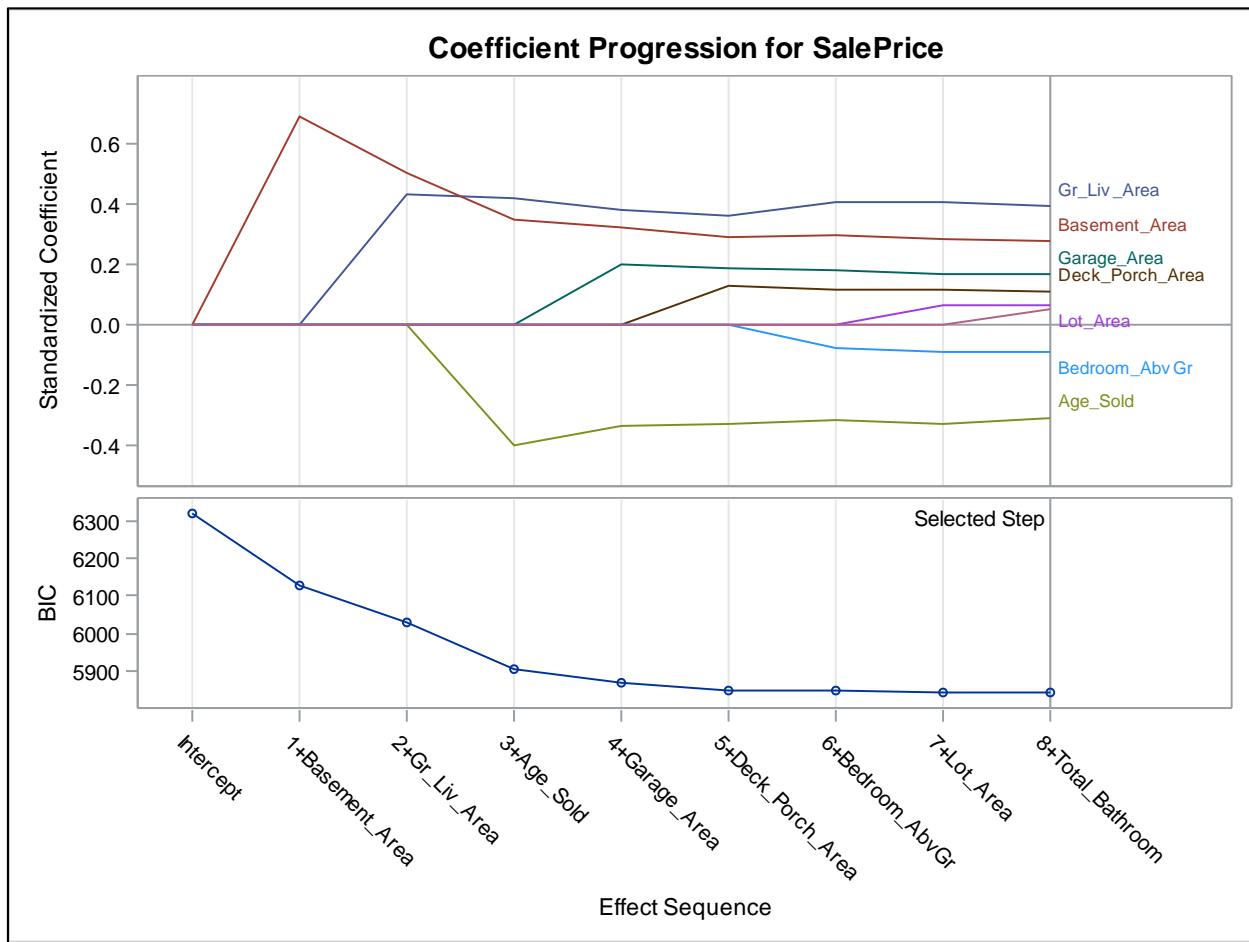
Root MSE	16590
Dependent Mean	137525
R-Square	0.8108
Adj R-Sq	0.8056
AIC	6140.79563
AICC	6141.55688
SBC	5872.12967

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	44347	6191.271944	7.16
Gr_Liv_Area	1	63.197764	5.585739	11.31
Basement_Area	1	28.692184	3.417034	8.40
Garage_Area	1	35.754191	6.445840	5.55
Deck_Porch_Area	1	31.370539	7.959436	3.94
Lot_Area	1	0.699495	0.316761	2.21
Age_Sold	1	-420.815037	44.219144	-9.52
Bedroom_AbvGr	1	-4834.848748	1688.858227	-2.86
Total_Bathroom	1	3022.124723	1920.839066	1.57

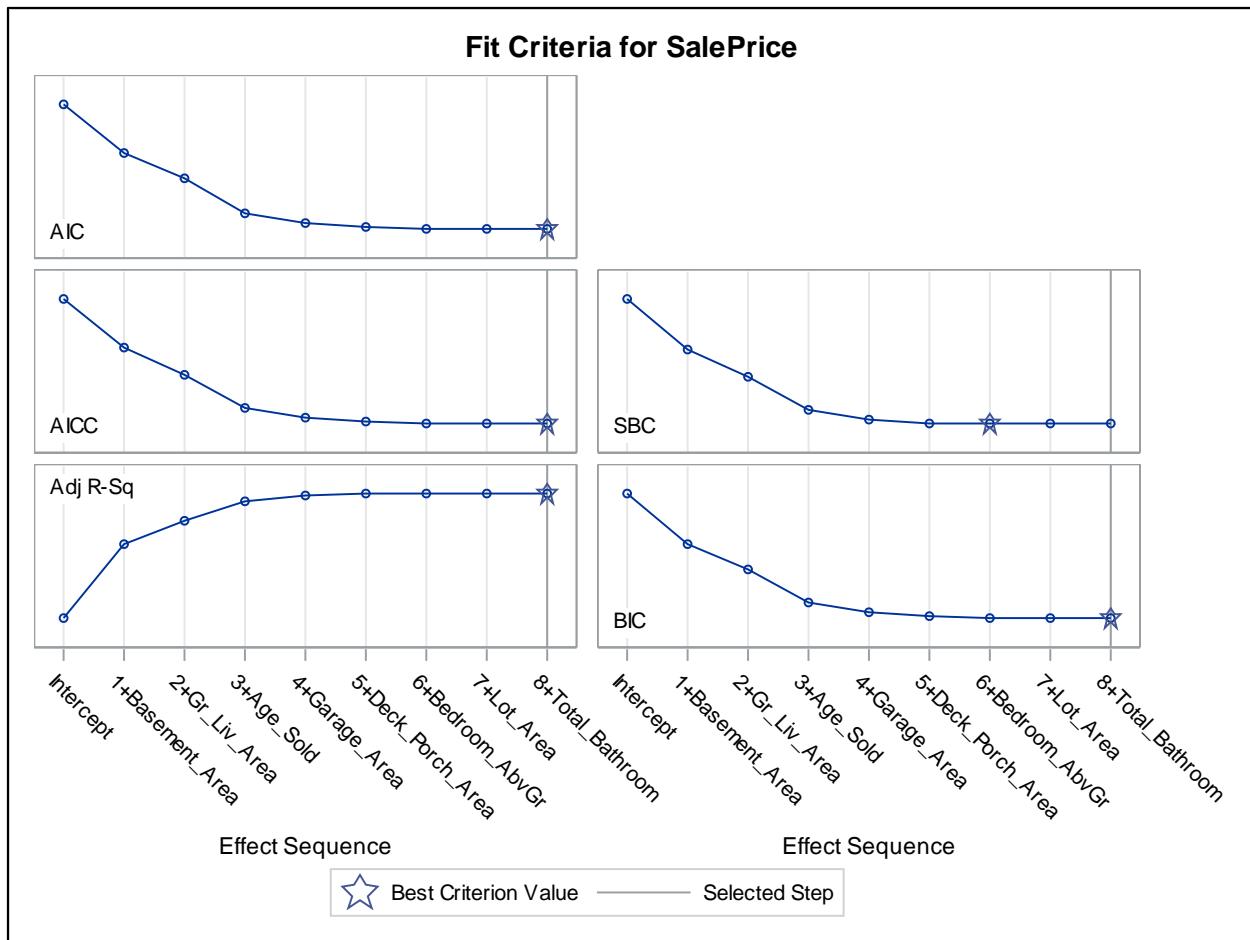
Partial PROC GLMSELECT (SELECT=BIC) Output

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	BIC
0	Intercept		1	6321.3096
1	Basement_Area		2	6129.3224
2	Gr_Liv_Area		3	6030.6866
3	Age_Sold		4	5903.1974
4	Garage_Area		5	5865.8247
5	Deck_Porch_Area		6	5848.6954
6	Bedroom_AbvGr		7	5844.4755
7	Lot_Area		8	5841.6915
8	Total_Bathroom		9	5841.3504*
* Optimal Value Of Criterion				

Similar to the AIC output, the selection process did not complete before all effects were added to the model.



The Coefficient Panel also mimics the patterns that you noticed from the AIC selection setup.



Recall that the Criteria Panel, by default, contains AIC, AICC, SBC, and adjusted R-squared. With the inclusion of SELECT=BIC, the BIC plot is included in the output. Again, you see similarities between this panel and the one generated by the AIC selection.

Effects: Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

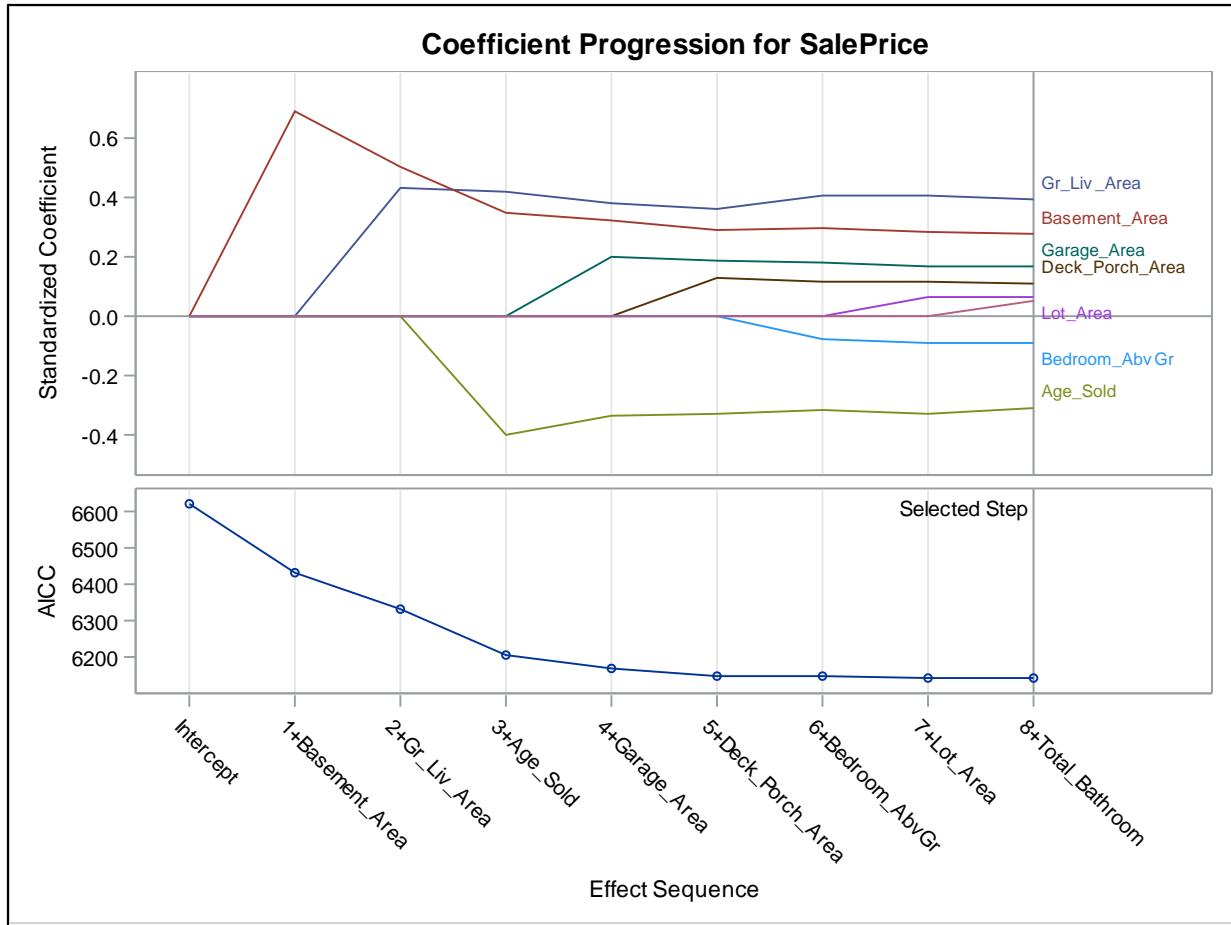
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	3.431321E11	42891512314	155.84
Error	291	80091420996	275228251	
Corrected Total	299	4.232235E11		

Root MSE	16590
Dependent Mean	137525
R-Square	0.8108
Adj R-Sq	0.8056
AIC	6140.79563
AICC	6141.55688
BIC	5841.35042
C(p)	9.00000
SBC	5872.12967

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	44347	6191.271944	7.16
Gr_Liv_Area	1	63.197764	5.585739	11.31
Basement_Area	1	28.692184	3.417034	8.40
Garage_Area	1	35.754191	6.445840	5.55
Deck_Porch_Area	1	31.370539	7.959436	3.94
Lot_Area	1	0.699495	0.316761	2.21
Age_Sold	1	-420.815037	44.219144	-9.52
Bedroom_AbvGr	1	-4834.848748	1688.858227	-2.86
Total_Bathroom	1	3022.124723	1920.839066	1.57

Partial PROC GLMSELECT (SELECT=AICC) Output

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	AICC
0	Intercept		1	6624.2555
1	Basement_Area		2	6432.7045
2	Gr_Liv_Area		3	6334.1618
3	Age_Sold		4	6205.0334
4	Garage_Area		5	6166.9140
5	Deck_Porch_Area		6	6149.2763
6	Bedroom_AbvGr		7	6144.8988
7	Lot_Area		8	6141.9575
8	Total_Bathroom		9	6141.5569*
* Optimal Value Of Criterion				



The results and plots from the AICC selection again mimic those of both AIC and BIC selection.

The parameter estimates table from the AICC selection match those from both AIC and BIC selection.

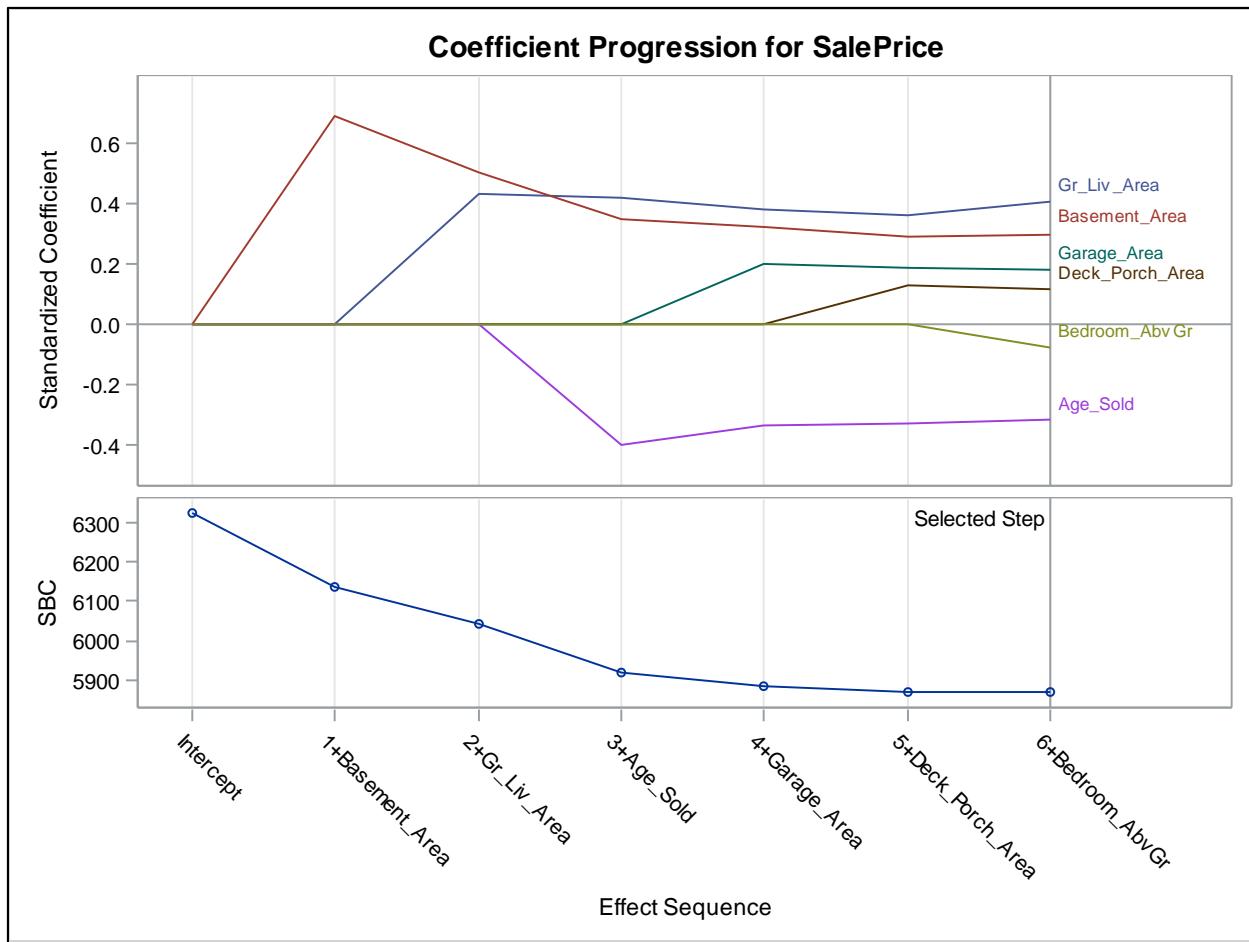
Partial PROC GLMSELECT (SELECT=SBC) Output

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	6325.9189
1	Basement_Area		2	6138.0310
2	Gr_Liv_Area		3	6043.1375
3	Age_Sold		4	5917.6444
4	Garage_Area		5	5883.1463
5	Deck_Porch_Area		6	5869.1154
6	Bedroom_AbvGr		7	5868.3305*

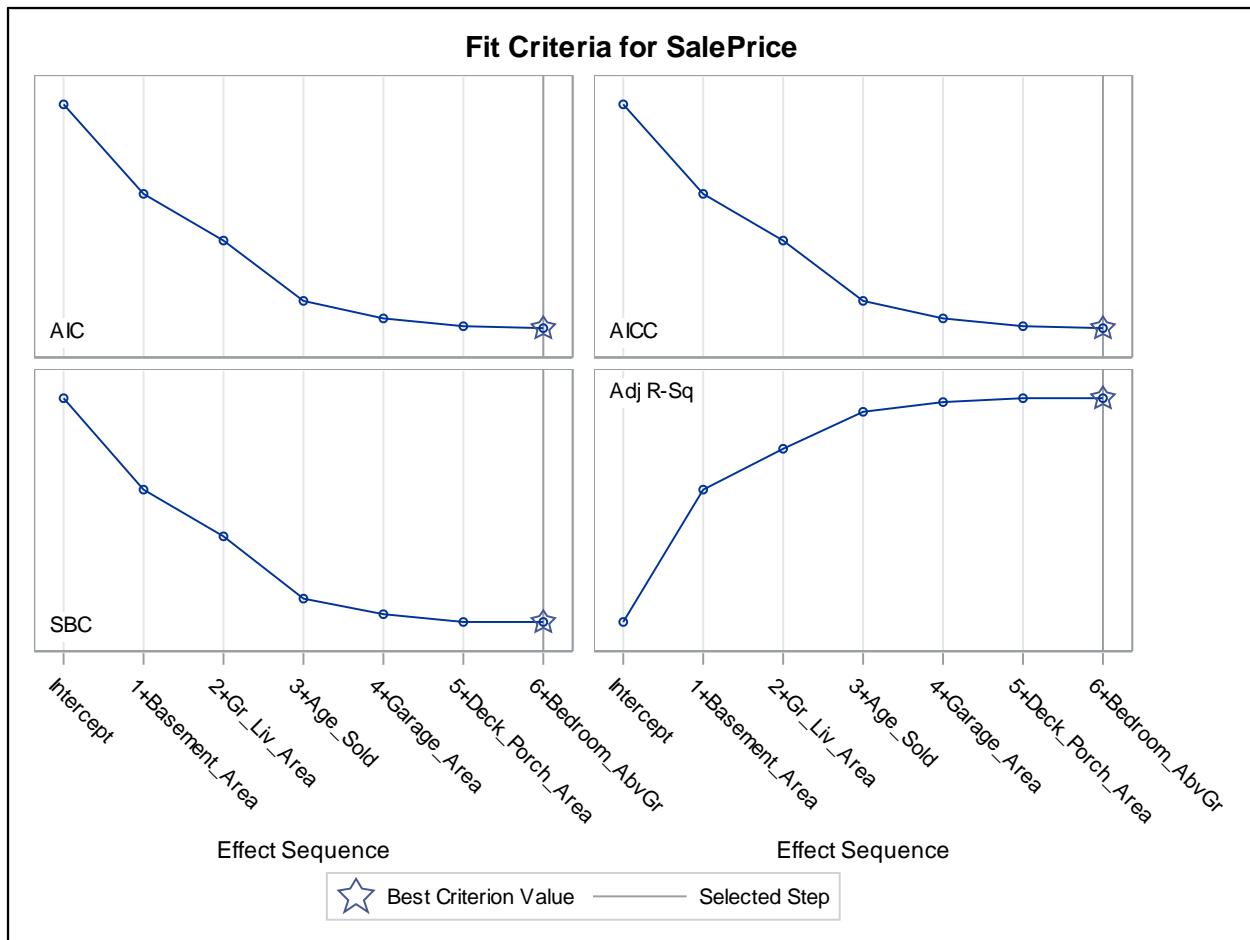
* Optimal Value Of Criterion

Selection stopped at a local minimum of the SBC criterion.

The SBC selection process stops after the sixth step. This is different than all the previous selection setups in this demonstration.



The Coefficient Panel shows similarities to those seen earlier. Larger improvements in SBC can be seen across steps one through three while minimal, compared to the rest, improvements in SBC occur over steps four through six. Like previous images, the standardized coefficients appear to stabilize after step four. One could entertain a four variable model as an option.



The Criteria Panel shows that, accounting only for the models viewed, the optimal fit statistics were obtained step six.

Effects: Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	3.410749E11	56845818595	202.75
Error	293	82148607939	280370676	
Corrected Total	299	4.232235E11		

Root MSE	16744
Dependent Mean	137525
R-Square	0.8059
Adj R-Sq	0.8019
AIC	6144.40398
AICC	6144.89882
SBC	5868.33046

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	48620	5897.324643	8.24
Gr_Liv_Area	1	65.097413	5.472624	11.90
Basement_Area	1	31.279351	3.305546	9.46
Garage_Area	1	38.728785	6.403565	6.05
Deck_Porch_Area	1	32.487956	8.019119	4.05
Age_Sold	1	-434.199118	40.877494	-10.62
Bedroom_AbvGr	1	-4189.095026	1655.065743	-2.53

Note: The decision of which selection methods and criteria to use typically depends on the area of research to which the problem is applied. Standards and practices that are common to your individual research/work area should be considered. When multiple methods are invoked yielding different models, honest assessment with hold-out data can be used and encouraged.

End of Demonstration

What Have You Learned?

- Significance level stepwise, backward, and forward selection (SLENTRY=SLSTAY=0.05) yield a model containing the same 7 effects.
- Stepwise selection using AIC, BIC, and AICC yield a model with all 8 effects included.
- Stepwise selection using SBC yields a model with only 6 effects.

The model selection strategies discussed over the past two sections generates a list of possible models from which you can choose. To aid in the decision of which model is “better,” consultation of a subject area expert can be incorporated. Another option is to perform honest assessment on the models in question using a hold-out data set. This option will be discussed in a later chapter of this course.



Exercises

2. Using Other Model Selection Techniques

Use the **STAT1.BodyFat2** data set to identify a set of “best” models.

- a. With the SELECTION=STEPWISE option, use SELECT=SBC to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.
- b. Try SELECT=AIC.

End of Exercises

4.3 All Possible Selection (Self-Study)

Objectives

- Explain the REG procedure options for all possible model selection.
- Describe model selection options and interpret output to evaluate the fit of several models.

47

Copyright © SAS Institute Inc. All rights reserved.



Model Selection

Data set contains eight interval variables as potential predictors.

Possible Option #1:

Use a form of Stepwise Selection by hand or with assistance from SAS.

Possible Option #2:

Explore all possible models and determine “best.”

48

Copyright © SAS Institute Inc. All rights reserved.



A process for selecting models might be to start with all the interval variables in the **STAT1.ameshousing3** data set and invoke some form of stepwise selection discussed in previous sections. This could be done by hand or with the assistance of SAS.

An alternative option would be to explore all possible models capable from the predictor variables provided and determine which is “best.” This method of all possible selection can be performed using PROC REG.

Model Selection Options

The SELECTION= option in the MODEL statement of PROC REG supports these model selection techniques:

Stepwise selection methods

- STEPWISE, FORWARD, or BACKWARD using significance level

All-possible regressions ranked using

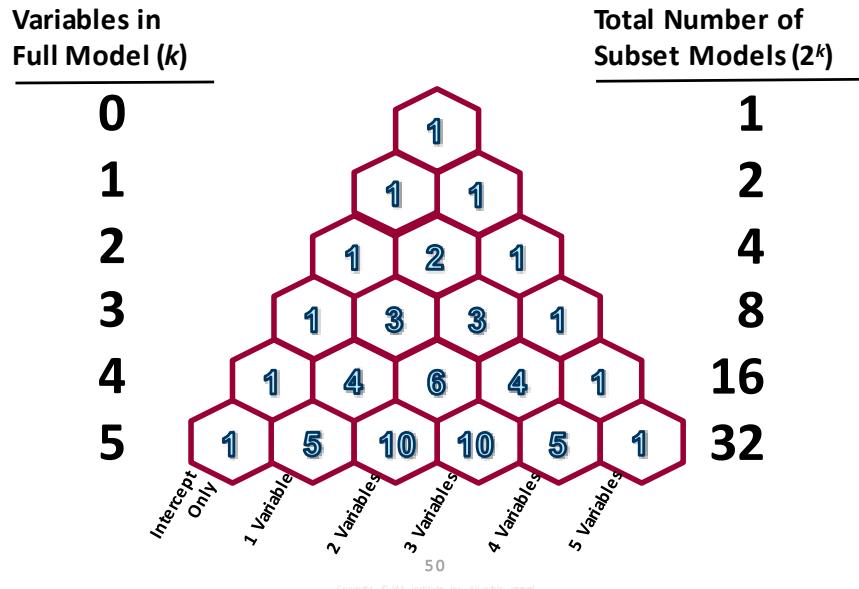
- RSQUARE, ADJRSQ, or CP

SELECTION=NONE is the default.

49

Copyright © 2017, SAS Institute Inc. All rights reserved.

RSQUARE, ADJRSQ, CP Selection Options


Copyright © 2017, SAS Institute Inc. All rights reserved.

In the **STAT1.ameshousing3** data set, there are eight possible independent variables. Therefore, there are $2^8=256$ possible regression models. There are eight possible one-variable models, 28 possible two-variable models, 56 possible three-variable models, and so on.

You can choose to only look at the best n (as measured by the model R^2 for $k=1, 2, 3, \dots, 7$) by using the **BEST=** option on the model statement. The **BEST=** option only reduces the output. All regressions are still calculated.

If there were 20 possible independent variables, there would be more than 1,000,000 models.

Mallows' C_p

- Mallows' C_p is a simple indicator of effective variable selection within a model.
- Look for models with $C_p \leq p$, where p equals the number of parameters in the model, including the intercept.

Mallows recommends choosing the first (fewest variables) model where C_p approaches p .

$$\text{Mallows' } C_p \text{ (1973) is estimated by } C_p = p + \frac{(MSE_p - MSE_{\text{full}})(n - p)}{MSE_{\text{full}}}$$

where

MSE_p is the mean squared error for the model with p parameters.

MSE_{full} is the mean squared error for the full model used to estimate the true residual variance.

n is the number of observations.

p is the number of parameters, including an intercept parameter, if estimated.

The choice of the best model based on C_p is debatable, as will be shown in the slide about Hocking's criterion. Many choose the model with the smallest C_p value. However, Mallows recommended that the best model will have a C_p value approximating p . The most parsimonious model that fits that criterion is generally considered to be a good choice, although subject-matter knowledge should also be a guide in the selection from among competing models.

Hocking's Criterion versus Mallows' C_p

Hocking (1976) suggests selecting a model based on the following:

- $C_p \leq p$ for prediction
- $C_p \leq 2p - p_{full} + 1$ for parameter estimation

Hocking suggested the use of the C_p statistic, but with alternative criteria, depending on the purpose of the analysis. His suggestion of ($C_p \leq 2p - p_{full} + 1$) is included in the REG procedure's calculations of criteria reference plots for best models.



All Possible Model Selection

Example: Invoke PROC REG to produce a regression of **SalePrice** on all the other interval variables in the **STAT1.ameshousing3** data set.

Note: Currently, **stepwise**, **forward**, and **backward** are the only three selection methods that can be chosen in the SAS Studio task. To perform model selection using a method other than these three, either manually edit the generated code or write the code directly.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
      Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d03.sas*/ /*Part A*/
ods graphics on;
proc reg data=STAT1.ameshousing3 plots(only)=(rsquare adjrsq cp);
  ALLPOSS: model SalePrice=&interval
            / selection=rsquare adjrsq cp;
  title "All Possible Model Selection for SalePrice";
run;
quit;
```

Selected MODEL statement options:

SELECTION= enables you to choose the different selection methods – RSQUARE, ADJRSQ, and CP. The first listed method is the one that determines the sorting order in the output.

Selected **SELECTION=** option methods:

RSQUARE tells PROC REG to use the model R-square to rank the model from best to worst for a given number of variables.

ADJRSQ prints the adjusted R-square for each model.

CP prints Mallows' C_p statistic for each model.

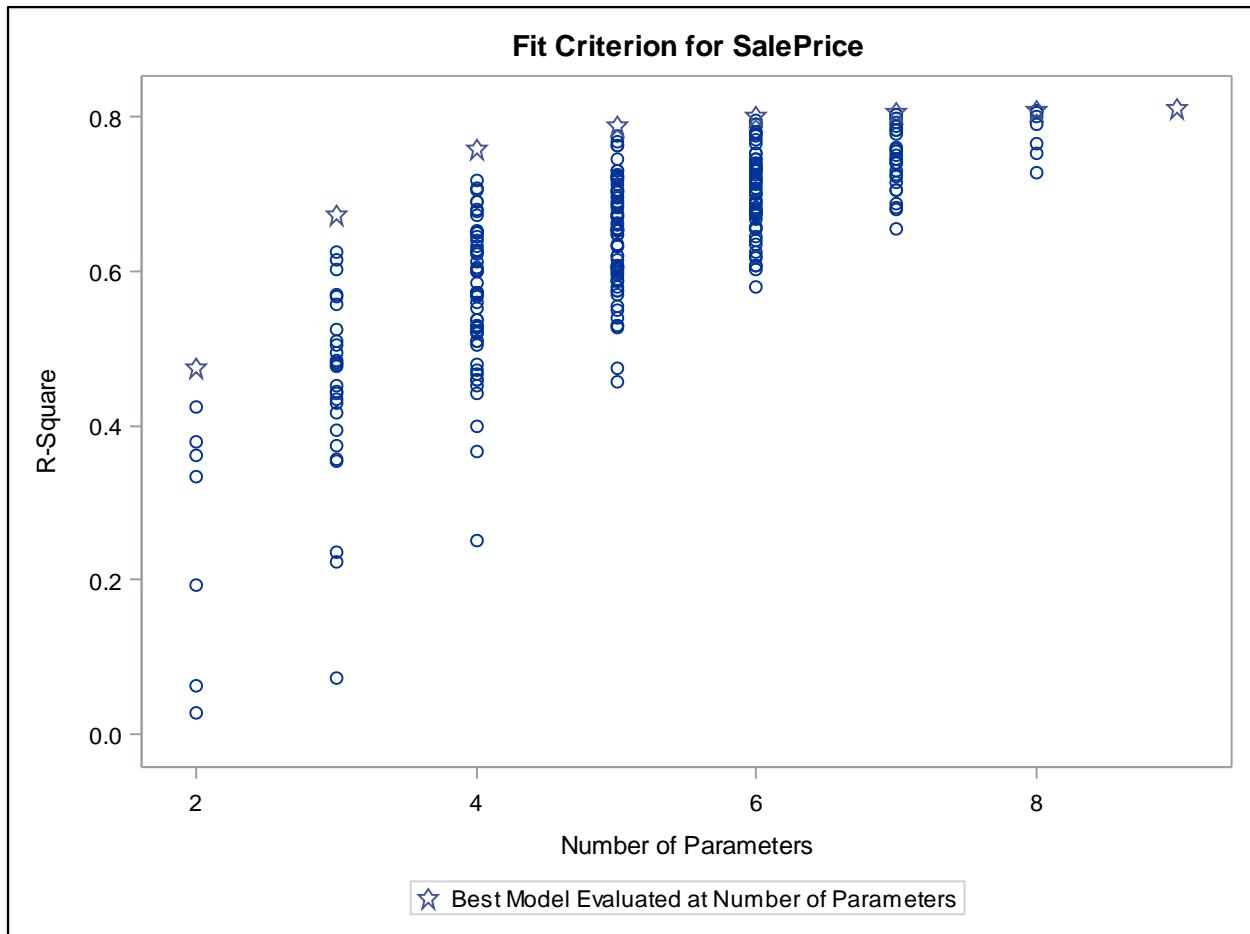
Partial PROC REG Output

Number of Observations Read	300
Number of Observations Used	300

Model Index	Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	1	0.4755	0.4737	510.5367	Basement_Area
2	1	0.4231	0.4212	591.1052	Gr_Liv_Area
3	1	0.3787	0.3767	659.3108	Age_Sold
4	1	0.3605	0.3584	687.3455	Total_Bathroom
5	1	0.3351	0.3329	726.3533	Garage_Area
6	1	0.1935	0.1908	944.1662	Deck_Porch_Area
7	1	0.0642	0.0610	1143.019	Lot_Area
8	1	0.0275	0.0243	1199.378	Bedroom_AbvGr

Model Index	Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
9	2	0.6725	0.6703	209.5529	Gr_Liv_Area Age_Sold
10	2	0.6249	0.6224	282.7594	Gr_Liv_Area Basement_Area
11	2	0.6148	0.6122	298.3135	Basement_Area Age_Sold
12	2	0.6027	0.6000	316.9559	Basement_Area Garage_Area
13	2	0.5708	0.5679	365.9609	Gr_Liv_Area Garage_Area
14	2	0.5680	0.5651	370.2769	Basement_Area Total_Bathroom

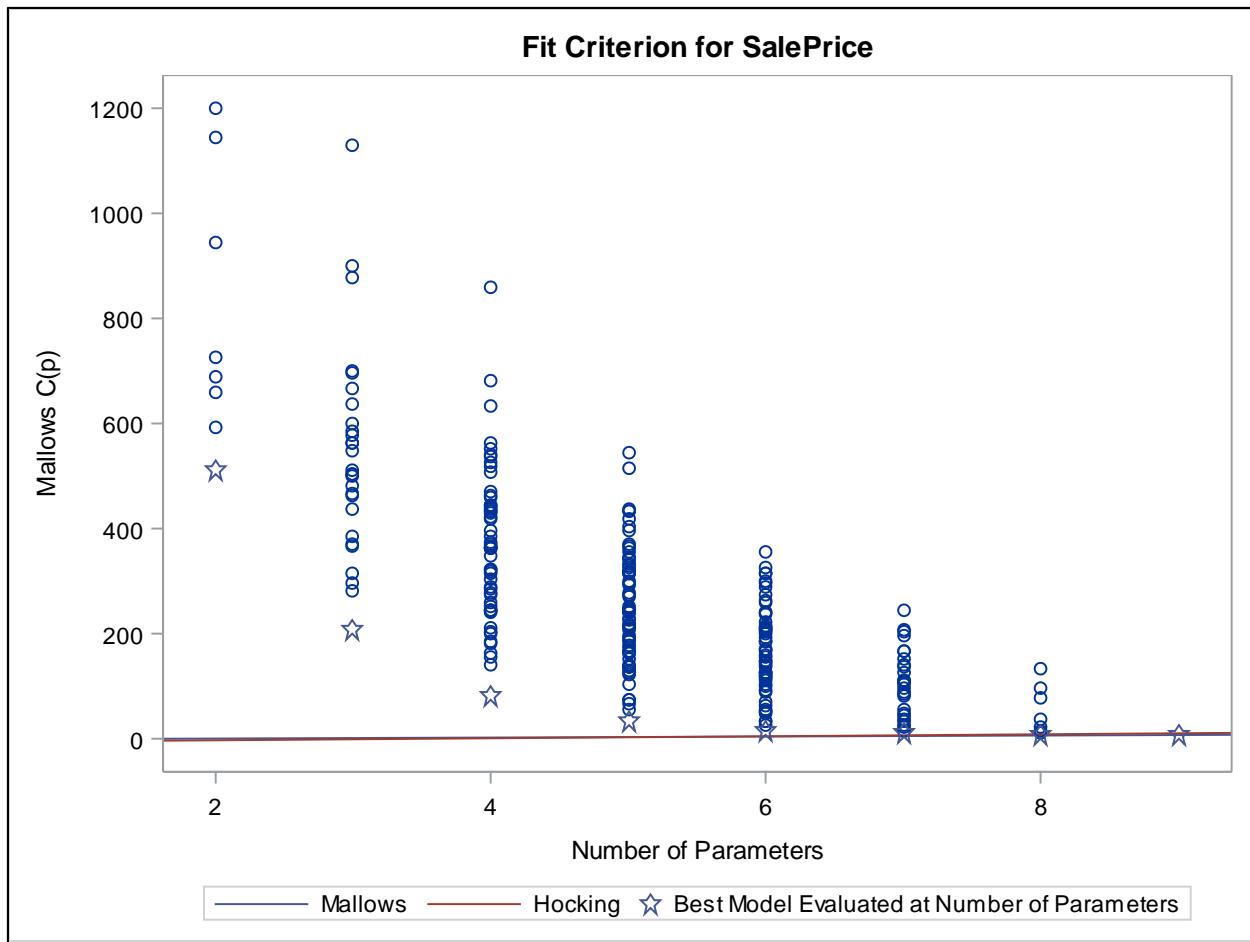
There are many models to compare. It would be unwieldy to try to determine the best model by viewing the output tables. Therefore, it is advisable to look at the ODS plots.



The R-square plot compares all models based on their R-square values. As noted earlier, adding variables to a model always increases R-square, and therefore the full model is always best. Therefore, you can only use the R-square value to compare models of equal numbers of parameters.



The adjusted R-square does not have the problem that the R-square has. You can compare models of different sizes. In this case, it is difficult to see which model has the higher adjusted R-square, the starred model for seven parameters or eight parameters.



The line $C_p=p$ is plotted to help you identify models that satisfy the criterion $C_p \leq p$ for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion $C_p \leq 2p - p_{\text{full}} + 1$ for parameter estimation.

Use the graph and review the output to select a relatively short list of models that satisfy the criterion appropriate for your objective. It is often the case that the best model is difficult to see because of the range of C_p values at the high end. These models are clearly not the best and therefore you can focus on the models near the bottom of the range of C_p .

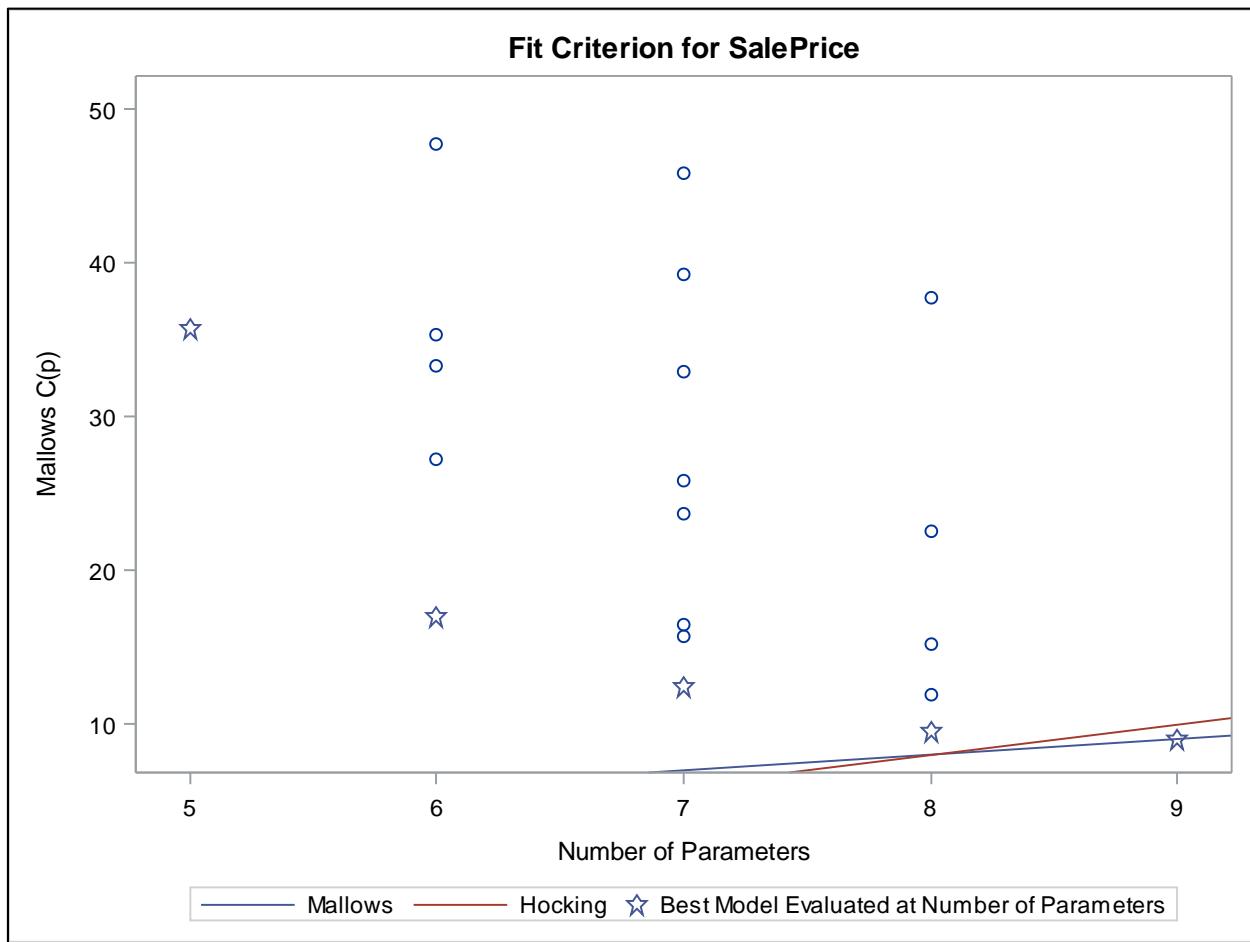
```
/*st104d03.sas*/ /*Part B*/
proc reg data=STAT1.ameshousing3 plots(only)=(cp);
  ALLPOSS: model SalePrice=&interval / selection=cp rsquare adjrsq
best=20;
  title "Best Models Using All Possible Selection for SalePrice";
run;
quit;
```

Selected SELECTION= option methods:

BEST=n limits the output to only the best n models.

Model Index	Number in Model	C(p)	R-Square	Adjusted R-Square	Variables in Model
1	8	9.0000	0.8108	0.8056	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom
2	7	9.4754	0.8091	0.8046	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
3	7	11.8765	0.8076	0.8030	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr Total_Bathroom
4	6	12.4745	0.8059	0.8019	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr
5	7	15.1956	0.8054	0.8008	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Total_Bathroom
6	6	15.7530	0.8038	0.7997	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold Total_Bathroom
7	6	16.4459	0.8033	0.7993	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold
8	5	17.0005	0.8017	0.7983	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold
9	7	22.5339	0.8007	0.7959	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom
10	6	23.7403	0.7986	0.7944	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold Bedroom_AbvGr
11	6	25.8313	0.7972	0.7931	Gr_Liv_Area Basement_Area Garage_Area Age_Sold Bedroom_AbvGr Total_Bathroom
12	5	27.1943	0.7950	0.7915	Gr_Liv_Area Basement_Area Garage_Area Age_Sold Bedroom_AbvGr
13	6	32.9173	0.7926	0.7884	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold Total_Bathroom
14	5	33.3028	0.7911	0.7875	Gr_Liv_Area Basement_Area Garage_Area Age_Sold Total_Bathroom
15	5	35.3618	0.7897	0.7861	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold
16	4	35.7387	0.7882	0.7853	Gr_Liv_Area Basement_Area Garage_Area Age_Sold
17	7	37.7677	0.7907	0.7857	Gr_Liv_Area Basement_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom
18	6	39.3019	0.7885	0.7841	Gr_Liv_Area Basement_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
19	6	45.8708	0.7842	0.7798	Gr_Liv_Area Basement_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr Total_Bathroom
20	5	47.7363	0.7817	0.7780	Gr_Liv_Area Basement_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr

Investigate the plot of Mallows' C(p).



In this example, the number of parameters in the full model, p_{full} , equals 9 (eight variables plus the intercept).

The smallest model that falls under the Hocking line has $p=9$, the full model. This model also has a Cp value that is equal to p exactly, falling directly on Mallows line. From this information, your full model appears to be a potential model for prediction and variable explanation. This result is likely to change if additional continuous predictors are included in the analysis.

If multiple models, sharing the same number of parameters, fall below these lines, there are several options that can be used to make a decision. First, the analyst can appeal to a subject matter expert who could potentially provide previous experiences that could "break the tie." Secondly, other fit statistics could be used as a comparison between the models. Perhaps one of the models has a higher adjusted R-square value. Thirdly, the models in question could be compared using a hold-out data set, especially when the focus is prediction.

End of Demonstration

4.02 Multiple Choice Poll

Which value tends to increase (can never decrease) as you add predictor variables to your regression model?

- a. R square
- b. Adjusted R square
- c. Mallows' C_p
- d. Both a and b
- e. F statistic
- f. All of the above



Exercises

3. Using All-Regression Techniques

Use the **STAT1.BodyFat2** data set to identify a set of “best” models.

- a. With the SELECTION=CP option, use an all-possible regression technique to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.

Hint: Select only the best 60 models based on C_p to compare.

End of Exercises

4.4 Solutions

Solutions to Exercises

1. Using Significance Level Model Selection Techniques

Use the **STAT1.BodyFat2** data set to identify a set of “best” models.

- a. With the SELECTION=STEPWISE option, use SELECT=SL to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**. Use the default values for SLENTRY and SLSTAY.
 - 1) Open the **Linear Regression** task under Statistics.
 - 2) On the DATA tab, select the appropriate **BodyFat2** data set and variables.
 - 3) On the MODEL tab, use the effect builder to specify the appropriate model.
 - 4) On the OPTIONS tab, suppress all the plots.
 - 5) On the SELECTION tab, choose **Stepwise selection** as the selection method and **Significance level** as the criterion to add/remove effects. Specify the significance level to add an effect to the model to **0.15** instead of 0.05.

Note: In PROC GLMSELECT, the default significance level is set to 0.15. However, in SAS Studio, the default significance level is set to 0.05 for all three methods.

- 6) Expand the SELECTION PLOTS property and select the options to plot both the criteria plots and the coefficient plots.
- 7) Run the code.

Note: Alternatively, you can write the code directly in SAS.

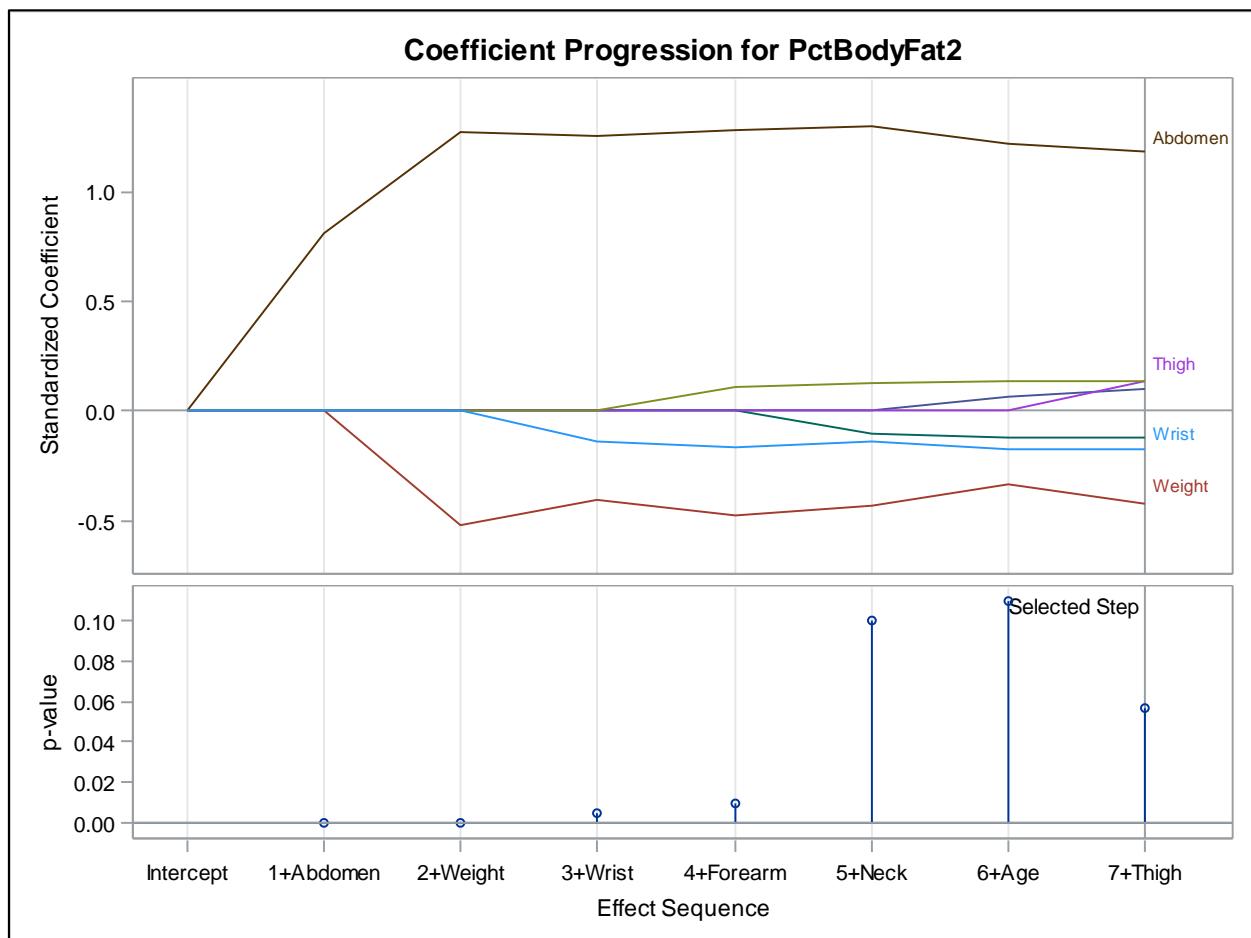
```
/*st104s01.sas*/ /*Part A*/
ods graphics on;
proc glmselect data=STAT1.bodyfat2 plots=all;
  STEPWISESEL: model PctBodyFat2=Age Weight Height Neck Chest
    Abdomen Hip Thigh Knee Ankle Biceps Forearm
    Wrist / SELECTION=STEPWISE SELECT=SL;
  title 'SL STEPWISE Selection with PctBodyFat2';
run;
quit;
```

Partial PROC GLMSELECT Output

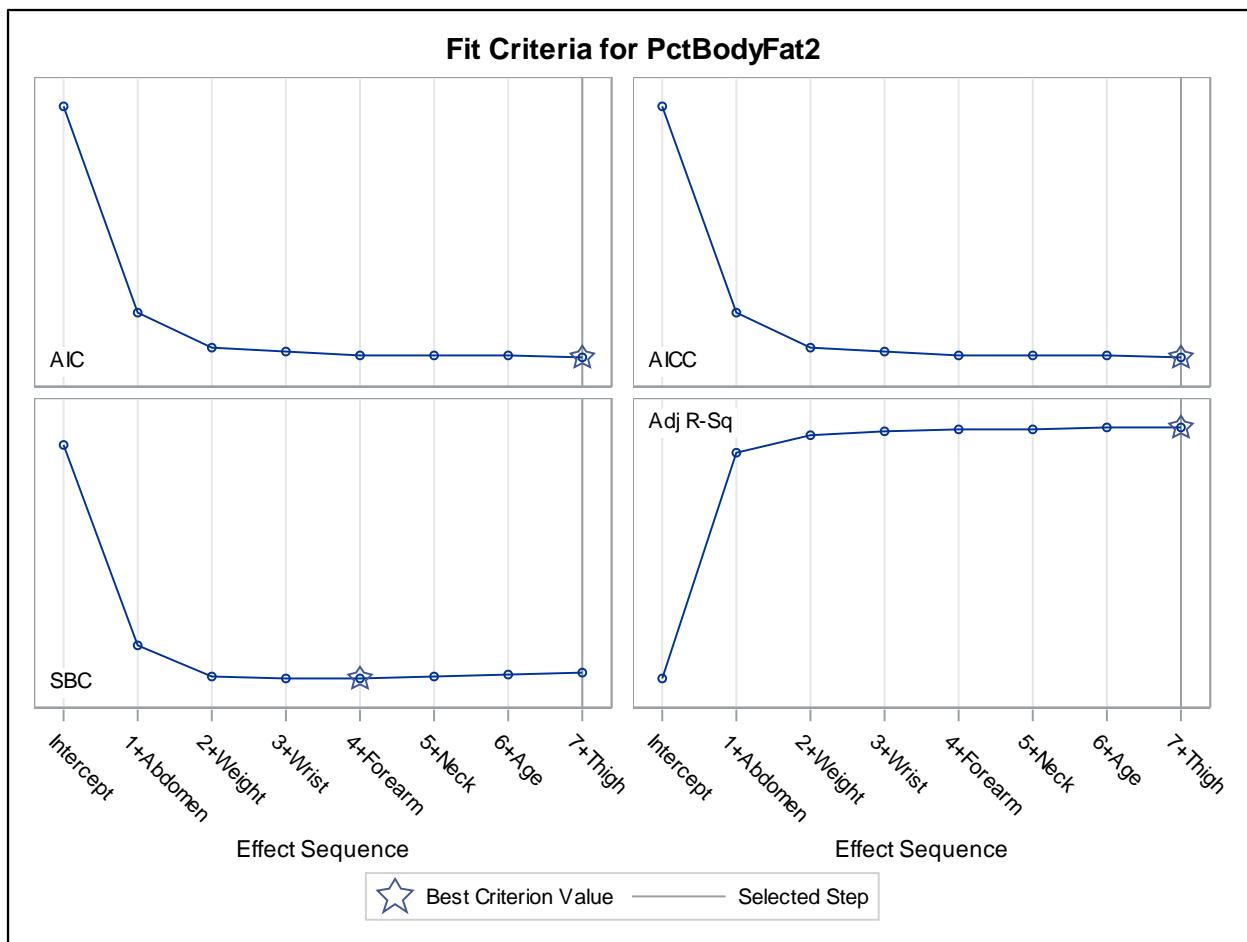
Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	Abdomen		2	488.93	<.0001
2	Weight		3	50.58	<.0001
3	Wrist		4	8.15	0.0047
4	Forearm		5	6.78	0.0098
5	Neck		6	2.73	0.1000
6	Age		7	2.58	0.1098
7	Thigh		8	3.66	0.0569

Selection stopped because the candidate for entry has SLE > 0.15 and the candidate for removal has SLS < 0.15.

The STEPWISE selection process, using significance level appears to select an eight effect model (including the intercept).



The Coefficient Panel shows that the standardized coefficients do not vary greatly as additional effects are added to the model.



The Fit Panel indicates that the best model, according to AIC, AICC, and Adjusted R-square, is the final model viewed during the selection process. SBC shows a minimum at step four.

Effects: Intercept Age Weight Neck Abdomen Thigh Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	13087	1869.59160	101.56
Error	244	4491.84861	18.40922	
Corrected Total	251	17579		

Root MSE	4.29060
Dependent Mean	19.15079
R-Square	0.7445
Adj R-Sq	0.7371
AIC	995.90881
AICC	996.65261
SBC	770.14425

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-33.257991	9.006812	-3.69
Age	1	0.068166	0.030792	2.21
Weight	1	-0.119441	0.034025	-3.51
Neck	1	-0.403802	0.220620	-1.83
Abdomen	1	0.917885	0.069499	13.21
Thigh	1	0.221960	0.116013	1.91
Forearm	1	0.553139	0.184788	2.99
Wrist	1	-1.532401	0.510415	-3.00

The parameter estimates from the selected model are presented in the Parameter Estimates table.

- b. Try FORWARD.
- 1) On the SELECTION tab, modify the selection method to **Forward selection**. Also change the significance level to **0.5**, the default value in PROC GLMSELECT.
 - 2) Rerun the linear regression task with the modified selection method.

Note: Alternatively, you can write the code directly in SAS.

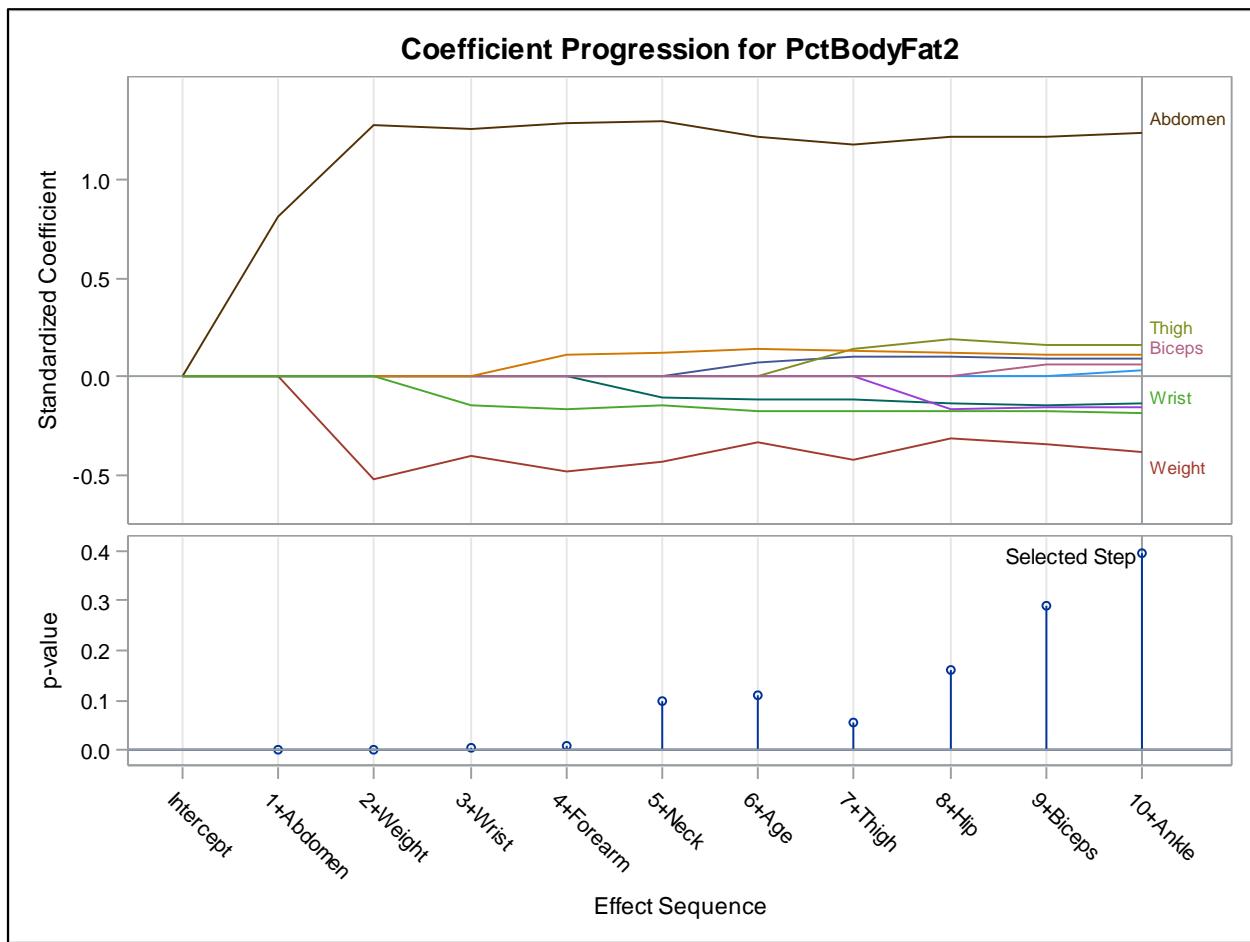
```
/*st104s01.sas*/ /*Part B*/
proc glmselect data=STAT1.bodyfat2 plots=all;
  FORWARDSL: model PctBodyFat2=Age Weight Height Neck Chest
              Abdomen Hip Thigh Knee Ankle Biceps Forearm
              Wrist / SELECTION=FORWARD SELECT=SL;
  title 'SL FORWARD Selection with PctBodyFat2';
run;
quit;
```

Partial PROC GLMSELECT Output

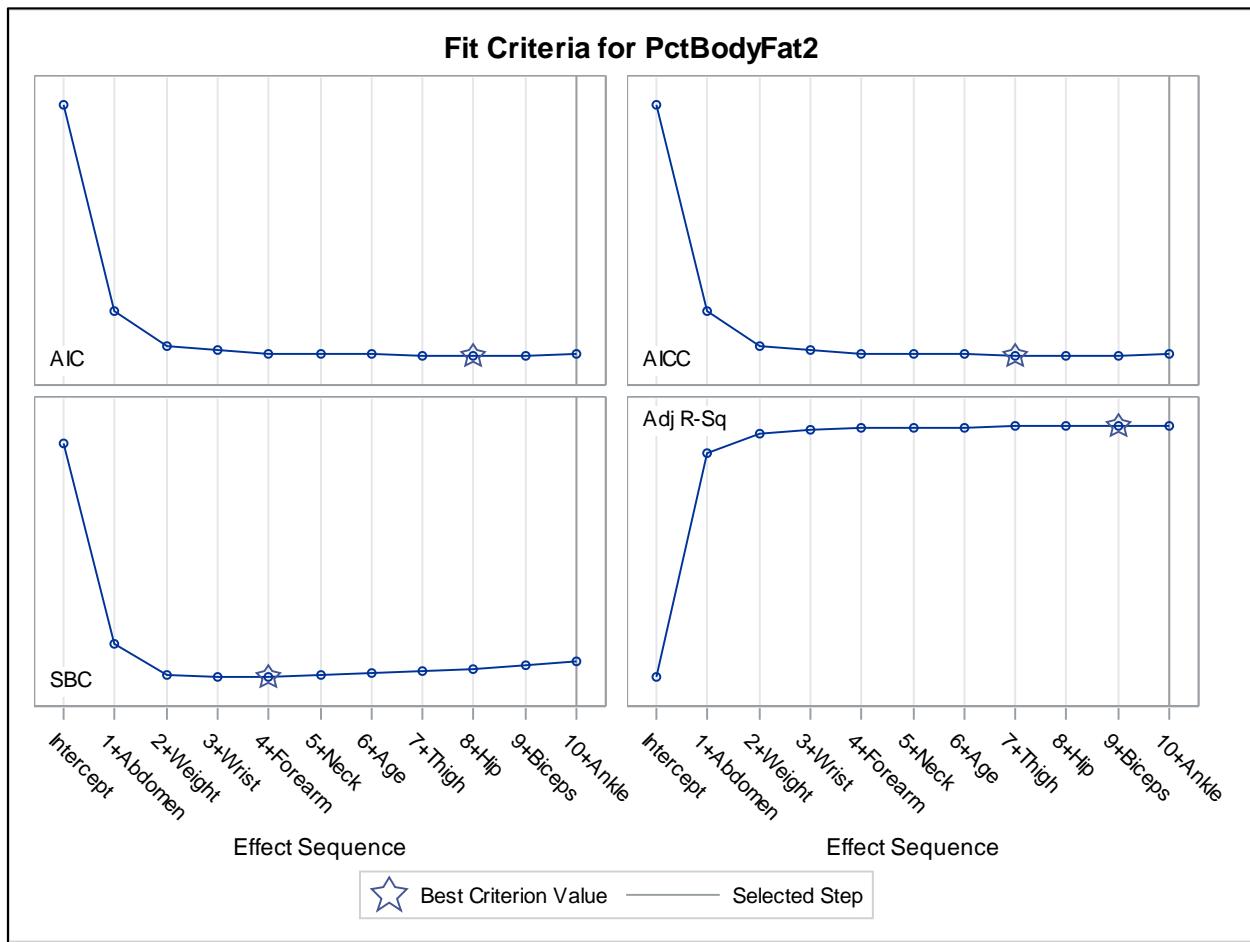
Forward Selection Summary				
Step	Effect Entered	Number Effects In	F Value	Pr > F
0	Intercept	1	0.00	1.0000
1	Abdomen	2	488.93	<.0001
2	Weight	3	50.58	<.0001
3	Wrist	4	8.15	0.0047
4	Forearm	5	6.78	0.0098
5	Neck	6	2.73	0.1000
6	Age	7	2.58	0.1098
7	Thigh	8	3.66	0.0569
8	Hip	9	1.99	0.1594
9	Biceps	10	1.13	0.2888
10	Ankle	11	0.72	0.3957

Selection stopped as the candidate for entry has SLE > 0.5.

The FORWARD selection process, using significance level appears to select an eleven effect model (including the intercept).



The Coefficient Panel shows that the standardized coefficients do not vary greatly as additional effects are added to the model.



The Fit Panel indicates that the best model, according to AIC, AICC, Adjusted R-square and SBC, are at various steps in the selection progression.

Effects: Intercept Age Weight Neck Abdomen Hip Thigh Ankle Biceps Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	10	13158	1315.76595	71.72
Error	241	4421.33035	18.34577	
Corrected Total	251	17579		

Root MSE	4.28320
Dependent Mean	19.15079
R-Square	0.7485
Adj R-Sq	0.7381
AIC	997.92124
AICC	999.22668
SBC	782.74496

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-25.999624	12.153156	-2.14
Age	1	0.065093	0.030919	2.11
Weight	1	-0.107396	0.042068	-2.55
Neck	1	-0.467490	0.228115	-2.05
Abdomen	1	0.957721	0.072760	13.16
Hip	1	-0.179124	0.139083	-1.29
Thigh	1	0.259259	0.133892	1.94
Ankle	1	0.184526	0.216864	0.85
Biceps	1	0.186171	0.168580	1.10
Forearm	1	0.453031	0.195932	2.31
Wrist	1	-1.656662	0.527061	-3.14

The parameter estimates from the selected model are presented in the Parameter Estimates table.

- c. How many variables would result from a model using FORWARD selection and a significance level for entry criterion of 0.05, instead of the default SLENTRY of 0.50?
 - 1) On the SELECTION tab, change the significance level to 0.05.

Note: Alternatively, you can modify the code.

