

# **Statistics 2: ANOVA and Regression**

Course Notes

*Statistics 2: ANOVA and Regression Course Notes* was developed by Lee Bennett, Chris Daman, Jill Tao, and Susan Walsh. Additional contributions were made by John Amrhein, David Dickey, Marc Huber, Kathleen Kiernan, Jay Laramore, Paul Marovich, Danny Modlin, Mike Patetta, Lorne Rothman, Eddie Routten, Roger Staum, and Catherine Truxillo. Editing and production support was provided by the Curriculum Development and Support Department.

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

### **Statistics 2: ANOVA and Regression Course Notes**

Copyright © 2015 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 E70445, course code LWST241/ST241, prepared date 16Sep2015.

LWST241\_001

ISBN 978-1-62959-935-9

## Table of Contents

Course Description .....	ix
Prerequisites .....	x
<b>Chapter 1     Multiple Linear Regression .....</b>	<b>1-1</b>
1.1    Review of General Linear Models.....	1-3
Demonstration: Multiple Linear Regression .....	1-13
1.2    Simple Polynomial Regression .....	1-22
Demonstration: Simple Polynomial Regression .....	1-26
Exercises .....	1-39
1.3    Polynomial Regression and Multicollinearity .....	1-41
Demonstration: Multicollinearity Diagnostics.....	1-47
Demonstration: Centering Variables .....	1-53
Exercises .....	1-57
1.4    Modeling Nonlinear Relationships .....	1-58
Demonstration: Initial Data Exploration.....	1-60
Demonstration: Select Candidate Models.....	1-70
Demonstration: Modeling with Splines .....	1-81
Exercises .....	1-84
1.5    Chapter Summary .....	1-85
1.6    Solutions .....	1-87
Solutions to Exercises .....	1-87
Solutions to Student Activities (Polls/Quizzes) .....	1-108
<b>Chapter 2    Regression Diagnostics and Remedial Measures .....</b>	<b>2-1</b>
2.1    Regression Model Diagnostics .....	2-3

Demonstration: Model Diagnostics – Normality, Constant Variance, Model Fit, Collinearity, and Influential Observations .....	2-15
Exercises .....	2-27
2.2 Remedial Measures.....	2-28
Demonstration: Fitting a Lognormal Regression Model .....	2-37
Exercises .....	2-43
2.3 Chapter Summary .....	2-44
2.4 Solutions .....	2-46
Solutions to Exercises .....	2-46
Solutions to Student Activities (Polls/Quizzes).....	2-64
<b>Chapter 3 Analysis of Variance .....</b>	<b>3-1</b>
3.1 ANOVA Review.....	3-3
Demonstration: Two-Way Analysis of Variance .....	3-12
Exercises .....	3-22
3.2 Postfitting Analyses .....	3-23
Demonstration: The LSMESTIMATE Statement .....	3-34
Exercises .....	3-37
3.3 Contrasts and Estimates (Self-Study) .....	3-38
Demonstration: Contrasts and Estimates .....	3-50
Exercises (Self-Study).....	3-53
3.4 Evaluations of Model Assumptions and Remedial Measures.....	3-54
Demonstration: Identifying Violations of ANOVA Assumptions .....	3-64
Demonstration: Accounting for Unequal Variances .....	3-74
Exercises .....	3-82
3.5 Chapter Summary .....	3-83
3.6 Solutions .....	3-85
Solutions to Exercises .....	3-85

Solutions to Student Activities (Polls/Quizzes) .....	3-97
<b>Chapter 4 Analysis of Covariance (ANCOVA).....</b>	<b>4-1</b>
4.1 Introduction to Analysis of Covariance (ANCOVA) .....	4-3
Demonstration: Conducting an Analysis of Covariance Using PROC GLM .....	4-12
Exercises .....	4-16
4.2 Least Squares Means for ANCOVA Models .....	4-17
Demonstration: Least Squares Means and Multiple Comparison Tests .....	4-19
Exercises .....	4-25
4.3 Diagnostics and Remedial Measures for ANCOVA Models .....	4-26
Demonstration: Diagnostics and Remedial Measures for ANCOVA Models.....	4-28
Exercises (Take-home) .....	4-34
4.4 Chapter Summary .....	4-35
4.5 Solutions .....	4-36
Solutions to Exercises .....	4-36
Solutions to Student Activities (Polls/Quizzes) .....	4-44
<b>Chapter 5 Introduction to Generalized Linear Models .....</b>	<b>5-1</b>
5.1 Introduction to Generalized Linear Models.....	5-3
5.2 Introduction to Poisson Regression and Negative Binomial Regression .....	5-8
Demonstration: Fitting a Poisson Regression Model for Count Data.....	5-16
Demonstration: Modeling Overdispersion By Using the Negative Binomial Distribution .....	5-28
Exercises .....	5-37
Demonstration: Fitting a Poisson Regression Model for Rate Data.....	5-44
5.3 Introduction to Gamma Regression .....	5-47
Demonstration: Fitting a Gamma Regression Model.....	5-51
Exercises .....	5-63

5.4	Chapter Summary .....	5-64
5.5	Solutions .....	5-66
	Solutions to Exercises .....	5-66
	Solutions to Student Activities (Polls/Quizzes) .....	5-79
<b>Chapter 6</b>	<b>Introduction to Linear Mixed Models.....</b>	<b>6-1</b>
6.1	Defining Linear Mixed Models .....	6-3
6.2	Using the GLIMMIX Procedure.....	6-13
	Demonstration: Fitting a Mixed Model Using PROC GLIMMIX.....	6-21
	Exercises .....	6-27
6.3	Solutions .....	6-29
	Solutions to Exercises .....	6-29
	Solutions to Student Activities (Polls/Quizzes) .....	6-34
<b>Appendix A</b>	<b>References .....</b>	<b>A-1</b>
A.1	References.....	A-3
<b>Appendix B</b>	<b>A Brief Review of Matrix Algebra .....</b>	<b>B-1</b>
B.1	Introducing Matrices.....	B-3
B.2	Basic Operations .....	B-3
B.3	Some Special Matrices .....	B-6
B.4	Determinant .....	B-10
B.5	Inverse Matrices.....	B-11
B.6	Linear Dependence and Rank .....	B-12
B.7	Generalized Inverse .....	B-14
B.8	Eigenvalues and Eigenvectors .....	B-15
B.9	Cholesky Root.....	B-17

**Appendix C Review of Simple Linear Regression and One-Way ANOVA..... C-1**

C.1	Univariate Analysis.....	C-3
	Demonstration: Exploring Data .....	C-7
C.2	Simple Linear Regression.....	C-13
	Demonstration: Simple Linear Regression .....	C-19
C.3	One-Way ANOVA Review.....	C-24
	Demonstration: One-Way ANOVA .....	C-31
	Demonstration: Multiple Comparison Tests .....	C-36
C.4	Chapter Summary .....	C-38

**Appendix D Additional Topics ..... D-1**

D.1	Nonlinear Regression.....	D-3
	Demonstration: Fitting a Nonlinear Regression Model .....	D-15
D.2	Local Regression .....	D-29
	Demonstration: Fitting a Local Regression Model .....	D-39
D.3	Modeling Data with Autocorrelation.....	D-51
	Demonstration: Detecting Autocorrelation .....	D-56
	Demonstration: Modeling Autocorrelation .....	D-62
D.4	Transforming the Dependent Variable as a Remedial Measure.....	D-67
	Demonstration: Transformations Based on Relationship between the Variance and Mean.....	D-71
	Demonstration: Box-Cox Transformation to Stabilize Nonconstant Variances.....	D-74
	Demonstration: Back-Transformation of the Model.....	D-83
	Demonstration: Transforming Variables as a Remedial Measure for Departures from Normality .....	D-88
	Exercises .....	D-96
D.5	Weighted Least Squares.....	D-98
	Demonstration: Weighted Least Squares Using PROC REG .....	D-100

D.6	Evaluating the Importance of Parameters .....	D-105
	Demonstration: Evaluating the Importance of Parameters .....	D-106
D.7	A Sample SAS Program for Comparing Model Fit .....	D-107
D.8	Incorrectly Treating Random Effects as Fixed .....	D-109
	Demonstration: Comparing Expected Mean Squares for Fixed and Random Effects .....	D-110
D.9	Solutions .....	D-114
	Solutions to Exercises .....	D-114
	Solutions to Polls and Quizzes .....	D-124

## Course Description

This course teaches you how to analyze continuous response data and discrete count data. Linear regression, Poisson regression, negative binomial regression, gamma regression, analysis of variance, linear regression with indicator variables, analysis of covariance, and mixed models ANOVA are presented in the course.

### 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](mailto: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](mailto: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.

## Prerequisites

Before attending this course, you should

- have some experience creating and managing SAS data sets, which you can gain from the SAS Programming 1: Essentials course
- be able to fit simple and multiple linear regression models using the REG procedure
- be able to analyze a one-way analysis of variance using the GLM procedure
- understand the statistical concepts of normal distribution, sampling distributions, hypothesis testing, and estimation
- have completed a graduate-level course in regression and analysis of variance methods or the Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression course.

Students should have completed the SAS Programming 1: Essentials and Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression courses, or have equivalent experience.

# Chapter 1    Multiple Linear Regression

<b>1.1   Review of General Linear Models .....</b>	<b>1-3</b>
Demonstration: Multiple Linear Regression .....	1-13
<b>1.2   Simple Polynomial Regression.....</b>	<b>1-22</b>
Demonstration: Simple Polynomial Regression .....	1-26
Exercises .....	1-39
<b>1.3   Polynomial Regression and Multicollinearity .....</b>	<b>1-41</b>
Demonstration: Multicollinearity Diagnostics .....	1-47
Demonstration: Centering Variables .....	1-53
Exercises .....	1-57
<b>1.4   Modeling Nonlinear Relationships .....</b>	<b>1-58</b>
Demonstration: Initial Data Exploration.....	1-60
Demonstration: Select Candidate Models .....	1-70
Demonstration: Modeling with Splines .....	1-81
Exercises .....	1-84
<b>1.5   Chapter Summary.....</b>	<b>1-85</b>
<b>1.6   Solutions .....</b>	<b>1-87</b>
Solutions to Exercises .....	1-87
Solutions to Student Activities (Polls/Quizzes) .....	1-108



# 1.1 Review of General Linear Models

## Objectives

- Review the properties of the general linear model.
- Do initial exploratory data analysis using the SGSCATTER procedure.
- Fit a multiple linear regression model using the GLMSELECT procedure.
- Examine linear regression assumptions using the diagnostic tools that are available in the REG and UNIVARIATE procedures.

6

## The General Linear Model

In the general linear model (GLM), a response Y is modeled as a linear function of predictor (X) variables:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\mathbf{Y}$  is a vector of continuous response values
- $\mathbf{X}$  is a matrix of predictor variables
- $\boldsymbol{\beta}$  is a vector of model parameters
- $\boldsymbol{\epsilon}$  is a vector random errors.

7

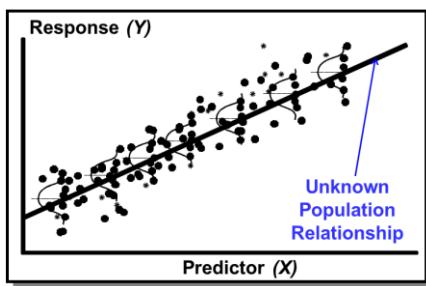
## Types of General Linear Models

Type of Predictor Variables in X		
Categorical	Continuous	Continuous and Categorical
Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)

8

General linear models are often classified according to the types of predictor variables that they include. *Analysis of variance (ANOVA)* models are those with only categorical predictor variables, but regression models have only continuous predictors. Models that include both categorical and continuous predictor variables are referred to as *analysis of covariance (ANCOVA)* models.

## Assumptions for GLMs



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

9

For illustrative purposes, the assumptions of GLMs are presented here for the simple case of a linear relationship between a continuous outcome variable ( $y$ ) and one predictor variable ( $x$ ).

The same types of assumptions are made for more complicated general linear models, such as those with more than one predictor variable, and those where the relationship between the outcome and predictor variables might not be linear (for example, polynomial regression models).

The assumptions on the error terms can be summarized in mathematical notation as  $\varepsilon \sim iid N(0, \sigma^2)$ . One suggested method for verifying the assumptions of GLMs is to plot the residuals versus the predicted values and versus the predictor variables. Another useful tool is a univariate analysis of the residuals. Both approaches are discussed in upcoming material.

## Violation of Model Assumptions

- *Normality* does not affect the parameter estimates, but it affects the test results.
- *Constant variance* does not affect the parameter estimates, but the standard errors are compromised.
- *Independent observations* do not affect the parameter estimates, but the standard errors are compromised.
- *Linear in the parameters* indicates a misspecified model, and therefore, the model results are not meaningful.

10

The assumption that the errors are normally distributed is not necessary for estimation of the model parameters and partitioning of the total sums of squares. The least squares estimates are still the best linear unbiased estimates (BLUE) if the other assumptions are met.

Normality is needed only for tests of significance and construction of confidence intervals of the parameters. The *t* test, *F* test, and chi-square test require the normality assumption of the residuals. Likewise, the confidence intervals also depend on the normality assumption (Rawlings, Pantula, and Dickey 1998).

## Linear Regression Models

- A simple linear regression model has one independent variable.
- A multiple linear regression model has more than one independent variable. For example,

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

where

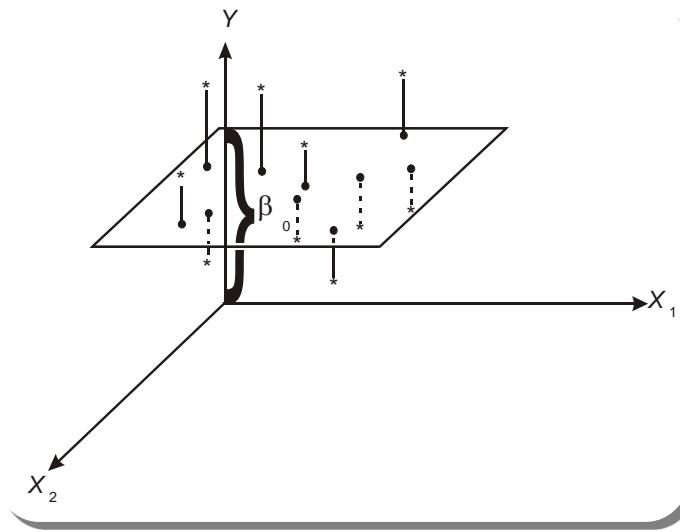
- $Y$  is the dependent variable
- $X_1$  and  $X_2$  are the independent or predictor variables
- $\varepsilon$  is the error term
- $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  are unknown parameters.

11

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 among the three variables (three dimensions) with a plane (two dimensions).

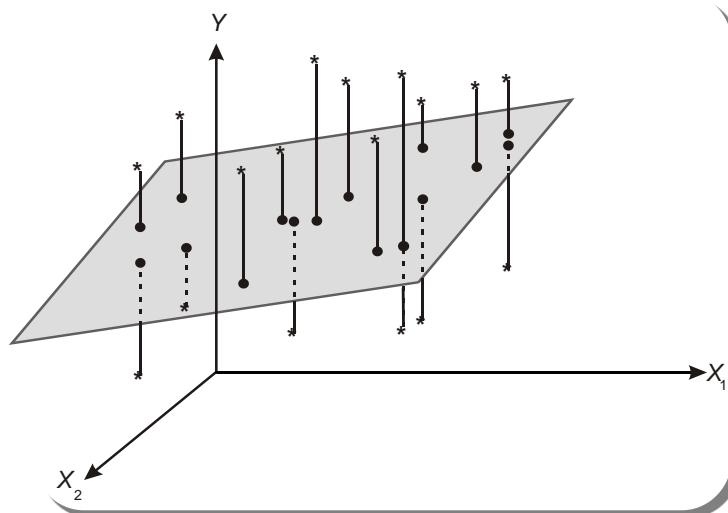
## Picturing the Model: No Relationship



12

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

## Picturing the Model: A Relationship



13

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

$$(Y = \beta_0, X_1 = 0, X_2 = 0)$$

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

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

## The Multiple Linear Regression Model

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

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

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

14

The multiple linear regression model refers to a GLM in which you have  $k$  continuous independent variables, or predictor variables. One of the objectives of the analysis might be to evaluate the significance of each of the  $k$  independent variables and how they relate to the dependent variable  $Y$ .

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

### Details

The multiple linear regression model shown on the previous page:

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

can be expressed in terms of four matrices:

$Y$ : the  $n \times 1$  column vector of values of the dependent variable  $Y$

$X$ : the  $n \times p$  matrix consisting of a column of ones, followed by  $k$  column vectors for the independent variables. Each column of  $X$  contains the values for a particular independent variable

$\beta$ : the  $p \times 1$  vector of parameters to be estimated, where  $p=k+1$

$\varepsilon$ : the  $n \times 1$  vector of random errors

The linear model can now be written in matrix notation as the following:

$$Y = X\beta + \varepsilon$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & & X_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In matrix notation, the normal equations are written as the following:

$$X'X\hat{\beta} = X'Y$$

The normal equations have a solution, given as the following:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

If  $X'X$  has an inverse, then the normal equations have a unique solution.

## The Multiple Linear Regression Model Hypothesis Test

**Null Hypothesis:**  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$

The regression model does not fit the data better than the mean model.

**Alternative Hypothesis:**  $H_1: \text{Not all } \beta_i\text{'s equal zero.}$

The regression model does fit the data better than the mean model.

15

If the estimated regression model does not fit the data better than the mean model, you fail to reject the null hypothesis. Thus, you do not have enough evidence to say that at least one slope is not zero. Therefore, you do not have enough evidence to say that the predictor variables explain a significant amount of variability in the response variable.

If the estimated regression model *does* fit the data better than the mean model, you reject the null hypothesis. Thus, you *do* have enough evidence to say that the slopes are *not* all equal to zero and that one or more predictor variables explain a significant amount of variability in the response variable.

### Details

The significance of the model can be tested using the  $F$  statistic shown in the ANOVA table below.

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Squares (MS=SS/df)	F value	p-value
Due to regression	$k$	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$\text{MSR}=\text{SSR}/\text{dfR}$	$\text{MSR}/\text{MSE}$	If $< \alpha$ (predefined, for example, 0.05), then significant model
Error	$n-k-1$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\text{MSE}=\text{SSE}/\text{dfE}$		
Total, corrected for mean $\bar{Y}$	$n-1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$			

## 1.03 Quiz

Suppose the regression model that you fit is the following:

$$\hat{y} = 3 + 5x$$

How do you interpret the slope for  $x$ , which is 5?

16

## The GLMSELECT Procedure

General form of the GLMSELECT procedure:

```
PROC GLMSELECT options OUTDESIGN(options)=  
      SAS-data-set;  
EFFECT effect-name=effect-type(<var-list></options>);  
MODEL dependent=model-effects / options;  
OUTPUT OUT=SAS-data-set  
      keyword=names;  
RUN;
```

18

The GLMSELECT procedure was introduced in SAS 9.1 and became a full production procedure in SAS 9.2. It combines features of PROC GLM and PROC REG for fitting general linear models, including the CLASS statement for categorical variables and automatic variable selection methods. Certain types of analyses, including selected regression diagnostics, are not yet available in PROC GLMSELECT. For these analyses, PROC REG is used.

Selected GLMSELECT statement option:

**OUTDESIGN** produces a SAS data set that contains the predictor ( $X$ ) variables for the model, including any effects constructed using an EFFECT statement. If the ADDINPUTVARS option is included, this data set also includes all variables in the original input data set.

Selected GLMSELECT procedure statements:

- MODEL** specifies the response and predictor variables. Model options include SELECTION= to perform automatic model selection. If SELECTION= is omitted, stepwise model selection is used. To fit a model without selection, specify SELECTION=NONE.
-  Additional model selection options are discussed later in the course.
- EFFECT** creates new effects for the model using predictor variables in the input data set. Examples include polynomial effects and spline effects.
- OUTPUT** creates a new SAS data set that saves diagnostic measures calculated after fitting the model. At least one *keyword=names* specification is required.

## The SGSCATTER Procedure

General form of the SGSCATTER procedure:

```
PROC SGSCATTER options;
  COMPARE X= variable | (variable-1...variable-n)
           Y= variable | (variable-1 ... variable-n) /options;
  MATRIX variable-1 < ... variable-n> / options;
  PLOT plot-request(s) / options;
RUN;
```

19

PROC SGSCATTER creates a paneled graph of scatter plots for multiple combinations of variables. You can use options to overlay fit plots and ellipses on your scatter plots. PROC SGSCATTER can create many different types of paneled graphs, such as paneled graphs of scatter plots, paneled graphs of scatter plots with shared axes, a scatter plot matrix with prediction ellipses, and a diagonal with histograms and density plots.

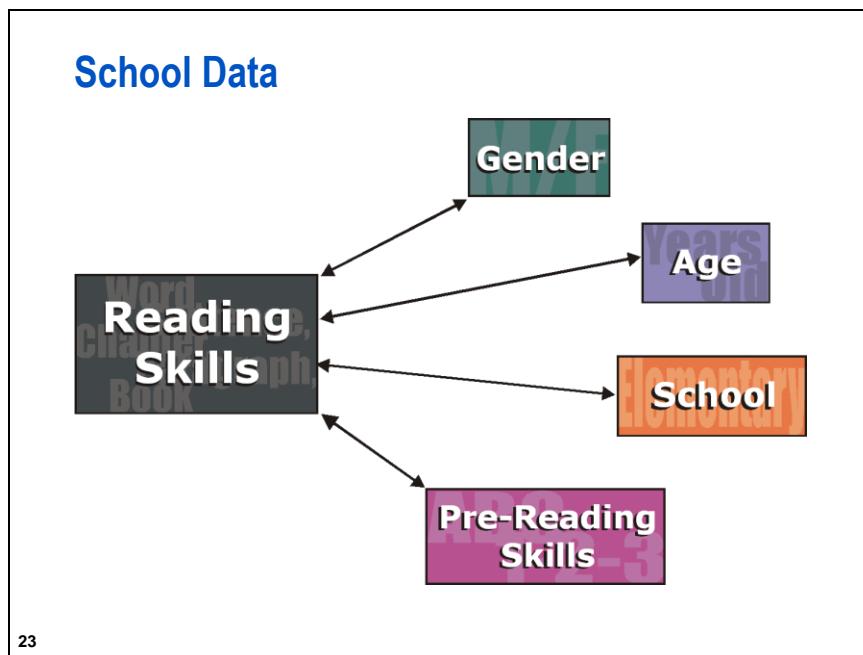
Selected SGSCATTER procedure statements:

**COMPARE** creates a comparative panel of scatter plots with shared axes.

**MATRIX** creates a scatter plot matrix.

**PLOT** creates a paneled graph that contains multiple independent scatter plots.

Notice that the graphs that you create with PROC SGSCATTER can have many individual graph cells. As the number of cells increases, the overall graph size does not automatically increase. To increase the graph size, use the HEIGHT= and WIDTH= options in the ODS GRAPHICS statement.



Data were collected by a school district to assess the reading skill progress of students in their first year of formal schooling. A simple random sample of students was selected from all the students in the district. The following are the variables in the **STAT2.school** data set:

<b>ID</b>	number of student
<b>GENDER</b>	gender of student ( <i>m,f</i> )
<b>AGE</b>	student's age (rounded to nearest tenth of a year)
<b>SCHOOL</b>	school student attends
<b>TEACHER</b>	name of student's teacher
<b>SEMESTERS</b>	number of semesters that student attended in the district
<b>LETTERS1</b>	score on letter identification test in the fall
<b>PHONICS1</b>	score on letter sound test in the fall
<b>WORDS1</b>	score on word identification test in the fall
<b>PHONICS2</b>	score on letter sound test in the winter
<b>WORDS2</b>	score on word identification test in the winter
<b>PHONICS3</b>	score on letter sound test in the spring
<b>WORDS3</b>	score on word identification test in the spring
<b>READING2</b>	score on reading test in the winter
<b>FLUENCY2</b>	score on reading fluency test in the winter
<b>READING3</b>	score on reading test in the spring
<b>FLUENCY3</b>	score on reading fluency test in the spring



## Multiple Linear Regression

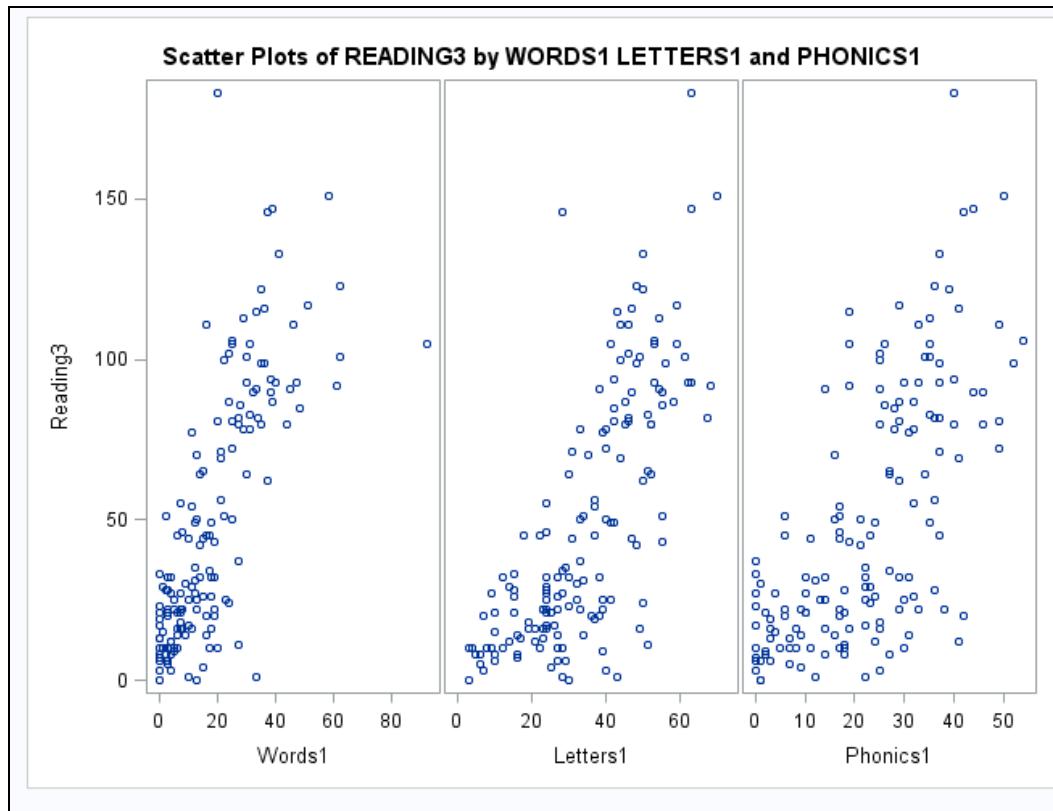
The school district is interested in predicting the spring reading scores (**reading3**) based on tests taken by students in the fall. Use the SGSCATTER procedure to generate a panel of scatter plots to examine the relationship between **reading3** and **words1**, **letters1**, and **phonics1**.

```
proc sgscatter data=STAT2.school;
  compare y=reading3 x=(words1 letters1 phonics1);
  title 'Scatter Plots of READING3 by WORDS1 LETTERS1 and PHONICS1';
run; *ST201d01.sas;
```

Selected SGSCATTER statement:

COMPARE creates a comparative panel of scatter plots with shared axes.

PROC SGSCATTER Output



All three predictor variables, **words1**, **letters1**, and **phonics1**, appear to have a positive relationship with **reading3** scores. It appears that **words1** has the strongest linear correlation and **phonics1** has the weakest.

Use the GLMSELECT procedure to generate a multiple linear regression with **reading3** as the response variable and **words1**, **letters1**, and **phonics1** as the predictor variables. Use the OUTPUT statement to save the residuals to a data set. Request the diagnostic panel of plots using PROC UNIVARIATE.

```
title 'School Data: Regression and Diagnostics';
proc glmselect data=STAT2.school;
  model reading3=words1 letters1 phonics1 / selection=none;
  output out=out r=residuals;
run;                                         *ST201d01.sas;
```

Selected system option:

PROC GLMSELECT Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	168543	56181	114.73	<.0001
Error	150	73453	489.68756		
Corrected Total	153	241996			

Root MSE	22.12889
Dependent Mean	49.24026
R-Square	0.6965
Adj R-Sq	0.6904
AIC	1113.78734
AICC	1114.19274
SBC	969.93515

The ANOVA table shows three degrees of freedom for the model. Recall that this is the number of parameters minus one. In this case, there are four parameters, the intercept and three slopes (one for each of the independent variables).

The *p*-value for the model is less than 0.0001. Because this is smaller than any reasonable alpha level, you reject the null hypothesis and conclude that at least one slope is not equal to zero. This model is better than the mean model for predicting **reading3**.

Given a significant model in multiple linear regression, you need to look further to decide which variables have slopes that are significantly different from zero. This can be discerned by looking at the parameter estimates with their *t* tests.

The table below the ANOVA table indicates that the R square is 0.6965, so the percent of the variability in **reading3** that is explained by the model is approximately 69.65%. In other words, the results of the three tests administered in the fall of the school year account for approximately 69% of the variability in the scores of the reading test administered in the spring.

The R square always increases as you include more terms in the model and, as a result, can be misleading when used to compare models. The adjusted R square is a measure similar to R square but takes into account the number of terms in the model. Therefore, when you compare models, it is more appropriate to compare the adjusted R square. The adjusted R square for this model is 0.6904.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr >  t
<b>Intercept</b>	1	-10.793658	4.605831	-2.34	0.0204
<b>Words1</b>	1	0.937366	0.180363	5.20	<.0001
<b>Letters1</b>	1	0.707867	0.181299	3.90	0.0001
<b>Phonics1</b>	1	0.847816	0.161333	5.26	<.0001

Looking at the parameter estimates, the estimated regression equation is as follows:

$$\text{reading3} = -10.79366 + 0.93737*\text{words1} + 0.70787*\text{letters1} + 0.84782*\text{phonics1}$$

The *t*-values and *p*-values in the Parameter Estimates table test the null hypothesis that the slope for each of the independent variables is equal to zero. Looking at these values, and presuming an alpha equal to 0.05, you reject the null hypothesis in each case. All of these variables are significant in predicting **reading3**.

You should be careful when you interpret the tests of hypothesis for the parameter estimates. They test the significance of each variable when it is added to a model that contains all of the other independent variables. As a result, if the independent variables in the model are correlated with one another, the significance of both variables can be hidden in these tests. Therefore, you should not remove more than one variable at a time from the model, based on these tests.

- ✍ The significance level of the test does not depend on the order in which you list the independent variables in the model. It does depend on the variables included in the model.

As with any statistical analysis, the assumptions of the analysis should be examined to ensure they were met. The OUTPUT statement of PROC GLMSELECT or the Output Delivery System (ODS) is used to output the residuals and save them to a data set. PROC UNIVARIATE can then be used to obtain an analysis of the normality of the residuals. Histograms and normal quantile plots of the residuals, as well as formal tests of normality, can be obtained using PROC UNIVARIATE with the HISTOGRAM and QQPLOT statements.

```
proc univariate data=out;
  var residuals;
  histogram residuals / normal kernel;
  qqplot residuals / normal(mu=est sigma=est);
run;                                              *ST201d01.sas;
```

Selected UNIVARIATE procedure statement:

**VAR** specifies numeric variables to analyze and the order in which they appear in the results. If no VAR statement is used, all numeric variables in the data set are analyzed.

Selected HISTOGRAM statement options:

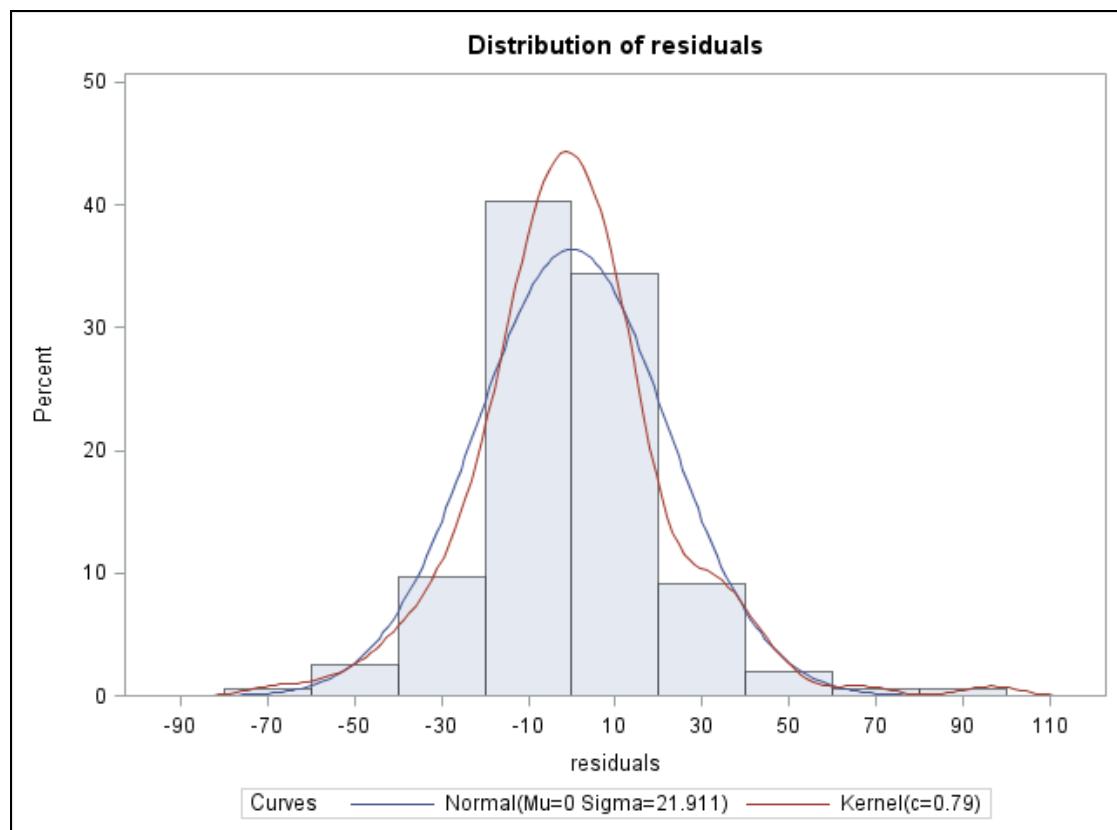
**NORMAL** requests an overlaid normal distribution plot as well as a series of goodness-of-fit tests based on the empirical distribution function. The table provides test statistics and  $p$ -values for the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Cramér-von Mises test. These tests can be requested without the histogram by using the NORMAL option in the PROC UNIVARIATE statement. This option does not apply if you use a WEIGHT statement.