```
/*st104s01.sas*/ /*Part C*/
proc glmselect data=STAT1.bodyfat2 plots=all;
  FORWARDSL: model PctBodyFat2=Age Weight Height Neck Chest
              Abdomen Hip Thigh Knee Ankle Biceps Forearm
              Wrist / SELECTION=FORWARD SELECT=SL
              SLENTRY=0.05;
  title 'SL FORWARD (0.05) Selection with PctBodyFat2';
run;
quit;
```

Partial PROC GLMSELECT Output

Forward Selection Summary				
Step	Effect Entered	Number Effects In	F Value	Pr > F
0	Intercept	1	0.00	1.0000
1	Abdomen	2	488.93	<.0001
2	Weight	3	50.58	<.0001
3	Wrist	4	8.15	0.0047
4	Forearm	5	6.78	0.0098

When the SLENTRY is changed from default to 0.05, the number of effects in the selected model reduces to five (including the intercept).

2. Using Other Model Selection Techniques

Use the STAT1.BodyFat2 data set to identify a set of “best” models.

- a. With the SELECTION=STEPWISE option, use SELECT=SBC to identify a set of candidate models that predict PctBodyFat2 as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.
 - 1) Open the **Linear Regression** task under Statistics.
 - 2) On the DATA tab, select the **BodyFat2** data set and assign the variables.
 - 3) On the MODEL tab, specify the appropriate model.
 - 4) On the OPTIONS tab, suppress all the plots.
 - 5) On the SELECTION tab, choose **Stepwise selection** as the selection method and choose **Schwarz Bayesian information criterion** to use **SBC** as the criterion.
 - 6) Expand the SELECTION PLOTS property and select the options to plot both the criteria plots and the coefficient plots.
 - 7) Run the code.

Note: Alternatively, you can write the code directly. In SAS.

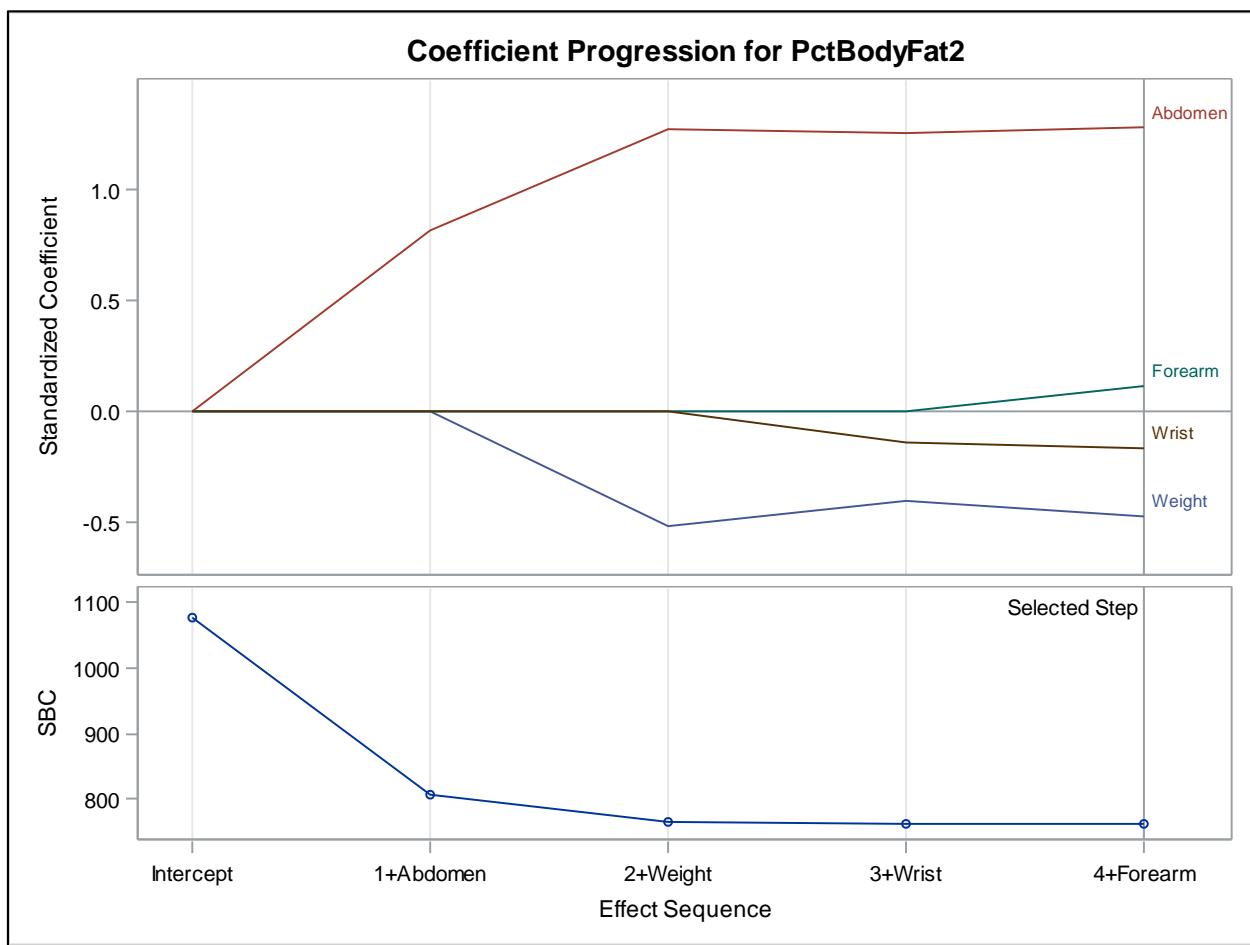
```
/*st104s02.sas*/ /*Part A*/
ods graphics on;
proc glmselect data=STAT1.bodyfat2 plots=all;
  STEPWISESBC: model PctBodyFat2=Age Weight Height Neck Chest
    Abdomen Hip Thigh Knee Ankle Biceps Forearm
    Wrist / SELECTION=STEPWISE SELECT=SBC;
  title 'SBC STEPWISE Selection with PctBodyFat2';
run;
quit;
```

Partial PROC GLMSELECT Output

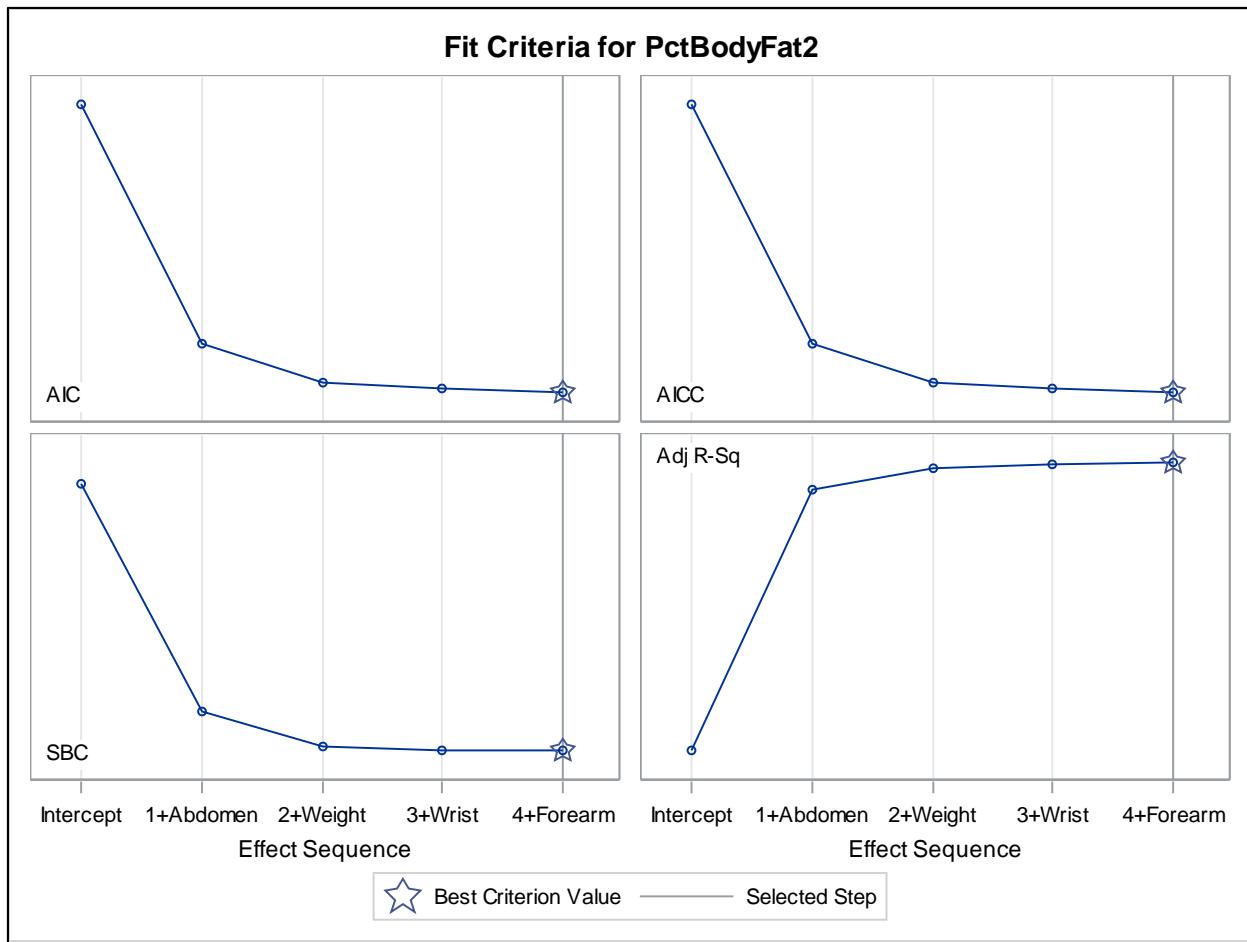
Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	1075.2771
1	Abdomen		2	807.7042
2	Weight		3	766.6280
3	Wrist		4	764.0139
4	Forearm		5	762.7218*

* Optimal Value Of Criterion

The STEPWISE selection process, using SELECT=SBC appears to select a five effect model (including the intercept).



The Coefficient Panel shows that the standardized coefficients do not vary greatly as additional effects are added to the model.



The Fit Panel indicates that the best model, according to AIC, AICC, Adjusted R-square and SBC, is the final model viewed during the selection process. Remember that this statement is made comparing only the models that were viewed in these steps of the selection process.

Effects: Intercept Weight Abdomen Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	12921	3230.18852	171.28
Error	247	4658.23577	18.85925	
Corrected Total	251	17579		

Root MSE	4.34272
Dependent Mean	19.15079
R-Square	0.7350
Adj R-Sq	0.7307
AIC	999.07467
AICC	999.41753
SBC	762.72182

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-34.854074	7.245005	-4.81
Weight	1	-0.135631	0.024748	-5.48
Abdomen	1	0.995751	0.056066	17.76
Forearm	1	0.472928	0.181661	2.60
Wrist	1	-1.505562	0.442666	-3.40

The parameter estimates from the selected model are presented in the Parameter Estimates table.

- b. Try SELECT=AIC.
 - 1) On the SELECTION tab, modify the criterion to **Akaike's information criterion**.
 - 2) Rerun the task.

Note: Alternatively, you can modify the code directly.

```
/*st104s02.sas*/ /*Part B*/
proc glmselect data=STAT1.bodyfat2 plots=all;
  STEPWISEAIC: model PctBodyFat2=Age Weight Height Neck Chest
    Abdomen Hip Thigh Knee Ankle Biceps Forearm
    Wrist / SELECTION=STEPWISE SELECT=AIC;
  title 'AIC STEPWISE Selection with PctBodyFat2';
run;
quit;
```

Partial PROC GLMSELECT Output

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	AIC
0	Intercept		1	1325.7477
1	Abdomen		2	1054.6453
2	Weight		3	1010.0398
3	Wrist		4	1003.8962
4	Forearm		5	999.0747
5	Neck		6	998.2968
6	Age		7	997.6612
7	Thigh		8	995.9088
8	Hip		9	995.8514*

* Optimal Value Of Criterion

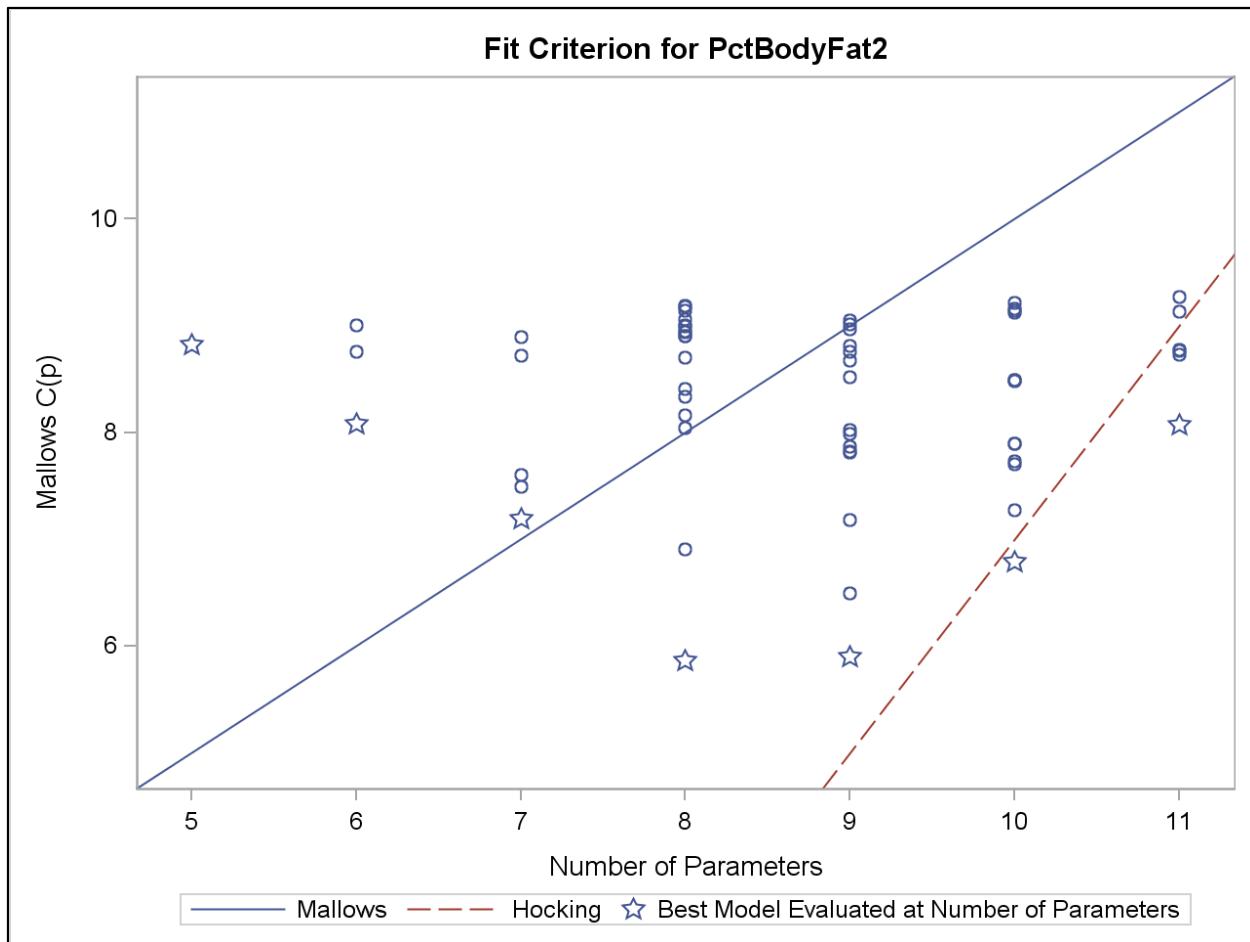
Using SELECT=AIC, the selected model contains nine effects (including the intercept).

3. Using All-Regression Techniques

- a. With the SELECTION=CP option, use an all-possible regression technique to identify a set of candidate models that predict PctBodyFat2 as a function of the variables Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, and Wrist.
Hint: Select only the best 60 models based on C_p to compare.

```
/*st104s03.sas*/ /*Part A*/
ods graphics / imagemap=on;

proc reg data=STAT1.BodyFat2 plots(only)=(cp);
  model PctBodyFat2=Age Weight Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
    / selection=cp best=60;
  title "Using Mallows Cp for Model Selection";
run;
quit;
```



The plot indicates that the best model according to Mallows' criterion is an eight-parameter (seven variables plus an intercept) model. The best model according to Hocking's criterion has 10 parameters (including the intercept).

A partial table of the 60 models, their $C(p)$ values, and the numbers of variables in the models is displayed.

Model Index	Number in Model	C(p)	R-Square	Variables in Model
1	7	5.8653	0.7445	Age Weight Neck Abdomen Thigh Forearm Wrist
2	8	5.8986	0.7466	Age Weight Neck Abdomen Hip Thigh Forearm Wrist
3	8	6.4929	0.7459	Age Weight Neck Abdomen Thigh Biceps Forearm Wrist
4	9	6.7834	0.7477	Age Weight Neck Abdomen Hip Thigh Biceps Forearm Wrist
5	7	6.9017	0.7434	Age Weight Neck Abdomen Biceps Forearm Wrist
6	8	7.1778	0.7452	Age Weight Neck Abdomen Thigh Ankle Forearm Wrist
7	6	7.1860	0.7410	Age Weight Abdomen Thigh Forearm Wrist
8	9	7.2729	0.7472	Age Weight Neck Abdomen Hip Thigh Ankle Forearm Wrist
9	6	7.4937	0.7406	Age Weight Neck Abdomen Forearm Wrist
10	6	7.6018	0.7405	Weight Neck Abdomen Biceps Forearm Wrist
11	9	7.7067	0.7468	Age Weight Neck Abdomen Thigh Ankle Biceps Forearm Wrist
12	9	7.7282	0.7467	Age Weight Height Neck Abdomen Hip Thigh Forearm Wrist
13	8	7.8146	0.7445	Age Weight Height Neck Abdomen Thigh Forearm Wrist
14	8	7.8246	0.7445	Age Weight Neck Chest Abdomen Thigh Forearm Wrist
15	8	7.8651	0.7445	Age Weight Neck Abdomen Thigh Knee Forearm Wrist
16	9	7.8966	0.7466	Age Weight Neck Abdomen Hip Thigh Knee Forearm Wrist
17	9	7.8986	0.7466	Age Weight Neck Chest Abdomen Hip Thigh Forearm Wrist
18	8	7.9907	0.7443	Age Weight Neck Abdomen Ankle Biceps Forearm Wrist

Note: Number in Model does not include the intercept in this table.

The best MALLOWS model is either the eight-parameter models, number 1 (includes the variables Age, Weight, Neck, Abdomen, Thigh, Forearm, and Wrist) or number 5 (includes the variables Age, Weight, Neck, Abdomen, Biceps, Forearm, and Wrist).

The best HOCKING model is number 4. It includes Hip, along with the variables in the best MALLOWS models listed above.

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

4.01 Poll – Correct Answer

The STEPWISE, BACKWARD, and FORWARD strategies result in the same final model if the same significance levels are used in all three.

- True
- False

37

Copyright © SAS Institute Inc. All rights reserved.

4.02 Multiple Choice Poll – Correct Answer

Which value tends to increase (can never decrease) as you add predictor variables to your regression model?

- a. R square
- b. Adjusted R square
- c. Mallows' C_p
- d. Both a and b
- e. F statistic
- f. All of the above

55

Copyright © SAS Institute Inc. All rights reserved.

Chapter 5 Model Post-Fitting for Inference

5.1 Examining Residuals.....	5-3
Demonstration: Residual Plots.....	5-10
Exercises.....	5-17
5.2 Influential Observations.....	5-18
Demonstration: Looking for Influential Observations.....	5-25
Exercises.....	5-36
5.3 Collinearity	5-37
Demonstration: Example of Collinearity.....	5-42
Demonstration: Collinearity Diagnostics	5-44
Exercises.....	5-50
5.4 Solutions	5-51
Solutions to Exercises	5-51
Solutions to Student Activities (Polls/Quizzes)	5-64

5.1 Examining Residuals

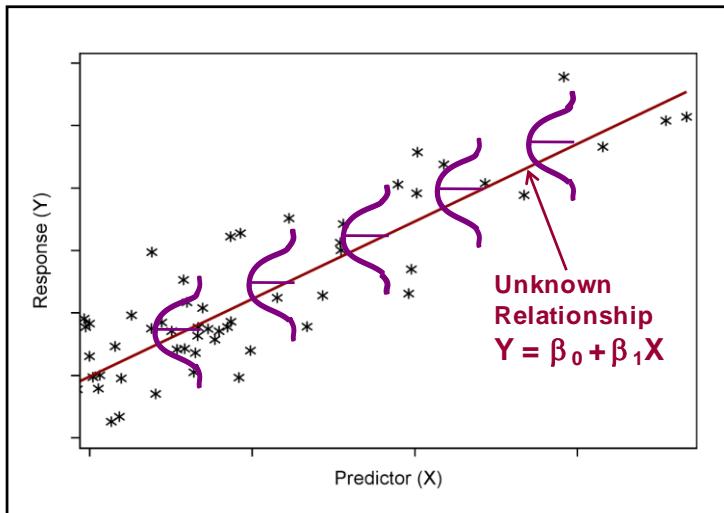
Objectives

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

3



Assumptions for Regression



4



Recall that the model for the linear regression has the form $Y=\beta_0+\beta_1X+\varepsilon$. 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

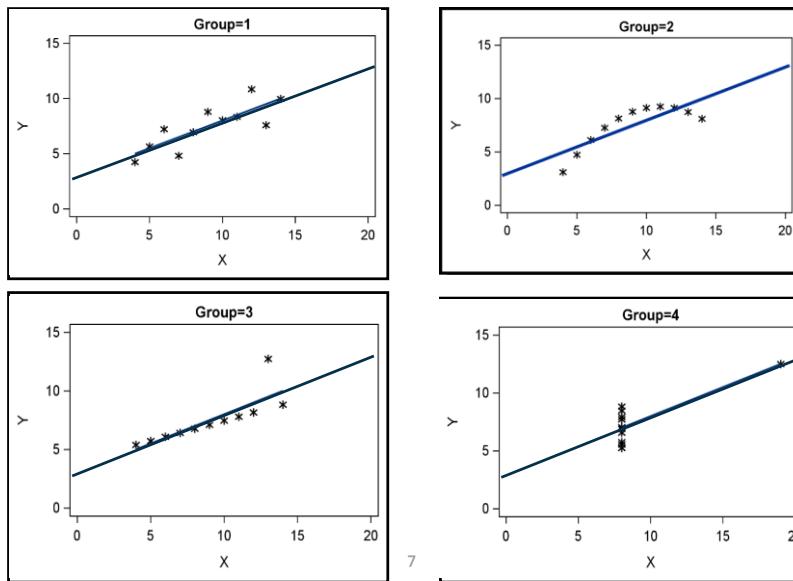
- have a mean of 0 at each value of the predictor variable
- are normally distributed at each value of the predictor variable
- have the same variance at each value of the predictor variable
- are independent.

5.01 Poll

Predictor variables are assumed to be normally distributed in linear regression models.

- True
- False

Importance of Plotting Data



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, $Y=3.0+0.5X$, and the R-square statistic, 0.67, are the same.

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

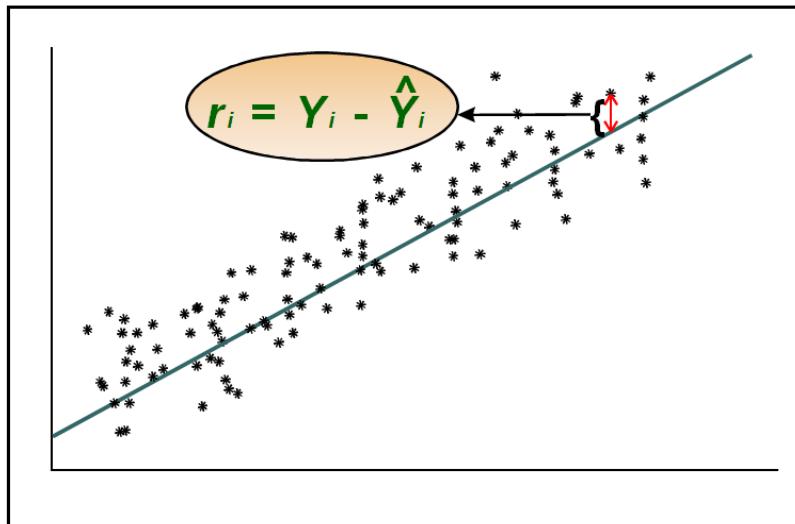
In the second plot, a simple linear regression model is not appropriate because you are fitting a straight line through a curvilinear relationship.

In the third plot, there seems to be an outlying data value that is affecting the regression line. This outlier is an influential data value in that it is substantially changing the fit of the regression line.

In the fourth plot, the outlying data point dramatically changes the fit of the regression line. In fact, the slope would be undefined without the outlier.

The four plots illustrate that relying on the regression output to describe the relationship between your variables can be misleading. The regression equations and the R-square 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



8

Sas

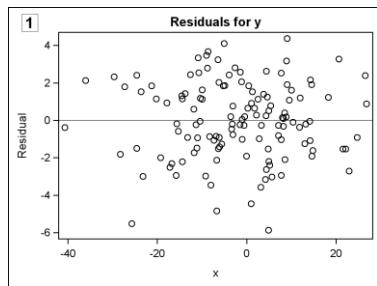
To verify the assumptions for regression, you can use the residual values from the regression analysis as your best estimates of the error terms. Residuals are defined as follows: $r_i = Y_i - \hat{Y}_i$

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

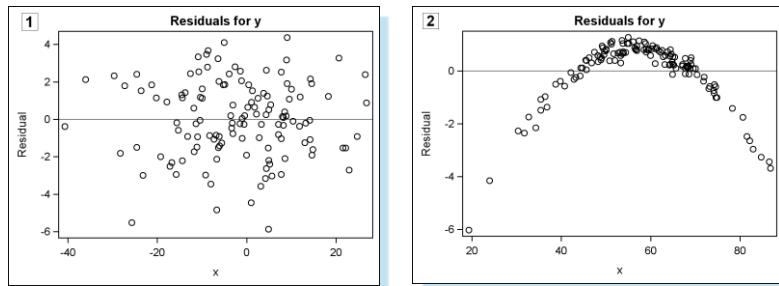
Examining Residual Plots



- Residuals are randomly scattered about zero reference line.
- No patterns found.
- Model form appears to be adequate.

9

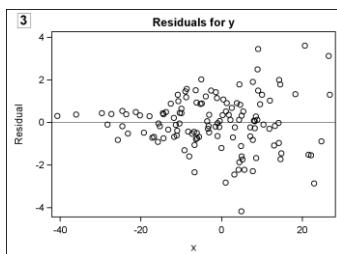
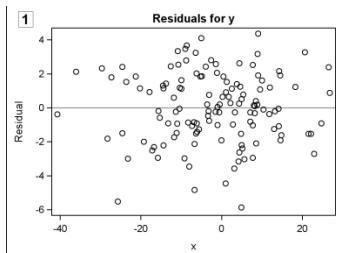
Examining Residual Plots



- Curvilinear pattern detected in residuals.
- Model form is incorrect.
- Possible remedies, depending on pattern, include polynomial terms, interactions, splines, and so on.

10

Examining Residual Plots



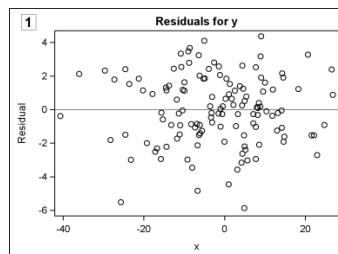
- Constant variance assumption is violated.
- Possible remedy is transforming variables to stabilize the variance.
- Procedures that model the non-constant variance can be used. (GENMOD, GLIMMIX)

11



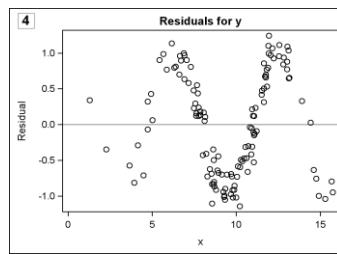
Copyright © SAS Institute Inc. All rights reserved.

Examining Residual Plots



- Remedy is to analyze using PROC AUTOREG.

- Observations not independent.
- Residuals follow cyclic pattern.
- Most evident when collected over time.

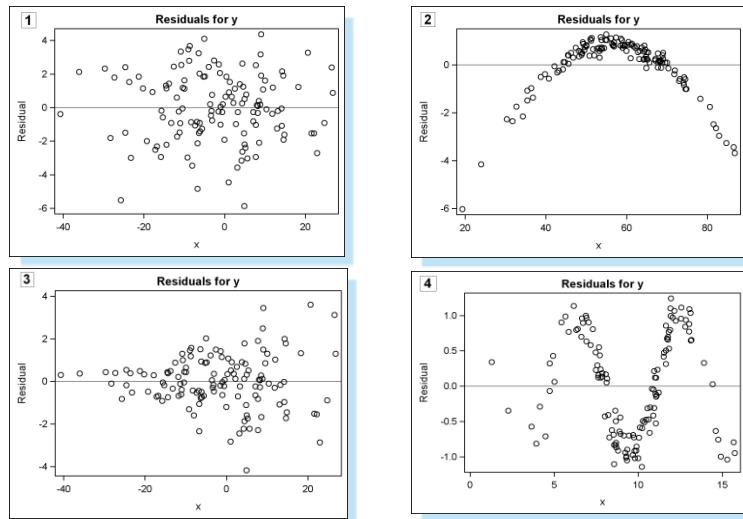


12



Copyright © SAS Institute Inc. All rights reserved.

Examining Residual Plots

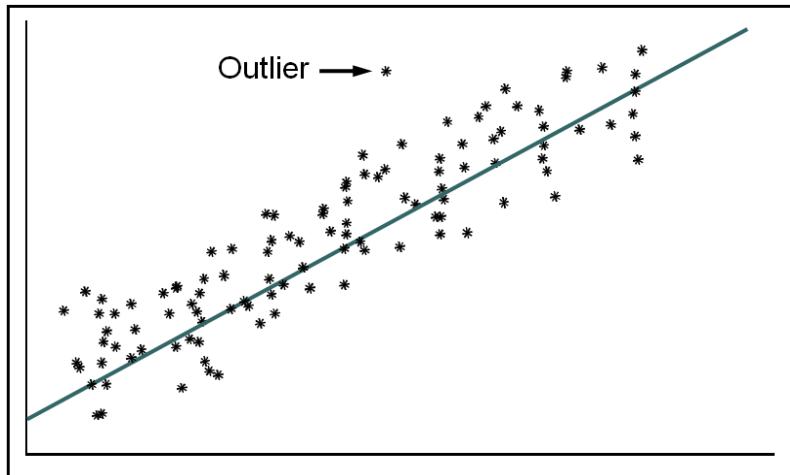


13

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. Another possible solution is to use either PROC GENMOD or PROC GLIMMIX, and choose a model that does not assume equal variances.
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 were collected over time. A possible solution is to use the AUTOREG procedure in SAS/ETS software.

Detecting Outliers



14

Copyright © SAS Institute Inc. All rights reserved.



Besides verifying assumptions, it is also important to check for outliers. Observations that are far away from the bulk of your data are outliers. These observations are often data errors or reflect unusual circumstances. In either case, it is good statistical practice to detect these outliers and find out why they occurred.



Residual Plots

Example: Invoke the REG procedure noticing the default graphics. Then use a PLOTS= option to produce full-sized ODS residual plots and diagnostic plots for the model including all interval predictor variables.

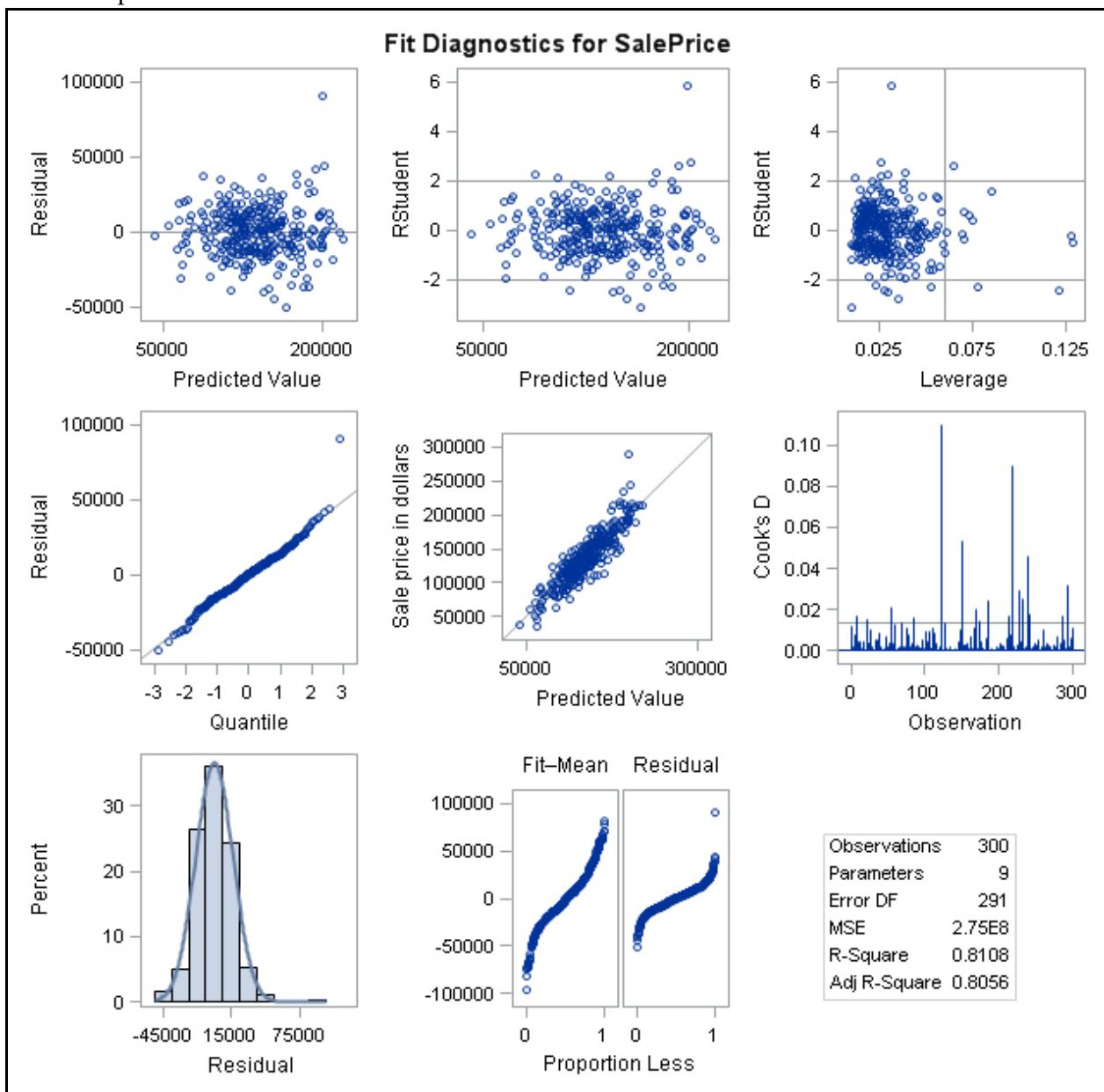
1. Open the **Linear Regression** task under **Statistics**.
2. Select the **Ameshousing3** data set and assign **SalePrice** as the dependent variable.
3. Assign **Gr_Liv_Area**, **Basement_Area**, **Garage_Area**, **Deck_Porch_Area**, **Lot_Area**, **Age_Sold**, **Bedroom_AbvGr**, and **Total_Bathroom** as the continuous variables.
4. On the MODEL tab, open the Model Effects Builder and include all the variables in the model.
5. On the OPTIONS tab, expand the **Scatter Plots** property and uncheck the option to display a scatter plot of **observed values by predicted values**
6. Run the code.

Note: Alternatively, you can write the code directly in SAS.

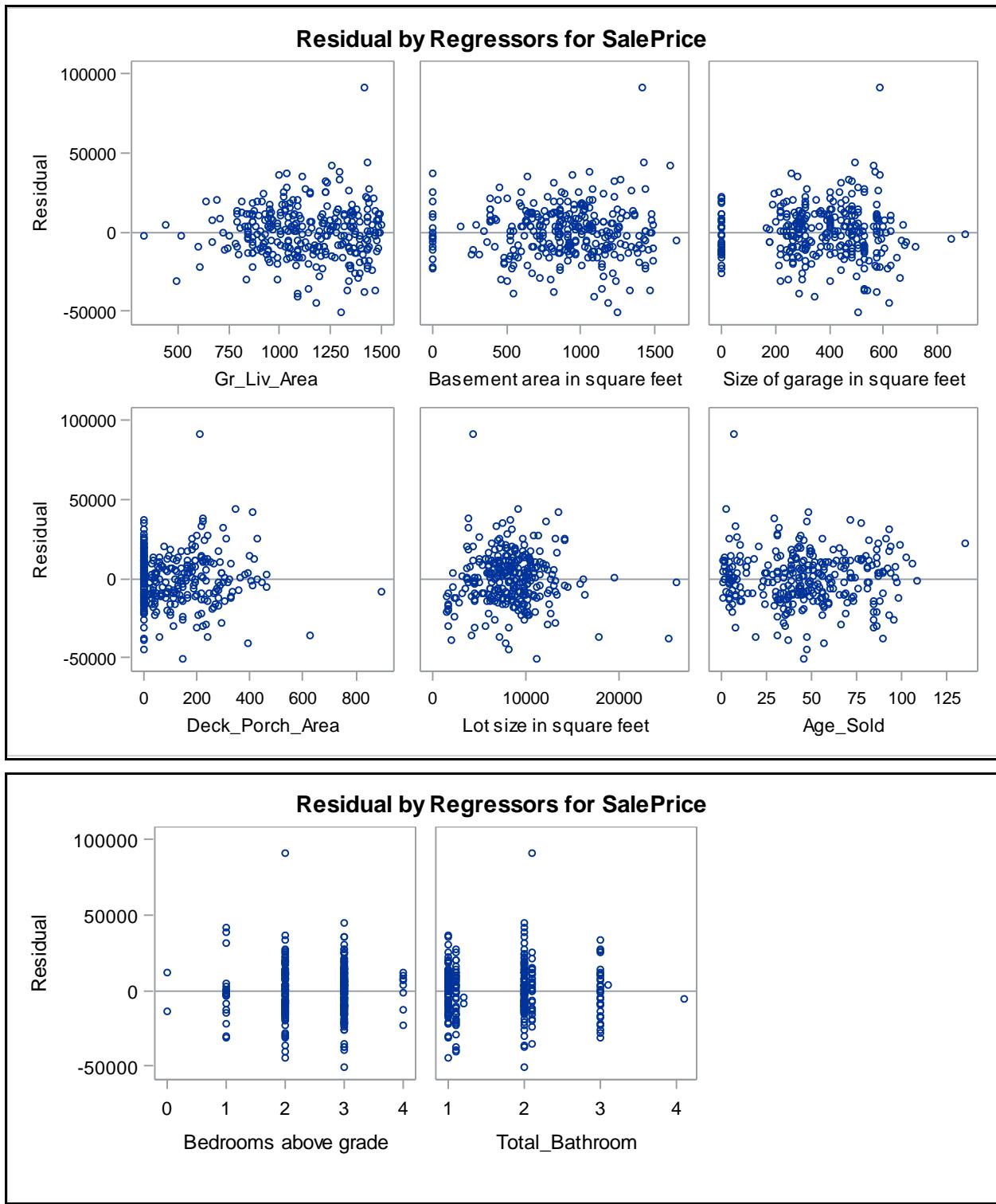
```
%let interval=Gr Liv Area Basement Area Garage Area Deck Porch_Area
          Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d01.sas*/ /*Part A*/
ods graphics on;
proc reg data=STAT1.ameshousing3;
  CONTINUOUS: model SalePrice = &interval;
  title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;
```

Partial Output



Residual and diagnostic plots are produced in the DIAGNOSTICS panel plot. (Several of these are discussed in more detail later in the chapter.)



The plot of the residuals versus the values of the interval predictor variables is shown above. They show 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 earlier).

Note: When visually inspecting residual plots, the distinction of whether a pattern exists is a matter of discretion for the viewer. If there is any question to the presence of a pattern, a further investigation for possible causes of potential patterns should be performed.

Hint: If you want to view the diagnostic plots separately using SAS Studio, use the drop-down menu under Diagnostic plots to Display as: **Individual plots**. Alternatively, you can edit the code and specify PLOTS=DIAGNOSTICS(UNPACK) in the PROC REG statement.

Note: SAS Studio does not offer all available procedure plots options using tasks. If you would like to display other plots, you must use the Program Editor and specify each plot using the appropriate plot option. The code below produces the Quantile - Quantile plot; the residual versus predicted values plot; and the residual versus regressor values plot. Individual plots are produced full sized.

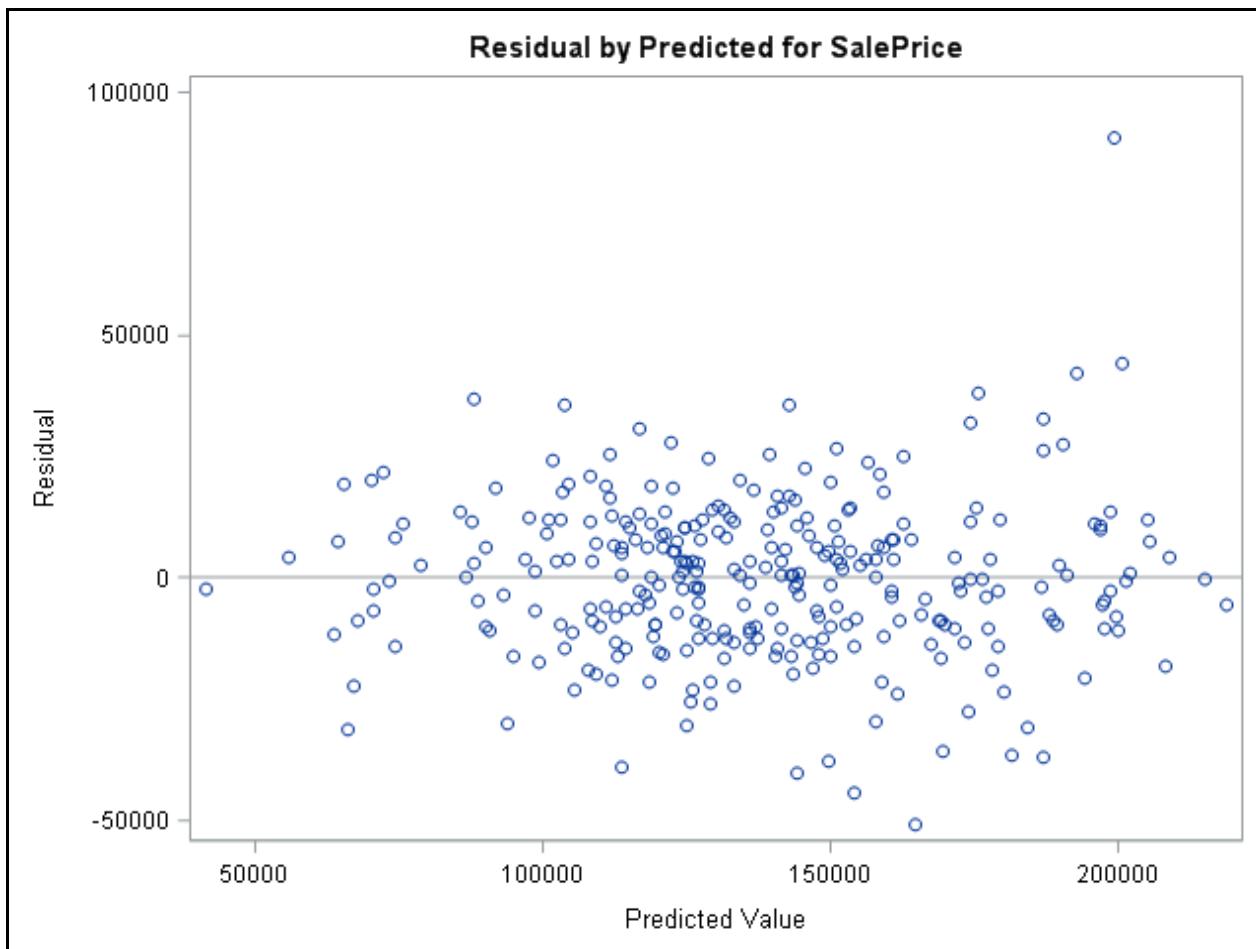
```
/*st105d01.sas*/ /*Part B*/
proc reg data=STAT1.ameshousing3
  plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);
  CONTINUOUS: model SalePrice = &interval;
  title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;
```

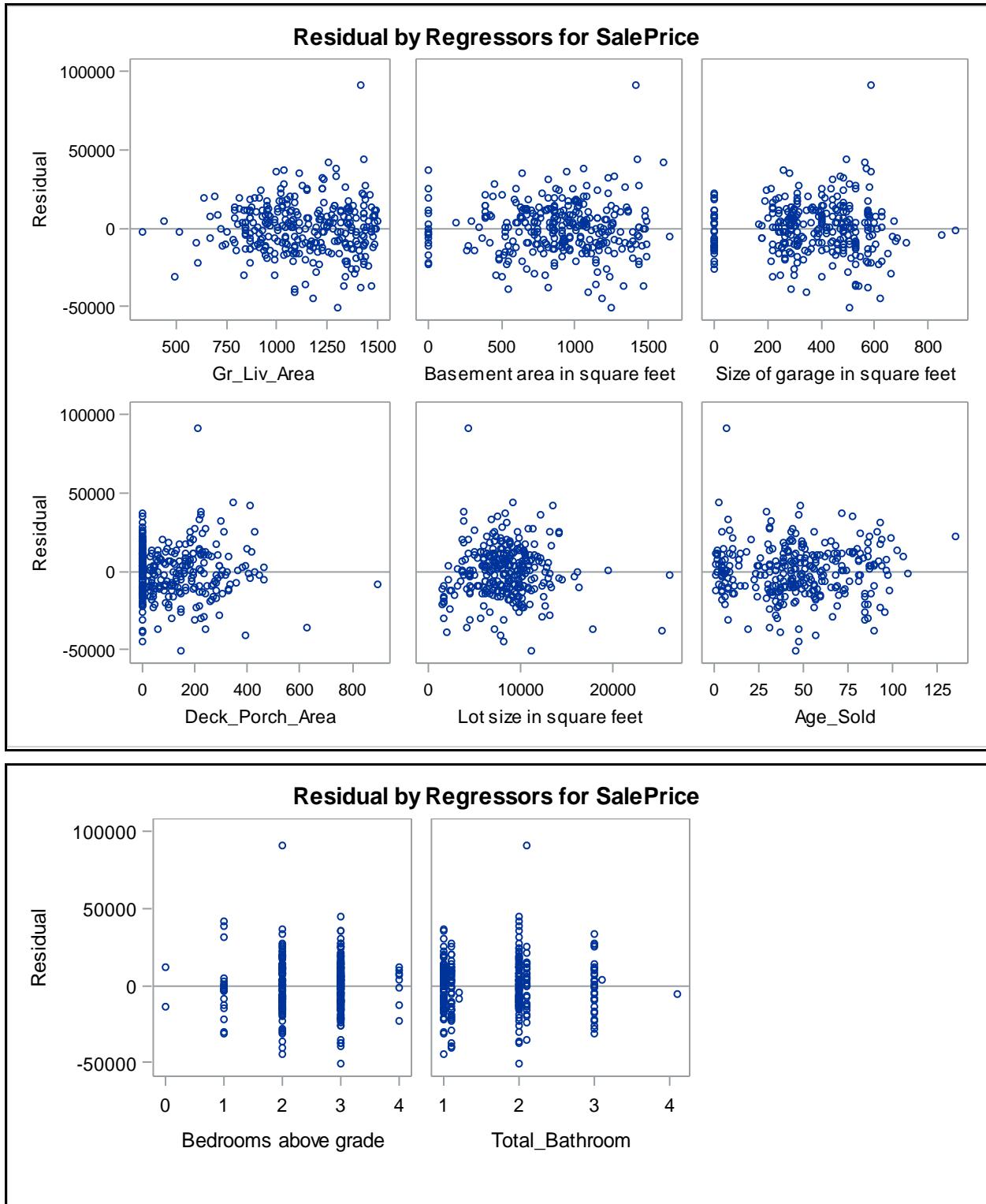
Selected REG statement PLOTS= options:

PLOTS(ONLY)=	produces only the plots listed and suppresses printing of default plots.
QQ	produces residual Quantile-Quantile plot to assess the normality of the residual error.
RESIDUALBYPREDICTED	produces residuals by predicted values.
RESIDUALS	produces residuals by predictor variable values.

Note: 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 plots of the residuals by predicted values of **SalePrice** and by each of the predictor variables are 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.

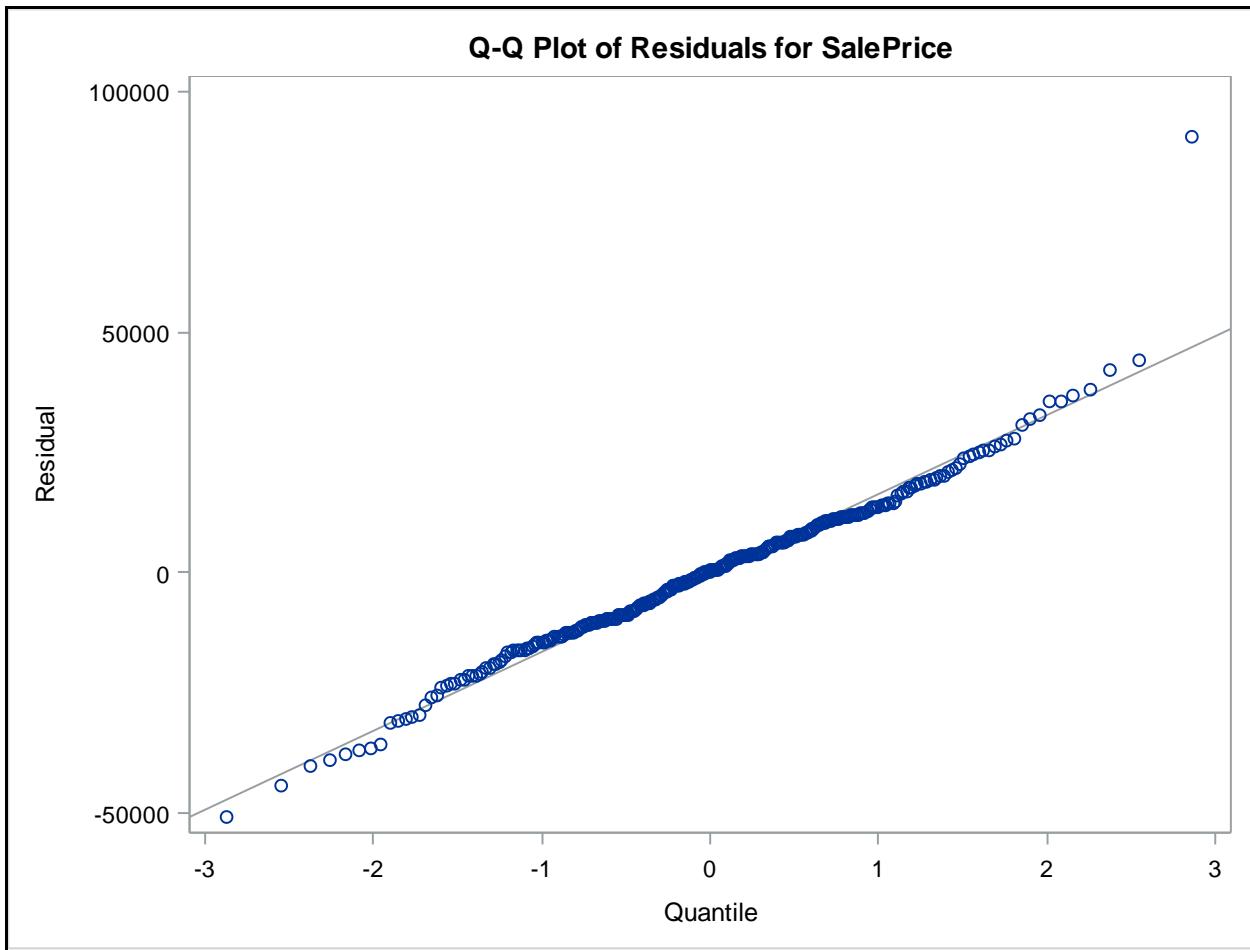




The plot of the residuals against the normal quantiles is shown below. If the residuals are normally distributed, the plot should appear to be a straight, diagonal line. If the plot deviates substantially from the reference line, then there is evidence against normality.

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)



End of Demonstration



Exercises

1. Examining Residuals

Assess the model obtained from the final forward stepwise selection of predictors for the **STAT1.BodyFat2** data set. Run a regression of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**. Create plots of the residuals by the four regressors 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 by the evidence in any of the residual plots?
- c. Does the Quantile-Quantile plot indicate any problems with the normality assumption?

End of Exercises

5.2 Influential Observations

Objectives

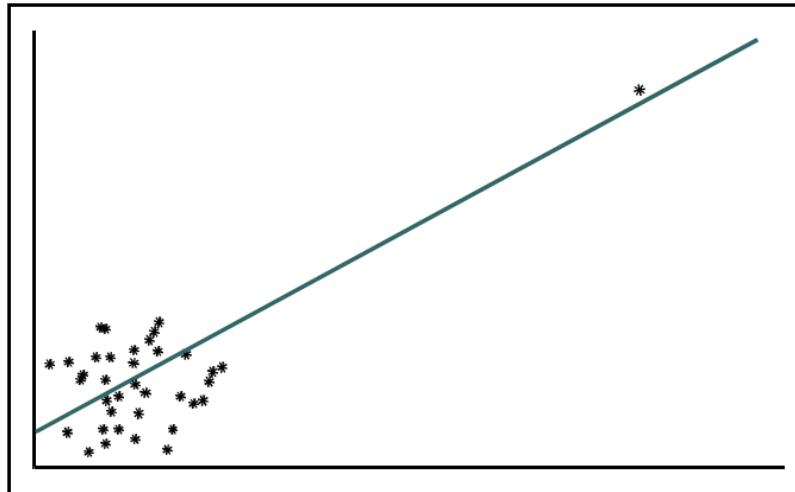
- Use statistics to identify potentially influential observations.

19

Copyright © SAS Institute Inc. All rights reserved.



Influential Observations



20

Copyright © SAS Institute Inc. All rights reserved.



Recall in the previous section that you saw examples of data sets where the simple linear regression model fits were essentially the same. However, plotting the data revealed that the model fits were different.

One of the examples showed a highly influential observation similar to the example above.

Identifying influential observations in multiple linear regression is more complex because you have more predictors to consider.

The REG procedure has options to calculate statistics to identify influential observations.

Diagnostic Statistics

Statistics that help identify influential observations are the following:

- Studentized residuals
- RSTUDENT residuals
- Cook's D
- DFFITS
- DFBETAS

The R option in the MODEL statement prints the studentized residuals and the Cook's D, as well as others discussed previously. The INFLUENCE option in the MODEL statement prints the RSTUDENT, DFFITS, and DFBETAS, as well as several others.

Studentized (Standardized) Residuals

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

One way to check for outliers is to use the studentized residuals. These are calculated by dividing the residual values by their standard errors. For a model that fits the data well and has no outliers, you can expect that 68% of the studentized residuals would be within [-1,1]. In general, studentized residuals that have an absolute value less than 2.0 could easily occur by chance. Studentized residuals that are between an absolute value of 2.0 to 3.0 occur infrequently and could be outliers. Studentized residuals that are larger than an absolute value of 3.0 occur rarely by chance alone and should be investigated.

Note: Studentized residuals are often referred to as “standardized residuals.” The cutoff values are chosen based on the tail probabilities from the normal probability distribution.

5.02 Multiple Choice Poll

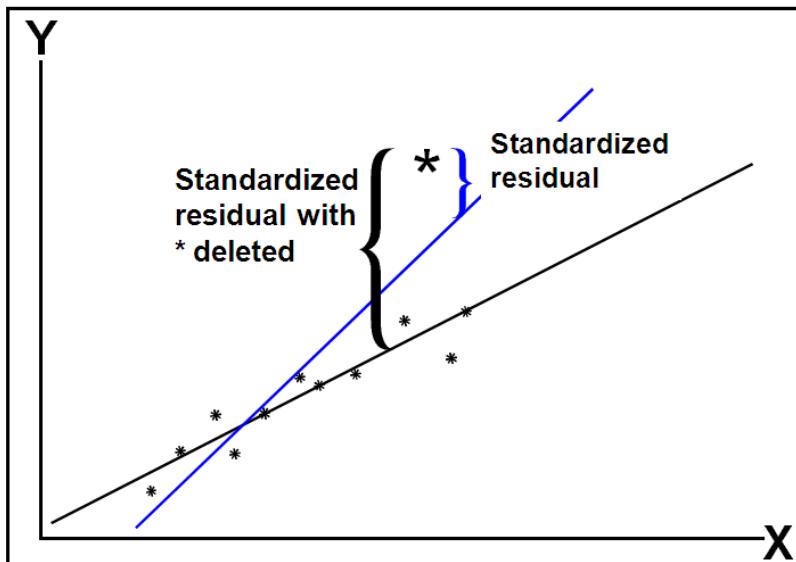
Given the properties of the standard normal distribution, you would expect about 95% of the studentized residuals to be between which two values?

- a. -3 and 3
- b. -2 and 2
- c. -1 and 1
- d. 0 and 1
- e. 0 and 2
- f. 0 and 3

23

Copyright © SAS Institute Inc. All rights reserved.

RSTUDENT



25

Copyright © SAS Institute Inc. All rights reserved.

Studentized residuals are the ordinary residuals divided by their standard errors. The RSTUDENT residuals are similar to the studentized residuals except that they are calculated after deleting the i^{th} observation. In other words, the RSTUDENT residual is the difference between the observed Y and the predicted value of Y excluding this observation from the regression.

Note: There is a difference between the labels used in SAS and in SAS Enterprise Guide.

SAS		SAS Enterprise Guide
Studentized residuals	⇒	Standardized residuals
RSTUDENT residuals (studentized residual with the i^{th} observation removed)	⇒	Studentized residuals

Cook's D Statistic

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when the i^{th} observation is deleted from the analysis.

A suggested cutpoint for influence is shown below:

$$\text{Cook's } D_i > \frac{4}{n}$$

To detect influential observations, you can use Cook's D statistic. This statistic measures the change in the parameter estimates that results from deleting each observation.

$$\text{Cook's } D_i = \left(\frac{1}{ps^2} \right) (\mathbf{b} - \mathbf{b}_{(i)})' (\mathbf{X}'\mathbf{X}) (\mathbf{b} - \mathbf{b}_{(i)})$$

p the number of regression parameters

s^2 mean squared error of the regression model

\mathbf{b} the vector of parameter estimates

$\mathbf{b}_{(i)}$ the vector of parameter estimates obtained after deleting the i^{th} observation

$\mathbf{X}'\mathbf{X}$ corrected sum of squares and cross-products matrix

Identify observations above the cutoff and investigate the reasons that they occurred.

DFFITS

DFFITS_i measures the impact that the i^{th} observation has on the predicted value.

A suggested cutoff for influence is shown below:

$$|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$$

27

Copyright © SAS Institute Inc. All rights reserved.



$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$$

\hat{Y}_i the i^{th} predicted value

$\hat{Y}_{(i)}$ the i^{th} predicted value when the i^{th} observation is deleted

$s(\hat{Y}_i)$ the standard error of the i^{th} predicted value

Belsey, Kuh, and Welsch (1980) provide this suggested cutoff: $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$, where p is the number of terms in the current model, including the intercept, and n is the sample size.

DFBETAS

- Measure of change in the j^{th} parameter estimate with deletion of the i^{th} observation
- One DFBETA per parameter per observation
- Helpful in explaining on which parameter coefficient the influence most lies
- A suggested cutoff for influence is shown below:

$$|\text{DFBETA}_{ij}| > 2\sqrt{\frac{1}{n}}$$

DFBETAS is abbreviated from Difference in Betas. They contain the standardized difference for each individual coefficient estimate resulting from the omission of the i^{th} observation. They are identified by column headings with the name of the corresponding predictor in the Output window and also by plots, if requested in the PROC REG statement. Because there are many DFBETAS, it might be useful to examine only those corresponding to a large Cook's D . Large DFBETAS indicate which predictor(s) might be the cause of the influence.

$$\text{DFBETA}_{ij} = \frac{b_j - b_{(i)j}}{s(b_j)}$$

b_j j^{th} regression parameter estimate

$b_{(i)j}$ j^{th} regression parameter estimate with observation i deleted

$s(b_j)$ standard error of b_j

Belsley, Kuh, and Welsch (1980) recommend 2 as a general cutoff value to indicate influential

observations and $2\sqrt{\frac{1}{n}}$ as a size-adjusted cutoff.



Looking for Influential Observations

Example: Using GLMSELECT, obtain the model selected using stepwise selection. Request SELECT=SL to ensure that the stepwise selection is using significance level as its criterion. Set the entry and stay criteria to be 0.05. Recall that the selected effects are stored in the macro variable &_GLSIND. Using REG, generate the RSTUDENT, DFFITS, DFBETAS, and Cook's D influence statistics and plots for the selected model. Save the statistics to an output data set and create a data set with only observations that exceed the suggested cutoffs of the influence statistics.

1. Open the **Linear Regression** task under **Statistics**.
2. Select the **Ameshousing3** data set, assign **SalePrice** as the dependent variable.
3. Assign the variables **Gr_Liv_Area**, **Basement_Area**, **Garage_Area**, **Deck_Porch_Area**, **Lot_Area**, **Age_Sold**, **Bedroom_AbvGr**, and **Total_Bathroom** as the continuous variable.
4. On the MODEL tab, specify the model by selecting all the variables and adding in model effects
5. On the OPTIONS tab, expand the **More Diagnostic Plots** property and check all the fields to display diagnostic plots and labels for influential observations. Uncheck boxes for Diagnostic plots, Residuals for explanatory variables, and Observed values by predicted values (after expanding Scatter Plots).
6. On the SELECTION tab, choose the **Stepwise selection** option and specify to use **Significance level** to add/remove effect.
7. Click **Edit** above the code to open a copy of the code for editing.
8. Type **cooksd** within the parentheses where plots are listed. Add **COOKSDPlot** to the list in the **ODS SELECT** statement. Include an **ODS OUTPUT** statement to output the data from the influence plots into data sets, as follows:

```
proc reg data=Work.reg_design alpha=0.05 plots(only
label)=(rstudentbypredicted cooksd
      dffits dfbetas);
ods select OutputStatistics ResidualStatistics
RStudentByPredicted COOKSDPlot DFFITSPlot
      DFBETASPanel;
ods output RSTUDENTBYPREDICTED=Rstud
      COOKSDPLOT=Cook
      DFFITSPLOT=Dffits
      DFBETASPANEL=Dfbs;
model SalePrice=&_GLSMOD / r;
run;
quit;
```

Note: Alternatively, you can write the code directly in SAS.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
      Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;
/*st105d02.sas*/ /*Part A*/
ods select none;
proc glmselect data=STAT1.ameshousing3 plots=all;
```

```

STEPWISE: model SalePrice = &interval / selection=stepwise
            details=steps select=SL slentry=0.05 slstay=0.05;
title "Stepwise Model Selection for SalePrice - SL 0.05";
run;
quit;
ods select all;

ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
    COOKSDPLOT=Cook
    DFFITSPLOT=Dffits
    DFBETASPANEL=Dfbs;
proc reg data=STAT1.ameshousing3
plots(only label)=
    (RSTUDENTBYPREDICTED
    COOKSD
    DFFITS
    DFBETAS);
MODEL1: model SalePrice = & GLSIND;
title 'MODEL1 Model - Plots of Diagnostic Statistics';
run;
quit;

```

Selected REG procedure PLOTS= options:

PLOTS(LABEL)=	labels extreme observations in the plot with either the observation number or the value of an ID variable, if there is an ID statement.
RSTUDENTBYPREDICTED	RStudent by predicted values.
COOKSD	Cook's <i>D</i> plot.
DFFITS	DFFITS plot.
DFBETAS	DFBETAS plots.

The ODS OUTPUT statement along with the PLOTS= option outputs the data from the influence plots into separate data sets.

PROC REG Output

Number of Observations Read	300
Number of Observations Used	300

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	3.424508E11	48921543221	176.86	<.0001
Error	292	80772716963	276618894		
Corrected Total	299	4.232235E11			

The following table is output from SAS Studio.

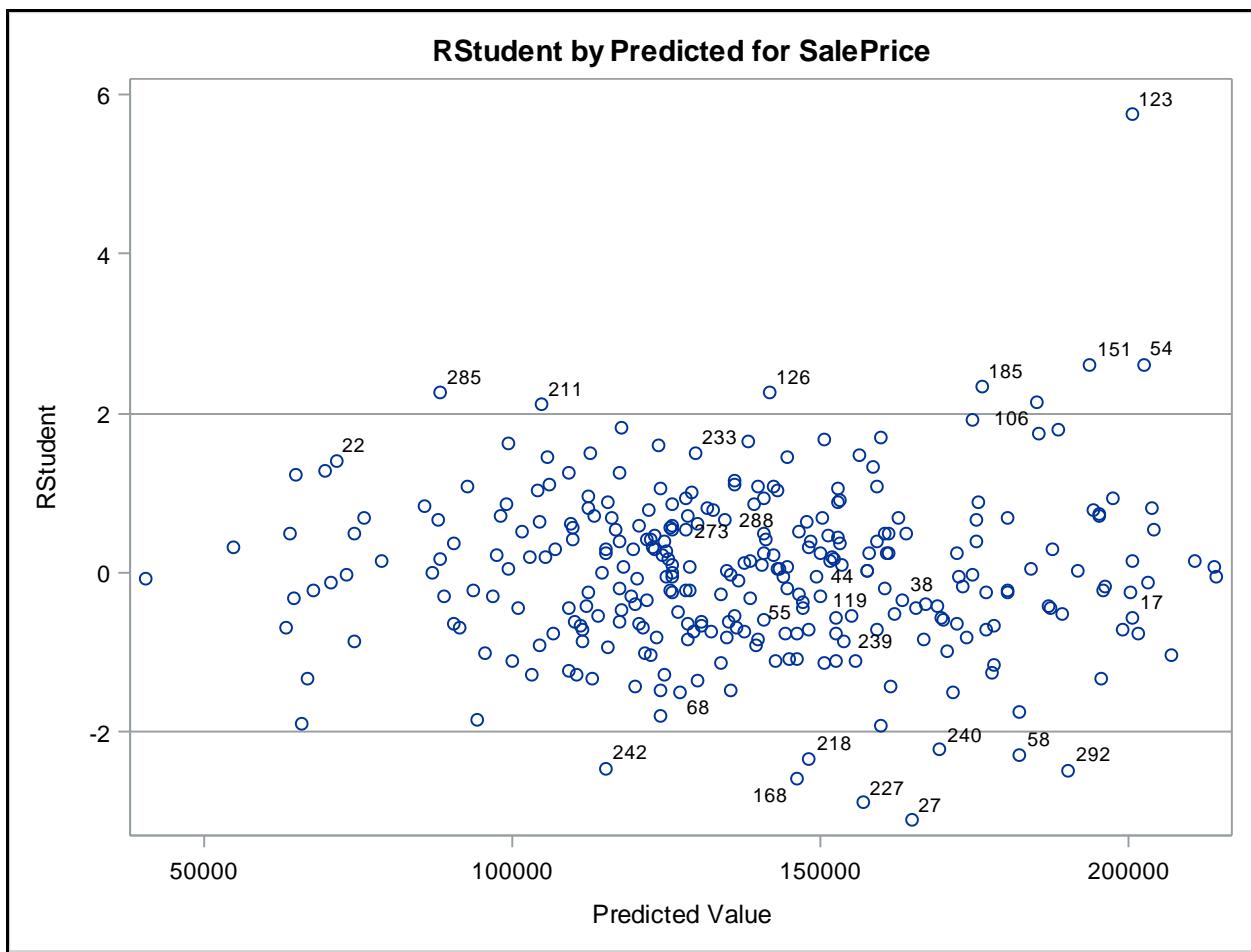
Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

The following table is output from Foundation SAS:

Root MSE	16632	R-Square	0.8091
Dependent Mean	137525	Adj R-Sq	0.8046
Coeff Var	12.09371		

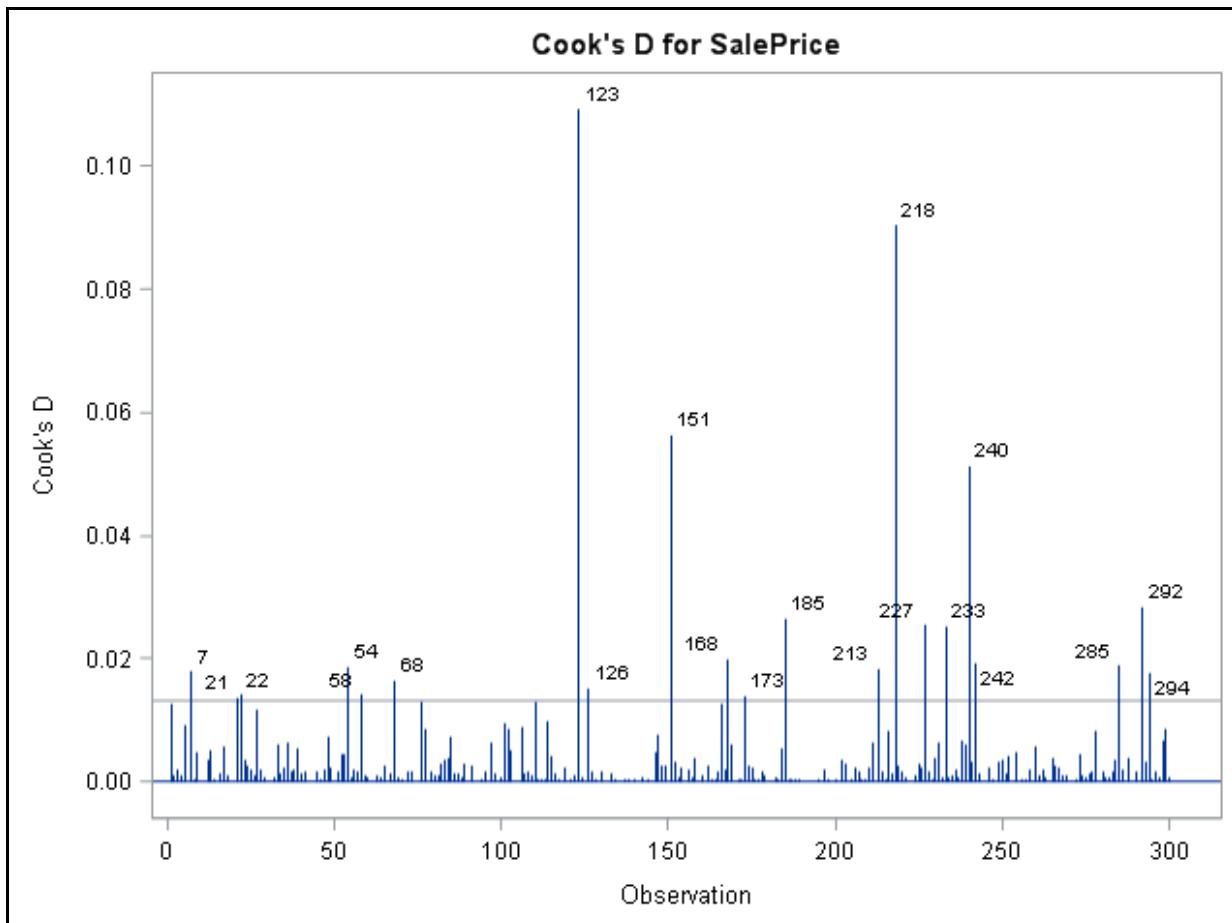
Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t 	
Intercept	Intercept	1	47463	5880.67404	8.07	<.0001	
Gr_Liv_Area	Above grade (ground) living area square feet	1	65.30372	5.43667	12.01	<.0001	
Basement_Area	Basement area in square feet	1	29.84908	3.34540	8.92	<.0001	
Garage_Area	Size of garage in square feet	1	36.30961	6.45241	5.63	<.0001	
Deck_Porch_Area	Total area of decks and porches in square feet	1	32.05255	7.96768	4.02	<.0001	
Lot_Area	Lot size in square feet	1	0.70813	0.31751	2.23	0.0265	
Age_Sold	Age of house when sold, in years	1	-447.19868	41.01931	-10.90	<.0001	
Bedroom_AbvGr	Bedrooms above grade	1	-5042.76650	1687.92817	-2.99	0.0031	

Partial PROC REG Output (Continued)

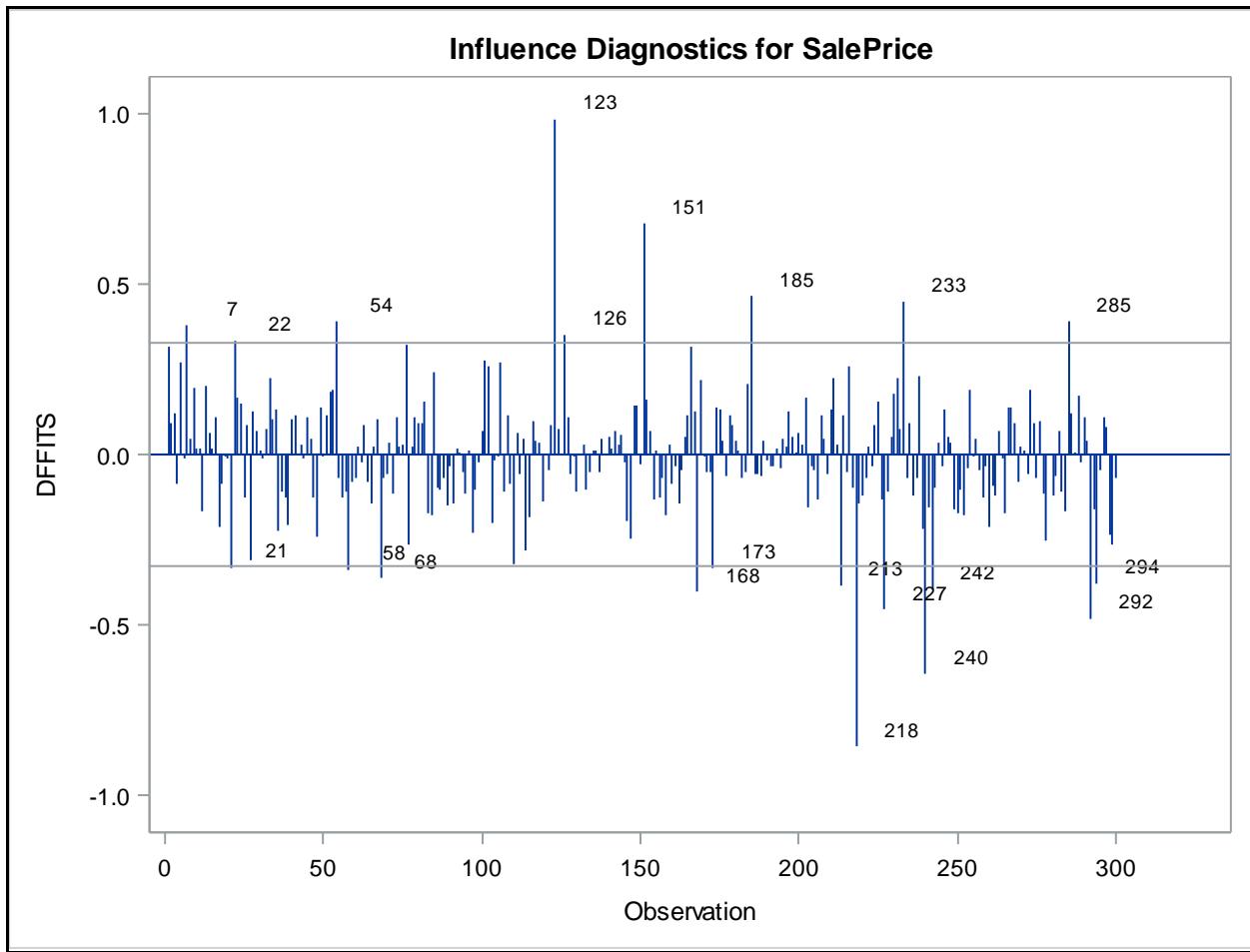


The RStudent plot shows sixteen observations beyond two standard errors from the mean of 0. Those are identified with their observation numbers. Because you expect 5% of values to be beyond two standard errors from the mean (remember that RStudent residuals are assumed to be normally distributed), the fact that you have sixteen that far outside the primary cluster gives no cause for concern. (Five percent of 300 is 15 expected observations.)