**KERNEL** requests an overlaid kernel distribution plot.

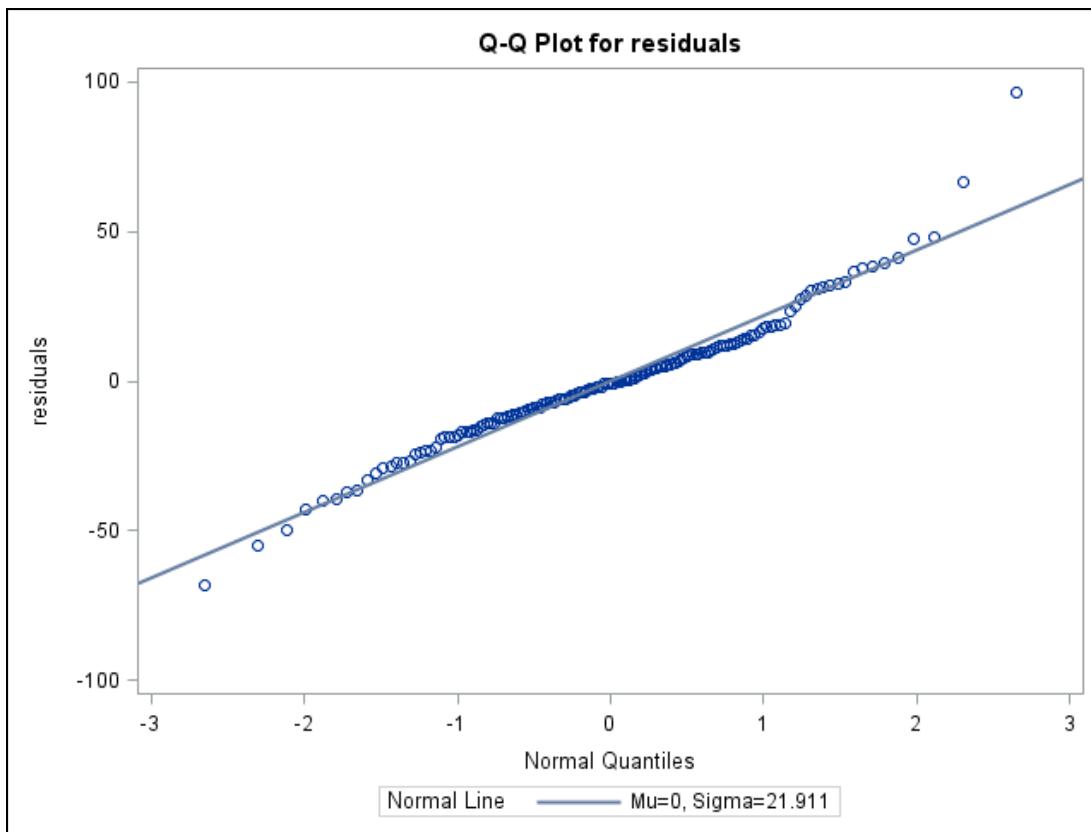
Selected QQPLOT statement option:

**NORMAL** along with the MU=est and SIGMA=est options, requests a reference line on the QQ-plot. This line represents the quantiles of a normal distribution having the same mean and standard deviation as the data.

Partial PROC UNIVARIATE ODS Graphics Output



The histogram of the residuals is shown with a normal density curve and a kernel density curve overlaid. The normal density, represented by the line with the lower peak, is constructed assuming that the data are from a normal distribution. The kernel density is represented by the line with the higher peak. It makes minimal assumptions about the functional form of the data and enables the data to determine the shape of the curve. The histogram of the residuals shows a fairly normal distribution.



The normal quantile plot shows that the residuals closely follow the reference line. There is a slight indication of an S-shaped curve and there seems to be one large error.

#### Partial PROC UNIVARIATE Output

The UNIVARIATE Procedure Variable: residuals (Residual)			
Moments			
<b>N</b>	154	<b>Sum Weights</b>	154
<b>Mean</b>	0	<b>Sum Observations</b>	0
<b>Std Deviation</b>	21.9108612	<b>Variance</b>	480.085839
<b>Skewness</b>	0.51306844	<b>Kurtosis</b>	2.68383889
<b>Uncorrected SS</b>	73453.1334	<b>Corrected SS</b>	73453.1334
<b>Coeff Variation</b>	.	<b>Std Error Mean</b>	1.76562751
Basic Statistical Measures			
Location		Variability	
<b>Mean</b>	0.00000	<b>Std Deviation</b>	21.91086
<b>Median</b>	-0.80052	<b>Variance</b>	480.08584
<b>Mode</b>	.	<b>Range</b>	164.76772
		<b>Interquartile Range</b>	22.69531

The skewness statistic is 0.51 and the mean is larger than the median, both of which indicate a possible skewed-to-the-right distribution of the residuals.

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.07746949	Pr > D	0.023
Cramér-von Mises	W-Sq	0.22750023	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.29873627	Pr > A-Sq	<0.005

All of the normality tests reject the null hypothesis of normally distributed residuals. These results should be evaluated in conjunction with the histogram and normal probability plots of the residuals.

-  All the normality tests depend on the sample size. As the sample size becomes larger, increasingly smaller departures from normality can be detected. A small sample size likely yields a less powerful test and you might want to use a higher alpha value.

## Details

When you specify the NORMAL option in the PROC UNIVARIATE statement, you obtain the Tests for Normality table. The table contains all three empirical distribution functions' (EDF) goodness-of-fit tests, plus the Shapiro-Wilk  $W$  statistic and test. The  $W$  statistic is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance (Shapiro and Wilk 1965). When  $n$  is greater than three, the coefficients to compute the linear combination of the order statistics are approximated by the method of Royston (1992). The  $W$  statistic is always greater than zero and less than or equal to one ( $0 < W \leq 1$ ). Small values of  $W$  lead to the rejection of the null hypothesis of normality. The distribution of  $W$  is highly skewed. Seemingly large values of  $W$  (such as 0.90) might be considered small and lead you to reject the null hypothesis. The method for computing the  $p$ -value (the probability of obtaining a  $W$  statistic less than or equal to the observed value) depends on  $n$ .

The Kolmogorov-Smirnov  $D$  statistic, the Cramér-von Mises  $W^2$  statistic, and the Anderson-Darling  $A^2$  statistic are based on the EDF. The EDF tests offer advantages over the traditional chi-square goodness-of-fit test, including improved power and invariance with respect to the histogram midpoints. For detailed information about the computation of each statistic, refer to the SAS online documentation. A thorough discussion about the EDF tests can be found from D'Agostino and Stephens (1986). Here are some highlights.

### Kolmogorov-Smirnov D Statistic and Test:

- It is the most well-known EDF statistic, but it is often less powerful than  $W^2$  and  $A^2$  statistics.
- It does not depend on the underlying cumulative distribution function being tested.
- It is an exact test. (The chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid.)
- It tends to be more sensitive near the center of the distribution than at the tails.

### Cramér-von Mises $W^2$ Statistic and Test

- It is a modification of the Kolmogorov-Smirnov (K-S) test.
- It allows a more sensitive and powerful test.

**Anderson-Darling A<sup>2</sup> Statistic and Test**

- It is a modification of the Kolmogorov-Smirnov (K-S) test. It uses the specific distribution in calculating critical values and gives more weight to the tails than does the K-S test.
- It behaves similarly to the Cramér-von Mises W<sup>2</sup> statistic. However, it is more powerful when departures from the true distribution are in the tails, especially when there appear to be too many outlying values from the data for the specified distribution.
- It is a recommended statistic when departures in the tails are important to detect.

The results of these tests should be considered in conjunction with the distribution of the variable, the histogram, and the normal probability plot to evaluate the normality.

## 1.04 Multiple Choice Poll

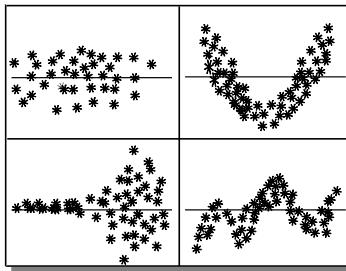
You learned from the demonstration that you should check the assumption that the error terms are normally distributed. How can you do this?

- examine the histogram and normal quantile plot of the residuals
- request formal tests of normality for the residuals in PROC UNIVARIATE
- either a or b

25

## Common Regression Problems

- The residual plots shown here identify some common regression problems:
  - misspecified model
  - nonconstant variance
  - correlated error terms
- Other regression problems include the following:
  - influential observations
  - collinearity



27

Some common regression problems include model misspecification, nonconstant variance, correlated error terms, influential observations, and collinearity. The first three of these are shown in the residual plots above.

The graph in the left column of the first row of the panel of charts above is a plot of the residuals versus the predicted values. This graph shows random scatter about a reference line placed at zero, indicating that the model is properly specified, and has constant variance. The remaining three graphs indicate the following problems with the model:

- The graph in the right column of the first row is a graph of the residuals versus the predicted values and indicates a discernable pattern. This can be an indication of *model misspecification*. A different model (for example, polynomial regression) might be needed for this situation.
- The graph in the left column of the second row is also a graph of the residuals versus the predicted values and indicates nonconstant variance. Nonconstant variance is a violation of the assumptions for linear regression and might arise in both simple and multiple linear regression. It can sometimes be corrected by a transformation of the response variable.
- The graph in the right column of the second row is a graph of the residuals versus time and indicates *correlated error terms*, which is a violation of the assumptions. Correlated error terms can result from clustered data, repeated measures data, and time series data. Different modeling tools are needed to model data with correlated errors.

*Influential observations* might arise in both simple linear regression and multiple regression and can be a violation of the assumptions, but not always. They could be an indication of erroneous data. Another possibility is that the observations, though valid, might be unusual. If you had a larger sample size, there might be more observations like the unusual ones. You might have to collect more data to confirm the relationship suggested by the influential observations. You can identify influential observations by using statistics such as Cook's D and DFFITS. These diagnostic tools are available as options in the MODEL statement in PROC REG and are discussed later in the course.

*Collinearity* can arise in multiple linear regression. It is not a violation of the assumptions, but it can cause significant problems in modeling. One possible solution to collinearity is to delete the collinear predictor variables from the model. Another possible solution is to use biased regression techniques such as ridge regression or principal components regression. You can use the variance inflation factor and other collinearity diagnostic statistics to identify collinear predictor variables. These are available as options in the MODEL statement in PROC REG. (More detailed discussions about each of these issues are provided later in the course.)

## 1.05 Multiple Choice Poll

The residual plots (residuals versus predicted values, and also residuals versus time, if applicable) for regression models are important because they help to

- a. identify lack of fit of the model
- b. display nonconstant variance
- c. evaluate normality of the residuals
- d. display correlated errors
- e. a, b, and d.

## 1.2 Simple Polynomial Regression

---

### Objectives

- Do an initial exploratory data analysis using PROC SGPlot.
- Fit a polynomial regression model.

32

### Polynomial Regression Models

- Quadratic Polynomial Model

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \varepsilon_j$$

- Cubic Polynomial Model

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3 + \varepsilon_j$$

- Polynomial Model with a Cross-Product Term

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{1j} X_{2j} + \varepsilon_j$$

33

A *polynomial regression model* is a special type of multiple linear regression where powers of variables and cross-product (or interaction) terms are included in the model. Some examples of polynomial regression models are shown above. Polynomial regression models fall into the category of general linear models because they are *linear in the parameters*.

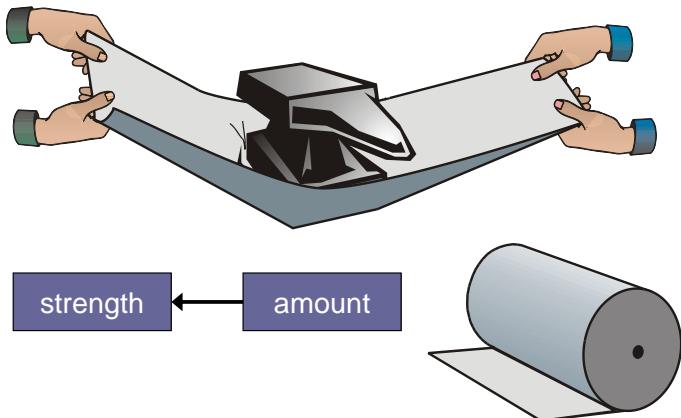
## 1.06 Multiple Choice Poll

Which of the following is **false**?

- a. Polynomial regression models belong to the category of nonlinear regression models.
- b. Polynomial regression models belong to the category of linear regression models.
- c. Polynomial regression models fit a curvilinear model to your data.
- d. all of the above
- e. none of the above

34

## Paper Example



36

A researcher is interested in studying the effect of a chemical additive on paper strength. Data are collected and stored in the **STAT2.paper** data set. The independent variable of interest is the amount of chemical additive (**amount**), and the dependent variable is the amount of force required to break the paper (**strength**).

## The Data

Obs	amount	strength
1	1	2.4
2	1	2.6
3	1	2.7
4	2	2.5
5	2	2.6
6	2	2.6
7	2	2.7
8	2	2.8
9	3	2.8
10	3	2.8
11	3	3.0
12	3	3.0
13	4	3.0
14	4	2.9
15	4	2.9
16	4	3.0
17	4	3.1
18	5	2.9
19	5	2.9
20	5	3.0
21	5	2.9
22	5	2.8

37

PROC SGPlot can be used for exploring the data graphically. The SGPlot procedure creates one or more plots and overlays them on a single set of axes. You can use the SGPlot procedure to create statistical graphics such as histograms and regression plots in addition to simple graphics such as scatter plots and line plots. To overlay plots, you request two or more plots in a single PROC SGPlot call.

## The SGPlot Procedure

General form of the SGPlot procedure:

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

38

Statements and options enable you to control the appearance of your graph and add additional features such as legends and reference lines.

Selected SGPlot procedure statements:

**HISTOGRAM** creates a histogram that displays the frequency distribution of a numeric variable.

REG creates a fitted regression line or curve.

SCATTER creates a scatter plot.

## The EFFECT Statement

```
EFFECT effect-name=effect-type(<var-list></options>);
```

```
effect qstress=polynomial(stress/degree=2) ;
effect t_exposure=spline(exposure) ;
```

39

The EFFECT statement in PROC GLMSELECT is used to create constructed predictor effects for the model. Each type of effect has its own options to customize the analysis. For example, the DEGREE= option specifies the order for a polynomial effect.

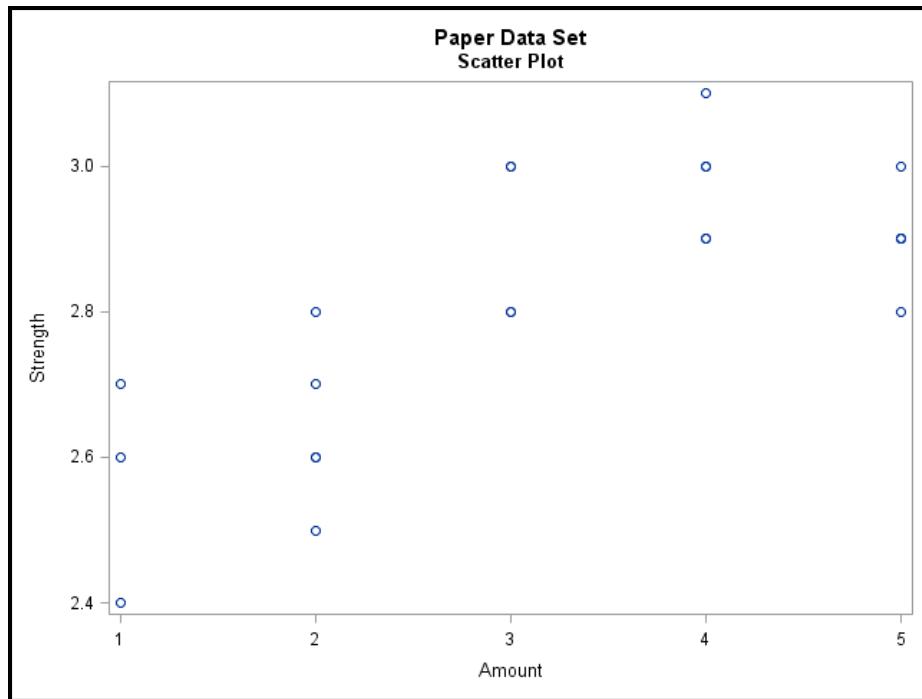
- ✍ Several types of constructed effects are available. See the SAS documentation for PROC GLMSELECT for a full list.



## Simple Polynomial Regression

You are interested in determining whether the amount of additive is related to the strength of the paper. Before generating a model for the data, explore the nature of the relationship between the variables by generating a scatter plot of the data.

```
title1 "Paper Data Set";
proc sgplot data=STAT2.paper;
  scatter x=amount y=strength;
  title2 "Scatter Plot";
run; *ST201d02.sas;
```



From the graph, it appears that the strength increases as the amount of additive increases up to a certain point. Then the strength of the paper begins to decrease.

This is an indication that a simple linear regression might not be appropriate and a polynomial regression might be necessary.

Explore polynomial relationships between **strength** and **amount** by fitting smoothed second-, third-, and fourth-degree curves to the scatter plot using the REG statement in PROC SGPlot. The plots can be overlaid by putting three REG statements in one PROC SGPlot call. However, the plots are easier to interpret if they are separate, so three PROC SGPlot calls are used instead.

```

proc sgplot data=STAT2.paper;
  reg x=amount y=strength / lineattrs=(color=brown pattern=solid)
    legendlabel="Linear";
title2 "Linear Model";
run;

proc sgplot data=STAT2.paper;
  reg x=amount y=strength / degree=2 lineattrs=(color=green
    pattern=mediumdash) legendlabel="2nd Degree";
title2 "Second Degree Polynomial";
run;

proc sgplot data=STAT2.paper;
  reg x=amount y=strength / degree=3 lineattrs =(color=red
    pattern=shortdash) legendlabel="3rd Degree";
title2 "Third Degree Polynomial";
run;

proc sgplot data=STAT2.paper;
  reg x=amount y=strength / degree=4 lineattrs =(color=blue
    pattern=longdash) legendlabel="4th Degree";
title2 "Fourth Degree Polynomial";
run;                                *ST201d02.sas;

```

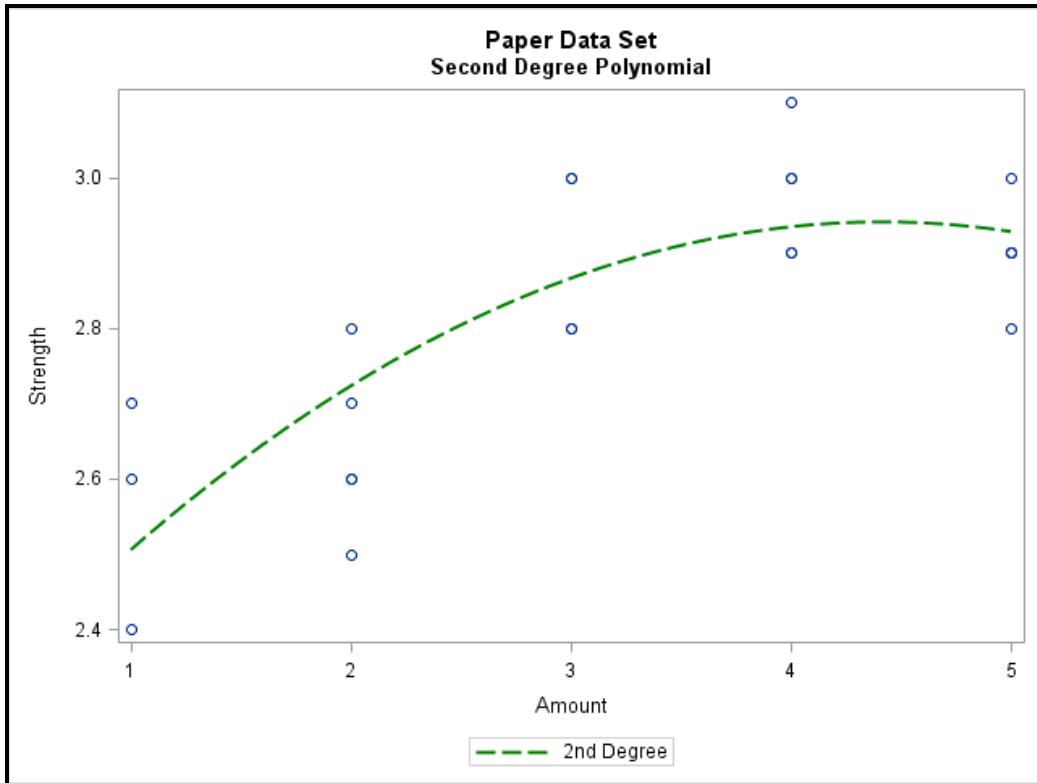
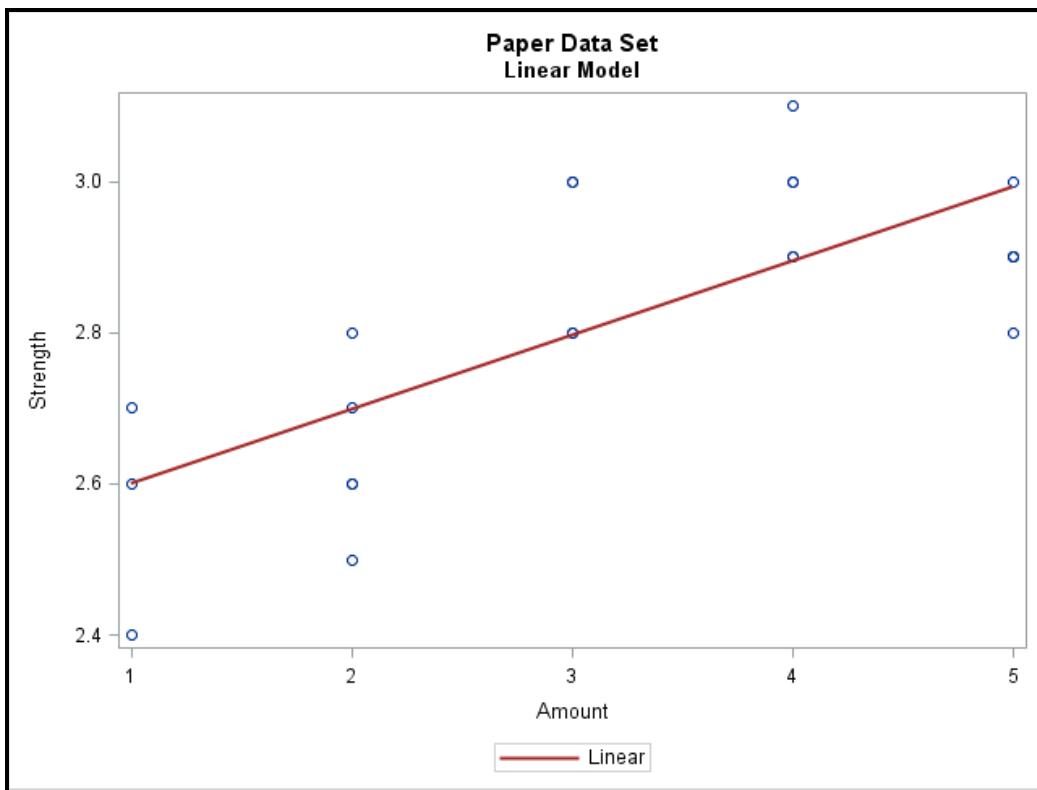
Selected REG statement options:

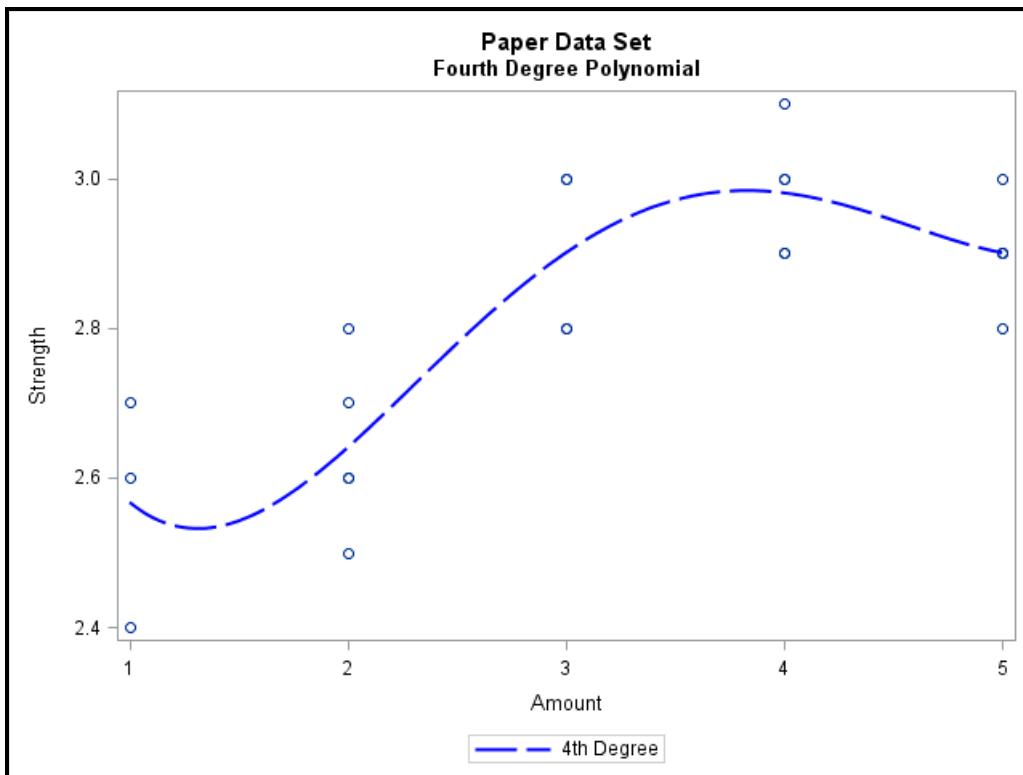
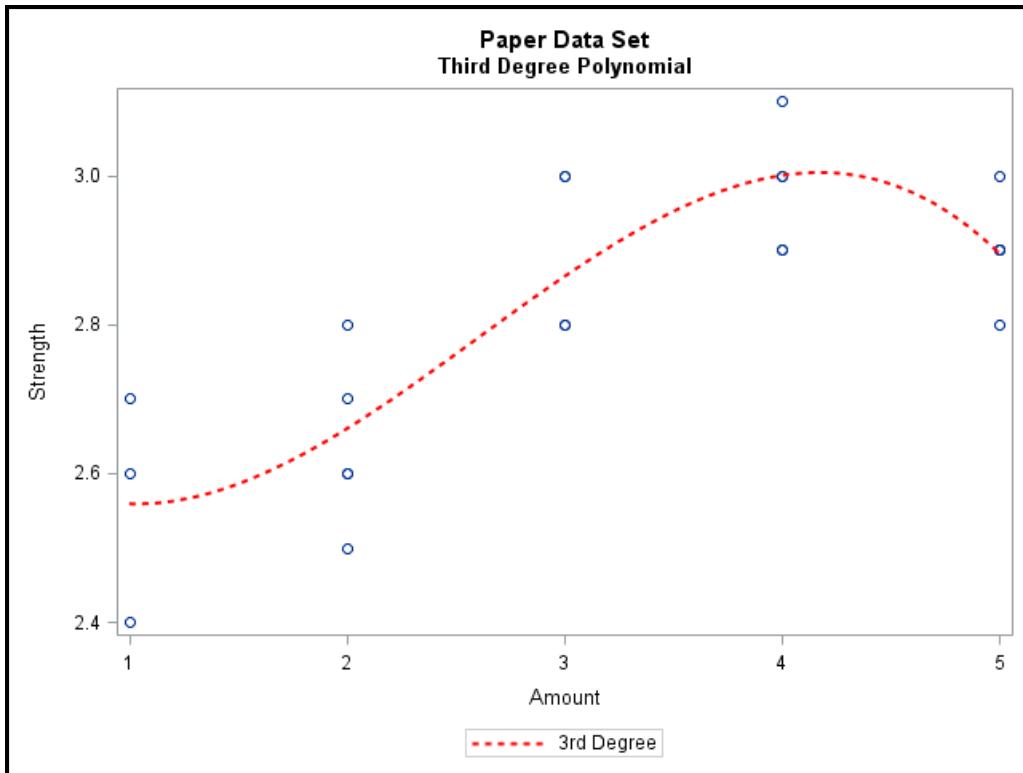
- DEGREE= specifies the degree of the polynomial fit. For example, DEGREE=1 specifies a linear fit, DEGREE=2 specifies a quadratic fit, and DEGREE=3 specifies a cubic fit.
- LINEATTRS= specifies the appearance of the fit line. You can specify the appearance by using a style element or by using suboptions. If you specify a style element, you can also specify suboptions to override specific appearance attributes.
- LEGENDLABEL= specifies a label that identifies the regression curve in the legend. By default, the label “Regression” is used.

Selected LINEATTRS= suboptions:

- COLOR specifies the color of the line.
- PATTERN specifies the line pattern for the line. You can reference SAS patterns by number or by name. See the SAS online documentation for a list of pattern numbers and names.

## PROC SGLOT Output





Although it is clear that a linear relationship does not capture the curvature represented in the data, it is unclear whether the relationship is captured best by a second-, third-, or fourth-degree polynomial. Therefore, you might want to start with a fourth-degree equation.

Rather than running a separate model for each polynomial, start with the highest-degree model and use sequential tests to refine the model. The polynomial terms are created by the EFFECT statement with the POLYNOMIAL keyword. The DEGREE=4 option is used to fit a fourth-degree polynomial for the variable **amount**. The OUTDESIGN= option is used to create a data set called **d\_paper** for subsequent analysis in PROC REG.

```
title;
title2;

proc glmselect data=STAT2.paper outdesign=d_paper;
  effect p_amount=polynomial(amount / degree=4);
  model strength = p_amount / selection=none;
  title "Paper Data Set: 4th Degree Polynomial";
run; *ST201d03.sas;
```

PROC GLMSELECT Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.53924	0.13481	12.28	<.0001
Error	17	0.18667	0.01098		
Corrected Total	21	0.72591			

Root MSE	0.10479
Dependent Mean	2.81364
R-Square	0.7429
Adj R-Sq	0.6823
AIC	-70.92841
AICC	-65.32841
SBC	-89.47320

The ANOVA table indicates that the model fits the data significantly better than the mean model. The R-square value is 0.7429, meaning that about 74% of the variation in breaking strength is explained by the model.

Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Pr >  t	
Intercept	1	3.433333	0.801698	4.28	0.0005	
Amount	1	-1.684444	1.459267	-1.15	0.2643	
Amount^2	1	1.023889	0.873803	1.17	0.2575	
Amount^3	1	-0.222222	0.209817	-1.06	0.3044	
Amount^4	1	0.016111	0.017428	0.92	0.3682	

The Parameter Estimates table provides the estimates for the intercept and slopes for **Amount**, **Amount^2**, **Amount^3**, and **Amount^4**, as well as the associated standard errors and *p*-values. It seems that none of the slopes is significantly different from zero at an alpha level of 0.05. You should be careful when you interpret the tests of hypothesis for the parameter estimates. They test the significance of each variable given that all of the other independent variables are already in the model. As a result, if the independent variables in the model are correlated with one another (a situation often referred to as *multicollinearity*), the significance of both variables can be hidden in these tests. In polynomial regressions, all the higher-ordered terms are correlated with the corresponding independent variable. Multicollinearity among **Amount**, **Amount^2**, **Amount^3**, and **Amount^4** is likely the cause of nonsignificant *p*-values for all the slopes.

It is usually desirable to construct hierarchically well-formulated models. This means that a model that includes a variable to a power should also include all lower powers of the variable. Likewise, a model that includes a cross-product term should also include each of the individual variable terms. An exception to this would be if there were an overriding physical reason that leads you to conclude that the true population model does not include the lower-order term. In PROC GLMSELECT, you can enforce model hierarchy by using the HIERARCHY= option in the MODEL statement.

Remove nonsignificant polynomial terms one at a time by using backward elimination, and output the design matrix to a data set so that additional diagnostics can be performed in PROC REG.

```
proc glmselect data=STAT2.paper outdesign=d_paper;
  effect p_amount=polynomial(amount / degree=4);
  model strength=p_amount / selection=backward select=sl
    slstay=0.05 hierarchy=single showpvalues;
  title "Paper Data Set: Model Selection";
run; *ST201d03.sas;
```

Selected MODEL statement options:

**SELECTION=** specifies the method used to perform automatic model selection. FORWARD, BACKWARD, STEPWISE, LAR, LASSO, or NONE (using the full model) can be specified. The default method is STEPWISE.

**SELECT=** specifies the criterion that PROC GLMSELECT uses to determine the order in which effects enter or leave (or do both) at each step of the specified selection method. To request the traditional approach where effects enter and leave the model based on the significance level, use **SELECT=SL**. In this example, the additional **SLSTAY=** option specifies the significance level that must be achieved for a variable to remain in the model.



Additional details about automatic model selection options in PROC GLMSELECT are provided later in the chapter.

HIERARCHY= specifies how model hierarchy is applied. When HIERARCHY=SINGLE is specified, only single terms can enter or leave the model at one time, following the principle of model hierarchy described above. The default is HIERARCHY=NONE.

Partial PROC GLMSELECT Output

**The selected model is the model at the last step (Step 1).**

Effects:	Intercept	Amount	Amount^2	Amount^3
----------	-----------	--------	----------	----------

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.52986	0.17662	16.22	<.0001
Error	18	0.19605	0.01089		
Corrected Total	21	0.72591			

Root MSE	0.10436
Dependent Mean	2.81364
R-Square	0.7299
Adj R-Sq	0.6849
AIC	-71.84939
AICC	-68.09939
SBC	-91.48522

Backward elimination drops **Amount^4** and fits a cubic model with **Amount**, **Amount^2**, and **Amount^3**. The cubic term is now significant, so the model selection stops at this point. The selected model has an adjusted R-square value of 0.6849, which is slightly higher than the previous fourth-degree model.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr >  t
Intercept	1	2.732803	0.260598	10.49	<.0001
Amount	1	-0.368996	0.322081	-1.15	0.2669
Amount^2	1	0.223389	0.116506	1.92	0.0712
Amount^3	1	-0.028619	0.012699	-2.25	0.0369

The regression equation is the following:

$$\text{Strength} = 2.73280 - 0.36900 * \text{Amount} + 0.22339 * \text{Amount}^2 - 0.02862 * \text{Amount}^3$$

To check the model assumptions, you use the built-in ODS Graphics capability of PROC REG to request a plot of the residuals versus the predicted values and a normal probability plot of the residuals. The input data set for PROC REG is **d\_paper4**, which was created by the OUTDESIGN= option of PROC GLMSELECT. The macro variable **&\_GLSMOD** is created automatically by PROC GLMSELECT. It contains all of the predictor effects (including those created by EFFECT statements) from the most recent run of the procedure. When automatic selection is used, **&\_GLSMOD** contains only the effects in the final model.

```
proc reg data=d_paper plots (unpack)=(diagnostics (stats=none));
  Cubic_Model: model strength=&_GLSMOD / lackfit;
  title "Paper Data Set: 3rd Degree Polynomial Model";
run;
quit;
*ST201d04.sas;
```

Selected PROC REG option:

PLOTS= controls the plots produced through ODS Graphics. By default, ODS Graphics are enabled in SAS 9.3 and higher. If you do not specify the PLOTS= option, then PROC REG produces a default set of plots. *Plot-requests* include panels of plots, such as the DIAGNOSTICS, RESIDUALPLOT, and FITPLOT panels. Individual plots might also be requested. See the SAS online documentation for a complete list of available plots. For general information about ODS Graphics, see the chapter about “Statistical Graphics Using ODS” in *SAS/STAT® User’s Guide 13.1*.

The *global-options* apply to all plots generated by the REG procedure, unless they are altered by *specific-plot-options*.

Selected PLOTS= options (*global-options*):

UNPACK as a *global option*, this suppresses paneling of all plots.

LABEL specifies that the LABEL option be applied to each plot that supports a LABEL option. See the descriptions of the specific plots for details.

Selected PLOTS= *plot-request* option:

DIAGNOSTICS produces a summary panel of fit diagnostics: residuals versus the predicted values, studentized residuals versus the predicted values, studentized residuals versus the leverage, normal quantile plot of the residuals, dependent variable values versus the predicted values, Cook’s D versus observation number, histogram of the residuals, “Residual-Fit” (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals, and a box plot of the residuals if you specify the STATS=NONE suboption.

Selected DIAGNOSTICS *plot-request* options:

STATS= requests statistics that are included on a diagnostics panel. STATS=ALL requests all these statistics; STATS=NONE suppresses them.

UNPACK as a *plot-request* option, this suppresses paneling of specific plots.

The STATS and UNPACK options are the only two options that are available for the DIAGNOSTICS panel of plots.

Selected MODEL statement option:

LACKFIT performs a lack-of-fit test. The test for lack of fit compares the variation around the model with “pure” variation within replicated observations. This test measures the adequacy of the specified model and can be specified only if some observations in your design are replicated.

### Details

The test for lack of fit compares the variation around the model with “pure” variation within replicated observations. In particular, if there are  $n_i$  replicated observations  $Y_{i1}, \dots, Y_{in_i}$  of the response all at the same values  $x_i$  of the regressors, then you can predict the true response at  $x_i$  either by using the predicted value  $\hat{Y}_i$  based on the model or by using the mean  $\bar{Y}_i$  of the replicated values. The test for lack of fit decomposes the residual error into a component due to the variation of the replications around their mean value (the “pure” error) and a component due to the variation of the mean values around the model prediction (the “bias” error).

$$\sum_i \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_i)^2 = \sum_i \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_i n_i (\bar{Y}_i - \hat{Y}_i)^2$$

If the model is adequate, then both components estimate the nominal level of error. However, if the bias component of error is much larger than the pure error, then this constitutes evidence that there is significant lack of fit.