Note: Other observations are also labeled in this plot for other reasons such as high leverage.

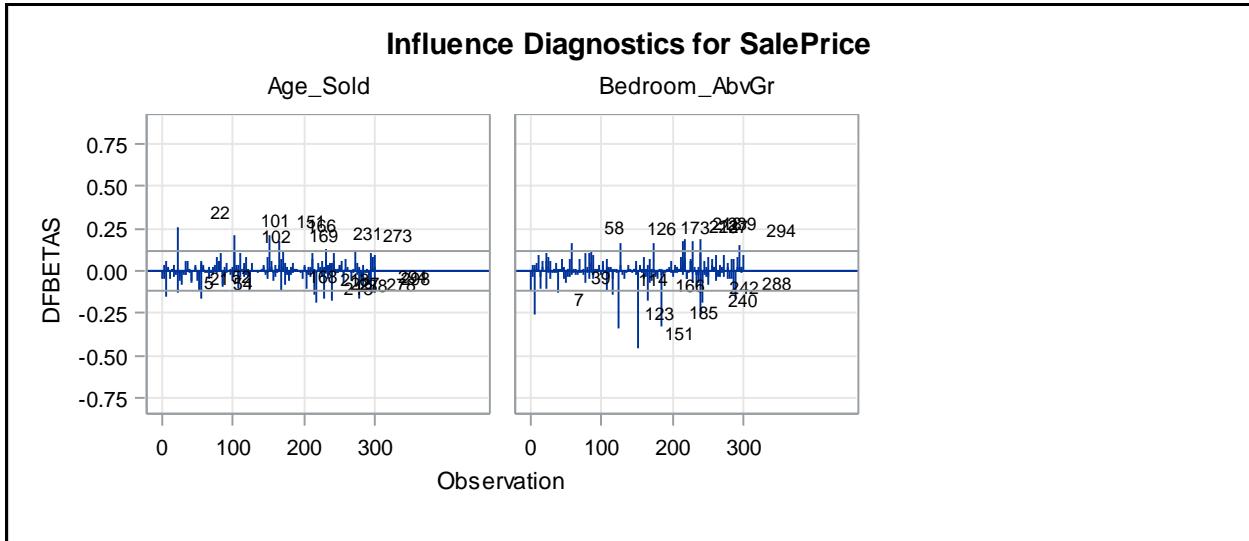
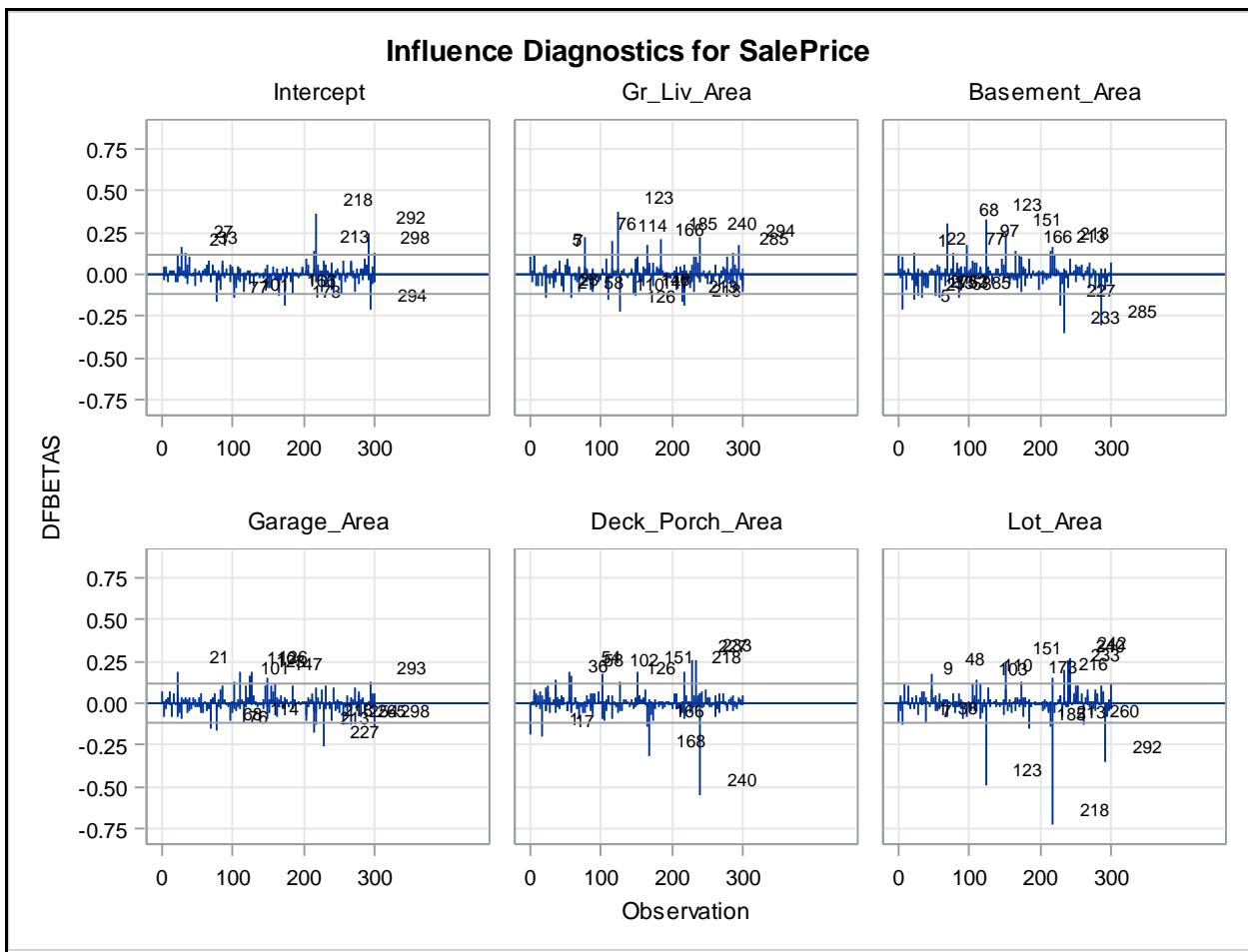


The Cook's D plot shows 21 points to be influential.



Once again, several observations have been flagged as an influential point based on DFFITS.

At this point, it might be helpful to see which parameters these observations might influence most. DFBETAS provides that information.



Detection of outliers or influential observations with plots is convenient for relatively small data sets, but for larger data sets, like the housing data, it can be very difficult to discern one observation from another. One method for extracting only the influential observations from a data set is to output the ODS plots data into data sets and then subset the influential observations.

The next part of the program prints the influential observations in the influence diagnostic data sets that were produced using ODS OUTPUT.

```
/*st105d02.sas*/ /*Part B*/
title;
proc print data=Rstud;
run;
```

Partial Output

Obs	Model	Dependent	RStudent	PredictedValue	outLevLabel	Observation
1	MODEL1	SalePrice	1.07416	13977.490	.	1
2	MODEL1	SalePrice	-1.26343	177720.93	.	2
3	MODEL1	SalePrice	0.48988	160512.06	.	3
4	MODEL1	SalePrice	0.69574	150148.52	.	4

The variable **outLevLabel** is nonmissing only for an observation labeled for any reason on the RStudent plot.

```
proc print data=Cook;
run;
```

Partial Output

Obs	Model	Dependent	CooksD	Observation	CooksDLabel
1	MODEL1	SalePrice	0.00221	1	.
2	MODEL1	SalePrice	0.00524	2	.
3	MODEL1	SalePrice	0.00130	3	.
4	MODEL1	SalePrice	0.00248	4	.

The variable **CooksDLabel** identifies observations that are deemed influential due to high Cook's *D* values; these are observations having influence on all the estimated parameters as a group.

```
proc print data=Dffits;
run;
```

Partial Output

Obs	Model	Dependent	Observation	DFFITS	DFFITSOUT
1	MODEL1	SalePrice	1	0.13306	.
2	MODEL1	SalePrice	2	-0.20493	.
3	MODEL1	SalePrice	3	0.10169	.
4	MODEL1	SalePrice	4	0.14082	.

The variable **DFFITSOUT** identifies observations that are deemed influential due to high DFFITS values; these are observations having influence on the predictions.

```
proc print data=Dfbs;
run;
```

Partial Output

Obs	Model	Dependent	Observation	_DFBETAS1	_DFBETASOUT1	_DFBETAS2	_DFBETASOUT2
1	MODEL1	SalePrice	1	0.03295	.	-0.02759	.
2	MODEL1	SalePrice	2	0.10779	.	0.01382	.
3	MODEL1	SalePrice	3	0.00131	.	0.05138	.
4	MODEL1	SalePrice	4	0.01183	.	0.06510	.

The variables **DFBETASOUT1** through **DFBETASOUT8** identify the observations whose DFBETA values exceed the threshold for influence. **DFBETASOUT1** represents the value for the intercept. The other seven variables show influential outliers on each of the predictor variables in the MODEL statement in PROC REG.

Note: As the number of predictor variables increases, additional panels are required to show all the information from DFBETAS. This will need to be handled prior to merging data sets. With the multiple panels for DFBETAS, the **DFBS** data set is effectively split. The first 300 observations display the DFBETAS information for the first panel, which includes the first six effects in the model (including the intercept). The information for the second panel, which includes the final two effects, is missing. Beginning at observation 301, this is reversed. The following code block splits the **DFBS** data set into two parts and combines them into one new data set (**DFBS2**) using the UPDATE statement.

```
data Dfbs01;
  set Dfbs (obs=300);
run;

data Dfbs02;
  set Dfbs (firstobs=301);
run;

data Dfbs2;
  update Dfbs01 Dfbs02;
  by Observation;
run;
```

The next DATA step merges the four data sets containing the influence data and outputs only the observations that exceeded the respective influence cutoff levels. The cutoff for RStudent has been augmented to 3 and -3.

The results are then displayed.

```
data influential;
/* Merge datasets from above.*/
merge Rstud
      Cook
      Dffits
      Dfbs2;
by observation;

/* Flag observations that have exceeded at least one cutpoint;*/
```

```

if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then
  flag=1;
array dfbetas{*} _dfbetasout: ;
do i=2 to dim(dfbetas);
  if dfbetas{i} then flag=1;
end;

/* Set to missing values of influence statistics for those*/
/* that have not exceeded cutpoints;*/
if ABS(Rstudent)<=3 then RStudent=.;
if Cooksdlabel eq ' ' then CooksD=.;

/* Subset only observations that have been flagged.*/
if flag=1;
drop i flag;
run;

title;
proc print data=influential;
  id observation;
  var Rstudent CooksD Dffitsout _dfbetasout: ;
run;

```

PROC PRINT Output

Observation	RStudent	CooksD	DFFITSOUT	_DFBETASOUT1	_DFBETASOUT2	_DFBETASOUT3
1	0.11744
5	0.12008	-0.21199
7	.	0.01782	0.37928	.	0.11635	.
9

Observation	_DFBETASOUT4	_DFBETASOUT5	_DFBETASOUT6	_DFBETASOUT7	_DFBETASOUT8
1	.	-0.18180	-0.12067	.	.
5	.	.	.	-0.15614	.
7	.	.	-0.13126	.	-0.25391
9	.	.	0.12030	.	.

This table is a summary of the plots displayed previously. From this output, flagged observations can then be investigated to try to determine what makes these points influential. As always, this would be after determining that the point was valid and not erroneous data.

End of Demonstration

How to Handle Influential Observations

Recheck the data to ensure that no transcription or data entry errors occurred.

If the data is valid, one possible explanation is that the model is not adequate.

Determine the robustness of the inference by running the analysis both with and without the influential observations.

Note: A model with higher-order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.

30

Copyright © SAS Institute Inc. All rights reserved.



If the unusual data are erroneous, correct the errors and reanalyze the data.

(In this course, time does not permit discussion of higher order models in any depth. This discussion is in Statistics 2:ANOVA and Regression.)

Another possibility is that the observation, although valid, could be unusual. If you had a larger sample size, there might be more observations similar to the unusual ones.

You might have to collect more data to confirm the relationship suggested by the influential observation.

In general, do not exclude data. In many circumstances, some of the unusual observations contain important information.

If you do choose to exclude some observations, include a description of the types of observations that you exclude and provide an explanation. Also discuss the limitation of your conclusions, given the exclusions, as part of your report or presentation.



Exercises

2. Generating Potential Outliers

Using the **STAT1.BodyFat2** data set, run a regression model of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**.

- a. Use plots to identify potential influential observations based on the suggested cutoff values.
- b. Output residuals to a data set, subset the data set by only those who are potentially influential outliers, and print the results.

End of Exercises

5.03 Multiple Choice Poll

How many observations did you find that might substantially influence parameter estimates as a group?

- a. 0
- b. 1
- c. 4
- d. 5
- e. 7
- f. 10

33



5.3 Collinearity

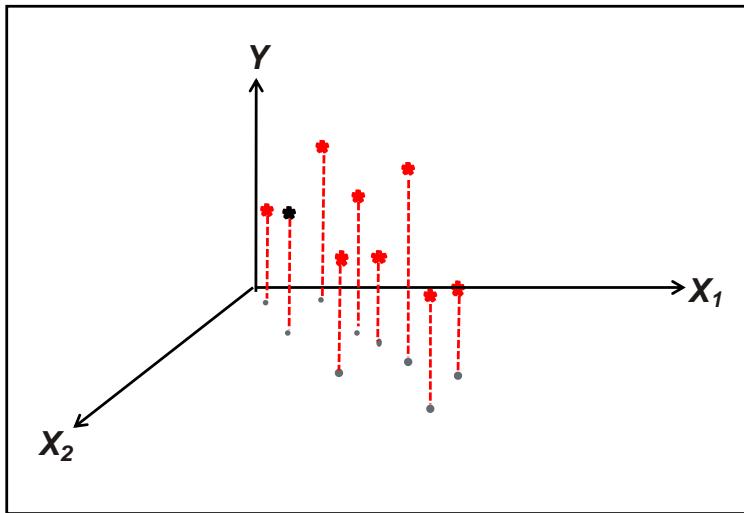
Objectives

- Determine whether collinearity exists in a model.
- Generate output to evaluate the strength of the collinearity and what variables are involved in the collinearity.
- Determine methods that can minimize collinearity in a model.

36



Illustration of Collinearity



37

Copyright © SAS Institute Inc. All rights reserved.

The goal of multiple linear regression is to find the best fit plane through the data to predict the response variable. Here is an example in three dimensions, two predictor variables, and a response variable. You can picture that the prediction plane that you are trying to build is similar to a tabletop, where the observations guide the angle of the tabletop, relative to the floor, in the same way as the legs for the table. If the legs line up with one another, then the plane built on top of it tends to be unstable.

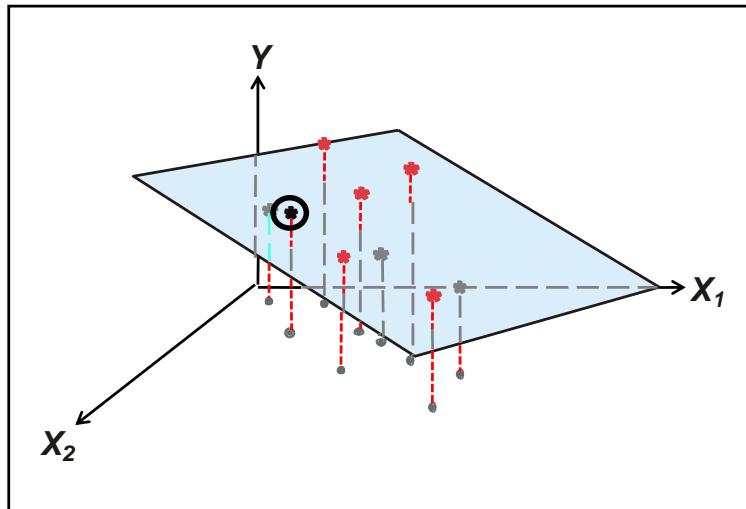
Where should the prediction plane be placed? The slopes of the prediction plane relative to each X and the Y are the parameter coefficient estimates.

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

Why is this a problem? Two reasons exist.

1. Neither might appear to be significant when both are in the model. However, either might be significant when only one is in the model. Thus, collinearity can hide significant effects. (The reverse can be true as well. Collinearity can increase the apparent statistical significance of effects.)
2. Collinearity tends to increase the variance of parameter estimates and consequently increase prediction error.

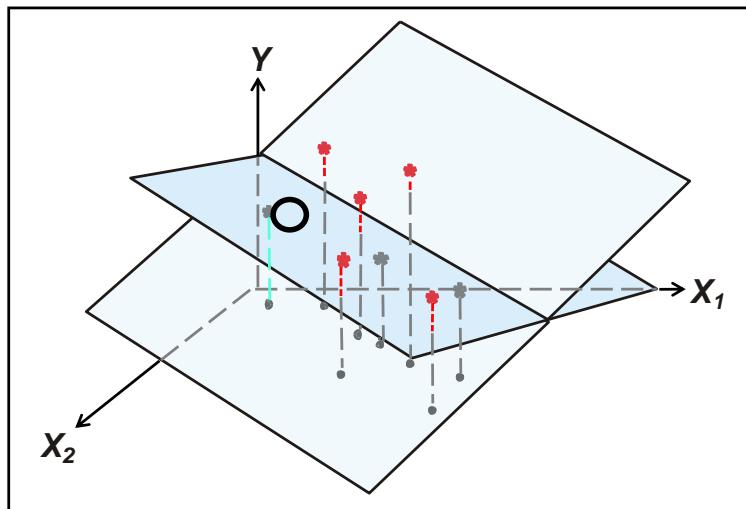
Illustration of Collinearity



38

This is a representation of a best-fit plane through the data.

Illustration of Collinearity



39

However, the removal of only one data point (or only moving the data point) results in a very different prediction plane (as represented by the lighter plane). This illustrates the variability of the parameter estimates when there is extreme collinearity.

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

Collinearity is ***not*** a violation of the assumptions of linear regression.

Collinearity Diagnostics

PROC REG offers these tools that help quantify the magnitude of the collinearity problems and identify the subset of Xs that is collinear:

- VIF
- COLLIN
- COLLINOINT

This course focuses on VIF.

Selected MODEL statement options:

VIF provides a measure of the magnitude of the collinearity (Variance Inflation Factor).

COLLIN includes the intercept vector when analyzing the $X'X$ matrix for collinearity.

COLLINOINT excludes the intercept vector when analyzing the $X'X$ matrix for collinearity.

Two options, COLLIN and COLLINOINT, also provide a measure of the magnitude of the problem as well as give information that can be used to identify the sets of Xs that are the source of the problem.

(COLLIN and COLLINOINT diagnostics are described in Statistics 2: ANOVA and Regression.)

Variance Inflation Factor (VIF)

The *VIF* is a relative measure of the increase in the variance because of collinearity. It can be thought of as this ratio:

$$VIF_i = \frac{1}{1 - R_i^2}$$

A $VIF_i > 10$ indicates that collinearity is a problem.

41

Copyright © SAS Institute Inc. All rights reserved.



You can calculate a VIF for each term in the model.

Marquardt (1990) suggests that a $VIF > 10$ indicates the presence of strong collinearity in the model.

$VIF_i = 1/(1 - R_i^2)$, where R_i^2 is the R-square of X_i , regressed on all the other Xs in the model.

For example, consider the model $Y=X1\ X2\ X3\ X4$.

To calculate the R-square for $X3$, referred to as R_3^2 , fit the model $X3=X1\ X2\ X4$. Take the R-square from the model with $X3$ as the dependent variable and replace it in the formula: $VIF_3=1/(1-R_3^2)$. If VIF_3 is greater than 10, $X3$ is possibly involved in collinearity.



Example of Collinearity

Example: Another research group has been working with the same Ames housing data but for different purposes. It has been brought to your attention that they found a variable useful in their analysis and it was merged onto the **STAT1.ameshousing3** data set. This new variable was called **score**. Using PROC CORR, investigate the correlations between the variable **score** and the other interval variables (using the macro variable **&interval**).

1. Open **st105d03.sas** and run the first two PROC SORTs and the first DATA step. This will combine the data from the other research group with the data we have been analyzing.

```
proc sort data=STAT1.ameshousing3;
  by PID;
run;
proc sort data=STAT1.amesaltuse;
  by PID;
run;

data amescombined;
  merge STAT1.ameshousing3 STAT1.amesaltuse;
  by PID;
run;
```

2. To investigate the correlations, open the **Correlation Analysis** task under **Statistics**.
3. Choose the **ameshousing3** data set and assign the interval variables, **Gr_Liv_Area**, **Basement_Area**, **Garage_Area**, **Deck_Porch_Area**, **Lot_Area**, **Age_Sold**, **Bedroom_AbvGr**, and **Total_Bathroom**, as the analysis variables and **score** as the correlate with variable.
4. Run the code.

Note: Alternatively, you can write the code directly in SAS.

```
title;
proc corr data=amescombined nosimple;
  var &interval;
  with score;
run;
```

Partial PROC CORR Output

Pearson Correlation Coefficients Prob > r under H0: Rho=0 Number of Observations							
	Gr_Liv_Area	Basement_Area	Garage_Area	Deck_Porch_Area	Lot_Area	Age_Sold	
score	-0.61394 <.0001 300	-0.97894 <.0001 300	-0.38872 <.0001 300	-0.35979 <.0001 300	-0.29249 <.0001 300	0.39125 <.0001 300	

Pearson Correlation Coefficients Prob > r under H0: Rho=0 Number of Observations		
	Bedroom_AbvGr	Total_Bathroom
score	-0.28357 <.0001 300	-0.51877 <.0001 300

The new variable **score** appears to be significantly correlated with all of the interval variables, but focus your attention on the actual correlations. Recall that closer to 1 or -1 implies a stronger correlation within the pairing. **Score** appears to be most correlated with **Basement_Area**, **Gr_Liv_Area**, and **Total_Bathroom**. These significant correlations are not enough to actually diagnose collinearity.

End of Demonstration



Collinearity Diagnostics

Example: Invoke PROC REG and use the VIF option to assess the magnitude of the collinearity problem and identify the terms involved in the problem. Use the predictor variables, **Gr_Liv_Area**, **Basement_Area**, **Garage_Area**, **Deck_Porch_Area**, **Lot_Area**, **Age_Sold**, **Bedroom_AbvGr**, **Total_Bathroom**, and **score**.

1. Open the **Linear Regression** task under **Statistics**.
2. Select the **Ameshousing3** data, assign **SalePrice** as the dependent variables, and the listed predictor variables as the continuous variables.
3. On the MODEL tab, open the Model Effects editor and select all the variables and add in the effects.
4. On the OPTIONS tab, use the drop-down menu for **Display statistics** and select **Default and selected statistics**.
5. Expand the **Collinearity** property and select the option to display **Variance inflation factors**.
6. Suppress all plots by unchecking the boxes under **Diagnostic and Residual Plots** and **Scatter Plots**.
7. Run the code.

DATA MODEL OPTIONS SE ▶ ▾

► METHODS

▲ STATISTICS

Display statistics:

Default and selected statistics ▾

Parameter Estimates

Standardized regression coefficients

Confidence limits for estimates

Sums of Squares

Sequential sum of squares (Type I)

Partial sum of squares (Type II)

Partial and Semipartial Correlations

Squared partial correlations

Squared semipartial correlations

Diagnostics

Analysis of influence

Analysis of residuals

Predicted values

► Multiple Comparisons

▲ Collinearity

Collinearity analysis

Tolerance values for estimates

Variance inflation factors

Note: Alternatively, you can write the code directly in SAS.

```
/*st105d03.sas*/ /*Part B*/
proc reg data=amescombined;
  model SalePrice = &interval score / vif;
  title 'Collinearity Diagnostics';
run;
quit;
```

Partial PROC REG Output

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	Intercept	1	-4254871	3419274	-1.24	0.2144	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	923.25717	684.04818	1.35	0.1782	27569
Basement_Area	Basement area in square feet	1	2178.35638	1709.68175	1.27	0.2036	411868
Garage_Area	Size of garage in square feet	1	35.01213	6.46640	5.41	<.0001	1.41398
Deck_Porch_Area	Total area of decks and porches in square feet	1	30.64725	7.97228	3.84	0.0001	1.21667
Lot_Area	Lot size in square feet	1	0.69964	0.31644	2.21	0.0278	1.20422
Age_Sold	Age of house when sold, in years	1	-422.21228	44.18905	-9.55	<.0001	1.60476
Bedroom_Abv Gr	Bedrooms above grade	1	-4888.35244	1687.71153	-2.90	0.0041	1.48233
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3047.94315	1919.03449	1.59	0.1133	1.73073
score		1	429.97552	341.96962	1.26	0.2096	533085

Some of the VIFs are much larger than 10. A severe collinearity problem is present. At this point there are many ways to proceed. However, it is always a good idea to use some subject-matter expertise. When subject-matter expertise is not available, another option is to systematically remove variables starting with the highest VIF and re-run the analysis. Much like *p-values*, the VIF values will need to be updated with each successive variable removal.

Reaching out to the researchers that provided the **score** variable, it was determined that **score** was a composite variable.

```
score=round(10000 - (2*Gr_Liv_Area + 5*Basement_Area),10);
```

The researchers, on the basis of prior literature, created a composite variable, which is a weighted function of the two variables, **Gr_Liv_Area**, and **Basement_Area**. This is not an uncommon occurrence and illustrates an important point. If a composite variable is included in a model along with some or all of its component measures, there is bound to be collinearity.

If the composite variable has meaning, it can be used as a stand-in measure for both components and you can remove the variables **Gr_Liv_Area** and **Basement_Area** from the analysis.

Composite measures have the disadvantage of losing some information about the individual variables. If this is of concern, then remove **score** from the analysis.

A decision was made to remove **score** from the analysis. Another check of collinearity is warranted.

8. Remove **score** from the model and rerun the task.

Note: Alternatively you can edit the code directly.

```
proc reg data=amescombined;
  NOSCORE: model SalePrice = &interval / vif;
  title2 'Removing Score';
run;
quit;
```

Partial PROC REG Output

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	Intercept	1	44347	6191.27194	7.16	<.0001	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	63.19776	5.58574	11.31	<.0001	1.83461
Basement_Area	Basement area in square feet	1	28.69218	3.41703	8.40	<.0001	1.64195
Garage_Area	Size of garage in square feet	1	35.75419	6.44584	5.55	<.0001	1.40220
Deck_Porch_Area	Total area of decks and porches in square feet	1	31.37054	7.95944	3.94	0.0001	1.21034
Lot_Area	Lot size in square feet	1	0.69950	0.31676	2.21	0.0280	1.20422
Age_Sold	Age of house when sold, in years	1	-420.81504	44.21914	-9.52	<.0001	1.60375
Bedroom_Abv Gr	Bedrooms above grade	1	-4834.84875	1688.85823	-2.86	0.0045	1.48138
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3022.12472	1920.83907	1.57	0.1167	1.73053

All VIF values are smaller than 2 now.

Collinearity can have a substantial effect on the outcome of a stepwise procedure for model selection. Because the significance of important variables can be masked by collinearity, the final model might not include very important variables. This is why it is advisable to deal with collinearity before using any automated model selection tool.

Note: There are other approaches to dealing with collinearity. Two techniques are *ridge regression* and *principal components regression*. In addition, *recentering* the predictor variables can sometimes eliminate collinearity problems, especially in a polynomial regression and in ANCOVA models. Another could be to make composite variables from the collinear variables by taking a ratio or perhaps an average.

End of Demonstration

5.04 Multiple Choice Poll

Which of the following assumptions does collinearity violate?

- a. Independent errors
- b. Constant variance
- c. Normally distributed errors
- d. None of the above

44

Copyright © SAS Institute Inc. All rights reserved.



5.05 Poll

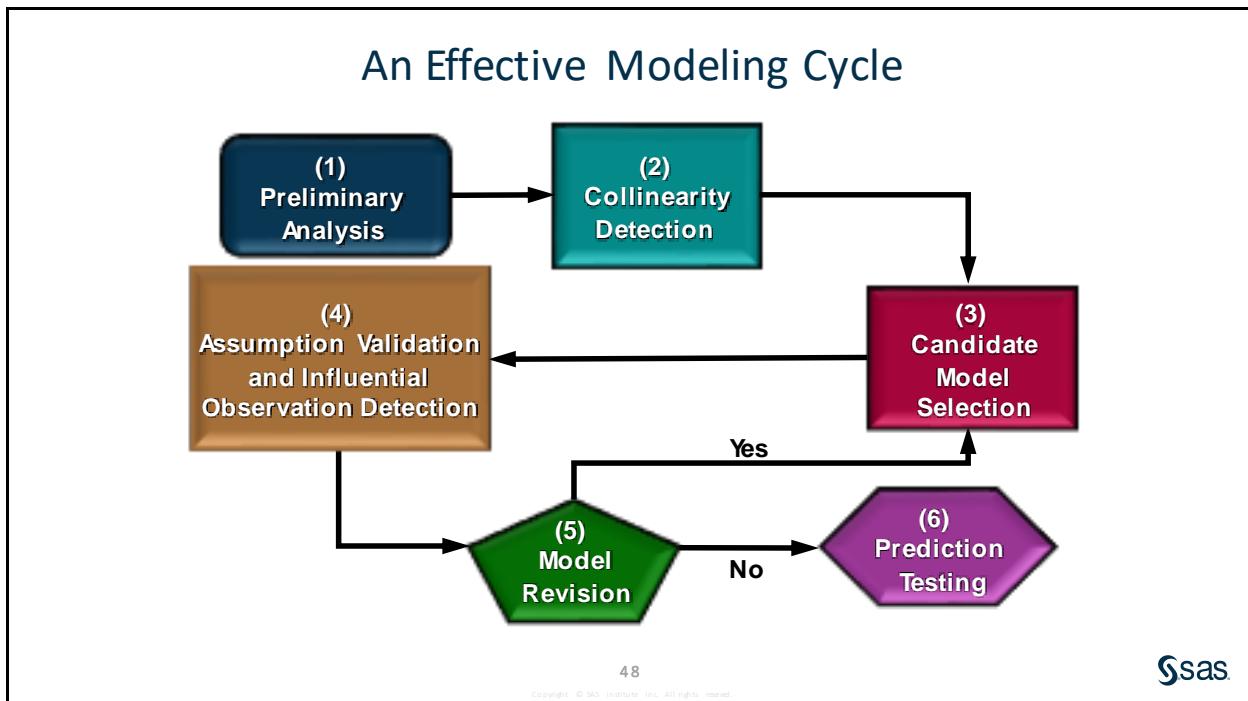
If there is no correlation among the predictor variables, can there still be collinearity in the model?

- Yes
- No

46

Copyright © SAS Institute Inc. All rights reserved.





- (1) **Preliminary Analysis:** This step includes the use of descriptive statistics, graphs, and correlation analysis.
- (2) **Collinearity Detection:** This step includes the use of the VIF statistic, condition indices, and variation proportions.
- (3) **Candidate Model Selection:** This step uses the numerous selection options in PROC REG or PROC GLMSELECT to identify one or more candidate models.
- (4) **Assumption Validation and Influential Observation Detection:** The former includes the plots of residuals and graphs of the residuals versus the predicted values. It also includes a test for equal variances. The latter includes the examination of R-Student residuals, Cook's D statistic, DFFITS, and DFBETAS statistics.
- (5) **Model Revision:** If steps (3) and (4) indicate the need for model revision, generate a new model by returning to these two steps.
- (6) **Prediction Testing:** If possible, validate the model with data not used to build the model.



Exercises

3. Assessing Collinearity

Using the **STAT1.BodyFat2** data set, run a regression of **PctBodyFat2** on all the other numeric variables in the file.

- a. Determine whether there is a collinearity problem.
- b. If so, decide what you would like to do about that. Will you remove any variables? Why or why not?

End of Exercises

5.4 Solutions

Solutions to Exercises

1. Examining Residuals

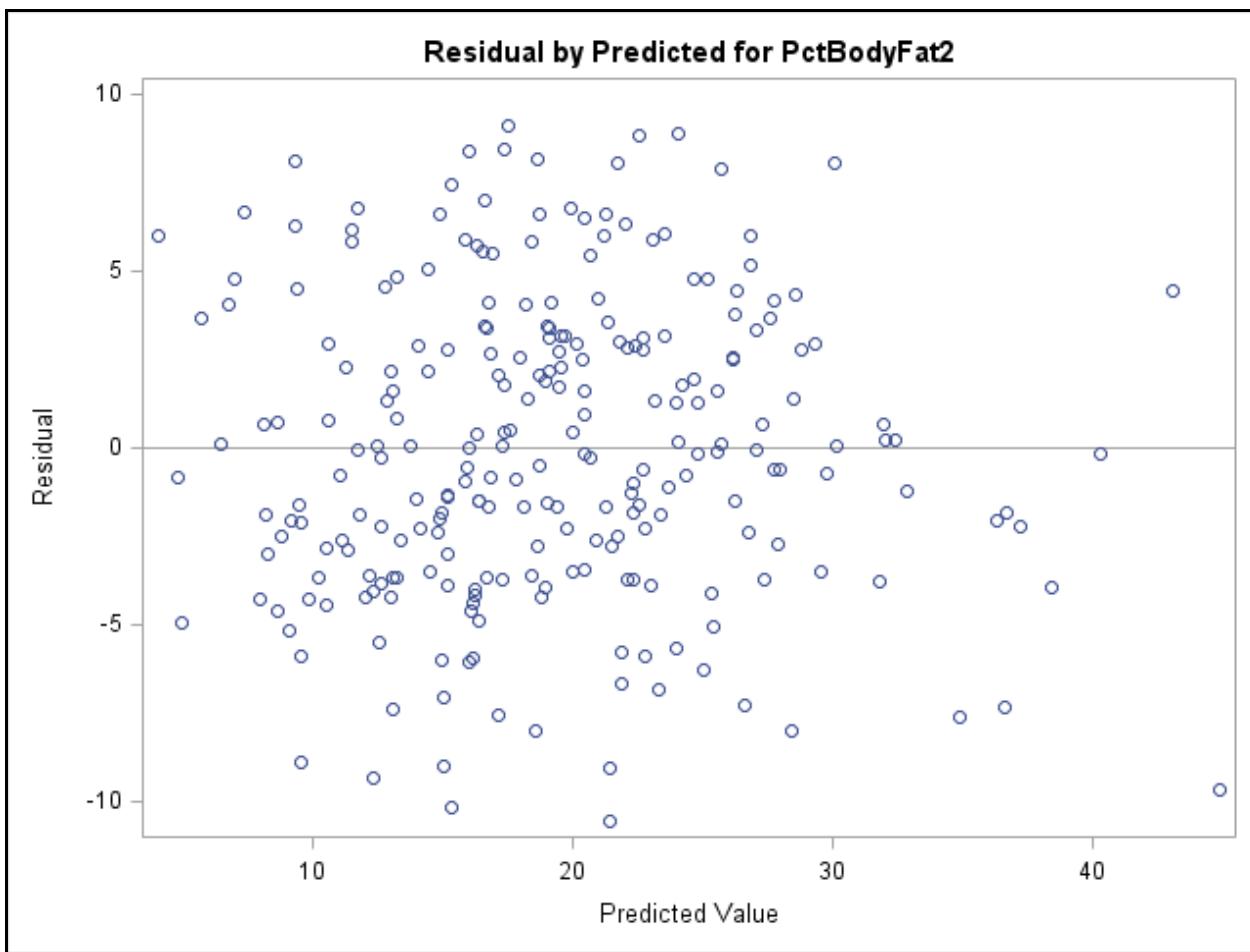
Assess the model obtained from the final forward stepwise selection of predictors for the **STAT1.BodyFat2** data set. Run a regression of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**. Create plots of the residuals by the four regressors and by the predicted values and a normal Quantile-Quantile plot.

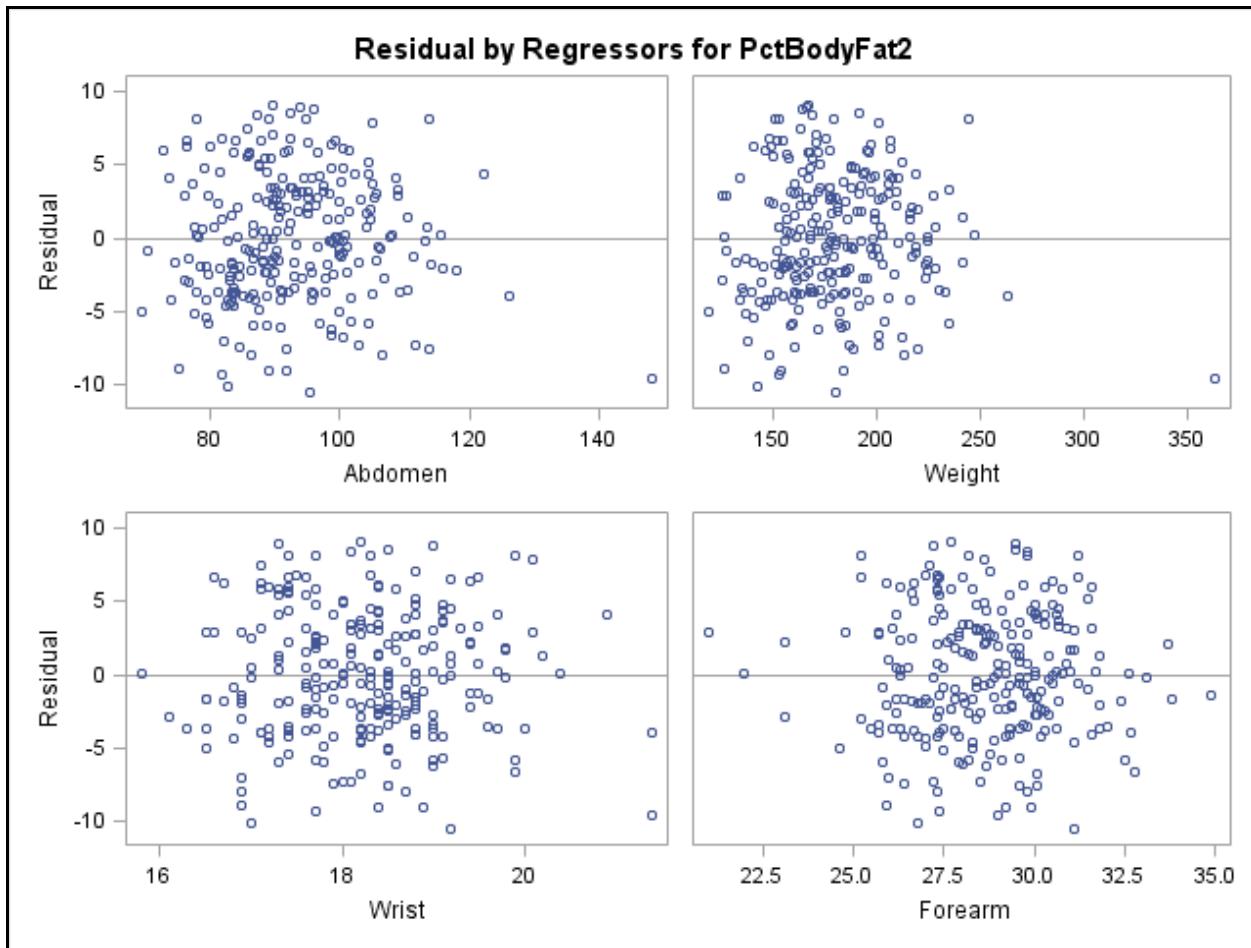
- 1) Open the **Linear Regression** task under **Statistics**.
- 2) On the DATA tab, select the BodyFat2 data set and assign the variables.
- 3) On the MODEL tab, specify the model by adding in the model effects.
- 4) On the OPTIONS tab, the default selections produce the residuals plot for each explanatory variable and all the default diagnostic plots, which include residuals versus predicted plot and the normal Quantile-Quantile plot. To specify which output plots to display, edit the code.
- 5) Run the code.

Note: Alternatively you can write the code directly in SAS.

```
/*st105s01.sas*/
ods graphics / imagemap=on;
proc reg data=STAT1.BodyFat2
plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);
FORWARD: model PctBodyFat2=
          Abdomen Weight Wrist Forearm;
id Case;
title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;
```

- a. Do the residual plots indicate any problems with the constant variance assumption?



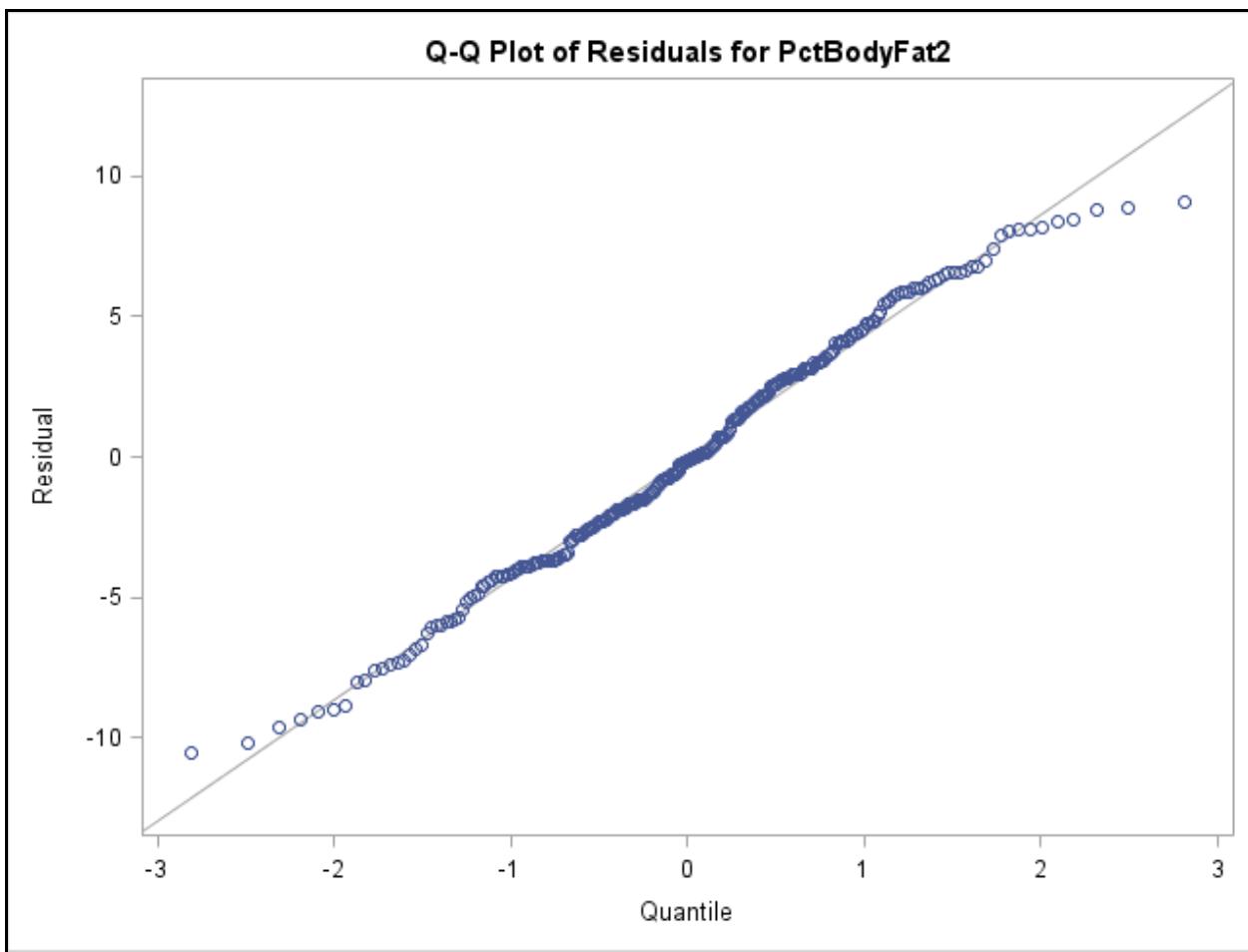


It does not appear that the data violate the assumption of constant variance. Also, the residuals show nice random scatter and indicate no problem with model specification.

- b. Are there any outliers indicated by the evidence in any of the residual plots?

There are a few (x-space) outliers for Wrist and Forearm and one clear outlier in each of Abdomen and Weight values.

- c. Does the Quantile-Quantile plot indicate any problems with the normality assumption?



The normality assumption seems to be met.

2. Generating Potential Outliers

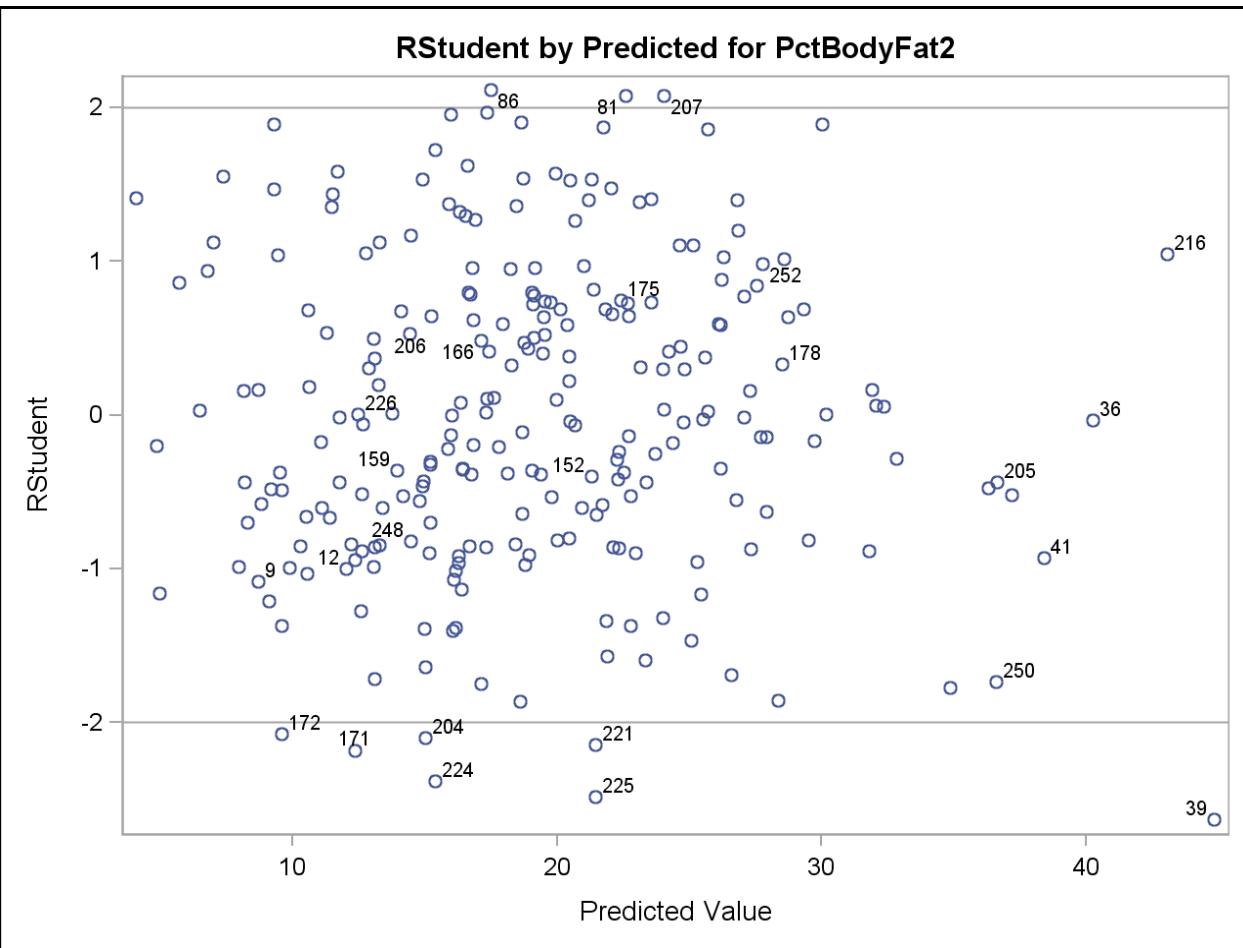
Using the **STAT1.BodyFat2** data set, run a regression model of **PctBodyFat2** on **Abdomen**, **Weight**, **Wrist**, and **Forearm**.

- Use plots to identify potential influential observations based on the suggested cutoff values.
 - Open the **Linear Regression** task under Statistics.
 - Select the **BodyFat2** data set, assign **PctBodyFat2** as the dependent variable, and assign the explanatory variables.
 - On the MODEL tab, specify the model.
 - On the OPTIONS tab, use the drop-down menu to specify displaying the diagnostic plots as individual plots.
 - Uncheck the option of displaying residual plot for each explanatory variable.
 - Expand the **More Diagnostic Plots** property and select the options to display diagnostic plots for potential influential observations with labeling for extreme points.
 - Modify the code to export the **RSTUDENT**, **DFFITS**, **DFBETAS**, and **Cook's D** influence statistics.

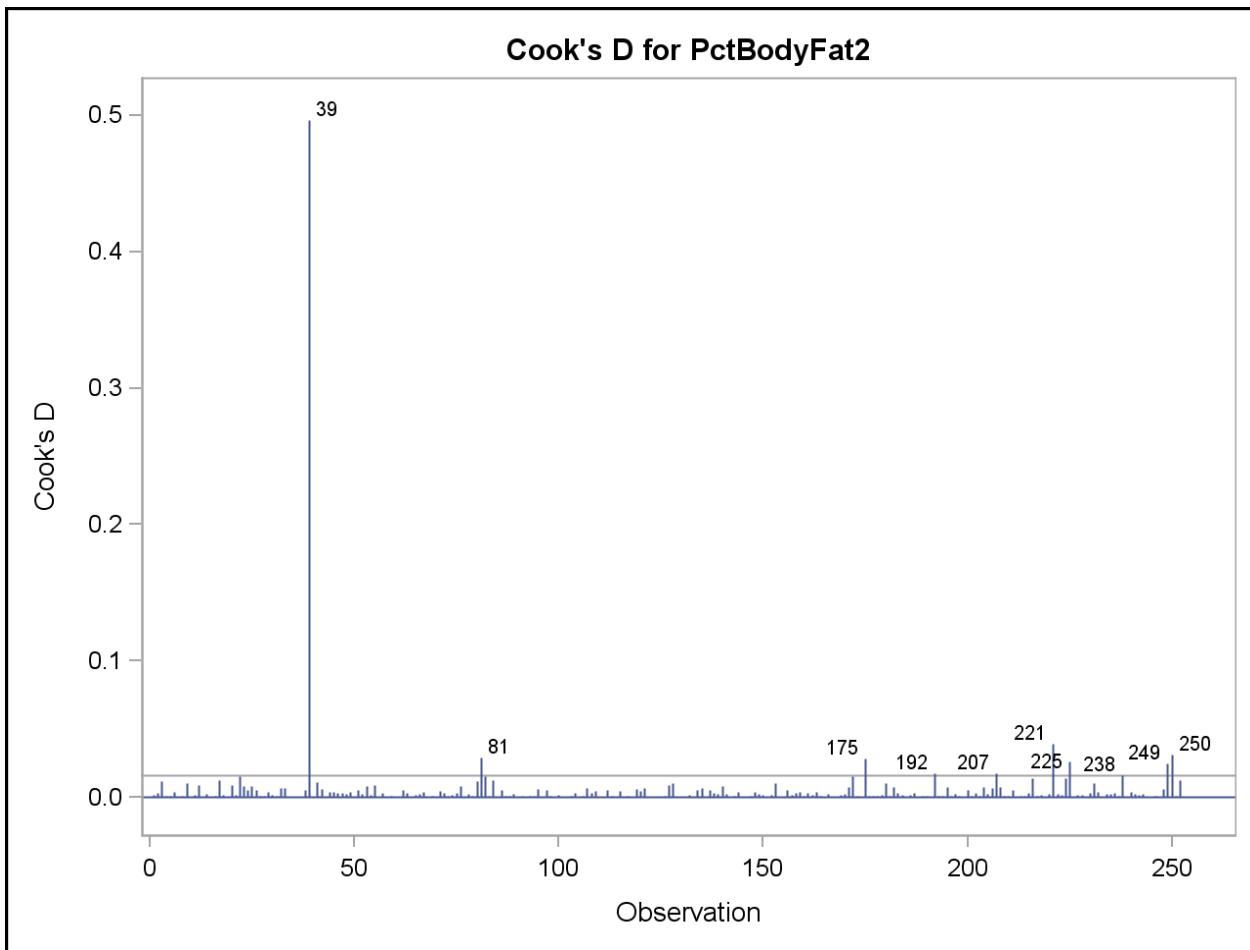
Note: Alternatively, you can write the code directly in SAS.

```
/*st105s02.sas*/ /*Part A*/
ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
    COOKSDPLOT=Cook
    DFFITSPLOT=Dffits
    DFBETASPANEL=Dfbs;
proc reg data=STAT1.BodyFat2
    plots(only label)=
        (RSTUDENTBYPREDICTED
        COOKSD
        DFFITS
        DFBETAS);
FORWARD: model PctBodyFat2=
    Abdomen Weight Wrist Forearm;
id Case;
title 'FORWARD Model - Plots of Diagnostic Statistics';
run;
quit;
```

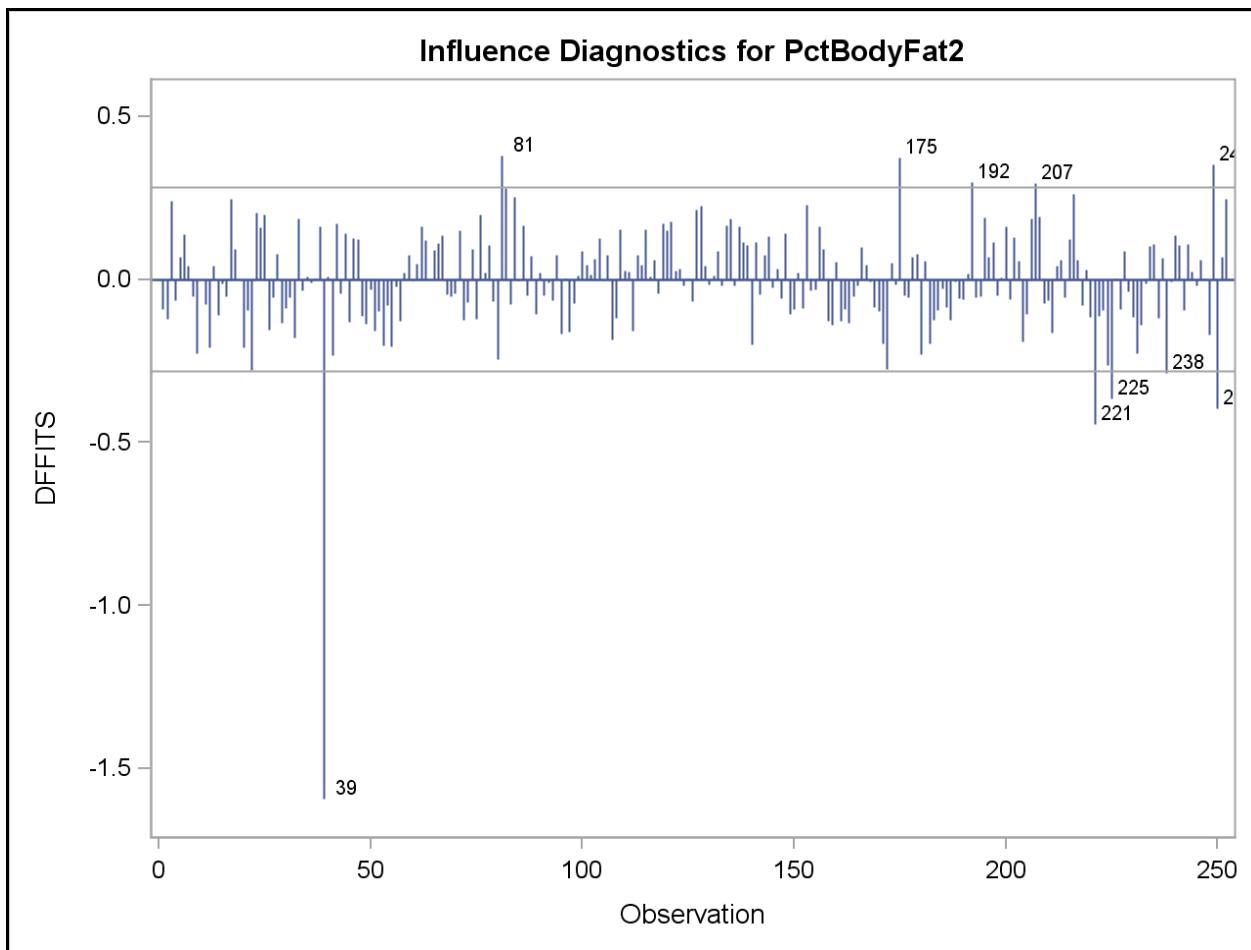
Note: The Linear Regression task produces all the default diagnostic plots. The output below only shows plots that are relevant to identifying potential influential observations.



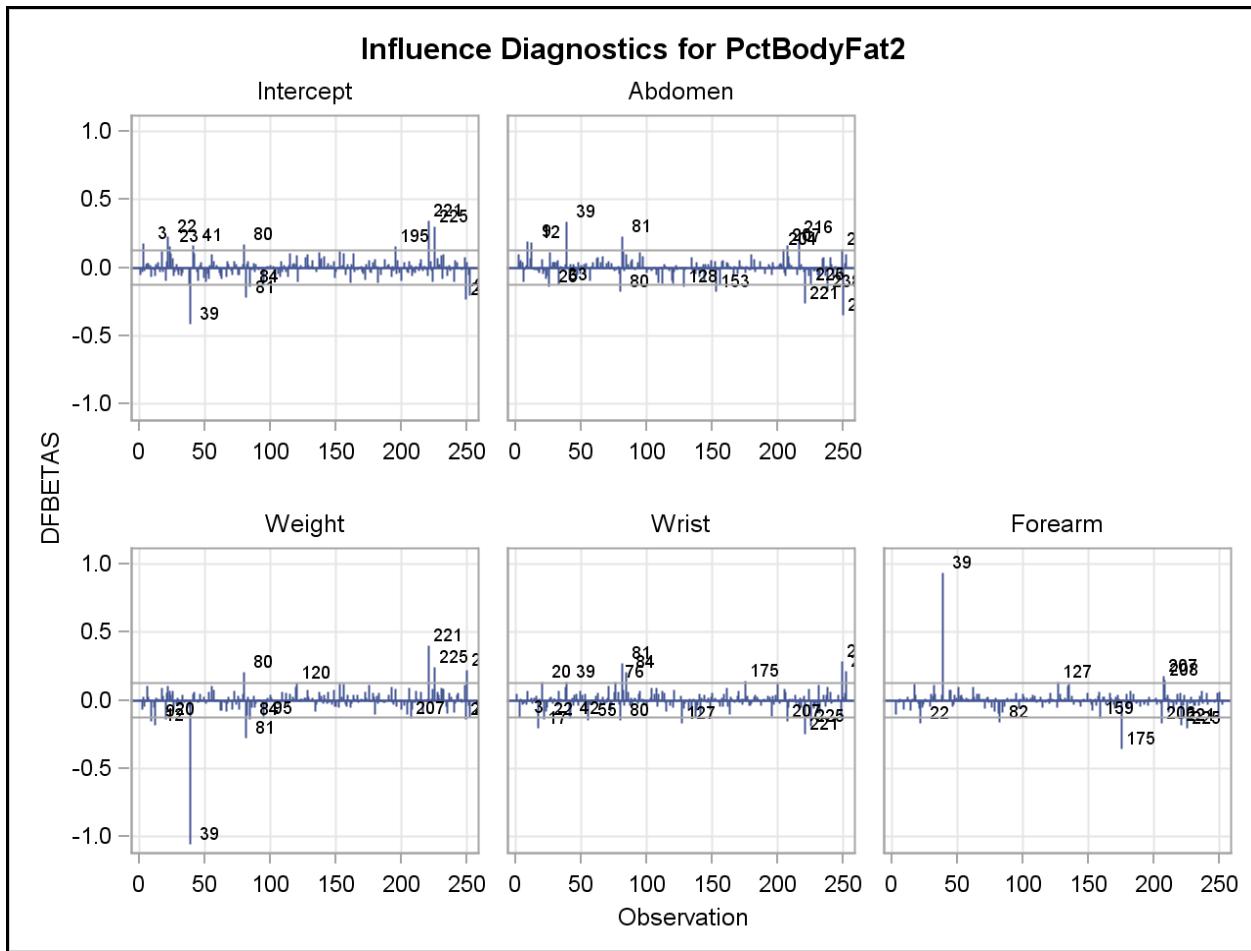
There are only a modest number of observations farther than two standard error units from the mean of 0.



There are 10 labeled outliers, but observation 39 is clearly the most extreme.



The same observations are shown to be influential by the DFFITS statistic.



DFBETAS are particularly high for observation 39 on the parameters for weight and forearm circumference.

- b. Output residuals to a data set, subset the data set by only those who are potentially influential outliers, and print the results.

```
/* st105s02.sas */ /* Part B */
data influential;
/* Merge data sets from above. */
merge Rstud
      Cook
      Dffits
      Dfbs;
by observation;

/* Flag observations that have exceeded at least one cutpoint; */
if (ABS(Rstudent)>3) or (Cooksdlable ne ' ') or Dffitsout then
flag=1;
array dfbetas{*} _dfbetasout: ;
do i=2 to dim(dfbetas);
   if dfbetas{i} then flag=1;
end;

/* Set to missing values of influence statistics for those */
/* who have not exceeded cutpoints; */
if ABS(Rstudent)<=3 then RStudent=.;
if Cooksdlable eq ' ' then CooksD=.;

/* Subset only observations that have been flagged. */
if flag=1;
drop i flag;
run;

proc print data=influential;
id observation ID1;
var Rstudent CooksD Dffitsout _dfbetasout: ;
run;
```

Observation	n	id1	RStudent	CooksD	DFFITSOUT	DFBETASOUT1	DFBETASOUT2	DFBETASOUT3	DFBETASOUT4	DFBETASOUT5
3	3	.	.	.	0.17943	.	.	-0.12815	.	.
9	9	0.18911	-0.15600	.	.	.
12	12	0.18169	-0.18076	.	.	.
17	17	-0.20902	.	.
20	20	-0.13786	0.13273	.	.
22	22	.	.	.	0.22887	.	.	-0.14080	-0.16797	.
25	25	-0.14080
33	33	-0.12765
39	39	.	0.49632	-1.59408	-0.41792	0.33576	-1.05761	0.13217	0.93125	.
42	42	-0.13688	.	.
55	55	-0.14907	.	.
76	76	0.13108	.	.
80	80	.	.	.	0.17122	-0.17507	0.20391	-0.14744	.	.
81	81	.	0.02858	0.38053	-0.22179	0.22631	-0.27484	0.26977	.	.
82	82	-0.16453	.
84	84	.	.	.	-0.14277	.	-0.13915	0.20279	.	.
95	95	-0.13519	.	.	.
120	120	0.12609	.	.	.
127	127	-0.16625	0.13285	.
128	128	-0.13838
153	153	-0.17467
159	159	-0.13278	.
175	175	.	0.02787	0.37296	.	.	.	0.14200	-0.35339	.
192	192	.	0.01752	0.29750
204	204	0.13453
206	206	-0.17242	.
207	207	.	0.01716	0.29490	.	0.16026	-0.13169	-0.15412	0.17410	.
208	208	0.14747	.
216	216	0.21712
221	221	.	0.03911	-0.44540	0.34282	-0.26106	0.39789	-0.24565	-0.18174	.
225	225	.	0.02633	-0.36660	0.30270	-0.12914	0.23904	-0.19078	-0.20840	.
238	238	.	0.01629	-0.28661	.	-0.17388
249	249	.	0.02463	0.35266	-0.23435	0.13125	-0.14344	0.28748	.	.
250	250	.	0.03108	-0.39579	.	-0.35320	0.21925	.	.	.
252	252	.	.	.	-0.20349	.	-0.12708	0.21088	.	.

The same observations appear on this listing as in the plots.

Note: Examine the values of observation 39 to see what is causing problems. You might find it interesting.

3. Assessing Collinearity

Using the **STAT1.BodyFat2** data set, run a regression of **PctBodyFat2** on all the other numeric variables in the file.