 A significant result for the lack-of-fit test indicates that the specified model is inadequate, so if this is a problem, you might want to refine the model.

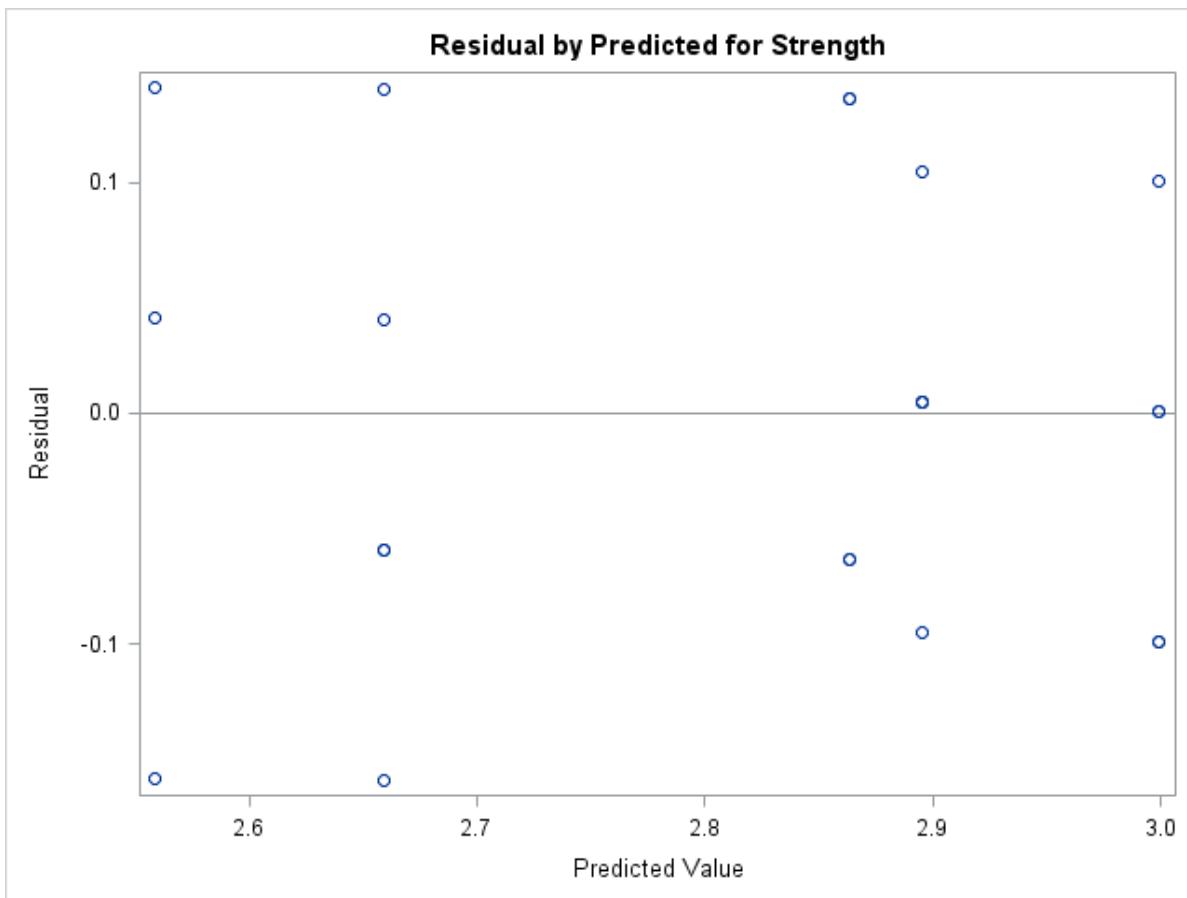
Partial PROC REG Output

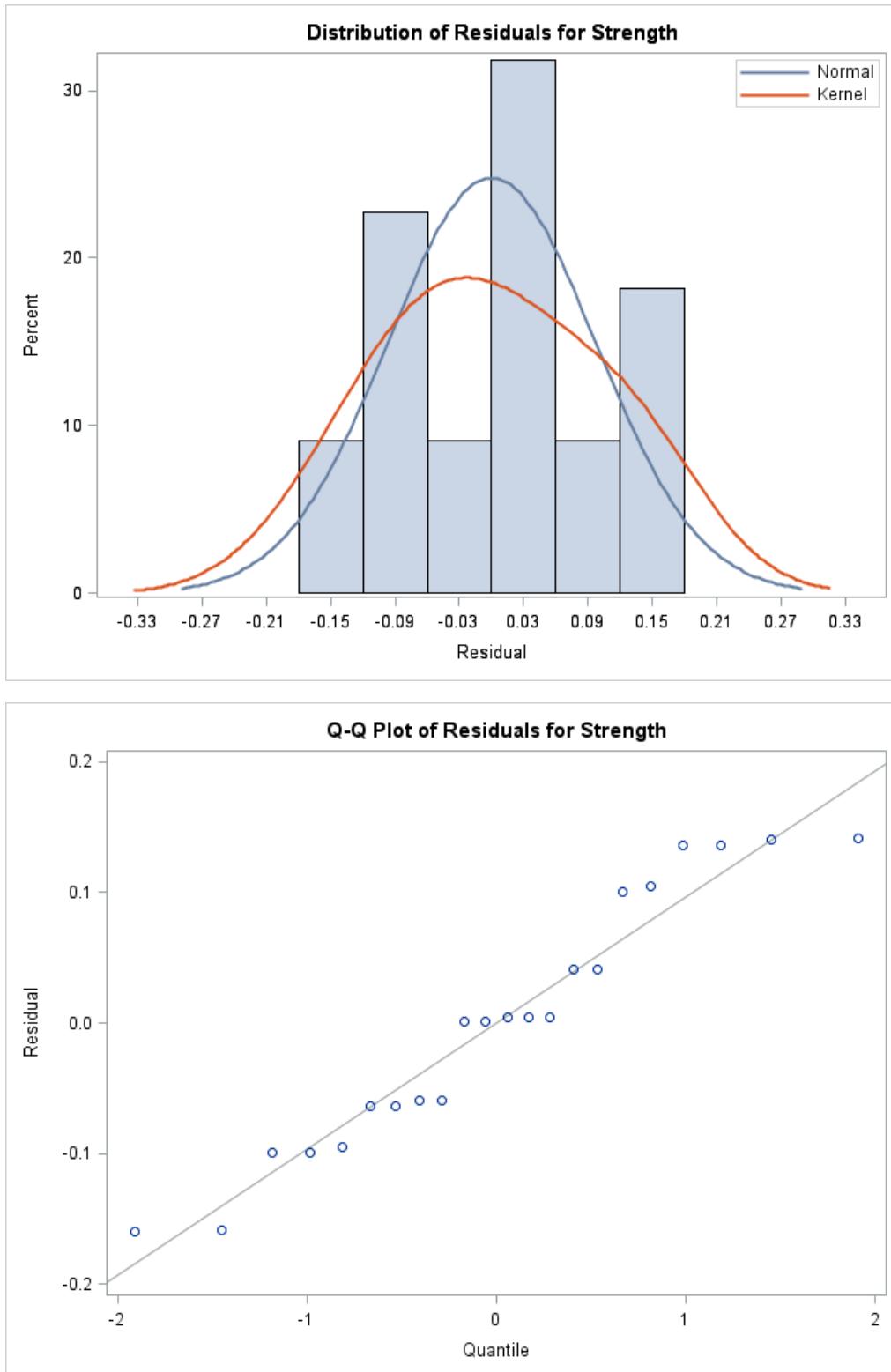
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.52986	0.17662	16.22	<.0001
Error	18	0.19605	0.01089		
Lack of Fit	1	0.00938	0.00938	0.85	0.3682
Pure Error	17	0.18667	0.01098		
Corrected Total	21	0.72591			

The Analysis of Variance table now contains additional rows for the lack-of-fit test. The test is not significant ( $F$ -value 0.85 and  $p$ -value 0.36820). It indicates that there is not enough evidence to conclude that the specified model is inadequate.

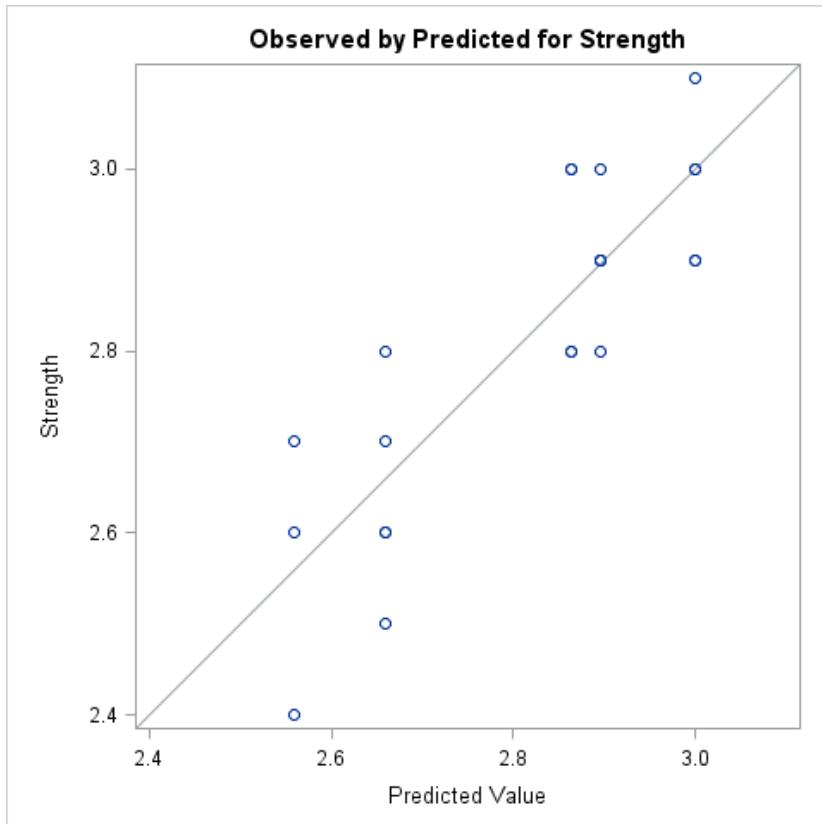
The following plots are produced by the PLOTS= DIAGNOSTICS option in the PROC REG statement.

PROC REG ODS Graphics Output





The plot of the residuals versus the predicted values appears to be a random scatter about the zero reference line. The histogram might indicate a problem with the normality assumption, but the normal quantile plot appears less problematic.



The plot of the observed values by the predicted values is also created as part of the DIAGNOSTICS panel of plots and does not indicate any serious lack of fit. Although the data set consists of 22 points, several points have the same values for both **amount** and **strength**. Therefore, predicted values for only the 15 unique combinations are shown.

## 1.07 Multiple Choice Poll

Which of the following is **false**?

- a. Polynomial regression is a nonlinear model, and therefore, you should not use PROC GLMSELECT.
- b. You should follow the principle of model hierarchy when you remove terms from a polynomial regression model.
- c. It is important to check model assumptions after the final polynomial regression model is chosen.



## Exercises

---

### 1. Fitting a Simple Polynomial Regression Model

An analyst for a cafeteria chain wanted to investigate whether the sales of coffee are related to the number of self-service coffee dispensers in a cafeteria line. Fourteen cafeterias that are similar in terms of volume of business, type of clientele, and location were chosen for the study. The number of self-service dispensers was assigned randomly at each cafeteria and the sales in hundreds of gallons of coffee were recorded.

Data are stored in the **STAT2.cafeteria** data set. These are the variables in the data set:

**cafeteria** cafeteria identification number

**dispensers** number of dispensers at each cafeteria

**sales** coffee sales in hundreds of gallons.

- a. Open the program **ST200d01.sas**. This program creates all SAS data sets for the demonstrations and exercises.

If you're running Base SAS or SAS Studio in the Virtual Lab, run the following code sequence to define the **STAT2** library:

```
%let homefolder=S:/workshop;
libname STAT2 "&homefolder";
```

If you're running SAS University Edition outside of the Virtual Lab environment, run the following code sequence to define the **STAT2** library:

```
%let homefolder=/folders/myfolders;
libname STAT2 "&homefolder";
```

Depending on how the folders were set up when SAS University Edition was installed, the path may be different.

- b. Use PROC SGSCATTER to plot **sales** versus **dispensers**. How is **sales** related to **dispensers**?
- c. Use PROC REG to fit a simple linear regression model to the data. Add the PLOTS (ONLY UNPACK)=DIAGNOSTICS options in the PROC REG statement to get the diagnostics panel of plots. Examine the plot of the residuals versus the predicted values and the plot of the observed versus the predicted values. Does your model seem to fit the data well?
- d. Fit a quadratic model by specifying an EFFECT statement in PROC GLMSELECT. Use the OUTDESIGN option to output the design matrix to a data set. Then use PROC REG with the &**\_GLSMOD** macro variable and the DIAGNOSTICS plot option to request diagnostic plots. Look at the plot of the residuals versus the predicted values, the plot of the observed versus the predicted values, and the normal quantile plot for residuals. Use PROC SGPLOT with a REG statement with the DEGREE=2 option to create a scatter plot of the data with a second-degree regression overlaid. From the plots, do you think the quadratic model fits your data better than the linear model?

### Advanced

- e. You can obtain plots of the residuals versus the regressors by adding the RESIDUALS keyword to your plots request in PROC REG. To help in detecting patterns, you can use the SMOOTH suboption of the RESIDUALS plots request to add LOESS smoothed curves to these residual plots. Consult the online documentation for PROC REG to see how to add the appropriate options to your code. Run PROC REG with two MODEL statements to compare the linear model to the quadratic model. Add the appropriate options to create the RESIDUALS plots with a smooth curve added. Compare the residual plots from the two models. Which model fits your data better?

### 1.08 Quiz

In the previous exercise, what did you find about the linear model when you examined the residual plot?

45

## 1.3 Polynomial Regression and Multicollinearity

---

### Objectives

- Define multicollinearity.
- Introduce diagnostics for multicollinearity.
- Discuss remedial measures for multicollinearity.

48

One of the difficulties encountered with polynomial regression is multicollinearity. As you add higher-ordered terms to the model, the terms are highly correlated with one another. What are the consequences of multicollinearity?

## Polynomial Regression and Multicollinearity

### *Multicollinearity*

- affects the parameter estimates
- affects the standard errors of the parameter estimates and predictions
- might not affect the predictions for future points
- can lead to serious rounding errors when the parameters are estimated.

49

*Multicollinearity* refers to the fact that some of the independent variables provide redundant information. It affects the parameter estimates because they depend on the correlated independent variables included in the regression model. In extreme cases, multicollinearity can cause the parameter estimates to have the wrong sign. It can also cloud the practical meaning of the true regression coefficients (Bowerman, O'Connell, and Dickey 1986).

Multicollinearity can inflate the standard errors of the parameter estimates and the standard errors of the predictions, and therefore, hinder your ability to determine the significance of the independent variables. However, it generally does not hinder the model's ability to predict a future value of the dependent variable based on a combination of future values of the independent variables, if the combination of future values of the independent variables is in the experimental region (Bowerman, O'Connell, and Dickey 1986). Multicollinearity can lead to serious rounding errors when you compute the point estimates of the parameters due to a nearly dependent  $\mathbf{X}$  matrix.

## Multicollinearity Diagnostics

- Correlation statistics (from PROC CORR)

$$\text{corr}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sqrt{\text{var}(x_1) \text{var}(x_2)}}$$

- Variance inflation factors (VIF) (from PROC REG)

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

- Condition index values (from PROC REG)

$$\eta_i = \sqrt{\frac{\lambda_{\max}}{\lambda_i}}$$

50

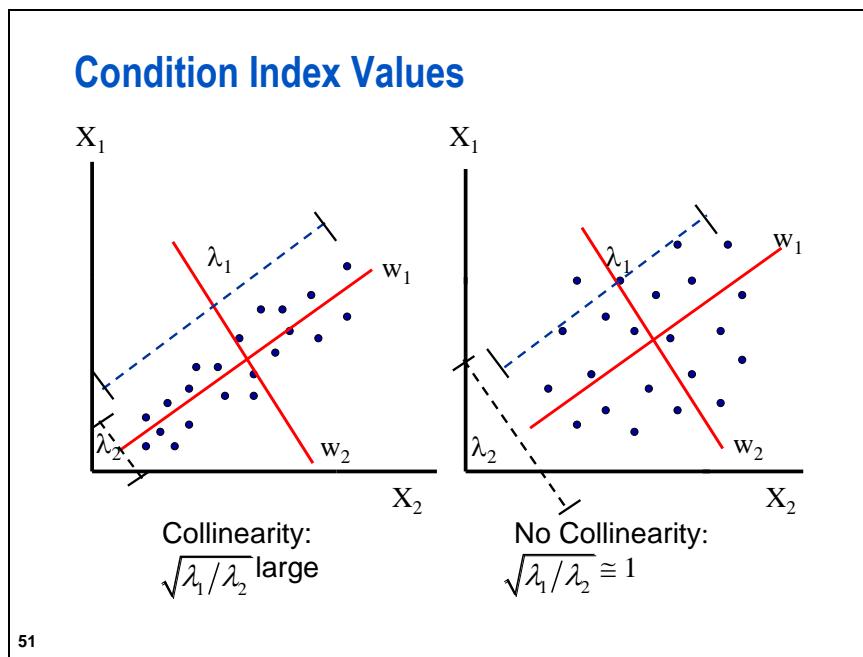
Correlation is a measure of the degree of linear relationship between two variables. The Pearson correlation coefficient is calculated as follows:

$$\text{corr}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sqrt{\text{var}(x_1) \text{var}(x_2)}} = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2}}$$

Correlation statistics are useful for evaluating collinearity between two independent variables. They cannot measure collinearity among more than two independent variables.

Variance inflation factors are useful for determining which variables might be involved in the multicollinearity. VIF is computed as  $1/(1-R_i^2)$ , where  $R_i^2$  is the coefficient of determination for the regression of the  $i^{\text{th}}$  independent variable on all other independent variables. You use the VIF option in the MODEL statement of PROC REG to obtain the variance inflation factor.

Condition index values are computed based on eigenanalysis. You can detect the presence of multicollinearity by examining the eigenvalues of the sums of squares and cross-products (SSCP) matrix  $\mathbf{X}'\mathbf{X}$ . You use the COLLIN or COLLINOINT option in the MODEL statement of PROC REG to obtain the condition index values.



A set of eigenvalues ( $\lambda$ s) of relatively equal magnitudes indicates little multicollinearity, and a wide variation in magnitudes indicates severe multicollinearity. Therefore, the ratio of the eigenvalues can be useful for examining multicollinearity.

#### Details

Collinearity occurs when linear combinations of some of the columns in the  $X$  matrix equal zero, or nearly zero. Geometrically, this occurs when at least one dimension of the  $X$ -space has very little dispersion (shown in the left graph above). When an independent variable has limited dispersion, its column in the  $X$  matrix is almost a multiple of a vector of ones, with the result that the variable is nearly collinear with the column for the intercept. The presence of collinearity is detected by *singular decomposition of  $X$*  or the *eigenanalysis of  $X'X$* .

A value  $\lambda$  is called the eigenvalue of the SSCP matrix  $X'X$  if there is a nonzero vector  $z$  such that  $(X'X)z = \lambda z$ . The nonzero vector  $z$  is called the *eigenvector*. Eigenvalues and eigenvectors of the SSCP matrix  $X'X$  are closely related to the principal components of the matrix  $X$ . Principal components ( $W = XZ$ ) are linear combinations of the independent variables such that

- the principal components  $w_i$  are uncorrelated. In other words, they are all pairwise orthogonal.
- the first principal component has the largest variance of any linear function of the original variables (subject to a scale constant). The second component has the second largest variance, and so on.

Principal components are obtained by computing the eigenvalues and eigenvectors of  $X'X$ . The eigenvalues are the variances of the components, and the eigenvectors are the coefficients,  $z_{ii}$ , of the linear equations that relate the principal components to the original variables, such that  $W = XZ$ .

Principal components with very small variances (which correspond to large condition indices) are of interest in identifying sources of multicollinearity.

## Multicollinearity Diagnostics Guidelines

- Correlation statistics
  - close to 1 or -1 indicate a high degree of linear relationship
  - close to 0 indicate no clear linear relationship.
- Variance inflation factors
  - greater than 10 indicate the presence of strong collinearity.

52

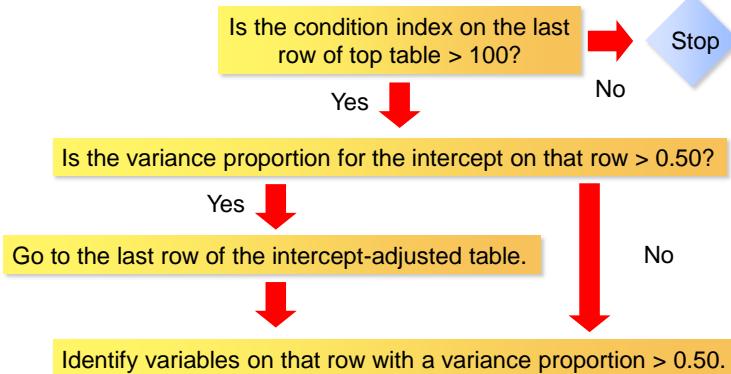
*continued...*

## Multicollinearity Diagnostics Guidelines

- Condition index values
  - between 10 and 30 suggest weak dependencies
  - between 30 and 100 indicate moderate dependencies
  - greater than 100 indicate strong collinearity.
- Proportion of variation explained by principal components
  - is greater than 0.5 for a large condition index.

53

## Using COLLIN and COLLINOINT Statistics



54

## Using COLLIN and COLLINOINT Statistics

- Multicollinearity is evident when one or more principal components have a large condition index.
- Predictors having large proportions of variation explained by high-index components are likely to be involved in the collinearity relationship.
- Collinearity between the intercept and continuous variables might be eliminated by centering.

55



## Multicollinearity Diagnostics

---

Use PROC CORR to compute the correlation statistics among the independent variables and to generate a matrix of scatter plots. Request the variance inflation factor (VIF) and condition index values using the VIF, COLLIN, and COLLINOINT options in the MODEL statement of PROC REG.

```
title 'Collinearity Diagnosis for the Cubic Model';
proc corr data=d_paper nosimple plots=matrix;
  var & _GLSMOD;
run;

proc reg data=d_paper;
  model strength=& _GLSMOD / vif collin collinoint;
run;
quit;
title;                                *ST201d05.sas;
```

Selected PROC CORR statement options:

**NOSIMPLE** suppresses printing simple descriptive statistics for each variable. However, if you request an output data set, the output data set still contains simple descriptive statistics for the variables.

**PLOTS** requests statistical graphics via the Output Delivery System (ODS). The **MATRIX** option requests a scatter plot matrix for variables. That is, the procedure displays a symmetric matrix plot with variables in the VAR list if a **WITH** statement is not specified. Otherwise, the procedure displays a rectangular matrix plot with the **WITH** variables appearing down the side and the VAR variables appearing across the top.

Selected MODEL statement options in PROC REG:

**VIF** produces variance inflation factors with the parameter estimates.

**COLLIN** requests a detailed analysis of collinearity among the regressors.

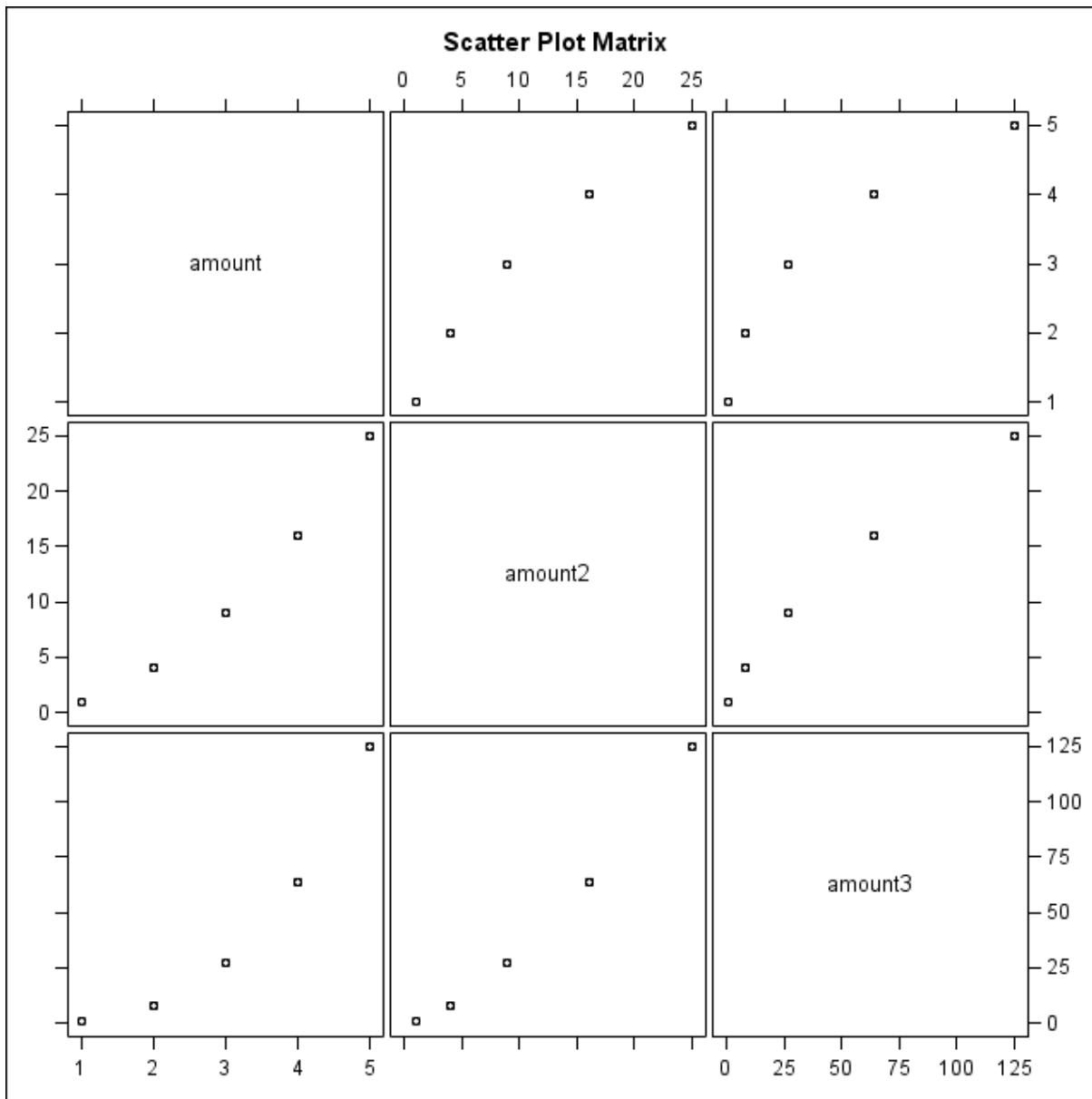
**COLLINOINT** requests the same analysis as the **COLLIN** option with the intercept variable adjusted out rather than included in the diagnostics.

## PROC CORR Output

Collinearity Diagnosis for the Cubic Model			
The CORR Procedure			
3 Variables:	Amount	Amount_2	Amount_3
Pearson Correlation Coefficients, N = 22 Prob >  r  under H0: Rho=0			
	Amount	Amount_2	Amount_3
Amount	1.00000	0.98278	0.94869
Amount		<.0001	<.0001
Amount_2	0.98278	1.00000	0.99036
Amount^2	<.0001		<.0001
Amount_3	0.94869	0.99036	1.00000
Amount^3	<.0001	<.0001	

It is obvious that very high correlations exist between **Amount** and **Amount^2** ( $r=0.98$ ), **Amount** and **Amount^3** ( $r=0.95$ ), and **Amount^2** and **Amount^3** ( $r=0.99$ ).

## PROC CORR ODS Graphics Output



The scatter plots created by PROC CORR illustrate the curvilinear relationships among **Amount**, **Amount<sup>2</sup>**, and **Amount<sup>3</sup>**.

## Partial PROC REG Output

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	Intercept	1	2.73280	0.26060	10.49	<.0001	0
Amount	Amount	1	-0.36900	0.32208	-1.15	0.2669	393.09634
Amount_2	Amount^2	1	0.22339	0.11651	1.92	0.0712	2048.80921
Amount_3	Amount^3	1	-0.02862	0.01270	-2.25	0.0369	699.53428

Collinearity Diagnostics							
Number	Eigenvalue	Condition Index	Proportion of Variation				
			Intercept	Amount	Amount_2	Amount_3	
1	3.68081	1.00000	0.00043562	0.00002910	0.00001194	0.00004451	
2	0.30838	3.45487	0.01447	0.00001443	0.00005095	0.00060756	
3	0.01072	18.53340	0.10613	0.01759	0.00072341	0.02002	
4	0.00009941	192.42582	0.87897	0.98237	0.99921	0.97932	

Collinearity Diagnostics (intercept adjusted)						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Amount	Amount_2	Amount_3	
1	2.94800	1.00000	0.00028442	0.00005614	0.00016067	
2	0.05168	7.55288	0.02697	0.00004444	0.01238	
3	0.00032033	95.93188	0.97274	0.99990	0.98746	

The variance inflation factors (VIF) are quite large for all three independent variables.

The COLLIN option in the MODEL statement requests that a collinearity analysis be performed. This includes eigenvalues, condition indices, and decomposition of the variances of the estimates with respect to each eigenvalue. The COLLINOINT option requests the same analysis as the COLLIN option with the intercept variable adjusted out rather than included in the diagnostics. Notice that the analysis in PROC REG is reported with eigenvalues of  $X'X$  rather than singular values of  $X$ . The eigenvalues of  $X'X$  are the squares of the singular values of  $X$ .

The condition indices are the square roots of the ratio of the largest eigenvalue to each individual eigenvalue. The largest condition index is the condition number of the scaled  $X$  matrix. The output shows a condition number of 192, which indicates a strong collinearity. For each variable, PROC REG produces the proportion of the variance of the estimate accounted for by each principal component. Variance proportions greater than 0.50 associated with a large condition index identify the subsets of the predictor variables that are collinear. Variables **Amount**, **Amount^2**, and **Amount^3** are all involved in the multicollinearity.

## Dealing with Multicollinearity

- Exclude redundant independent variables.
- Redefine variables.
- Use biased regression techniques such as ridge regression or principal component regression.
- Center the independent variables in polynomial regression models.

57

If you can identify redundancy among independent variables in the model, you might be able to remove one or more of the highly correlated independent variables to lessen the multicollinearity. However, you must be careful not to remove an independent variable that is important when you describe and predict the dependent variable. You can also redefine some of the variables as some related quantities. For example, in a study to examine the manpower needed to operate a U.S. Navy Bachelor Officers' Quarters (BOQ) (Myers 1990), the variables (such as monthly man-hours, average daily occupancy, monthly average number of check-ins, and square feet of common use area) seem to be collinear with the variable number of rooms. You can refine these variables to measure per-room characteristics. That is, estimate the relationship of the man-hour requirement per room to the occupancy rate per room, and so on.

In the presence of multicollinearity, the minimum variance of the unbiased parameter estimates might be unacceptably large. Biased regression techniques provide biased parameter estimates but smaller standard errors, compared with ordinary least squares estimates and standard errors. You use PROC REG with the RIDGE= option in the MODEL statement to perform the ridge regression analysis. You use the PCOMIT= option in the MODEL statement in PROC REG to perform principal component regression analysis.

For polynomial regression models, the effective and easy way of reducing multicollinearity is to center the variables that are involved in higher-ordered terms (Marquardt, D.W. 1980). This centering makes the independent variables orthogonal to the intercept column and removes any collinearity with the intercept. An alternative to centering the independent variables before creating the higher-order terms is to use orthogonal polynomials (Rawlings, Pantula, and Dickey 1998).

## Center Variables

- Use the STDIZE procedure with the METHOD=mean option.
- Use SAS DATA steps to subtract the means from the variables.
- Use the STANDARDIZE option within the EFFECT statement in PROC GLMSELECT.



## Centering Variables

---

Use PROC GLMSELECT to create a new centered cubic polynomial effect.

```
proc glmselect data=STAT2.paper outdesign=dc_paper;
  effect qc_amount=polynomial(amount /
    degree=3 standardize(method=moments)=center) ;
  model strength = qc_amount / selection=none;
  title "Paper Data Set: Centered Cubic Model";
run;                                              *ST201d06.sas;
```

Selected EFFECT statement option:

STANDARDIZE specifies that the variables that define the polynomial be standardized. By default, the standardized variables receive prefix “s\_” in the variable names. The CENTER option specifies that variables be centered, but not scaled.

Selected STANDARDIZE suboption:

METHOD= specifies that the center be estimated by the variable mean and the scale be estimated by the standard deviation. Only observations that are used in performing the analysis are used for the standardization.

As an alternative, use PROC SQL to compute the mean value for **amount** and create a macro variable **&mamount** to store this mean value. Finally, use a DATA step to center the variable **amount** and compute the quadratic and cubic terms for the centered variable **mcamount**. This method is useful when the mean value used to center the data is needed for subsequent analyses (for example, obtaining predictions for new data using a model fit to using the centered variables).

```
proc sql;
  select mean(amount) into: mamount
  from STAT2.paper;
run;

data paper2;
  set STAT2.paper;
  mcamount=amount-&mamount;
  mcamount2 = mcamount**2;
  mcamount3 = mcamount**3;
run;                                              *ST201d06.sas;
```

SQL Query Results

---

3.181818

The average value for **amount** is 3.181818. This value was subtracted from the variable **amount** to create a centered variable **mcamount** and the associated higher-ordered terms.

## PROC GLMSELECT Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.52986	0.17662	16.22	<.0001
Error	18	0.19605	0.01089		
Corrected Total	21	0.72591			

The Analysis of Variance table is exactly the same as the table for the model with the original uncentered variables. This is because you changed only the location of the data values by centering. You did not change the variation of the data values.

Use PROC REG to check the multicollinearity diagnostics for the centered model.

```
ods select ParameterEstimates CollinDiag CollinDiagNoInt;
proc reg data=dc_paper;
    model strength = &_GLSMOD / vif collin collinoint;
    title 'Diagnostics for Centered Cubic Model';
run;
title;
quit; *ST201d06.sas;
```

Selected ODS statement:

SELECT      enables the user to choose specific components of a procedure's output to be displayed. Each table and graphic produced by a procedure have an associated ODS table or graph name. A full listing of these can be found in the SAS documentation for each procedure.

## PROC REG Output

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	Intercept	1	2.89841	0.03495	82.93	<.0001	0
s_Amount	s_Amount	1	0.18335	0.04372	4.19	0.0005	7.24339
s_Amount_2	s_Amount^2	1	-0.04979	0.01500	-3.32	0.0038	1.18088
s_Amount_3	s_Amount^3	1	-0.02862	0.01270	-2.25	0.0369	7.66067

Collinearity Diagnostics							
Number	Eigenvalue	Condition Index	Proportion of Variation				
			Intercept	s_Amount	s_Amount_2	s_Amount_3	
1	2.04973	1.00000	0.02343	0.02199	0.03190	0.02368	
2	1.64309	1.11691	0.09627	0.01486	0.06824	0.00960	
3	0.24015	2.92150	0.80943	0.02321	0.70431	0.00537	
4	0.06702	5.53008	0.07087	0.93994	0.19554	0.96135	

Collinearity Diagnostics (intercept adjusted)					
Number	Eigenvalue	Condition Index	Proportion of Variation		
			s_Amount	s_Amount_2	s_Amount_3
1	2.00352	1.00000	0.03095	0.03215	0.03094
2	0.92755	1.46969	0.01078	0.83355	0.00205
3	0.06893	5.39133	0.95827	0.13431	0.96701

The variance inflation factors are greatly reduced by using the centered variable. The largest condition index is also reduced to less than 10. In addition, the standard errors of the parameter estimates are also much smaller than the original model. These are all indications that this model is more stable. Also, all of the variables in the model are now significant at the 0.05 alpha level.

When you interpret the parameter estimates, remember that they are for the centered variables, not the original variables.

- ✍ In regression models, you can only validate the model within the range of the data. Predictions based on independent variable values outside of the range of the data should be done with care because the relationship might be different outside of this range.

## 1.09 Quiz

Is the statement below true or false? Explain why.

In general, you can remove the nonsignificant terms from a model all at once to obtain a parsimonious model.

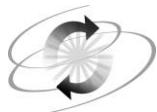
61

## 1.10 Poll

When you fit a straight line to data that shows a curvilinear relationship, it is very likely that the residual plot shows a curvilinear pattern.

- True
- False

63



## Exercises

---

### 1. Fitting a Simple Polynomial Regression Model (Continued)

An analyst for a cafeteria chain wanted to investigate whether the sales of coffee are related to the number of self-service coffee dispensers in a cafeteria line. Fourteen cafeterias that are similar in terms of volume of business, type of clientele, and location were chosen for the study. The number of self-service dispensers was assigned randomly at each cafeteria and the sales in hundreds of gallons of coffee were recorded.

Data are stored in the **STAT2.cafeteria** data set. These are the variables in the data set:

**cafeteria** cafeteria identification number

**dispensers** number of dispensers at each cafeteria

**sales** coffee sales in hundreds of gallons

 Use the data set that you created in Exercise **1.d** that contains **dispensers** and **dispensers squared**.

- f. Use PROC CORR to compute the Pearson correlation coefficient between **dispensers** and **dispensers squared**. Use the NOSIMPLE and PLOTS( )=MATRIX options in the PROC CORR statement. Examine the tabular and graphical output. What do you conclude?
- g. Use the appropriate options in the MODEL statement in PROC REG to compute the variance inflation factor (VIF) and the collinearity diagnostics statistics. Is there collinearity among the independent variables? If so, which ones?
- h. Use PROC GLMSELECT with an EFFECT statement to create a new, centered, quadratic effect for **Dispensers** and fit a model with the centered polynomial terms. Use the OUTDESIGN option to create a new model design data set. Using the design data set, obtain collinearity diagnostics from PROC REG. Does the centered model appear to have multicollinearity among the independent variables?

### 1.11 Quiz

Is the statement below true or false? Explain why.

In the previous exercise, after you centered the variable **dispenser** and refit the polynomial model using the centered variables, both the ANOVA table and the parameter estimate table changed.

66

## 1.4 Modeling Nonlinear Relationships

---

### Objectives

- Perform an initial exploratory data analysis.
- Conduct polynomial regression with more than one independent variable.
- Evaluate polynomial models with model selection criteria.
- Compare and evaluate potential models.
- Introduce splines for modeling nonlinear relationships.

69

## Car Example



70

Consider data that were collected to examine the prices of cars. The SAS data set **STAT2.cars** contains information about a sample of 1993 model cars. Eighty-one 4-, 5-, and 6-passenger models were selected from the *1993 Cars Annual Auto Issue* published by *Consumer Reports* and from *Pace New Car and Truck 1993 Buying Guide*. Although newer data sets are available, this data set has some unique characteristics that make it an excellent set for regression analysis.

You are interested in examining the relationships between the price of the vehicles and various other car attributes. These are the variables in the data set:

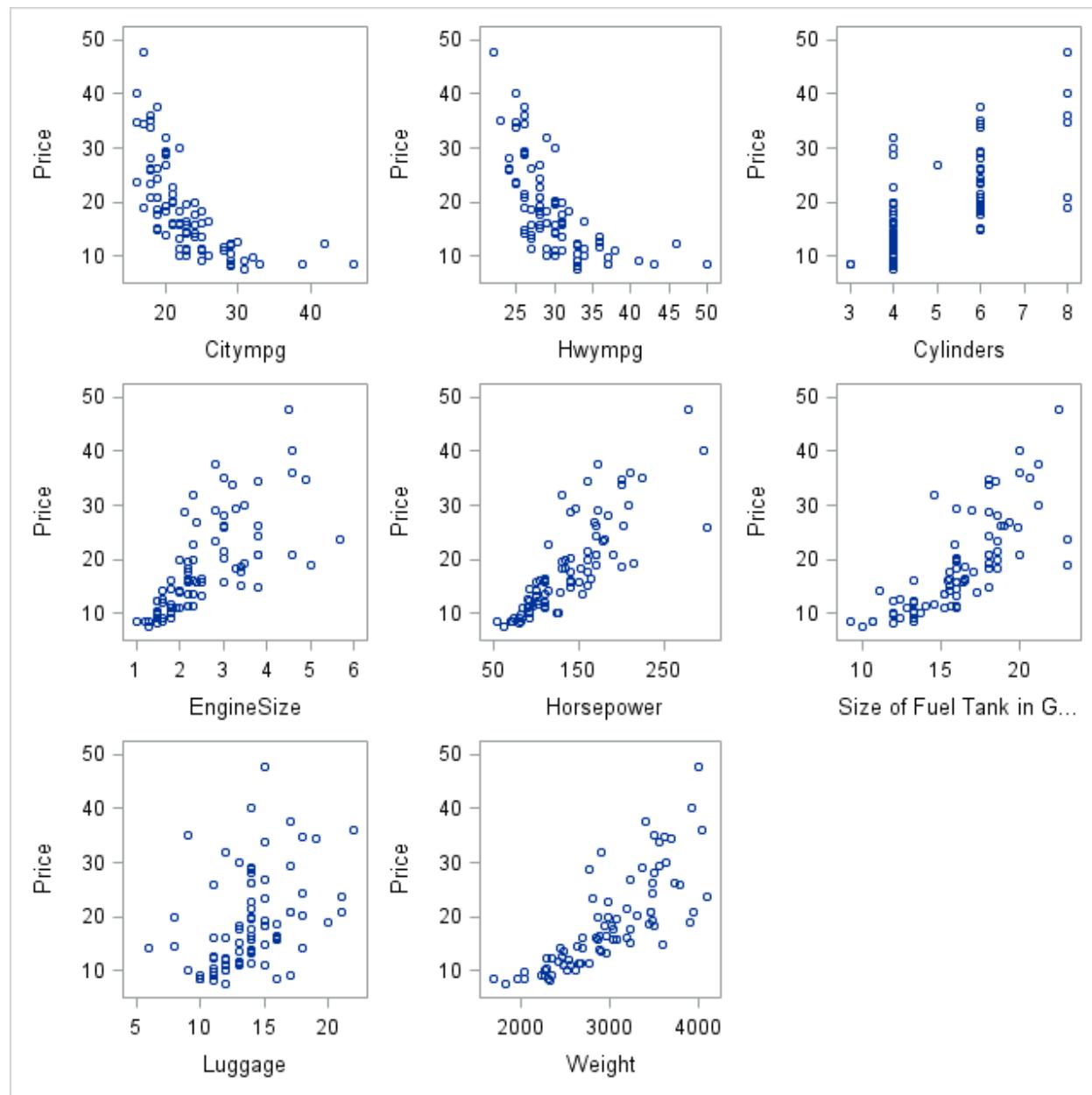
<b>Manufacturer</b>	name of the manufacturer
<b>Model</b>	name of the model
<b>Type</b>	type of vehicle ( <i>Compact, Large, Midsize, Small, or Sporty</i> )
<b>Price</b>	average of the maximum and minimum price
<b>Citympg</b>	average city miles per gallon (EPA rating)
<b>Hwympg</b>	average highway miles per gallon (EPA rating)
<b>Cylinders</b>	number of cylinders
<b>EngineSize</b>	engine displacement size (in liters)
<b>Horsepower</b>	maximum horsepower
<b>FuelTank</b>	fuel tank capacity (in gallons)
<b>Passengers</b>	passenger capacity
<b>Luggage</b>	luggage capacity (in cubic feet)
<b>Weight</b>	weight of the vehicle (in pounds)
<b>Origin</b>	origin of the vehicle ( <i>US or non-US</i> )



## Initial Data Exploration

Examine the relationship between **Price** and the other continuous variables in the data set by generating plots of **Price** versus the other variables.

```
proc sgscatter data=STAT2.cars;
  plot price*(citympg hwympg cylinders enginesize horsepower fueltank
    luggage weight);
run; *ST201d07.sas
```



Both **Citympg** and **Hwympg** appear to have a negative relationship with **Price**, but the relationships do not appear to be linear. A quadratic term might be appropriate for these two variables. **Cylinders** might not be a true continuous variable because it only takes on a few values. However, the values are ordinal in nature, so it could be used as a numeric variable in a regression. Each of the variables (**EngineSize**, **Horsepower**, and **FuelTank**) appears to have a positive relationship with **Price**. The relationship between **FuelTank** and **Price** might be curvilinear, and there does not appear to be a relationship between **Luggage** and **Price**. **Weight** seems to have a positive relationship with **Price** and curvature might be present. The patterns in several of the plots might point toward nonconstant variance.

For the variables where curvature is seen, it might be useful to add a smooth curve to the plots. One way to do this is by fitting a penalized B-spline curve by adding the PBSPLINE option to the PLOT statement in PROC SGSCATTER. In addition, you can use the IMAGEMAP=ON option of the ODS GRAPHICS statement to turn on data-tips generation for HTML output.

```
ods graphics / imagemap=on;

ods html style=statistical;
proc sgscatter data=STAT2.cars;
  plot price*(citympg hwympg fueltank weight) / pbspline;
run;
```

Selected ODS GRAPHICS options:

ON|OFF                 enables and disables ODS to create graphics. In SAS 9.3 and later, ODS Graphics is enabled by default. The ODS GRAPHICS OFF statement disables ODS Graphics. The ODS GRAPHICS ON or ODS GRAPHICS statement enables ODS Graphics again.

IMAGEMAP=ON|OFF controls data-tips generation. Data tips are pieces of explanatory text that appear when you move your mouse pointer the data portions of a graph contained in an HTML page.

To generate output from SAS, a valid ODS destination must be open. By default, in SAS 9.3 and later, the HTML destination is open. You can use an ODS destination statement, such as ODS LISTING or ODS RTF, to open a different destination. You can also specify options, such as the filename or the path to an output directory, in the ODS destination statement.

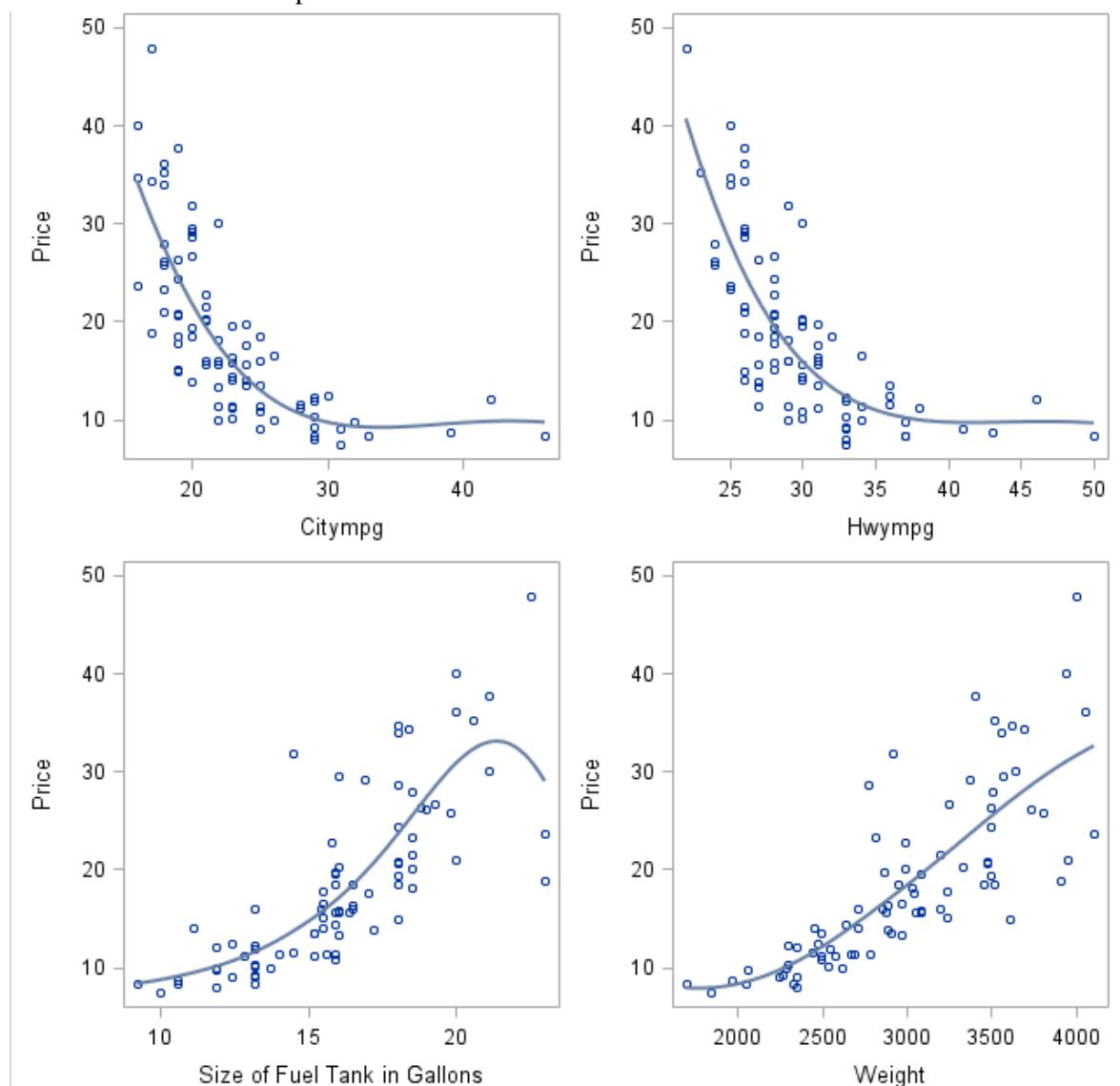
ODS *destination* statement option:

STYLE=    specifies a style to be used for the output. Each ODS destination has a default style for the formatting of output. The style specifies a collection of visual attributes that are used for the rendering of the output. The STYLE= option is valid for all ODS destinations except the Document destination and the Output destination. HTMLBLUE is the style developed for the default HTML destination. For full color output, the recommended styles include DEFAULT, ANALYSIS, STATISTICAL, and LISTING. For black and white output, JOURNAL is recommended. JOURNAL2 is recommended for gray scale. For complete documentation about the STYLE= option, see the ODS statements in *SAS® Output Delivery System: User's Guide*. For more information about using the STYLE= option with SAS/GPGRAPH output, see the *SAS®9 Online Documentation*.

Selected PLOT statement option:

PBSPLINE adds a fitted, penalized B-spline curve to the scatter plot.

## PROC SGSCATTER Output



With the curves added to the plots, you can see that **Citympg** and **Hwympg** seem to exhibit a quadratic relationship with **Price**, so quadratic terms for these two variables are added to the model. The curve in the scatter plot of **Price** versus **FuelTank** shows a positive curvilinear relationship up to the values of **FuelTank** at approximately 20 gallons. Then the curve turns downward. This might be due to the two lower points in the graph. Because tooltips are turned on in the graphical output, you can see that these two points have a value of 23 for **FuelTank**. The top data point has a **Price** of 23.7 and the lower data point has a **Price** of 18.8. (These two points might be influential and are examined later in the course.) For a polynomial model, it might be appropriate to start with cubic and quadratic terms for **FuelTank** in the model.

Evidence of increasing variability might be present in the plot of **Price** versus **Weight**. The apparent curvature for **Weight** might reflect either increasing variability in the data or a curvilinear relationship. You might choose to include both the linear and quadratic terms of **Weight**, or because the curvature is slight, only include the linear term. Therefore, proceed with your regression and check the assumption of homogeneity of variances during your model diagnostics. (Model diagnostics are discussed in further detail in a later chapter.)

The next step in data exploration is to generate the correlations between the variables using PROC CORR. In addition to the correlations, you can examine a matrix of scatter plots that show the relationships between the independent predictor variables.

```
proc corr data=STAT2.cars nosimple;
  var price citympg hwympg cylinders enginesize horsepower fueltank
    luggage weight;
run; *ST201d08.sas;
```

### PROC CORR Output

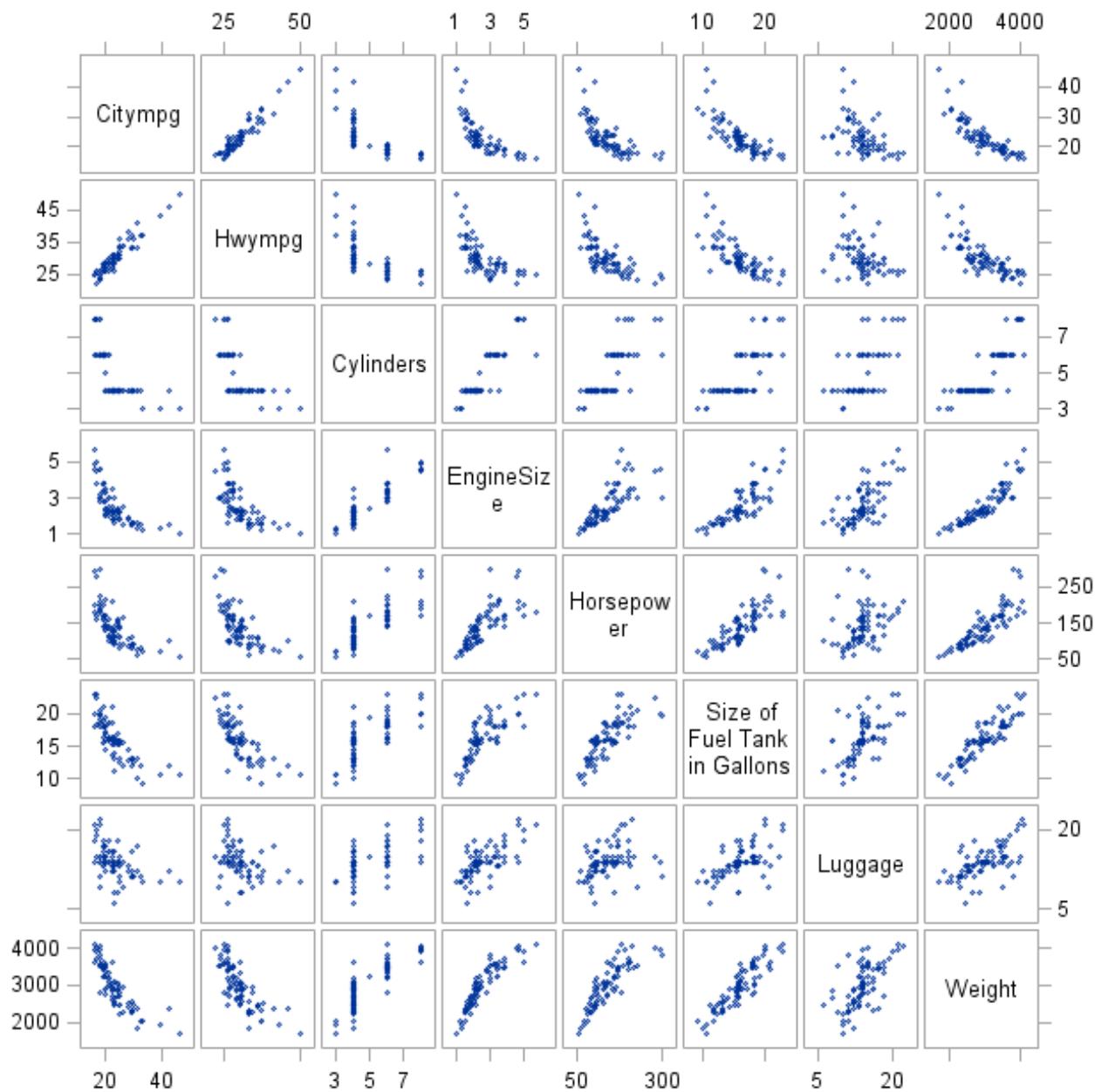
The CORR Procedure										
9 Variables:	Price	Citympg	Hwympg	Cylinders	EngineSize	Horsepower	FuelTank	Luggage	Weight	
<b>Pearson Correlation Coefficients, N = 81</b>										
<i>Prob &gt;  r  under H0: Rho=0</i>										
<b>Price</b>	1.00000	-0.66596 <.0001	-0.65672 <.0001	0.71848 <.0001	0.69586 <.0001	0.81667 <.0001	0.76059 <.0001	0.39548 0.0003	0.78695 <.0001	
<b>Citympg</b>	-0.66596 <.0001	1.00000	0.94518 <.0001	-0.67818 <.0001	-0.74586 <.0001	-0.70294 <.0001	-0.79558 <.0001	-0.49358 <.0001	-0.83455 <.0001	
<b>Hwympg</b>	-0.65672 <.0001	0.94518 <.0001	1.00000	-0.63872 <.0001	-0.66457 <.0001	-0.68257 <.0001	-0.73645 <.0001	-0.36963 0.0007	-0.77411 <.0001	
<b>Cylinders</b>	0.71848 <.0001	-0.67818 <.0001	-0.63872 <.0001	1.00000	0.88770 <.0001	0.78812 <.0001	0.72227 <.0001	0.58568 <.0001	0.83014 <.0001	
<b>EngineSize</b>	0.69586 <.0001	-0.74586 <.0001	-0.66457 <.0001	0.88770 <.0001	1.00000	0.77345 <.0001	0.82035 <.0001	0.68017 <.0001	0.91680 <.0001	
<b>Horsepower</b>	0.81667 <.0001	-0.70294 <.0001	-0.68257 <.0001	0.78812 <.0001	0.77345 <.0001	1.00000	0.79511 <.0001	0.35766 0.0010	0.85875 <.0001	
<b>FuelTank</b> Size of Fuel Tank in Gallons	0.76059 <.0001	-0.79558 <.0001	-0.73645 <.0001	0.72227 <.0001	0.82035 <.0001	0.79511 <.0001	1.00000	0.61270 <.0001	0.89932 <.0001	
<b>Luggage</b>	0.39548 0.0003	-0.49358 <.0001	-0.36963 0.0007	0.58568 <.0001	0.68017 <.0001	0.35766 0.0010	0.61270 <.0001	1.00000	0.63697 <.0001	
<b>Weight</b>	0.78695 <.0001	-0.83455 <.0001	-0.77411 <.0001	0.83014 <.0001	0.91680 <.0001	0.85875 <.0001	0.89932 <.0001	0.63697 <.0001	1.00000	

**Horsepower** has the strongest correlation with **Price** (0.81667), so you conclude that this variable would be one of the best variables to include in a regression model. However, recall that the Pearson correlation statistic measures the linear relationship between variables. **Citympg**, **Hwympg**, and **FuelTank** appeared to have a relationship with **Price** that was not linear. These relationships are missed by a correlation analysis. It might be that these variables are better predictors of **Price** if the nature of their relationships is considered.

Also many of the variables that are potentially independent variables are highly correlated with one another. Correlation of the independent variables can cause model instability. This should be taken into consideration when you develop the model. The scatter plots of the independent variables might help you visualize these relationships.

```
proc sgscatter data=STAT2.cars;
  matrix citympg hwympg cylinders enginesize horsepower fueltank
    luggage weight;
run; *ST201d08.sas;
```

## PROC SGSCATTER Output



Most of the independent variables are related one another, so the majority of the plots indicate strong linear or curvilinear patterns. (For example, look at the strong linear relationship that is evident between **Citympg** and **Hwympg**, or the strong curvilinear relationship among **Citympg** and **EngineSize**, **Horsepower**, **FuelTank**, and **Weight**.) Because most of the plots indicate strong relationships, it is easier to look for plots with the weakest indicated patterns. The series of plots that seem to indicate the weakest relationships are those for **Luggage** versus the other predictor variables. Even some of those plots, such as the plots of **Luggage** versus **FuelTank** and the plots of **Luggage** versus **Weight** and **EngineSize**, indicate some relationship might be present. Notice that many of the relationships between the independent variables are curvilinear.

## 1.12 Multiple Choice Poll

Which of the following statements regarding the EDA is *false*?

- There is a curvilinear relationship between **Price** and **Hwympg**, **Citympg**, and **FuelTank**.
- Price** and **Hwympg** are negatively correlated.
- Luggage** seems to be an important independent variable.

72

## Model Selection Statistics

- Significance level for individual variables
- Coefficient of determination ( $R^2$ )
- Adjusted coefficient of determination (adjusted  $R^2$ )
- Mallows'  $C_p$  statistic
- Akaike's information criteria (AIC, AICC)
- Schwarz's Bayesian criteria (SBC)

74

Many different statistics can be used in evaluating competing models to determine which model is most appropriate. Most of these criteria are based on the presumption that you want to create a model that minimizes the unexplained variability (the mean square error) with the fewest number of variables that are possible (the principle of parsimony).

## R<sup>2</sup> and Adjusted R<sup>2</sup>

$$R^2 = 1 - \frac{SSE}{SST}$$

$$\begin{aligned} AdjR^2 &= 1 - \frac{SSE / df_E}{SST / df_T} = 1 - \frac{SSE / (n - p)}{SST / (n - 1)} \\ &= 1 - \frac{(n - 1)}{(n - p)} (1 - R^2) \end{aligned}$$

75

In the formulas above, the following are defined:

SSE      is sum of the squared residuals.

SST      is total sum of squares corrected for the mean.

$df_E$       is degrees of freedom associated with sum of squared residuals.

$df_T$       is degrees of freedom associated with sum of squared total.

$n$       is the total number of observations.

$p$       is the total number of parameters in the model, including the intercept.

The coefficient of determination (R square) is a measure of the proportion of variability in the response variable explained by the predictor variables. R square never decreases when you include more terms in the model. Therefore, it is not necessarily a good measure to use to compare models with a different number of predictor variables.

The adjusted R square is similar to R square, but it is adjusted for the number of terms in the model. Therefore, when you compare models with a different number of predictor variables, it is more appropriate to use the adjusted R square.

## Mallows' $C_p$ Statistic

$$C_p = p + \frac{(MSE_p - MSE_{full})(n - p)}{MSE_{full}}$$

- $p$  is the number of parameters in the model being evaluated, including the intercept.
- $n$  is the total number of observations.
- Models with  $C_p > p$  are underspecified.
- Mallows recommends choosing the first model where  $C_p \leq p$ .

76

In the formula above, the following are defined

- $p$  is the number of parameters in the model being evaluated; the number of variables plus one.
- $MSE_p$  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.

Mallows'  $C_p$  is a simple indicator of model misspecification. When  $C_p$  is much larger than  $p$ , it usually indicates model underspecification. Refer to Mallows, C.L. (1973) for more details.

### Details

Another representation of the computational formula for  $C_p$  is  $C_p = \frac{SS(\text{Residual})}{MS(\text{Residual})} + 2p - n$ , where

$SS(\text{Residual})$  is the residual sum of squares for the model with  $p-1$  variables and  $MS(\text{Residual})$  is the residual mean square when using all the independent variables.

When the model is correctly specified the residual sum of squares is an unbiased estimate of  $(n - p)\sigma^2$ , and  $C_p$  is an unbiased estimate of  $\frac{(n - p)\sigma^2}{\sigma^2} + 2p - n = p$ . So  $C_p$  is approximately equal to  $p$  when the model is correctly specified. When important variables are omitted from the model, the residual sum of squares is increased by the amount of variability that can be explained by those terms if they were included in the model. Therefore,  $C_p$  increases and  $C_p > p$ . (Rawlings, Pantula, and Dickey 1998)

## Information Criteria

- Akaike's information criterion (AIC)

$$AIC = (n) \ln\left(\frac{SSE}{n}\right) + 2p + n + 2$$

- Finite-sample corrected AIC (AICC)

$$AICC = (n) \ln\left(\frac{SSE}{n}\right) + \frac{n(n+p)}{n-p-2}$$

- Schwarz's Bayesian criterion (SBC)

$$SBC = (n) \ln\left(\frac{SSE}{n}\right) + p \ln(n)$$

Smaller values indicate a better model.

77

In the formulas above, the following are defined

$n$  is the number of observations.

$p$  is the number of parameters in the model, including the intercept.

$SSE$  is sum of the squared residuals.

These criteria are designed to assess the precision of fit of the model against the number of parameters in the model. For model selection, the model with the smallest information criterion indicates a better model. The AIC tends to select models with a larger number of parameters. To overcome this tendency, the SBC has a larger penalty for additional parameters in the model.



PROC GLMSELECT uses the definitions of AIC and AICC described in Hurvich and Tsai (1989). PROC REG uses an earlier definition of AIC (Akaike 1969 and Judge 1980).

## Select Candidate Models

Candidate models can be identified by using

- your subject-matter knowledge
- information gathered from data exploration
- automatic selection methods available in the GLMSELECT procedure:
  - forward, backward, or stepwise selection
  - least angle regression (LAR)
  - LASSO and adaptive LASSO
  - elastic net selection
- residual plots to evaluate model fit and model assumptions.

78

There are several automatic selection criteria available in PROC GLMSELECT. These include the following:

- Sequential selection techniques including forward, backward, stepwise. At each step, candidate variables are evaluated for inclusion in or exclusion from the model. They use the specified model selection statistics described previously. When significance levels are used, the selection is controlled with the SLENTRY= and SLSTAY= options in the MODEL statement. The default significance levels are as follows:

forward	SLENTRY=.50
backward	SLSTAY=.10
stepwise	SLENTRY=.15 and SLSTAY=.15

- Least angle regression, or LAR (Efron, 2006).
- The LASSO method (Tibshirani, 1996) and adaptive LASSO.
- Elastic net selection (Zhou, 2005).

 Details about the LAR, LASSO, and elastic net selection methods can be found in the SAS®9 documentation.

Additional model selection methods are available in PROC REG. These include all-possible model selection using R square and adjusted R square.

Residual plots can be used to evaluate model assumptions and model fit. A normal quantile plot of the residuals can help you determine whether they appear to be normally distributed. A plot of the residuals versus the predicted values can help you determine whether you have the correct model form and whether the variances appear to be equal. The plot should be a random scatter of points about a zero reference line. These plots are available from PROC REG in the ODS Graphics output.



## Select Candidate Models

Recall that there are three variables, **Citympg**, **Hwympg**, and **FuelTank**. They appear to have a polynomial relationship with **Price**. (**Weight** is not included because the potential curvilinear relationship between **Price** and **Weight** is most likely due to increasing variability in the data.) To reduce collinearity problems with the models, it might be wise to center these variables when polynomial effects are created for modeling.

You can use PROC GLMSELECT to identify candidate models. It is often helpful to use more than one of the selection options in attempting to identify the models. For example, you might choose to do forward and backward regression using multiple model selection statistics.

```
ods graphics on / reset=all;

title 'Model Selection Cars2 Data Set';

%macro p_eff;
effect p_city = polynomial(citympg /degree=2
    standardize(method=moments)=center);
effect p_hwy = polynomial(hwympg /degree=2
    standardize(method=moments)=center);
effect p_fuel = polynomial(fueltank /degree=3
    standardize(method=moments)=center);
%mend;

proc glmselect data=STAT2.cars2 plots=criteria;
    title2 'Backward elimination with significance levels';
    %p_eff;
    model price = p_city p_hwy cylinders enginesize horsepower p_fuel
        luggage weight / selection=backward select=sl
        slstay=0.05 hierarchy = single;
run;

proc glmselect data=STAT2.cars2;
    title2 'Forward selection with significance levels';
    %p_eff;
    model price = p_city p_hwy cylinders enginesize horsepower p_fuel
        luggage weight / selection=forward select=sl
        slentry=0.1 hierarchy=single;
run;

proc glmselect data=STAT2.cars2;
    title2 'Backward elimination using SBC';
    %p_eff;
    model price = p_city p_hwy cylinders enginesize horsepower p_fuel
        luggage weight / selection=backward select=sbc
        hierarchy=single;
run;
```

```

proc glmselect data=STAT2.cars2 plots=criteria;
  title2 'Backward elimination using adjusted R-square';
  %p_eff;
  model price = p_city p_hwy cylinders enginesize horsepower p_fuel
    luggage weight / selection=backward select=adjrsq
    hierarchy=single;
run;
ods graphics off;                                *ST201d09.sas;

```

Selected ODS LISTING statement option:

**RESET** resets one or more ODS GRAPHICS options to their default settings. The RESET and RESET=ALL options are equivalent. To reset more than one option, but not all of the options, the RESET= must be specified separately in order for each option to be reset.

Selected PLOTS= options:

**ONLY** suppresses the default plots. Only specifically requested plots are displayed.

**CRITERIA | CRITERIONPANEL** (*criteria-options*) produces a panel of fit criteria for the models that are examined when you request variable selection with the SELECTION= option in the MODEL statement. The displayed fit criteria are adjusted R square, Akaike's information criterion (AIC), finite-sample corrected AIC (AICC), Schwarz's Bayesian information criterion (SBC), and any additional criteria that are specified in the CHOOSE=, SELECT=, STOP=, or STATS= options.

The following *criteria-options* are available:

**STEPAXIS** specifies the horizontal axis to be used on the plots. Choices include the step number (NUMBER) or the effect entering or leaving the model at each step (EFFECT). The default is STEPAXIS=EFFECT.

**UNPACK** suppresses paneling. Separate plots are produced for each of the six fit statistics.

## Partial PROC GLMSELECT Output for the Backward Elimination Model Using Significance Levels

Effects: Intercept s_Hwympg s_Hwympg^2 Horsepower				
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	4448.69341	1482.89780	65.73
Error	77	1737.20536	22.56111	
Corrected Total	80	6185.89877		

Root MSE	4.74985
Dependent Mean	18.64321
R-Square	0.7192
Adj R-Sq	0.7082
AIC	339.31229
AICC	340.11229
SBC	265.89009

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	4.039486	2.170242	1.86
s_Hwympg	1	-0.804069	0.213779	-3.76
s_Hwympg^2	1	0.043504	0.014299	3.04
Horsepower	1	0.097301	0.016142	6.03

Backward elimination results in a model with three variables, **Hwympg**, **Hwympg<sup>2</sup>**, and **Horsepower**. The model has an adjusted R square of 0.7082 and an SBC of 265.89. Because the analysis specified HIERARCHY=SINGLE, the final model is hierarchically well formulated.

The summary of the backward selection process shows the effect removed at each step.