- a. Determine whether there is a collinearity problem.
 - 1) Open the **Linear Regression** task under **Statistics**.
 - 2) On the DATA tab, select the **BodyFat2** data set and assign **PctBodyFat2** as the dependent variable and assign **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist** as the continuous variables.
 - 3) On the MODEL tab, specify the appropriate model.
 - 4) On the OPTIONS tab, change the option from displaying default statistics to default and selected statistics. Expand the **Collinearity** property and check the box to display **Variance inflation factors**.
 - 5) Suppress all plots by unchecking the boxes for all the different graphic output options.
 - 6) Run the code.

Note: Alternatively, you can write the code directly.

```
/*st105s03.sas*/ /*Part A*/
ods graphics off;
proc reg data=STAT1.BodyFat2;
  FULLMODL: model PctBodyFat2=
    Age Weight Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
    / vif;
  title 'Collinearity -- Full Model';
run;
quit;
ods graphics on;
```

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	13	13159	1012.22506	54.50	<.0001
Error	238	4420.06401	18.5710		
Corrected Total	251	17579			

Root MSE	4.30949	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7348
Coeff Var	22.50293		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-21.35323	22.18616	-0.96	0.3368	0
Age	1	0.06457	0.03219	2.01	0.0460	2.22447
Weight	1	-0.09638	0.06185	-1.56	0.1205	44.65251
Height	1	-0.04394	0.17870	-0.25	0.8060	2.93911
Neck	1	-0.47547	0.23557	-2.02	0.0447	4.43192
Chest	1	-0.01718	0.10322	-0.17	0.8679	10.23469
Abdomen	1	0.95500	0.09016	10.59	<.0001	12.77553
Hip	1	-0.18859	0.14479	-1.30	0.1940	14.54193
Thigh	1	0.24835	0.14617	1.70	0.0906	7.95866
Knee	1	0.01395	0.24775	0.06	0.9552	4.82530
Ankle	1	0.17788	0.22262	0.80	0.4251	1.92410
Biceps	1	0.18230	0.17250	1.06	0.2917	3.67091
Forearm	1	0.45574	0.19930	2.29	0.0231	2.19193
Wrist	1	-1.65450	0.53316	-3.10	0.0021	3.34840

There seems to be high collinearity associated with Weight, Hip, and Abdomen. Chest and Thigh are below the cutoff but are larger than the others that do not exceed 5.

- b. If so, decide what you would like to do about that. Will you remove any variables? Why or why not?

The answer is not so easy. Weight is collinear with some set of the other variables, but as you saw before in your model-building process, Weight is a relatively significant predictor in the “best” models. The answer is for a subject-matter expert to determine.

If you want to remove Weight, simply run the model again without that variable.

- 1) Modify the model using the Model Effects Builder and rerun the task.

Note: Alternatively, you can write the code directly in SAS.

```
/*st105s03.sas*/ /*Part B*/
ods graphics off;
proc reg data=STAT1.BodyFat2;
  NOWT: model PctBodyFat2=
    Age Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
    / vif;
  title 'Collinearity -- No Weight';
run;
quit;
ods graphics on;
```

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	12	13114	1092.81860	58.49	<.0001
Error	239	4465.16664	18.68271		
Corrected Total	251	17579			

Root MSE	4.32235	R-Square	0.7460
Dependent Mean	19.15079	Adj R-Sq	0.7332
Coeff Var	22.57008		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation	
Intercept	1	10.44939	8.73019	1.20	0.2325	0	
Age	1	0.07376	0.03174	2.32	0.0210	2.14988	
Height	1	-0.22096	0.16836	-1.60	0.1116	1.75152	
Neck	1	-0.60041	0.22217	-2.70	0.0074	3.91858	
Chest	1	-0.09400	0.09096	-1.03	0.3025	7.90070	
Abdomen	1	0.91038	0.08575	10.62	<.0001	11.48744	
Hip	1	-0.30384	0.12485	-2.43	0.0157	10.74814	
Thigh	1	0.21896	0.14538	1.51	0.1334	7.82619	
Knee	1	-0.02664	0.24711	-0.11	0.9142	4.77198	
Ankle	1	0.10706	0.21858	0.49	0.6247	1.84391	
Biceps	1	0.12481	0.16901	0.74	0.4610	3.50299	
Forearm	1	0.45808	0.19989	2.29	0.0228	2.19181	
Wrist	1	-1.77201	0.52937	-3.35	0.0009	3.28143	

Some collinearity still exists in the model. If Abdomen, the remaining variable with the highest VIF, is removed and then the R-square (and adjusted R-square) value is reduced by approximately 0.13.

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

5.01 Poll – Correct Answer

Predictor variables are assumed to be normally distributed in linear regression models.

- True
- False

6

Copyright © SAS Institute Inc. All rights reserved.

5.02 Multiple Choice Poll – Correct Answer

Given the properties of the standard normal distribution, you would expect about 95% of the studentized residuals to be between which two values?

- a. -3 and 3
- b. -2 and 2
- c. -1 and 1
- d. 0 and 1
- e. 0 and 2
- f. 0 and 3

24

Copyright © SAS Institute Inc. All rights reserved.

5.03 Multiple Choice Poll – Correct Answer

How many observations did you find that might substantially influence parameter estimates as a group?

- a. 0
- b. 1
- c. 4
- d. 5
- e. 7
- f. 10

34

Copyright © SAS Institute Inc. All rights reserved.

5.04 Multiple Choice Poll – Correct Answer

Which of the following assumptions does collinearity violate?

- a. Independent errors
- b. Constant variance
- c. Normally distributed errors
- d. None of the above

45

Copyright © SAS Institute Inc. All rights reserved.

5.05 Poll – Correct Answer

If there is no correlation among the predictor variables, can there still be collinearity in the model?

- Yes
- No

Chapter 6 Model Building and Scoring for Prediction

6.1 Brief Introduction to Predictive Modeling.....	6-3
Demonstration: Predictive Model Building.....	6-9
Exercises.....	6-15
6.2 Scoring Predictive Models	6-16
Demonstration: Scoring Using PROC PLM and PROC GLMSELECT	6-19
Exercises.....	6-21
6.3 Solutions	6-22
Solutions to Exercises	6-22
Solutions to Student Activities (Polls/Quizzes)	6-29

6.1 Brief Introduction to Predictive Modeling

Objectives

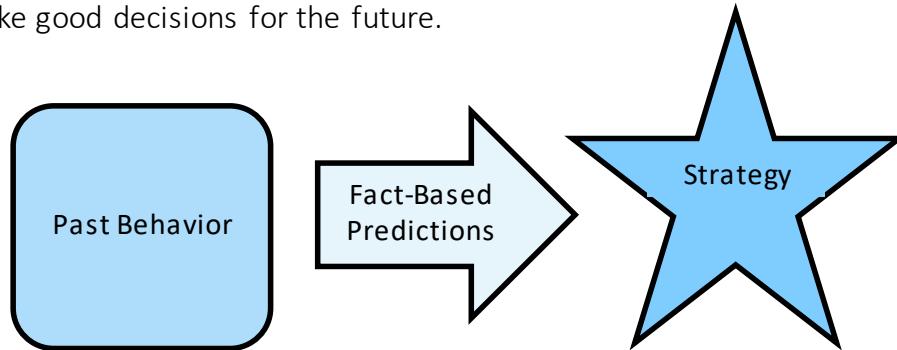
- Explain the concepts of predictive modeling.
- Illustrate the modeling essentials of a predictive model.
- Explain the importance of data partitioning.

3



From Descriptive to Predictive Modeling

Predictive modeling techniques, paired with scoring and good model management, enable you to use your data about the past and the present to make good decisions for the future.



4

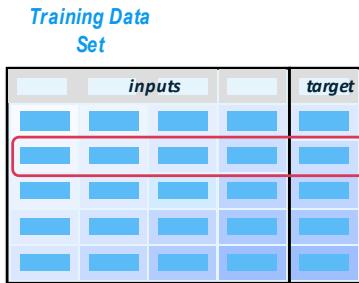


There are many business applications of predictive modeling. *Database marketing* uses customer databases to improve sales promotions and product loyalty. In *target marketing*, the cases are customers, the inputs are attributes such as previous purchase history and demographics, and the target is often a binary variable indicating a response to a past promotion. The aim is to find segments of customers that are likely to respond to some offer so that they can be targeted. Historic customer databases can also be used to predict who is likely to switch brands or cancel services (churn). Loyalty promotions can then be targeted at new cases that are at risk.

Credit scoring is used to decide whether to extend credit to applicants. The cases are past applicants. Most input variables come from the credit application or credit reports. A relevant binary target is whether the case defaulted (charged off) or the debt was paid. The aim is to reduce defaults and serious delinquencies on new applicants for credit.

In *fraud detection*, the cases are transactions (for example, telephone calls and credit card purchases) or insurance claims. The inputs are the particulars and circumstances of the transaction. The binary target is whether that case was fraudulent. The aim is to anticipate fraud or abuse on new transactions or claims so that they can be investigated or impeded.

Predictive Modeling Terminology

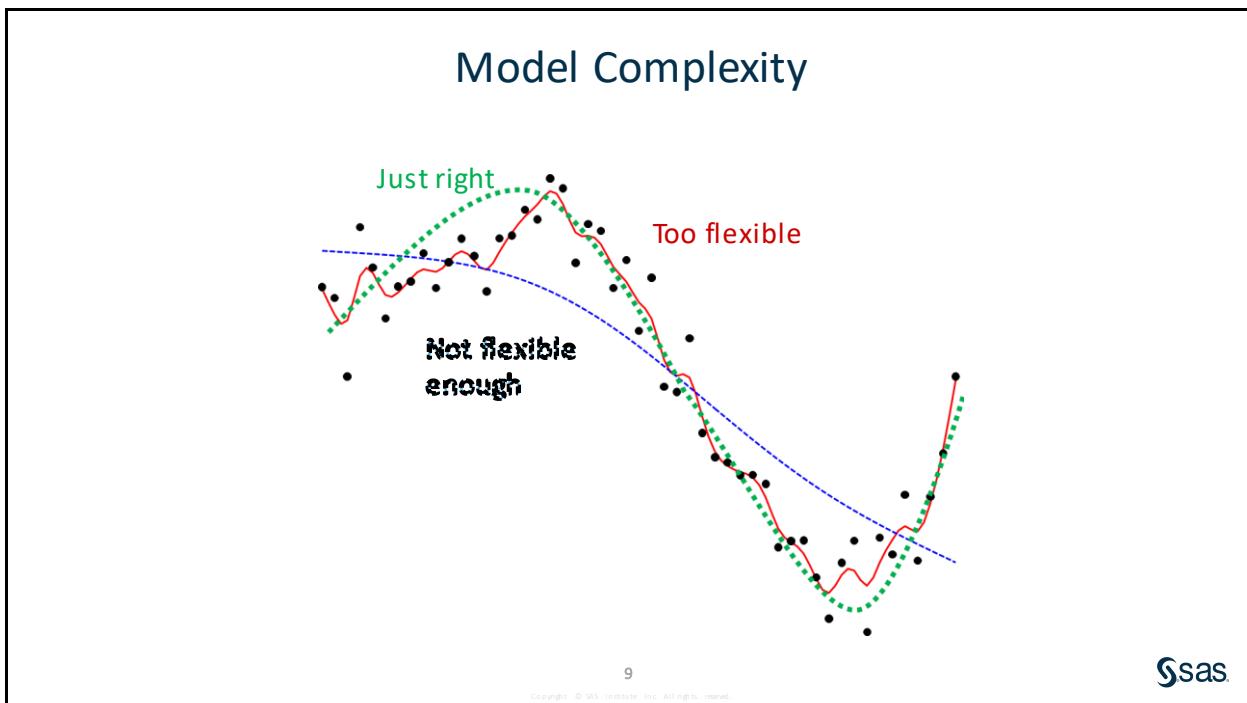


The variables are called *inputs* and *targets*.

The observations in a training data set are known as *training cases*.

Predictive modeling starts with a *training data set*. The observations in a training data set are known as *training cases* (also known as *examples*, *instances*, or *records*). The variables are called *inputs* (also known as *predictors*, *features*, *explanatory variables*, or *independent variables*) and *targets* (also known as a *response*, *outcome*, or *dependent variable*). For a given case, the inputs reflect your state of knowledge before measuring the target.

The measurement scale of the inputs and the target can be varied. The inputs and the target can be numeric variables, such as income. They can be nominal variables, such as occupation. They are often binary variables, such as a positive or negative response concerning home ownership.

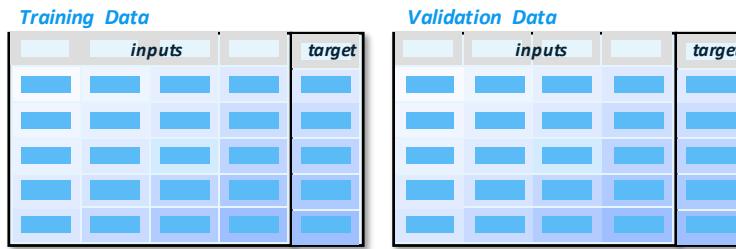


Fitting a model to data requires searching through the space of possible models. Constructing a model with good generalization requires choosing the right complexity. For regression, including more terms in the model increases complexity.

Selecting model complexity involves a tradeoff between bias and variance. An insufficiently complex model might not be flexible enough. This leads to *underfitting* – that is, systematically missing the *signal* (the true relationships). This leads to biased inferences, which are inferences that are not the true ones in the population.

A naive modeler might assume that the most complex model should always outperform the others, but this is not the case. An overly complex model might be too flexible. This leads to *overfitting* – that is, accommodating nuances of the *random noise* (chance relationships) in the particular sample. This leads to models that have higher variance when applied to a population. A model with just enough flexibility gives the best generalization.

Honest Assessment and Data Partitioning



Partition available data into training and validation sets.

The model is fit on the training data set, and model performance is evaluated on the validation data set.

The strategy for choosing model complexity in data mining is to use *honest assessment*. With honest assessment, you select the model that performs best on a **validation** data set, which is not used to fit the model. Assessing performance on the same data set that was used to develop the model leads to selecting too complex a model (overfitting).

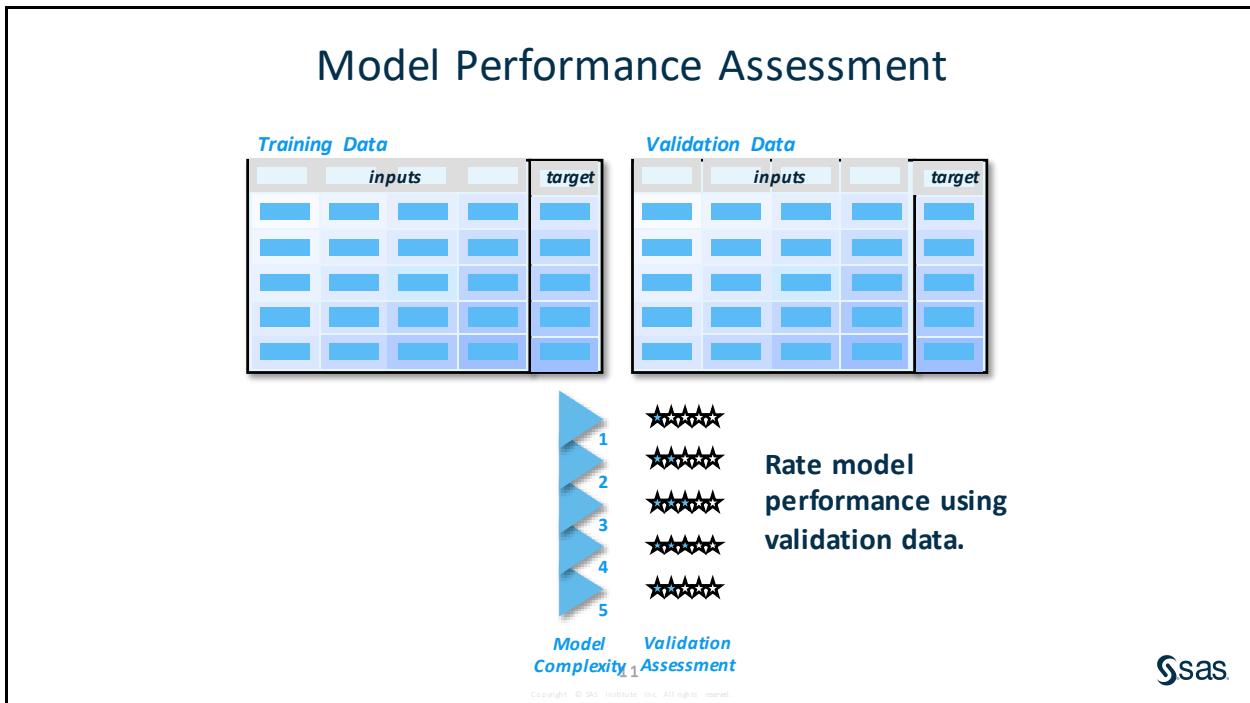
Note: The classic example of this is selecting linear regression models based on R square.

In predictive modeling, the standard strategy for honest assessment of model performance is data splitting. A portion is used for fitting the model – that is, the training data set. The remaining data are separated for empirical validation.

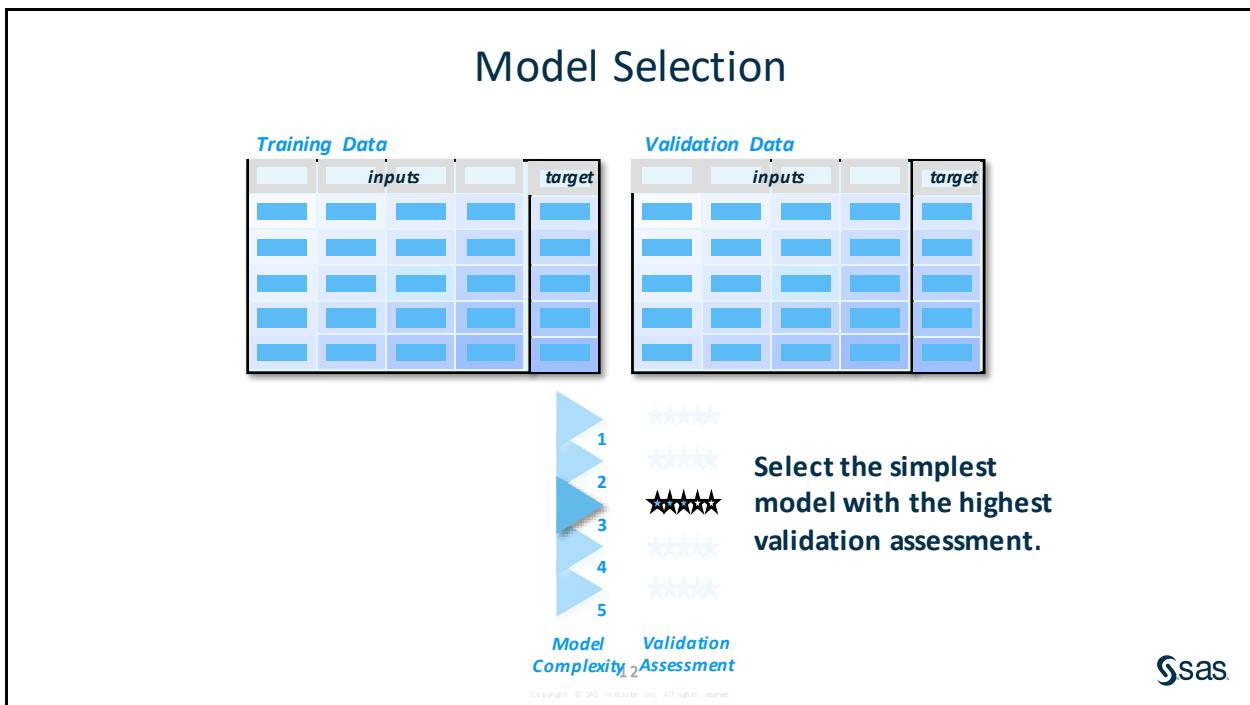
The validation data set is used for monitoring and tuning the model to improve its generalization. The tuning process usually involves selecting among models of different types and complexities. The tuning process optimizes the selected model on the validation data.

Note: Because the validation data are used to select from a set of related models, reported performance will be overstated, on the average. Consequently, a further holdout sample is needed for a final, unbiased assessment. The *test data set* has only one use, which is to give a final honest estimate of generalization. Cases in the test set must be treated in the same way that new data would be treated. The cases cannot be involved in any way in the determination of the fitted prediction model. In practice, many analysts see no need for a final honest assessment of generalization. An optimal model is chosen using the validation data, and the model assessment measured on the validation data is reported as an upper bound on the performance expected when the model is deployed.

With small or moderate data sets, data splitting is inefficient; the reduced sample size can severely degrade the fit of the model. Computer-intensive methods, such as the crossvalidation and bootstrap methods, were developed so that all the data can be used for both fitting and honest assessment. However, data mining usually has the luxury of massive data sets.



Using performance on the training data set usually leads to selecting a model that is too complex. (The classic example is selecting linear regression models based on R square.) To avoid this problem, PROC GLMSELECT can select the model based on validation data performance, from the sequence of models selected based on training data measures.



In keeping with Occam's razor, the best model is the simplest model with the highest validation performance.

6.01 Multiple Choice Poll

When using honest assessment, which of the following would be considered the best model?

- a. The simplest model with the best performance on the training data
- b. The simplest model with the best performance on the validation data
- c. The most complex model with the best performance on the training data
- d. The most complex model with the best performance on the validation data

PROC GLMSELECT for Predictive Modeling

If a validation data set is available:

```
PROC GLMSELECT DATA=Training data-set
    VALDATA=Validation data-set;
    MODEL targets=inputs </ options>;
RUN;
```

If a validation data set is not available:

```
PROC GLMSELECT DATA=Training data-set
    <SEED=number>;
    MODEL targets=inputs </ options>;
    PARTITION FRACTION(<TEST=fraction>
        <VALIDATE=fraction>
    );
RUN;
```

PROC GLMSELECT can perform model building with honest assessment with a holdout (validation) data set in two ways. If a holdout data set has already been created, then it can be referred to in the PROC GLMSELECT statement as the VALDATA. If there is only one data set, it can be randomly partitioned into training and validation data (as well as test data, if required) using a FRACTION option in the PARTITION statement. A nonzero seed in the PROC GLMSELECT statement will assure replicability.



Predictive Model Building

Example: Use the GLMSELECT procedure to build a predictive linear regression model of **SalePrice** from both categorical (**Fireplaces** **Garage_Type_2** **Foundation_2** **Heating_QC** **Masonry_Veneer** **Lot_Shape_2** **Central_Air** **Season_Sold** **House_Style2** **Overall_Qual2** **Overall_Cond2**) and interval (**Gr_Liv_Area** **Bedroom_AbvGr** **Total_Bathroom** **Deck_Porch_Area** **Age_Sold** **Lot_Area** **Basement_Area** **Garage_Area**) predictors. Use the STORE statement to create an item store to use in subsequent processes. Use **AmesHousing3** as the training data set and **AmesHousing4** as the validation data set. Use backward elimination with SBC for the training data as the model building criterion and choose the model with the smallest average squared error for the validation data set.

1. Open the **Predictive Regression Models** task under Statistics.
 2. On the DATA tab, select the **AmesHousing3** data set.
 3. Assign **SalePrice** as the dependent variable. Assign the classification and continuous variables as listed.
 4. On the DATA tab, expand **Parameterization of Effects** and select the **GLM coding** option.
 5. On the MODEL tab, specify the model using the model editor.
 6. On the SELECTION tab, specify using backward elimination as the selection method and Schwarz Bayesian information criterion as the criterion.
 7. Expand the SELECTION PLOTS property and select the option to include **Coefficient plots**.
 8. Expand the DETAILS property and then expand the Model Effects Hierarchy property. Modify the model effects hierarchy to **do not maintain hierarchy of effects**.
 9. Open the editor.
 10. Manually specify **AmesHousing4** as the test data set using **valdata=STAT1.AMESHOUSING4**.
 11. Type **choose=validate** within the parentheses containing **select=sbc** in the **MODEL** statement.
- Note:** Currently SAS Studio does not include the option to specify a separate data set as the test data.
12. In the editor, type **ref=first** as an option in the **CLASS** statement in order to treat the first level of each variable in the classification variable list as the reference level.
 13. Add the **choose=validate** option after the **selection=sbc** option to use the average square error of the validation data set as the model selection tool.
 14. Add in the **STORE** statement to save the analysis results in a SAS item store.

```
proc glmselect data=STAT1.AMESHOUSING3
               plots=( criterionpanel coefficientpanel)
               valdata=STAT1.AMESHOUSING4;
   class House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
        Season_Sold Garage_Type_2 Foundation_2 Heating_QC
        Masonry_Veneer Lot_Shape_2 Central_Air / param=glm
        ref=first;
   model SalePrice= Gr_Liv_Area Basement_Area Garage_Area
              Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
```

```

Total_Bathroom House_Style2 Overall_Qual2 Overall_Cond2
Fireplaces Season_Sold Garage_Type_2 Foundation_2
Heating_QC Masonry_Veneer Lot_Shape_2 Central_Air /
selection=backward
(select=sbc choose=validate) hierarchy=none;
store out=STAT1.amesstore;
run;

```

15. Submit the SAS code.

Note: Alternatively, you can write the code directly in SAS.

```

/*st106d01.sas*/

%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
    Lot_Area Age Sold Bedroom AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
    Season_Sold Garage_Type_2 Foundation_2 Heating_QC
    Masonry_Veneer Lot_Shape_2 Central_Air;

ods graphics;

proc glmselect data=STAT1.ameshousing3
    plots=all
    valdata=STAT1.ameshousing4;
class &categorical / param=glm ref=first;
model SalePrice= &interval &categorical /
    selection=backward
    select=sbc
    choose=validate;
store out=STAT1.amesstore;
title "Selecting the Best Model using Honest Assessment";
run;

```

Selected CLASS statement options:

PARAM=criterion Specifies the parameterization method for the classification variable or variables.
 In this case, GLM parameterization will be used, which is the same as that used in PROC GLM. One dummy variable is produced per level of the CLASS variable.

REF= Specifies the reference level for PARAM=EFFECT, PARAM=REFERENCE, PARAM=GLM, and their orthogonalizations. For PARAM=GLM, REF=FIRST causes the first level of the CLASS variable to be treated as the last level (the redundant level) and the parameter estimates of all other levels will be comparisons to that (reference) level.

Partial Output

Data Set	STAT1.AMESHOUING3
Validation Data Set	STAT1.AMESHOUING4
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	SBC

Stop Criterion	SBC
Choose Criterion	Validation ASE
Effect Hierarchy Enforced	None

Observation Profile for Analysis Data	
Number of Observations Read	300
Number of Observations Used	294
Number of Observations Used for Training	294

Observation Profile for Validation Data	
Number of Observations Read	300
Number of Observations Used	293

Class Level Information		
Class	Levels	Values
House_Style2	5	1Story 2Story SFoyer SLvl 1.5Fin
Overall_Qual2	3	5 6 4
Overall_Cond2	3	5 6 4
Fireplaces	3	1 2 0
Season_Sold	4	2 3 4 1
Garage_Type_2	3	Detached NAAttached
Foundation_2	3	Cinder Block Concrete/Slab Brick/Tile/Stone
Heating_QC	4	Fa Gd TA Ex
Masonry_Veneer	2	Y N
Lot_Shape_2	2	Regular Irregular
Central_Air	2	Y N

Dimensions	
Number of Effects	20
Number of Parameters	43

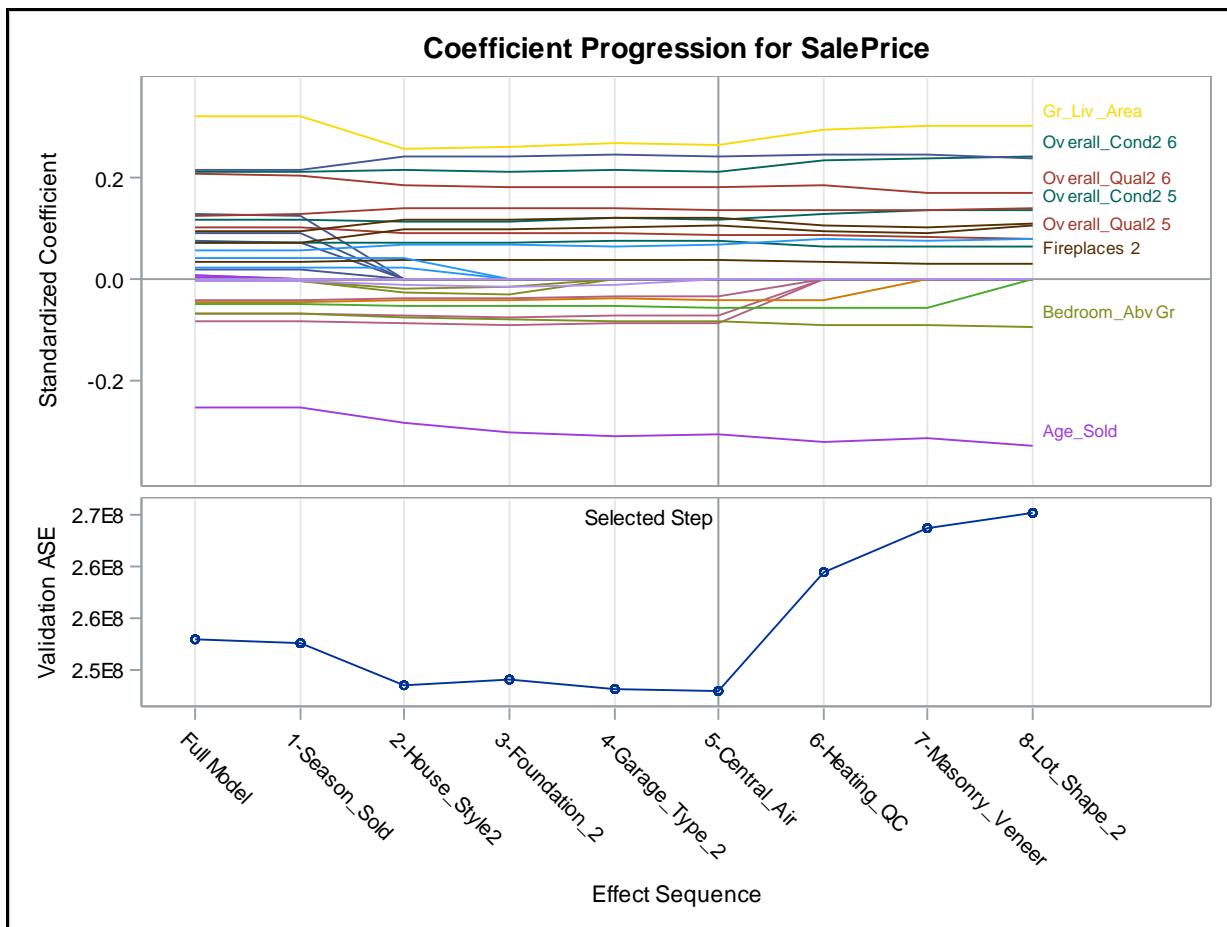
Backward Selection Summary						
Step	Effect Removed	Number Effects In	NumberParms In	SBC	ASE	Validation ASE
0		20	32	5779.6460	185773538	252878776
1	Season_Sold	19	29	5762.6753	185824120	252480746
2	House_Style2	18	25	5750.8247	192832172	248469026
3	Foundation_2	17	23	5740.3830	193440101	248951925

Backward Selection Summary						
Step	Effect Removed	Number Effects In	NumberParms In	SBC	ASE	Validation ASE
4	Garage_Type_2	16	21	5730.0735	194137231	247966687
5	Central_Air	15	20	5724.5490	194242334	247854963*
6	Heating_QC	14	17	5721.3123	203586891	259432895
7	Masonry_Veneer	13	16	5718.5873	205646000	263660934
8	Lot_Shape_2	12	15	5717.9317*	209193215	265159474

* Optimal Value Of Criterion

Selection stopped at a local minimum of the SBC criterion.

Stop Details			
Candidate For	Effect	Candidate SBC	Compare SBC
Removal	Deck_Porch_Area	5718.6683	> 5717.9317



The Coefficient Progression plot shows the history of the model selection. The vertical reference line at 5-Central_Air shows that the model at that step had the optimal level of Validation ASE (average squared error) on the validation data set compared with any other model in the progression.

...

Selected Model

The selected model, based on Validation ASE, is the model at Step 5.

Effects:	Intercept Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Masonry_Veneer Lot_Shape_2 Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom
-----------------	---

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	19	3.566452E11	18770797693	90.06
Error	274	57107246191	208420607	
Corrected Total	293	4.137524E11		

Root MSE	14437
Dependent Mean	137179
R-Square	0.8620
Adj R-Sq	0.8524
AIC	5946.87742
AICC	5950.27448
SBC	5724.54902
ASE (Train)	194242334
ASE (Validate)	247854963

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	51207	7079.121457	7.23
Gr_Liv_Area	1	42.972194	5.709351	7.53
Basement_Area	1	25.491273	3.170869	8.04
Garage_Area	1	29.698556	5.913131	5.02
Deck_Porch_Area	1	20.952561	7.235245	2.90
Lot_Area	1	1.199858	0.307660	3.90
Age_Sold	1	-422.187733	47.675825	-8.86
Bedroom_AbvGr	1	-4541.124997	1523.500120	-2.98

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Total_Bathroom	1	3806.351237	1714.333548	2.22
Overall_Qual2_5	1	6782.080263	3104.469941	2.18
Overall_Qual2_6	1	13659	3414.565419	4.00
Overall_Qual2_4	0	0	.	.
Overall_Cond2_5	1	8996.618020	4137.937302	2.17
Overall_Cond2_6	1	15909	4025.283609	3.95
Overall_Cond2_4	0	0	.	.
Fireplaces_1	1	9716.205925	2044.560791	4.75
Fireplaces_2	1	7235.661619	4540.159269	1.59
Fireplaces_0	0	0	.	.
Heating_QC_Fa	1	-11668	4315.812370	-2.70
Heating_QC_Gd	1	-3178.918390	2496.841385	-1.27
Heating_QC_TA	1	-6689.247126	2133.424223	-3.14
Heating_QC_Ex	0	0	.	.
Masonry_Veneer_Y	1	-3369.652622	2079.343731	-1.62
Masonry_Veneer_N	0	0	.	.
Lot_Shape_2-Regular	1	-4507.715447	2036.544994	-2.21
Lot_Shape_2-Irregular	0	0	.	.

End of Demonstration



Exercises

1. Predictive Model Building Using PROC GLMSELECT to Partition

Build a model predicting **SalePrice** starting with all of the variables used in the previous demonstration. Use the STORE statement to create an item store to use in subsequent exercises. Partition the **AmesHousing3** data set into a training data set of 200 and a validation data set of 100. Use stepwise selection with AIC as the selection criterion and validation average squared error for the model choice criterion.

End of Exercises

6.2 Scoring Predictive Models

Objectives

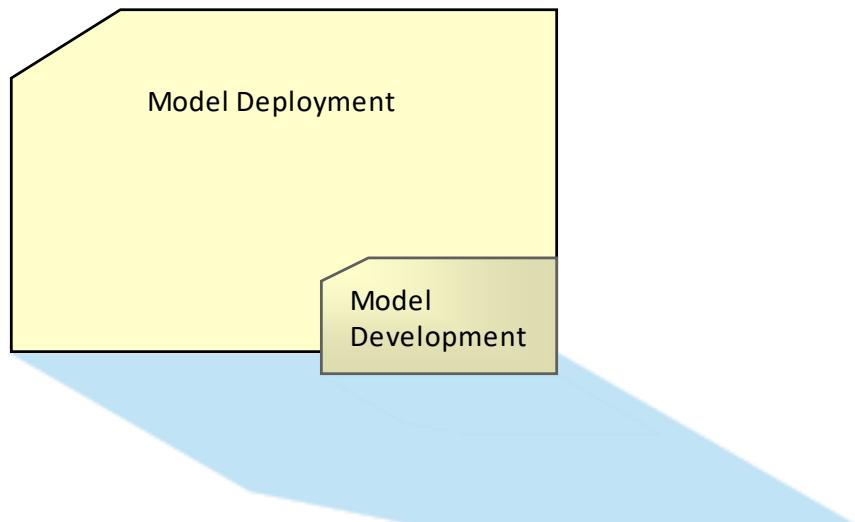
- Explain the concepts of scoring.
- Score new data using both PROC GLMSELECT and PROC PLM.

19

Copyright © SAS Institute Inc. All rights reserved.



Scoring



20

Copyright © SAS Institute Inc. All rights reserved.



The predictive modeling task is not completed when a model and allocation rule is determined. The model must be practically applied to new cases. This process is called *scoring*.

In database marketing, this process can be tremendously burdensome because the data to be scored might be many times more massive than the data used to develop the model. Moreover, the data might be stored in a different format on a different system using different software.

In other applications, such as fraud detection, the model might need to be integrated into an online monitoring system.

Scoring Recipe

- The model results in a formula or rules.
- The data require modifications.
 - Derived inputs
 - Transformations
 - Missing value imputation
- The scoring code is deployed.
 - To score, you do not rerun the algorithm; apply score code (equations) obtained from the final model to the scoring data.

Any modifications that you make to the training data (imputing missing values, transformations, standardization) should be applied to the validation and the scoring data in the same way. This means that if you have subtracted the mean of $x(\text{training})$ from the training data, then the mean of $x(\text{training})$ should also be subtracted from the validation and the scoring data. This practice keeps the different data sets comparable.

Ways to Score Using PROC GLMSELECT

There are several options:

- Use a SCORE statement in PROC GLMSELECT.
- Use a STORE statement in PROC GLMSELECT and then a SCORE statement in PROC PLM.
- Use a STORE statement in PROC GLMSELECT and then a CODE statement in PROC PLM to output SAS code and then score in a DATA step.

It might seem strange to go through two or three steps to score new data when there is a way to do it in one step. There is a SCORING tab in the Linear Regression task and the equivalent SCORE statement in PROC GLMSELECT enable you to produce a model and score data in one step. However, this method is inefficient if you will want to score more than once or model using a large data set. You can score with PROC PLM using the item store created in a STORE statement in PROC GLMSELECT. One potential problem with this method is that others might not be able to use this code with earlier versions of SAS or you might not want to share the entire item store. If so, you can produce detailed scoring code by checking the option to Add SAS scoring code to the log on the SCORING tab or using the CODE statement in PROC PLM.



Scoring Using PROC PLM and PROC GLMSELECT

Example: Use an item store created by PROC GLMSELECT to score a new data set in PROC PLM. Also, create SAS score code and use it in a DATA step to score a new data set. Compare the results using PROC COMPARE. Use the **AmesHousing4** data set for scoring.

```
/*st106d02.sas*/ /*Part A*/
proc plm restore=STAT1.amesstore;
  score data=STAT1.ameshousing4 out=scored;
  code file="&homefolder\scoring.sas";
run;
```

The CODE statement creates a file with SAS DATA step programming code.

Note: The macro variable &homefolder was created when you ran the data creation files at the beginning of the course. If the macro variable cannot be resolved, then re-run the data creation files.

In order to use the scoring code, you will need to write a DATA step and include the code within it.

```
data scored2;
  set STAT1.ameshousing4;
  %include "&homefolder\scoring.sas";
run;
```

The %include statement should reference the same file as that created in PROC PLM.

To check to see whether both methods scored the data set the same way, use PROC COMPARE.

```
proc compare base=scored compare=scored2 criterion=0.0001;
  var Predicted;
  with P_SalePrice;
run;
```

Selected PROC COMPARE statement option:

CRITERION= Specifies the criterion for judging the equality of numeric values. Normally, the value of γ is positive. In that case, the number itself becomes the equality criterion. If you use a negative value for γ , then PROC COMPARE uses an equality criterion proportional to the precision of the computer on which SAS is running.

Note: The default scored variable name from the SCORE statement in PROC PLM is Predicted. In the DATA step code the variable will have the same name as the target with “P_” prepended.

Partial Output

The COMPARE Procedure					
Comparison of WORK.SCORED with WORK.SCORED2					
(Method=RELATIVE(2.22E-10), Criterion=0.0001)					
Data Set Summary					
Dataset	Created	Modified	NVar	NObs	Label
WORK.SCORED	30AUG14:13:21:21	30AUG14:13:21:21	33	300	Scoring Results for DATA=STAT1.AMESHOUSING4
WORK.SCORED2	30AUG14:13:21:21	30AUG14:13:21:21	33	300	
Variables Summary					
Number of Variables in Common: 32.					
Number of Variables in WORK.SCORED but not in WORK.SCORED2: 1.					
Number of Variables in WORK.SCORED2 but not in WORK.SCORED: 1.					
Number of VAR Statement Variables: 1.					
Number of WITH Statement Variables: 1.					
Observation Summary					
Observation	Base	Compare			
First Obs	1	1			
Last Obs	300	300			
Number of Observations in Common: 300.					
Total Number of Observations Read from WORK.SCORED: 300.					
Total Number of Observations Read from WORK.SCORED2: 300.					
Number of Observations with Some Compared Variables Unequal: 0.					
Number of Observations with All Compared Variables Equal: 300.					
Values Comparison Summary					
Number of Variables Compared with All Observations Equal: 1.					
Number of Variables Compared with Some Observations Unequal: 0.					
Total Number of Values which Compare Unequal: 0.					
Total Number of Values not EXACTLY Equal: 296.					
Maximum Difference Criterion Value: 2.2837E-15.					

Use of the DATA step code can result in a small reduction in precision. However, the predictions are essentially the same.

End of Demonstration



Exercises

2. Scoring Using the SCORE Statement in PROC GLMSELECT

Rerun the code that you used from the previous exercise and add a STORE statement to create an item store. SCORE the **AmesHousing4** data using both a SCORE statement in PROC GLMSELECT and in PROC PLM. Be sure to use different names for the scored data sets produced. Use PROC COMPARE to compare the scoring results.

Hint: The scored variable produced in PROC GLMSELECT is named similarly to the scored variable produced with output DATAstep code.

End of Exercises

6.3 Solutions

Solutions to Exercises

1. Predictive Model Building Using PROC GLMSELECT to Partition

Build a model predicting **SalePrice** starting with all of the variables used in the previous demonstration. Partition the **AmesHousing3** data set into a training data set of approximately 2/3 and a validation data set of approximately 1/3. Use stepwise selection with AIC as the selection criterion and validation average squared error for the model choice criterion.

- a. Open the **Predictive Regression Models** task under Statistics.
- b. On the DATA tab, select the **AmesHousing3** data set and assign the variables.
- c. Select the **Validation data** option and specify the proportion of validation cases to 0.3333 and set the random seed as 8675309.
- d. Expand **Parameterization of Effects** and choose **Reference Coding**.
- e. On the MODEL tab, specify the model.
- f. On the SELECTION tab, specify using stepwise regression as the selection method and AIC as the selection criterion.
- g. Specify using average square error for validation data to select the best model.
- h. Expand the SELECTION PLOTS property and check the option to plot both the criteria plots and the coefficient plots.
- i. Click the EDIT button the manually edit the code. On the CLASS line, add the option **ref=first** after the **param=ref** option.
- j. Run the code.

Note: You can also write the code directly.

```
/*st106s01.sas*/
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
    Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House Style2 Overall Qual2 Overall Cond2 Fireplaces
    Season_Sold Garage_Type_2 Foundation_2 Heating_QC
    Masonry_Veneer Lot_Shape_2 Central_Air;

ods graphics;
proc glmselect data=STAT1.ameshousing3
    plots=all
    seed=8675309;
class &categorical / param=ref ref=first;
model SalePrice=&categorical &interval /
    selection=stepwise
    (select=aic
    choose=validate) hierarchy=single;
partition fraction(validate=0.3333);
title "Selecting the Best Model using Honest Assessment";
run;
```

Note: Your results will likely differ somewhat, depending on the seed that you choose.

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	AIC
Stop Criterion	AIC
Choose Criterion	Validation ASE
Effect Hierarchy Enforced	None
Random Number Seed	8675309

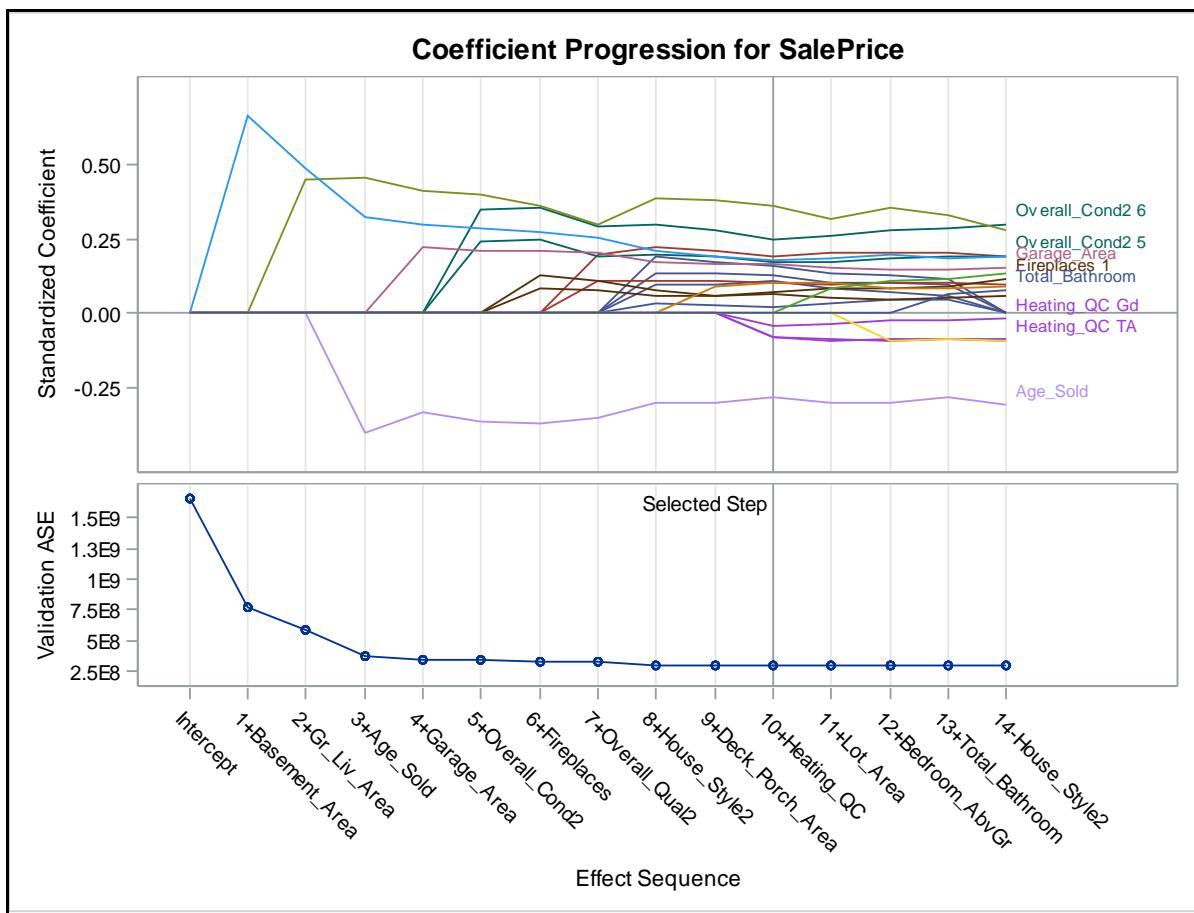
Number of Observations Read	300
Number of Observations Used	294
Number of Observations Used for Training	197
Number of Observations Used for Validation	97

Stepwise Selection Summary							
Step	Effect Entered	Effect Removed	Number Effects In	NumberParms In	AIC	ASE	Validation ASE
0	Intercept		1	1	4335.7651	1303938780	1656501303
1	Basement_Area		2	2	4222.6053	726746007	767937080
2	Gr_Liv_Area		3	3	4153.7335	507157741	590152215
3	Age_Sold		4	4	4070.6947	329360476	379123329
4	Garage_Area		5	5	4040.9787	280383339	349351979
5	Overall_Cond2		6	7	4017.8121	244265684	348031039
6	Fireplaces		7	9	4001.1755	219972414	328829426
7	Overall_Qual2		8	11	3991.0799	204782951	328466410
8	House_Style2		9	15	3981.7659	187553153	302046363
9	Deck_Porch_Area		10	16	3975.3902	179746298	298786920
10	Heating_QC		11	19	3971.6360	171063090	290197323*
11	Lot_Area		12	20	3966.5960	165057936	290656975
12	Bedroom_AbvGr		13	21	3961.0693	158870625	291293258
13	Total_Bathroom		14	22	3959.6794	156160207	292267671
14	House_Style2		13	18	3958.4479*	161618790	302466608

* Optimal Value Of Criterion

Selection stopped at a local minimum of the AIC criterion.

Stop Details			
Candidate For	Effect	Candidate AIC	Compare AIC
Entry	Masonry_Veneer	3959.1313	> 3958.4479
Removal	Total_Bathroom	3961.4810	> 3958.4479



The selected model, based on Validation ASE, is the model at Step 10.

Effects:	Intercept House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold
-----------------	---

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	18	2.231765E11	12398695049	65.49
Error	178	33699428801	189322634	
Corrected Total	196	2.568759E11		

Root MSE	13759
Dependent Mean	133582
R-Square	0.8688
Adj R-Sq	0.8555
AIC	3971.63597
AICC	3976.40870
SBC	3835.01684
ASE (Train)	171063090
ASE (Validate)	290197323

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	27334	10120	2.70
House_Style2 1Story	1	12267	4203.159135	2.92
House_Style2 2Story	1	2456.477699	4386.235156	0.56
House_Style2 SFoyer	1	20779	7050.033468	2.95
House_Style2 SLvl	1	17117	5527.649598	3.10
House_Style2 1.5Fin	0	0	.	.
Overall_Qual2 5	1	7841.596393	3417.138088	2.29
Overall_Qual2 6	1	14024	3806.928311	3.68
Overall_Qual2 4	0	0	.	.
Overall_Cond2 5	1	12475	4949.669709	2.52
Overall_Cond2 6	1	17766	4841.031305	3.67
Overall_Cond2 4	0	0	.	.
Fireplaces 1	1	5832.276234	2471.249968	2.36
Fireplaces 2	1	10886	4999.141012	2.18
Fireplaces 0	0	0	.	.
Heating_QC Fa	1	-13782	5544.767861	-2.49
Heating_QC Gd	1	-3687.706899	2867.792984	-1.29
Heating_QC TA	1	-5944.139856	2467.507946	-2.41
Heating_QC Ex	0	0	.	.
Gr_Liv_Area	1	54.360524	6.486247	8.38
Basement_Area	1	18.329197	3.964241	4.62
Garage_Area	1	33.820604	6.692579	5.05
Deck_Porch_Area	1	27.291527	8.243101	3.31
Age_Sold	1	-379.483707	54.640384	-6.95

2. Scoring Using the SCORE Statement in PROC GLMSELECT

Rerun the code that you used from the previous exercise and add a STORE statement to create an item store. SCORE the **AmesHousing4** data using both a SCORE statement in PROC GLMSELECT and in PROC PLM. Be sure to use different names for the scored data sets produced. Use PROC COMPARE to compare the scoring results.

- a. Use the editor to manually add in the STORE statement to generated code.

Note: Alternatively, you can write the code directly.

```

/*st106s02.sas*/

proc glmselect data=STAT1.ameshousing3
               seed=8675309
               noprint;
  class &categorical / param=ref ref=first;
  model SalePrice=&categorical &interval /
    selection=stepwise
    (select=aic
     choose=validate) hierarchy=single;
  partition fraction(validate=0.3333);
  score data=STAT1.ameshousing4 out=score1;
  store out=store1;
  title "Selecting the Best Model using Honest Assessment";
run;

proc plm restore=store1;
  score data=STAT1.ameshousing4 out=score2;
run;

proc compare base=score1 compare=score2 criterion=0.0001;
  var P_SalePrice;
  with Predicted;
run;

```

Note: Your results will likely differ somewhat, depending on the seed that you choose.

Partial Output

The COMPARE Procedure					
Comparison of WORK.SCORE1 with WORK.SCORE2					
(Method=RELATIVE(2.22E-10), Criterion=0.0001)					
Data Set Summary					
Dataset	Created	Modified	NVar	NObs	
WORK.SCORE1	30AUG14:13:21:21	30AUG14:13:21:21	33	300	
WORK.SCORE2	30AUG14:13:21:21	30AUG14:13:21:21	33	300	
Variables Summary					
Number of Variables in Common: 32.					
Number of Variables in WORK.SCORE1 but not in WORK.SCORE2: 1.					
Number of Variables in WORK.SCORE2 but not in WORK.SCORE1: 1.					
Number of VAR Statement Variables: 1.					
Number of WITH Statement Variables: 1.					

Observation Summary

	Observation	Base	Compare
First Obs		1	1
Last Obs		300	300

Number of Observations in Common: 300.

Total Number of Observations Read from WORK.SCORED: 300.

Total Number of Observations Read from WORK.SCORED2: 300.

Number of Observations with Some Compared Variables Unequal: 0.

Number of Observations with All Compared Variables Equal: 300.

Values Comparison Summary

Number of Variables Compared with All Observations Equal: 1.

Number of Variables Compared with Some Observations Unequal: 0.

Total Number of Values which Compare Unequal: 0.

Total Number of Values not EXACTLY Equal: 174.

Maximum Difference Criterion Value: 4.5574E-16.

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

6.01 Multiple Choice Poll – Correct Answer

When using honest assessment, which of the following would be considered the best model?

- a. The simplest model with the best performance on the training data
- b.** The simplest model with the best performance on the validation data
- c. The most complex model with the best performance on the training data
- d. The most complex model with the best performance on the validation data

Chapter 7 Categorical Data Analysis

7.1 Describing Categorical Data	7-3
Demonstration: Examining Distributions.....	7-10
7.2 Tests of Association.....	7-12
Demonstration: Chi-Square Test	7-12
Demonstration: Detecting Ordinal Associations	7-12
Exercises.....	7-12
7.3 Introduction to Logistic Regression	7-12
Demonstration: Simple Logistic Regression Model.....	7-12
Exercises.....	7-12
7.4 Logistic Regression with Categorical Predictors.....	7-12
Demonstration: Multiple Logistic Regression with Categorical Predictors.....	7-12
Exercises.....	7-12
7.5 Stepwise Selection with Interactions and Predictions	7-12
Demonstration: Logistic Regression: Backward Elimination with Interactions	7-12
Demonstration: Logistic Regression: Predictions Using PROC PLM	7-12
Exercises.....	7-12
7.6 Solutions	7-12
Solutions to Exercises	7-12
Solutions to Student Activities (Polls/Quizzes)	7-12

7.1 Describing Categorical Data

Objectives

- Examine the distribution of categorical variables.
- Do preliminary examinations of associations between variables.

3

Copyright © SAS Institute Inc. All rights reserved.

Examining Categorical Variables

By examining the distributions of categorical variables, you can do the following:

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

4

Copyright © SAS Institute Inc. All rights reserved.

Categorical Variables Association

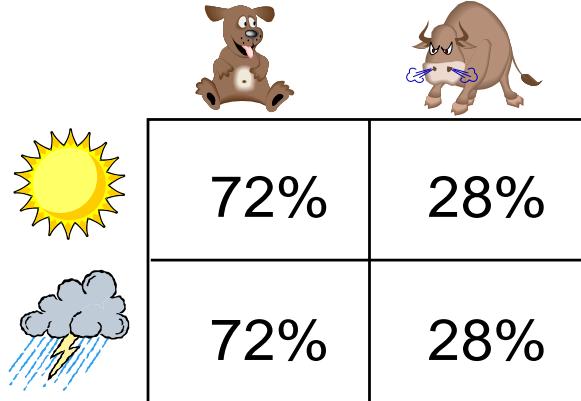
- An association exists between two categorical variables if the distribution of one variable changes when the level (or value) of the other variable changes.
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

5



Copyright © SAS Institute Inc. All rights reserved.

No Association



Is your manager's mood associated
with the weather?

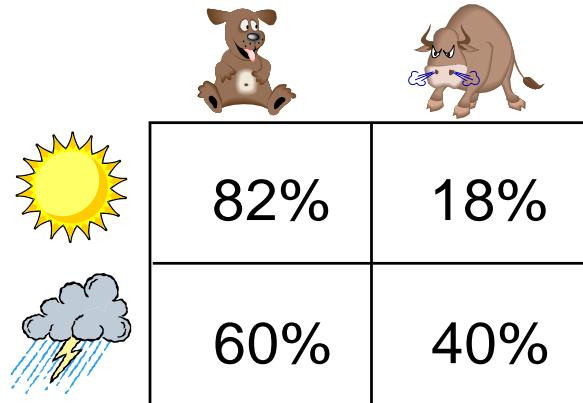
6



Copyright © SAS Institute Inc. All rights reserved.

There appears to be no association between your manager's mood and the weather here because the row percentages are the *same* in each column.

Association



Is your manager's mood associated with the weather?

7



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 occur 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

8



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 | 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}

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;
```

10

Copyright © SAS Institute Inc. All rights reserved.



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 one-way frequency tables, simply list the variable names individually. For two-way crosstabulation tables (first*second), the first variable represents the rows and the second variable represents the columns.

Note: PROC FREQ can generate large volumes of output as the number of variables or the number of variable levels (or both) increases.

Ames Housing Example – Bonus Eligible Sale

Realtors in Ames, Iowa receive the standard 3% commission on homes sales. One particular realty company offers an additional bonus for homes that sell for more than \$175,000. Are there attributes of the home that can predict whether it will be bonus eligible?



11

Copyright © SAS Institute Inc. All rights reserved.

The SAS logo, which consists of a stylized lowercase 's' followed by the word 'sas' in a serif font.

Example: The data are stored in the **STAT1.ameshousing3** data set.

These are the variables in the data set:

Bonus	bonus eligibility status (1= Bonus Eligible , 0= Not Bonus Eligible)
Fireplaces	number of fireplaces within house (0, 1, 2)
Lot_Shape_2	orientation of lot shape (Irregular , Regular)
Basement_Area	square footage of included basement
SalePrice	selling price of house (This variable is used to create the variable Bonus .)

7.01 Multiple Answer Poll

Which of the following would likely not be considered categorical in the data?

- a. Bonus
- b. Fireplaces
- c. Basement_Area
- d. Lot_Shape_2
- e. SalePrice



Examining Distributions

Example: Invoke PROC FREQ and create one-way frequency tables for the variables **Bonus**, **Fireplaces**, and **Lot_Shape_2** and create two-way frequency tables for the variables **Bonus** by **Fireplaces**, and **Bonus** by **Lot_Shape_2**. For the continuous variable, **Basement_Area**, create histograms for each level of **Bonus**. Use a CLASS statement in PROC UNIVARIATE.

One-way frequency tables

1. Open the **One-Way Frequencies** task under Statistics.
2. Select the **AmesHousing3** data set and assign the analysis variables.
3. Run the code.

Two-way frequency tables

4. Open the **Table Analysis** task under Statistics.
5. On the DATA tab, assign **Fireplaces** and **Lot_Shape_2** as the Row variables and **Bonus** as the Column variables.
6. On the OPTIONS tab, check the options to display **Cell**, **Row**, and **Column** percentages in the Frequency table. Uncheck the option to display **Chi-square statistics**.
7. Run the code.

Histograms for the continuous variable

8. Open the **Summary Statistics** task under Statistics.
9. Assign **Basement_Area** as the analysis variable and **Bonus** as the classification variable.
10. On the OPTIONS tab, uncheck the box to display Number of observations.
11. Expand the PLOTS property and select the option to display Histogram along with inset statistics.
12. Run the code.

Note: Alternatively, you can write the code directly and combine the three tasks and use the FORMAT procedure to format the values of **Bonus**.