Backward Selection Summary				
Step	Effect Removed	Number Effects In	F Value	Pr > F
0		13		
1	Luggage	12	0.02	0.8947
2	s_Citympg^2	11	0.06	0.8103
3	s_FuelTank^3	10	0.82	0.3686
4	Cylinders	9	0.97	0.3291
5	s_Citympg	8	1.18	0.2810
6	s_FuelTank^2	7	0.51	0.4774
7	EngineSize	6	0.89	0.3492
8	Weight	5	0.33	0.5658
9	s_FuelTank	4	3.60	0.0614

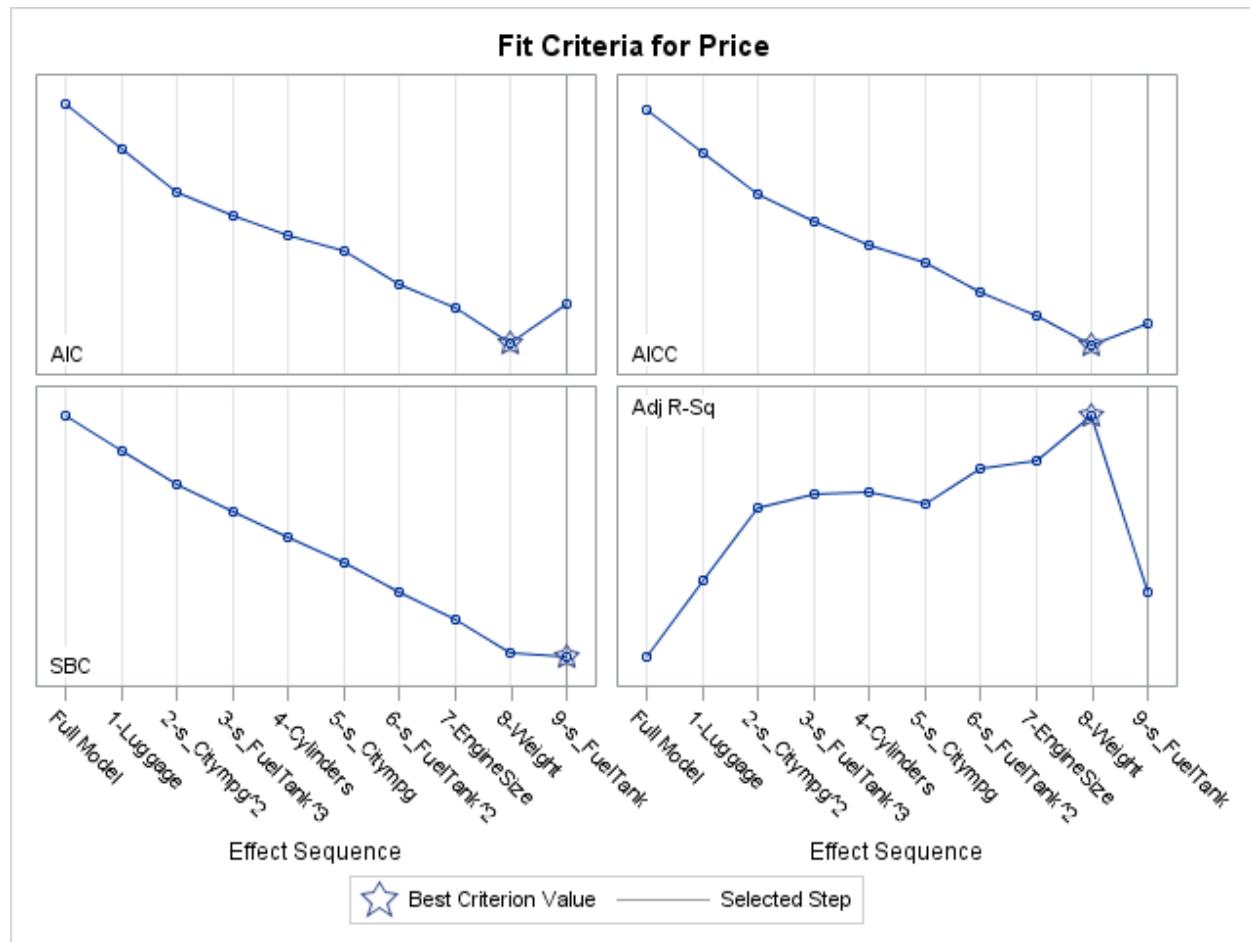
  

Selection stopped because the next candidate for removal has SLS < 0.05.
--

Stop Details				
Candidate For	Effect	Candidate Significance	Compare Significance	
Removal	s_Hwympg^2	0.0032	<	0.0500 (SLS)

## PROC GLMSELECT ODS Graphics Output



This panel of criteria plots from PROC GLMSELECT shows the changes in the model-fit statistics during the backward elimination process. The star symbol on each graph represents the model with the best fit according to the specific statistic. The horizontal axis shows the effect that is removed from the model at each step. The adjusted R square increases as variables are removed from the model and then starts to drop after step eight. The AIC and AICC statistics also select the model at step 8 of the elimination process. The SBC statistic selects the more parsimonious final model as the best model that is evaluated.

## Partial PROC GLMSELECT Output for the Forward Selection Model Using Significance Levels

Effects: Intercept Horsepower s_FuelTank				
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	4333.86956	2166.93478	91.26
Error	78	1852.02921	23.74396	
Corrected Total	80	6185.89877		

Root MSE	4.87278
Dependent Mean	18.64321
R-Square	0.7006
Adj R-Sq	0.6929
AIC	342.49663
AICC	343.02295
SBC	266.67998

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	4.735345	2.524644	1.88
Horsepower	1	0.100057	0.017740	5.64
s_FuelTank	1	0.881558	0.297741	2.96

The forward selection procedure results in a model with two variables, **Horsepower** and **FuelTank**. The adjusted R square is 0.6929, and the SBC is 266.68. Based on these two statistics, this model appears to be inferior to the backward elimination model.

The forward selection and backward elimination procedures resulted in models with completely different variables. This is often the case when the variables being considered for the model are highly correlated with one another.

## Partial PROC GLMSELECT Output for the Backward Elimination Model Using SBC

<b>Effects:</b>	Intercept s_Hwympg s_Hwympg^2 Horsepower
-----------------	--

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	4448.69341	1482.89780	65.73
Error	77	1737.20536	22.56111	
Corrected Total	80	6185.89877		

Root MSE	4.74985
Dependent Mean	18.64321
R-Square	0.7192
Adj R-Sq	0.7082
AIC	339.31229
AICC	340.11229
SBC	265.89009

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	4.039486	2.170242	1.86
s_Hwympg	1	-0.804069	0.213779	-3.76
s_Hwympg^2	1	0.043504	0.014299	3.04
Horsepower	1	0.097301	0.016142	6.03

Backward elimination using SBC chooses the same model as backward elimination using significance levels.

## Partial PROC GLMSELECT Output for the Backward Elimination Model Using Adjusted R Square

<b>Effects:</b>	Intercept s_Citympg s_Hwympg s_Hwympg^2 EngineSize Horsepower s_FuelTank s_FuelTank^2 Weight
-----------------	--

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
<b>Model</b>	8	4591.69736	573.96217	25.92
<b>Error</b>	72	1594.20141	22.14169	
<b>Corrected Total</b>	80	6185.89877		

Root MSE	4.70550
Dependent Mean	18.64321
R-Square	0.7423
Adj R-Sq	0.7136
AIC	342.35400
AICC	345.49686
SBC	280.90405

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	12.947767	4.987854	2.60
s_Citympg	1	-0.477262	0.439348	-1.09
s_Hwympg	1	-0.284251	0.359185	-0.79
s_Hwympg^2	1	0.046174	0.016790	2.75
EngineSize	1	-2.237441	1.542386	-1.45
Horsepower	1	0.070174	0.021417	3.28
s_FuelTank	1	0.298353	0.427658	0.70
s_FuelTank^2	1	0.063630	0.057657	1.10
Weight	1	0.004698	0.003842	1.22

Backward elimination using adjusted R square selects a final model that has eight predictor variables. However, the AIC and SBC values for this model are larger than previous models.

Backward Selection Summary			
Step	Effect Removed	Number Effects In	Adjusted R-Square
0		13	0.7047
1	Luggage	12	0.7089
2	s_Citympg^2	11	0.7128
3	s_FuelTank^3	10	0.7135
4	Cylinders	9	0.7136*
* Optimal Value Of Criterion			

Selection stopped at a local maximum of the AdjRSq criterion.

Stop Details			
Candidate For	Effect	Candidate Adj-RSq	Compare Adj-RSq
Removal	s_Citympg	0.7129	< 0.7136

An examination of the selection summary shows that the process was stopped when the adjusted R square reached a local maximum. If the next candidate variable (**Citympg**) were removed from the model, the adjusted R square would drop. It is important to note that the stepwise selection processes do not guarantee that the model having the optimum value of the selection criterion statistic is selected as the final model.

## 1.13 Multiple Choice Poll

Which of the following is **false**?

- Using different model selection statistics might result in different models.
- Using different model selection methods might result in different models.
- Backward and forward model selection methods might result in very different models if multicollinearity exists among predictor variables.
- Using SELECT=AIC always chooses a final model having the minimum value of the AIC among all possible models.

80

## Select Candidate Models

VARIABLES	METHOD	R <sup>2</sup>	ADJUSTED R <sup>2</sup>	AIC	SBC
Hwympg Hwympg <sup>2</sup> Horsepower	Backward (SL, SBC)	0.7192	0.7082	339.31	265.89
Horsepower FuelTank	Forward SL	0.7006	0.6929	342.50	266.68
Citympg Hwympg Hwympg <sup>2</sup> EngineSize Horsepower FuelTank FuelTank <sup>2</sup> Weight	Backward adjusted R <sup>2</sup>	0.7423	0.7136	342.35	266.53

82

Based on the statistics shown, it appears that the best of the models under consideration is the three-variable model with **Hwympg**, **Hwympg<sup>2</sup>**, and **Horsepower**. The two-variable model with **Horsepower** and **FuelTank** has the smallest R square and the highest AIC and SBC values. The eight-variable model has the highest value of the adjusted R-square statistic, but the AIC and SBC suggest that it is likely more complex than necessary.

It is important to note that there are many potential models that were not examined, and a few are almost as good in terms of these statistics. There are times when a subject-matter expert might be compelled to include certain variables in the model. In addition, monetary considerations, such as the cost to collect specific information, patterns of missing data, and management decisions might be a consideration when you select a model.

## Using Splines for Nonlinear Relationships

- A *spline* is a smooth function consisting of piecewise polynomials joined at points called *knots*.
- Splines can take a number of forms depending on the degree of the polynomial and the number of knots that are used.
- In regression modeling, splines are useful for modeling complex nonlinear relationships for which simple polynomials are not sufficient.
- The EFFECT statement in PROC GLMSELECT can be used to create spline effects.

83

Some nonlinear relationships between independent variables and a response variable might be too complex for a simple polynomial function to model adequately. Splines provide another method to incorporate these predictor variables into a regression model.

Splines consist of polynomial functions that connect smoothly at points called *knots*. In a regression model, a linear combination of polynomial basis functions are fit to the response, and the coefficients for these linear combinations are estimated using ordinary least squares. The EFFECT statement in PROC GLMSELECT uses a penalized B-spline basis (Eilers and Marx, 1996) by default.



## Modeling with Splines

Recall from the exploratory analysis that the variable **FuelTank** had a nonlinear relationship with **Price**, as indicated by the penalized B-spline fit that we produced using PROC SGLOT. Previous models incorporated **FuelTank** using a cubic polynomial, but a spline might provide a better approximation of the relationship between this variable and **Price**.

Consider a model with **FuelTank** as the only predictor variable. First, a cubic polynomial effect for **FuelTank** is used.

```
title 'Spline Effect with Cars Dataset';

proc glmselect data=STAT2.cars2;
  title2 'Cubic polynomial for Fueltank';
  effect p_fuel = polynomial(fueltank /degree=3
                               standardize(method=moments)=center);
  model price = p_fuel / selection=none;
run;                                *ST201d10.sas;
```

Partial PROC GLMSELECT Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3804.58522	1268.19507	41.01	<.0001
Error	77	2381.31354	30.92615		
Corrected Total	80	6185.89877			

Root MSE	5.56113
Dependent Mean	18.64321
R-Square	0.6150
Adj R-Sq	0.6000
AIC	364.85763
AICC	365.65763
SBC	291.43542

The model with a cubic polynomial effect for **FuelTank** has three predictor terms and an adjusted R square of 0.6000.

Now fit a model using a spline effect for **FuelTank**.

```
proc glmselect data=STAT2.cars2;
  title2 'Spline for Fueltank';
  effect sp_fuel = spline(fueltank);
  model price = sp_fuel;
run;                                *ST201d10.sas;
```

Partial PROC GLMSELECT Output

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	4080.24441	680.04073	23.90
Error	74	2105.65436	28.45479	
Corrected Total	80	6185.89877		

Root MSE	5.33430
Dependent Mean	18.64321
R-Square	0.6596
Adj R-Sq	0.6320
AIC	360.89252
AICC	362.89252
SBC	294.65367

The spline model has six predictor terms, and an adjusted R square of 0.6320. The AIC for this model is 360.89, which is slightly smaller than the AIC for the cubic polynomial model. This indicates a slightly better fit.

In this analysis, there are seven spline basis terms that serve as predictors in the model, and the coefficients for these terms are estimated as model parameters, shown below.

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-102.703306	46.498832	-2.21
sp_fuel 1	1	93.787625	65.991401	1.42
sp_fuel 2	1	114.159573	48.940332	2.33
sp_fuel 3	1	111.239552	45.694144	2.43
sp_fuel 4	1	120.802684	47.689484	2.53
sp_fuel 5	1	127.554800	44.203846	2.89
sp_fuel 6	1	160.565066	55.653163	2.89
sp_fuel 7	0	0	.	.

Interpretation of the parameter estimates associated with a spline is not straightforward. In addition, models that contain spline effects are prone to overfitting, so tuning the shape of a spline by adjusting the number, location, or both number and location of the the knots can be important.

-  For more information about fitting models with splines with the EFFECT statement, consult the SAS®9 documentation.



## Exercises

---

### 2. Generating Candidate Models

In the **STAT2.cars4** data set, the dependent variable is **logprice**. The independent variables are the same as the ones in the **STAT2.cars** data set. Complete the following exercise to generate candidate models:

- a. Use PROC SGSCATTER to generate plots of **logprice** versus all other predictor variables. Based on these plots, which variables appear to have a curvilinear relationship with **logprice**? Create new scatter plots of the variables that exhibit curvature using PROC SGSCATTER with the PBSPLINE option. Which variables might need to be squared in a regression model?
- b. Use the EFFECT statement to create centered polynomial effects for the variables identified in step a. Use the model selection methods shown below to generate candidate models with **logprice** as the dependent variable. Make sure that ODS Graphics is enabled. Add the code **plots=criteria** to the PROC GLMSELECT statement to request the selection criteria panel of plots for these model selection methods.
  - 1) backward elimination method using significance levels
  - 2) stepwise selection method using AICC
  - 3) forward selection using adjusted R square
- c. Which variables appear to be appropriate for the regression model?

### Advanced

- d. The plot of **EngineSize** that you created in Exercise 2.a might indicate a more complex relationship than a quadratic one and might be overfitting the data. Look up the NKNOTS= option in the online documentation. Add the NKNOTS=5 option to the PBSPLINE option in your program for Exercise 2.a. What happens to the graph?

## 1.14 Quiz

In the previous exercise, what is your model of choice for **logprice**?

87

# 1.5 Chapter Summary

---

Multiple linear regression enables you to investigate the relationship between a response variable and several predictor variables simultaneously. The null hypothesis is that the slopes for all of the predictor variables are equal to zero. If you reject the null hypothesis, you must determine which of the independent variables have nonzero slopes and are, therefore, useful in the model.

The assumptions for linear regression are the following:

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

A polynomial regression model is a type of multiple linear regression model where powers of variables or cross-product (or interaction) terms (or both) are included in the model. Model development is a process that involves a series of steps. The process should always begin with an initial data exploration that does the following:

- involves examining correlation statistics
- plotting the data to identify
  - those variables that might be useful in the regression model
  - variables that might need higher-ordered terms or other types of transformations

The tests of the parameter estimates help you determine which slopes are nonzero, but they must be considered carefully. They test the significance of each variable when it is added to a model that already contains all of the other independent variables. Therefore, if independent variables in the model are correlated with one another, the significance of both variables can be hidden in these tests. In simple polynomial regression models, you can examine the Type I test to determine the significance of the independent variables in the model. You also want to fit a hierarchically well-formulated model.

For multiple polynomial regression, candidate models can be identified by trial and error or by the use of automatic model selection techniques. Potential models can be evaluated by comparing the adjusted coefficients of determination and information criteria. The criteria panel of plots can provide visual summaries of these statistics when the automatic selection methods are used in PROC GLMSELECT.

When polynomial effects cannot adequately model the nonlinear relationship between a predictor and the response, a spline effect can be used instead. Interpretation of the model parameters associated with splines can be difficult, however.

# 1.6 Solutions

---

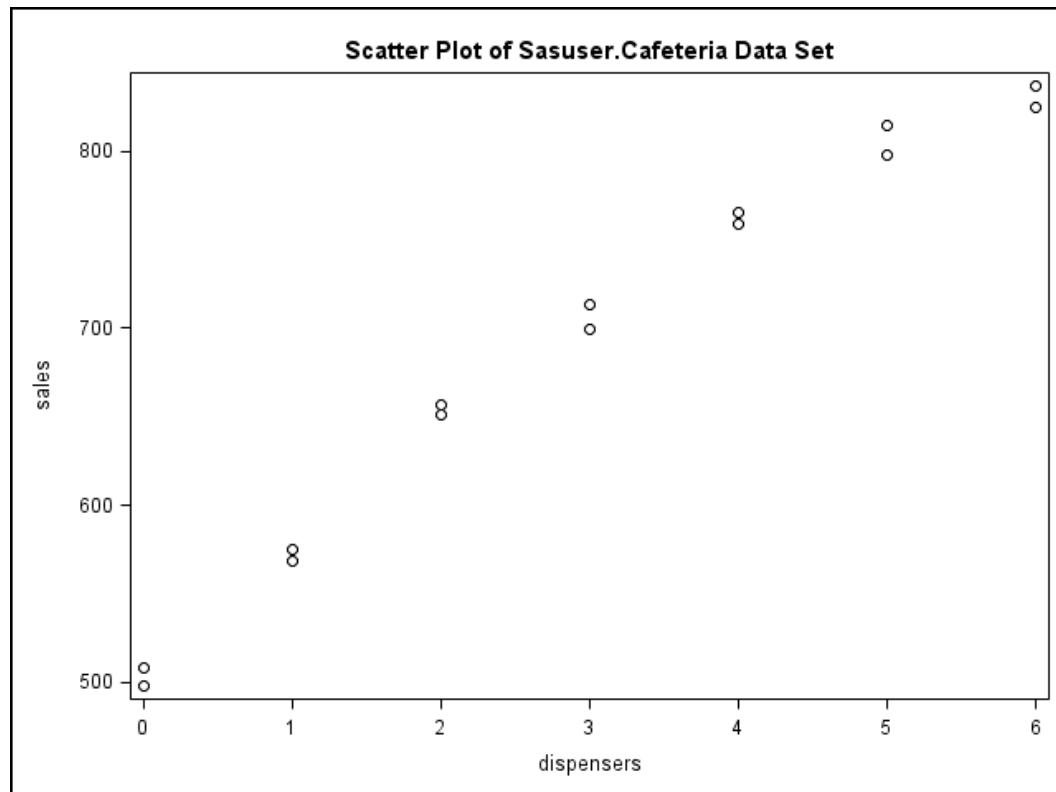
## Solutions to Exercises

### 1. Fitting a Simple Polynomial Regression Model

- Open and run ST200d01.sas.
- Use PROC SGSCATTER to plot sales versus dispensers. How is sales related to dispensers?

```
title "Scatter Plot of STAT2.Cafeteria Data Set";
proc sgscatter data=STAT2.cafeteria;
  plot sales * dispensers;
run;
quit;                                         *ST201s01.sas;
```

PROC SGSCATTER Output



As the number of dispensers increases, the sales increase as well. The relationship appears to be slightly curvilinear.

- Use PROC REG to fit a simple linear regression model to the data. Make sure that ODS Graphics is enabled and add the PLOTS (ONLY UNPACK)=DIAGNOSTICS option in the PROC REG statement to create the diagnostics panel of plots. Examine the plot of the residuals versus the predicted values and the plot of the observed versus the predicted values. Does your model seem to fit the data well?

```
proc reg data=STAT2.cafeteria plots (only unpack)=diagnostics;
  model sales=dispensers;
run;
*ST201s01.sas;
```

PROC REG Output

The REG Procedure  
Model: MODEL1  
Dependent Variable: Sales

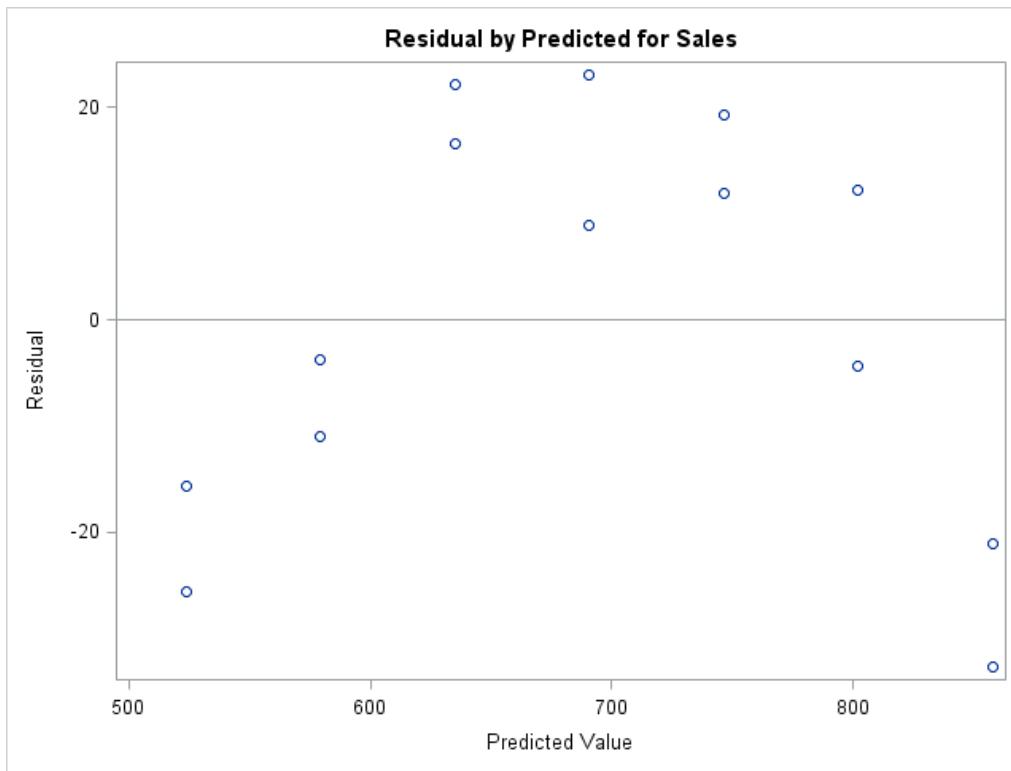
Number of Observations Read	14
Number of Observations Used	14

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	173806	173806	452.52	<.0001
Error	12	4609.05357	384.08780		
Corrected Total	13	178415			

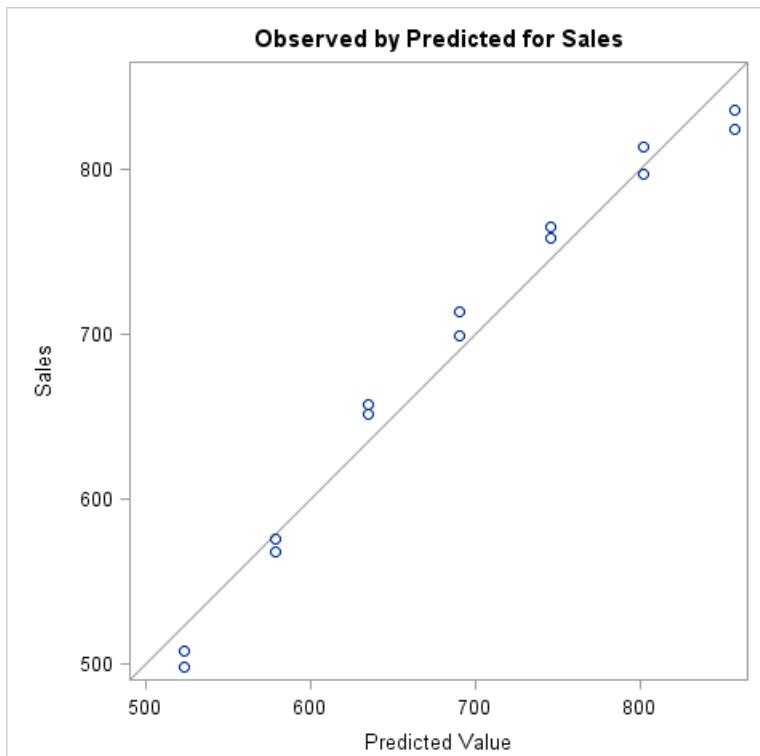
Root MSE	19.59816	R-Square	0.9742
Dependent Mean	690.70000	Adj R-Sq	0.9720
Coeff Var	2.83743		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	523.56786	9.44263	55.45	<.0001
Dispensers	1	55.71071	2.61891	21.27	<.0001

The model is significant with a *p*-value of <0.0001. The variable **dispensers** has a positive slope of 55.71. As the number of dispensers increases by 1, the sales of coffee are expected to increase by 55.71 in hundreds of gallons. The R-square value is 0.974. The model explains about 97% of variation in **sales**.



The plot of the residuals versus the predicted values shows a curvilinear relationship between the residuals and the predicted values. The linear model might not fit your data well.



Although the regression model seems to be close to the observed data, the data points do not seem to be randomly dispersed around the regression line. The linear model might not fit your data well.

- d. Fit a quadratic model by specifying an EFFECT statement in PROC GLMSELECT. Use the OUTDESIGN option to output the design matrix to a data set. Then use PROC REG with the **&\_GLSMOD** macro variable to request Type I tests and the DIAGNOSTICS panel of plots. Look at the plot of the residuals versus the predicted values, the plot of the observed versus the predicted values and the normal quantile plot for residuals. Use PROC SGPlot with a REG statement with the DEGREE=2 option to create a scatter plot of the data with a second-degree regression overlaid. From the plots, do you think the quadratic model fits your data better than the linear model?

```

proc glmselect data=STAT2.cafeteria outdesign=d_disp;
  effect q_disp=polynomial(dispensers / degree=2);
  model sales = q_disp / selection=none;
run;

proc reg data=d_disp plots (only unpack) = diagnostics;
  model sales=&_GLSMOD / scorrl(tests);
run;
quit;

proc sgplot data=STAT2.cafeteria;
  reg y=sales x=dispensers / degree=2;
run;
*ST201s01.sas;

```

Partial PROC GLMSELECT Output

The GLMSELECT Procedure Least Squares Model (No Selection)					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	177741	88870	1448.87	<.0001
Error	11	674.71429	61.33766		
Corrected Total	13	178415			

Root MSE	7.83184
Dependent Mean	690.70000
R-Square	0.9962
Adj R-Sq	0.9955
AIC	76.25325
AICC	80.69769
SBC	62.17042

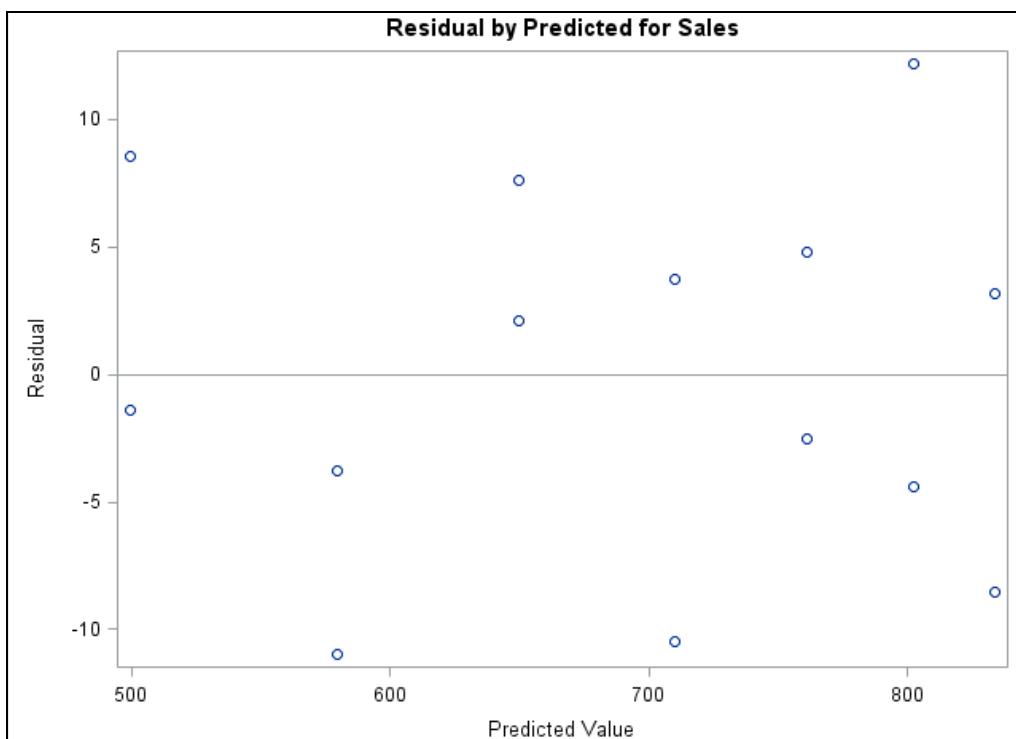
The model is significant. The adjusted R-square value increases to 0.9955.

## Partial PROC REG Output

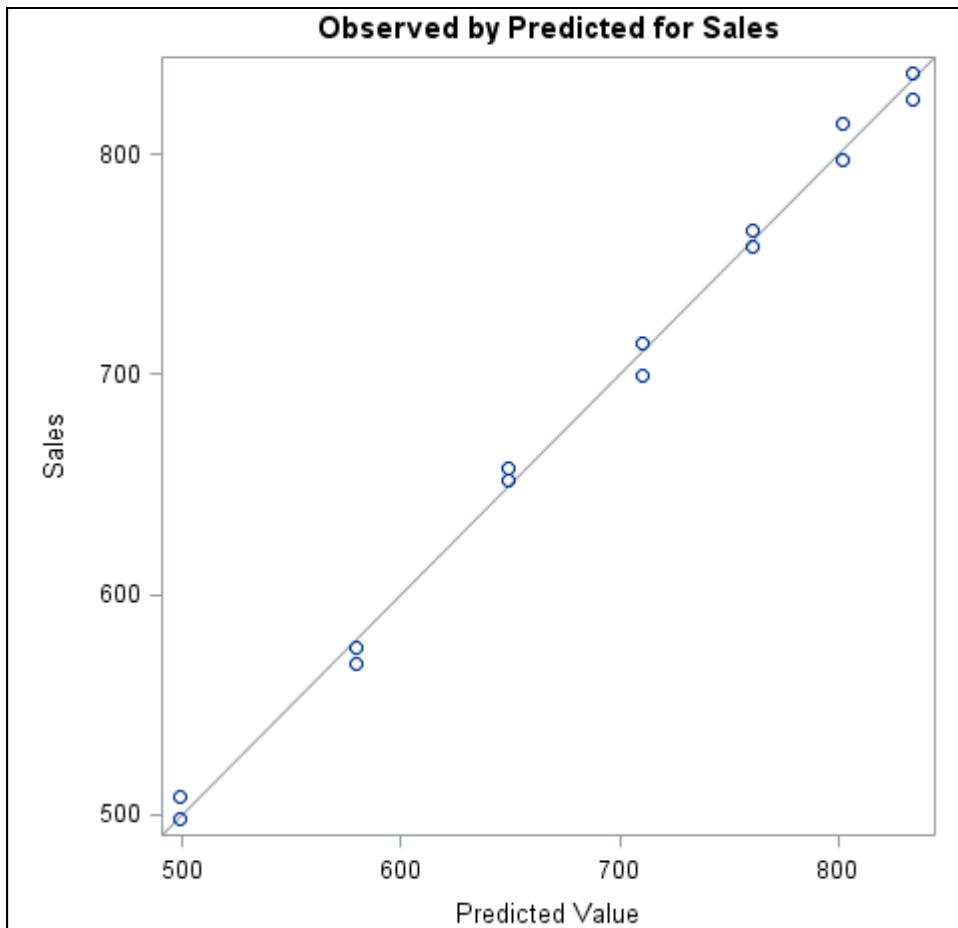
Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Squared Semi-partial Corr Type I	Cumulative R-Square	Type I	
									F Value	Pr > F
Intercept	Intercept	1	499.37143	4.83391	103.31	<.0001	.	0	.	.
Dispensers	Dispensers	1	84.74643	3.77347	22.46	<.0001	0.97417	0.97417	2833.60	<.0001
Dispensers_2	Dispensers^2	1	-4.83929	0.60424	-8.01	<.0001	0.02205	0.99622	64.14	<.0001

The Type I test indicates that both **dispensers** and **dispensers^2** are significant factors.

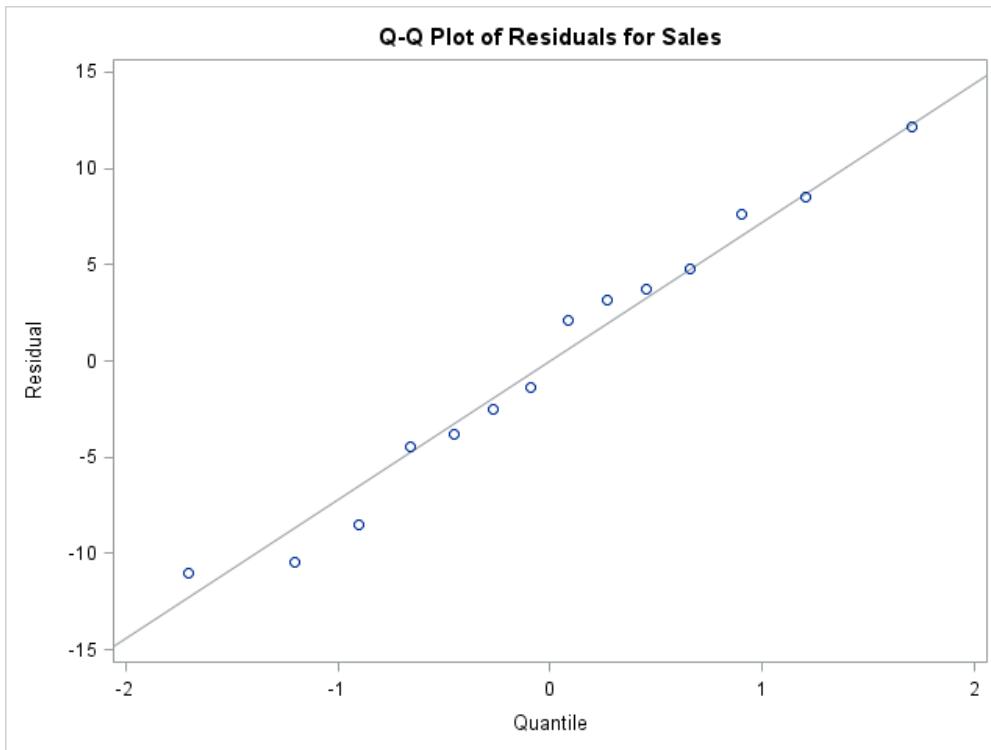
## Partial PROC REG ODS Graphics Output



The residual plots show a random scatter around the reference line.

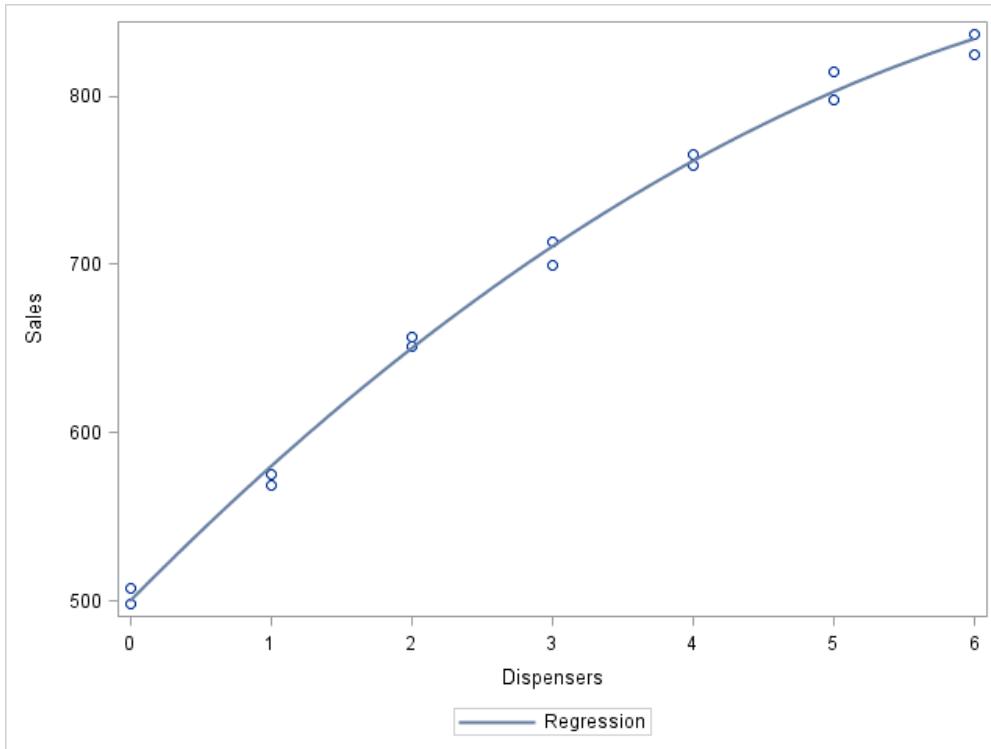


The plot of the observed values versus the predicted values indicates a better model fit than the linear model.



From the Q-Q plot, the residuals seem to be normally distributed.

#### PROC SGPLOT Output

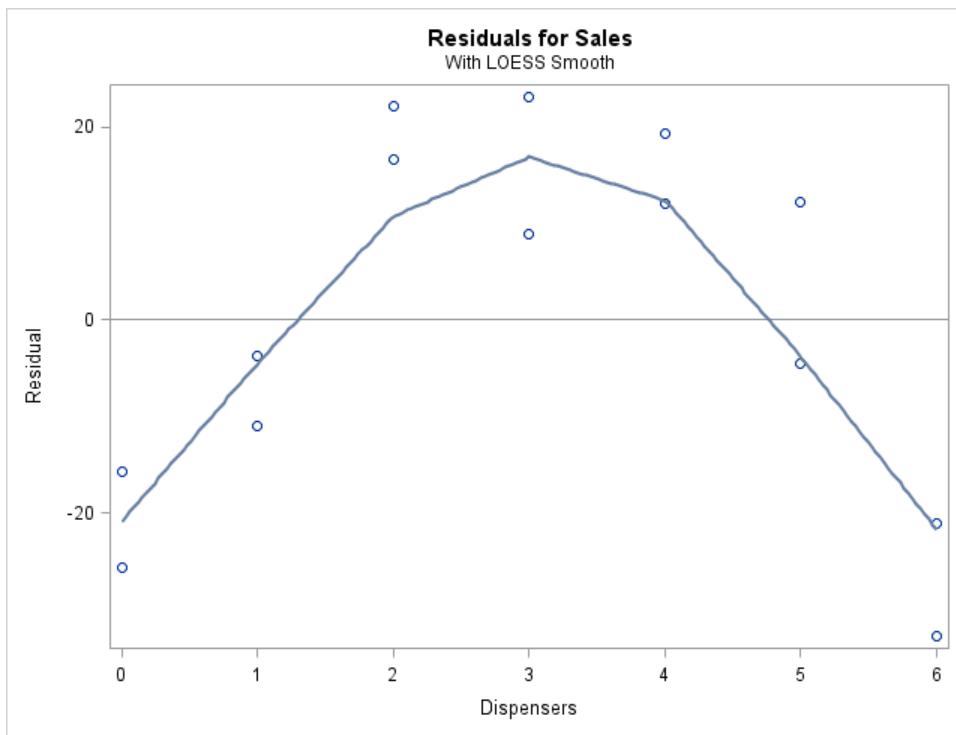


The scatter plot with the overlaid quadratic model indicates that this model fits the data better than the linear model.

- e. You can obtain plots of the residuals versus the regressors by adding the RESIDUALS keyword to your plots request in PROC REG. To help in detecting patterns, you can use the SMOOTH= suboption of the RESIDUALS plots request to add LOESS smoothed curves to these residual plots. Consult the online documentation for PROC REG to see how to add the appropriate options to your code. Run PROC REG with two MODEL statements to compare the linear model to the quadratic model. Add the appropriate options to create the RESIDUALS plots with a smooth curve added. Compare the residual plots from the two models. Which model fits your data better?

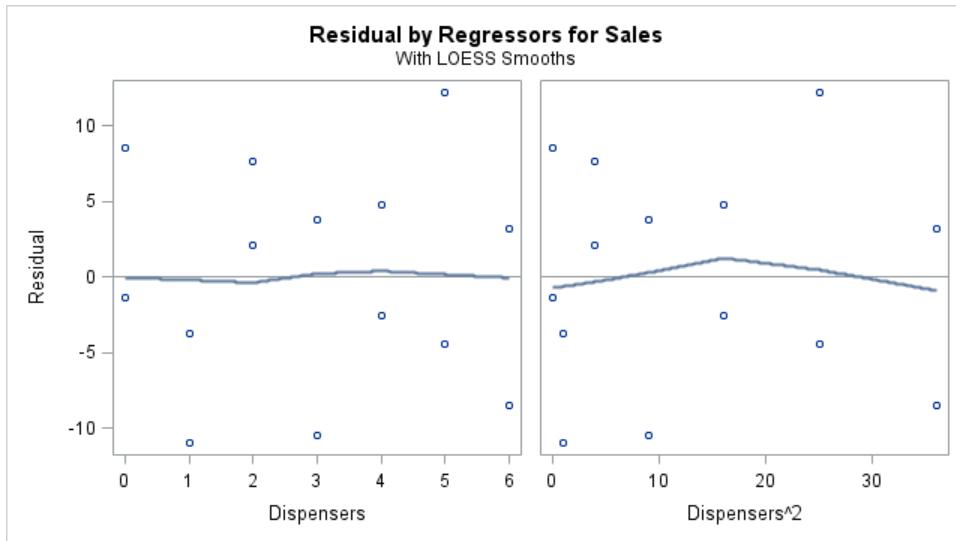
```
proc reg data=d_disp plots(only label) = (residuals (smooth));
  LINEAR: model sales=dispensers;
  QUADRATIC: model sales=&_GLSMOD;
run;
quit;
```

Partial PROC REG ODS Graphics Output



The plot of the residuals versus the regressor (**dispensers**) for the linear model shows a definite pattern of curvature, indicating that the model might need a quadratic term.

## Partial PROC REG ODS Graphics Output



For the quadratic model, the plots of the residuals versus **dispensers** and the residuals versus **dispensers\_2** indicate a much better fit.

- f. Use PROC CORR to compute the Pearson correlation coefficient between **dispensers** and **dispensers\_2** (dispensers squared). Recall that the automatic macro variable created from the previous run of PROC GLMSELECT, &\_GLSMOD, resolves to the two variables of interest. Use the NOSIMPLE and PLOTS()=MATRIX options in the PROC CORR statement. Examine the tabular and graphical output. What do you conclude?

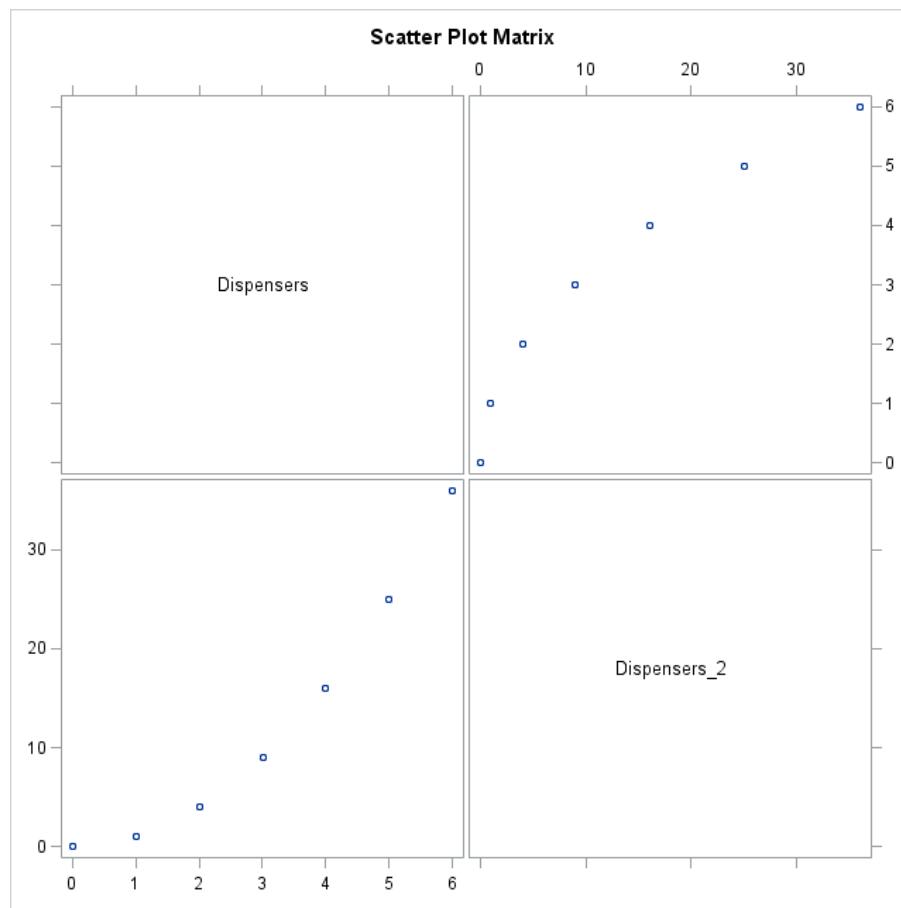
```
proc corr data=d_disp nosimple plots()=matrix;
  var &_GLSMOD;
run; *ST201s01.sas;
```

PROC CORR Output

The CORR Procedure		
2 Variables:		Dispensers Dispensers_2
<b>Pearson Correlation Coefficients, N = 14</b>		
Prob >  r  under H0: Rho=0		
	Dispensers	Dispensers_2
Dispensers	1.00000	0.96077
Dispensers		<.0001
Dispensers_2	0.96077	1.00000
Dispensers^2	<.0001	

The variables **dispensers** and **dispensers\_2** are highly correlated. The correlation coefficient is 0.96 and is significantly different from zero.

## PROC CORR ODS Graphics Output



The scatter plot between **dispensers** and **dispensers\_2** shows a strong curvilinear relationship, as expected.

- g. Use the appropriate options in the MODEL statement in PROC REG to compute the variance inflation factor (VIF) and the collinearity diagnostics statistics. Is there collinearity among the independent variables? If so, which ones?

```
proc reg data=d_disp;
  model sales=&_GLSMOD / vif collin collinoint;
run;
quit; *ST201s01.sas;
```

Partial PROC REG Output

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
<b>Intercept</b>	Intercept	1	499.37143	4.83391	103.31	<.0001	0
<b>Dispensers</b>	Dispensers	1	84.74643	3.77347	22.46	<.0001	13.00000
<b>Dispensers_2</b>	Dispensers^2	1	-4.83929	0.60424	-8.01	<.0001	13.00000

Collinearity Diagnostics					
Number	Eigenvalue	Condition Index	Proportion of Variation		
			Intercept	Dispensers	Dispensers_2
1	2.68573	1.00000	0.02084	0.00321	0.00464
2	0.30018	2.99115	0.41154	0.00350	0.03651
3	0.01409	13.80533	0.56762	0.99329	0.95886

Collinearity Diagnostics (intercept adjusted)				
Number	Eigenvalue	Condition Index	Proportion of Variation	
			Dispensers	Dispensers_2
1	1.96077	1.00000	0.01962	0.01962
2	0.03923	7.06965	0.98038	0.98038

The VIF for **dispensers** and **dispensers\_2** are both 13, which indicates strong collinearity. However, the largest condition index value is 13.08, suggesting weak collinearity. It is not uncommon to reach inconsistent conclusions about collinearity based on different statistics. It might be a good idea to reduce the possible collinearity for more reliable inferences.

- h. Use PROC GLMSELECT with an EFFECT statement to create a new, centered, quadratic effect for **dispensers** and fit a model with the centered polynomial terms. Use the OUTDESIGN option to create a new model design data set. Using the design data set, obtain collinearity diagnostics from PROC REG. Does the centered model appear to have multicollinearity among the independent variables?

```

proc glmselect data=STAT2.cafeteria outdesign=d_dispc;
  effect q_dispc=polynomial(dispensers / degree=2
                             standardize(method=moments)=center);
  model sales = q_dispc / selection=none;
run;

ods html select ParameterEstimates CollinDiag CollinDiagNoInt;
proc reg data=d_dispc;
  model sales = &_GLSMOD / vif collin collinoint;
  title 'Centered Quadratic Model';
run;
title;
quit; *ST201s01.sas;

```

## Partial PROC REG Output

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	Intercept	1	710.05714	3.19733	222.08	<.0001	0
s_Dispensers	s_Dispensers	1	55.71071	1.04657	53.23	<.0001	1.00000
s_Dispensers_2	s_Dispensers^2	1	-4.83929	0.60424	-8.01	<.0001	1.00000

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	s_Dispensers	s_Dispensers_2	
1	1.75593	1.00000	0.12204	0	0.12204	
2	1.00000	1.32511	0	1.00000	0	
3	0.24407	2.68223	0.87796	0	0.87796	

Collinearity Diagnostics (intercept adjusted)				
Number	Eigenvalue	Condition Index	Proportion of Variation	
			s_Dispensers	s_Dispensers_2
1	1.00000	1.00000	1.00000	0
2	1.00000	1.00000	0	1.00000

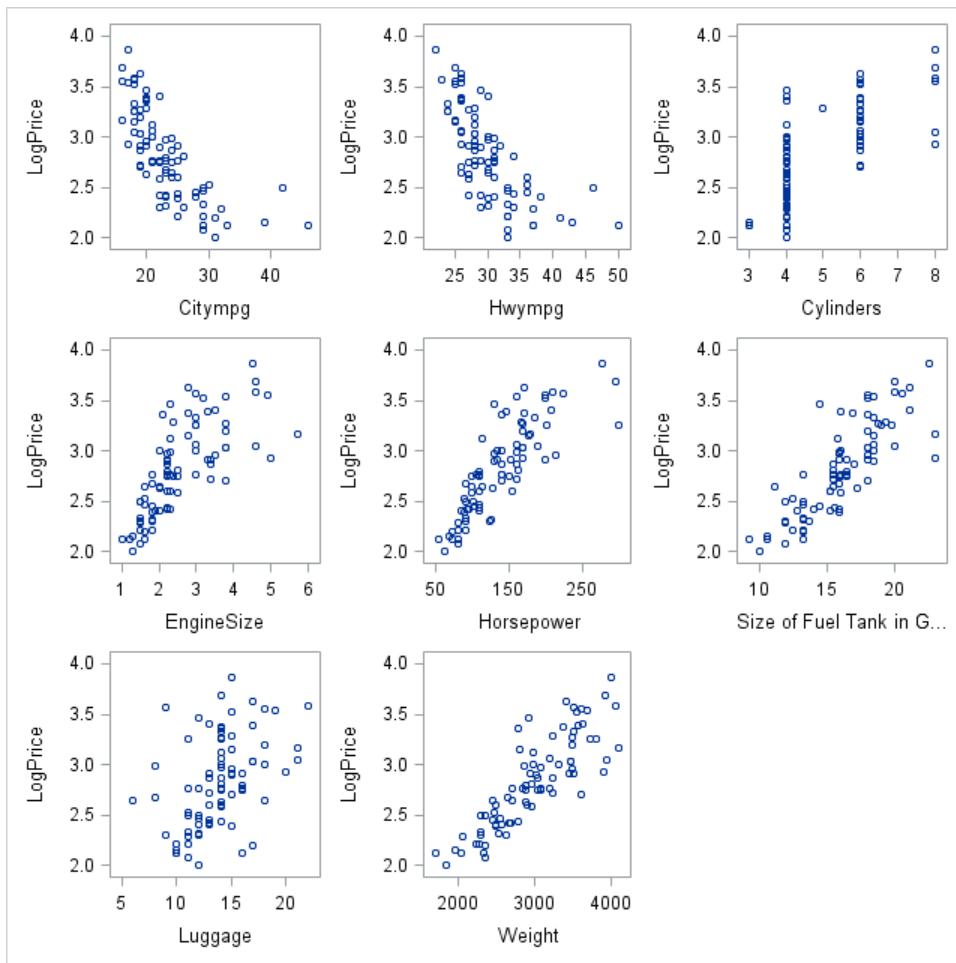
After you center the independent variable **dispensers**, the VIF is reduced to 1 for both centered variables. The largest condition index value is only 2.68. There does not seem to be any collinearity among the (centered) independent variables.

## 2. Generating Candidate Models

- Use PROC SGSCATTER to generate plots of **logprice** versus all other predictor variables. Based on these plots, which variables appear to have a curvilinear relationship with **logprice**? Create new scatter plots of the variables that exhibit curvature using PROC SGSCATTER with the PBSPLINE option. Which variables might need to be squared in a regression model?

```
proc sgscatter data=STAT2.cars4;
  plot logprice*(citympg hwympg cylinders enginesize horsepower
    fueltank luggage weight);
run;                                         *ST201s02.sas;
```

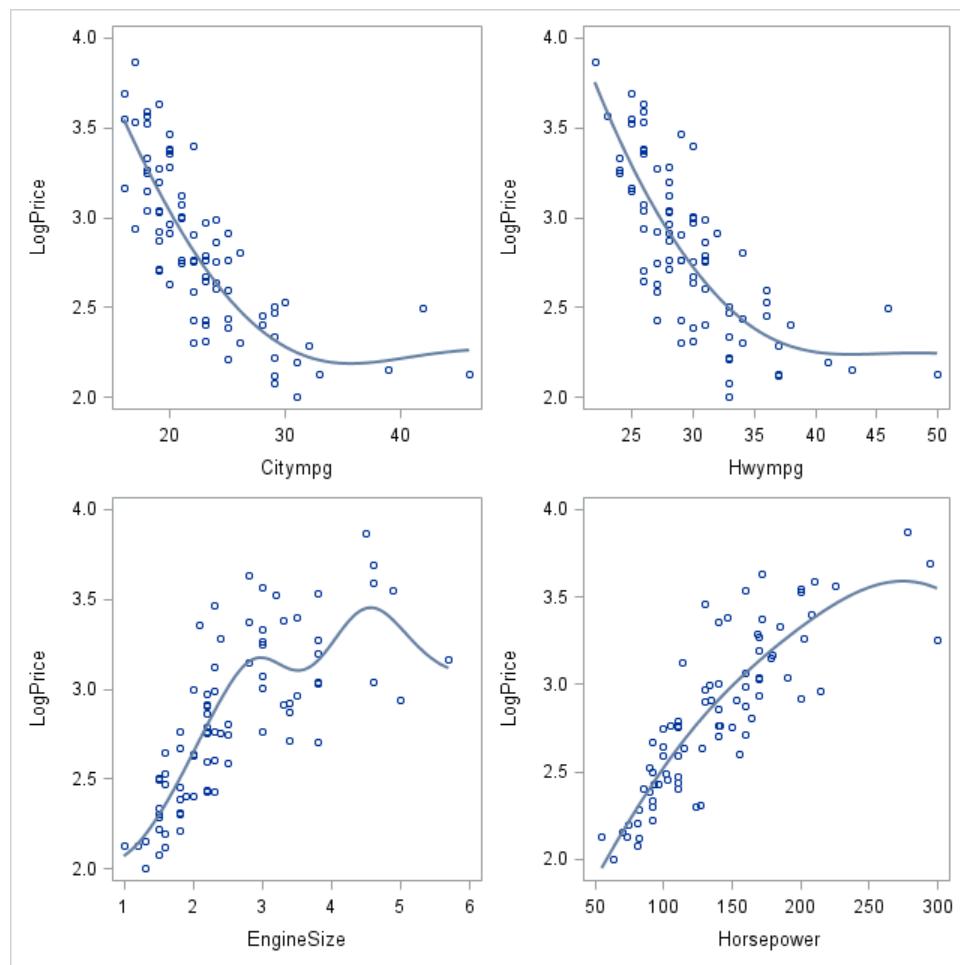
## PROC SGSCATTER Output



Based on these plots, **Citympg**, **Hwympg**, **EngineSize**, and **Horsepower** might appear to have curvilinear relationships with **logprice**. Use PROC SGSCATTER with the PBSPLINE option to create new scatter plots of these variables.

```
proc sgscatter data=STAT2.cars4;
  plot logprice*(citympg hwympg enginesize horsepower) / pbspline;
run; *ST201s02.sas;
```

## PROC SGSCATTER Output



Based on these graphics, it appears that **Citympg**, **Hwympg**, and **Horsepower** might need to be squared in the regression model. The variable **EngineSize** might need higher-order polynomial terms.

- b. Use the EFFECT statement to create centered polynomial effects for the variables identified in step a. Use the model selection methods shown below to generate candidate models with **logprice** as the dependent variable; make sure that ODS Graphics is enabled. Add the code **plots=criteria** to the PROC GLMSELECT statement to request the selection criteria panel of plots for these model selection methods.
- 1) backward elimination method using significance levels
  - 2) stepwise selection method using AICC
  - 3) forward selection using adjusted R square

```
%macro e_poly;

effect p_city=polynomial(citympg / degree=2
                           standardize(method=moments)=center);
effect p_hwy=polynomial(hwympg / degree=2
                           standardize(method=moments)=center);
effect p_engine=polynomial(engine size / degree=2
```

```
standardize(method=moments)=center);
effect p_hp=polynomial(horsepower / degree=2
                        standardize(method=moments)=center);

%mend;

proc glmselect data=STAT2.cars4 plots= criteria;
  title 'Backward elimination using p-values';
  %e_poly;
  model logprice=p_city p_hwy cylinders p_engine p_hp fueltank
    luggage weight / selection=backward select=sl hierarchy=single;
run;

proc glmselect data=STAT2.cars4 plots= criteria;
  title 'Stepwise selection using AICC';
  %e_poly;
  model logprice=p_city p_hwy cylinders p_engine p_hp fueltank
    luggage weight / selection=stepwise select=aicc
    hierarchy=single;
run;

proc glmselect data=STAT2.cars4 plots= criteria;
  title 'Forward selection using adjusted R-squared';
  %e_poly;
  model logprice=p_city p_hwy cylinders p_engine p_hp fueltank
    luggage weight / selection=forward select=adjrsq
    hierarchy=single;
run;                                              *ST201s02.sas;
```

## Partial PROC GLMSELECT Backward Elimination Output

The selected model is the model at the last step (Step 6).

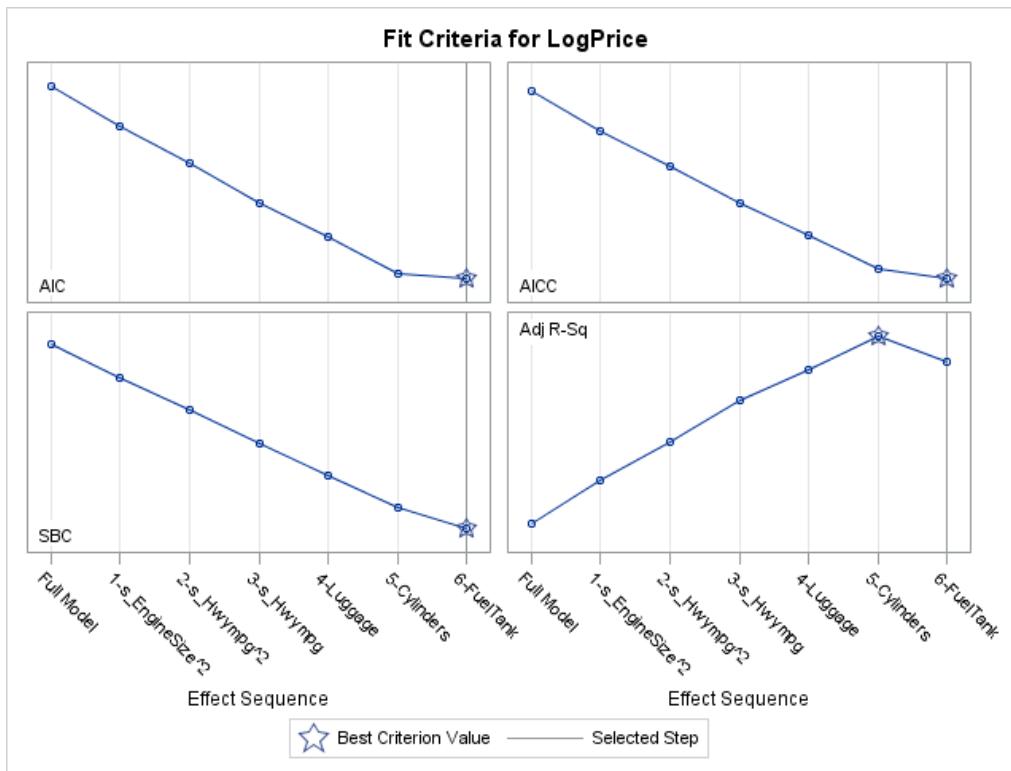
**Effects:** Intercept s\_Citympg s\_Citympg^2 s\_EngineSize s\_Horsepower s\_Horsepower^2 Weight

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	13.13371	2.18895	53.14
Error	74	3.04807	0.04119	
Corrected Total	80	16.18179		

Root MSE	0.20295
Dependent Mean	2.82388
R-Square	0.8116
Adj R-Sq	0.7964
AIC	-168.67506
AICC	-166.67506
SBC	-234.91391

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	1.581640	0.458723	3.45
s_Citympg	1	-0.038770	0.012692	-3.05
s_Citympg^2	1	0.001968	0.000587	3.35
s_EngineSize	1	-0.170680	0.060525	-2.82
s_Horsepower	1	0.004417	0.001233	3.58
s_Horsepower^2	1	-0.000014810	0.000007506	-1.97
Weight	1	0.000409	0.000152	2.69

## PROC GLMSELECT Backward Elimination ODS Graphics Output



The fit statistics AIC, AICC, and SBC decrease with each step of the backward elimination. The adjusted R square increases until step 5 and then decreases. The AIC, AICC, and SBC statistics select the model that is generated at step 6 to be the “best” model. The adjusted R square selects the model at step 5.

## PROC GLMSELECT Stepwise Selection Output

The selected model is the model at the last step (Step 5).

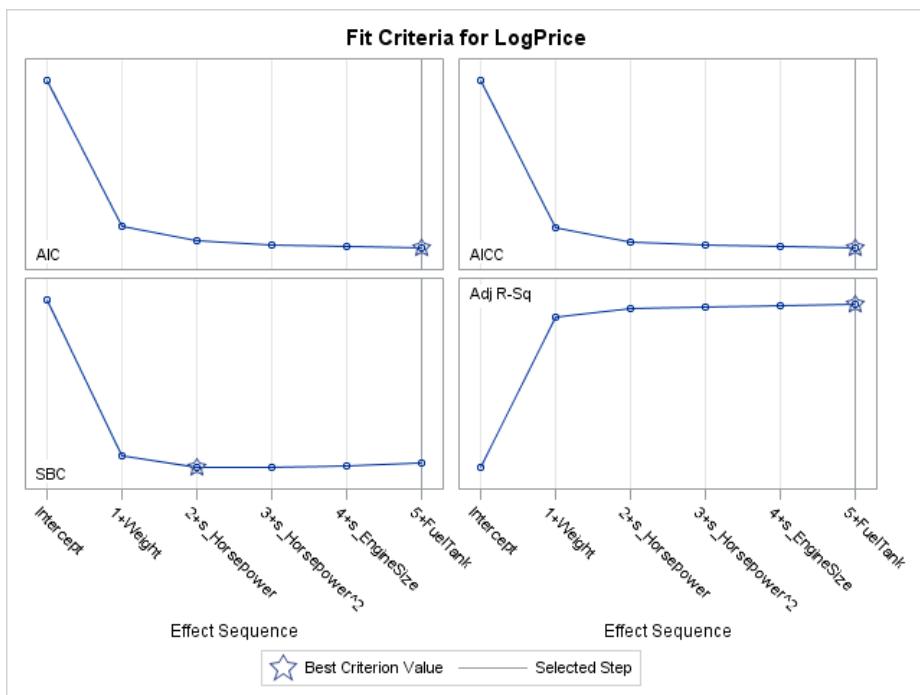
<b>Effects:</b>	Intercept s_EngineSize s_Horsepower s_Horsepower^2 FuelTank Weight
-----------------	--

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	12.79036	2.55807	56.57
Error	75	3.39143	0.04522	
Corrected Total	80	16.18179		

Root MSE	0.21265
Dependent Mean	2.82388
R-Square	0.7904
Adj R-Sq	0.7764
AIC	-162.02910
AICC	-160.49486
SBC	-230.66241

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	1.276910	0.450509	2.83
s_EngineSize	1	-0.104482	0.059656	-1.75
s_Horsepower	1	0.004659	0.001283	3.63
s_Horsepower^2	1	-0.0000013128	0.0000007615	-1.72
FuelTank	1	0.031462	0.018190	1.73
Weight	1	0.000360	0.000165	2.19

## PROC GLMSELECT Stepwise Selection ODS Graphics Output



The adjusted R-square statistic rises initially and then begins to level off. The five-predictor model selected by AICC is the same as the model selected by AIC and the adjusted R square. However, SBC selects the two-predictor model with **Weight** and **Horsepower** only.

## PROC GLMSELECT Forward Selection Output

The selected model is the model at the last step (Step 6).

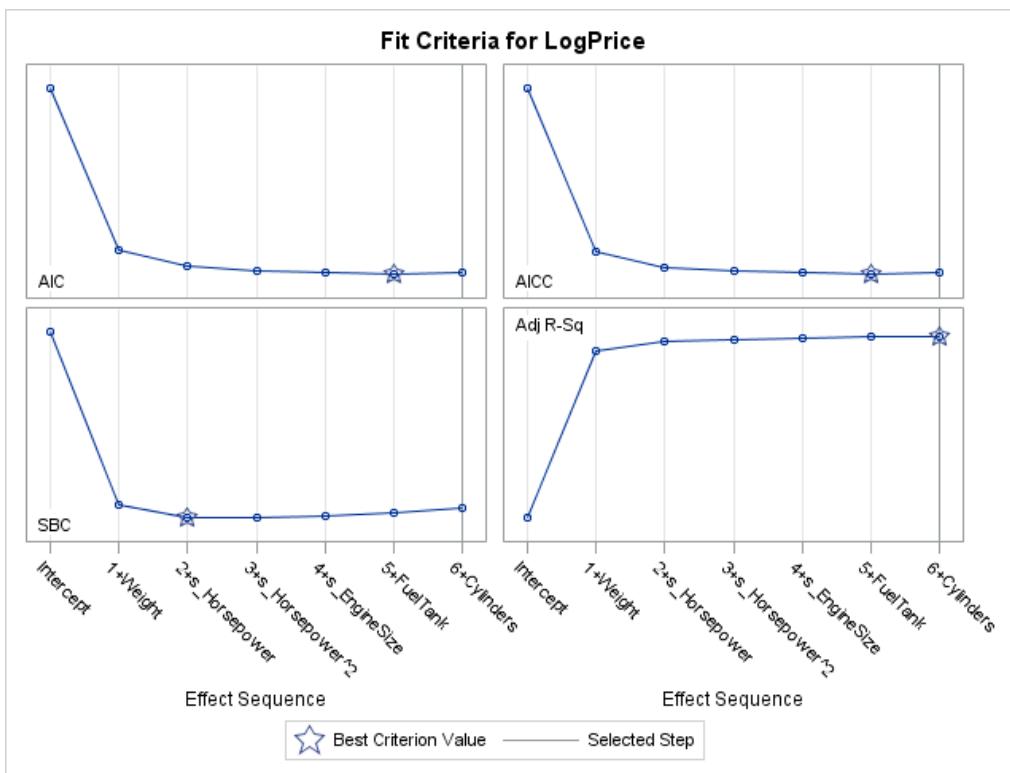
Effects:	Intercept Cylinders s_EngineSize s_Horsepower s_Horsepower^2 FuelTank Weight
----------	--

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
<b>Model</b>	6	12.83652	2.13942	47.33
<b>Error</b>	74	3.34527	0.04521	
<b>Corrected Total</b>	80	16.18179		

Root MSE	0.21262
Dependent Mean	2.82388
R-Square	0.7933
Adj R-Sq	0.7765
AIC	-161.13915
AICC	-159.13915
SBC	-227.37800

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	1.018811	0.517822	1.97
Cylinders	1	0.044192	0.043733	1.01
s_EngineSize	1	-0.150302	0.074926	-2.01
s_Horsepower	1	0.004375	0.001313	3.33
s_Horsepower^2	1	-0.000014030	0.000007666	-1.83
FuelTank	1	0.034125	0.018377	1.86
Weight	1	0.000362	0.000165	2.20

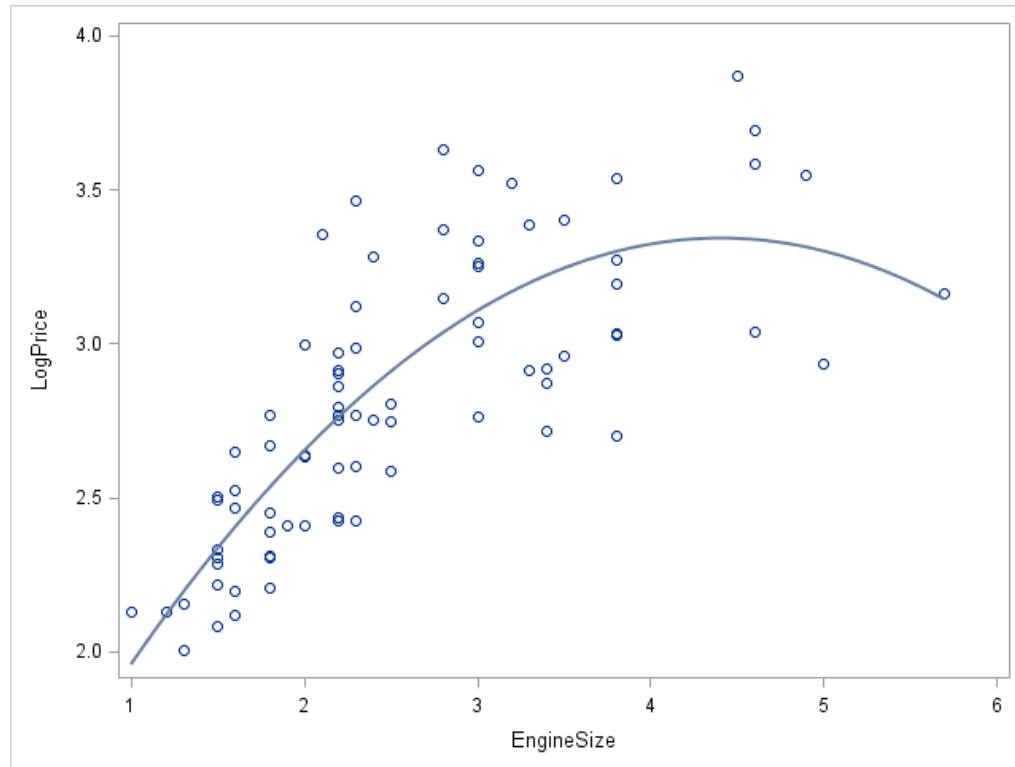
## Partial PROC GLMSELECT Forward Selection ODS Graphics



The criteria panel of plots for the adjusted R-square forward selection indicates (with a star) the best fitting model among models with a given number of parameters.

- The stepwise model based on AICC and the forward selection model based on adjusted R square choose similar models with **EngineSize**, **Horsepower**, **Horsepower<sup>2</sup>**, **FuelTank**, and **Weight** in common. The forward model adds **Cylinders**. These two models have nearly identical adjusted R-square values (0.7764 and 0.7765). The backward elimination model includes **EngineSize**, **Citympg**, **Citympg<sup>2</sup>**, **Horsepower**, **Horsepower<sup>2</sup>**, and **Weight**. This model has a larger adjusted R-square value (0.7964) and smaller information criteria values. Among the three models considered here, it would be favored.
- The plot of **EngineSize** that you created in Exercise 2.a might indicate a more complex relationship than a quadratic one and might be overfitting the data. Look up the NKNOTS= option in the online documentation. Add the NKNOTS=5 option to the PBSPLINE option in your program for Exercise 2.a. What happens to the graph?

```
proc sgscatter data=STAT2.cars4;
  plot logprice*enginesize / pbspline = (nknots=5);
run; *ST201s02.sas;
```



The curve becomes smoother and looks like a quadratic.

## Solutions to Student Activities (Polls/Quizzes)

### 1.03 Quiz – Correct Answer

Suppose the regression model that you fit is the following:

$$\hat{y} = 3 + 5x$$

How do you interpret the slope for  $x$ , which is 5?

**For every one-unit increase in  $x$ , the predicted value for  $y$  increases by 5.**

## 1.04 Multiple Choice Poll – Correct Answer

You learned from the demonstration that you should check the assumption that the error terms are normally distributed. How can you do this?