```
/*st107d01.sas*/
title;
proc format;
  value bonusfmt 1="Bonus Eligible"
                0="Not Bonus Eligible"
                ;
run;
proc freq data=STAT1.ameshousing3;
  tables Bonus Fireplaces Lot_Shape_2
```

```

Fireplaces*Bonus Lot_Shape_2*Bonus /
plots(only)=freqplot(scale=percent);
format Bonus bonusfmt. ;
run;

proc univariate data=STAT1.ameshousing3 noint;
class Bonus;
var Basement_Area;
histogram Basement_Area;
inset mean std median min max / format=5.2 position=nw;
format Bonus bonusfmt. ;
run;

```

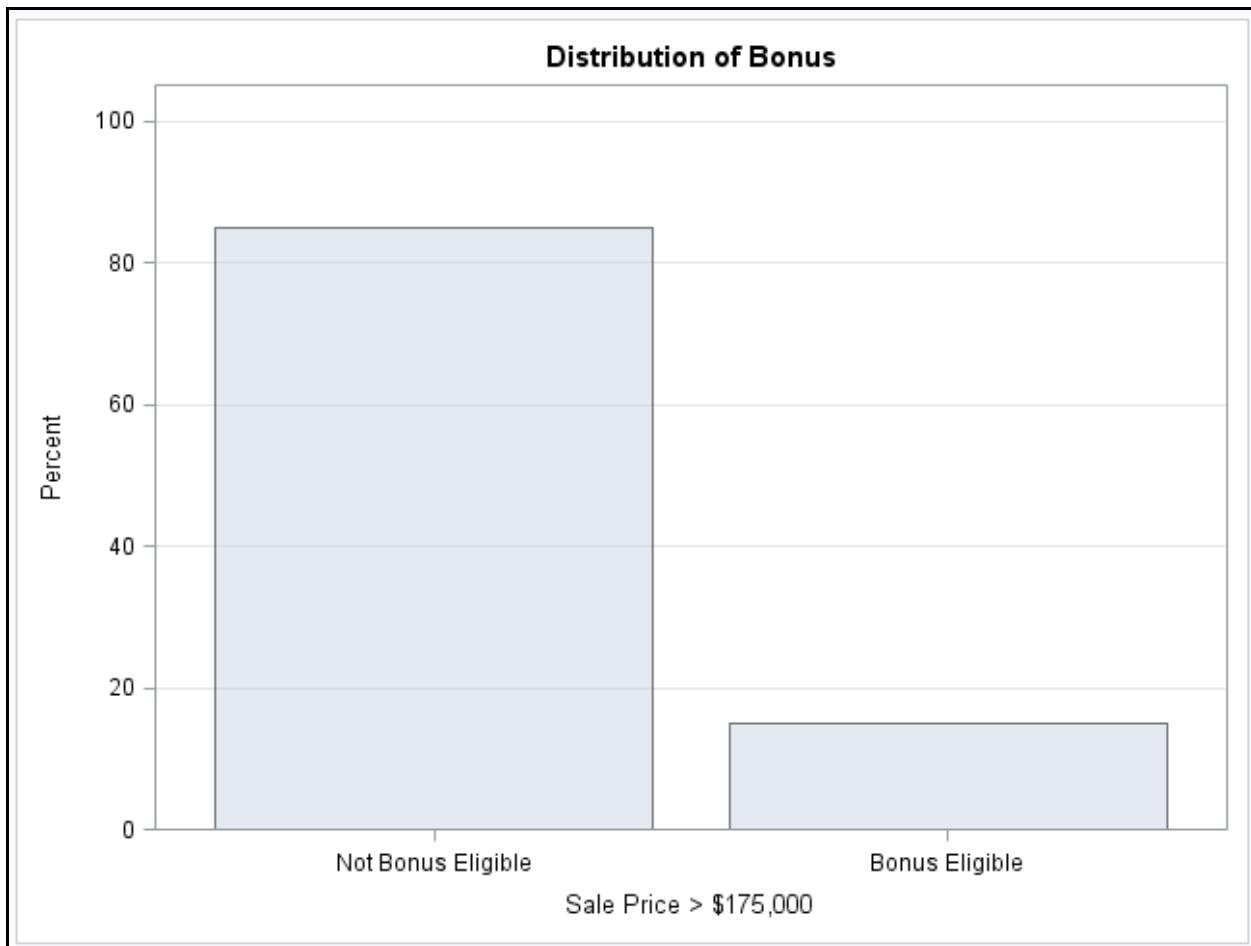
Selected TABLES statement PLOTS option and suboptions:

FREQPLOT(<suboptions>) requests a frequency plot. Frequency plots are available for frequency and crosstabulation tables. For multiway tables, PROC FREQ provides a two-way frequency plot for each stratum.

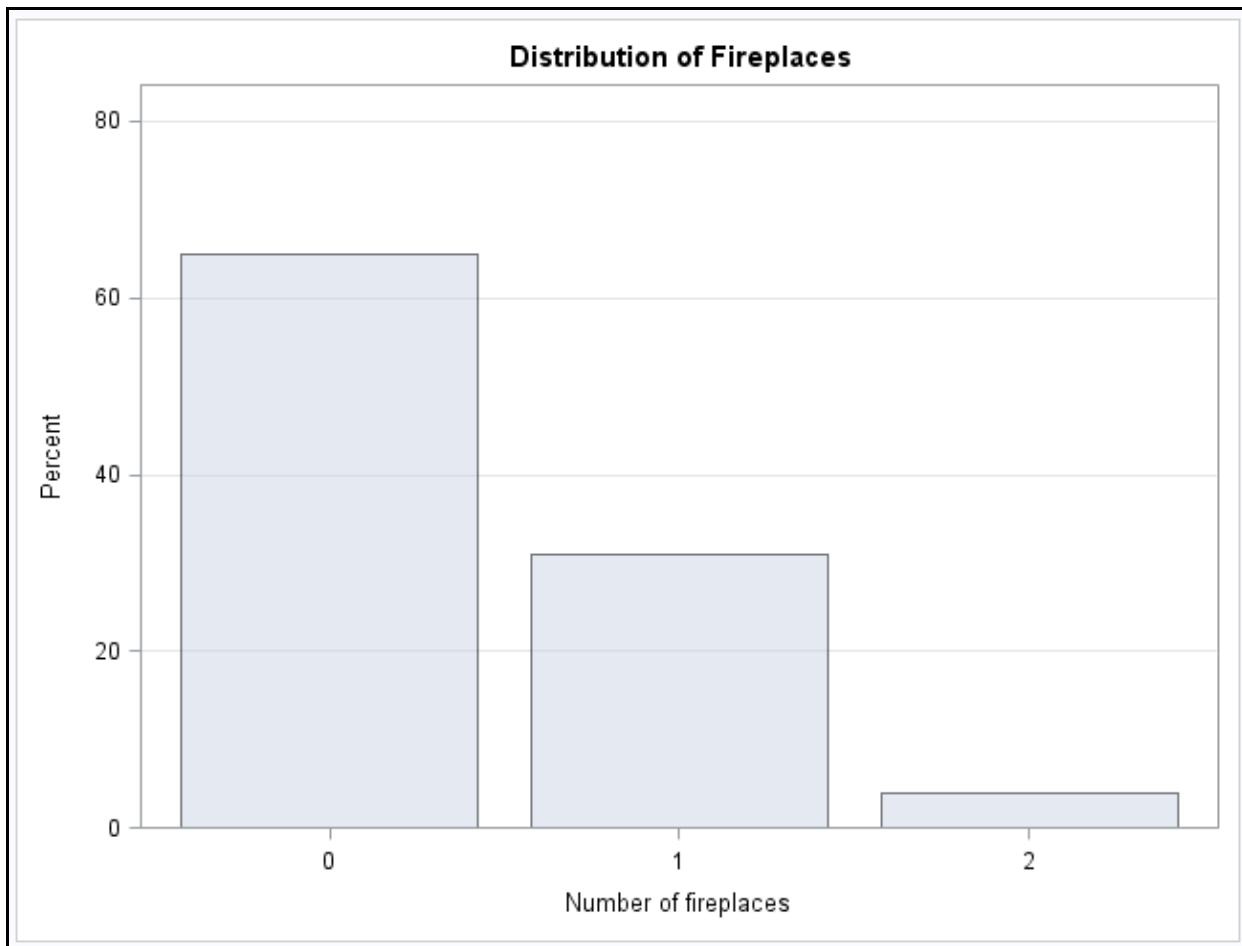
(SCALE=) specifies the scale of the frequencies to display. The default is SCALE=FREQ, which displays unscaled frequencies. SCALE=PERCENT displays percentages (relative frequencies).

PROC FREQ Output

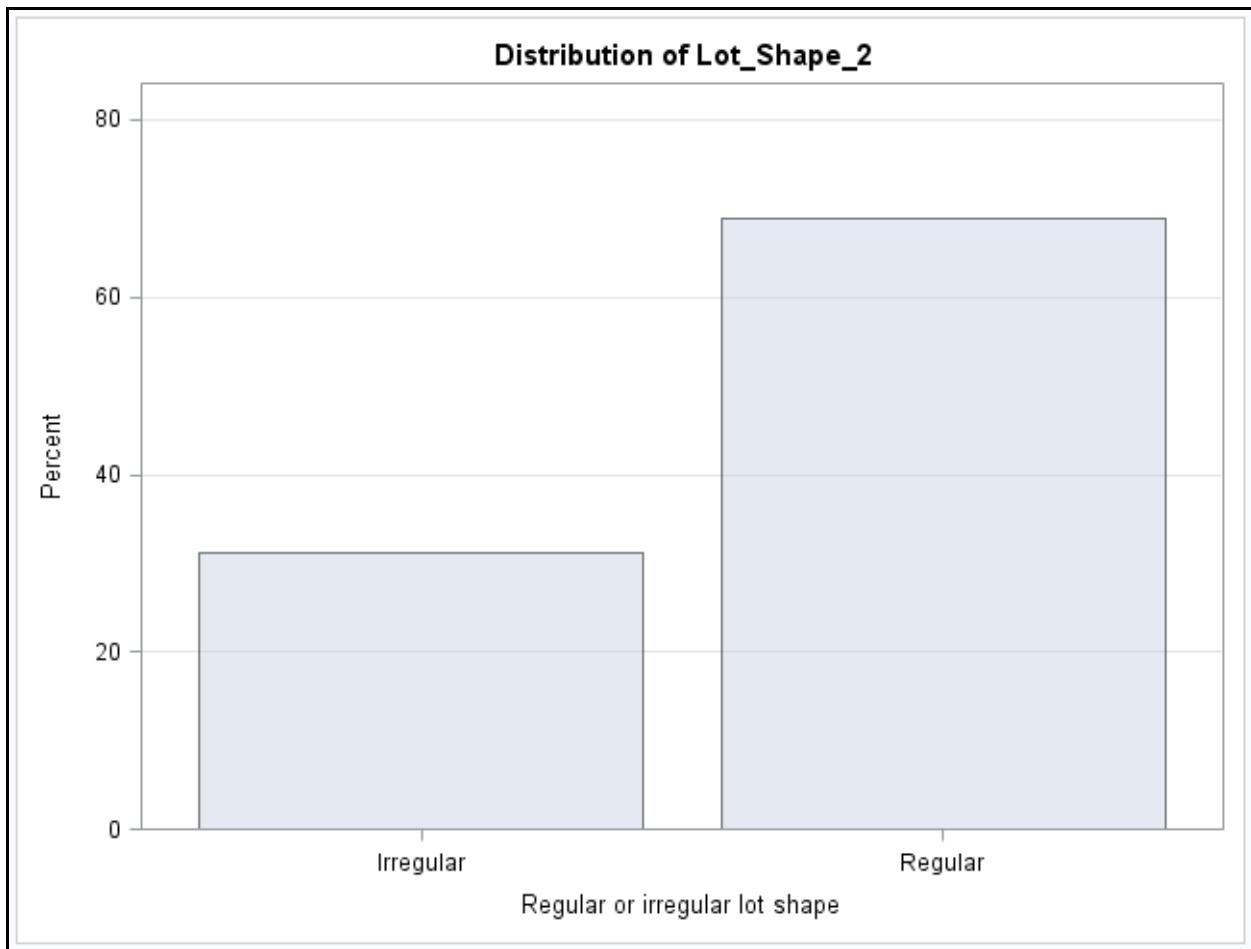
Sale Price > \$175,000				
Bonus	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Bonus Eligible	255	85.00	255	85.00
Bonus Eligible	45	15.00	300	100.00



Number of fireplaces					
Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
0	195	65.00	195	65.00	
1	93	31.00	288	96.00	
2	12	4.00	300	100.00	



Regular or irregular lot shape					
Lot_Shape_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
Irregular	93	31.10	93	31.10	
Regular	206	68.90	299	100.00	
Frequency Missing = 1					

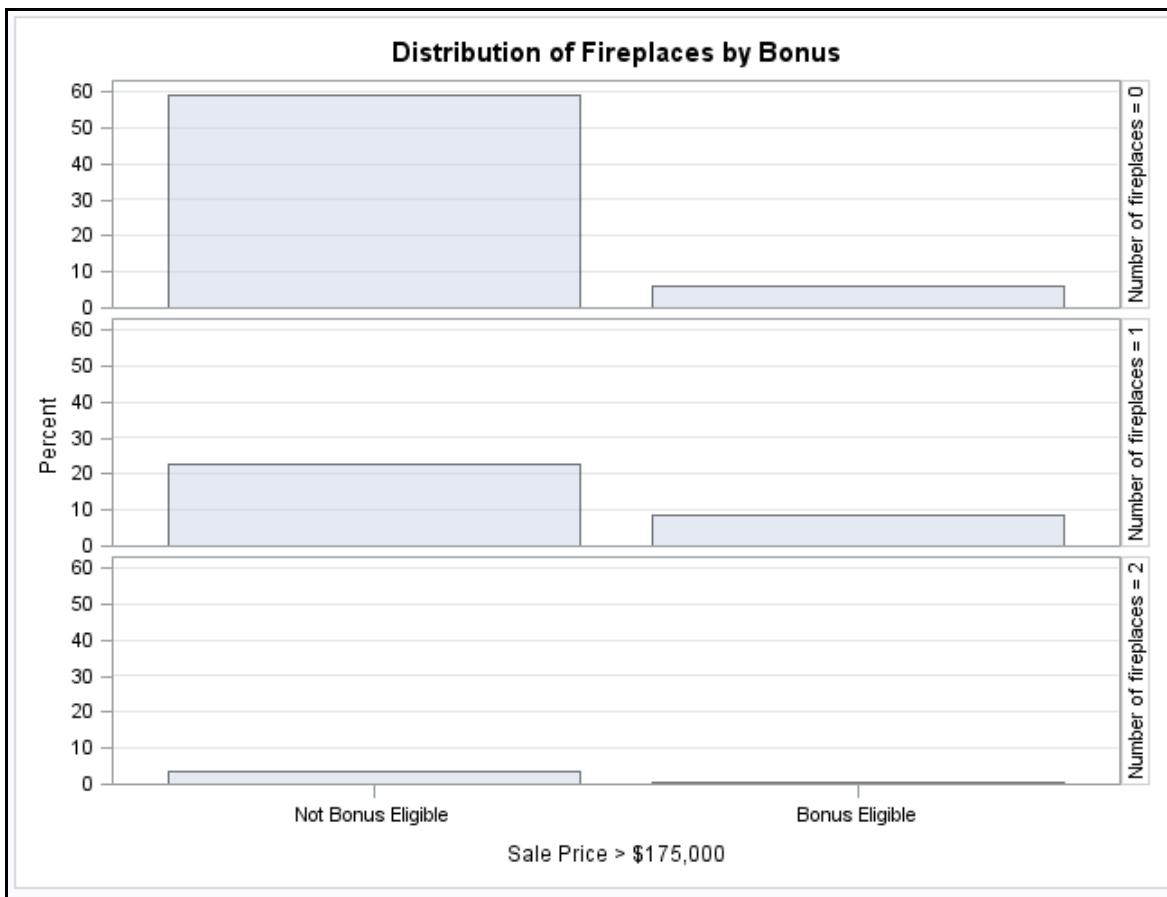


There seem to be no unusual data values that could be due to coding errors for any of the categorical variables.

The requested two-way frequency tables follow. You can get a preliminary idea whether there are associations between the outcome variable, **Bonus**, and the predictor variables, **Fireplaces** and **Lot_Shape_2**, by examining the distribution of **Bonus** at each value of the predictors.

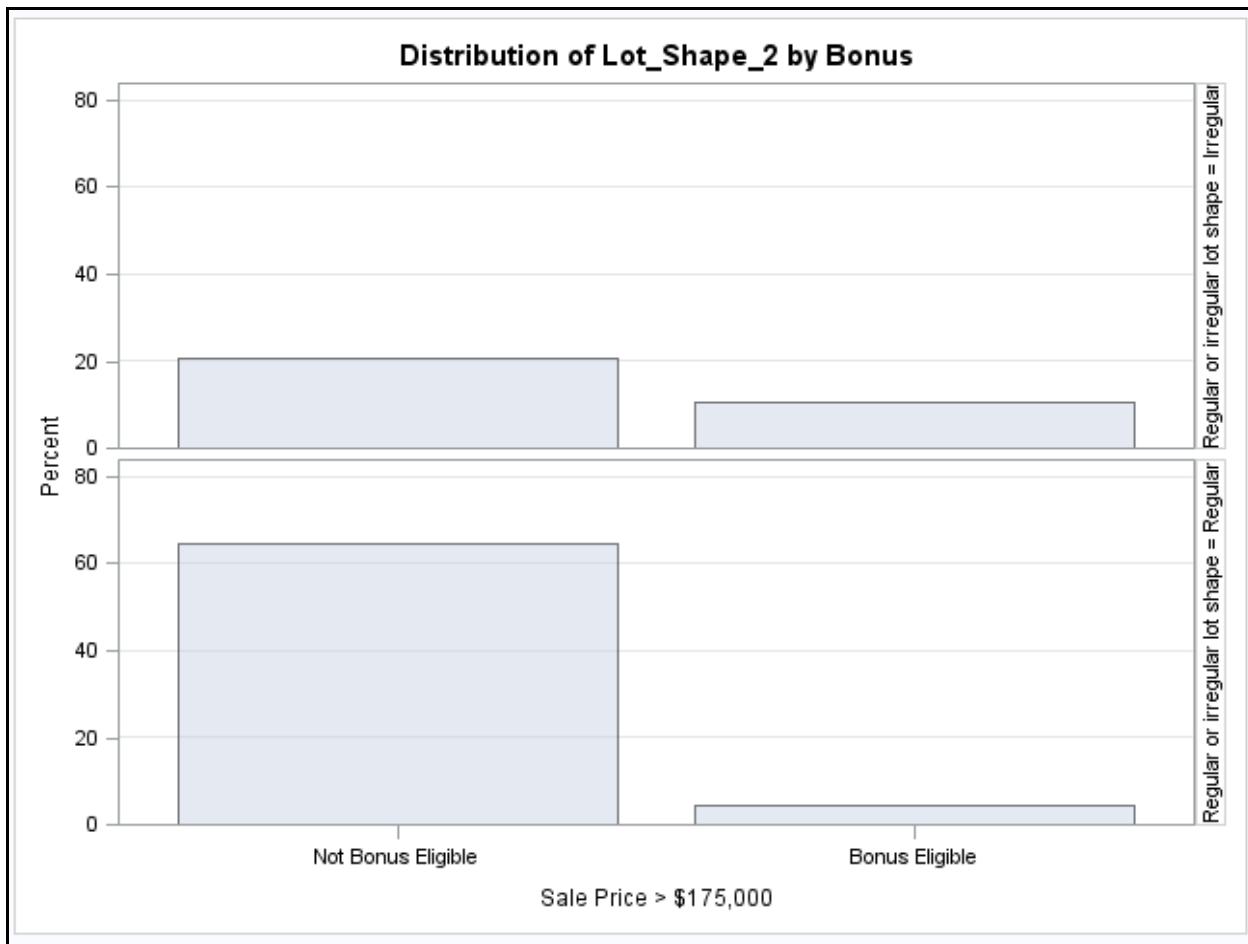
Note: The following output excludes the mosaic plots created as part of the SAS Studio Table Analysis task default output.

Table of Fireplaces by Bonus				
Fireplaces(Number of fireplaces)	Bonus(Sale Price > \$175,000)			
Frequency Percent Row Pct Col Pct	Not Bonus Eligible	Bonus Eligible	Total	
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00	
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00	
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00	
Total	255 85.00	45 15.00	300 100.00	



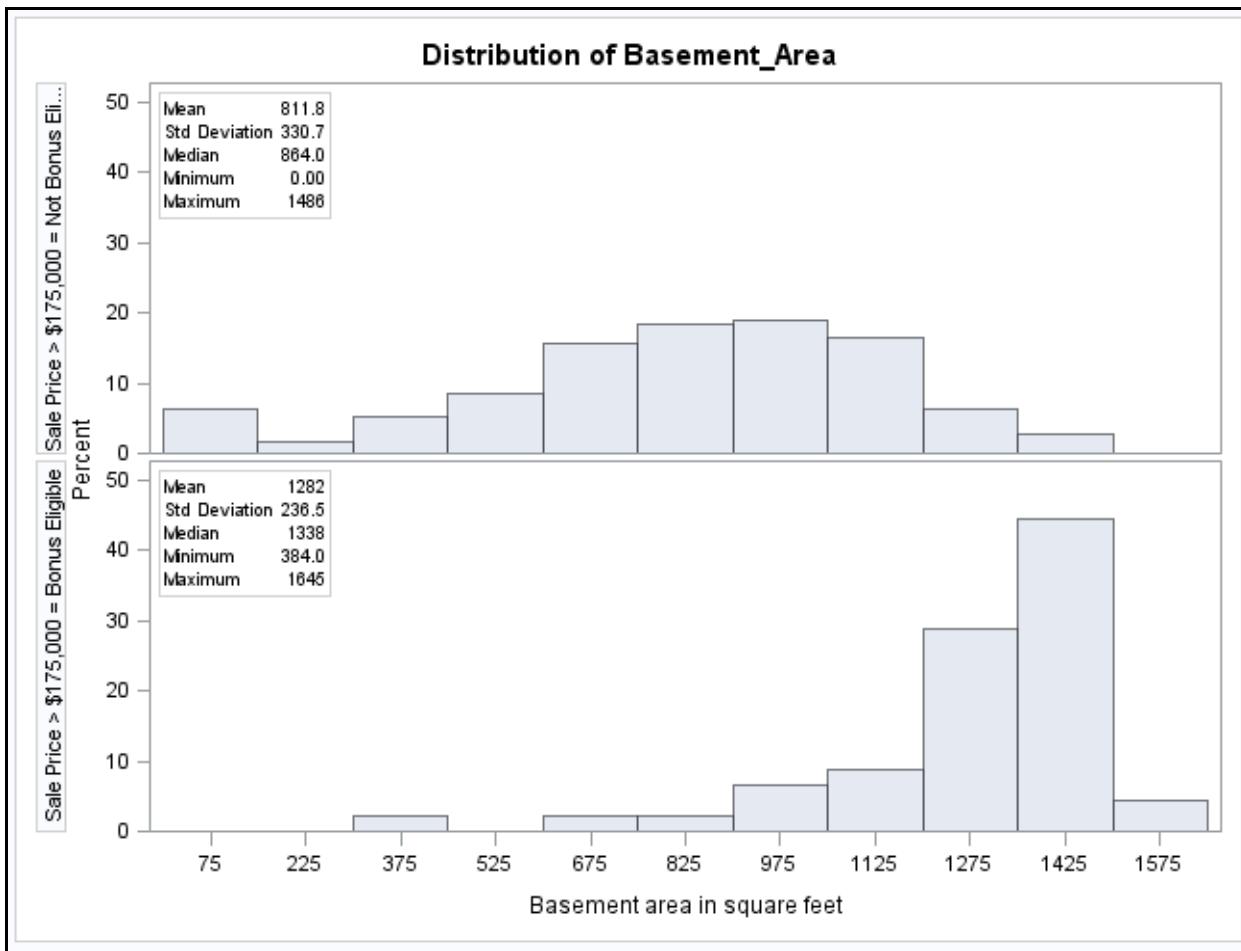
With the unequal group sizes, the row percentages might not easily display if **Fireplaces** is associated with **Bonus**.

Table of Lot_Shape_2 by Bonus			
Lot_Shape_2(Regular or irregular lot shape)	Bonus(Sale Price > \$175,000)		
	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 20.74 66.67 24.31	31 10.37 33.33 70.45	93 31.10
Regular	193 64.55 93.69 75.69	13 4.35 6.31 29.55	206 68.90
Total	255 85.28	44 14.72	299 100.00
Frequency Missing = 1			



There seems to be an association between **Bonus** and **Lot_Shape_2**, with a greater chance of not being bonus eligible when lot shape is regular.

The plot below shows the distribution of the continuous variable, **Basement_Area**, by bonus status. The distribution of houses that are not bonus eligible appears to be more variable as evident by the larger standard deviation. The mean of not bonus eligible houses is over 400 square feet smaller than houses that are bonus eligible.



End of Demonstration

7.2 Tests of Association

Objectives

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

17

Copyright © SAS Institute Inc. All rights reserved.



Overview

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

18

Copyright © SAS Institute Inc. All rights reserved.



Introduction

Table of Lot_Shape_2 by Bonus				
Lot_Shape_2	Bonus			
	Row Pct	Not Bonus Eligible	Bonus Eligible	Total
Irregular	66.67%	33.33%	N=93	
Regular	93.69%	6.31%	N=206	
Total	N=255	N=44	N=299	

19

Copyright © SAS Institute Inc. All rights reserved.



There appears to be an association between **Lot_Shape_2** and **Bonus** because the row probabilities are different in each column. To test for this association, you assess whether the difference between the probabilities of irregular lots being bonus eligible (33.33%) and regular lots being bonus eligible (6.31%) is greater than would be expected by chance.

Null Hypothesis

- There is no association between **Lot_Shape_2** and **Bonus**.
- The probability of a home sale being bonus eligible *is* the same regardless of lot shape.

Alternative Hypothesis

- There is an association between **Lot_Shape_2** and **Bonus**.
- The probability of a home sale being bonus eligible is not the same for irregular and regular lot shapes.

20

Copyright © SAS Institute Inc. All rights reserved.



Chi-Square Test

NO ASSOCIATION

observed frequencies=expected frequencies

ASSOCIATION

observed frequencies \neq expected frequencies

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

21

Copyright © SAS Institute Inc. All rights reserved.



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.

Note: Under the null hypothesis of no association between the Row and Column variables, 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 only that expected percentage times the total sample size. The expected count=(R/T)*(C/T)*T=(R*C)/T.

Chi-Square Tests

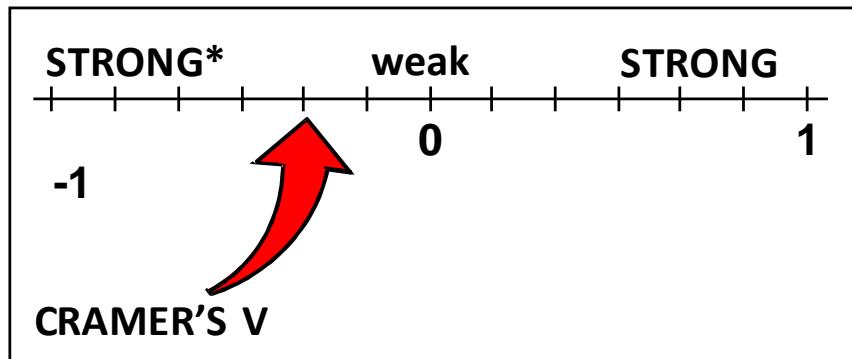
Chi-square tests and the corresponding p-values

- determine whether an association exists
- do not measure the strength of an association
- depend on and reflect the sample size.

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

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.

Measures of Association



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

One measure of the strength of the association between two nominal variables is Cramer's V statistic. It has a range of -1 to 1 for 2-by-2 tables and 0 to 1 for larger tables. Values farther from 0 indicate stronger association. Cramer's V statistic is derived from the Pearson chi-square statistic.

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 irregular lot shapes being bonus eligible compare to those of regular lot shapes?

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

24

Copyright © SAS Institute Inc. All rights reserved.



The odds ratio can be used as a measure of 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
	Yes	No	
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

$$\frac{\text{Total Yes outcomes in Group B}}{\text{Total outcomes in Group B}}$$

$$\text{Probability of a Yes in Group B} = 90 \div 100 = 0.9$$

25

Copyright © SAS Institute Inc. All rights reserved.



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

Probability versus Odds of an Outcome

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

$$\frac{\text{Probability of Yes in Group B} = 0.90}{\text{Probability of No in Group B} = 0.10} \quad \bullet \quad \bullet$$

$\text{Odds of Yes in Group B} = 0.90 : 0.10 = 9$



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, which indicate that you expect nine times as many occurrences as non-occurrences in group B.

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

Odds Ratio

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

$$\frac{\text{Odds of Yes in Group A}=3}{\text{Odds of Yes in Group B}=9} \cdot \cdot$$

$$\text{Odds Ratio, A to B} = 3 : 9 = 0.3333$$

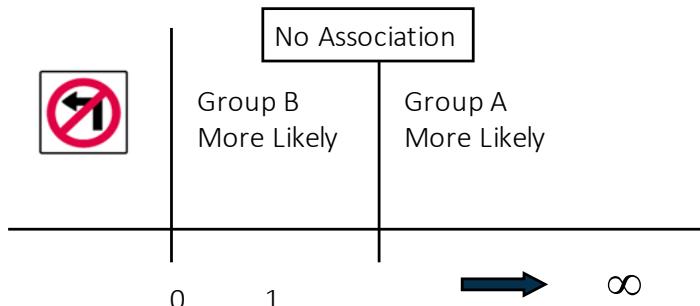
27



Copyright © SAS Institute Inc. All rights reserved.

The odds ratio of group A to group B equals 1/3, or 0.3333, which indicates that the odds of getting the outcome in group A are one third those in group B. If you were interested in the odds ratio of group B to group A, you would simply take the multiplicative inverse (or reciprocal) of 1/3 to arrive at 3.

Properties of the Odds Ratio, A to B



28



Copyright © SAS Institute Inc. All rights reserved.

The odds ratio shows the strength of the association between the predictor variable and the outcome variable. If the odds ratio is 1, then there is no association between the predictor variable and the outcome. If the odds ratio is greater than 1, then group A, the numerator group, is more likely to have the outcome. If the odds ratio is less than 1, then group B, the denominator group, is more likely to have the outcome.



Chi-Square Test

Example: Use the FREQ procedure to test for an association between the variables **Lot_Shape_2** and **Bonus** as well as **Fireplaces** and **Bonus**. Generate the expected cell frequencies and the cell's contribution to the total chi-square statistic.

1. Open the **Table Analysis** task under Statistics.
2. On the DATA tab, select the **AmesHousing3** data set.
3. Assign **Lot_Shape_2** and **Fireplaces** as the row variables and **Bonus** as the column variables.
4. On the OPTIONS tab, expand PLOTS and select the option to **suppress plots**.
5. Expand the FREQUENCY TABLE property and select the options to include observed frequencies, expected frequencies, row percentages, and cell's contribution to the total chi-square statistic.
6. Expand the STATISTICS property and select the option to display **Odds ratio and relative risk** in addition to Chi-square statistics.
7. Run the code.

Note: Alternative, the code below produces equivalent outputs.

```
/*st107d02.sas*/
ods graphics off;
proc freq data=STAT1.ameshousing3;
  tables (Lot Shape 2 Fireplaces)*Bonus
    / chisq expected cellchi2 nocol nopercnt
      relrisk;
  format Bonus bonusfmt. ;
  title 'Associations with Bonus';
run;
ods graphics on;
```

Selected TABLES statement options:

- | | |
|-----------|--|
| CHISQ | produces the chi-square test of association and the measures of association based on 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. |
| RELRISK | prints a table with risk ratios (probability ratios) and odds ratios. |

The frequency table is shown below.

Table of Lot_Shape_2 by Bonus			
Lot_Shape_2(Regular or irregular lot shape)	Bonus(Sale Price > \$175,000)		
	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 79.314 3.7797 66.67	31 13.686 21.905 33.33	93
Regular	193 175.69 1.7064 93.69	13 30.314 9.8893 6.31	206
Total	255	44	299
Frequency Missing = 1			

It appears that the cell for **Lot_Shape_2=Irregular** and **Bonus=1 (Bonus Eligible)** contributes the most to the chi-square statistic. The Cell Chi-Square value is 21.905.

Note: The cell chi-square is calculated using the formula
(observed frequency – expected frequency)²/expected frequency.

The overall chi-square statistic is calculated by adding the cell chi-square values over all rows and columns: $\sum \sum ((\text{observed}_{rc} - \text{expected}_{rc})^2 / \text{expected}_{rc})$.

Below is the table that shows the chi-square test and Cramer's V.

Statistic	DF	Value	Prob
Chi-Square	1	37.2807	<.0001
Likelihood Ratio Chi-Square	1	34.4226	<.0001
Continuity Adj. Chi-Square	1	35.1587	<.0001
Mantel-Haenszel Chi-Square	1	37.1561	<.0001
Phi Coefficient		-0.3531	
Contingency Coefficient		0.3330	
Cramer's V		-0.3531	

Because the *p*-value for the chi-square statistic is <.0001, which is below 0.05, you reject the null hypothesis at the 0.05 level and conclude that there is evidence of an association between **Lot_Shape_2** and **Bonus**. Cramer's V of -0.3531 indicates that the association detected with the chi-square test is relatively weak.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	62
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000

Fisher's Exact Test	
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Exact tests are often useful where asymptotic distributional assumptions are not met. The usual guidelines for the asymptotic chi-square test are generally 20-25 total observations for a 2*2 table, with 80% of the table cells having counts greater than 5. Fisher's Exact Test is provided by PROC FREQ when tests of association are requested for 2*2 tables. Otherwise, the exact test must be requested using an EXACT statement.

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.1347	0.0664	0.2735
Relative Risk (Column 1)	0.7116	0.6137	0.8251
Relative Risk (Column 2)	5.2821	2.9002	9.6202

The Odds Ratio and Relative Risk table shows another measure of strength of association.

The odds ratio is shown in the first row of the table, along with the 95% confidence limits. To interpret the odds ratio, refer to the contingency table at the beginning of the output. The top row (**Irregular**, in this case) is the numerator of the ratio while the bottom row (**Regular**) is the denominator. The interpretation is stated in relation to the left column of the contingency table (**Not Bonus Eligible**). The value of 0.1347 says that an irregular lot has about 13.5% of the odds of not being bonus eligible, compared with a regular lot. This is equivalent to saying that a regular lot has about 13.5% of the odds of being bonus eligible, compared with an irregular lot.

Relative Risk estimates for each column are interpreted as probability ratios, rather than odds ratios. You get a choice of assessing probabilities of the left column (Column1) or the right column (Column2). For example, the Column1 Relative Risk shows the ratio of the probabilities of irregular lots to regular lots being in the left column ($66.67/93.69=0.7116$).

It is often easier to report odds ratios by first transforming the decimal value to a percent difference value. The formula for doing that is $(OR-1) * 100$. In the example, you have $(0.1347-1)*100=-86.53\%$. In other words, regular lots have 86.53 percent lower odds of being bonus eligible compared with irregular lots.

The 95% odds ratio confidence interval goes from 0.0664 to 0.2735. That interval does not include 1. This confirms the statistically significant (at alpha=0.05) result of the Pearson chi-square test of association. A confidence interval that included the value 1 (equality of odds) would be a non-significant result.

Table of Fireplaces by Bonus			
Fireplaces(Number of fireplaces)	Bonus(Sale Price > \$175,000)		
Frequency Expected Cell Chi-Square Row Pct	Not Bonus Eligible	Bonus Eligible	Total
0	177 165.75 0.7636 90.77	18 29.25 4.3269 9.23	195
1	68 79.05 1.5446 73.12	25 13.95 8.7529 26.88	93
2	10 10.2 0.0039 83.33	2 1.8 0.0222 16.67	12
Total	255	45	300

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

There also seems to be an association between **Fireplaces** and **Bonus** (Chi-Square(2 df)=15.4141, p=0.0004). Cramer's V for that association is 0.2267.

Note: Mantel-Haenszel chi-square is a test of an ordinal association between **Fireplaces** and **Bonus**.

End of Demonstration

7.02 Multiple Answer Poll

What tends to happen when sample size decreases?

- a. The chi-square value increases.
- b. The p-value increases.
- c. Cramer's V increases.
- d. The Odds Ratio increases.
- e. The width of the CI for the Odds Ratio increases.

30

Copyright © SAS Institute Inc. All rights reserved.



Association among Ordinal Variables

Is



associated
with



?

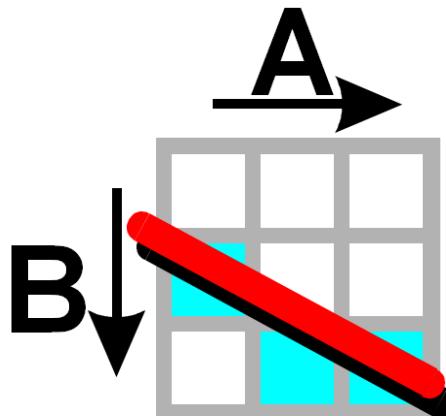
32

Copyright © SAS Institute Inc. All rights reserved.



You already saw that **Bonus** and **Fireplaces** have a significant general association. Another question that you can ask is whether **Bonus** and **Fireplaces** have a significant ordinal association. The appropriate test for ordinal associations is the Mantel-Haenszel chi-square test.

Mantel-Haenszel Chi-Square Test



Test Ordinal Association

33

sas

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. In order 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

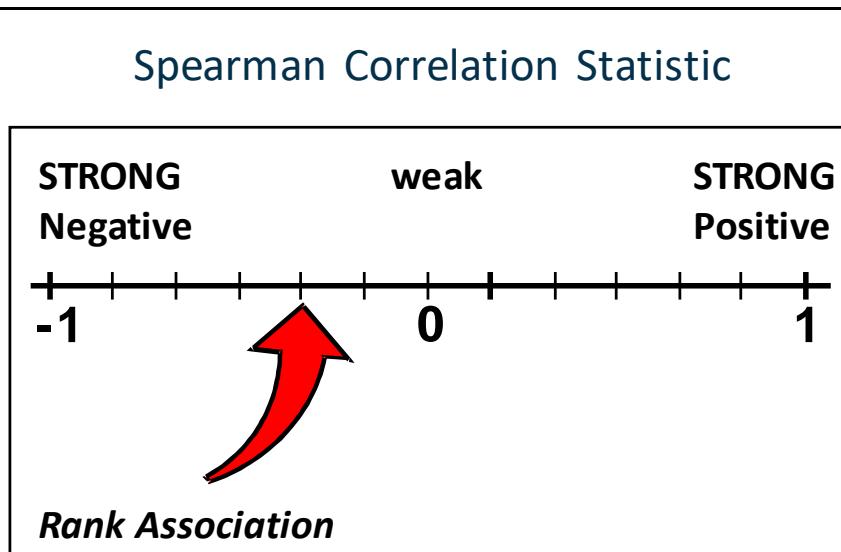
- Determines whether an ordinal association exists
- Does not measure the strength of the ordinal association
- Depends on and reflects the sample size

34

sas

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.



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.

The Spearman statistic can be interpreted as the Pearson correlation between the ranks on variable X and the ranks on variable Y.

For character values, SAS assigns, by default, a 1 to column 1, a 2 to column 2, and so on. You can change the default with the SCORES= option in the TABLES statement.



Detecting Ordinal Associations

Example: Use PROC FREQ to test whether an ordinal association exists between **Bonus** and **Fireplaces**.

1. Open the **Table Analysis** task under Statistics.
2. On the DATA tab, select the **AmesHousing3** data set.
3. Assign **Fireplaces** as the row variable and **Bonus** as the column variable.
4. On the OPTIONS tab, expand PLOTS property and suppress all plots.
5. Select the options to display cell, row, and column percentages.
6. Under STATISTICS, check the box to display **Measures of association** statistics.
7. To include confidence bounds in the table, open the editor and specify **cl** in the options list (after the slash /) in the TABLES statement.
8. Run the code.

The edited SAS code is shown below.

```
ods noproctitle;

proc freq data=STAT1.AMESHOUSING3;
  tables(Fireplaces) *(Bonus) / chisq measures nocum plots=none
  cl;
run;
```

Note: Alternatively, you can write the code directly.

```
/*st107d03.sas*/
ods graphics off;
proc freq data=STAT1.ameshousing3;
  tables Fireplaces*Bonus / chisq measures cl;
  format Bonus bonusfmt. ;
  title 'Ordinal Association between FIREPLACES and BONUS?';
run;
ods graphics on;
```

Selected TABLES statement options:

- | | |
|----------|--|
| CHISQ | produces the Pearson chi-square, the likelihood-ratio chi-square, and the Mantel-Haenszel chi-square. It also produces measures of association based on chi-square such as the phi coefficient, the contingency coefficient, and Cramer's V. |
| MEASURES | produces the Spearman correlation statistic along with other measures of association. |
| CL | produces confidence bounds for the MEASURES statistics. |

The crosstabulation is shown below.

Table of Fireplaces by Bonus			
Fireplaces(Number of fireplaces)	Bonus(Sale Price > \$175,000)		
Frequency	Not Bonus Eligible	Bonus Eligible	Total
Percent			
Row Pct			
Col Pct			
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00

The results of the Mantel-Haenszel chi-square test are shown below.

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Because the *p*-value of the Mantel-Haenszel chi-square is 0.0010, you can conclude at the 0.05 significance level that there is evidence of an ordinal association between **Bonus** and **Fireplaces**.

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

Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.4964	0.1111	0.2786	0.7143
Kendall's Tau-b	0.2072	0.0585	0.0926	0.3218
Stuart's Tau-c	0.1449	0.0433	0.0600	0.2298
Somers' D C R	0.1510	0.0451	0.0626	0.2395
Somers' D R C	0.2842	0.0786	0.1301	0.4383
Pearson Correlation	0.1896	0.0591	0.0737	0.3054
Spearman Correlation	0.2107	0.0594	0.0943	0.3272
Lambda Asymmetric C R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R C	0.0667	0.0603	0.0000	0.1849
Lambda Symmetric	0.0467	0.0424	0.0000	0.1298
Uncertainty Coefficient C R	0.0571	0.0298	0.0000	0.1156
Uncertainty Coefficient R C	0.0313	0.0167	0.0000	0.0640
Uncertainty Coefficient Symmetric	0.0404	0.0213	0.0000	0.0823

The Spearman Correlation (0.2107) indicates that there is a moderate, positive ordinal relationship between **Fireplaces** and **Bonus** (that is, as **Fireplaces** levels increase, **Bonus** tends to increase).

The ASE is the asymptotic standard error (0.0594), which is an appropriate measure of the standard error for larger samples.

Because the 95% confidence interval (0.0943, 0.3272) 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.

End of Demonstration



Exercises

1. Performing Tests and Measures of Association

An insurance company wants to relate the safety of vehicles to several other variables. A score is given to each vehicle model, using the frequency of insurance claims as a basis. The data are in the **STAT1.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 1000s of pounds
Size	trichotomized version of Type (1=Small or Sports, 2=Medium, 3=Large or Sport/Utility).

a. Invoke the FREQ procedure and create one-way frequency tables for the categorical variables.

1) What is the measurement scale of each variable?

<u>Variable</u>	<u>Measurement Scale</u>
Unsafe	_____
Type	_____
Region	_____
Weight	_____
Size	_____

2) What is the proportion of cars made in North America?

3) For the variables **Unsafe**, **Size**, **Region**, and **Type**, are there any unusual data values that warrant further investigation?

b. Use PROC FREQ to examine the crosstabulation of the variables **Region** by **Unsafe**. Generate a temporary format to clearly identify the values of **Unsafe**. Along with the default output, generate the expected frequencies, the chi-square test of association, and the odds ratio.

Use the following code for the format:

```
proc format;
  value safefmt 0='Average or Above'
                1='Below Average';
run;
```

- 1) For the cars made in Asia, what percentage had a below-average safety score?
 - 2) For the cars with an average or above safety score, what percentage was made in North America?
 - 3) Do you see a statistically significant (at the 0.05 level) association between **Region** and **Unsafe**?
 - 4) What does the odds ratio compare and what does this one say about the difference in odds between Asian and North American cars?
- c. Use the variable named **Size**. Examine the ordinal association between **Size** and **Unsafe**. Use PROC FREQ.
- 1) What statistic should you use to detect an ordinal association between **Size** and **Unsafe**?
 - 2) Do you reject or fail to reject the null hypothesis at the 0.05 level?
 - 3) What is the strength of the ordinal association between **Size** and **Unsafe**?
 - 4) What is the 95% confidence interval around that statistic?

End of Exercises

7.03 Multiple Answer Poll

A researcher wants to measure the strength of an association between two binary variables. Which statistic(s) can he use?

- a. Hansel and Gretel Correlation
- b. Mantel-Haenszel Chi-Square
- c. Pearson Chi-Square
- d. Odds Ratio
- e. Spearman Correlation

40

Copyright © SAS Institute Inc. All rights reserved.

7.3 Introduction to Logistic Regression

Objectives

- Define the concepts of logistic regression.
- Fit a binary logistic regression model using the LOGISTIC procedure.
- Describe the standard output from the LOGISTIC procedure with one continuous predictor variable.
- Read and interpret odds ratio tables and plots.

44

Copyright © SAS Institute Inc. All rights reserved.

Overview

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

45

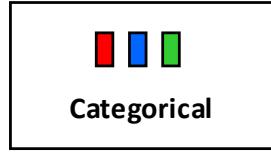


Overview

Response



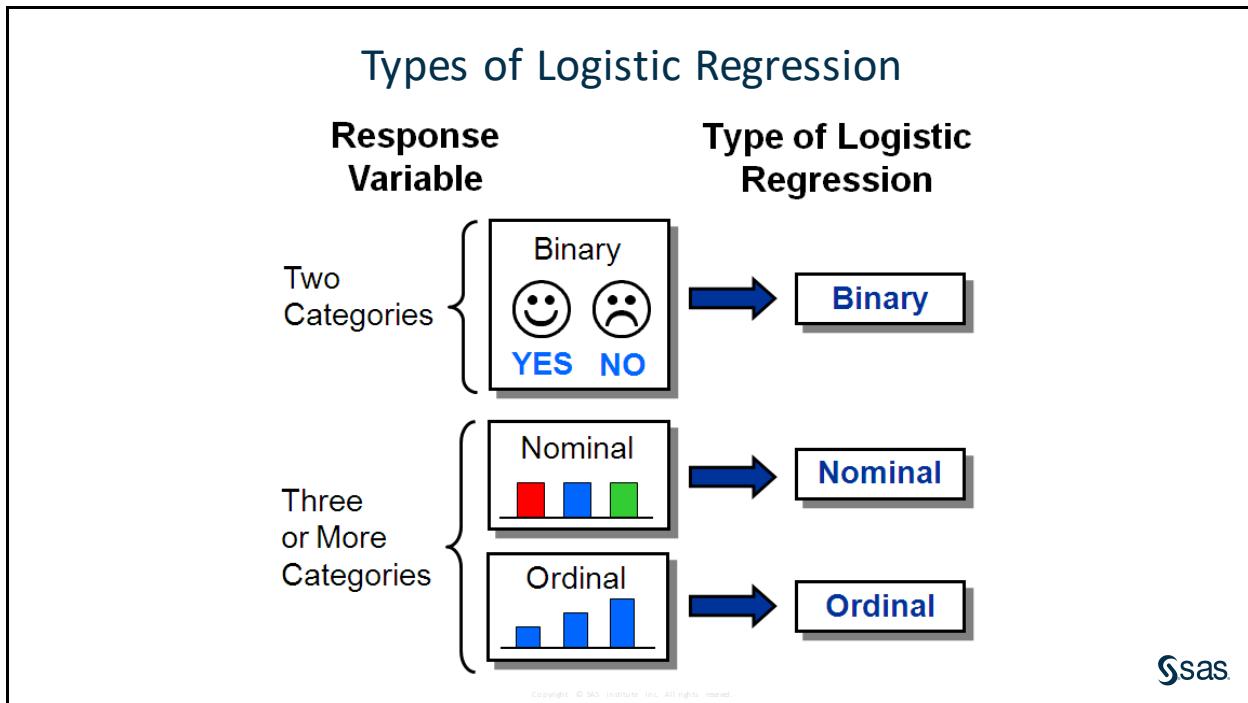
Analysis



46



Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables. In *linear regression*, the response variable is continuous. In *logistic regression*, the response variable is categorical.



If the response variable is dichotomous (two categories), the appropriate logistic regression model is binary logistic regression.

If you have more than two categories (levels) within the response variable, then there are two possible logistic regression models:

1. If the response variable is nominal, you fit a nominal logistic regression model.
2. If the response variable is ordinal, you fit an ordinal logistic regression model.

Why Not Ordinary Least Squares Regression?

- 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?

$$\text{OLS Regression: } Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

48

Copyright © SAS Institute Inc. All rights reserved.



You might be tempted to analyze a regression model with a binary response variable using PROC GLMSELECT, PROC REG or PROC GLM. 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 regard 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?

- 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?

$$\text{Linear Probability Model: } p_i = \beta_0 + \beta_1 X_{1i}$$

49

Copyright © SAS Institute Inc. All rights reserved.



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

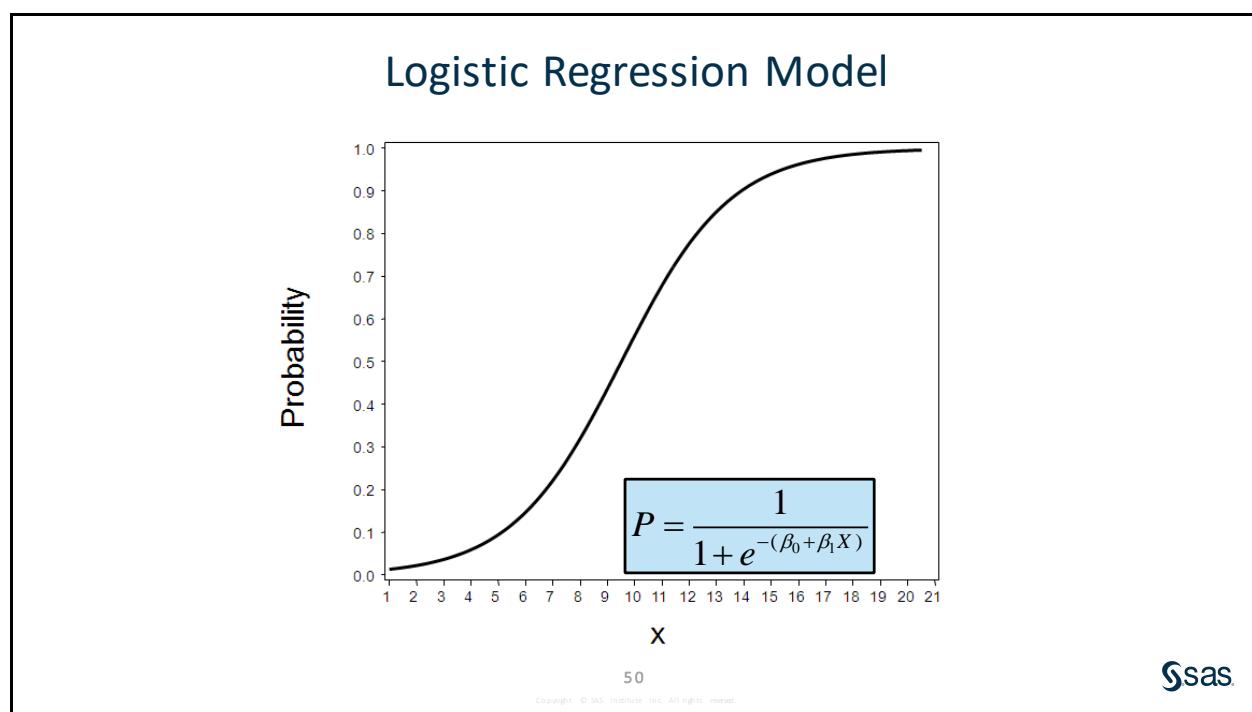
One problem is that the predicted values from a linear model can assume, theoretically, any value. However, probabilities are by definition bounded between 0 and 1.

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.

Note: As mentioned above, probabilities have a binomial error of the form $p^*(1-p) = (p-p^2)$. Taking the derivative of this expression with respect to p yields the expression $1-2*p$. Setting the derivative equal to zero and solving for p returns a value of 0.5. This binomial error equation is a downward facing parabola, which means that the greatest value is at 0.5 and lowest values are near 0 and 1.

Finally, there is no such thing as an “observed probability” and therefore least squares methods cannot be used. The response variable is always either 0 or 1 and therefore the probability of the event is either 0% or 100%. This is another reason why it is untenable to assume a normal distribution of error.



This plot shows a model of the relationship between a continuous predictor and the probability of an event or outcome. The linear model clearly does 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 S-shape of the function is known as a *sigmoid*.

The rate of change parameter of this function (β_1) determines the rate of increase or decrease of the curve. When the parameter value is greater than 0, the probability of the outcome increases as the predictor variable values increase. When the parameter is less than 0, the probability decreases as the predictor variable values increase. As the absolute value of the parameter increases, the curve has a steeper rate of change. When the parameter value 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 for this model cannot be estimated in PROC GLMSELECT, PROC REG, or PROC GLM because this is not a linear model.

Logit Transformation

Logistic regression models transformed 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 that the event (for example, a sale) occurs in the i^{th} case

\ln is the natural log (to the base e).

* The logit is the natural log of the odds.

51



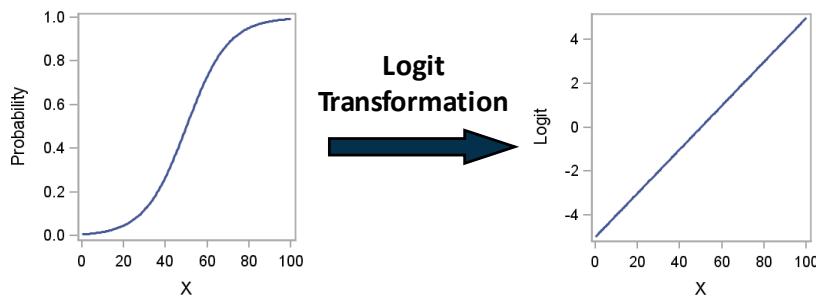
A logistic regression model applies a logit transformation to the probabilities. Two of the problems that you saw with modeling the probability directly were that probabilities were bounded between 0 and 1, and that there was not likely a straight line relationship between predictors and probabilities.

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 $\ln(p/(1-p))$ goes to infinity. As p approaches its minimum value of 0, $p/(1-p)$ approaches 0. The natural log of something approaching 0 is something that goes to negative infinity. So, the logit has no upper or lower bounds.

If you can model the logit, then simple algebra enables 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. The odds and odds ratios were discussed in a previous section. This relationship between the odds and the logit will become important later in this section.

Assumption



52



Note: Assumption in logistic regression: The logit has a linear relationship with the predictor variables.

If the hypothesized nature of the direct relationship between X and p are correct, then the logit has a linear relationship with X through the parameters. In other words, a linear function of X , additive in relation to the parameters, 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 the appendix.)

Logistic Regression Model

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$$

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

53



For a binary response variable, the linear logistic model with one predictor variable has the form above.

Unlike linear regression, the logit is not normally distributed and the variance is not constant. Therefore, logistic regression requires a more computationally complex estimation method, named the *Method of Maximum Likelihood*, to estimate the parameters. This method finds the values of the parameters that make the observed data most likely. This is accomplished by maximizing the *likelihood function* that expresses the probability of the observed data as a function of the unknown parameters.

7.04 Multiple Choice Poll

What are the upper and lower bounds for a logit?

- a. Lower=0, Upper=1
- b. Lower=0, No upper bound
- c. No lower bound, No upper bound
- d. No lower bound, Upper=1

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

54

Copyright © SAS Institute Inc. All rights reserved.



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;
```

56

Copyright © SAS Institute Inc. All rights reserved.



Selected LOGISTIC procedure statements:

- CLASS** names the classification variables to be used in the analysis. The CLASS statement must precede the MODEL statement. By default, these variables will be analyzed using effects coding parameterization. This can be changed with the PARAM= option.
- 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 (for example, 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.



Simple Logistic Regression Model

Example: Fit a binary logistic regression model in PROC LOGISTIC. Select **Bonus** as the outcome variable and **Basement_Area** as the predictor variable. Use the EVENT= option to model the probability of being bonus eligible and request profile likelihood confidence intervals around the estimated odds ratios.

1. Open the **Binary Logistic Regression** task under Statistics.
2. On the DATA tab, select the **AmesHousing3** data set.
3. Assign **Bonus** as the response variable and use the drop-down menu to specify **1** as the event of interest.
4. Assign **Basement_Area** as the continuous variable under Explanatory Variables.
5. On the MODEL tab, use the radio buttons and select **Main effects model**.
6. On the OPTIONS tab, select the option to display default and additional statistics under the Select statistics to display area of the STATISTICS list.
7. Expand the Parameter Estimates property. Use the drop-down menu to request the profile likelihood confidence intervals for odds ratios.
8. Under PLOTS and within the Select the options to display section, choose **Default and Additional Plots** and specify to include the **Effect plot** and the **Odds ratio plots**.
9. Run the code.

Note: Alternatively, you can write the code directly in SAS.

```
/*st107d04.sas*/
ods graphics on;
proc logistic data=STAT1.ameshousing3 alpha=.05
            plots(only)=(effect oddsratio);
model Bonus(event='1')=Basement_Area / clodds=pl;
title 'LOGISTIC MODEL (1):Bonus=Basement_Area';
run;
```

Selected PLOTS options:

- EFFECT requests a plot of the predicted probability on the Y axis by the predictor on the X axis. If there is more than one predictor variable in the model, the partial effect plot can be requested using the option (X=<variable>).
- ODDSRATIO requests a plot of the odds ratios, along with its (1-ALPHA) confidence limits. The width of the confidence limits can be changed from the default of 95% using an ALPHA= option in the PROC LOGISTIC statement. The chosen alpha level applies to all confidence intervals produced in all tables and plots in that run of PROC LOGISTIC.

Selected MODEL statement options:

- (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 quotation marks 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.
- 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.

SAS Output

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > \$175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

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 Optimization Technique is the iterative numerical technique that PROC LOGISTIC uses to estimate the model parameters.

The model is assumed to be “binary logit” when there are exactly two response levels.

Number of Observations Read	300
Number of Observations Used	300

The Number of Observations Used is the count of all observations that are nonmissing for all variables specified in the MODEL statement.

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	45

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 being bonus eligible (**Bonus=1**). The Response Profile table also shows frequencies of response values.

Probability modeled is Bonus=1.

It is advisable to check that the modeled response level is the one that you intended.

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

The Model Convergence Status simply informs you that the convergence criterion was met. There are a number of options to control the convergence criterion.

The optimization technique does not always converge to a maximum likelihood solution. When this is the case, the output after this point cannot be trusted. Always check to see that the Convergence criterion is satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	255.625	161.838
SC	259.329	169.246
-2 Log L	253.625	157.838

The Model Fit Statistics provides three measures:

- AIC is Akaike's 'A' information criterion.
- SC is the Schwarz criterion.
- -2 Log L is -2 times the natural log of the likelihood.

$-2 \log L$, AIC, and SC are goodness-of-fit measures that you can use to compare one model to another. **These statistics measure relative fit among models, but they do not measure absolute fit of any single model.** Smaller values for all of these measures indicate better fit. However, $-2 \log L$ can be reduced by simply adding more regression parameters to the model. Therefore, it is not used to compare the fit of models that use different numbers of parameters except for comparisons of nested models via likelihood ratio tests. 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.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	95.7870	1	<.0001
Score	65.5624	1	<.0001
Wald	48.0617	1	<.0001

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.

A significant p -value for these tests provides evidence that at least one of the regression coefficients for an explanatory variable is significantly different from 0. In this way, they are similar to the overall F test in linear regression. The Likelihood Ratio Chi-Square is calculated as the difference between the $-2 \log L$ value of the baseline model (Intercept Only) and the $-2 \log L$ value of the hypothesized model (Intercept and Covariates). The statistic is a distributed asymptotically chi-square with degrees of freedom equal to the difference in number of parameters between the hypothesized model and the baseline model. 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). All three tests are asymptotically equivalent and often give very similar values.

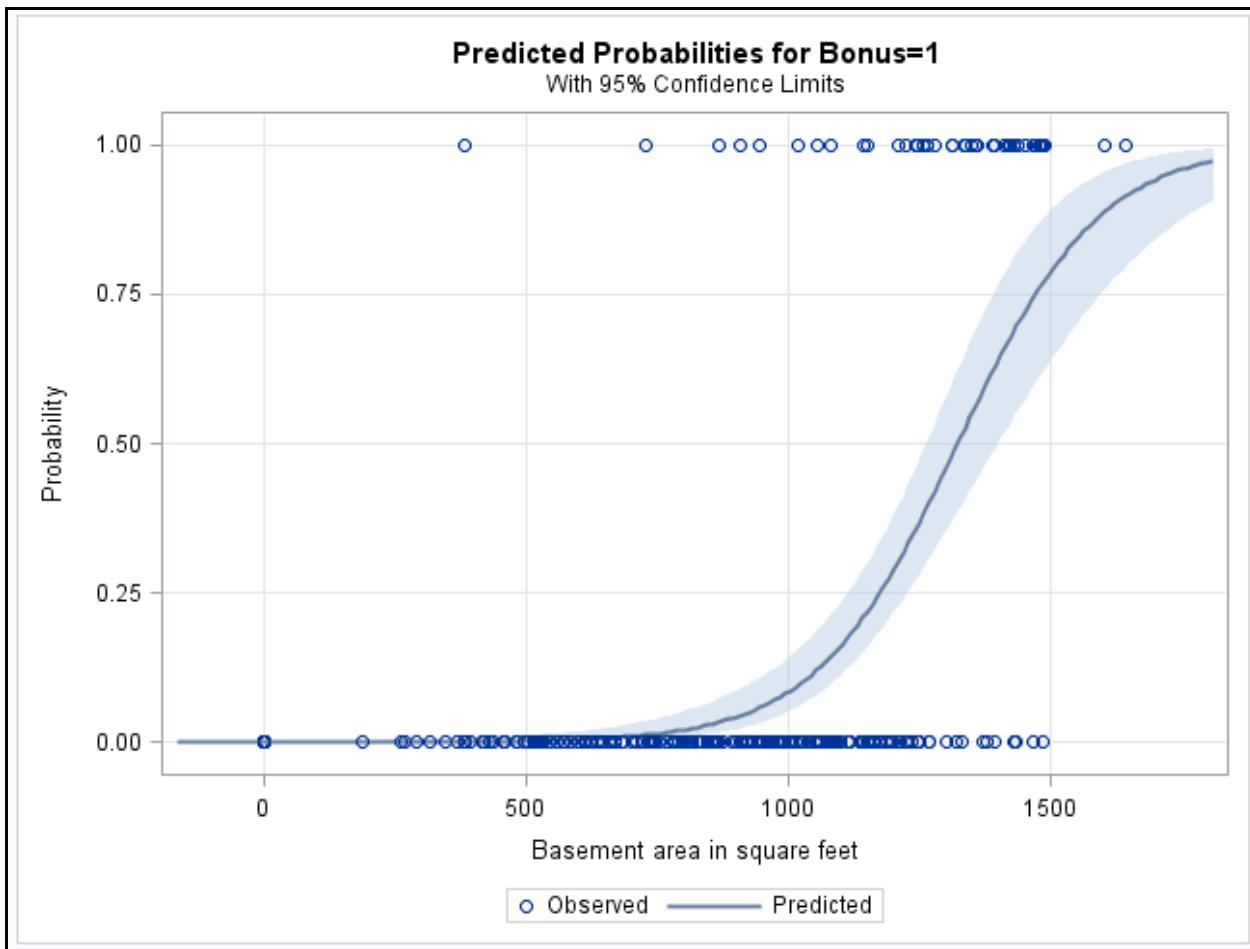
Note: Wald statistics (p -values and confidence limits) require fewer computations to perform and are therefore the default for most output in PROC LOGISTIC.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.7854	1.2896	57.5758	<.0001
Basement_Area	1	0.00739	0.00107	48.0617	<.0001

The Analysis of Maximum Likelihood Estimates table lists the estimated model parameters, their standard errors, Wald Chi-Square values, and p -values.

The parameter estimates are the estimated coefficients of the fitted logistic regression model. The logistic regression equation is $\text{logit}(\hat{p}) = -9.7854 + (0.00739) * \text{Basement_Area}$ for this example.

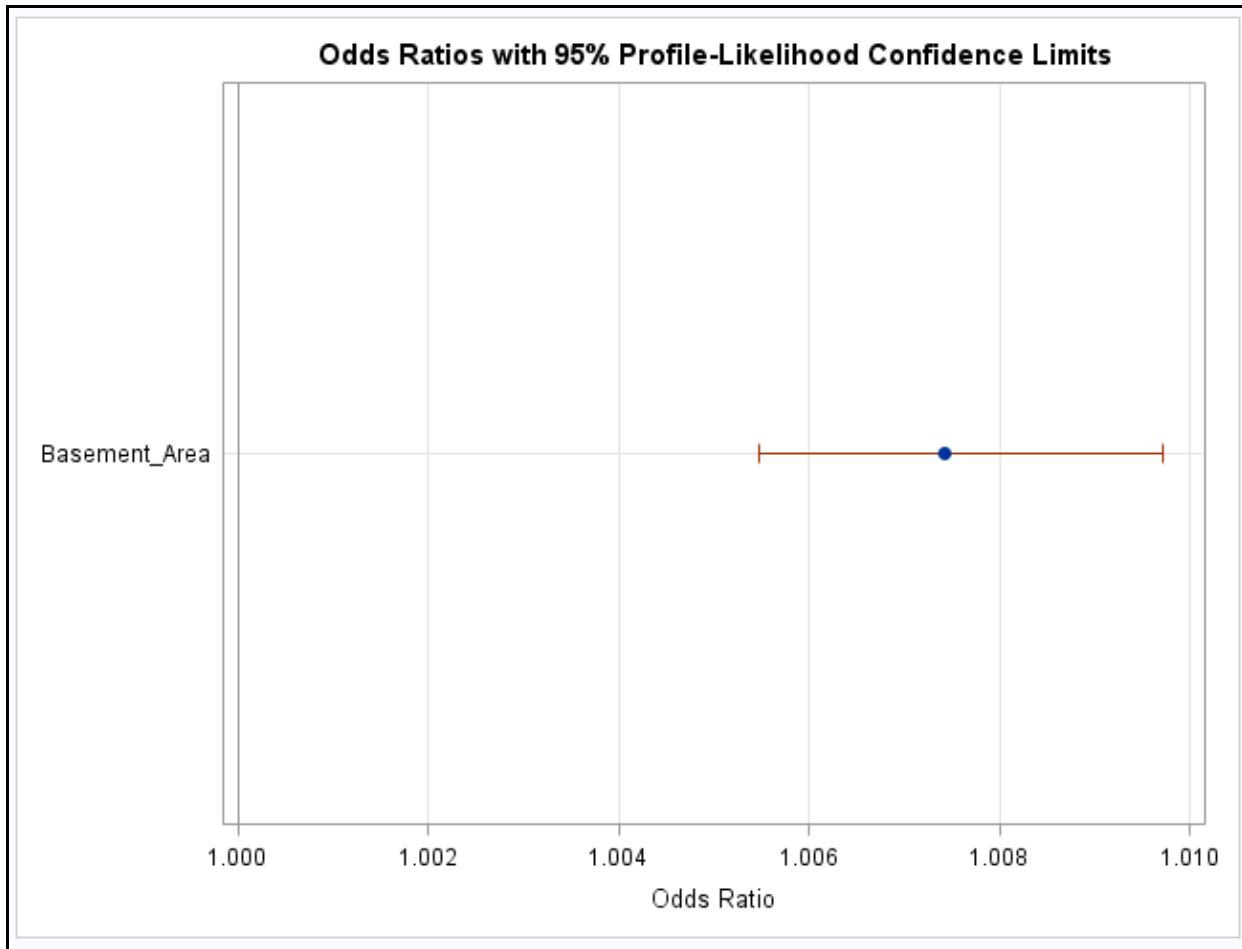
The Wald chi-square and its associated p -value tests whether the parameter estimate is significantly different from 0. For this example, the p -value for the variable **Basement_Area** is significant at the 0.05 significance level ($p < .0001$).



The estimated model is displayed on the probability scale in the Effect plot. The observed values are plotted at probabilities 1.00 and 0.00.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	1.0000	1.007	1.005	1.010



The above tables and plots are described in detail in the next slides.

End of Demonstration

Odds Ratio Calculation from the Current Logistic Regression Model

Logistic regression model:

$$\text{logit}(\hat{p}) = \log(\text{odds}) = \beta_0 + \beta_1 * (\text{Basement_Area})$$

Odds ratio (1 sq. ft. increase in Basement_Area):

$$\text{odds}_{\text{larger}} = e^{\beta_0 + \beta_1 * (\text{Basement_Area} + 1)}$$

$$\text{odds}_{\text{smaller}} = e^{\beta_0 + \beta_1 * (\text{Basement_Area})}$$

$$\begin{aligned}\text{Odds Ratio} &= \frac{e^{\beta_0 + \beta_1 * (\text{Basement_Area} + 1)}}{e^{\beta_0 + \beta_1 * (\text{Basement_Area})}} = e^{\beta_1} \\ &= e^{(0.00739)} = 1.007\end{aligned}$$

58

Copyright © SAS Institute Inc. All rights reserved.

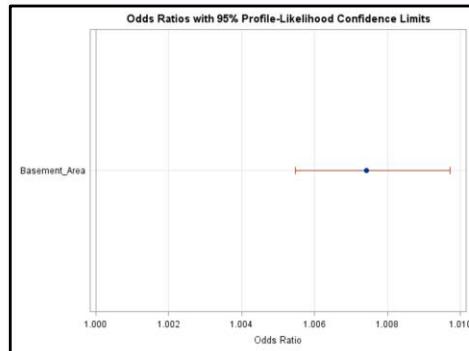


The odds ratio for a continuous predictor calculates the estimated relative odds for subjects that are one unit apart on the continuous measure. For example, in the Housing Bonus example, **Basement_Area** is the continuous measure. If you remember, the logit is the natural log of the odds. Because you can calculate an estimated logit from the logistic model, the odds can be calculated by simply exponentiating that value. An odds ratio for a one-unit difference is then the ratio of the exponentiated predicted logits for two people who are one unit apart.

The odds ratio for basement_area indicates that the odds of being bonus eligible increase by 0.7% for each increase in 1 square foot of basement area.

Odds Ratio for a Continuous Predictor

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	1.0000	1.007	1.005	1.010



59

The 95% confidence limits indicate that you are 95% confident that the true odds ratio is between 1.005 and 1.010. Because the 95% confidence interval does not include 1.000, the odds ratio is significant at the 0.05 alpha level.

Note: 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 0.05 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 error approximation, whereas the profile likelihood confidence intervals are based on the value of the log-likelihood. These likelihood-ratio confidence intervals require a much greater number of computations, but are generally preferred to the Wald confidence intervals, especially for sample sizes less than 50 (Allison 1999).

The Odds Ratio plot displays the results of the Odds Ratio table graphically. A reference line shows the null hypothesis. When the confidence interval crosses the reference line, the effect of the variable is not significant.

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.

60

Copyright © SAS Institute Inc. All rights reserved.

Comparing Pairs

To find concordant, discordant, and tied pairs, compare houses that had the outcome of interest against houses that did not.

Not Bonus Eligible



Bonus Eligible



61

Copyright © SAS Institute Inc. All rights reserved.

Concordant Pair

Compare a 1200-square foot basement that was bonus eligible with an 800-square foot basement that was not.

Not Eligible, 800 sq. ft.



$P(\text{Eligible})=.0204$

Bonus Eligible, 1200 sq. ft.



$P(\text{Eligible})=.2865$

The actual sorting agrees with the model. This is a **concordant** pair.

62

sas

For all pairs of observations with different values of the response variable, a pair is *concordant* if the observation with the outcome has a **higher** predicted outcome probability (based on the model) than the observation without the outcome.

Discordant Pair

Compare a 1400-square foot basement that was bonus eligible with a 1600-square foot basement that was not.

Not Eligible, 1600 sq. ft.



$P(\text{Eligible})=.8855$

Bonus Eligible, 1400 sq. ft.



$P(\text{Eligible})=.6379$

The actual sorting disagrees with the model. This is a **discordant** pair.

63

sas

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

Tied Pair

Compare two 1350-square foot basements. One was bonus eligible and the other not.

Not Eligible, 1350 sq. ft.



$$P(\text{Eligible}) = .5490$$

Bonus Eligible, 1350 sq. ft.



$$P(\text{Eligible}) = .5490$$

The model cannot distinguish between the two. This is a **tied pair**.

64

sas

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

Model: Concordant, Discordant, and Tied Pairs

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.2	Somers' D	0.790
Percent Discordant	10.2	Gamma	0.795
Percent Tied	0.6	Tau-a	0.202
Pairs	11475	c	0.895

65

sas

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.

The number of pairs used to calculate the values of this table is equal to the product of the counts of observations with positive responses and negative responses. In this example, that value is $255 \times 45 = 11,475$.

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 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* (concordance) statistic estimates the probability of an observation with the outcome having a higher predicted probability than an observation without the outcome. It is calculated as the percent concordant plus one half the percent tied. The range of possible values is 0.500 (no better predictive power than flipping a fair coin) to 1.000 (perfect prediction). The value of 0.895 shows a very strong ability of **Basement_Area** to discriminate between houses that were bonus eligible and houses that were not.



Exercises

2. Performing a Logistic Regression Analysis

Fit a simple logistic regression model using **STAT1.safety** with **Unsafe** as the outcome variable and **Weight** as the predictor variable. Use the **EVENT=** option to 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 global null hypothesis that all regression coefficients of the model are 0?
- b. Write the logistic regression equation.
- c. Interpret the odds ratio for **Weight**.

End of Exercises

7.4 Logistic Regression with Categorical Predictors

Objectives

- State how a logistic model with categorical predictors does and does not differ from one with continuous predictors.
- Describe what a CLASS statement does.
- Define the standard output from the LOGISTIC procedure with categorical predictor variables.

69

Copyright © SAS Institute Inc. All rights reserved.



Overview

Type of Predictors \\	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

70

Copyright © SAS Institute Inc. All rights reserved.



What Does a CLASS Statement Actually Do?

- The CLASS statement creates a set of “design variables” representing the information in the categorical variables.
 - Character variables cannot be used, as is, in a model.
 - The design variables are the ones actually used in model calculations.
 - There are several “parameterizations” available in PROC LOGISTIC.

71

Copyright © SAS Institute Inc. All rights reserved.



The CLASS statement creates a set of “design variables” representing the information contained in any categorical variables. These design variables are incorporated into the model calculations rather than the original categorical variables. Character variables cannot be used, as is, in the model. SAS cannot use a variable with values such as ‘yes’ or ‘no’ adequately in the determination of a model.

Even if categorical variables are represented by numbers such as 1, 2, 3, the CLASS statement tells SAS to set up design variables to represent the categories. This is necessary because the numeric values that are assigned to the levels of the categorical variable are generally arbitrary and might not truly reflect distances between levels.

Effect (Default) Coding: Three Levels

Design Variables					
<u>CLASS</u>	<u>Value</u>	<u>Label</u>	<u>1</u>	<u>2</u>	
IncLevel	1	Low Income	1	0	
	2	Medium Income	0	1	
	3	High Income	-1	-1	

For *effect coding* (also called *deviation from the mean coding*), the number of design variables created is the number of levels of the CLASS variable minus 1. For example, consider a variable **IncLevel**, which has three levels. In this case, two design variables were created. For the last level of the CLASS variable (**High Income**), all the design variables have a value of -1. Parameter estimates of the CLASS main effects using this coding scheme estimate the **difference** between the effect of each level and the average effect over all levels.

Effect Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

β_0 =the average value of the logit across all categories

β_1 =the difference between the logit for Low income and the average logit

β_2 =the difference between the logit for Medium income and the average logit

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	
Intercept	1	-0.5363	0.1015	27.9143	<.0001	
IncLevel 1	1	-0.2259	0.1481	2.3247	0.1273	
IncLevel 2	1	-0.2200	0.1447	2.3111	0.1285	

73



If you use Effect Coding for a CLASS variable, then the parameter estimates and *p*-values reflect differences from the mean logit value over all levels. So, for **IncLevel**, the Estimate shows the estimated difference in logit values between **IncLevel**=1 (Low Income) and the average logit across all income levels.

Reference Cell Coding: Three Levels

Design Variables

<u>CLASS</u>	<u>Value</u>	<u>Label</u>	<u>1</u>	<u>2</u>
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	0	0

74



For *reference cell coding*, parameter estimates of the CLASS main effects estimate the difference between the effect of each level and the last level, called the *reference level*. For example, the effect for the level **Low** estimates the logit difference between **Low** and **High**. You can choose the reference level in the CLASS statement.

Reference Cell Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

β_0 =the value of the logit when income is High

β_1 =the difference between the logits for Low and High income

β_2 =the difference between the logits for Medium and High income

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	
Intercept	1	-0.0904	0.1608	0.3159	0.5741	
IncLevel	1	-0.6717	0.2465	7.4242	0.0064	
IncLevel	2	-0.6659	0.2404	7.6722	0.0056	

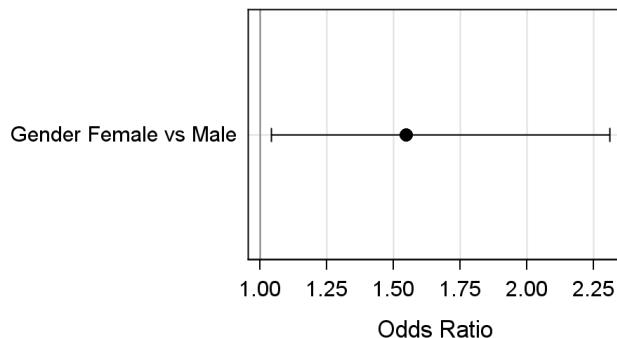
Notice the difference between this table and the previous parameter estimates table. Because you used Reference Cell Coding, instead of Effect Coding, the meanings of the parameter estimates and *p*-values are different. Now, the parameter estimate and *p*-value for **IncLevel**=1 reflect the difference between **IncLevel**=1 and **Inclevel**=3 (the reference level).

Note: It is important to know what type of parameterization you are using in order to interpret and report the results of this table.

Odds Ratio for Categorical Predictor

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	Unit	Estimate	95 % Confidence Limits	
Gender Female vs Male	1.0000	1.549	1.043	2.312

Odds Ratios with 95% Profile-Likelihood Confidence Limits



Sas

Copyright © SAS Institute Inc. All rights reserved.

Odds ratios for categorical predictors are reported for bi-group comparisons in PROC LOGISTIC, no matter which parameterization is chosen. Thus, even if Effect Coding is selected for the **Gender** variable, the odds ratio tables display odds comparisons between females and males (and not females versus the average of both). The same holds true for variables with more than two levels; comparisons will not be group versus the average of all.

Setup for the Poll

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard	Wald	Pr > ChiSq
			Error	Chi-Square	
Intercept	1	-0.0904	0.1608	0.3159	0.5741
IncLevel 1	1	-0.6717	0.2465	7.4242	0.0064
IncLevel 2	1	-0.6659	0.2404	7.6722	0.0056

Sas

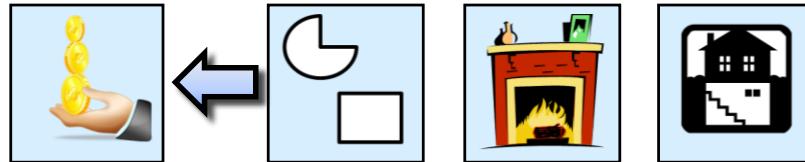
Copyright © SAS Institute Inc. All rights reserved.