- a. examine the histogram and normal quantile plot of the residuals
- b. request formal tests of normality for the residuals in PROC UNIVARIATE
- c. either a or b

26

## 1.05 Multiple Choice Poll – Correct Answer

The residual plots (residuals versus predicted values, and also residuals versus time, if applicable) for regression models are important because they help to

- a. identify lack of fit of the model
- b. display nonconstant variance
- c. evaluate normality of the residuals
- d. display correlated errors
- e. a, b, and d.

30

## 1.06 Multiple Choice Poll – Correct Answer

Which of the following is **false**?

- a. Polynomial regression models belong to the category of nonlinear regression models.
- b. Polynomial regression models belong to the category of linear regression models.
- c. Polynomial regression models fit a curvilinear model to your data.
- d. all of the above
- e. none of the above

35

## 1.07 Multiple Choice Poll – Correct Answer

Which of the following is **false**?

- a. Polynomial regression is a nonlinear model, and therefore, you should not use PROC GLMSELECT.
- b. You should follow the principle of model hierarchy when you remove terms from a polynomial regression model.
- c. It is important to check model assumptions after the final polynomial regression model is chosen.

42

## 1.08 Quiz – Correct Answer

In the previous exercise, what did you find about the linear model when you examined the residual plot?

**Because the residual plot showed a pattern (curvilinear), your model might not be adequately specified (missing a quadratic term).**

46

## 1.09 Quiz – Correct Answer

Is the statement below true or false? Explain why.

In general, you can remove the nonsignificant terms from a model all at once to obtain a parsimonious model.

**The answer is false. Multicollinearity might hide significant terms. You need to eliminate them one at a time.**

62

## 1.10 Poll – Correct Answer

When you fit a straight line to data that shows a curvilinear relationship, it is very likely that the residual plot shows a curvilinear pattern.

- True  
 False

64

## 1.11 Quiz – Correct Answer

Is the statement below true or false? Explain why.

In the previous exercise, after you centered the variable **dispenser** and refit the polynomial model using the centered variables, both the ANOVA table and the parameter estimate table changed.

**False. The ANOVA table did not change, but the parameter estimate table changed.**

67

## 1.12 Multiple Choice Poll – Correct Answer

Which of the following statements regarding the EDA is **false**?

- a. There is a curvilinear relationship between **Price** and **Hwympg**, **Citympg**, and **FuelTank**.
- b. **Price** and **Hwympg** are negatively correlated.
- c. **Luggage** seems to be an important independent variable.

73

## 1.13 Multiple Choice Poll – Correct Answer

Which of the following is **false**?

- a. Using different model selection statistics might result in different models.
- b. Using different model selection methods might result in different models.
- c. Backward and forward model selection methods might result in very different models if multicollinearity exists among predictor variables.
- d. Using **SELECT=AIC** always chooses a final model having the minimum value of the AIC among all possible models.

81



# Chapter 2 Regression Diagnostics and Remedial Measures

<b>2.1 Regression Model Diagnostics .....</b>	<b>2-3</b>
Demonstration: Model Diagnostics – Normality, Constant Variance, Model Fit, Collinearity, and Influential Observations .....	2-15
Exercises .....	2-27
<b>2.2 Remedial Measures .....</b>	<b>2-28</b>
Demonstration: Fitting a Lognormal Regression Model.....	2-37
Exercises .....	2-43
<b>2.3 Chapter Summary.....</b>	<b>2-44</b>
<b>2.4 Solutions .....</b>	<b>2-46</b>
Solutions to Exercises .....	2-46
Solutions to Student Activities (Polls/Quizzes).....	2-64



## 2.1 Regression Model Diagnostics

---

### Objectives

- Evaluate model assumptions.
- Evaluate model fit.
- Evaluate multicollinearity.
- Identify influential observations.

3

### Evaluate Model Assumptions

- The assumptions for linear regression are that the error terms are independent and normally distributed with equal variance.

$$\varepsilon \sim iid N(0, \sigma^2)$$



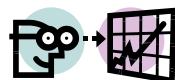
- Therefore, evaluating model assumptions for linear regression includes checking for the following:
  - independent observations
  - normally distributed error terms
  - constant variance

4

Recall that the assumptions of regression are that the residuals are independently and normally distributed with equal variance. These assumptions can be written in statistical notation as  $\varepsilon \sim N(0, \sigma^2)$ . If these assumptions are met, plots of the residuals versus the predicted values are a random scatter about a zero reference line. Additional tools are available to evaluate each of the assumptions.

## Evaluate Model Assumptions: Independence

- Know the source of your data. Correlated errors can arise from data gathered over time, repeated measures, clustered data, or data from complex survey designs.
- For time series data, check that the errors are independent by examining the following:
  - plots of residuals versus time or other ordering component
  - Durbin-Watson statistic or the first-order autocorrelation statistic for time series data



5

One of the assumptions in linear regression is independent errors. Error terms are dependent if the values of the errors depend on the other values of errors. Knowing how your data are generated helps evaluate the assumption of independence. Correlated error terms can arise from data from a complex survey design, or from repeated measures on a given subject, any type of clustered data, or data gathered over time.

For time series data, tools are available to evaluate whether the error terms are correlated. These tools include the following:

- plots of the residuals versus time to examine whether there seems to be any positive or negative autocorrelations
- Durbin-Watson statistic or the first-order autocorrelation statistic available from PROC REG

## Evaluate Model Assumptions: Normality

Check that the error terms are normally distributed by examining these items:

- a histogram of the residuals
- a normal probability plot of the residuals
- tests for normality



6

You can check the assumption of normality of the error terms by examining a histogram of the residuals with a normal curve overlaid or a normal quantile plot of the residuals. These can be obtained either from the ODS Graphics output of many statistical procedures, including PROC REG and PROC UNIVARIATE, in addition to PROC SGPlot. Formal tests of normality are available from PROC UNIVARIATE.

## Evaluate Model Assumptions: Constant Variance

Check for constant variance of the error terms by examining the following items:



- plot of residuals versus predicted values
- plots of residuals versus the independent variables
- test for heteroscedasticity
- Spearman rank correlation coefficient between absolute values of the residuals and predicted values

7

For the constant variance assumption, you can examine plots of the residuals versus the predicted values, and plots of the residuals versus the independent variables. Both of these plots are available via the ODS Graphics output from PROC REG. In addition, you can perform a test for heteroscedasticity under certain conditions. This test is available as an option in the MODEL statement in PROC REG.

You can also compute the Spearman rank correlation coefficient between the absolute value of the residuals and the predicted values to evaluate homoscedasticity.

## Spearman Rank Correlation Coefficient

- The Spearman rank correlation coefficient is available as an option in PROC CORR.
- If the Spearman rank correlation coefficient between the absolute value of the residuals and the predicted values is
  - close to zero, then the variances are approximately equal
  - positive, then the variance increases as the mean increases
  - negative, then the variance decreases as the mean increases.

8

One way to determine whether the variance is stable is to compute the Spearman rank correlation coefficient between the absolute value of the residuals and the predicted values (Carroll and Ruppert 1988).

This statistic, available as an option in PROC CORR, measures the correlation between the size of the ordered predicted values and the absolute value of their associated residuals. If this quantity is close to zero, it means that there is no correlation between the size of the predicted value and the magnitude of the residual. This indicates that the variances are equal.

Positive values mean that the magnitude of the residuals increases as the predicted values increase. This indicates that the variance increases as the mean increases. Negative values indicate that the variance decreases as the mean increases.

## Evaluate Model Fit

- Use the diagnostic plots available via the ODS Graphics output of PROC REG to evaluate the model fit.
  - plots of residuals and studentized residuals versus predicted values
  - “residual-fit spread” (or R-F) plot
  - plots of the observed values versus the predicted values
  - partial regression leverage plots



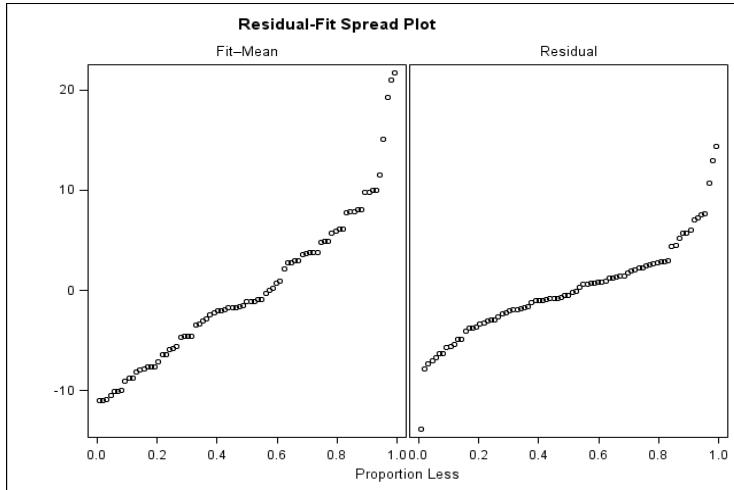
9

Several diagnostic plots are available from PROC REG via the ODS Graphics output. This output helps you assess the model fit. For models that are fit with PROC GLMSELECT, the OUTDESIGN= option can be used to create an input data set for PROC REG.

- Plots of the residuals and studentized residuals (obtained using the RSTUDENT option in SAS) versus the predicted values are useful to visually evaluate the goodness of the model fit. In both of these plots, a random scatter of points about a zero reference line indicates a good model fit. Studentized residuals can also be used to identify influential observations.
- “Residual-Fit Spread” (or R-F) plots compare the quantiles of the fitted values minus their mean to the quantiles of the residuals. If the spread of the residual distribution is considerably smaller than that of the fitted values, the fit explains most of the variation in the data. This plot, together with the adjusted R-square statistic from your model can evaluate the “explanatory power” of the model.
- Plots of the observed values versus the predicted values give a visual tool for examining how close the fitted values are to the observed values.
- Partial leverage plots are an attempt to isolate the effects of a single variable on the residuals. The slope of the linear regression line in the partial regression leverage plot is the regression coefficient for that independent variable in the full model.

 Studentized residuals produced by the RSTUDENT option are externally studentized residuals in which the error variance for the  $i^{\text{th}}$  observation is estimated without including the  $i^{\text{th}}$  observation.

## Residual-Fit Spread Plots

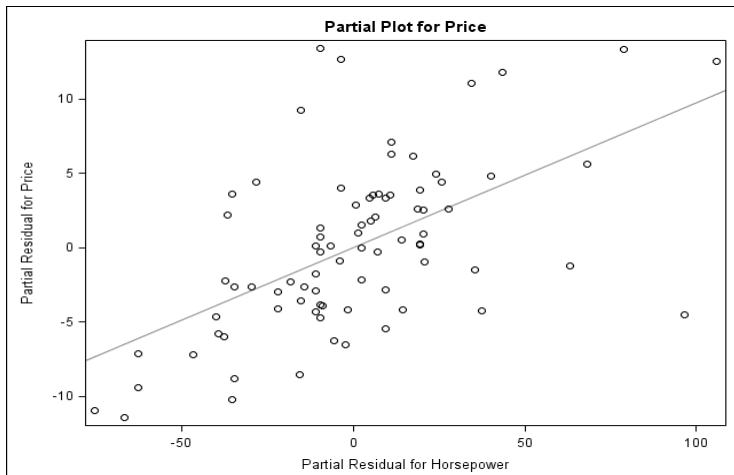


10

The Residual-Fit Spread Plot provides a visual summary of the amount of variability accounted for by a model. It consists of two panels. The left panel shows the quantile plot of the predicted values minus their mean. The right panel is a quantile plot of the residuals. For comparison purposes, the scales are identical on both plots.

The plot (in addition to the R square of your model) is used to evaluate the explanatory power of the model. If the range of the fit-mean plot is substantially larger than the range of the residual plot, then the model explains most of the variability in the data.

## Partial Leverage Plots



11

A partial regression leverage plot is the plot of the residuals for the dependent variable against the residuals for a selected regressor. The residuals for the dependent variable come from regressing the dependent variable on all independent variables, except for the selected regressor. The residuals for the selected regressor are calculated from a model where the selected regressor is regressed on the remaining independent variables. A line fit to the points has a slope equal to the parameter estimate in the full model. (Sall 1990)

Patterns in these plots (for example, curvature) would be evidence of a relationship not accounted for by the full model. These plots are also helpful in detecting outliers and influential points. (Rawlings, Pantula, and Dickey)

## Evaluate Model Fit

- Examine model-fitting statistics such as R square, adjusted R square, AIC/AICC, SBC, and Mallows'  $C_p$ .
- Use the LACKFIT option in the MODEL statement in PROC REG to test for lack of fit for models that have replicates for each value of the combination of the independent variables.

12

In addition to examining diagnostic plots, you can examine several statistics mentioned previously to evaluate model fit. If you have multiple observations (replicates) for each value of the combination of the independent variables, then you can use the LACKFIT option in the MODEL statement in PROC REG to perform a lack-of-fit test for the regression model.

## Evaluate Multicollinearity

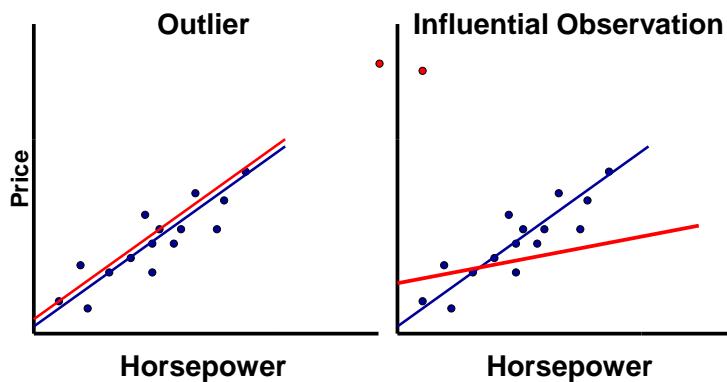
- Correlation statistics (PROC CORR)
- Variance inflation factors (VIF option in the MODEL statement in PROC REG)
- Condition index values (COLLIN and COLLINOUT options in the MODEL statement in PROC REG)

13

The presence of multicollinearity can cause the model to be unstable and can make interpreting the relationships between the dependent and independent variables almost impossible.

As suggested earlier, you can compute the correlation statistics to measure the linear relationship between pairs of the independent variables. The variance inflation factors can help determine the presence of multicollinearity. In addition, the collinearity statistics available in PROC REG are useful in identifying sets of variables involved in the multicollinearity.

## Influential Observations versus Outliers



14

*Outliers* are data points that differ from the general trend of the data by more than is expected. An outlier might be an erroneous data point, or one that is atypical compared with the rest of the data. Influential observations are the ones that affect the model statistics (parameter estimates, standard errors of the parameter estimates, predicted values, studentized residuals, and so on) when they are excluded from the analysis. Outliers might or might not be influential observations, and vice versa.

## Identifying Influential Observations

### RSTUDENT residual

estimates the residual for an observation based on a model fit with that observation deleted.

### Leverage

measures how far an observation is from the cloud of observed data points.

### Cook's D

measures the simultaneous change in the parameter estimates when an observation is deleted.

### DFFITS

measures the change in predicted values when an observation is deleted from the model.

15

Influential observations might be difficult to detect from simple scatter plots in a multiple regression setting. Several statistics are designed to assist in identifying influential observations, and diagnostic plots of these statistics provide a visual method to identify influential observations.

The RSTUDENT option computes the standardized residual for an observation based on a model that is fit without that observation in the data. If the absolute value of the RSTUDENT option is greater than 2, the observation might be influential or an outlier.

The leverage statistic for an observation measures how far the observation is from the centroid of the sample X-space. Observations far from the centroid tend to be influential points. If their leverage values are greater than  $\frac{2p}{n}$ , where  $p$  is the number of model parameters and  $n$  is the sample size, they might be influential.

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis. An observation might have an adverse effect on the analysis if the Cook's D statistic is greater than  $\frac{4}{n}$ .

DFFITS measures the impact that the observation has on the predicted value. If the absolute value of the DFFITS statistic is greater than  $2\sqrt{\frac{p}{n}}$ , the observation is considered influential.

## Details

Consider a regression model with two independent variables,  $X_1$  and  $X_2$ . The vector of observed values  $\mathbf{Y}$  is not in the plane defined by the vectors  $X_1$  and  $X_2$ . The perpendicular projection from  $\mathbf{Y}$  onto the plane of  $X_1$  and  $X_2$  defines the vector  $\hat{\mathbf{Y}}$  of predicted values. The projection matrix, also called the *hat matrix*, is  $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . The leverage values are the diagonal elements of the projection matrix.

### Identifying Influential Observations: DFBETAs

$$DFBETA_{j(i)} = \frac{b_j - b_{j(i)}}{\hat{\sigma}(b_j)}$$

measures the change in each parameter estimate when an observation is deleted from the model.

- $b_j$  is the parameter estimate for the  $j^{\text{th}}$  independent variable.
- $b_{j(i)}$  is the parameter estimate for the  $j^{\text{th}}$  independent variable with the  $i^{\text{th}}$  observation deleted from the analysis.
- $\hat{\sigma}(b_j)$  is the standard error of the  $j^{\text{th}}$  parameter estimate when all observations are included in the analysis.

16

DFBETAs measure the change in the parameter estimates when an observation is deleted from the analysis. Separate DFBETAs are calculated for each observation for each variable. Therefore, the statistic assesses the impact of the individual observation on the parameter estimate for a particular variable. The suggested cutoff is when the absolute value of the DFBETAs is greater than  $\frac{2}{\sqrt{n}}$ .

## Identifying Influential Observations: The Covariance Ratio

$$COVRATIO_i = \frac{|s_i^2 (X'_i X_i)^{-1}|}{|s^2 (X X)^{-1}|}$$

measures the change in the precision of the parameter estimates when an observation is deleted from the model.

17

The *covariance ratio* reflects the impact of the  $i^{\text{th}}$  observation on the precision of the regression estimates. It is computed as the ratio of the determinants of two covariance matrices. The covariance ratio measures the impact of deleting an observation on the estimated variance-covariance matrix of the parameter estimates.

Values of the covariance ratio

- greater than 1 indicate that the presence of the  $i^{\text{th}}$  observation increases the precision of the estimates
- less than 1 indicate that the presence of the  $i^{\text{th}}$  observation decreases the precision of the estimates
- near 1 indicate that the  $i^{\text{th}}$  observation has little effect on the precision of the estimates.

Suggested cutoff values for the covariance ratio are given by the following:

$$COVRATIO_i < 1 - \frac{3p}{n}$$

$$COVRATIO_i > 1 + \frac{3p}{n}$$

## Identifying Influential Observations – Summary of Suggested Cutoffs

Influential Statistics	Cutoff Values
RSTUDENT Residuals	$ RSTUDENT  > 2$
LEVERAGE	$LEVERAGE > \frac{2p}{n}$
Cook's D	$CooksD > \frac{4}{n}$
DFFITS	$ DFFITS  > 2\sqrt{\frac{p}{n}}$
DFBETAS	$ DFBETAS  > \frac{2}{\sqrt{n}}$
COVRATIO	$COVRATIO < 1 - \frac{3p}{n}$ or $COVRATIO > 1 + \frac{3p}{n}$

$p$  = number of model parameters,  $n$  = sample size

18

Identifying observations that fall outside the range of these cutoffs can be easily accomplished by writing the statistics and using a screening program to identify the influential observations, or by examining diagnostic plots. The ODS Graphics output from PROC REG provides plots of values of RSTUDENT, LEVERAGE, Cook's D, DFFITS, DFBETAS, and COVRATIO, as well as plots of the studentized residuals plotted against the predicted values and against the leverage statistics.

### 2.01 Multiple Choice Poll

Regression model diagnostics might include checking for which of the following?

- a. model fit
- b. model assumptions (independent normal errors with constant variance)
- c. multicollinearity
- d. influential observations
- e. a and b
- f. all of the above.

19



## Model Diagnostics – Normality, Constant Variance, Model Fit, Collinearity, and Influential Observations

Presume that you chose to use the model with three variables, **Hwympg**, **Hwympg<sup>^2</sup>**, and **Horsepower**. Evaluate this model by checking for violations of the assumptions, model fit, and collinearity. Also identify any observations that appear to be influential.

```
ods graphics / imagemap=on;
ods html style=analysis;

proc glmselect data=STAT2.cars outdesign(addinputvars)=d_carfinal;
  effect q_hwympg = polynomial(hwympg / degree=2
                                standardize(method=moments)=center);
  model price = q_hwympg horsepower / selection=none;
run;

proc reg data=d_carfinal plots(unpack label)=all;
  model price = &_GLSMOD
    / vif collin collinoint influence spec partial;
  id model;
  output out=check r=residual p=pred rstudent=rstudent h=leverage;
run;
quit;                                              ST202d01.sas;
```

Selected REG procedure statements:

- |        |  |
|--------|--|
| OUTPUT | creates a new SAS data set that saves diagnostic measures and residuals that are calculated after fitting the model.   |
| ID     | lists variables to use to identify each observation when one of the MODEL statement options (CLI, CLM, P, R, or INFLUENCE) is requested. These variables can be used to identify each observation. If the ID statement is omitted, the observation number is used to identify the observations. Although there are no restrictions on the length of ID variables, PROC REG might truncate ID values to 16 characters for display purposes. |

Selected MODEL statement options:

- |           |   |
|-----------|---|
| INFLUENCE | requests that a detailed analysis of the influence of each observation on the estimates and the predicted values be printed.  |
| SPEC      | performs a test so that the first and second moments of the model are correctly specified. The null hypothesis for this test maintains that the errors are homoscedastic, independent of the regressors, and that several technical assumptions about the model specification are valid. For details, see theorem 2 and assumptions 1 through 7 of White (1980). When the model is correctly specified and the errors are independent of the regressors, the rejection of this null hypothesis is evidence of heteroscedasticity. |

**PARTIAL** requests partial regression leverage plots for each regressor. You can use the PARTIALDATA option to obtain a tabular display of the partial regression leverage data. If ODS Graphics is in effect (see the "ODS Graphics" section in the SAS®9 online documentation), then these partial plots are produced in panels with up to six plots per panel. See the "Influence Statistics" section in the SAS®9 online documentation for more information.

Partial PROC REG Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	3	4448.69341	1482.89780	65.73	<.0001
<b>Error</b>	77	1737.20536	22.56111		
<b>Corrected Total</b>	80	6185.89877			

Root MSE	4.74985	R-Square	0.7192
Dependent Mean	18.64321	Adj R-Sq	0.7082
Coeff Var	25.47766		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
<b>Intercept</b>	Intercept	1	4.03949	2.17024	1.86	0.0665	0
<b>s_Hwympg</b>	s_Hwympg	1	-0.80407	0.21378	-3.76	0.0003	4.06937
<b>s_Hwympg_2</b>	s_Hwympg^2	1	0.04350	0.01430	3.04	0.0032	2.26764
<b>Horsepower</b>	Horsepower	1	0.09730	0.01614	6.03	<.0001	2.36905

As expected, the overall *F* test is significant and the adjusted R square is 0.7082. The model equation is as follows:

$$\text{Price} = 4.03949 - 0.80407 * \text{Hwympg} + 0.0435 * \text{Hwympg}^2 + 0.09730 * \text{Horsepower}$$

The largest variance inflation factor is 4.06937, which is less than 10.

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	s_Hwympg	s_Hwympg_2	Horsepower
1	2.17779	1.00000	0.01125	0.00133	0.03345	0.00845
2	1.52787	1.19389	0.00125	0.09241	0.06816	0.00352
3	0.26890	2.84585	0.03127	0.32286	0.68972	0.00003866
4	0.02544	9.25245	0.95623	0.58341	0.20867	0.98800

Collinearity Diagnostics (intercept adjusted)					
Number	Eigenvalue	Condition Index	Proportion of Variation		
			s_Hwympg	s_Hwympg_2	Horsepower
1	2.06018	1.00000	0.05330	0.05815	0.05768
2	0.79624	1.60854	0.00001184	0.28493	0.25738
3	0.14358	3.78796	0.94669	0.65693	0.68494

None of the condition index values is greater than 10. Multicollinearity does not appear to be a problem with this model.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
8	16.49	0.0359

The SPEC option performs a test so that the first and second moments of the model are correctly specified. The null hypothesis for this test includes the following:

- The errors are homoscedastic.
- The errors are independent of the regressors.
- Several technical assumptions about the model specification are valid. For example, the correct model is specified.

For details, see theorem 2 and assumptions 1 through 7 of White (1980).

However, the following warning message appears in the Log window.

WARNING: The average covariance matrix for the SPEC test has been deemed singular, which violates an assumption of the test. Use caution when interpreting the results of the test.

This warning message tends to appear when you have dummy variables or higher-ordered terms in the model. You might consider an alternative to evaluate constant variance assumption, such as computing the Spearman rank correlation coefficient between the absolute values of the residuals and the predicted values.

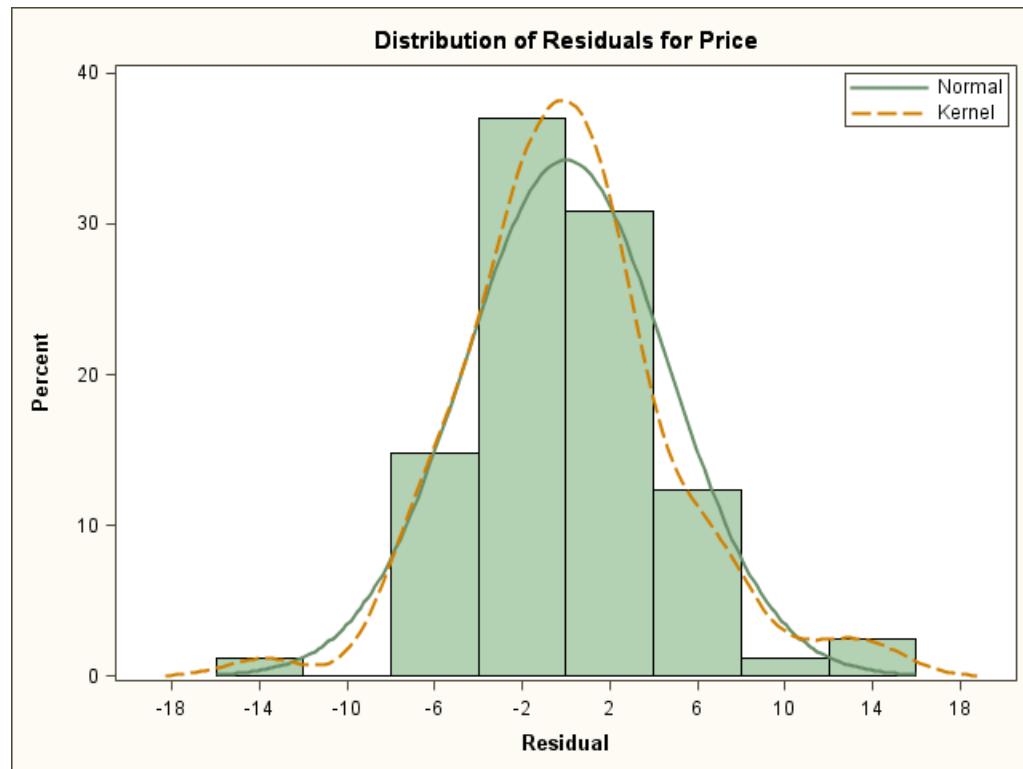
-  The SPEC test is a simultaneous test of multiple null hypotheses. If you reject the test, then you cannot identify, without further investigation, which particular null hypothesis caused the rejection.

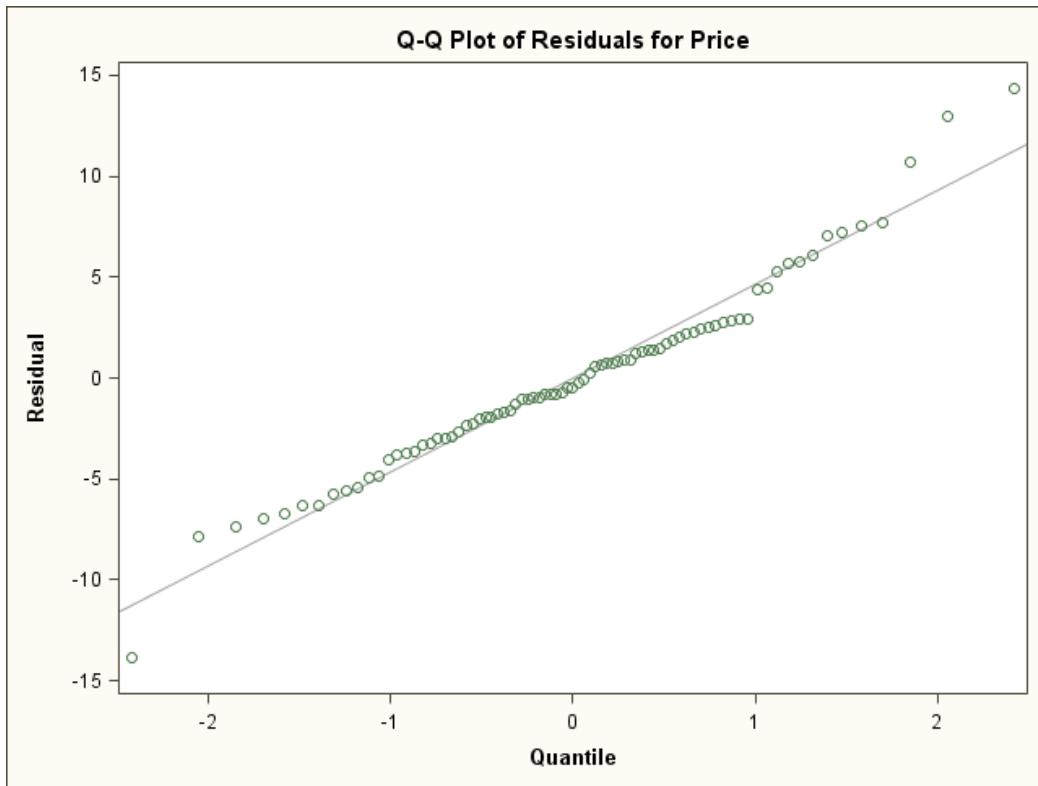
## Partial INFLUENCE Option Output

Obs	Model	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS			
							Intercept	s_Hwympg	s_Hwympg_2	Horsepower
1	Integra	-1.0276	-0.2177	0.0244	1.0774	-0.0344	0.0059	-0.0218	0.0232	-0.0151
2	Legend	5.2465	1.1274	0.0367	1.0236	0.2199	-0.0266	-0.0837	0.0843	0.0424
3	100	12.9697	2.8930	0.0239	0.7108	0.4527	0.0885	-0.2313	0.1656	-0.0353
4	90	4.3697	0.9304	0.0239	1.0317	0.1456	0.0285	-0.0744	0.0533	-0.0113
5	535i	5.6921	1.2598	0.0882	1.0639	0.3917	-0.2901	0.3016	-0.2385	0.3528
6	Century	1.6914	0.3578	0.0208	1.0688	0.0522	0.0257	0.0081	-0.0212	-0.0120
7	LeSabre	-1.5990	-0.3379	0.0189	1.0675	-0.0469	0.0114	-0.0112	0.0147	-0.0223

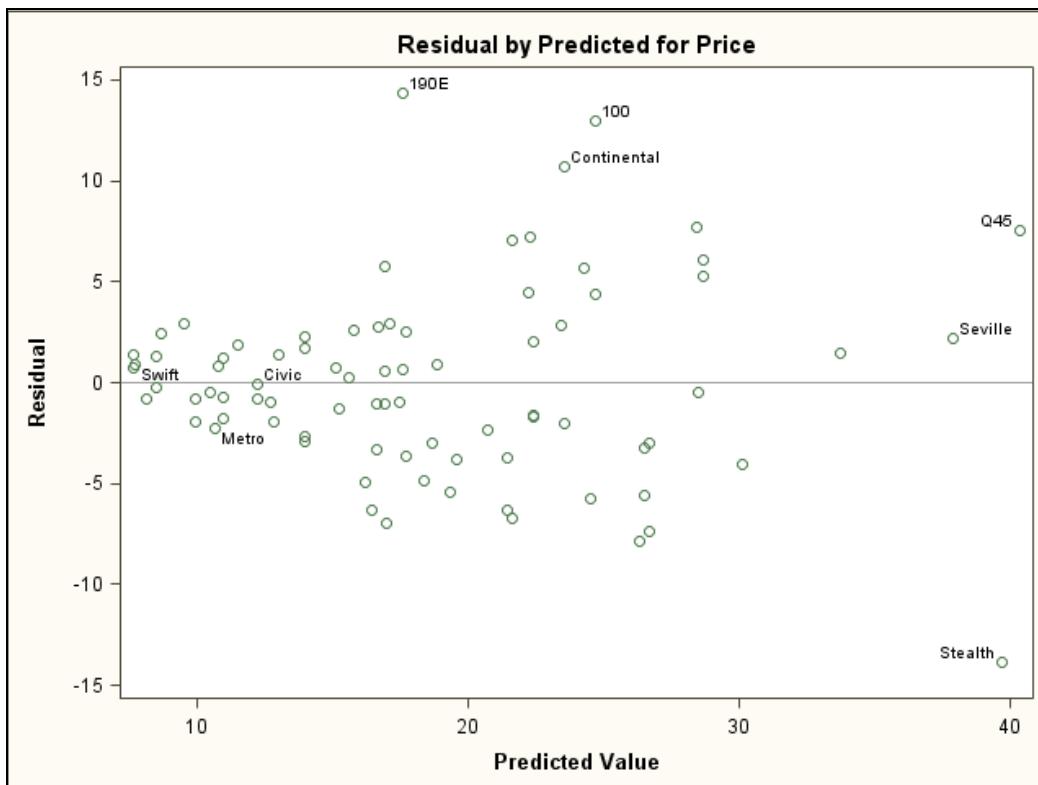
Influential observations can be identified easily by examining the plots that are produced with the INFLUENCE option in the MODEL statement. They can also be identified by comparing the influential statistics with the corresponding cutoff values. By using the ID statement in PROC REG, the value of the variable **Model** is used to identify influential observations instead of the observation number.

Examine the graphs of the residuals to evaluate the normality assumption.





The histogram of the residuals and the normal quantile plot do not indicate any serious problems with the normality assumption. To further evaluate the normality assumption, you can use the output data set and generate a test of normality using the UNIVARIATE procedure.



The plot of the residuals versus the predicted values might cause some concern about the model. The smaller predicted values appear to have less variability than the larger predicted values. This is a violation of the assumption of equal variances. (Remedial measures are discussed later in the course.) Because the ID statement was used and the LABEL option was specified in the PLOTS= option of the PROC REG statement, points identified as outliers or influential are labeled with their **Model** (in the plots that support the LABEL option). The Mercedes-Benz 190E and the Dodge Stealth have the most extreme outliers.