7.05 Multiple Choice Poll

In the Analysis of Maximum Likelihood table, using effect coding, what is the estimated logit for someone at **IncLevel=2**?

- a. -.5363
- b. -.6717
- c. -.6659
- d. -.7563
- e. Cannot tell from the information provided

Multiple Logistic Regression



$$\text{logit}(p) = \beta_0 + \beta_1 X_{\text{irregular}} + \beta_2 X_{\text{fireplace}=1} + \beta_3 X_{\text{fireplace}=2} + \beta_4 X_{\text{Basement_Area}}$$

Each design variable is assigned its own beta value. The number of parameters in the logistic model take into account the intercept, the number of continuous predictors, and the number of design variables assigned to CLASS variables.



Multiple Logistic Regression with Categorical Predictors

Example: Fit a binary logistic regression model in PROC LOGISTIC. Select **Bonus** as the outcome variable and **Basement_Area**, **Fireplaces**, and **Lot_Shape_2** as the predictor variables. Specify reference cell coding and specify **Regular** as the reference group for **Lot_Shape_2** and **0** as the reference level for **Fireplaces**. Use the EVENT= option to specify 1 as the level of **Bonus** to model. Request profile likelihood confidence intervals around the estimated odds ratios. Request a report of odds ratios for 100 units for the **Basement_Area** variable.

1. Open the **Binary Logistic Regression** task under Statistics.
2. On the DATA tab, select the **AmesHousing3** data set.
3. Assign **Bonus** as the response variable and set **1** as the event of interest.
4. Under Explanatory Variables, assign **Fireplaces** and **Lot_Shape_2** as the classification variables and expand the Parameterization of Effects property to specify using **Reference coding**.
5. Assign **Basement_Area** as the continuous variable.
6. On the MODEL tab, specify Main effects model using the radio buttons.
7. On the OPTIONS tab, select the option to display default and additional statistics then expand the Parameter Estimates property. Select the option to include the **profile likelihood** confidence intervals for odds ratios.
8. On the OPTIONS tab in the PLOTS area, set to display default and additional plots and select the **Effect plot** and the **Odds ratio plot** options.
9. Edit a copy of the generated code as follows:
 - a. To specify specific levels of each class variable to use as reference levels, add the options (REF='Regular') immediately after **Lot_Shape_2** and (REF='0') immediately after **Fireplaces** in the CLASS statement.
 - b. Add the statement, "UNITS Basement_Area=100;" right before the RUN statement.
10. Run the code

Edited code is shown below.

```
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=STAT1.AMESHOUSING3 plots=(effect
    oddsratio(cldisplay=serifarrow) );
  class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') /
    param=ref;
  model Bonus(event='1')=Basement_Area Fireplaces Lot_Shape_2 /
    link=logit technique=fisher clodds=pl;
  units Basement_Area=100;
run;
```

Note: Alternatively, you can write the code directly.

```
/*st107d05.sas*/
ods graphics on;
proc logistic data=STAT1.ameshousing3
  plots(only)=(effect oddsratio);
  class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
  model Bonus(event='1')=Basement_Area Fireplaces Lot_Shape_2 /
    clodds=pl;
  units Basement_Area=100;
  title 'LOGISTIC MODEL (2): Bonus= Basement_Area Fireplaces
    Lot_Shape_2';
run;
```

Selected PROC LOGISTIC statement:

UNITS enables you to specify units of change for the continuous explanatory variables so that customized odds ratios can be estimated.

Selected CLASS statement options:

(REF='level') specifies the event category chosen as the reference level when using Reference or Effect parameterization. You can specify the value (formatted if a format is applied) of the reference category in quotation marks or you can specify one of the following keywords. The default is REF=LAST.

FIRST designates the first ordered category as the reference level.

LAST designates the last ordered category as the reference level.

PARAM= specifies the parameterization. This value can be specified for each CLASS variable by typing it within parentheses after the variable name, or for all CLASS variables, by typing it after the options slash (/) at the end of the list of CLASS variables.

NOTE: If there are numerous levels in the CLASS variable, you might want to use subject-matter knowledge to reduce the number of levels. This is especially important when the levels have few or no observations.

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > \$175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	299

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	44

Probability modeled is Bonus=1.

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

The Class Level Information table includes the predictor variable in the CLASS statement. Because you used the PARAM=REF and REF='Regular' options, this table reflects your choice of **Lot_Shape_2='Regular'** as the reference level. The design variable is 1 when **Lot_Shape_2='Irregular'** and 0 when **Lot_Shape_2='Regular'**. The reference level for **Fireplaces** is 0, so there are two design variables, each coded 0 for observations where **Fireplaces**=0.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	140.499
SC	255.513	159.001
-2 Log L	249.812	130.499

The SC value in the **Basement_Area** only model was 169.246. Here it is 159.001. Recalling that smaller values imply better fit, you can conclude that this model is better fitting.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	119.3133	4	<.0001
Score	91.7250	4	<.0001
Wald	49.8671	4	<.0001

This model is statistically significant, indicating at least one of the predictors in the model is useful in predicting **Bonus**.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Basement_Area	1	38.1356	<.0001
Fireplaces	2	5.2060	0.0741
Lot_Shape_2	1	16.9421	<.0001

The Type 3 Analysis of Effects table is generated when a predictor variable is used in the CLASS statement. This analysis is similar to the individual tests in the GLMSELECT procedure parameter estimates table. Just as in PROC GLMSELECT and PROC REG, these are adjusted effects.

Fireplaces is not statistically significant at the 0.05 level while the other remaining predictors are statistically significant.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-11.0882	1.5384	51.9467	<.0001
Basement_Area		1	0.00744	0.00120	38.1356	<.0001
Fireplaces	1	1	0.8810	0.4658	3.5770	0.0586
Fireplaces	2	1	-0.7683	0.9654	0.6335	0.4261
Lot_Shape_2	Irregular	1	1.9025	0.4622	16.9421	<.0001

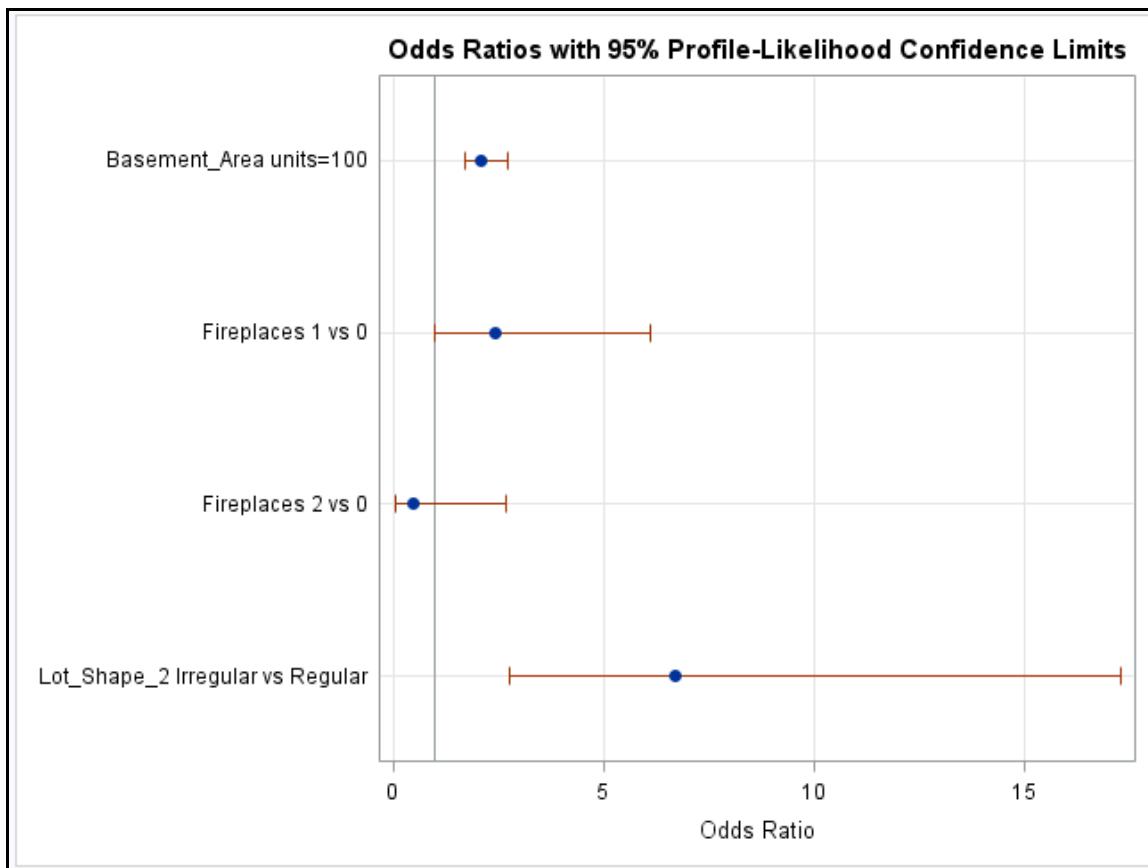
For CLASS variables, effects are displayed for each of the design variables. Because reference cell coding was used, each effect is measured against the reference level. For example, the estimate for **Lot_Shape_2 | Irregular** shows the difference in logits between houses with irregular and regular lot shapes. **Fireplaces | 1** shows the logit difference between houses with 1 fireplace and 0 fireplaces while **Fireplaces | 2** shows the difference in logits between houses with 2 fireplaces and 0 fireplaces. Not all of these contrasts are statistically significant.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	92.9	Somers' D	0.859
Percent Discordant	7.0	Gamma	0.860
Percent Tied	0.1	Tau-a	0.216
Pairs	11220	c	0.930

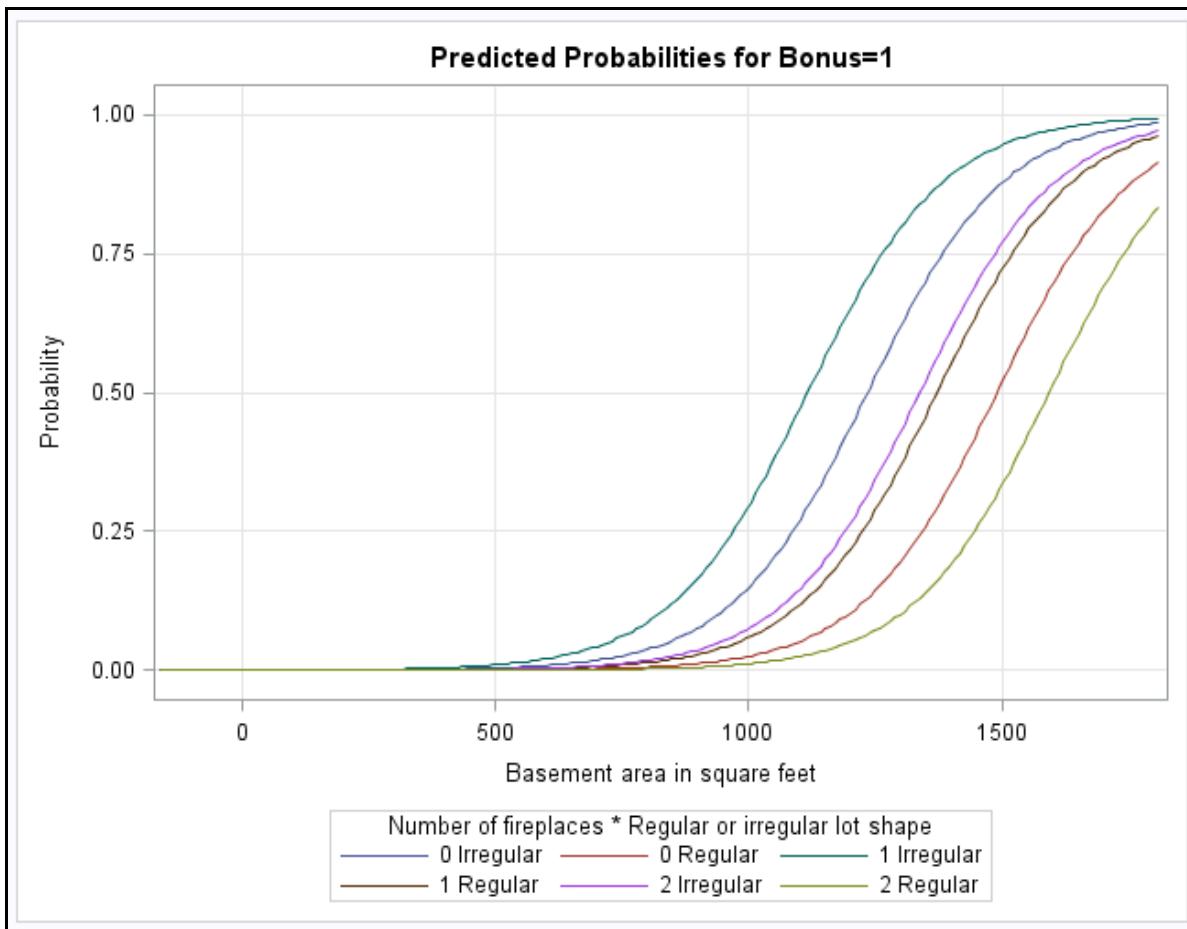
The c (Concordance) statistic value is 0.930 for this model, indicating that 93% of the positive and negative response pairs are correctly sorted using **Basement_Area**, **Fireplaces**, and **Lot_Shape_2**.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	100.0	2.105	1.696	2.727
Fireplaces 1 vs 0	1.0000	2.413	0.973	6.127
Fireplaces 2 vs 0	1.0000	0.464	0.054	2.703
Lot_Shape_2 Irregular vs Regular	1.0000	6.703	2.786	17.301

The odds ratios show that, adjusting for the other predictor variables, houses with irregular plots had 6.703 times the houses with regular plots odds of being bonus eligible. Houses with 1 fireplace had nearly 2.5 times the odds (2.413) of houses with 0 fireplaces and houses with 2 fireplaces had 53.6% lower odds than houses with 0 fireplaces. The UNITS statement applies to the odds ratio table requested by the CLODDS=PL option. The table shows that a 100 square foot larger basement is associated with a 110.5% increase in bonus eligibility odds. The ODDSRATIO plot displays these values graphically.



Finally, the Effects plot shows the probability of survival across all combinations of categories and levels of all three predictor variables.



This plot is obtained by applying the parameter estimates from the logistic model to values of the predictors and then converting the predictions to the probability scale.

End of Demonstration



Exercises

3. Performing a Multiple Logistic Regression Analysis Including Categorical Variables

Fit a logistic regression model using **STAT1.safety** with **Unsafe** as the outcome variable and **Weight**, **Region**, and **Size** as the predictor variables. Request reference cell coding with **Asia** as the reference level for **Region** and **3** (large cars) as the reference level for **Size**. Use the **EVENT=** option to 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. If you do reject the global null hypothesis, then which predictors significantly predict safety outcome?
- c. Interpret the odds ratio for significant predictors.

End of Exercises

7.06 Multiple Choice Poll

A variable coded 1, 2, 3, and 4 is parameterized with effect coding, with 2 as the reference level. The parameter estimate for level 1 tells you which of the following?

- a. The difference in the logit between level 1 and level 2
- b. The odds ratio between level 1 and level 2
- c. The difference in the logit between level 1 and the average of all levels
- d. The odds ratio between level 1 and the average of all levels
- e. Both a and b
- f. Both c and d

7.5 Stepwise Selection with Interactions and Predictions

Objectives

- Fit a multiple logistic regression model with main effects and interactions using the backward elimination method.
- Explain interactions using graphs.
- Calculate predictions in a logistic setting using PROC PLM.

Overview

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

89



Copyright © SAS Institute Inc. All rights reserved.

Stepwise Methods – Default Selection Criteria

	PROC REG/ PROC GLMSELECT			PROC LOGISTIC	
	SLENTRY	SLSTAY		SLENTRY	SLSTAY
FORWARD	0.50	-----		0.05	
BACKWARD	-----	0.10			0.05
STEPWISE	0.15	0.15		0.05	0.05

90



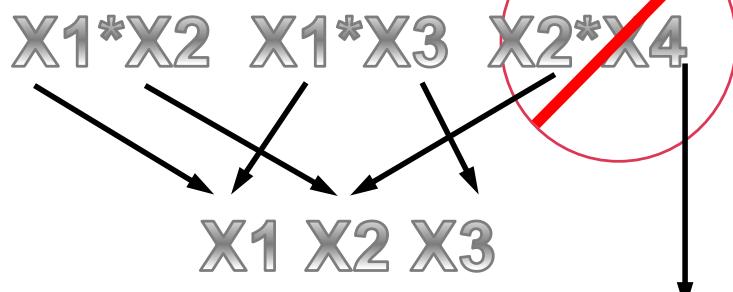
Copyright © SAS Institute Inc. All rights reserved.

If you are doing exploratory analysis and want to find a best subset model, PROC LOGISTIC provides the three stepwise methods that are available in PROC REG or PROC GLMSELECT. However, the default selection criteria are not the same. Remember that you can always change the selection criteria using the SLENTRY= and SLSTAY= options in the MODEL statement.

If you have a large number of variables, you might first need to try a variable reduction method such as variable clustering.

Stepwise Hierarchy Rules

By default, at each step model hierarchy is retained. This means that higher level effects cannot be in a model when any of its lower level composite effects are not present.



Model hierarchy refers to the requirement that, for any term to be in the model, all effects contained in the term must be present in the model. For example, in order for the interaction $X2*X4$ to enter the model, the main effects $X2$ and $X4$ must be in the model. Likewise, neither effect $X2$ nor $X4$ can leave the model while the interaction $X2*X4$ is in the model.

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 consist of only significant interactions, the main effects involved in those interactions, and any other significant main effects.

Note: For a more customized analysis, the `HIERARCHY=` option specifies whether the 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® 9.4 User's Guide* in the SAS online documentation for more information about 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.



Logistic Regression: Backward Elimination with Interactions

Example: Fit a multiple logistic regression model using the backward elimination method. The full model should include all the main effects and two-way interactions.

1. Open the **Binary Logistic Regression** task under Statistics.
2. On the DATA tab, select the **AmesHousing3** data set.
3. Assign **Bonus** as the response variable and set the level **1** as the event of interest.
4. Assign **Fireplaces** and **Lot_Shape_2** as the classification variables and choose Reference coding to parameterize the effects.
5. Assign **Basement_Area** as the continuous variable.
6. On the MODEL tab, click the radio button for **Custom model**. Then click the edit button under model effects. Use the model effects builder to specify the model. To specify the interaction terms, either use the **Cross** button under Single Effects to specify each interaction term or select all the variables and use the **N-way Factorial** button with **N=2** to specify a model with all the variables and the associated two-way interaction terms.
7. On the SELECTION tab, specify using **Backward elimination** method for model selection with 0.1 significance level.
8. On the OPTIONS tab, select the option to display default and additional statistics then expand the Parameter Estimates property. Select the option to include the **profile likelihood** confidence intervals for odds ratios.
9. Select the option to display default and additional plots and select the option to include the effect plot.
10. Edit a copy of the generated code as follows:
 - a. To specify specific levels of each class variable to use as reference levels, add the options (REF='Regular') immediately after **Lot_Shape_2** and (REF='0') immediately after **Fireplaces** in the CLASS statement.
 - b. Add the statement, "UNITS Basement_Area=100;" right before the RUN statement.
11. Run the code.

Note: Alternatively you can write the code directly.

```
/*st107d06.sas*/ /*Part A*/
proc logistic data=STAT1.ameshousing3
  plots(only)=(effect oddsratio);
  class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
  model Bonus(event='1')=Basement_Area|Fireplaces|Lot_Shape_2 @2 /
    selection=backward clodds=pl slstay=0.10;
  units Basement_Area=100;
  title 'LOGISTIC MODEL (3): Backward Elimination '
    'Bonus=Basement_Area|Fireplaces|Lot_Shape_2';
run;
```

Note: The bar notation with the @2 constructs a model with all the main effects and the two-factor interactions. If you increase it to @3, then you 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.

Selected MODEL statement option:

SELECTION= specifies the method to select the variables in the model. BACKWARD requests backward elimination, FORWARD requests forward selection, NONE fits the complete model specified in the MODEL statement, STEPWISE requests stepwise selection, and SCORE requests best subset selection. The default is NONE.

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > \$175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	299

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	44

Probability modeled is Bonus=1.

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

All information to this point is the same as that from the previous model.

Backward Elimination Procedure

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

The Model Fit Statistics and Testing Global Null Hypothesis tables at Step 0 are presented.

Step 0. The following effects were entered:

**Intercept Basement_Area Fireplaces Basement_Area*Fireplaces Lot_Shape_2
Basement_Area*Lot_Shape_2 Fireplaces*Lot_Shape_2**

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	141.737
SC	255.513	178.741
-2 Log L	249.812	121.737

Testing Global Null Hypothesis: BETA=0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	128.0756	9	<.0001	
Score	109.4005	9	<.0001	
Wald	40.8304	9	<.0001	

Step 1. Effect Fireplaces*Lot_Shape_2 is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	141.166
SC	255.513	170.769
-2 Log L	249.812	125.166

Testing Global Null Hypothesis: BETA=0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	124.6462	7	<.0001	
Score	106.9810	7	<.0001	
Wald	42.3266	7	<.0001	

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
3.4592	2	0.1774

Step 2. Effect Basement_Area*Fireplaces is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	138.872
SC	255.513	161.074
-2 Log L	249.812	126.872

Testing Global Null Hypothesis: BETA=0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	122.9405	5	<.0001	
Score	102.6370	5	<.0001	
Wald	42.9826	5	<.0001	

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
5.5364	4	0.2365

Note: No (additional) effects met the 0.1 significance level for removal from the model.

The procedure stops after the two interactions involving **Fireplaces** are removed.

Summary of Backward Elimination						
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Fireplace*Lot_Shape_	2	5	3.2305	0.1988	
2	Basement_*Fireplaces	2	4	1.7237	0.4224	

Joint Tests				
Effect	DF	Wald Chi-Square	Pr > ChiSq	
Basement_Area	1	18.2896	<.0001	
Fireplaces	2	4.7171	0.0946	
Lot_Shape_2	1	5.0247	0.0250	
Basement_*Lot_Shape_	1	3.1127	0.0777	

Note: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-15.3017	3.2407	22.2952	<.0001
Basement_Area		1	0.0109	0.00254	18.2896	<.0001
Fireplaces	1	1	0.7671	0.4687	2.6781	0.1017
Fireplaces	2	1	-0.9405	0.9503	0.9795	0.3223
Lot_Shape_2	Irregular	1	8.0362	3.5850	5.0247	0.0250
Basement_*Lot_Shape_2	Irregular	1	-0.00503	0.00285	3.1127	0.0777

Notice that when a CLASS statement is used, new rows are added to the parameter estimates table. These represent design variables that SAS creates in order to test the interactions.

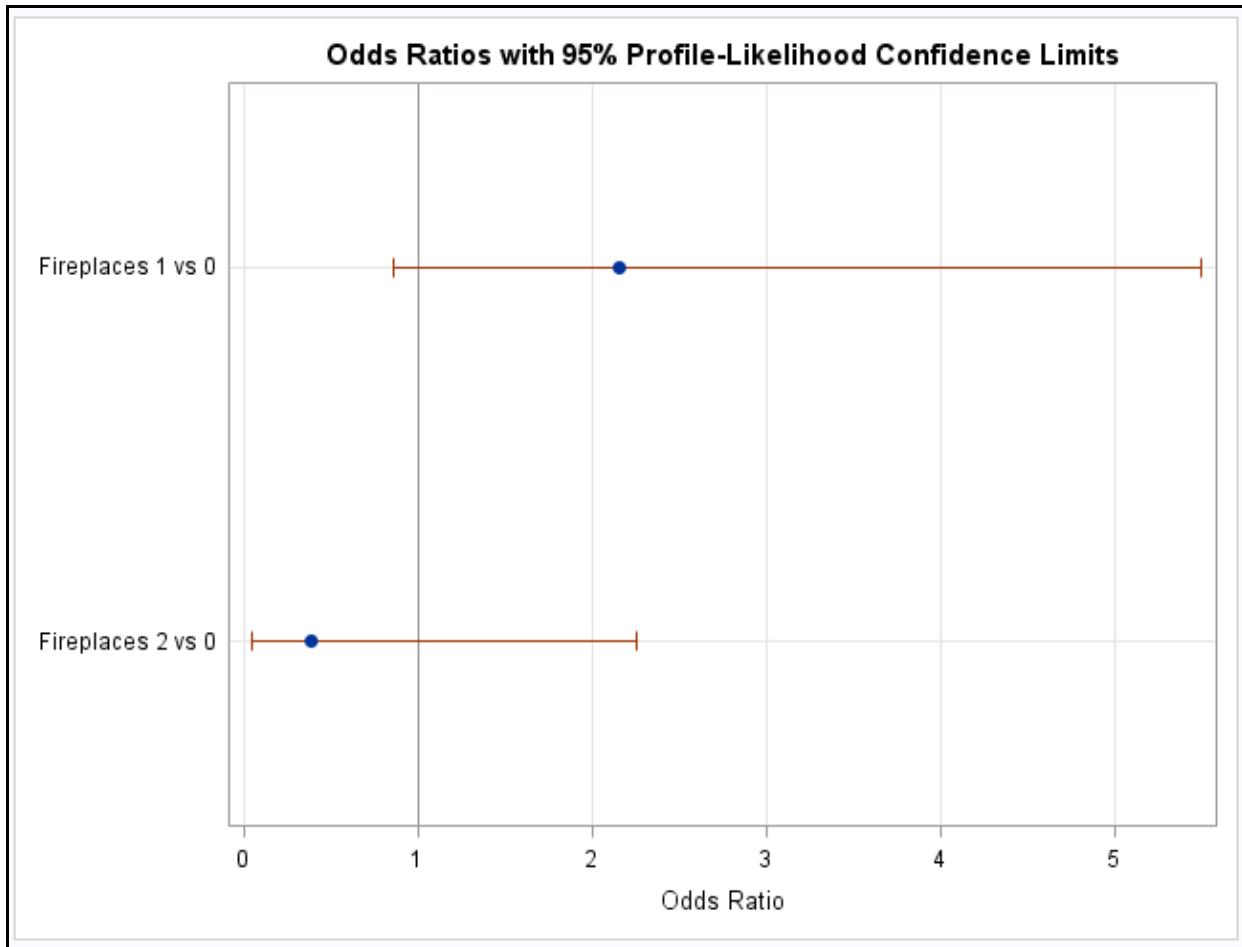
As described in the ANOVA chapter, an interaction between two variables means that the effect of one variable is different at different values of the other variable. This makes the model more complex to interpret.

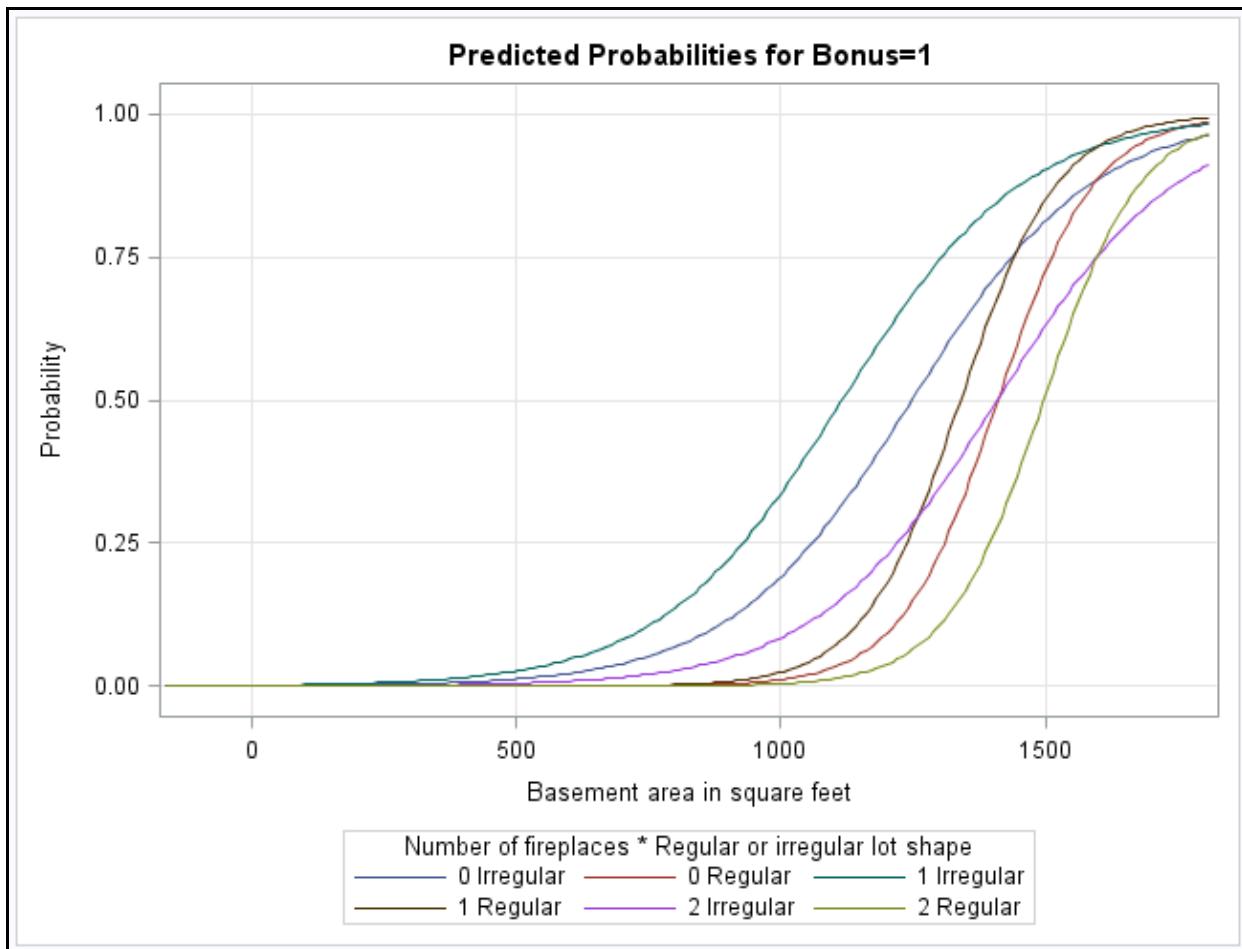
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.8	Somers' D	0.876
Percent Discordant	6.2	Gamma	0.876
Percent Tied	0.1	Tau-a	0.221
Pairs	11220	c	0.938

The c value is a slight improvement over the previous model (c=0.930) that only included the main effects.

Odds ratios are not calculated for effects involved in interactions. Any single odds ratio for **Basement_Area** or for **Lot_Shape_2** would be misleading because the effects vary for each at different levels of the other variable.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Fireplaces 1 vs 0	1.0000	2.153	0.865	5.500
Fireplaces 2 vs 0	1.0000	0.390	0.047	2.251





In order to estimate and plot odds ratios for the simple effects of variables involved in an interaction, an ODDSRATIO statement with the AT= option can be used. An EFFECTSPLOT statement can help display the interaction, as well.

Note: SAS Studio task currently need manual modification for the ODDSRATIO statement. Open the editor and modify the code or you can write the directly.

```
/*st107d06.sas*/ /*Part B*/
proc logistic data=STAT1.ameshousing3
plots(only)=oddsratio(range=clip);
class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
units Basement Area=100;
oddsratio Basement_Area / at (Lot_Shape_2=ALL) cl=pl;
oddsratio Lot_Shape_2 / at (Basement Area=1000 1500) cl=pl;
title 'LOGISTIC MODEL (3.1): Bonus=Basement_Area|Lot_Shape_2
Fireplaces';
run;
```

Selected PROC LOGISTIC statement PLOTS option:

RANGE= with suboptions (*<min><max>*) | CLIP, specifies the range of the displayed odds ratio axis. The RANGE=CLIP option has the same effect as specifying the minimum odds ratio as *min* and the maximum odds ratio as *max*. By default, all odds ratio confidence intervals are displayed. This option is helpful when one or more odds ratio confidence intervals are so large that the smaller ones become difficult to see on the scale required to show the larger ones.

Selected statement:

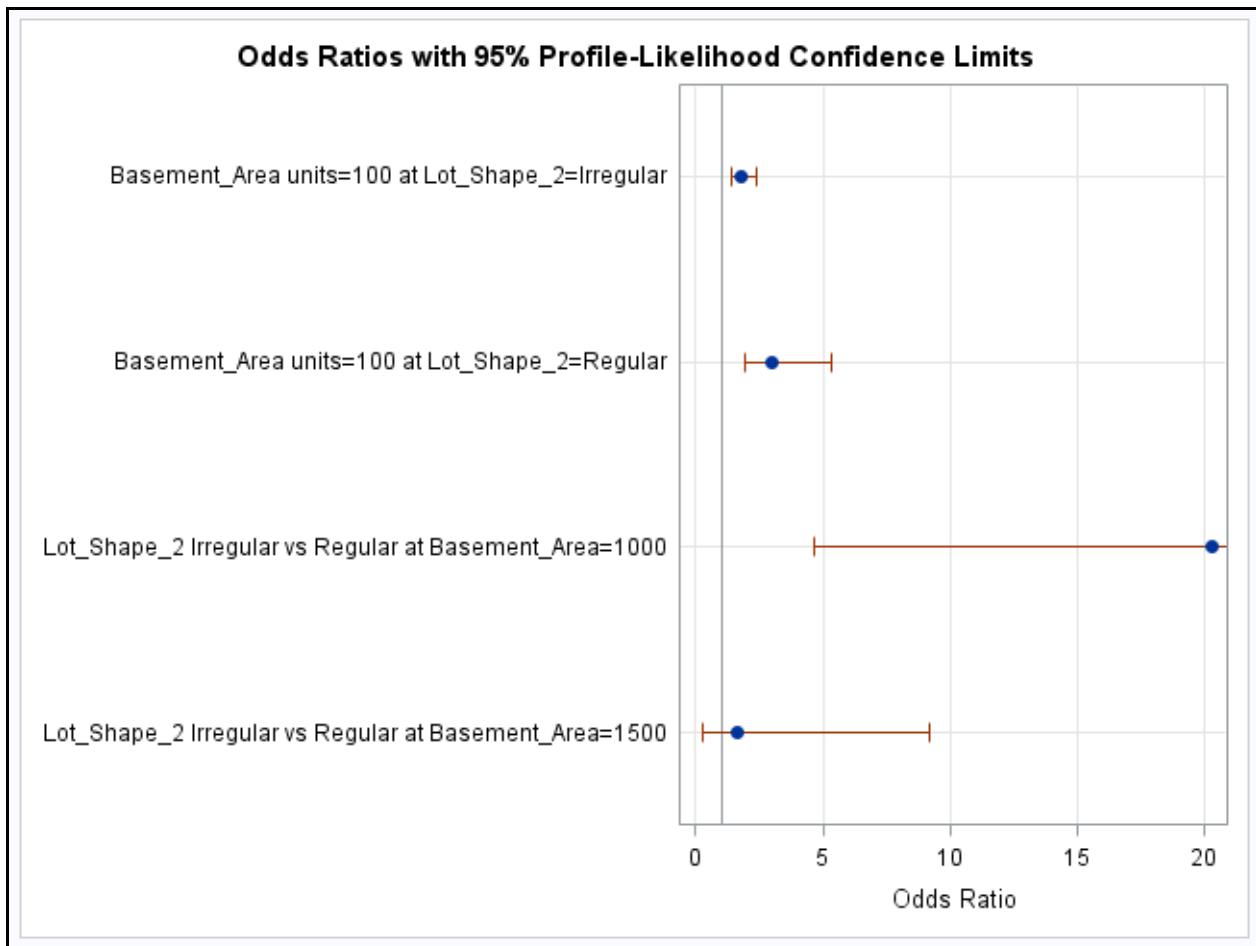
ODDSRATIO produces odds ratios for a variable even when the variable is involved in interactions with other covariates, and for classification variables that use any parameterization. You can also specify variables on which constructed effects are based, in addition to the names of COLLECTION or MULTIMEMBER effects.

Selected options for the ODDSRATIO statement:

AT specifies fixed levels of the interacting covariates. If a specified covariate does not interact with the variable, then its AT list is ignored. For continuous interacting covariates, you can specify one or more numbers in the value-list. For classification covariates, you can specify one or more formatted levels of the covariate enclosed in single quotation marks (for example, A='cat' 'dog'), you can specify the keyword REF to select the reference-level, or you can specify the keyword ALL to select all levels of the classification variable. By default, continuous covariates are set to their means, while CLASS covariates are set to ALL.

Partial PROC LOGISTIC Output

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
Basement_Area units=100 at Lot_Shape_2=Irregular	1.791	1.421	2.396
Basement_Area units=100 at Lot_Shape_2=Regular	2.960	1.932	5.315
Lot_Shape_2 Irregular vs Regular at Basement_Area=1000	20.278	4.623	146.987
Lot_Shape_2 Irregular vs Regular at Basement_Area=1500	1.643	0.283	9.145



Notice the effect of the RANGE=CLIP suboption. The Odds Ratio axis is clipped just beyond the odds ratio estimate of **Lot_Shape_2 Irregular** versus **Regular** at **Basement_Area=1000**. The upper bound of the associated 95% confidence interval is 146.987.

From this plot it is clear that the lot shape effect is different at different values of basement area.

End of Demonstration



Logistic Regression: Predictions Using PROC PLM

Example: Using the model selected from backward selection including main effects and two-way interactions, generate predictions for bonus eligibility for new data.

Note: Currently, the SAS Studio binary logistic regression task does not have the option to generate predictions for new data. To save the context and results of the statistical analysis, edit the generated code and use PROC PLM for predictions.

```
/*st107d07.sas*/
ods select none;
proc logistic data=STAT1.ameshousing3;
  class Fireplaces (ref='0') Lot_Shape_2 (ref='Regular') / param=ref;
  model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
  units Basement Area=100;
  store out=isbonus;
run;
ods select all;

data newhouses;
  length Lot_Shape_2 $9;
  input Fireplaces Lot_Shape_2 $ Basement_Area;
  datalines;
  0 Regular 1060
  2 Regular 775
  2 Irregular 1100
  1 Irregular 975
  1 Regular 800
  ;
run;

proc plm restore=isbonus;
  score data=newhouses out=scored_houses / ILINK;
  title 'Predictions using PROC PLM';
run;

proc print data=scored_houses;
run;
```

Selected SCORE statement option:

ILINK= requests that predicted values be inversely linked to produce predictions on the data scale. By default, predictions are produced on the linear scale where covariate effects are additive.

Store Information	
Item Store	WORK.ISBONUS
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC LOGISTIC
Date Created	04SEP14:09:27:06

Store Information	
Response Variable	Bonus
Link Function	Logit
Distribution	Binary
Class Variables	Fireplaces Lot_Shape_2 Bonus
Model Effects	Intercept Basement_Area Lot_Shape_2 Basement_*Lot_Shape_Fireplaces

Obs	Lot_Shape_2	Fireplaces	Basement_Area	Predicted
1	Regular	0	1060	0.02192
2	Regular	2	775	0.00040
3	Irregular	2	1100	0.14210
4	Irregular	1	975	0.30608
5	Regular	1	800	0.00286

The PROC PLM output shows that the house with the highest predicted probability (0.306) of being bonus eligible has an irregular lot shape, 1 fireplace, and a basement area of 975 square feet. The house with the lowest predicted probability (0.0004) has a regular lot shape, 2 fireplaces, and a basement area of 775.

Note: Care should be taken to ensure that predictions are made only for new data records that fall within the range of the training data. If not, predictions could be invalid due to extrapolation.

End of Demonstration



Exercises

4. Performing Backward Elimination and Prediction

Using the **STAT1.safety** data set, run PROC LOGISTIC and use backward elimination. Start with a model using **only main effects**. 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. Use the **SIZEFMT** format for the variable **Size**. Specify **Region** and **Size** as classification variables using reference cell coding and specify **Asia** as the reference level for **Region** and **Small** as the reference level for **Size**. Use a **UNITS** statement with -1 as the units for weight, so that you can see the odds ratio for lighter cars over heavier cars. Request any relevant plots.

- Which terms appear in the final model?
- Do you think this is a better model than the one fit with only **Region**?
- Using the final model, chosen by backward elimination, and the STORE statement, generate predictive probabilities for the cars in the following DATA step code.

```
data checkSafety;
length Region $9.;
input Weight Size Region $ 5-13;
datalines;
4 1 N America
3 1 Asia
5 3 Asia
5 2 N America
;
run;
```

Note: 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.

```
value sizefmt 1='Small'
      2='Medium'
      3='Large';
```

End of Exercises

7.6 Solutions

Solutions to Exercises

1. Performing Tests and Measures of Association

An insurance company wants to relate the safety of vehicles to several other variables. A score is given to each vehicle model, using the frequency of insurance claims as a basis. The data are in the **STAT1.safety** data set.

- a. Invoke the FREQ procedure and create one-way frequency tables for the categorical variables.
 - 1) Open the **One-Way Frequencies** task under Statistics.
 - 2) Select the **Safety** data set.
 - 3) Assign **Unsafe**, **Type**, **Region**, and **Size** as the analysis variables.
 - 4) On the OPTIONS tab, expand PLOTS and select the option to suppress plots.
 - 5) Run the code.

Note: The code below produce all the necessary tables.

```
/*st107s01.sas*/ /*Part A*/
ods graphics off;
proc freq data=STAT1.safety;
  tables Unsafe Type Region Size;
  title "Safety Data Frequencies";
run;
ods graphics on;
```

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

Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	35	36.46	35	36.46
2	29	30.21	64	66.67
3	32	33.33	96	100.00

- a) What is the measurement scale of each variable?

<u>Variable</u>	<u>Measurement Scale</u>
Unsafe	Nominal, Ordinal, Binary
Type	Nominal
Region	Nominal
Weight	Ratio (Continuous)
Size	Ordinal

- b) What is the proportion of cars made in North America?

63.54 %

- c) For the variables **Unsafe**, **Size**, **Region**, and **Type**, are there any unusual data values that warrant further investigation?

No.

- b.** Use PROC FREQ to examine the crosstabulation of the variables **Region** by **Unsafe**. Along with the default output, generate the expected frequencies, the chi-square test of association and the odds ratio. (Optional: Generate a temporary format to clearly identify the values of **Unsafe**.)
- 1) Open the **Table Analysis** task under **Statistics**.
 - 2) Assign **Region** as the row variable and **Unsafe** as the column variable.
 - 3) On the OPTIONS tab, expand the PLOTS property and select the option to suppress plots.
 - 4) Select the option to display the observed frequencies, expected frequencies, cell percentages, row percentages, and column percentages. Also select the option to display the chi-square test of association and the odds ratio.
 - 5) Run the code.

Note: Alternatively, you can write the code directly.

```
/*st107s01.sas*/ /*Part B*/
proc format;
  value safefmt 0='Average or Above'
            1='Below Average';
run;

proc freq data=STAT1.safety;
  tables Region*Unsafe / expected chisq relrisk;
  format Unsafe safefmt. ;
  title "Association between Unsafe and Region";
run;
```

Table of Region by Unsafe				
Region	Unsafe			
	Average or Above	Below Average	Total	
Asia	20	15	35	
	24.063	10.938		
	20.83	15.63	36.46	
	57.14	42.86		
	30.30	50.00		
N America	46	15	61	
	41.938	19.063		
	47.92	15.63	63.54	
	75.41	24.59		
	69.70	50.00		
Total	66	30	96	
	68.75	31.25	100.00	

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	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	20
Left-sided Pr <= F	0.0525
Right-sided Pr >= F	0.9809
Table Probability (P)	0.0334
Two-sided Pr <= P	0.0718

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.4348	0.1790	1.0562
Relative Risk (Column 1)	0.7578	0.5499	1.0443
Relative Risk (Column 2)	1.7429	0.9733	3.1210

- a) For the cars made in Asia, what percentage had a below-average safety score?

Region is a row variable, so look at the Row Pct value in the Below Average cell of the Asia row. That value is 42.86.

- b) For the cars with an average or above safety score, what percentage was made in North America?

The Col Pct value for the cell for North America in the column for Average or Above is 69.70.

- c) Do you see a statistically significant (at the 0.05 level) association between Region and Unsafe?

The association is not statistically significant at the 0.05 alpha level. The p-value is 0.0631.

- d) What does the odds ratio compare and what does this one say about the difference in odds between Asian and North American cars?

The odds ratio compares the odds of below average safety for North America versus Asia. The odds ratio of 0.4348 means that cars made in North America have 56.52 percent lower odds for being unsafe than cars made in Asia.

Note: Recall that odds ratios given in the Estimates of Relative Risk table are calculated comparing row1/row2 for column1. In this problem, this comparison is **Asia to N America** whose outcome is **Average or Above** in safety.

The value 0.4348

is interpreted as the odds of having an **Average or Above** car made in **Asia** is 0.4348 times the odds for American-made cars. If you wished to compare **N America to Asia**, still using **Average or Above** for safety, the odds ratio would be the inverse of 0.4348, or approximately 2.3. This is interpreted as cars made in North America have 2.3 times the odds for being safe than cars made in Asia. This single inversion would also create the odds ratio for comparing **Asia**

to **N America** but **Below Average** in safety. If you wished to compare **N America** to **Asia** using **Below Average** in safety, you would invert your odds ratio twice returning to the value 0.4348.

- c. Use the variable named **Size**. Examine the ordinal association between **Size** and **Unsafe**. Use PROC FREQ.
 - a. Edit the **Table Analysis** task by assigning **Size** as the row variables.
 - b. On the OPTIONS tab, clear the option to include expected frequencies and select the option to display statistics for the **Measures of association**.
 - c. To include confidence level for the estimates, edit the generated code and specify **cl** after the slash (/) in the TABLES statement.

Note: Alternatively, you can write the code directly.

```
/*st107s01.sas*/ /*Part C*/
proc freq data=STAT1.safety;
  tables Size*Unsafe / chisq measures cl;
  format Unsafe safefmt.;
  title "Association between Unsafe and Size";
run;
```

Table of Size by Unsafe				
Size	Unsafe			
	Average or Above	Below Average	Total	
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

- 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?

Reject

- c) What is the strength of the ordinal association between **Size** and **Unsafe**?

The Spearman correlation is -0.5425.

- d) What is the 95% confidence interval around that statistic?

The CI is (-0.6932, -0.3917).

2. Performing a Logistic Regression Analysis

Fit a simple logistic regression model using **STAT1.safety** with **Unsafe** as the outcome variable and **Weight** as the predictor variable. Use the **EVENT=** option to model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

- a. Open the **Binary Logistic Regression** task under Statistics.
- b. Select the **Safety** data set.
- c. Assign **Unsafe** as the response variable and set 1 as the event of interest.
- d. Assign **Weight** as the continuous variable.
- e. On the MODEL tab, specify the model.
- f. On the OPTIONS tab, select the option to display default and additional statistics and expand the Parameter Estimates property. Select the option to include profile likelihood confidence intervals for odds ratios.
- g. Select the option to display default and additional plots and select the options to include the Effect plot and the Odds ratio plot.
- h. Run the code.

Note: Alternatively, you can write the code directly.

```
/*st107s02.sas*/
ods graphics on;
proc logistic data=STAT1.safety plots(only)=(effect oddsratio);
  model Unsafe(event='1')=Weight / clodds=pl;
  title 'LOGISTIC MODEL (1) :Unsafe=Weight';
run;
```

Model Information	
Data Set	STAT1.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	Unsafe	Total Frequency
1	0	66
2	1	30

Probability modeled is Unsafe=1.

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

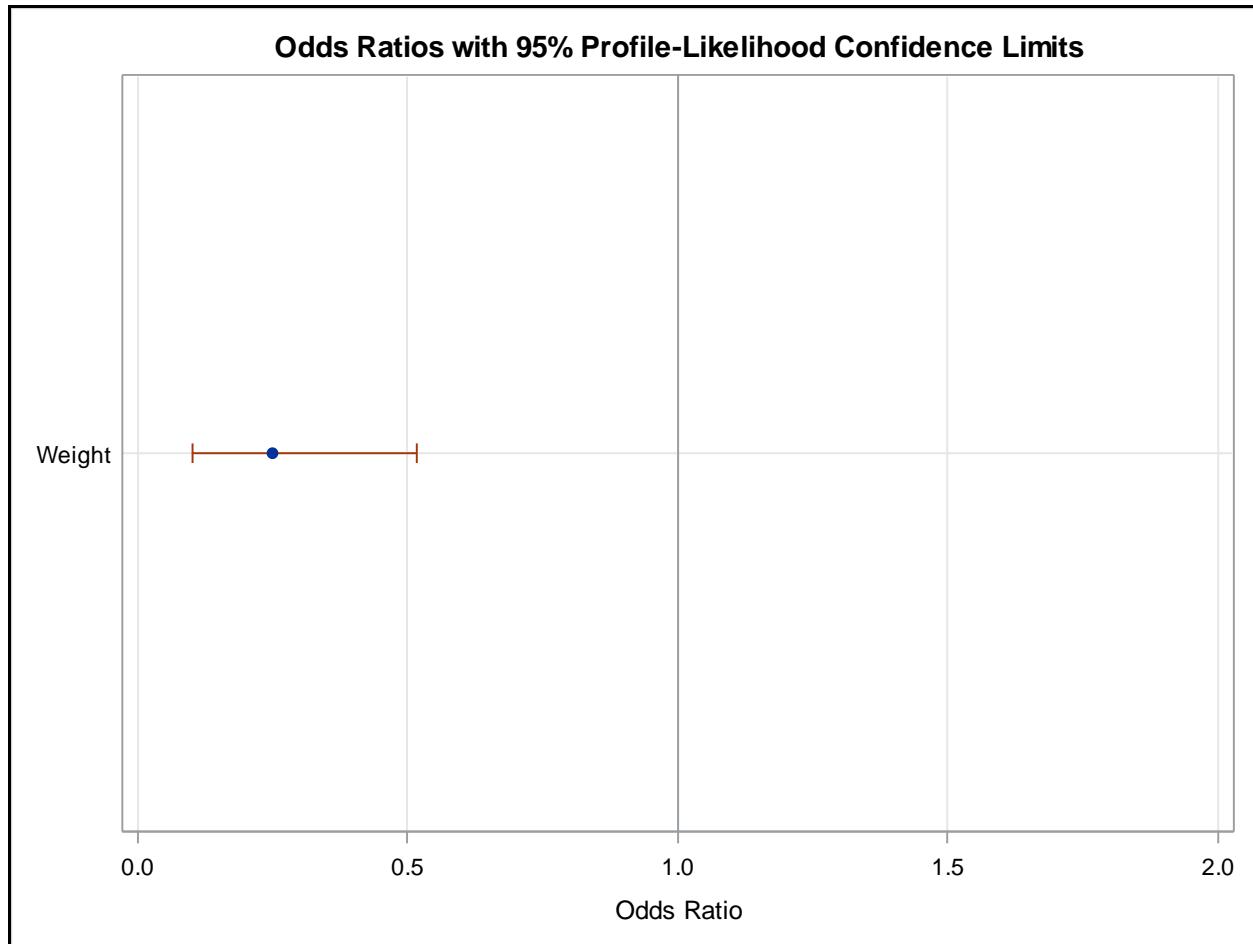
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	121.249	106.764
SC	123.813	111.893
-2 Log L	119.249	102.764

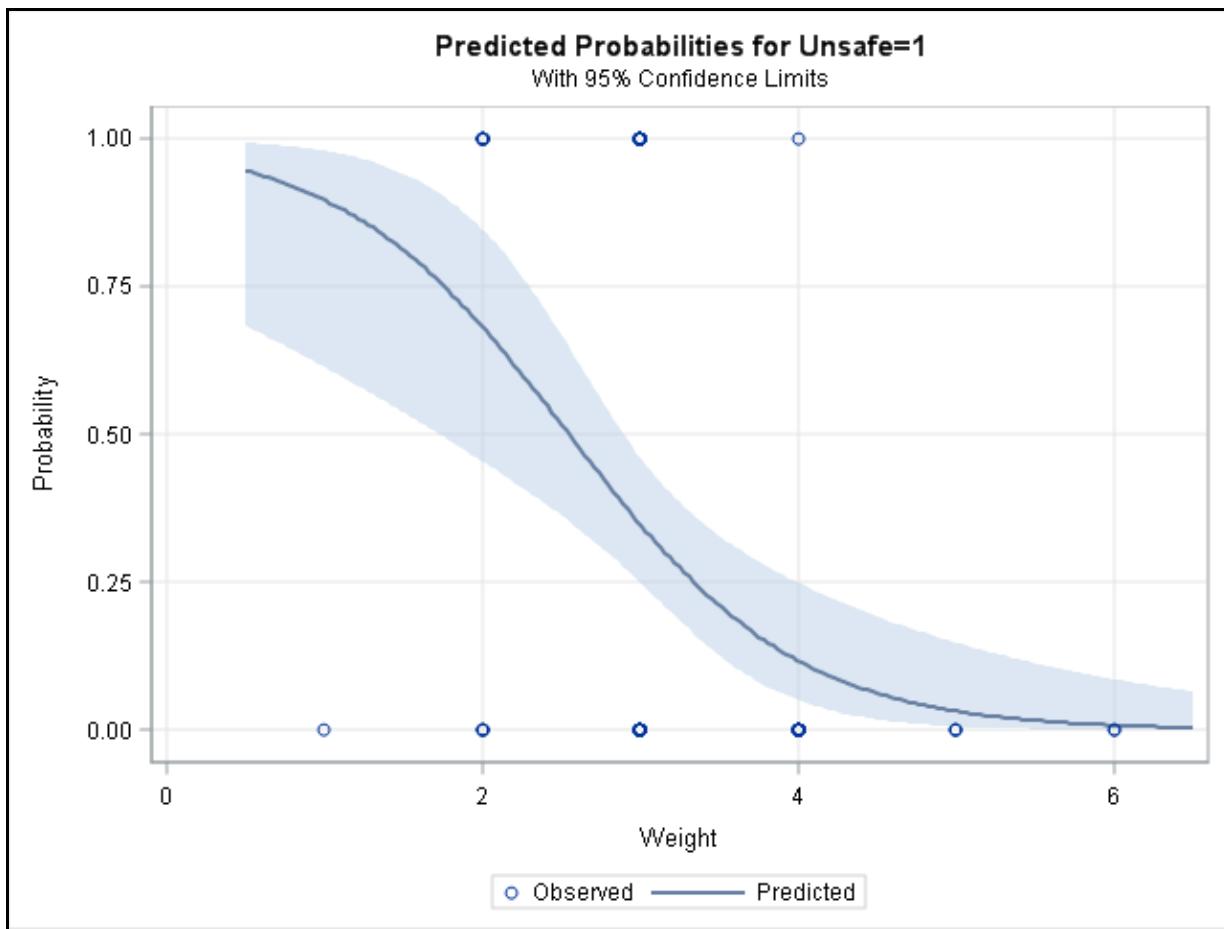
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.4845	1	<.0001
Score	13.7699	1	0.0002
Wald	11.5221	1	0.0007

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.5422	1.2601	7.9023	0.0049
Weight	1	-1.3901	0.4095	11.5221	0.0007

Association of Predicted Probabilities and Observed Responses				
Percent Concordant	55.2	Somers' D	0.474	
Percent Discordant	7.7	Gamma	0.754	
Percent Tied	37.1	Tau-a	0.206	
Pairs	1980	c	0.737	

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Weight	1.0000	0.249	0.102	0.517





- Do you reject or fail to reject the global null hypothesis that all regression coefficients of the model are 0?
The p-value for the Likelihood Ratio test is <.0001 and therefore the global null hypothesis is rejected.
- Write the logistic regression equation.
The regression equation is as follows:
 $\text{Logit(Unsafe)} = 3.5422 + (-1.3901) * \text{Weight}$.
- Interpret the odds ratio for Weight.
The odds ratio for Weight (0.249) says that the odds for being unsafe (having a below average safety rating) are 75.1% lower for each thousand pound increase in weight. The confidence interval (0.102 , 0.517) does not contain 1, indicating that the odds ratio is statistically significant.

3. Performing a Multiple Logistic Regression Analysis Including Categorical Variables

Fit a logistic regression model using **STAT1.safety** with **Unsafe** as the outcome variable and **Weight**, **Region**, and **Size** as the predictor variables. Request reference cell coding with **Asia** as the reference level for **Region** and **3** (large cars) as the reference level for **Size**. Use the **EVENT=** option to model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

- a. Open the **Binary Logistic Regression** task under Statistics.
- b. On the DATA tab, select the **Safety** data set.
- c. Assign **Unsafe** as the response variable and the event of interest to **1**.
- d. Assign **Region** and **Size** as the classification variables and select Reference coding to parameterize effects.
- e. Assign **Weight** as the continuous variable.
- f. On the MODEL tab, specify the model.
- g. On the OPTIONS tab, select the option to display default and additional statistics and expand the Parameter Estimates property to request the profile likelihood confidence intervals for odds ratios based.
- h. Select the option to display default and additional plots and select the options to include the **Effect plot** and the **Odds ratio plot**.
- i. Open the editor to specify the reference level explicitly for the two classification variables.
- j. Run the code.

Note: Alternatively, you can write the code directly.

```
/*st107s03.sas*/
ods graphics on;
proc logistic data=STAT1.safety plots(only)=(effect oddsratio);
  class Region (param=ref ref='Asia')
    Size (param=ref ref='3');
  model Unsafe(event='1')=Weight Region Size / clodds=pl;
  title 'LOGISTIC MODEL (2):Unsafe=Weight Region Size';
run;
```

Partial PROC LOGISTIC Output

Class Level Information			
Class	Value	Design Variables	
Region	Asia	0	
	N America	1	
Size	1	1	0
	2	0	1
	3	0	0

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and 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	

- a. Do you reject or fail to reject the null hypothesis that all regression coefficients of the model are 0?

You reject the null hypothesis with a p<.0001.

Type 3 Analysis of Effects				
Effect	DF	Wald		Pr > ChiSq
		Chi-Square	Pr > ChiSq	
Weight	1	2.1176	0.1456	
Region	1	0.4506	0.5020	
Size	2	15.3370	0.0005	

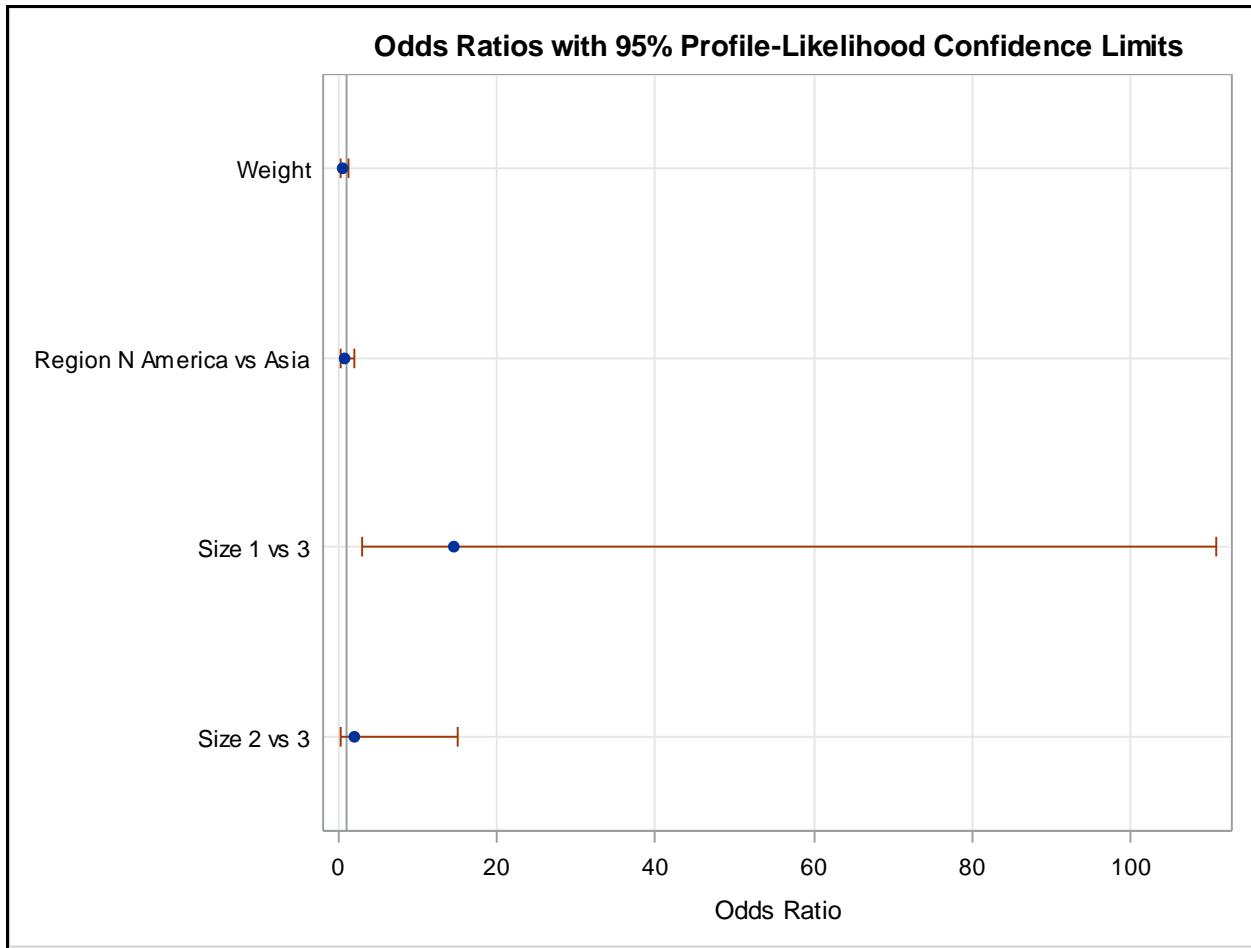
- b. If you do reject the global null hypothesis, then which predictors significantly predict safety outcome?

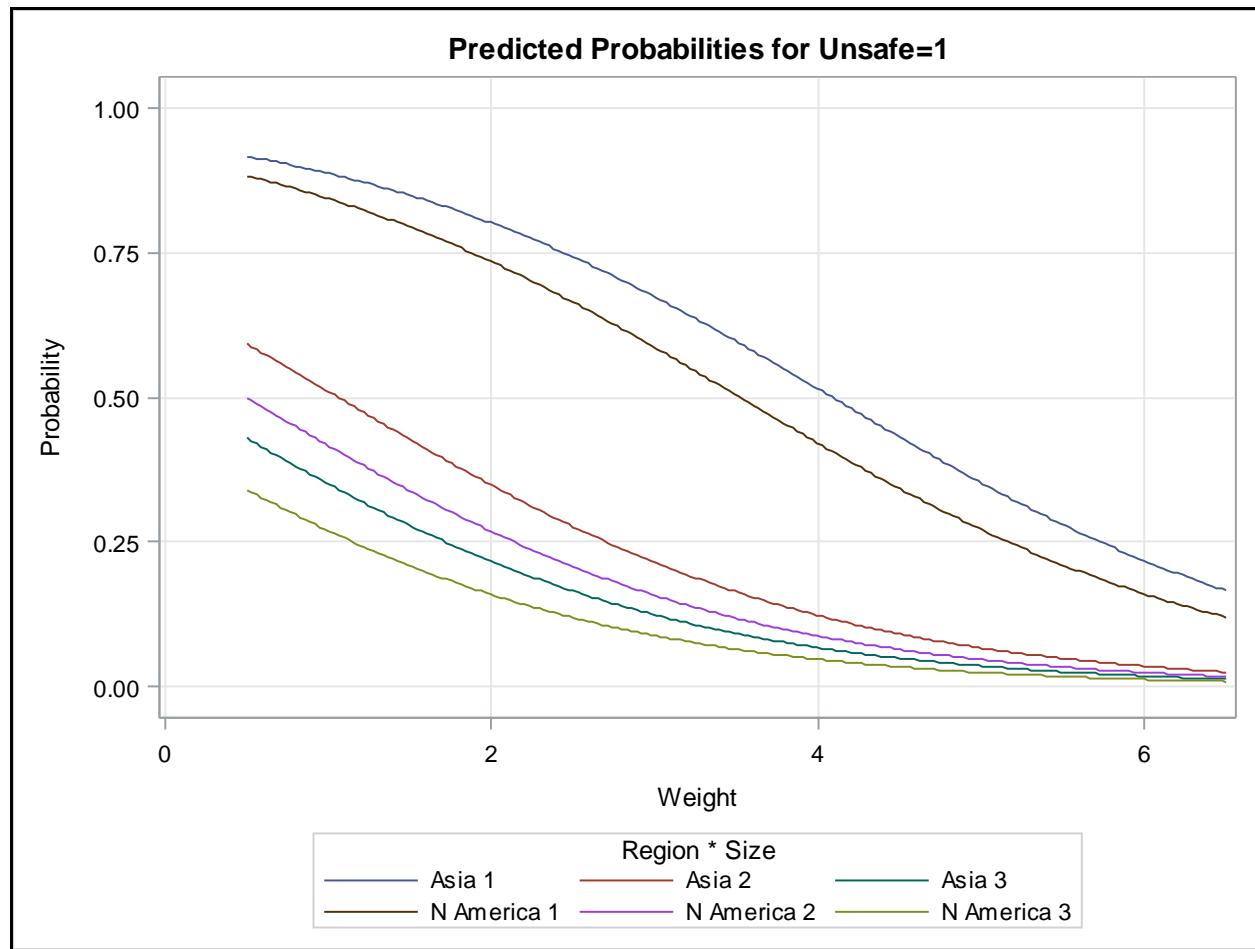
Only Size is significantly predictive of Unsafe.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
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	1	1	2.6783	0.8810	9.2422	0.0024
Size	2	1	0.6582	0.9231	0.5085	0.4758

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

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
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 1 vs 3		1.0000	14.560	3.018 110.732
Size 2 vs 3		1.0000	1.931	0.343 15.182





- c. Interpret the odds ratio for significant predictors.

Only Size is significant. The design variables show that Size=1 (Small or Sports) cars have 14.560 times the odds of having a below-average safety rating compared to the reference category, 3 (Large or Sport/Utility). The 95% confidence interval (3.018, 110.732) does not contain 1, implying that the contrast is statistically significant at the 0.05 level. The contrast from the second design variable is 1.931 (Medium versus Sport/Utility), implying a trend toward greater odds of low safety as size decreases. However, the 95% confidence interval (0.343, 15.182) contains 1 and therefore the contrast is not statistically significant.

4. Performing Backward Elimination and Prediction

Using the **STAT1.safety** data set, run PROC LOGISTIC and use backward elimination. Start with a model using **only main effects**. 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. Use the **SIZEFMT** format for the variable **Size**. Specify **Region** and **Size** as classification variables using reference cell coding and specify **Asia** as the reference level for **Region** and **Small** as the reference level for **Size**. Use a **UNITS** statement with -1 as the units for weight, so that you can see the odds ratio for lighter cars over heavier cars. Request any relevant plots.

- Open the **Binary Logistic Regression** task under Statistics.
- On the DATA tab, select the **Safety** data set.

- c. Assign **Unsafe** as the response variable and the event of interest to **1**.
- d. Assign **Region** and **Size** as the classification variables and select Reference coding to parameterize effects.
- e. Assign **Weight** as the continuous variable.
- f. On the MODEL tab, specify the model with only main effects.
- g. On the SELECTION tab, select **Backward elimination** as the selection method.
- h. Open the editor to specify using ‘**Asia**’ and ‘**1**’ as the reference levels for Region and Size, respectively.
- i. Add a **UNITS** statement for odds ratio calculations and a **STORE** statement to save the analysis results.
- j. Run the code.

Note: Alternatively, you can write the code directly.

Note: Notice that the reference level for **Size** is set to ‘Small’ in the program below, rather than ‘1’. When a format is applied to a CLASS statement variable, the reference level option should refer to the formatted value and not the internal value.

```
/*st107s04.sas*/
ods graphics on;
proc logistic data=STAT1.safety plots(only)=(effect oddsratio);
  class Region (param=ref ref='Asia')
    Size (param=ref ref='Small');
  model Unsafe(event='1') = Weight Region Size
    / clodds=pl selection=backward;
  units Weight = -1;
  store isSafe;
  format Size sizefmt. ;
  title 'Logistic Model: Backwards Elimination';
run;
```

- k. Which terms appear in the final model? **Only Size appears in the final model.**

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Region	1	2	0.4506	0.5020
2	Weight	1	1	2.1565	0.1420

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Size	2	24.2875	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.6506	0.3561	3.3377	0.0677
Size	Large	1	-3.3585	0.8125	17.0880	<.0001
Size	Medium	1	-2.2192	0.6070	13.3654	0.0003

- a. Do you think this is a better model than the one fit with only Region? Comparing the model fit statistics, you see that the AIC (92.629) and SC (100.322) are both smaller in the model fit by the backward elimination method, 119.854 and 124.982 respectively. This indicates that the Size only model is doing better than the Region only model. Using the c statistics, you can also see improvement beyond the Region only model, 0.818 previously 0.598.
- b. Using the final model, chosen by backward elimination, and the STORE statement, generate predictive probabilities for the cars in the following DATA step code.

```
data checkSafety;
length Region $9.;
  input Weight Size Region $ 5-13;
  datalines;
4 1 N America
3 1 Asia
5 3 Asia
5 2 N America
;
run;

proc plm restore=isSafe;
  score data=checkSafety out=scored_cars / ILINK;
  title 'Safety Predictions using PROC PLM';
run;

proc print data=scored_cars;
run;
```

Obs	Region	Weight	Size	Predicted
1	N America	4	Small	0.65714
2	Asia	3	Small	0.65714
3	Asia	5	Large	0.06251
4	N America	5	Medium	0.17241

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

7.01 Multiple Answer Poll – Correct Answer

Which of the following would likely not be considered categorical in the data?

- a. Bonus
- b. Fireplaces
- c. Basement_Area
- d. Lot_Shape_2
- e. SalePrice

7.02 Multiple Answer Poll – Correct Answers

What tends to happen when sample size decreases?

- a. The chi-square value increases.
- b. The p-value increases.
- c. Cramer's V increases.
- d. The Odds Ratio increases.
- e. The width of the CI for the Odds Ratio increases.

7.03 Multiple Answer Poll – Correct Answers

A researcher wants to measure the strength of an association between two binary variables. Which statistic(s) can he use?

- a. Hansel and Gretel Correlation
- b. Mantel-Haenszel Chi-Square
- c. Pearson Chi-Square
- d. Odds Ratio
- e. Spearman Correlation

7.04 Multiple Choice Poll – Correct Answer

What are the upper and lower bounds for a logit?

- a. Lower=0, Upper=1
- b. Lower=0, No upper bound
- c. No lower bound, No upper bound
- d. No lower bound, Upper=1

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

7.05 Multiple Choice Poll – Correct Answer

In the Analysis of Maximum Likelihood table, using effect coding, what is the estimated logit for someone at `IncLevel=2`?

- a. -.5363
- b. -.6717
- c. -.6659
- d. -.7563
- e. Cannot tell from the information provided



7.06 Multiple Choice Poll – Correct Answer

A variable coded 1, 2, 3, and 4 is parameterized with effect coding, with 2 as the reference level. The parameter estimate for level 1 tells you which of the following?

- a. The difference in the logit between level 1 and level 2
- b. The odds ratio between level 1 and level 2
- c. The difference in the logit between level 1 and the average of all levels
- d. The odds ratio between level 1 and the average of all levels
- e. Both a and b
- f. Both c and d



Appendix A References

A.1 References.....	A-3
---------------------	-----

A.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, NC: 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, "ANote 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.
- Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.
- Welch, B. L. 1951. "On the Comparison of Several Mean Values: An Alternative Approach." *Biometrika* 38:330–336.

Appendix B Sampling from SAS Data Sets

B.1 Random Samples B-3

B.1 Random Samples

Selecting Random Samples

The SURVEYSELECT procedure selects a random sample from a SAS data set.

```
PROC SURVEYSELECT DATA=name-of-SAS-data-set
    OUT=name-of-output-data-set
    METHOD=method-of-random-sampling
    SEED=seed-value
    SAMPSIZE=number of observations desired in sample
    ;
    <STRATA stratification-variable(s)>;
RUN;
```

Selected PROC SURVEYSELECT statement options:

- DATA= identifies the data set to be selected from.
- OUT= indicates the name of the output data set.
- METHOD= specifies the random sampling method to be used. For simple random sampling without replacement, use METHOD=SRS. For simple random sampling with replacement, use METHOD=URS. For other selection methods and details about sampling algorithms, see the SAS online documentation for PROC SURVEYSELECT.
- SEED= specifies the initial seed for random number generation. If no SEED option is specified, SAS uses the system time as its seed value. This creates a different random sample every time the procedure is run.
- SAMPSIZE= indicates the number of observations to be included in the sample. To select a certain fraction of the original data set rather than a given number of observations, use the SAMPRATE= option.

Selected SURVEYSELECT procedure statement:

- STRATA enables the user to specify one or more stratification variables. If no STRATA statement is specified, no stratification takes place.

Other statements and options for the SURVEYSELECT procedure can be found in the SAS online documentation.

Part A shows how to select a certain sample size using the SAMPSIZE= option.

1. Open the **Select Random Sample** task under Data.
2. On the DATA tab, specify to draw the sample from the **Safety** data set.
3. Change the name of the output data set to **SafetySample** under the OUTPUT DATASET property.
4. On the OPTIONS tab, specify the sample size as **12 Rows** and set the random seed **31475**.
5. Run the code.

Note: Alternatively, you can write the code directly in SAS.

```
/* st10bd01.sas */ /*Part A*/
proc surveyselect
  data= STAT1.Safety /* sample from data table */
  seed=31475           /* recommended that you use this option */
  method=srs          /* simple random sample */
  sampsize=12          /* sample size */
  out=work.SafetySample /* sample stored in this data set */
;
run;

proc print data=work.SafetySample;
run;
```

Note: If you do not provide a seed, you will not be able to reproduce the sample. It is recommended that you always include a seed when using PROC SURVEYSELECT.

Selection Method	Simple Random Sampling
-------------------------	------------------------

Input Data Set	SAFETY
Random Number Seed	31475
Sample Size	12
Selection Probability	0.125
Sampling Weight	8
Output Data Set	SAFETYSAMPLE

Obs	Unsafe	Size	Weight	Region	Type
1	0	2	3	N America	Medium
2	0	2	3	N America	Medium
3	0	2	3	Asia	Medium
4	0	2	3	N America	Medium
5	0	3	4	N America	Large
6	0	3	6	N America	Sport/Utility
7	1	1	3	Asia	Sports
8	0	3	4	N America	Large
9	0	1	4	Asia	Sports
10	0	2	3	N America	Medium
11	1	1	2	Asia	Small
12	0	3	5	Asia	Sport/Utility

Part B shows how to select a certain percentage of the original sample.

1. Open the **SELECTION RANDOM SAMPLE** task from part A.
2. On the OPTIONS tab, select the **Percent of rows** option and specify to sample **5** percent of the original data.
3. Run the code.

Note: Alternatively, you can write the code directly and use the SAMPRATE= option.

```
/* st10bd01.sas */ /*Part B*/
proc surveyselect
  data= STAT1.Safety /* sample from data table */
  seed=31475           /* recommended that you use this option */
  method=srs          /* simple random sample */
  samprate=0.05        /* sample size */
  out=work.SafetySample /* sample stored in this data set */
;
run;
proc print data=work.SafetySample;
run;
```

Selection Method	Simple Random Sampling
------------------	------------------------

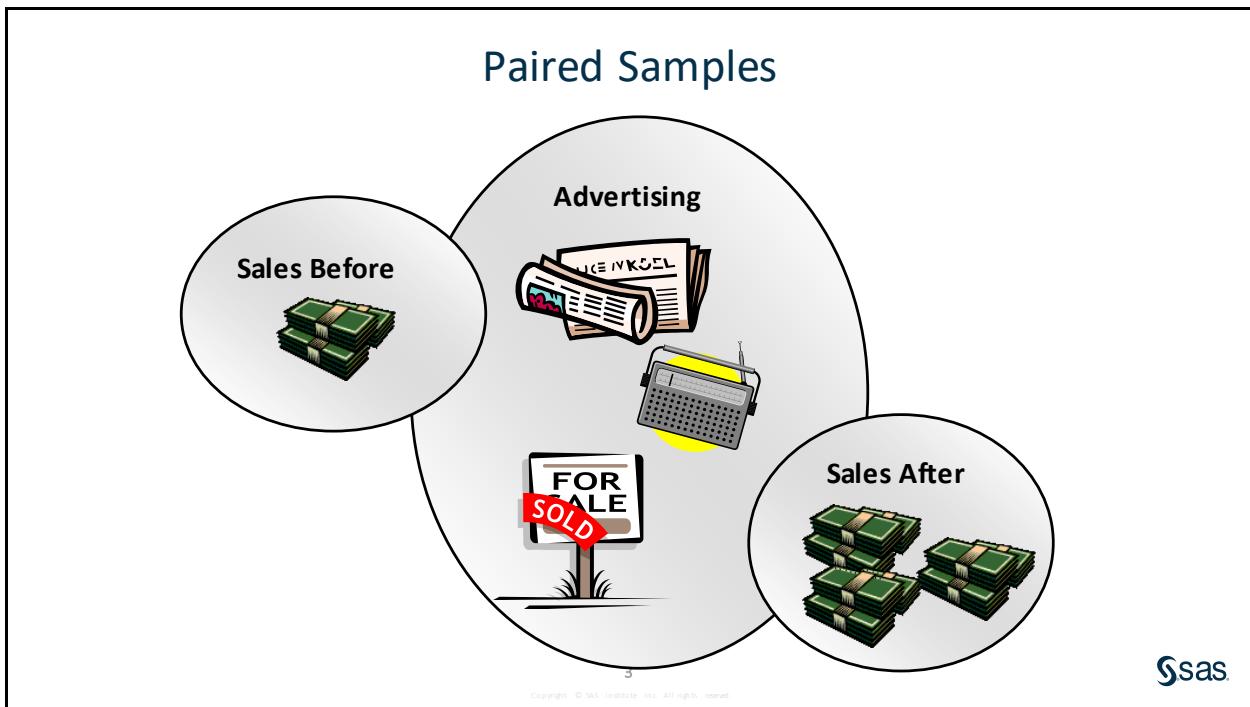
Input Data Set	SAFETY
Random Number Seed	31475
Sampling Rate	0.05
Sample Size	5
Selection Probability	0.052083
Sampling Weight	19.2
Output Data Set	SAFETYSAMPLE

Obs	Unsafe	Size	Weight	Region	Type
1	0	2	3	N America	Medium
2	0	2	3	N America	Medium
3	1	1	3	Asia	Sports
4	0	2	3	N America	Medium
5	1	1	4	N America	Sports

Appendix C Additional Topics

C.1 Paired <i>t</i>-Tests.....	C-3
Demonstration: Paired <i>t</i> -Test.....	C-5
C.2 One-Sided <i>t</i>-Tests.....	C-11
Demonstration: One-Sided <i>t</i> -Test.....	C-13
C.3 Nonparametric ANOVA.....	C-15
Demonstration: The NPAR1WAY Procedure for Hospice Referral Data.....	C-21
Demonstration: The NPAR1WAY Procedure for Small Samples.....	C-31
C.4 Partial Regression Plots.....	C-34
Demonstration: Partial Regression Plots	C-36
C.5 Exact Tests for Contingency Tables	C-40
Demonstration: Fisher's Exact <i>p</i> -Values for the Pearson Chi-Square Test.....	C-45
C.6 Empirical Logit Plots.....	C-47
Demonstration: Fisher's Exact <i>p</i> -Values for the Pearson Chi-Square Test.....	C-51
Solutions to Student Activities (Polls/Quizzes)	C-57

C.1 Paired t-Tests



For many types of data, repeat measurements are taken on the same subject throughout a study. The simplest form of this study is often referred to as the *paired t-test*.

In this study design,

- subjects are exposed to a treatment, for example, an advertising strategy
- a measurement is taken of 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 the following:

$$H_0: \mu_{\text{POST}} = \mu_{\text{PRE}}$$

$$H_1: \mu_{\text{POST}} \neq \mu_{\text{PRE}}$$

The TTEST Procedure

General form of the TTEST procedure:

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

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 were collected both before and after a particular advertising campaign. You are interested in determining the effect of the campaign on sales. You 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.

1. Open the **List Data** task under Data.
2. Select the **Market** data set.
3. On the OPTIONS tab, select the option to display only the first n rows and set n to **20**.
4. Run the code.

```
/*st10cd01.sas*/ /*Part A*/
proc print data=STAT1.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

Paired t- Test

5. Open the **t Tests** task under Statistics.
6. On the DATA tab, select the **Market** data set.
7. Specify to conduct a **Paired test**.
8. Select **post** as Group 1 variable and **pre** as Group 2 variable.
9. On the OPTIONS tab, clear the option to conduct tests for normality.
10. Select the option to display selected plots and check to display all the plots.
11. Run the code.

Note: Alternatively, you can write the code directly and use the PAIRED statement to test whether the mean of post-sales is significantly different from the mean of pre-sales.

```
/*st10cd01.sas*/ /*Part B*/
proc ttest data=STAT1.MARKET plots(only) showh0=(summaryPlot
    intervalPlot qqplot agreement profiles);
    paired post*pre;
run;
```

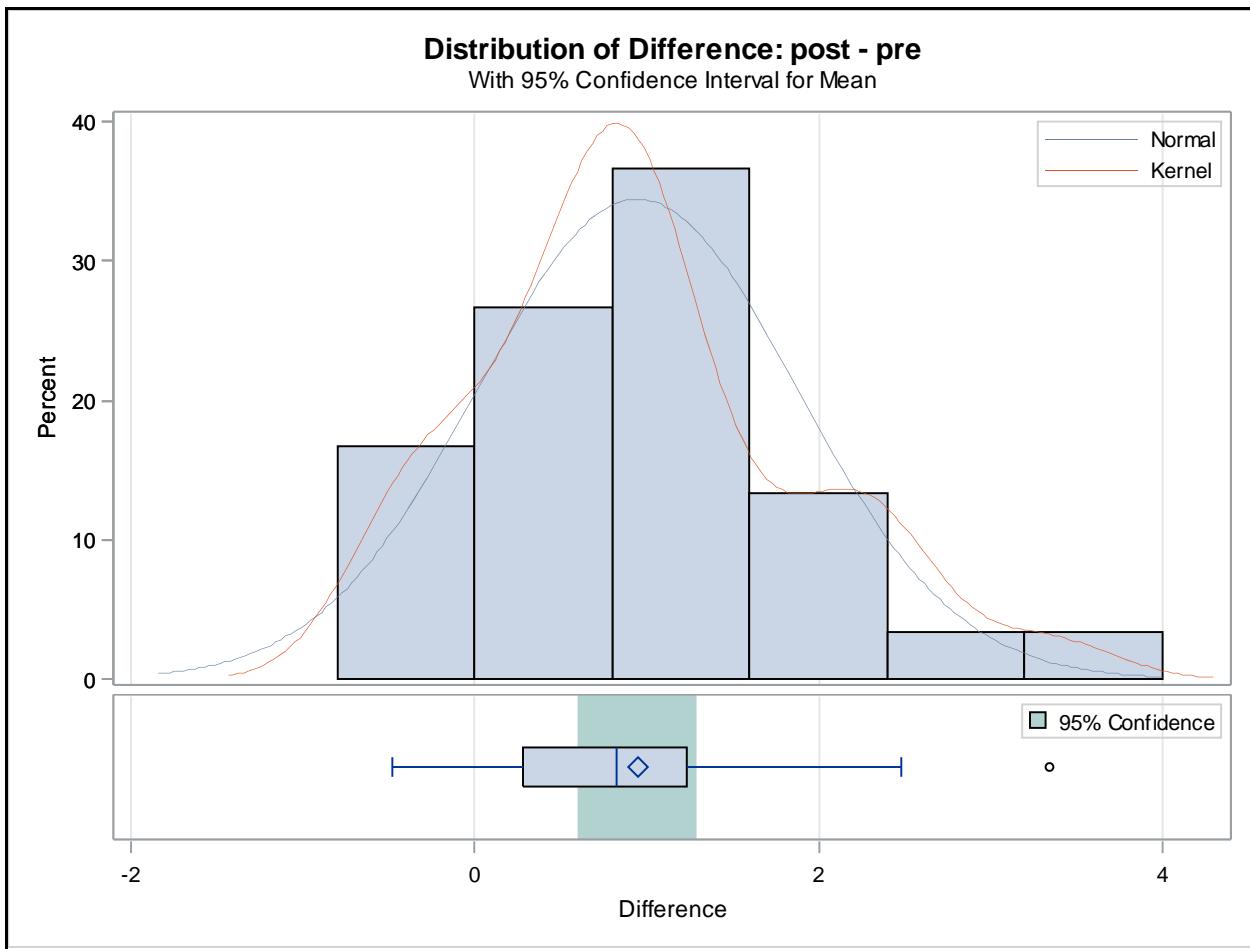
N	Mean	Std Dev	Std Err	Minimum	Maximum
30	0.9463	0.9271	0.1693	-0.4800	3.3400

The Mean in this table refers to the difference of **Post** minus **Pre**. That ordering is specified in the PAIRED statement.

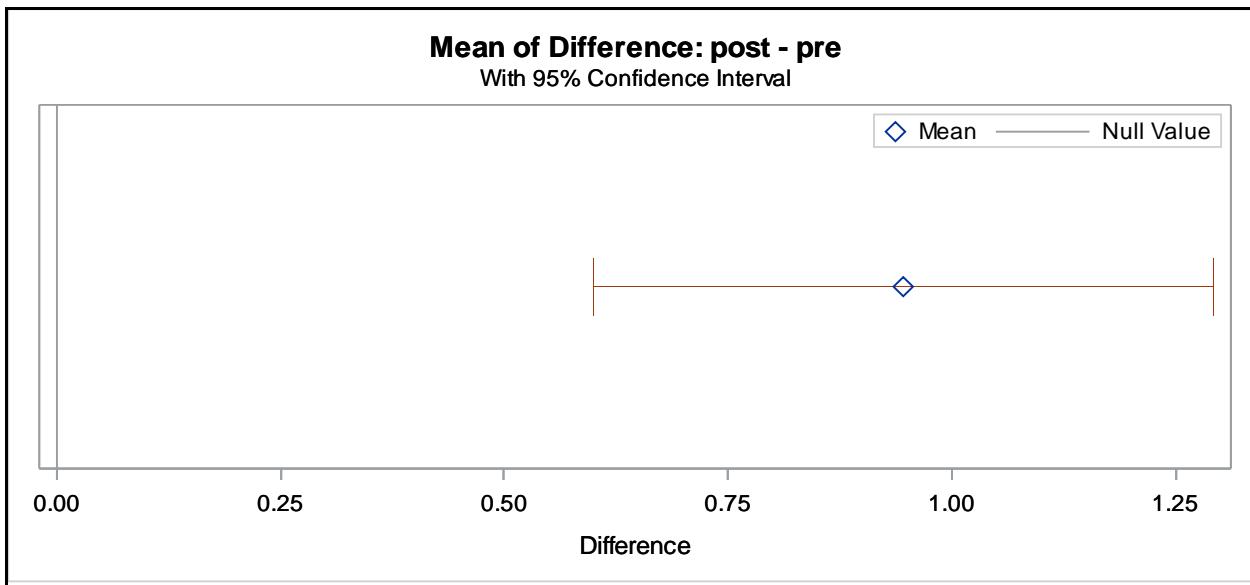
Mean	95% CL Mean	Std Dev	95% CL Std Dev
0.9463	0.6001	1.2925	0.9271 0.7384 1.2464

DF	t Value	Pr > t
29	5.59	<.0001

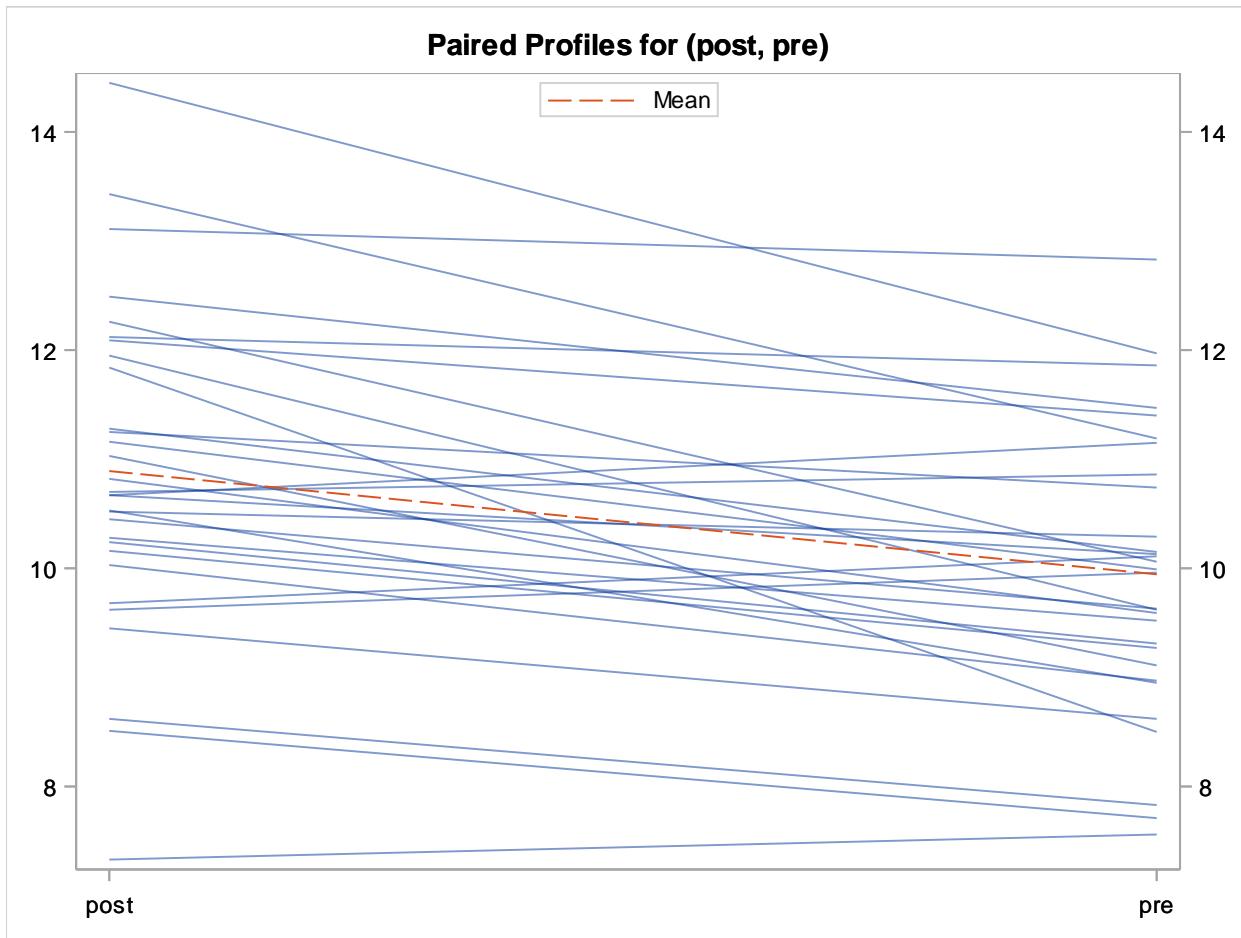
The T Tests table provides the requested analysis. The *p*-value for the difference **post–pre** is less than 0.0001. Assuming that you want a 0.01 level of significance, you reject the null hypothesis and conclude that there is a change in the average sales after the advertising campaign. Also, based on the fact that the mean is positive 0.9463, there appears to be an increase in the average sales after the advertising campaign.



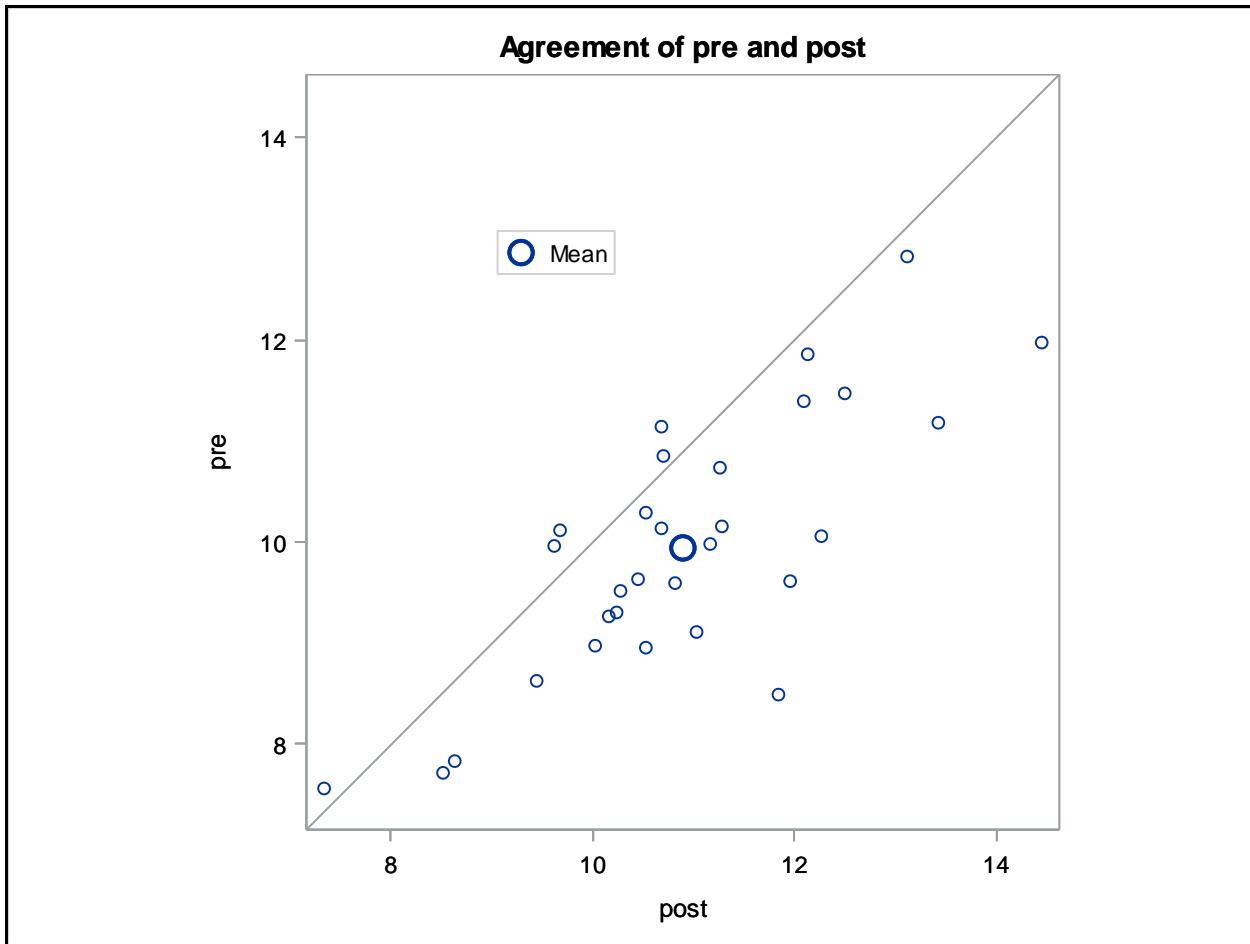
The difference scores seem approximately normally distributed in the histogram.



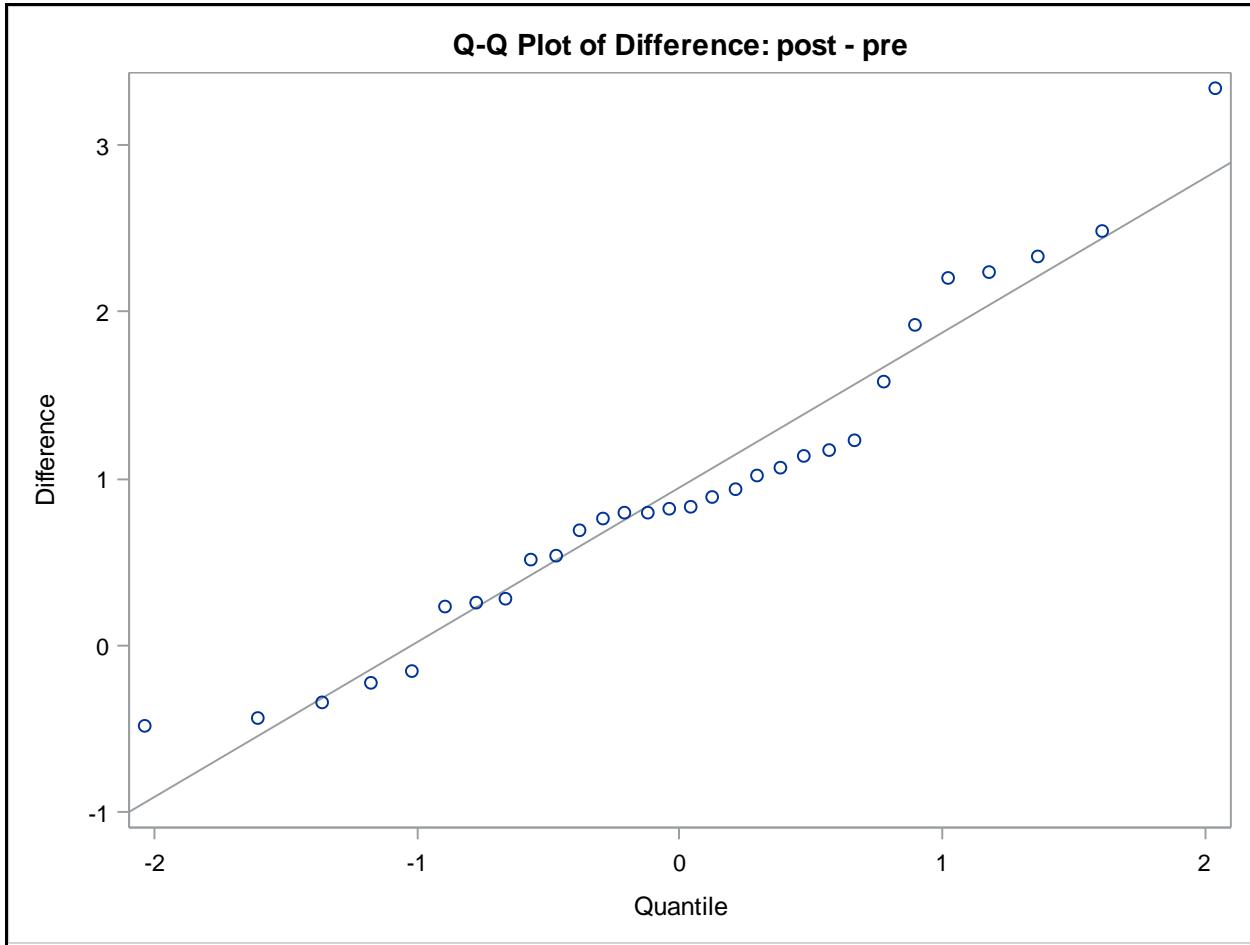
The difference is statistically significant from the (default) value of 0.



The Paired Profiles plot shows each observation pair as well as the pair of means.



The agreement plot shows most pairs lie to the lower left of the diagonal reference line, representing equality between pre and post measurements. This plot shows that not only is the mean greater for post than for pre, but that relationship holds true in most pairs.



The q-q plot confirms the assumption of normality of difference scores.

End of Demonstration

C.2 One-Sided t-Tests

One-Sided Tests and Confidence Intervals

- Used when the null hypothesis is one of these forms:
 - $H_0: \mu \leq k$
 - $H_0: \mu \geq k$
- Can increase power
- Tests and confidence intervals produced in PROC TTEST using the following:
 - SIDES=U for Upper Tail Tests ($\mu_0 \leq k$) and CIs
 - SIDES=L for Lower Tail Tests ($\mu_0 \geq k$) and CIs

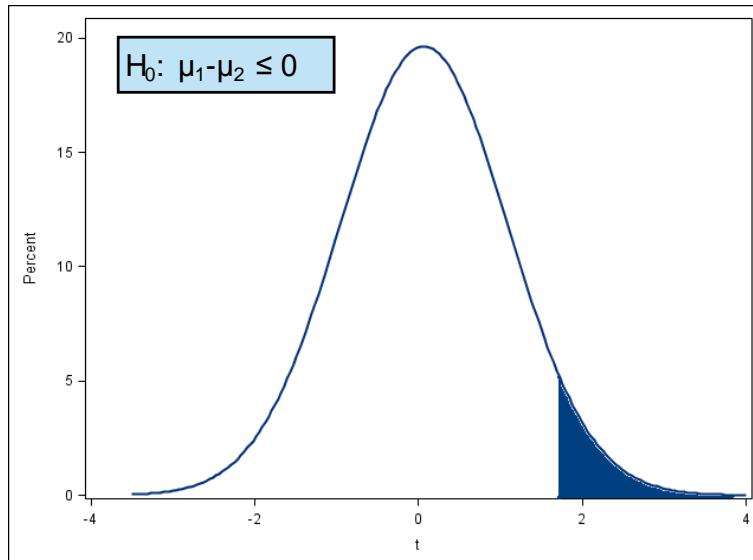
7



In many situations that you might decide that rejection on only one side of the mean is important. For example, a drug company might only want to test for positive differences between a new drug and a placebo and not negative differences. One-sided tests are a way doing this.

In the exercise data, the researcher might have been only curious in seeing the improvement in change scores due to the intervention and not considered the possibility that intervention would actually harm student performance.

One-Sided t-Test (Upper Tail)



sas

For two-sample upper-tail t -tests, the null hypothesis is one not only of equivalence, but also of difference between two means. If you believe that the mean change of the treatment group is strictly greater than the mean change of the control group, this implies that you believe that the difference between the mean changes for (Treatment-Control) is strictly greater than zero. That would then be your alternative hypothesis, $H_1: \mu_1 - \mu_2 > 0$. The null hypothesis is then, $H_0: \mu_1 - \mu_2 \leq 0$. Only t values above zero can achieve statistical significance. The critical t value for significance on the upper end will be smaller than it would have been in a two-sample test. Therefore, if you are correct about the direction of the true difference, you would have more power to detect that significance using the one-sided test. Confidence intervals for one-sided upper-tail tests always have an upper bound of infinity (no upper bound).

The $H_0=$ option in PROC TTEST allows other values for the null hypothesis.



One-Sided t-Test

1. Open the **t Tests** task under **Statistics**.
2. On the DATA tab, select the **German** data set.
3. Select the option to conduct **Two-sample** test.
4. Assign **Change** as the analysis variable and **Group** as the groups variable.
5. On the OPTIONS tab, use the drop-down menu to specify the direction of the one-sided tests.
6. Clear the selection for conducting tests for normality.
7. Select the option to display only selected plots and select to only display the confidence interval plot.
8. Run the code.

Note: Alternatively, you can write the code directly in SAS.

```
/*st10cd02.sas*/
proc ttest data=STAT1.German
            plots(only shownull)=interval h0=0 sides=L;
  class Group;
  var Change;
  title "One-Sided t-Test Comparing Treatment to Control";
run;
```

Note: H0=0 is the default, but is written here explicitly for completeness. SIDES=L declares this to be a lower one-sided *t*-test. Because **Control** comes before **Treatment** in the alphabet, the difference score in PROC TTEST will be for **Control** minus **Treatment** by default.

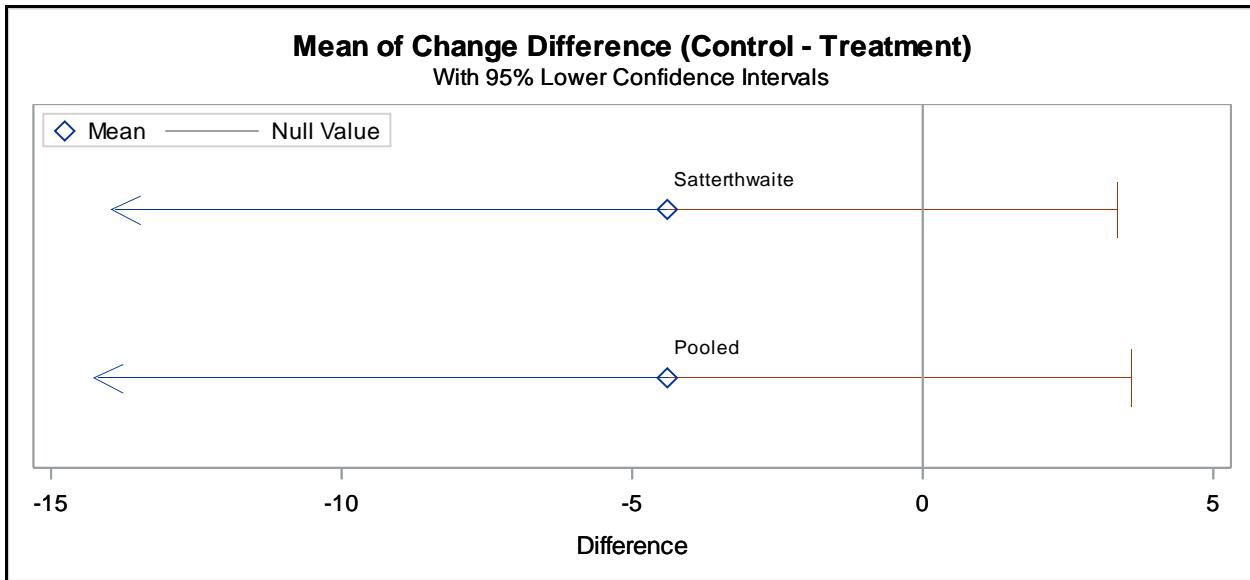
Group	N	Mean	Std Dev	Std Err	Minimum	Maximum
Control	13	6.9677	8.6166	2.3898	-6.2400	19.4100
Treatment	15	11.3587	14.8535	3.8352	-17.3300	32.9200
Diff (1-2)		-4.3910	12.3720	4.6882		

Group	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
Control		6.9677	1.7607	12.1747	8.6166
Treatment		11.3587	3.1331	19.5843	14.8535
Diff (1-2)	Pooled	-4.3910	-Infty	3.6052	12.3720
Diff (1-2)	Satterthwaite	-4.3910	-Infty	3.3545	

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	26	-0.94	0.1788
Satterthwaite	Unequal	22.947	-0.97	0.1707

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	12	2.97	0.0660

Notice that the confidence limits for the difference between **Control** and **Treatment** are different from that in the exercise, even though the Mean Diff is exactly the same. The lower confidence bound for the difference is now Infty (Infinity). For right-sided tests, the lower bound would be infinite in the positive direction. The *p*-value for the pooled variance test of the difference between **Control** and **Treatment** is now 0.1788, which is half of what it was in the two-sided test.



The Difference Interval plot reflects the one-sided nature of the analysis. The arrows pointing left represent the infinite confidence bound.

Note: The determination of whether to perform a one-sided test or a two-sided test should be made before any analysis or glancing at the data, and should be made bases on subject-matter considerations and not statistical power considerations.

End of Demonstration

C.01 Multiple Choice Poll

What justifies the choice of a one-sided test versus a two-sided test?

- a. The need for more statistical power
- b. Theoretical and subject-matter considerations
- c. A two-sided test that is nonsignificant
- d. The need for an unbiased test statistic

10

Copyright © SAS Institute Inc. All rights reserved.

C.3 Nonparametric ANOVA

This section addresses nonparametric options in 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.

13

Copyright © SAS Institute Inc. All rights reserved.

Nonparametric tests are most often used when the normality assumption required for analysis of variance is in question. Although ANOVA is robust with regard to minor departures from normality, extreme departures can make the test less sensitive to differences between means. Therefore, when the data is markedly 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.

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				

Sas

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					

15



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.

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 did not identify 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;
```

17


Copyright © SAS Institute Inc. All rights reserved.

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?



18


Copyright © SAS Institute Inc. All rights reserved.

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?



19

Copyright © SAS Institute Inc. All rights reserved.

Sas

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 cabinets. 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 for Hospice Referral Data

Example: A portion of Ms. Skarzynski's data about the hospice marketing visits is in the **STAT1.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 before the visit
ref2p	the number of referrals two months before the visit
ref1p	the number of referrals one month before 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 before the visit
diff1	the difference between the number of referrals one month after the visit and the average number of referrals before the visit
diff2	the difference between the number of referrals two months after the visit and the average number of referrals before the visit
diff3	the difference between the number of referrals three months after the visit and the average number of referrals before the visit
diffbys1	the difference between the number of referrals one month after the visit and the number of referrals three months before the visit
diffbys2	the difference between the number of referrals two months after the visit and the number of referrals three months before the visit
diffbys3	the difference between the number of referrals three months after the visit and the number of referrals three months before the visit.

Print a subset of the variables for the first 10 observations in the data set.

1. Open the **List Data** task under Data.
2. Select the **Hosp** data set.
3. Select the variables **visit**, **code**, and **diffby3** as the variables to list.
4. On the OPTIONS tab, modify the option to list First 10 rows.

5. Run the code.

Note: Alternatively, you can write the code directly. The code below includes additional FORMAT options.

```
/*st10cd03.sas*/ /*Part A*/
proc format;
  value vstfmt
    0='staff only'
    1='physician';
  value spcfmt
    1='oncologist'
    2='internal med'
    3='family prac'
    4='pulmonolgist'
    5='other special';
run;

proc print data= STAT1.hosp (obs=10);
  var visit code diffbys3;
  format visit vstfmt. code spcfmt. ;
run;
```

Note: The table listed here displays the formatted values.

Obs	visit	code	diffbys3
1	physician	familyprac	0
2	physician	familyprac	1
3	physician	oncologist	-1
4	physician	familyprac	-3
5	physician	oncologist	1
6	physician	familyprac	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.

1. Open the **Distribution Analysis** task under Statistics.
2. Assign **diffbys3** as the analysis variable.
3. On the OPTIONS tab, in the EXPLORING DATA list, under Histogram, assign **code** as the classification variable.
4. Select the options to include a normal curve, a kernel density estimate, and inset statistics to the histogram.

5. Expand the Inset Statistics sub-menu under Add inset statistics and select the mean, standard deviation, skewness, and kurtosis in the histogram.
6. Also check the box for **Normal probability plot** in the menu for CHECKING FOR NORMALITY.
7. A box will appear for **Add inset statistics**. Check that box.
8. Below **Normal quantile-quantile plot**, expand **Inset Statistics** and check **Mean, Standard deviation, Skewness, and Kurtosis** in the expanded list.
9. Open the editor and add goodness-of-fit test results as part of the inset statistics to the histograms (the code listed below in the INSET statement in the first PROC UNIVARIATE step, following the word, **kurtosis**).
10. Run the code.

Note: Edited code is shown below, adding a format for **code**, an NCOLS= option to show the panel of histograms as one row with 3 columns, and a title.

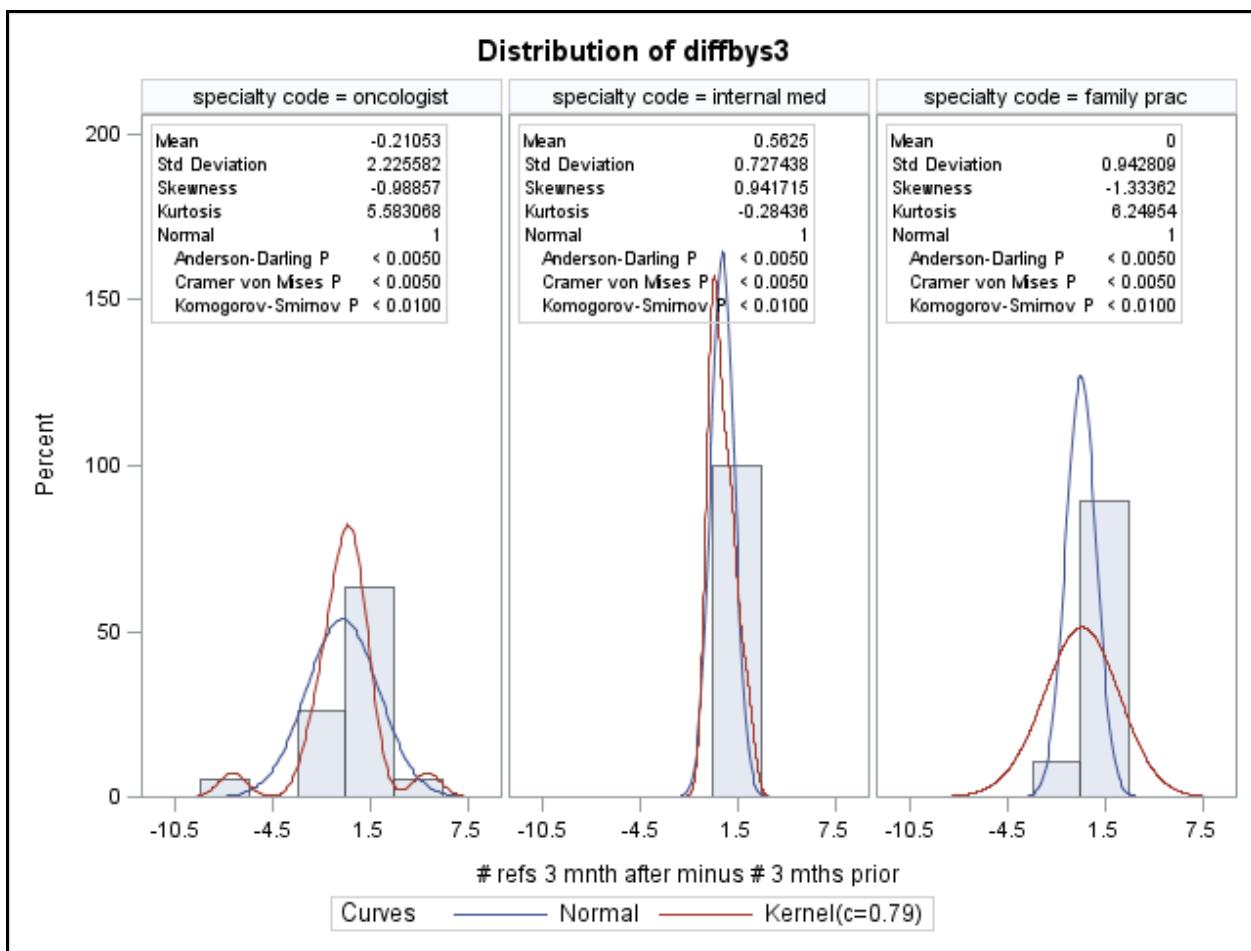
Note: Note that the code can be combined to run in one PROC UNIVARIATE step.

```
/*st10cd03.sas*/ /*Part B*/
proc univariate data=STAT1.hosp noprint;
ods select Histogram probplot;
class code;
var diffbys3;
histogram diffbys3 / normal kernel ncols=3;
inset mean std skewness kurtosis
    normal(adpval="Anderson-Darling P"
           cvmpval="Cramer von Mises P"
           ksdpval="Komogorov-Smirnov P");
probplot diffbys3 / normal ncols=3;
inset mean std skewness kurtosis;
title 'Descriptive Statistics for Hospice Data';
format code spcfmt.;

run;
```

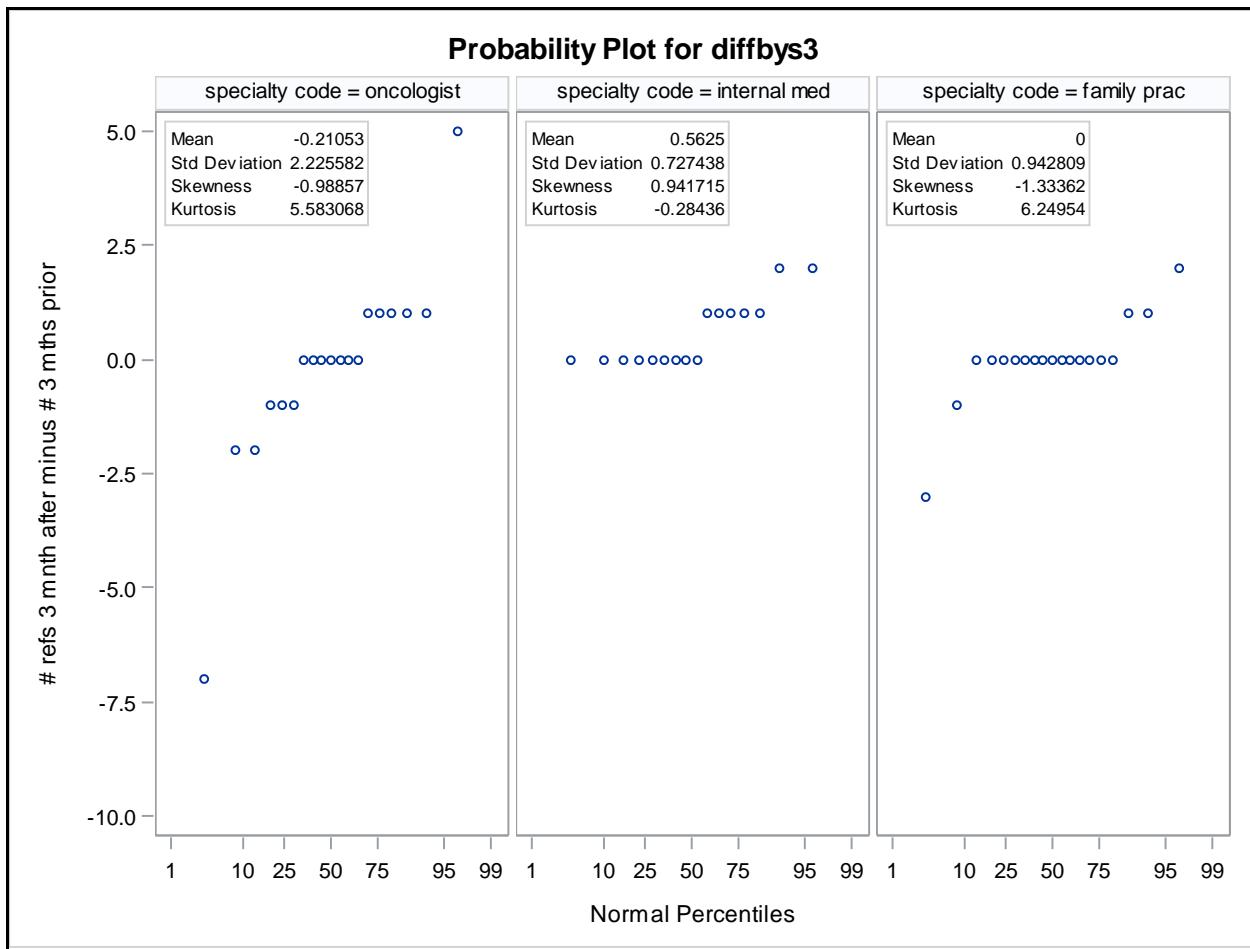
Examine the histograms and normal probability plots for each group.

Partial Output



Based on skewness and kurtosis, the oncologists and family practice doctors data might not be normal. All three goodness-of-fit tests reject the null hypothesis that the data is normal.

...



Internal medicine doctors appear to have only three values: 0, 1, and 2. The plots indicate that the data is not normal.

Family practice doctors appear to have mostly 0 values.

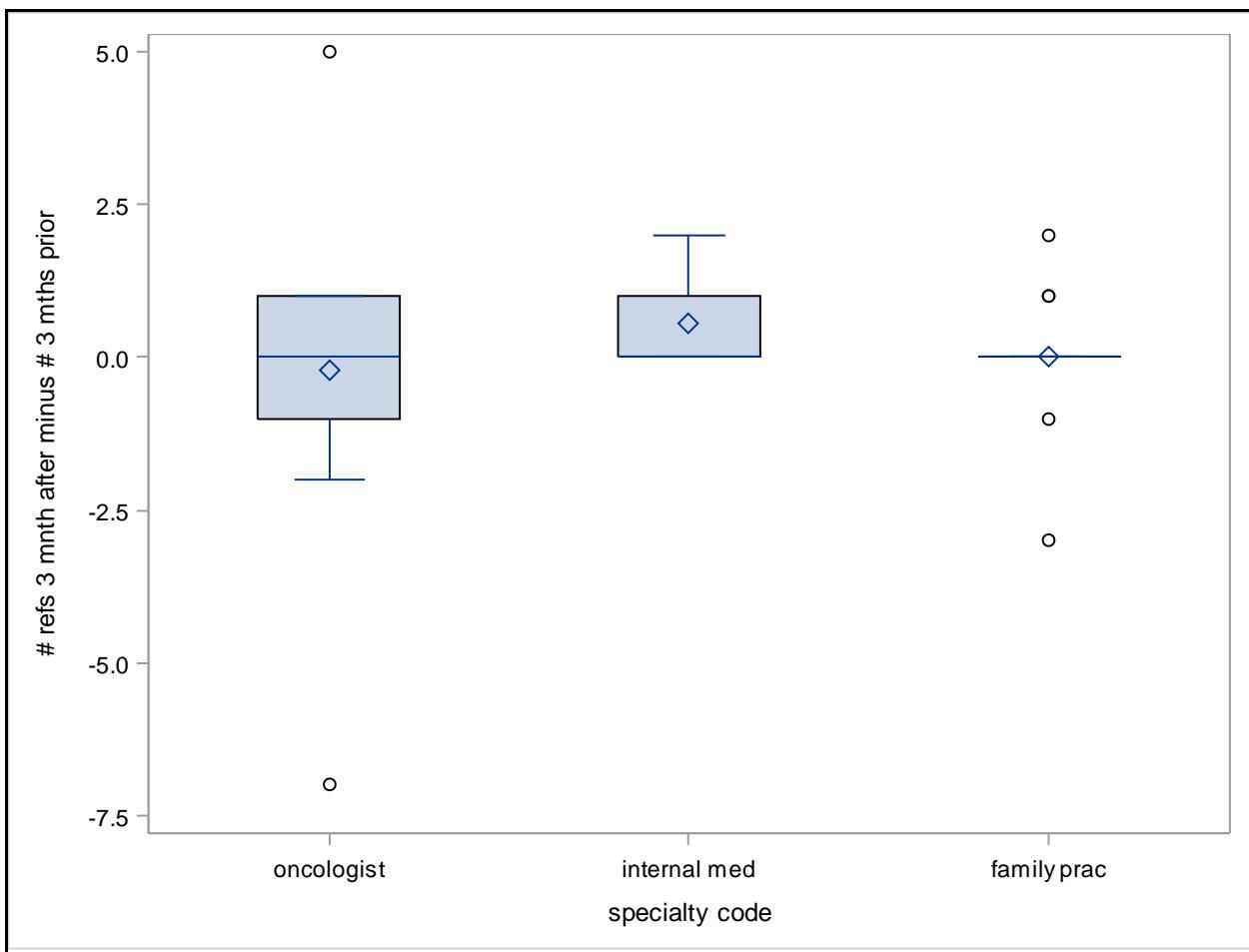
Both family practice doctors and oncologists have highly kurtotic distributions.

11. Open the **Box Plot** task under Graph.
12. On the DATA tab, select the **Hosp** data set.
13. Assign **diffbys3** as the analysis variable and **code** as the category variable.
14. Run the code.

Note: Alternatively, you can write the code directly.

```
/*st10cd03.sas*/ /*Part C*/
proc sgplot data=STAT1.hosp;
  vbox diffbys3 / category=code;
  format code spcfmt.;
run;
```

Now examine the PROC SGLOT output.



The box plots strongly support the conclusion 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.

15. Open the **Nonparametric One-Way ANOVA** task under Statistics.
16. Select the **Hosp** data set.
17. Assign **diffbys3** as the dependent variable and **code** as the classification variable.
18. On the OPTIONS tab, select the options to use the **Median scores** to perform the Median test in addition to the default Wilcoxon scores.
19. Run the code.

Note: Alternatively, you can write the code directly.

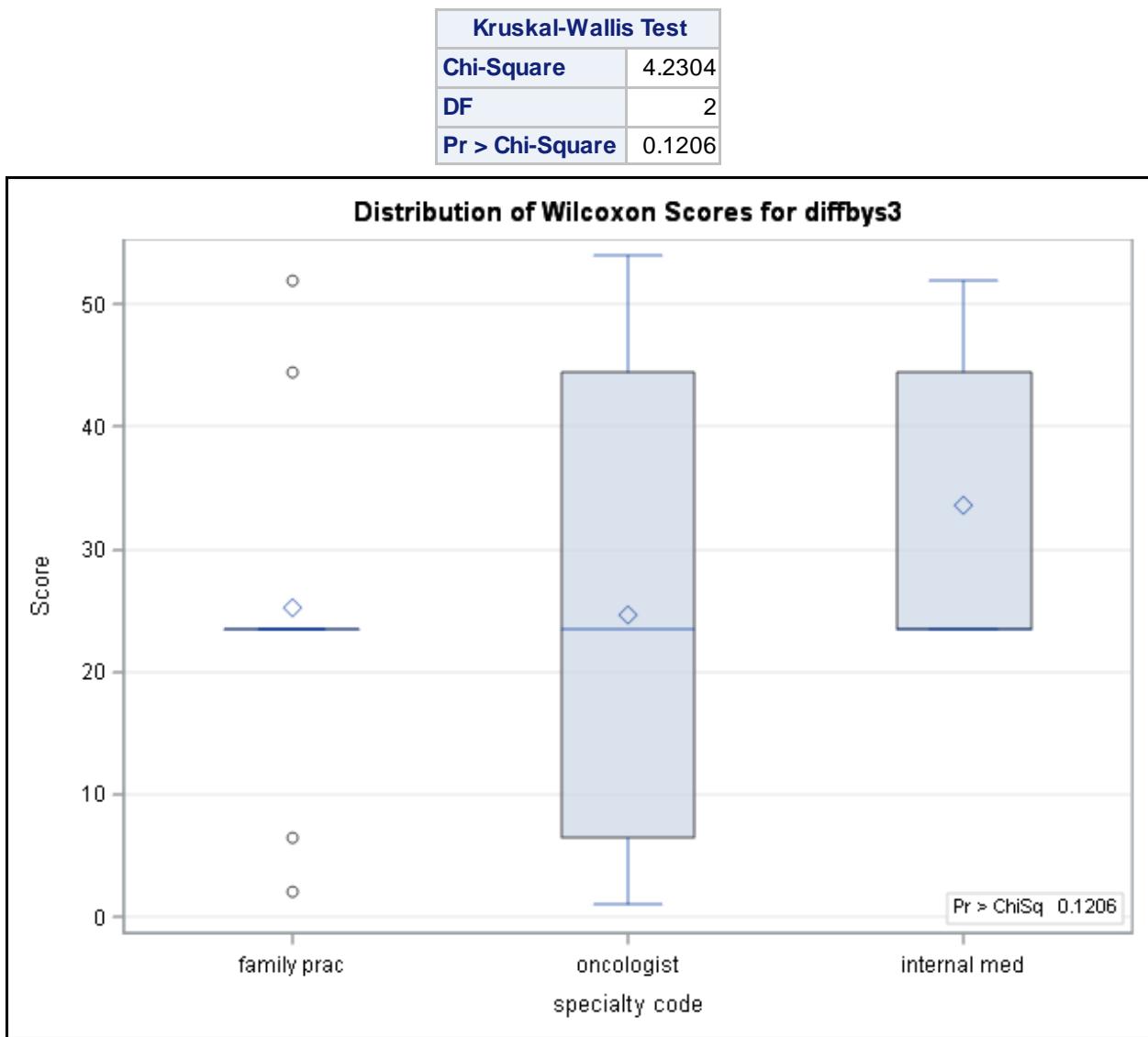
```
/*st10cd03.sas*/ /*Part D*/
proc npar1way data=STAT1.hosp wilcoxon median;
  class code;
  var diffbys3;
  format code spcfmt.;
run;
```

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.

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
family prac	19	478.50	522.50	49.907208	25.184211
oncologist	19	468.50	522.50	49.907208	24.657895
internal med	16	538.00	440.00	47.720418	33.625000
Average scores were used for ties.					



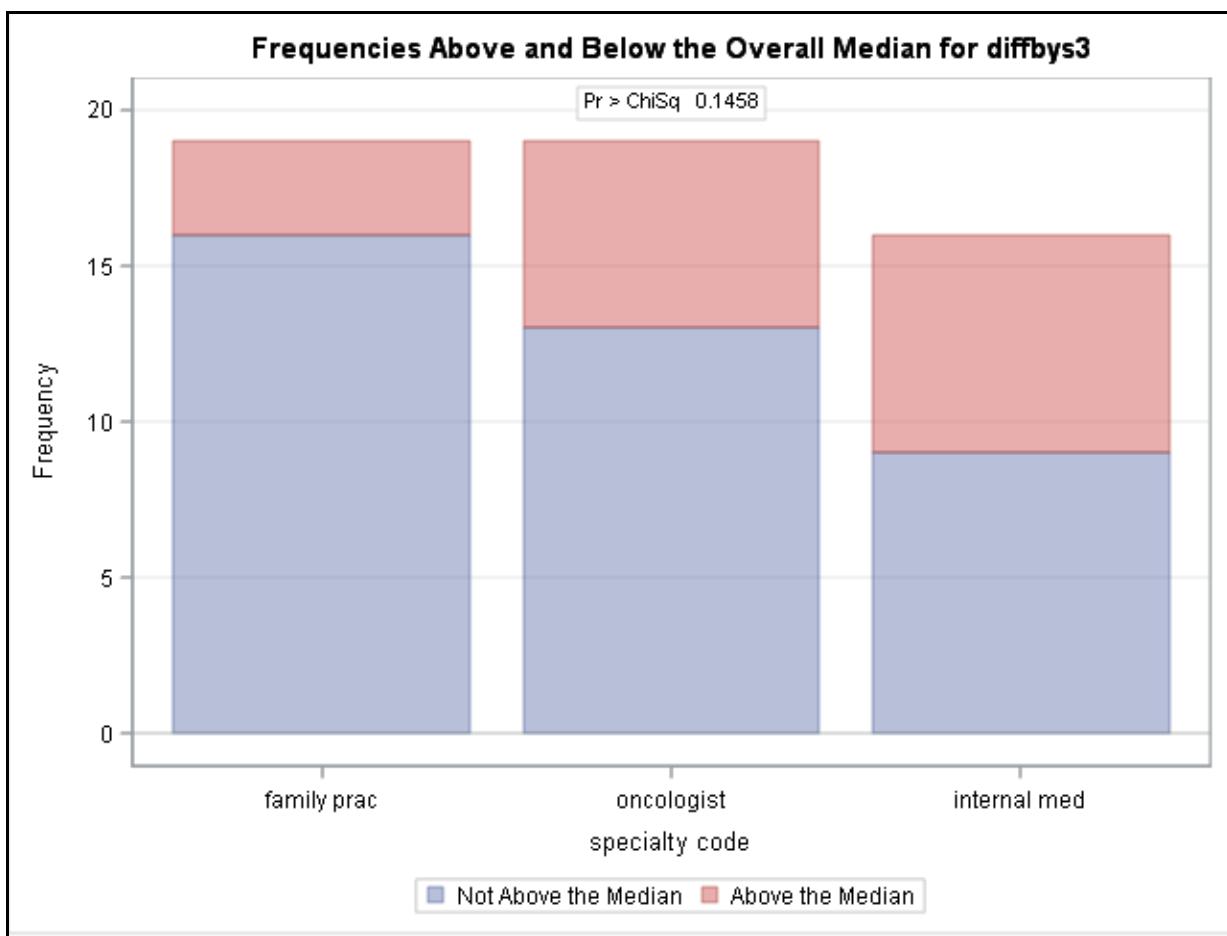
The PROC NPAR1WAY output from the WILCOXON option shows the actual sums of the rank scores and the expected sums of the rank scores if the null hypothesis is true. From the Kruskal-Wallis test (chi-square approximation), the p -value is .1206. Therefore, at the 5% level of significance, you do not reject 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

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
family prac	19	8.133333	9.50	1.232093	0.428070
oncologist	19	8.566667	9.50	1.232093	0.450877
internal med	16	10.300000	8.00	1.178106	0.643750

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 and shows the number of observations above and below the median for each group:

End of Demonstration



The NPAR1WAY Procedure for Small Samples

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 **STAT1.ven** data set.

1. Open the **List Data** task under Data.
2. Select the **Ven** data set.
3. Run the code.

Note: Equivalent code is shown below.

```
/*st10cd04.sas*/
proc print data=STAT1.ven;
  title 'Wood Veneer Wear Data';
run;
```

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 **Break** button in the SAS windowing environment or the **Stop** button in SAS Enterprise Guide, and choosing to stop computations.

4. Open the **Nonparametric One-Way ANOVA** task under Statistics.

5. Assign **wear** as the dependent variable and **brand** as the classification variable.
6. On the OPTIONS tab, use the drop-down menu and select the option to conduct Asymptotic and exact tests under the TESTS property.
7. Run the code.

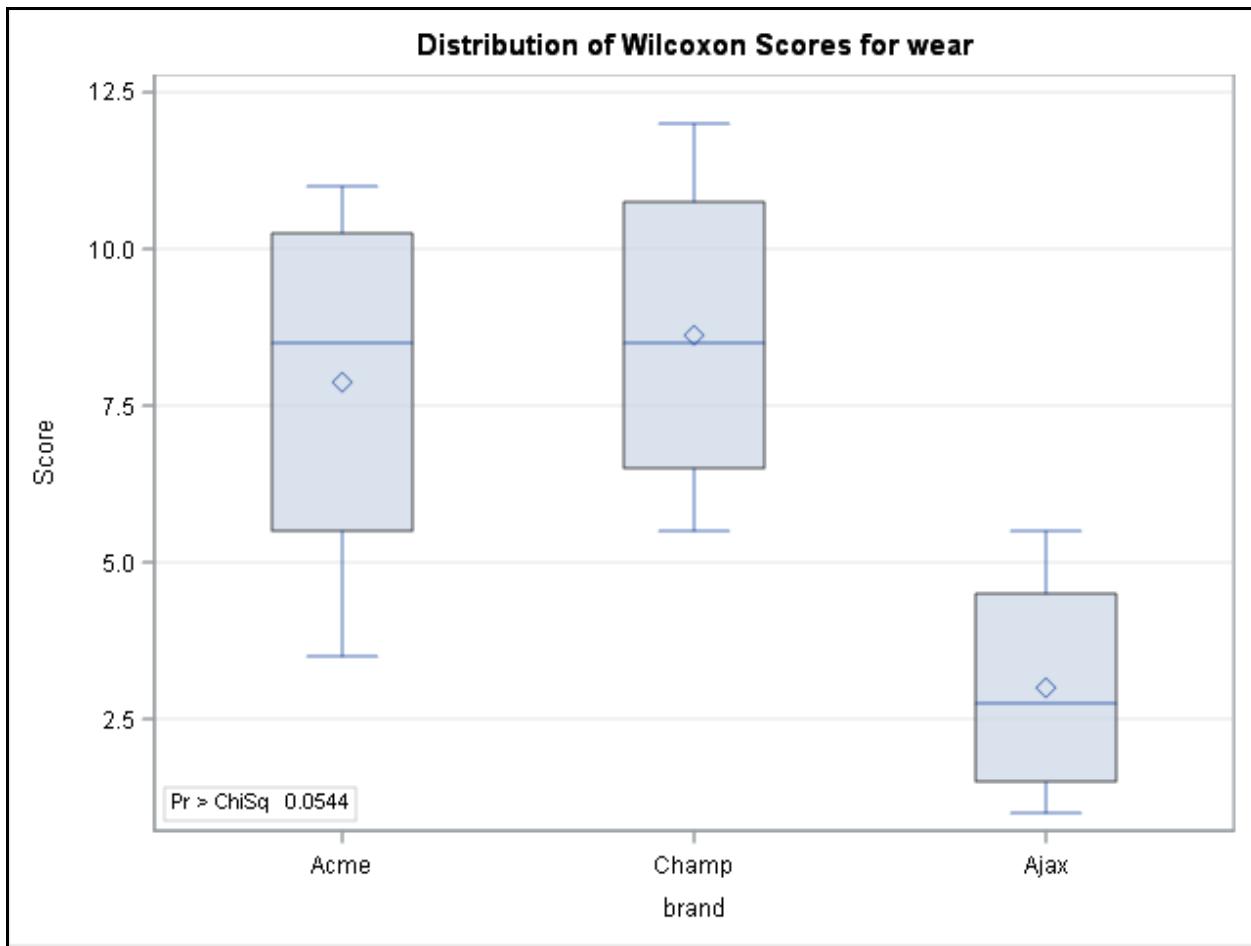
Note: Alternatively, you can write the code directly in SAS.

```
/*st10cd04.sas*/ /*Part B*/
proc npar1way data=STAT1.ven wilcoxon;
  class brand;
  var wear;
  exact;
run;
```

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 $\alpha=.05$. Notice the difference between the exact *p*-value and the (asymptotic) *p*-value based on the chi-square approximation.



End of Demonstration

C.4 Partial Regression Plots

Partial Regression Plots

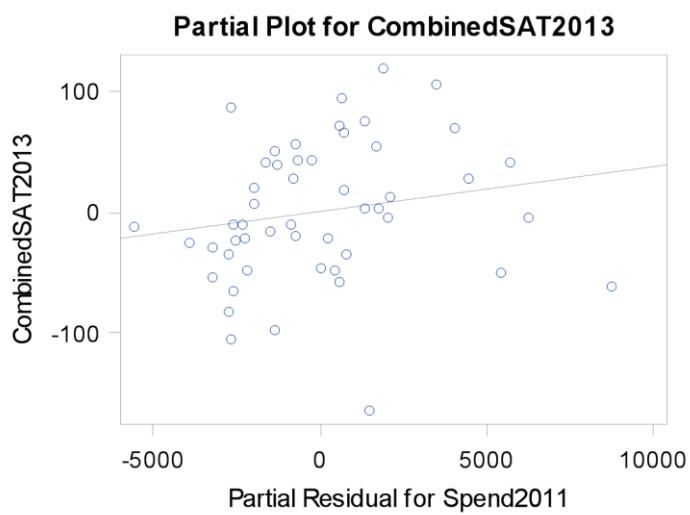
- Producing scatter plots of the response (Y) versus each of the possible predictor variables (the Xs) is recommended.
- However, in the multiple regression situation, these plots can be somewhat misleading because Y might depend on the other Xs not accounted for in the plot.
- Partial leverage plots compensate for this limitation of the scatter plots.

23



A *partial regression plot* is a graphical method for visualizing the test of significance for the parameter estimates in the full model. It is a plot of the residuals from two regression analyses.

Example of a Partial Regression Plot



24



Partial regression leverage plots are graphical methods that enable you to see the effect of a single variable in a multiple regression setting, controlling for the effect of all other variables.

Note: Partial regression plots are produced automatically with ODS Statistical Graphics when you specify the PLOTS=PARTIAL option in the PROC REG statement.

Partial Regression 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 regression plot for X2, do the following:

- 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.

In the example shown, there are three partial regression plots, one for each independent variable.

In general terms, for a partial regression plot of the independent variable X_r ,

- the vertical axis is the residuals from a regression of Y regressed on all Xs except X_r
- the horizontal axis is the residuals from a regression of X_r regressed on all other Xs.



Partial Regression Plots

Example: Generate and interpret partial regression plots for the full model and compare them to the fit plot from the simple regression model with **RunTime**.

1. Open the **Linear Regression** task under Statistics.
2. Select the **SAT** data set.
3. Assign **CombinedSAT2013** as the dependent variable and **Spend2011** as the continuous variable.
4. On the MODEL tab, specify the model with Spend2011 alone.
5. On the OPTIONS tab, clear the options to plot diagnostic plots and residual plots.
6. Expand the Scatter Plots property and clear the option to include the plot for observed values by predicted values.
7. Run the code.

Note: Alternatively, you can write the code directly in SAS.

```
/*st10cd05.sas*/ /*Part A*/
proc reg data=STAT1.SAT
plots(only)=fitplot;
model CombinedSAT2013 = Spend2011;
title 'Simple Regression';
run;
quit;
```

Selected PLOTS= options:

NOLIMITS

suppresses the display of confidence and prediction limits.

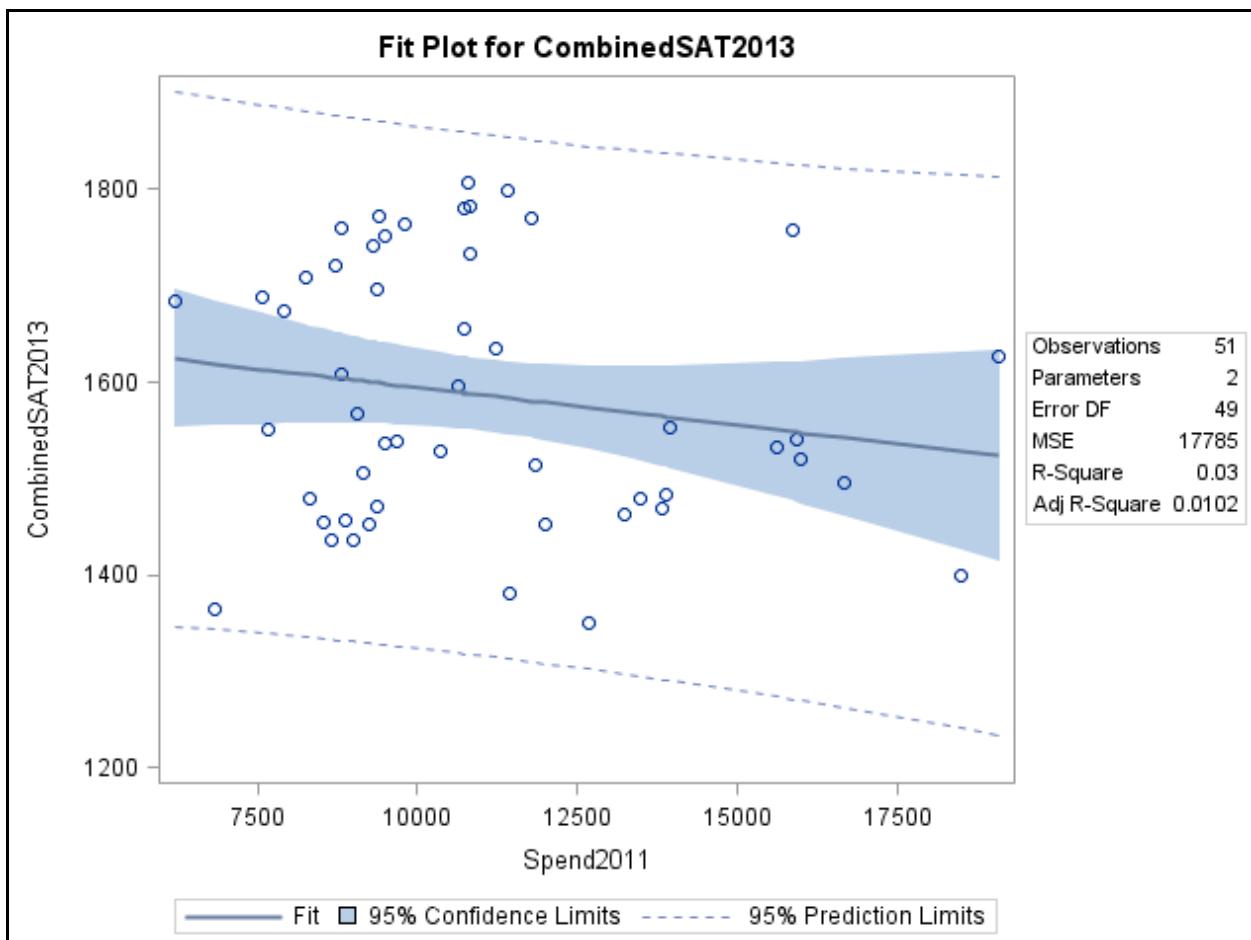
STATS=NONE

suppresses the display of the model statistics box.

Partial Output

Simple Regression
Model: MODEL1
Dependent Variable: CombinedSAT2013

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	1672.65384	72.34052	23.12	<.0001
Spend2011	1	-0.00782	0.00636	-1.23	0.2243



On its own, **Spend2011** is not a significant predictor of **CombinedSAT2013** with a *p*-value 0.2243. The observations seem highly variable about the regression line and the slope is negative (-0.00782).

Next run the same model, except add the **Participation2013** variable, which is the proportion of eligible high school seniors who have taken the SAT, as a regressor.

8. On the DATA tab of the same task, assign **Participation2013** as a second continuous variable.
9. On the MODEL tab, include **Participation2013** in the model effects.
10. On the OPTIONS tab, expand the Scatter Plots property and check the option to include partial regression plots for each explanatory variable.
11. Select the option to display the plots as individual plots.
12. Run the code

```
/*st10cd05.sas*/ /*Part B*/
proc reg data=STAT1.SAT
plots(only)=partial(unpack);
model CombinedSAT2013=Spend2011 Participation2013 / partial;
title 'Partial Regression Plots';
run;
quit;
```

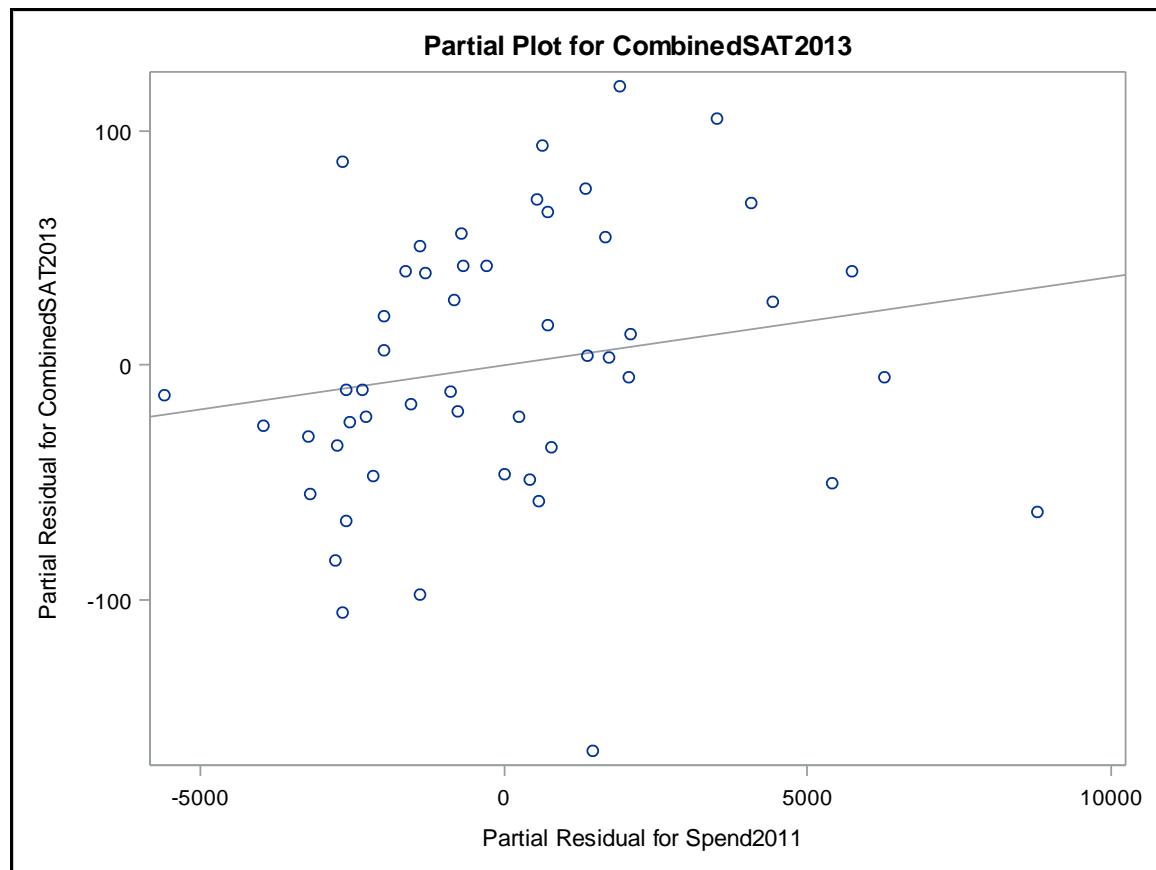
Selected MODEL statement option:

PARTIAL generates partial regression plots for all predictor variables in the model. If you also specify PLOTS=PARTIAL in the PROC REG statement, ODS Graphics are produced.

Partial Output

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	1693.46454	31.37863	53.97	<.0001
Spend2011	1	0.00375	0.00287	1.31	0.1967
Participation2013	1	-366.95940	25.14546	-14.59	<.0001

The parameter estimate for **Spend2011** is not significant in this model, either. The *p*-value is 0.1967. However, the sign of the parameter estimate has changed.



The plot shows this relationship graphically. The Y axis is now the partial residuals from regressing **CombinedSAT2013** on **Participation2013**. The X axis is the partial residuals from regressing **Spend2011** on **Participation2013**. The variance is much greater for observations around the partial regression line than for the simple regression line shown previously.

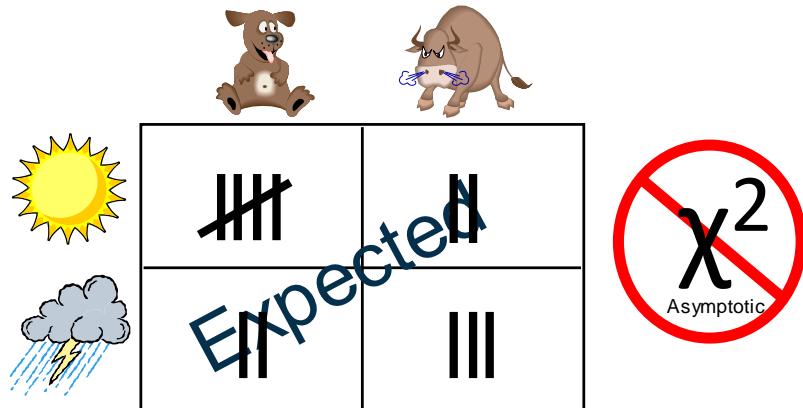
Note: In addition to enabling you to visualize the adjusted relationships in a multiple regression model, partial residual plots can help you detect potential outliers. For example, a potential influential outlier can be seen in the upper right corner of the plot. None was seen in the simple regression fit plot.

Note: If you use 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.

End of Demonstration

C.5 Exact Tests for Contingency Tables

When Not to Use the Asymptotic χ^2



When more than 20% of cells have expected counts less than five

28

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				
Row	Column			
Frequency		1	2	3
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

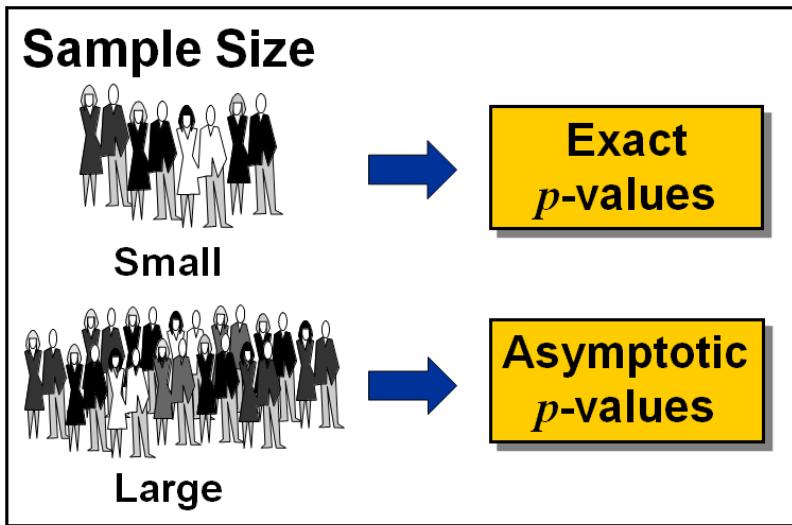
29



Copyright © SAS Institute Inc. All rights reserved.

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, has an *observed* count less than 5. However, 4 out of 9, or 44%, of the cells have *expected* counts less than 5. Therefore, the chi-square test might not be valid.

Small Samples – Exact p-Values



30



Copyright © SAS Institute Inc. All rights reserved.

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 this 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 as or more extreme than the one observed, just by chance.

Could the underlined sample values occur just by chance?

31

Copyright © SAS Institute Inc. All rights reserved.

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.

Note: Recall that expected count within each cell is calculated by expected count=(R*C)/T.

Exact p -Values for Pearson Chi-Square

Observed Table	Possible Table 2	Possible Table 3																											
<table border="1"> <tr><td>0</td><td>3</td><td>3</td></tr> <tr><td>2</td><td>2</td><td>4</td></tr> <tr><td>2</td><td>5</td><td>7</td></tr> </table>	0	3	3	2	2	4	2	5	7	<table border="1"> <tr><td>1</td><td>2</td><td>3</td></tr> <tr><td>1</td><td>3</td><td>4</td></tr> <tr><td>2</td><td>5</td><td>7</td></tr> </table>	1	2	3	1	3	4	2	5	7	<table border="1"> <tr><td>2</td><td>1</td><td>3</td></tr> <tr><td>0</td><td>4</td><td>4</td></tr> <tr><td>2</td><td>5</td><td>7</td></tr> </table>	2	1	3	0	4	4	2	5	7
0	3	3																											
2	2	4																											
2	5	7																											
1	2	3																											
1	3	4																											
2	5	7																											
2	1	3																											
0	4	4																											
2	5	7																											
$\chi^2=2.100$ prob=0.286	$\chi^2=0.058$ prob=0.571	$\chi^2=3.733$ prob=0.143																											
<i>Most likely, given marginal values</i>																													

32



Copyright © SAS Institute Inc. All rights reserved.

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 (0.571) 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	Possible Table 2	Possible Table 3																											
<table border="1"> <tr><td>0</td><td>3</td><td>3</td></tr> <tr><td>2</td><td>2</td><td>4</td></tr> <tr><td>2</td><td>5</td><td>7</td></tr> </table>	0	3	3	2	2	4	2	5	7	<table border="1"> <tr><td>1</td><td>2</td><td>3</td></tr> <tr><td>1</td><td>3</td><td>4</td></tr> <tr><td>2</td><td>5</td><td>7</td></tr> </table>	1	2	3	1	3	4	2	5	7	<table border="1"> <tr><td>2</td><td>1</td><td>3</td></tr> <tr><td>0</td><td>4</td><td>4</td></tr> <tr><td>2</td><td>5</td><td>7</td></tr> </table>	2	1	3	0	4	4	2	5	7
0	3	3																											
2	2	4																											
2	5	7																											
1	2	3																											
1	3	4																											
2	5	7																											
2	1	3																											
0	4	4																											
2	5	7																											
$\chi^2=2.100$ prob=0.286	$\chi^2=0.058$ prob=0.571	$\chi^2=3.733$ prob=0.143																											
<i>Exact p-value is the sum of probabilities of all tables with χ^2 values as great as or greater than that of the Observed Table:</i>																													

$$p\text{-value} = 0.286 + 0.143 = 0.429$$



Copyright © SAS Institute Inc. All rights reserved.

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 two-sided exact p -value.

The exact p -value would be 0.286 (Observed Table)+0.143 (Possible Table 3)=0.429. 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.



Fisher's 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 **STAT1.exact** data set, which has the data from the previous example.

1. Open the **Table Analysis** task under Statistics.
2. Select the **Exact** data set.
3. Assign variable **A** as the row variable and **B** as the column variable.
4. On the OPTIONS tab, expand the PLOTS property and select the option to suppress plots.
5. Select the options to include Observed frequencies, Expected frequencies, Row percentages, and Cell contributions to the chi-square statistics.
6. Run the code.

Note: Alternatively, you can write the code directly.

```
/*st10cd06.sas*/
ods graphics off;

proc freq data=STAT1.exact;
  tables A*B / chisq expected cellchi2 nocol nopercnt;
  title "Exact P-Values";
run;
```

The frequency table is shown below.

Table of A by B				
Frequency Expected Cell Chi-Square Row Pct	B			
	1	2	Total	
1	0	3	3	
	0.8571	2.1429		
	0.8571	0.3429		
	0.00	100.00		
2	2	2	4	
	1.1429	2.8571		
	0.6429	0.2571		
	50.00	50.00		
Total		2	5	7

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

Statistic	DF	Value	Prob
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. Chi-Square may not be a valid test.			

The warning tells you that you should not trust the reported *p*-value in this table.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	0
Left-sided Pr <= F	0.2857
Right-sided Pr >= F	1.0000
Table Probability (P)	0.2857
Two-sided Pr <= P	0.4286

The Two-sided Pr <= P value is the one that you will report. Notice the difference between the exact *p*-value (0.4286) and the asymptotic *p*-value (0.1473) in the Pearson chi-square test table. The exact *p*-values are larger. Exact tests tend to be more conservative than asymptotic tests.

Note: For tables larger than 2*2, an EXACT statement must be submitted to obtain exact *p*-values. For large tables, this can take a long time and use a great deal of computational resources.

End of Demonstration

C.6 Empirical Logit Plots

Objectives

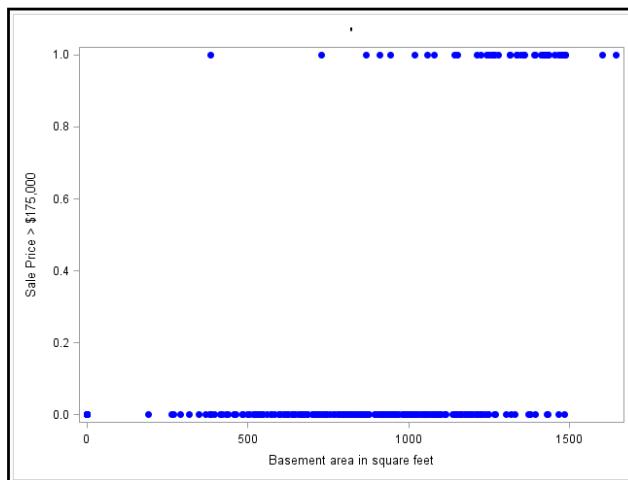
- Explain the concept of empirical logit plots.
- Plot empirical logits for continuous and ordinal predictor variables.

36



Copyright © SAS Institute Inc. All rights reserved.

Scatter Plot of Binary Response Data



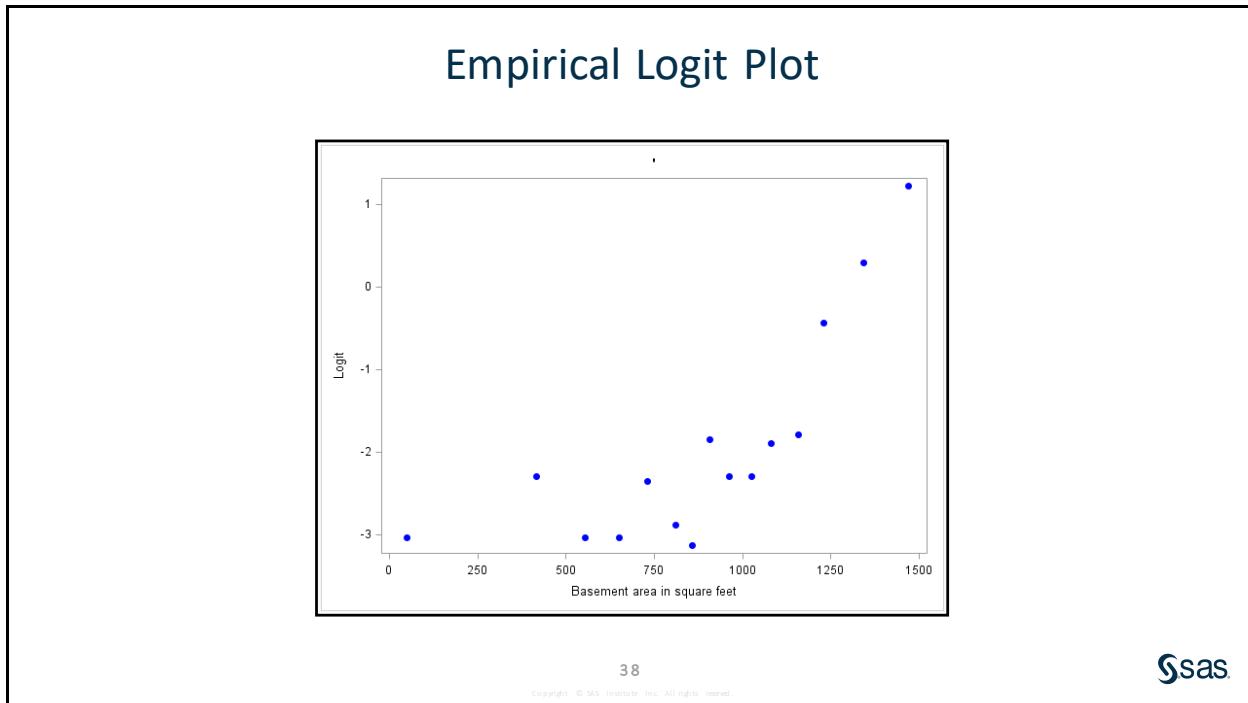
37



Copyright © SAS Institute Inc. All rights reserved.

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 whether there might be some nonlinear trend.

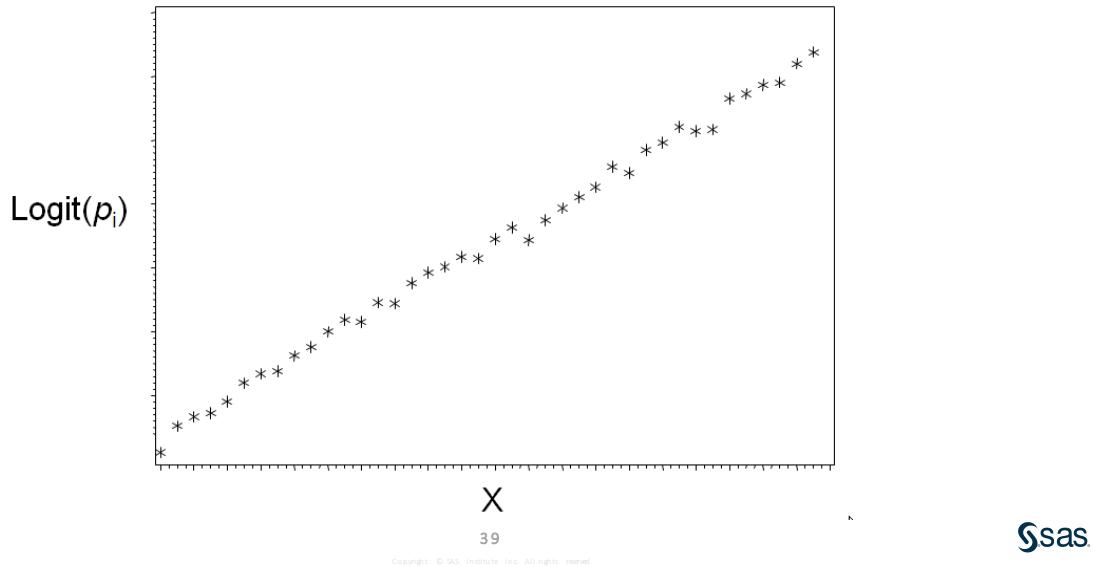
that should be modeled in the analysis. For binary response variables, a scatter plot contributes little to these ends.



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. You 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 $\ln(\text{proportion}/(1-\text{proportion}))$.

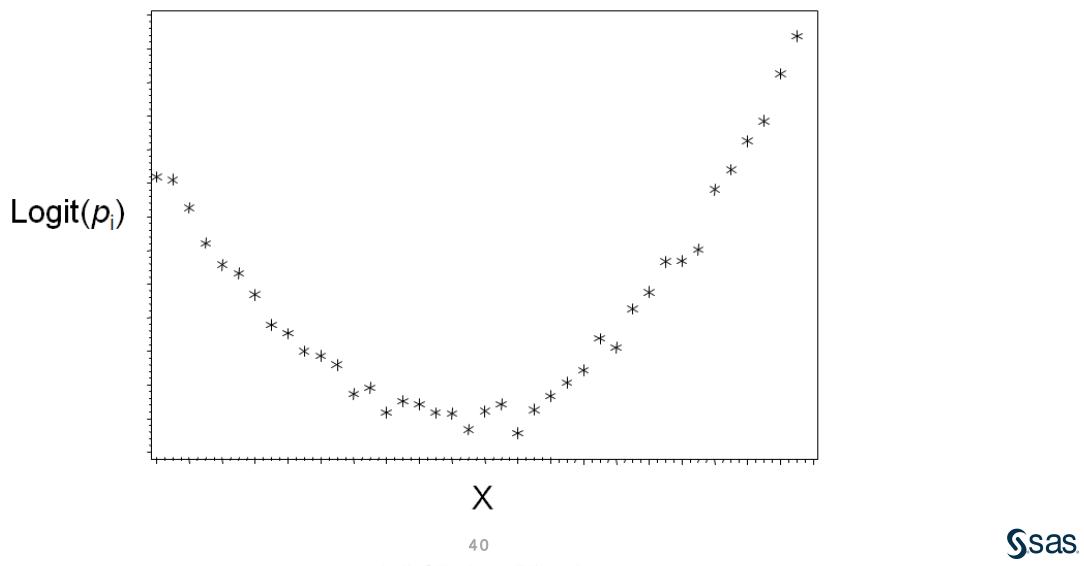
Note: If the predictor variable is a nominal variable, then there is no need to create a logit plot.

Logit Plot Implying Linear Relationship



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.

Logit Plot Implying Quadratic Relationship



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.

Empirical Logit Estimation

$$\ln\left(\frac{E_i + 0.5}{C_i - E_i + 0.5}\right)$$

where

E_i = number of events in bin

C_i = number of cases in bin

A common approach when 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. C represents the total number in the bin and E represents the total number of positive events in the bin.

$$\left(\frac{P_i}{(1-P_i)}\right) = \left(\frac{\frac{E_i}{C_i}}{\left(\frac{C_i-E_i}{C_i}\right)}\right) = \left(\frac{E_i}{(C_i-E_i)}\right)$$

The logit is undefined 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).



Fisher's Exact p-Values for the Pearson Chi-Square Test

Example: Plot the estimated logits of the outcome variable **Bonus** versus the predictor variable **Fireplaces**. To construct the estimated logits, the number of bonus eligible houses and the total number of houses by each level of **Fireplaces** must be computed.

1. Open the **Summary Statistics** task under Statistics.
2. On the DATA tab, select the **Ameshousing3** data set.
3. Assign **Bonus** as the analysis variable and **Fireplaces** as the classification variable.
4. On the OPTIONS tab, clear all the options under basic statistics except for number of observations.
5. Expand the additional statistics and select the option to calculate the Sum for the variable **Bonus**.
6. On the OUTPUT tab, name the new data set **STAT1.bins**.
7. Run the code.
8. To create the logit variable, open the **Transform Data** task under Data.
9. Select **Bins** data set that was just created.
10. Under the **Transform 1** tab, set **Bonus_N** as variable 1.
11. Use the drop-down menu and select the option to **Specify custom transformation**.
12. In the Custom transform field, enter the transformation **log((Bonus_Sum+0.5)/(Bonus_N-Bonus_Sum+0.5))**.
13. Expand the OUTPUT DATA SET property and name the new data set.
14. Run the code.
15. To plot the estimated logits, open the **Scatter Plot** task under Graph.
16. Select the data set with the transformed data.
17. Set **Fireplaces** as the X variable and the transformed variable, **tr1_Bonus_N**, as the Y variable.
18. On the OPTIONS tab, expand the X AXIS and Y AXIS properties and clear the options to show grid lines.
19. Run the code.

Note: In SAS Studio task, the names of the new variables are automatically generated. The following code produces all the necessary output with the addition of user-specified variable names.

Note: The Transform Data task in SAS Studio does not permit overwriting the existing data set, which is not the case in DATA step.

```
/*st10cd07.sas*/ /*Part A*/
proc means data=STAT1.ameshousing3 noprint nway;
  class Fireplaces;
  var Bonus;
  output out=bins sum(Bonus)=NEvents n(Bonus)=NCases;
run;

data bins;
  set bins;
  Logit=log( (NEvents+0.5) / (NCases-NEvents+0.5) );
run;

proc sgplot data=bins;
  scatter Y=Logit X=Fireplaces /
    markerattrs=(symbol=asterisk color=blue size=15);
  xaxis integer;
  title "Estimated Logit Plot of Fireplaces";
run;
```

Selected PROC MEANS statement option:

NWAY specifies that the output data set only contain statistics broken down by all combinations of levels of the CLASS variables

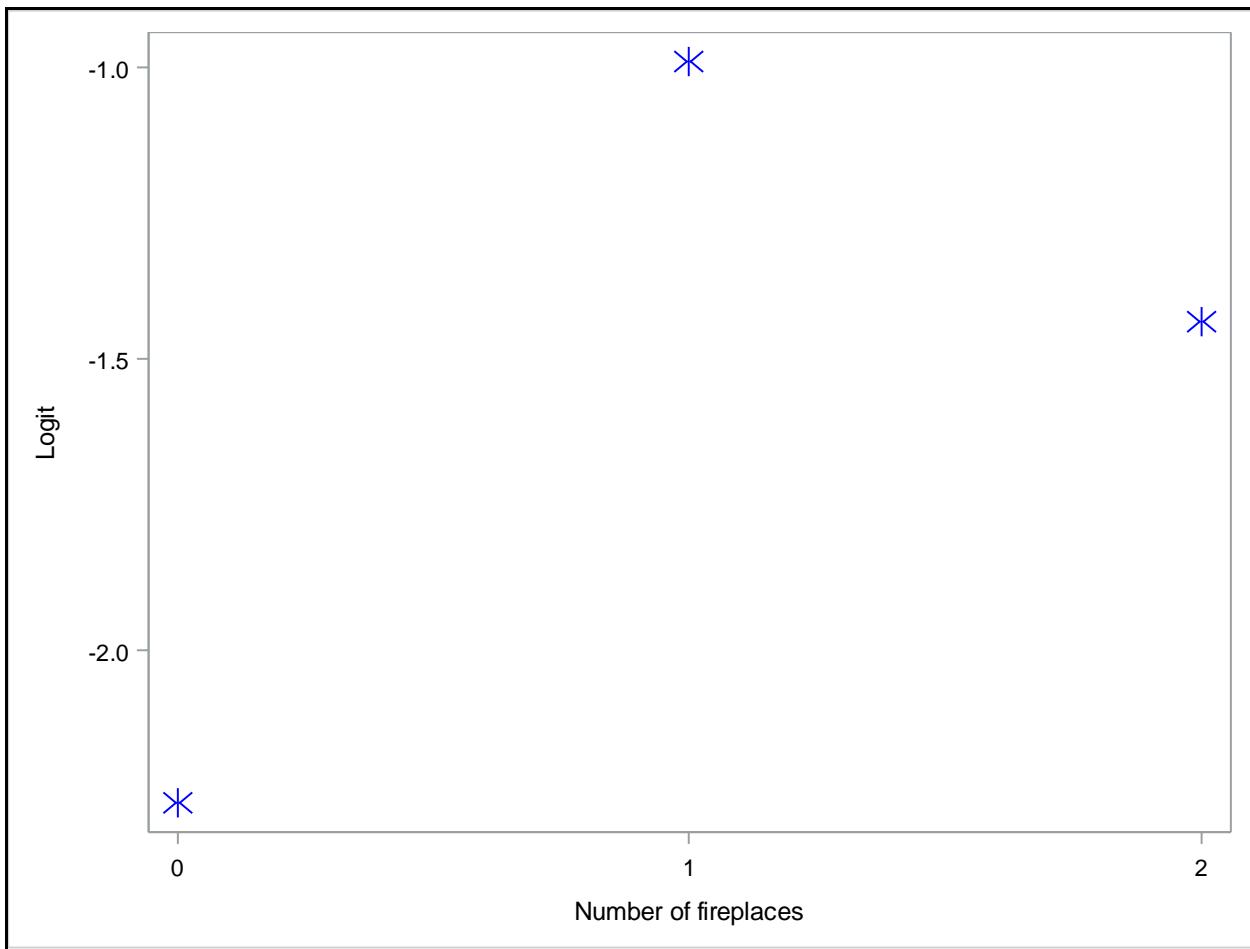
Selected options for SCATTER statement:

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 **Bonus** is coded 0/1, SUM(**Bonus**) returns the value for the count of ones within each level of **Fireplaces**. N(**Bonus**) returns the number of nonmissing values of **Bonus**, 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, C is represented by **NCases** and E is represented by **NEvent**.

PROC SGLOT shows the data.



The trend shown in the empirical logit plot for this ordinal variable does not appear linear.

Note: In some cases, when a linear pattern is detected in a logit plot for an ordinal variable, the variable can be removed from the CLASS statement, implying that it would be considered the same as a 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.

Example: Plot the estimated logits of the outcome variable **Bonus** versus the predictor variable **Basement_Area**. Because **Basement_Area** is a continuous variable, bin the observations into 20 groups to ensure that an adequate number of observations are used to compute the estimated logit.

1. Open the **Rank Data** task under Data.
2. Select the **AmesHousing3** data set.

3. Assign **Basement_Area** as the column to rank.
4. Open the editor and specify to bin the observations into groups using groups= option.
5. Run the code.
6. Open the **Summary Statistics** task under Statistics.
7. On the DATA tab, select the **Rank** data set created from the previous task.
8. Assign **Bonus** as the analysis variable and **rank_Basement_Area** as the classification variable.
9. On the OPTIONS tab, clear all the options under basic statistics except for number of observations.
10. Expand the additional statistics and select the option to calculate the Sum for the variable Bonus.
11. On the OUTPUT tab, name the new data set **STAT1.bins**.
12. Run the code.
13. To create the logit variable, open the **Transform Data** task under Data.
14. Select **Bins** data set that was just created.
15. Under the **Transform 1** tab, set **rank_Basement_Area** as variable 1.
16. Use the drop-down menu and select the option to **Specify custom transformation**.
17. In the Custom transform field, enter the transformation **log((Bonus_Sum+0.5)/(Bonus_N-Bonus_Sum+0.5))**.
18. Expand the OUTPUT DATA SET property and name the new data set.
19. Run the code.
20. To plot the estimated logits, open the **Scatter Plot** task under Graph.
21. Select the data set with the transformed data.
22. Set **rank_Basement_Area** as the X variable and the transformed variable, **tr1_rank_Basement_Area**, as the Y variable.
23. Expand the FIT PLOTS property and select the options to include a fitted regression line and a fitted Loess curve in the plot.
24. On the OPTIONS tab, expand the X AXIS and Y AXIS properties and clear the options to show grid lines.
25. Run the code.

```

/*st10cd07.sas*/ /*Part B*/
proc rank data=STAT1.ameshousing3 groups=20 out=Ranks;
  var Basement_Area;
  ranks Rank;
run;

proc means data=Ranks noprint nway;
  class Rank;
  var Bonus Basement_Area;
  output out=Bins sum(Bonus)=NEvents n(Bonus)=NCases
    mean(Bonus)=Basement_Area;
run;

data bins;
  set bins;
  Logit=log( (NEvents+0.5) / (NCases-NEvents+0.5) );
run;

proc sgplot data=bins;
  reg Y=Logit X=Basement_Area /
    markerattr=(symbol=asterisk color=blue size=15);
  loess Y=Logit X=Basement_Area / nomarkers;
  title "Estimated Logit Plot of Basement_Area";
run;

```

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 SGPOINT statements:

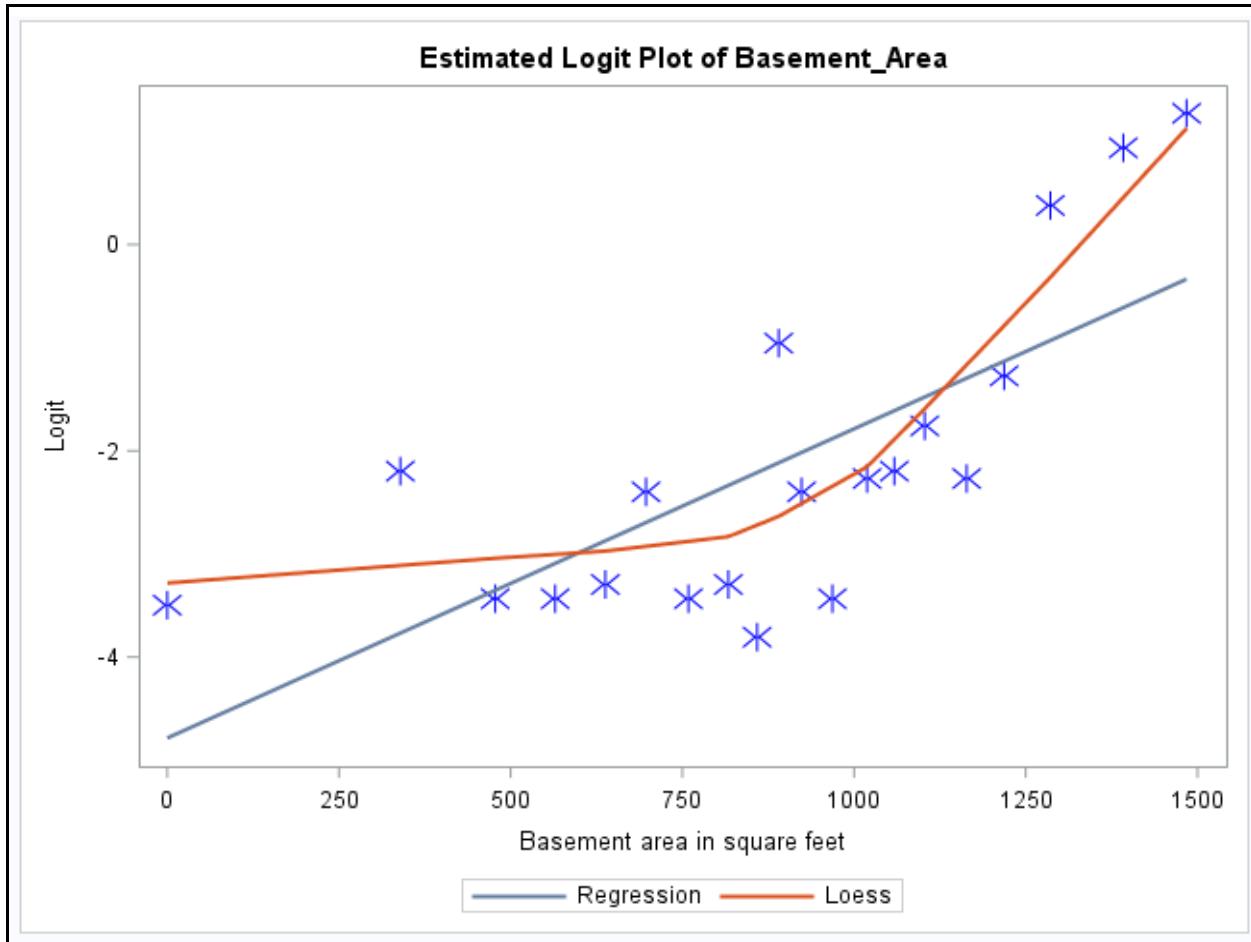
REG creates a fitted regression line or curve.

LOESS creates a fitted LOESS curve.

Selected options for LOESS statement:

NOMARKERS suppresses the scatter markers from the plot.

In the case of **Basement_Area**, 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 300 observations. It is recommended that you have approximately 20 to 30 observations per bin. If you divide the sample size by the desired per bin count, you can estimate the value to use for the GROUPS= option of PROC RANK. In this case,, with 20 observations per bin, the number of bins should be 300/20 or 15 bins.



The empirical logit plot shows a deviation from linearity. One possibility is to add a quadratic (squared) term for **Basement_Area**.

The empirical logit plot is univariate and therefore can be misleading in the presence of interactions and partial associations in a logistic regression model. (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.

End of Demonstration

Solutions to Student Activities (Polls/Quizzes)

C.01 Multiple Choice Poll – Correct Answer

What justifies the choice of a one-sided test versus a two-sided test?

- a. The need for more statistical power
- b. Theoretical and subject-matter considerations
- c. A two-sided test that is nonsignificant
- d. The need for an unbiased test statistic

Appendix D Percentile Definitions

D.1 Calculating Percentiles	D-3
-----------------------------------	-----

D.1 Calculating Percentiles

Using the UNIVARIATE Procedure

Example: Calculate the 25th percentile for the following data using the five definitions available in PROC UNIVARIATE:

1 3 7 11 14

For all of these calculations (except definition 4), you use the value $np = (5)(0.25) = 1.25$. This can be viewed as an observation number. However, there is obviously no observation 1.25.

Definition 1 returns a weighted average. The value returned is 25% of the distance between observations 1 and 2. (The value of 25% is the fractional part of 1.25 expressed as a percentage.)

$$\text{percentile} = 1 + (0.25)(3 - 1) = 1.5$$

Definition 2 rounds to the nearest observation number. Thus, the value 1.25 is rounded to 1 and the first observation, 1, is taken as the 25th percentile. If np were 1.5, then the second observation would be selected as the 25th percentile.

Definition 3 always rounds up. Thus, 1.25 rounds up to 2 and the second data value, 3, is taken as the 25th percentile.

Definition 4 is a weighted average similar to definition 1, except instead of using np , definition 4 uses $(n + 1)p = 1.5$.

$$\text{percentile} = 1 + (0.5)(3 - 1) = 2$$

Definition 5 rounds up to the next observation number unless np is an integer. In that case, an average of the observations represented by np and $(np + 1)$ is calculated. In this example, definition 5 rounds up, and the 25th percentile is 3.

Appendix E Writing and Submitting SAS® Programs in SAS® Enterprise Guide®

E.1 Writing and Submitting SAS Programs in SAS Enterprise Guide.....	E-3
Demonstration: Adding a SAS Program to a Project.....	E-11

E.1 Writing and Submitting SAS Programs in SAS Enterprise Guide

Objectives

- Create and submit new SAS programs.
- Insert existing programs into a project.
- List programming statements to avoid.
- Generate a combined project program and log.

2



Copyright © SAS Institute Inc. All rights reserved.

SAS Enterprise Guide Program Editor

SAS Enterprise Guide includes a programming editor similar to the Enhanced Program Editor.

Additional functionality in the SAS Enterprise Guide 4.3 Program Editor includes the following:

- autocomplete
- dynamic syntax tooltips
- formatting programs to provide consistent spacing
- analyzing program flow

```

Program Log Output Data
Save Run Stop Selected Server: Local (Connected) Analyze Program ...
1 libname orion "s:\workshop";
2
3 proc means data=orion.products su
4
SUM
SUMSIZE=
SUMWGT
T
THREADS
UCLM
USS
VAR
VARDEF=

```

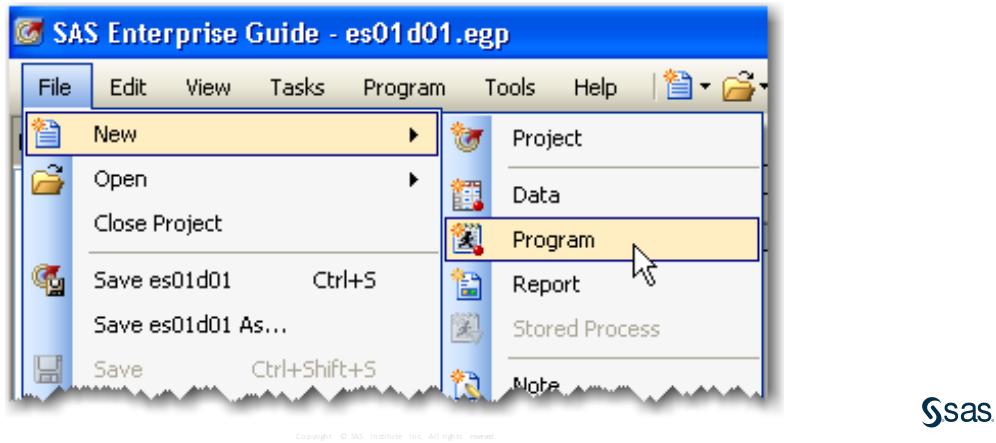
3



Copyright © SAS Institute Inc. All rights reserved.

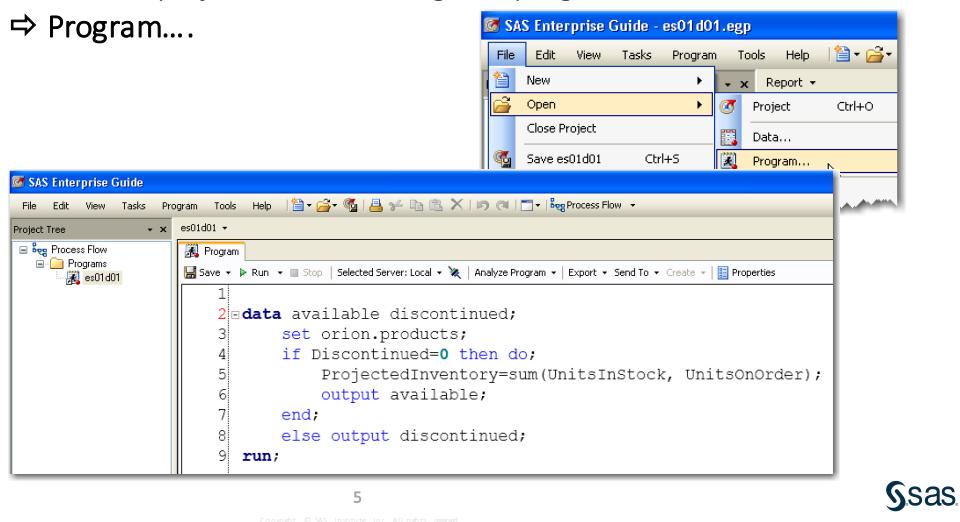
Writing a New SAS Program

To build a SAS program, select **File** ⇒ **New** ⇒ **Program** to create a new code node in the project. The new program is saved along with the project.



Adding Existing Code

To add a shortcut in the project to an existing SAS program, select **File** ⇒ **Open** ⇒ **Program....**



When you insert code, a shortcut to the file is added in the project, which means that changes made to the code in the project are also saved to the .sas file that you inserted. Also, if you make changes to the .sas file outside of SAS Enterprise Guide, the changes are reflected when you open or run the project again.

Running SAS Code

A SAS program can be submitted using one of these techniques:

- Select **Run** or **Run Selection** from the toolbar.
- Select **Program** ⇒ **Run** or **Run Selection** from the menu bar.
- Right-click on the program and select **Run** or **Run Selection**.
- Press F8 or F3.



If SAS is available on multiple servers, you can select **Select Server** and designate the server on which the program should run.

If the data for a task is located on a server that is different from the server where the SAS code is run, then SAS Enterprise Guide copies the data to the server where the code actually runs. Because moving large amounts of data over a network can be time- and resource-intensive, it is recommended that the server that you choose to process the code be the same server on which the data resides.

Accessing Program, Log, and Results

The code, log, output data, and results are accessible via separate tabs.

The screenshot shows the SAS Enterprise Guide interface with the 'es01a01' project open. The top navigation bar includes tabs for Program, Log, Output Data, and Results. The 'Log' tab is circled in red. The main editor area displays SAS code, including various options and notes about supported procedures. The SAS logo is visible in the bottom right corner.

```

1 /*;
2   *;;
3   1      ;*/*;/quit;run;
4   2      OPTIONS PAGENO=MIN;
5   3      %LET _CLIENTTASKLABEL='es01a01';
6   4      %LET _CLIENTPROJECTPATH='';
7   5      %LET _CLIENTPROJECTNAME='';
8   6      %LET _SASPROGRAMFILE='S:\workshop\es01a01.sas';
9   7
10  8      ODS _ALL_ CLOSE;
11  9      OPTIONS DEV=ACTIVE;
12  NOTE: Procedures may not support all options or statements for all devices. For details, see
13  the documentation for each procedure.
14  10     GOPTIONS XPIXELS=0 YPIXELS=0;
15  11     FILENAME EGSR TEMP;
16  12     ODS tagsets.sasreport12(ID=EGSR) FILE=EGSR STYLE=Analysis
17  12     ! STYLESHEET=(URL="file:///C:/Program%20Files/SAS/EnterpriseGuide/4.3/Styles/Analysis.c
18  12     ! ss") NOGTITLE NOFOOTNOTE GPATH=%sasworklocation ENCODING=UTF8 options(rolap="on");
19  NOTE: Writing TAGSETS.SASREPORT12(EGSR) Body file: EGSR

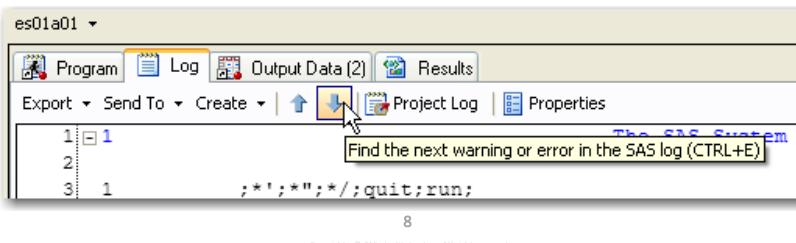
```

Identifying Warning and Errors in the Log

The code icons in the project indicate whether there are warnings or errors in the SAS log.



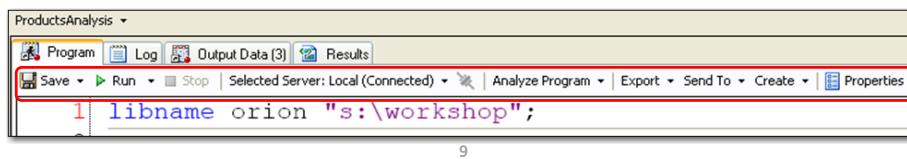
Arrows on the Log tab enable quick navigation to the next warning or error.



Using the Program Toolbar

A toolbar above the program offers easy access to common actions, such as the following:

- saving the program
- running or stopping a program
- selecting the execution server
- analyzing the program for flow or grid computing
- exporting and emailing
- creating a stored process
- modifying program properties



The Analyze Program button enables you to select one of these two options:

Analyze Program Flow	SAS Enterprise Guide can create a process flow from a program. Using this process flow, you can quickly identify the different parts of the program and see how the parts are related.
Analyze Program for Grid Computing	When analyzing a program for grid computing, SAS Enterprise Guide identifies the parts of the program that are not dependent on one another. These parts can run simultaneously on multiple computers, which means that SAS Enterprise Guide returns the results more quickly. When SAS analyzes a program, lines of SAS/CONNECT code are added to your original program. Therefore, you must have a license for SAS Grid Manager or SAS/CONNECT to analyze a program for grid computing.

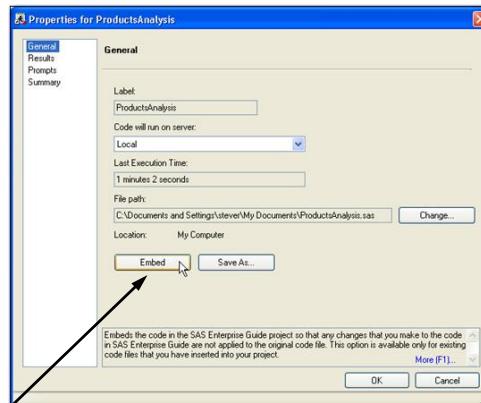
Note: Both options run the code behind the scenes to complete the analysis. If a data set is open in the SAS Enterprise Guide session, the analysis might fail. To view and close any open data sets, select **Tools** \Rightarrow **View Open Data Sets**.

Embedding Programs in a Project

New SAS programs are embedded in the project so that it is saved as part of the .epg file.

When an existing SAS program is added to a project, a shortcut to the program file is created. You can also embed the program so that it is stored as part of the project file.

Select **Properties** from the Program toolbar and select **Embed**.



Using Autocomplete

In SAS Enterprise Guide 4.3, the Program Editor includes an autocomplete feature. The editor can suggest

- SAS statements
- procedures
- macro programs
- macro variables
- functions
- formats
- librefs
- SAS data sets.



The autocomplete feature automatically suggests appropriate keywords. You can also manually open the Autocomplete window by using the following shortcut keys:

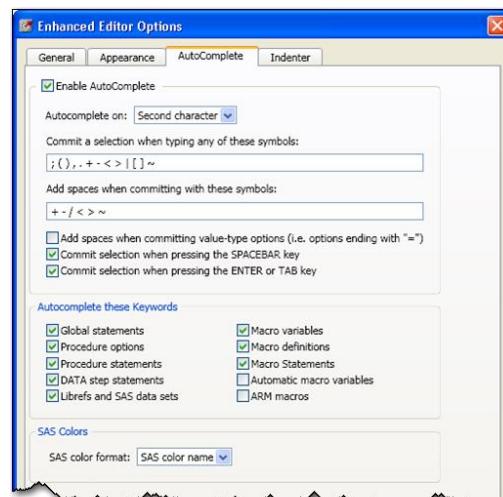
Action	Keyboard Shortcut
Open the Autocomplete window for the keyword on which the cursor is currently positioned. In a blank program, this shortcut displays a list of global statements.	Ctrl + spacebar
Open the Autocomplete window that contains a list of the SAS libraries that are available with the current server connection.	Ctrl + L
Open the Autocomplete window that contains a list of data sets that were created by using the DATA statement.	Ctrl + D
Open the Autocomplete window that contains a list of SAS functions.	Ctrl + Shift + F1
Open the Autocomplete window that contains a list of macro functions.	Ctrl + Shift + F2
Open the Autocomplete window that contains a list of SAS formats.	Ctrl + Shift + F
Open the Autocomplete window that contains a list of SAS informats.	Ctrl + Shift + I
Open the Autocomplete window that contains a list of statistics keywords.	Ctrl + Shift + K

Open the Autocomplete window that contains a list of SAS colors.	Ctrl + Shift + C
Open the Autocomplete window that contains a list of style attributes.	Ctrl + Shift + F4
Open the Autocomplete window that contains a list of style elements.	Ctrl + Shift + F3

Customizing the Program Editor

The Program Editor can be customized by selecting **Program** ⇒ **Editor Options**.

Autocomplete can be customized or disabled on the Autocomplete tab.

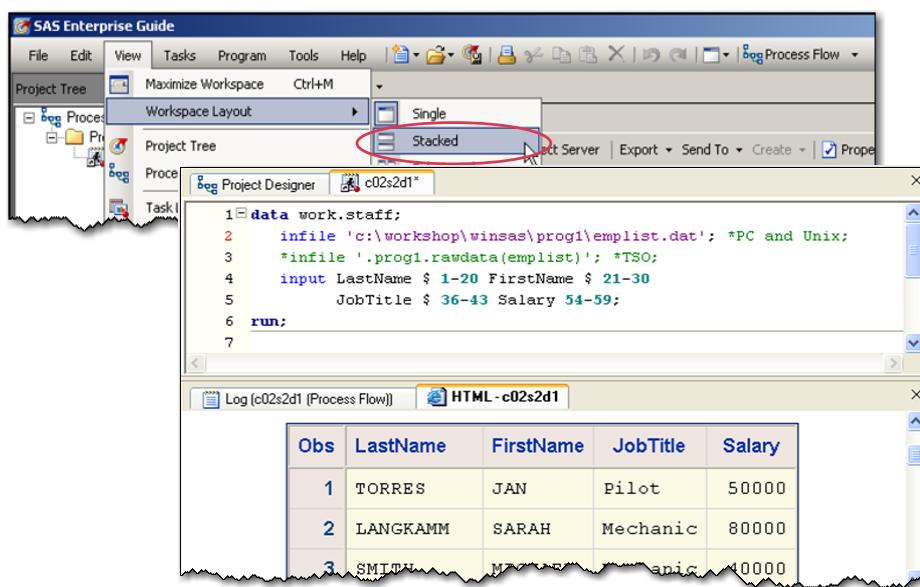


sas

12

Copyright © SAS Institute Inc. All rights reserved.

Rearranging Windows

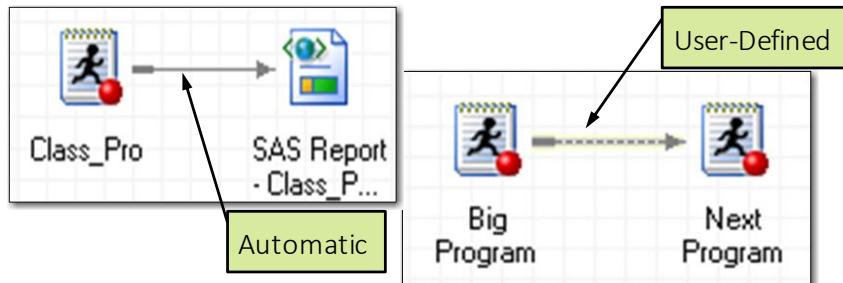


sas

Linking Items in the Process Flow

Links define the directional relationship between objects in SAS Enterprise Guide to create a process flow. Links can be either automatic or user-defined.

Links can enable you to force a particular flow between programs and point-and-click tasks in the project.





Adding a SAS Program to a Project

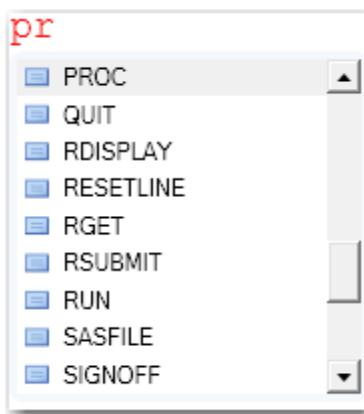
1. Create a new project.
2. To open an existing SAS program, select **File** \Rightarrow **Open** \Rightarrow **Program**.... Navigate to the location of the course data and select **st100d05.sas** \Rightarrow **Open**. A shortcut to the program is added to the project.
3. There is no indenting in this program to make it easier to read. Also, statements flow over onto new lines. Select **Edit** \Rightarrow **Format Code** to improve the spacing and organization of the code, or you can right-click on the program and select **Format Code**.

Note: To modify the rules for formatting code, select **Program** \Rightarrow **Editor Options** \Rightarrow **Indenter**.

4. To execute the SAS program, select **Run** on the toolbar. A report is generated and lists the products in the **saleswomen** data set. Twenty-one are added to the project. Because **TestScores** was the first data set created, it is automatically placed on a new tab. All other data sets are accessible from the process flow.

Obs	Purchase	Gender	Income	Age
1	0	Female	Low	40
2	0	Female	Low	46
3	1	Female	Low	41

5. To include a frequency report to analyze the distribution of Purchase in the **Sales** data set, use the FREQ procedure in the SAS program. At the end of the program, type **pr**. A list of keywords is provided. Press the spacebar to select the word **PROC** for the program.



6. A list of procedure names is automatically provided. Type **fr** and press the spacebar again to select **freq** for the program. Next, a list of valid options for the PROC FREQ statement is provided. Type **d** and press the spacebar to select **data=**.
7. A list of data sets in the project and defined libraries is provided. Select **STAT1**, press the spacebar, and then select **SALES** for the data set.

8. The list of valid options for the PROC FREQ statement appears again. Type **o**, select **order=**, and press the spacebar. Type **fr** to select **FREQ** and then enter a semicolon to complete the statement that appears as follows:

```
proc freq data=STAT1.sales order=freq;
```

9. Continue to use the autocomplete feature to write the remainder of the step:

```
proc freq data=STAT1.sales order=freq;
  tables Gender*Purchase / chisq relrisk;
run;
```

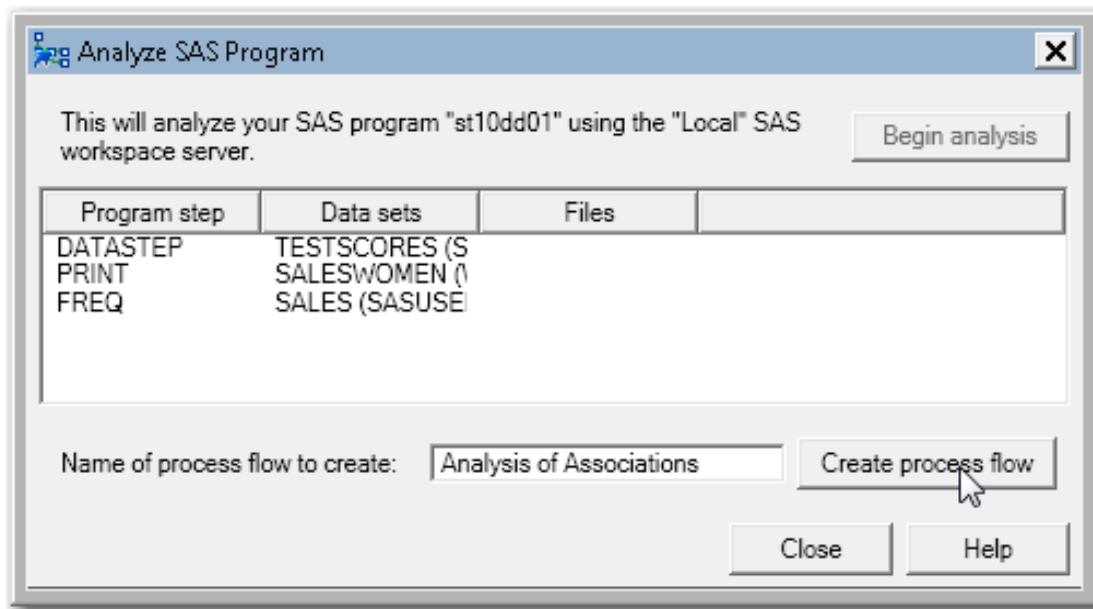
10. Highlight the PROC FREQ step in the program and select **Run** ⇒ **Run Selection**. Select **Yes** when you are prompted to replace the results.

Partial Results

Table of Gender by Purchase			
Gender	Purchase		
Frequency			
Female	0	1	Total
	139 32.25 57.92 51.67	101 23.43 42.08 62.35	240 55.68
Male	130 30.16 68.06 48.33	61 14.15 31.94 37.65	191 44.32
Total	269 62.41	162 37.59	431 100.00

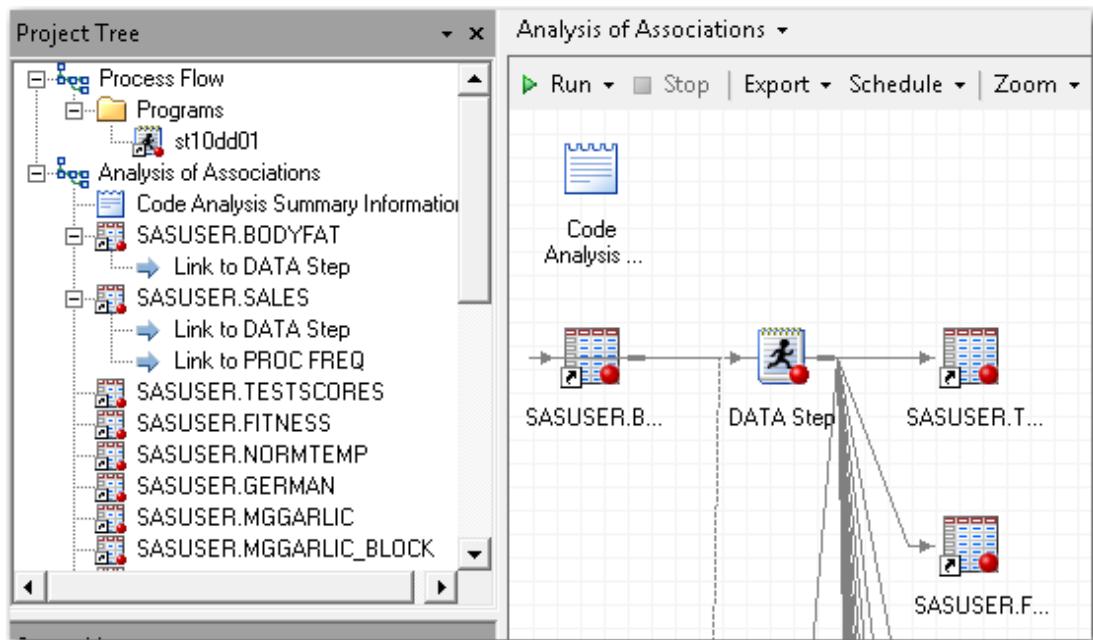
11. The program now includes three steps and creates multiple data sets and reports. To better visualize the flow of the program, return to the Program tab and select **Analyze Program** ⇒ **Analyze Program Flow**.
12. In the Analyze SAS Program window, select **Begin analysis**. Then type **Analysis of Associations** in the **Name of process flow to create** field and select **Create process flow** ⇒ **Close**.

Note: If a data set is open in the SAS Enterprise Guide session, the analysis might fail.
To view and close any open data sets, select **Tools** ⇒ **View Open Data Sets....**



A new process flow is added to the project, and illustrates the flow of the steps in the program.

Note: To delete a process flow, right-click on the process flow in the Project Tree and select **Delete**.



13. The Program Editor also includes syntax tooltips. Double-click on the **st10dd01** program in the Project Tree or Process Flow window. Hold the mouse pointer over any keyword in the program. A tooltip displays syntax details for that particular step or statement.

Note: The F1 key also displays syntax help.

Note: You can view syntax tooltips by holding the mouse pointer over items in the autocomplete windows.

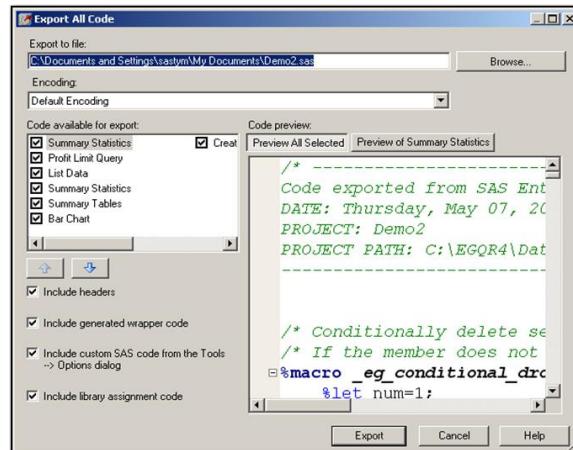
14. Save the modified program by returning to the Program tab and selecting **Save** ⇒ **Save As....**
Save the program as **st100d05s** and select **Save**.

End of Demonstration

Exporting Code

All SAS code within a project can be exported to a file that can be edited and executed in other SAS environments.

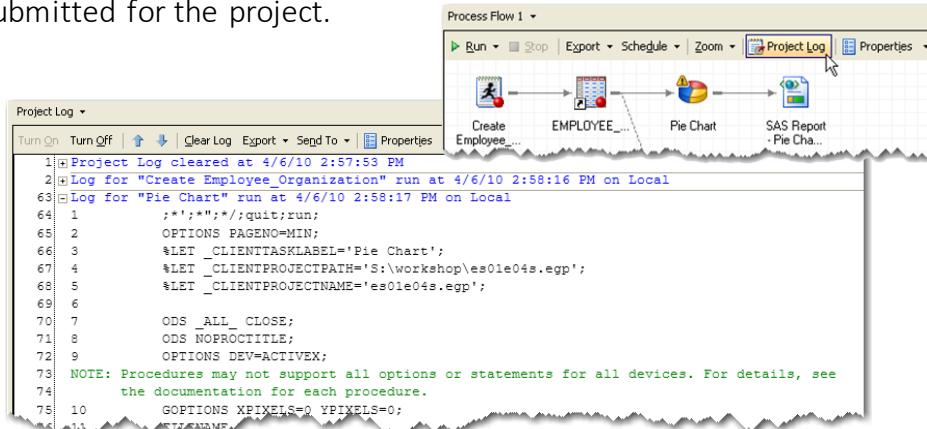
Select
File \Rightarrow Export \Rightarrow
Export All Code
in Project.



16

Project Log

The project log can be used to maintain and export an aggregated log of all code submitted for the project.



17

Programming Statements to Avoid

Programs that run in the SAS windowing environment can also run successfully in SAS Enterprise Guide. Be aware of the following exceptions:

- Code that calls X commands or SYSTASK might not work unless this permission is granted by the administrator.
- Code that would normally cause a window or prompt to appear in the SAS windowing environment (DEBUG, PROC FSLIST, AF applications) does not work in SAS Enterprise Guide.
- Code that terminates the SAS process with ABORT or ENDSAS calls terminates the connection between SAS Enterprise Guide and the SAS server.

For more information about enabling X and SYSTASK commands, go to blogs.sas.com/sasdummy/index.php?/archives/136-Using-the-X-and-SYSTASK-commands-from-SAS-Enterprise-Guide.html.