The constant variance assumption can be evaluated with the Spearman rank correlation coefficient between the absolute values of the residuals and the predicted values.

```
data check;
  set check;
  abserror=abs(residual);
run;

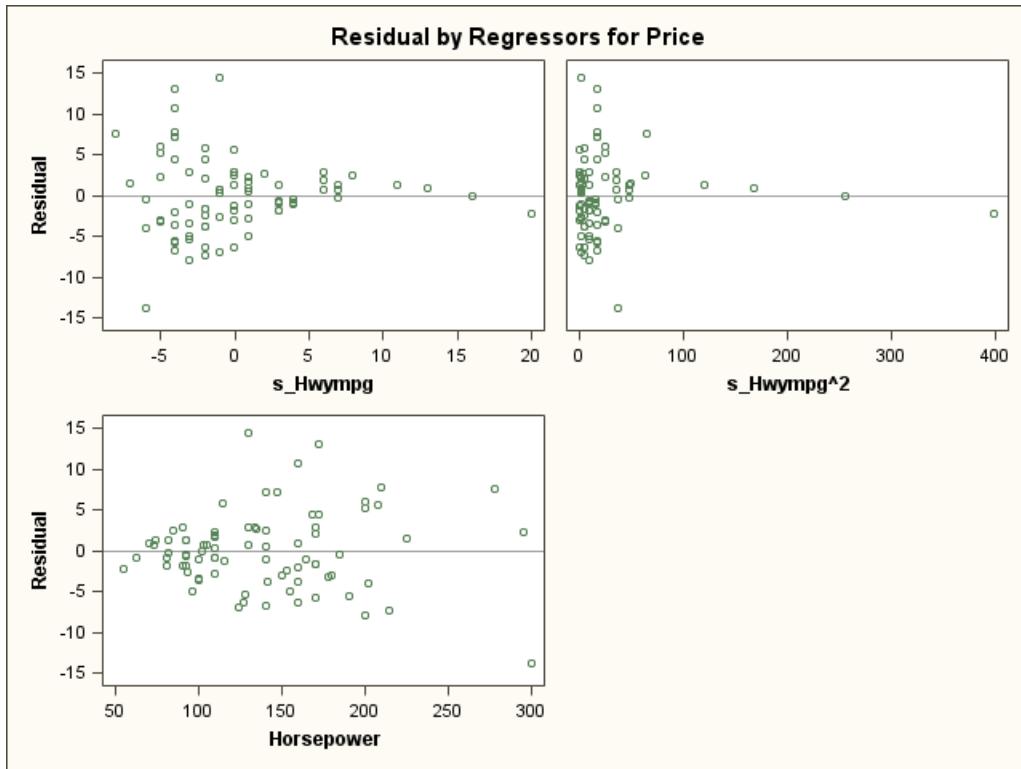
proc corr data=check spearman nosimple;
  var abserror pred;
run;                                         *ST202d01.sas;
```

PROC CORR Output

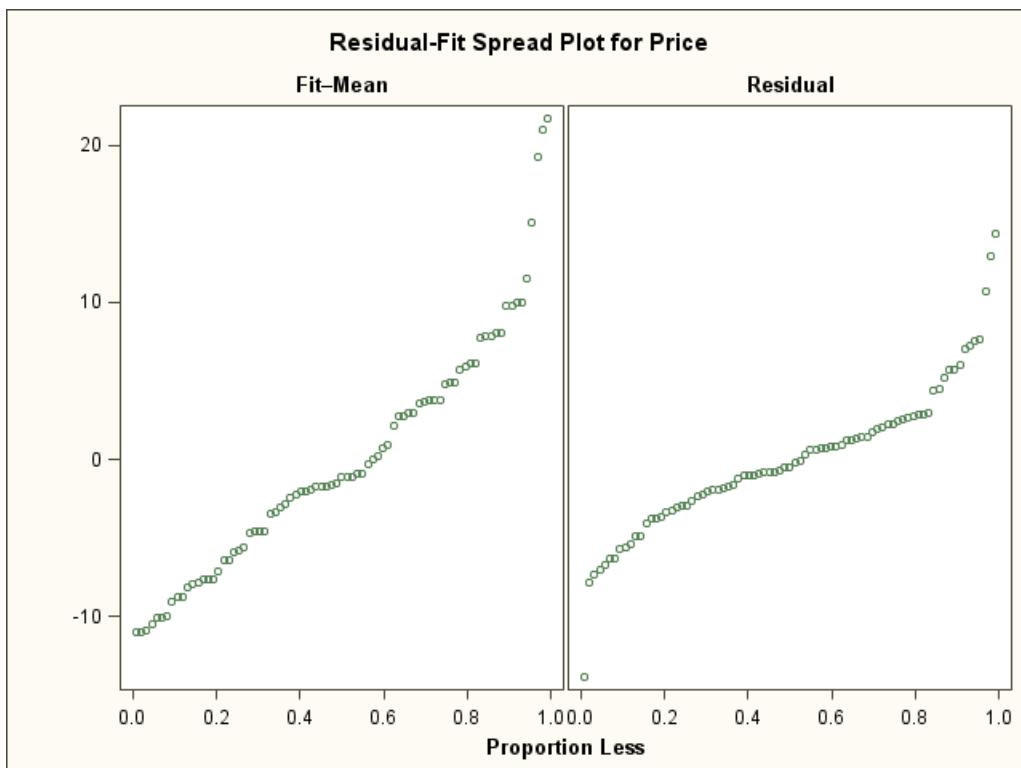
Spearman Correlation Coefficients, N = 81 Prob >  r  under H0: Rho=0		
	abserror	pred
abserror	1.00000	0.60274 <.0001
pred Predicted Value of Price	0.60274 <.0001	1.00000

The Spearman rank correlation coefficient between the absolute values of the residuals and the predicted values is about 0.603. The highly significant *p*-value (<.0001) indicates a strong correlation between the absolute values of the residuals and the predicted values. The positive correlation coefficient indicates that the residuals increase as the predicted values increase.

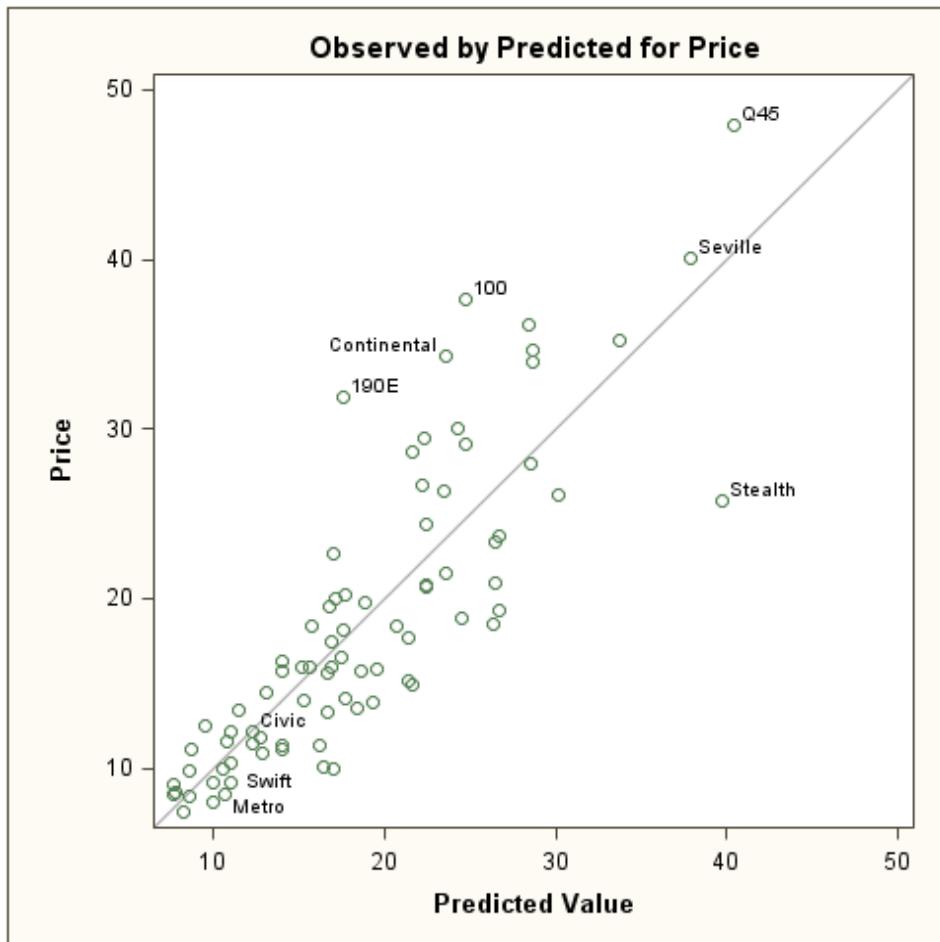
To assess the model fit, examine plots of residuals versus the independent variables, the R-F plot, and the plot of the observed values versus the predicted values.



The plots of the residuals by **s\_Hwympg** and **s\_Hwympg<sup>2</sup>** show a linear pattern for large values of the predictor variable. This indicates that the model does not fit well for cars with higher gas efficiency. These two plots also confirm the nonconstant variance. Evidence of the nonconstant variance might also be seen in the plot of the residuals by **Horsepower**.

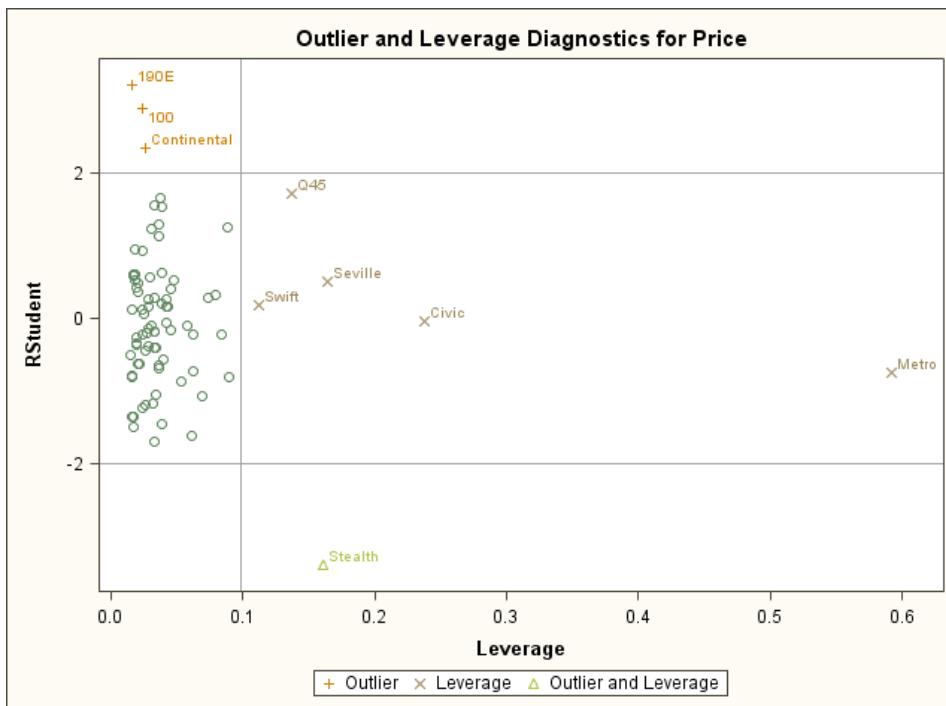


The R-F plot indicates that the Fit-Mean distribution ranges from about -10 to 25 for a range of approximately 35. The Residual distribution ranges from about -10 to 15 for a range of approximately 25. Because the spread of the residual distribution is smaller than that of the fitted values, and the model has an R square of 0.7192, these indicate that the model explains most of the variation in the data.

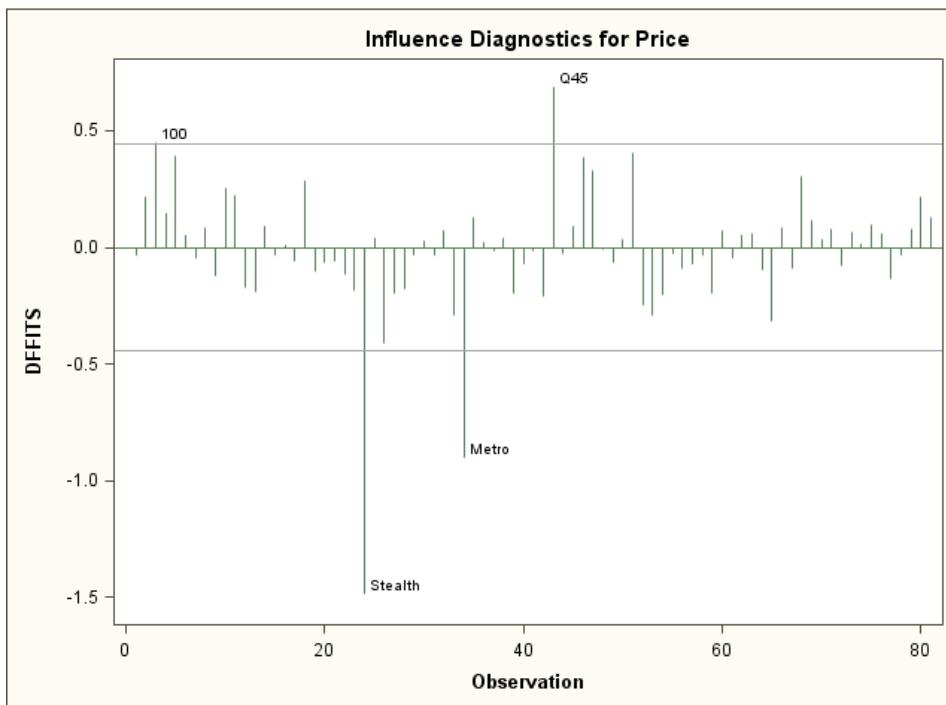


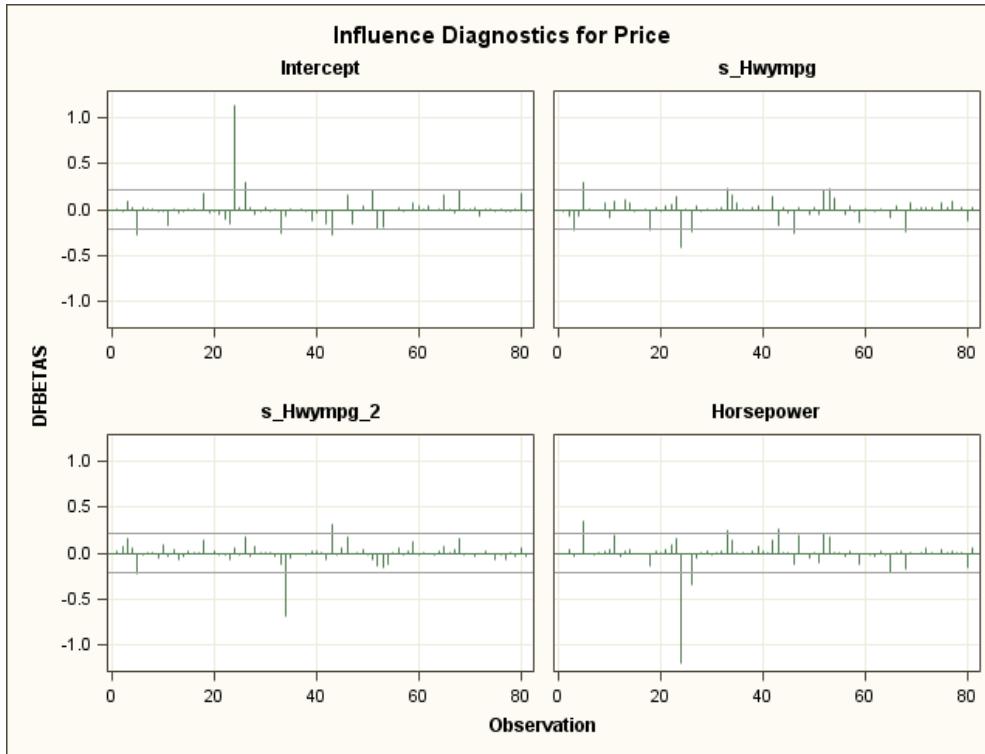
The plot of the observed values by the predicted values suggests that the model fits the data better for the cars with lower predicted values. As the predicted value increases, the variability around the 45-degree reference line increases. As was previously seen in the plot of the residuals versus the predicted values, the Dodge Stealth, Mercedes-Benz 190E, and the Audi 100 lie the farthest from the fitted line.

The output from the INFLUENCE option can be voluminous, so the following plots can help you easily identify influential observations (plots of the Cook's D, DFFITS, and DFBETA statistics, as well as a plot of RSTUDENT versus LEVERAGE). Because each of these statistics has a recommended cutoff value, observations lying beyond those cutoffs are labeled in the output. Partial regression plots can also be used to identify outlying points that might exert influence on the model.

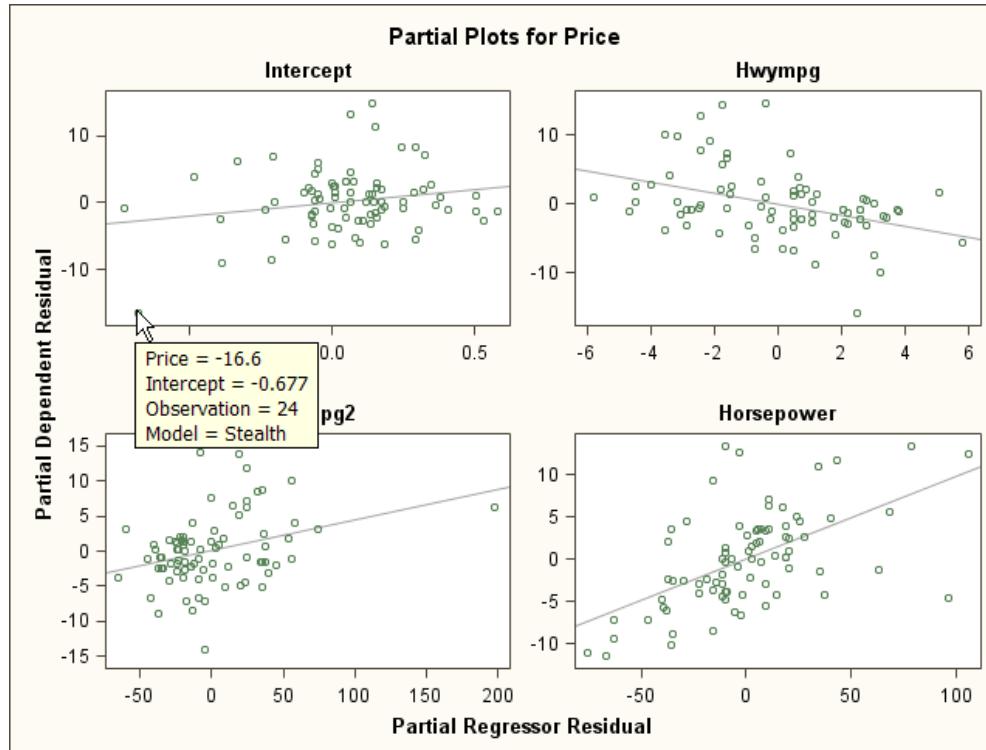


The plot of RSTUDENT versus LEVERAGE indicates that the Mercedes-Benz 190E, the Audi 100, and the Lincoln Continental are beyond the cutoff recommended for RSTUDENT. This indicates that these points might be influential. The Infiniti Q45, Cadillac Seville, Suzuki Swift, and Honda Civic are beyond the cutoff for the LEVERAGE statistic and might be influential. The Geo Metro is not only beyond the cutoff for the LEVERAGE statistic, but is also much larger than the other potentially influential points. It might warrant further investigation. The Dodge Stealth was flagged in the earlier plots and appears again. It is beyond the cutoff values of both RSTUDENT and LEVERAGE.





The plots of Cooks' D, DFITTS, and DFBETAS flag several points as influential. Most of these points were previously flagged as either outliers or influential points, with the exception of the Eagle Vision, the Ford Tempo, the Chrysler Imperial, the Mercury Cougar, the Geo Metro, the Saab 900, and the BMW 535i. The Dodge Stealth exerts a positive influence on the intercept and a negative influence on the parameter estimates for **Hwympg** and **Horsepower**.



Partial leverage plots isolate the effects of a single variable on the residuals and can detect outliers and influential points. Although these plots do not flag observations of interest, when tooltips are turned on (via the IMAGEMAP= option of the ODS GRAPHICS statement) points of interest are identified as you move the mouse pointer over the points. The Dodge Stealth appears as outlying and potentially influential in all four of the partial plots. The most extreme data point in the partial regression plot for **hwympg2** is the Geo Metro.

When you use the cutoff values of RSTUDENT and LEVERAGE, the following program requests a printout of outliers and influential data points as identified by the plot of RSTUDENT versus LEVERAGE.

```
/*set the values of the macro variables based on your data and model*/
%let numparms = 4;
%let numobs = 81;
data influence;
  set check;
  absrstud=abs(rstudent);
  if absrstud ge 2 then output;
  else if leverage ge (2*&numparms /&numobs) then output;
run;
proc print data=influence;
  var manufacturer model price hwympg horsepower;
run; *ST202d02.sas;
```

PROC PRINT Output

Obs	Manufacturer	Model	Price	Hwympg	Horsepower
1	Audi	100	37.7	-4.0370	172
2	Cadillac	Seville	40.1	-5.0370	295
3	Dodge	Stealth	25.8	-6.0370	300
4	Geo	Metro	8.4	19.9630	55
5	Honda	Civic	12.1	15.9630	102
6	Infiniti	Q45	47.9	-8.0370	278
7	Lincoln	Continental	34.3	-4.0370	160
8	Mercedes-Benz	190E	31.9	-1.0370	130
9	Suzuki	Swift	8.6	12.9630	70

There are nine observations that appear to be influential based on LEVERAGE or RSTUDENT statistics. Notice that all of the observations were flagged by at least two, and some by three or four, influence statistics. The data for these observations should be checked to ensure that no transcription or data entry errors occurred. If the data are erroneous, correct the errors and re-analyze the data.

It is possible that the model is not adequate. Notice that most of these cars fall at the high or low end of the range for **Price**. There might be another variable, such as one indicating whether the car is a luxury, midrange, or economy car that would be important in explaining these unusual observations.

Another possibility is that the observation, though valid, might be unusual. If you had a larger sample size, there might be more observations like the unusual ones. You might need 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, you should include a description of the types of observations that you exclude and provide an explanation. You should also discuss the limitations of your conclusions, given the exclusions, as part of your report or presentation.

## 2.02 Multiple Choice Poll

In the previous demonstration, which of the following did you discover?

- a. The residuals appear to be normally distributed.
- b. The residual variances might not be constant.
- c. There does not seem to be apparent multicollinearity.
- d. There can be some influential observations.
- e. all of the above
- f. none of the above



## Exercises

---

### 1. Performing Model Diagnostics on the Revised Model

In Chapter 1, the model selected with **LogPrice** as the dependent variable resulted in different predictor variables than the model with **Price** as the dependent variable. Perform model diagnostics on the revised model. Use **LogPrice** as the dependent variable, and **Citympg**, **Citympg<sup>2</sup>**, **EngineSize**, **Horsepower**, **Horsepower<sup>2</sup>**, and **Weight** as the independent variables. Use the **STAT2.cars4** data set to fit the model in PROC GLMSELECT. Include the OUTDESIGN= option with ADDINPUTVARS to create a data set for doing the diagnostics in PROC REG.

- a. Use the appropriate options in the MODEL statement of PROC REG to evaluate multicollinearity, influential observations, and the constant variance assumption. Generate the necessary plots to check the assumptions of regression. Is multicollinearity a problem for this model? Does this model meet the assumptions for linear regression?
- b. To assess the model fit, examine the R-F plot, the plot of the observed values versus the predicted value, and the plots of residuals versus the independent variables. What are your conclusions?
- c. Examine the plots of the Cook's D, DFFITS, and DFBETA statistics, the plot of RSTUDENT versus LEVERAGE, and the partial leverage plots to identify any outlying or influential observations. Output the potentially influential observations to a data set. What are your conclusions?

### Advanced

- d. Use the output data set containing the influence statistics that you created in the previous exercise. Use PROC SGPlot to create a scatter plot of the residuals versus the predicted values. Look in the SGPlot documentation for how to use the GROUP option. Plot the residuals versus the predicted values. Use the variable **Manufacturer** as the grouping variable. Do you see any patterns?

### 2.03 Multiple Choice Poll

What did you find out about the model diagnostics in the previous exercise?

- a. The residuals are normally distributed.
- b. There seems to be apparent multicollinearity.
- c. The error variance might not be constant.
- d. a and c
- e. all of the above

25

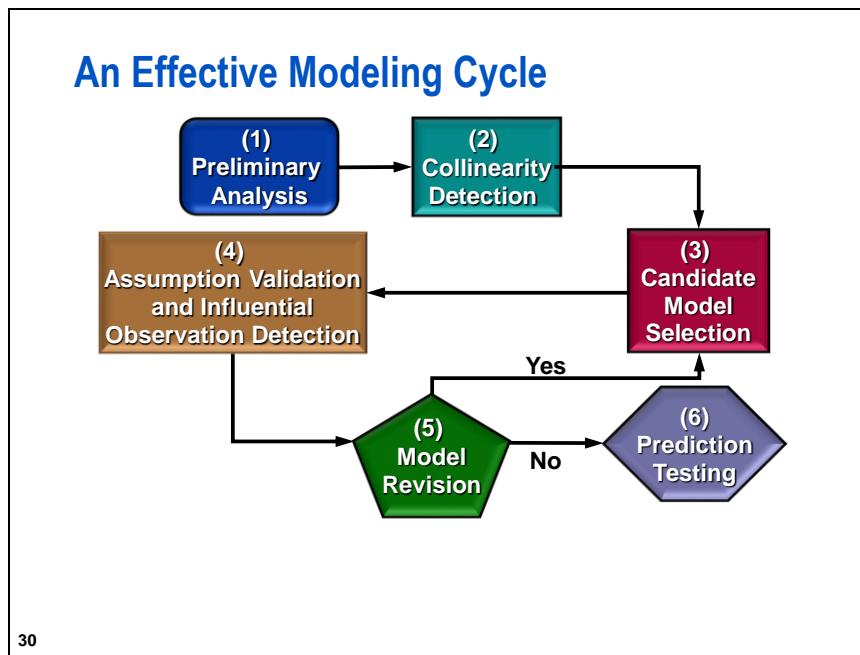
## 2.2 Remedial Measures

---

### Objectives

- List remedial measures for violation of model assumptions.
- Use PROC GLIMMIX to model a response with nonconstant variance.

29



- (1) **Preliminary Analysis:** This step includes the use of descriptive statistics, graphs, and correlation analysis to identify those variables that might be useful in the regression model.
- (2) **Collinearity Detection:** The presence of multicollinearity among the variables that you identify in step (1) can be detected by the use of the VIF statistic, condition indices, and variation proportions.
- (3) **Candidate Model Selection:** This step uses the information gathered from the exploratory data analysis and numerous model selection options in PROC GLMSELECT to identify one or more candidate models. Potential models can be evaluated by comparing the adjusted coefficients of determination, Mallows'  $C_p$  statistic, and information criteria statistics. You can also use PROC REG to produce the plot of residuals versus the predicted values, plots of the residuals versus the regressors, and the R-F spread plot to assess the model fit.
- (4) **Assumption Validation and Influential Observation Detection:** This step includes examination of graphs of the residuals versus the predicted values. It also includes tests for normality of the residuals, constant variance, and independent observations. Influential observations can be detected by examining plots of RSTUDENT residuals, Cook's D statistics, DFFITS statistics, DFBETAS statistics, covariance ratio statistics, leverage statistics, and partial leverage plots.
- (5) **Model Revision:** If step (4) indicates the need for model revision, generate a new model that is more appropriate. Based on the nature of the refinement, you might need to return to step (3) to identify new candidate models.
- (6) **Prediction Testing:** This final step (not discussed in this course) is to evaluate the model's predictive capability with data that is not used to build the model. In other words, you build the model with part of your data and use the remainder of the data to determine how well the model fits the data.

## When the Normality Assumption Is Violated

- Transform the dependent variable.
- Use PROC GENMOD or PROC GLIMMIX with the appropriate DIST= and LINK= option to fit a generalized linear model.

31

When the normality assumption is violated, you can transform the dependent variable to normalize the distribution. You might want to use other procedures such as PROC GENMOD or PROC GLIMMIX to fit a model that is appropriate for the distribution exhibited by your data. You must specify the appropriate distribution and link function for the data that you have.

 Information about variable transformations, including sample programs, is included in an appendix.

## When the Constant Variance Assumption Is Violated

- Request tests using the heteroscedasticity-consistent variance estimates.
- Transform the dependent variable.
- Model the nonconstant variance by using the following:
  - PROC GENMOD or PROC GLIMMIX with the appropriate DIST= option
  - PROC MIXED with the GROUP= option and TYPE= option
  - SAS SURVEY procedures for survey data
  - SAS/ETS procedures for time series data
  - weighted least squares regression model

32

If your data seems to show heteroscedasticity (nonconstant variance), you can request tests using both the usual covariance matrix and the heteroscedasticity-consistent covariance matrix when you specify the SPEC, ACOV, HCC, or WHITE option in the MODEL statement of PROC REG. You can also transform the dependent variable to stabilize the variance, or use different tools to model the nonconstant variance. These can include the following:

- PROC GENMOD or PROC GLIMMIX with the appropriate DIST= option
- PROC MIXED with the GROUP= option or the power-of-mean models
- PROC SURVEYREG for survey data
- procedures provided in SAS/ETS for time series data
- a weighted least squares regression model

 Information about weighted least squares, including a sample program, is included in an appendix.

## When the Independence Assumption Is Violated

Use the appropriate modeling tools to account for correlated observations.

- PROC MIXED, PROC GENMOD, or PROC GLIMMIX for repeated measures data
- PROC AUTOREG or PROC ARIMA in SAS/ETS for time series data
- PROC SURVEYREG for survey data

33

One of the assumptions in linear regression is independent errors. Error terms are correlated if the values of the errors depend on the other values of errors. They often arise from time series data, repeated measures on a given subject, data gathered from a nested design, or from a complex survey design.

Correlated errors can affect the standard errors of the parameter estimates, and therefore, affect the confidence intervals and the significance tests for the parameters.

You should use other tools to model data with correlated errors.

- procedures in SAS/ETS, such as PROC AUTOREG or PROC ARIMA to model time series data
- PROC MIXED, PROC GENMOD, or PROC GLIMMIX to model the correlations arising with data that have repeated measures from each subject
- the SAS SURVEY procedures to model data gathered from a complex sample design

 A section about analyzing data gathered over time is included in an appendix.

## When a Straight Line Is Inappropriate

- Fit a regression model with polynomial effects or splines.
- Transform the independent variables to obtain linearity.
- Fit a nonlinear regression model using PROC NLIN if appropriate.
- Fit a nonparametric regression model using PROC LOESS.

34

When the relationship between the dependent variable and one or more predictor variables does not follow a linear relationship, you might consider transforming the predictor variables to obtain the linearity (Neter, Wasserman, and Kutner 1990). Sometimes a regression model that includes polynomial terms or spline effects might be a better solution.

For applications in physical fields such as pharmacokinetics, chemistry, and biology, the relationship might not be linear in terms of parameters. In that case, you might want to fit a nonlinear regression model using PROC NLIN. When the parametric form of the relationship is difficult or impossible to define, you might want to fit a nonparametric regression model using PROC LOESS.

 PROC NLIN and PROC LOESS are discussed in an appendix.

## When There Is Multicollinearity

- Exclude redundant independent variables.
- Redefine independent variables.
- Use biased regression techniques such as ridge regression or principal component regression.
- Center the independent variables in polynomial regression models.

35

Biased estimations produced by ridge regression or (incomplete) principal component regression are alternatives to ordinary least squares estimation. Ordinary least squares estimators provide unbiased estimates of parameters, and the estimates have minimum variance among all unbiased estimators. In the presence of multicollinearity, the minimum variance of the parameter estimates might be unacceptably large. Therefore, it might be better to use a biased estimator that has a smaller variance.

The basic idea behind ridge regression is to reduce the variances of the parameter estimates by considering a matrix  $\mathbf{X}'\mathbf{X}+k\mathbf{I}$ , where  $k$  is usually a small positive quantity sometimes referred to as a *shrinkage parameter*. The choice of  $k$  is a compromise between decreasing variance and increasing bias. The most popular method is to compute ridge regression estimates for a set of values of  $k$  starting with  $k=0$  (the unbiased estimate). In many applications, a plot of the coefficients against  $k$  (also called *ridge traces*) enables you to choose the value of  $k$  where most of the changes in the parameter estimates were realized. You use the RIDGE= option in the MODEL statement in PROC REG to specify the range of values of  $k$ . You use the PLOT or RIDGE PLOT statement in PROC REG to obtain the ridge traces. The selection of  $k$  can be subjective. Although more objective procedures were proposed, none gained universal acceptance. The parameter estimates and other statistics corresponding to the chosen  $k$  are the results of the ridge regression model fit to the data.

Principal component regression is another biased regression technique. With principal component regression, you drop linear combinations of independent variables (principal components) from the model instead of dropping individual variables. Each principal component is a linear combination of the predictor variables. The first principal component has the largest variance of any linear function of the original variables (subject to a scaling constant). The second component has the second largest variance, and so on. You can use the PCOMIT=list option in the MODEL statement in PROC REG to compute parameter estimates using all but the last  $m$  principal components. (The value  $m$  is one of the values specified in the *list*.) The principal component variables are jointly uncorrelated, and therefore, eliminate the multicollinearity problem. The interpretation of the resulting model might be difficult.

## When There Are Influential Observations

- Make sure that there are no data errors.
- Perform a sensitivity analysis and report results from different scenarios.
- Investigate the cause of the influential observations and redefine the model if appropriate.
- Delete the influential observations if appropriate and document the situation.
- Limit the influence of outliers by performing robust regression analysis using PROC ROBUSTREG.

36

Sensitivity analysis refers to performing the analysis for different scenarios and comparing the results. For example, you might perform the analysis for data with and without the influential observations, and evaluate how the results are affected by the inclusion and exclusion of the influential observations.

You can use PROC ROBUSTREG to perform a regression analysis that is robust to outliers. The main purpose of robust regression is to detect outliers and provide resistant (stable) results in the presence of outliers.

## The Lognormal Distribution

- The lognormal distribution describes a response variable  $Y$  having the property that  $\log(Y)$  follows a normal distribution.

$$\log(Y) \sim N(\mu, \sigma^2)$$

$$E[Y] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$Var[Y] = (e^{\sigma^2} - 1)(E[Y])^2$$

- The variance of  $Y$  is proportional to the square of the mean, so a lognormal model explicitly accounts for nonconstant variance.

37

When nonconstant variance is evident, it might be appropriate to consider modeling the response using a probability distribution that can accommodate the heteroscedasticity. An appropriate distribution can be chosen based on theoretical knowledge, previous research or both, or by examining the nature of the relationship between the mean and the variance of the residuals.

One distribution that is often used to analyze cost or price data is the lognormal distribution. A variable is said to follow this distribution when its logarithm follows a normal distribution. A lognormal random variable has the property that its variance is proportional to the square of its mean.

## GLIMMIX Procedure

General form of the GLIMMIX procedure:

```
PROC GLIMMIX <options>;
  EFFECT effect-name=effect-type<(effect options)>;
  MODEL response<(response options)>=<fixed-effects>
    < DIST=keyword options>;
  OUTPUT OUT=SAS-data-set keyword=names;
RUN;
```

38

A generalized linear model is one that enables the distribution of the response variable to be something other than the normal distribution. These models can be fit in SAS using PROC GENMOD or PROC GLIMMIX. The lognormal distribution, however, is available only in PROC GLIMMIX.

 Generalized linear models are discussed in more depth in Chapter 5.

The syntax for the GLIMMIX procedure is similar to that used in other SAS modeling procedures, such as PROC GLMSELECT.

## Interpretation of the Lognormal Model

- Parameters and standard errors estimated by PROC GLIMMIX are on the log scale.
- Applying the inverse log (exponential) function to the predicted values from the lognormal model ( $X\hat{\beta}$ ) produces unbiased estimates of the **median** of the original data rather than the **mean**.
- The formula for the mean of a lognormal distribution provides the correct low-bias back-transformation:

$$E[Y] \approx \exp\left(X\hat{\beta} + \frac{\hat{\sigma}^2}{2}\right)$$

39

To fit a lognormal model, PROC GLIMMIX applies a log transformation to the response variable and models that transformed response using a normal distribution. As a result, the parameter estimates and standard errors reported by PROC GLIMMIX are on the log-transformed scale. It is usually preferable to obtain predicted means on the original scale of the data, so back-transformation is necessary.

The formula for the mean of a lognormal distribution, shown previously, provides the appropriate reverse transformation that can be used with the log-scale predicted values produced by PROC GLIMMIX.



## Fitting a Lognormal Regression Model

Use PROC GLIMMIX to fit a lognormal regression model. Use the three predictor variables chosen previously and create an output data set with the predicted values and residuals.

```
title 'Lognormal model for CARS dataset';

ods output ParameterEstimates=params;
proc glimmix data=STAT2.cars;
  effect q_hwympg = polynomial(hwympg / degree=2
                                standardize(method=moments)=center);
  model price = q_hwympg horsepower / dist=lognormal solution;
  output out=out pred=pred resid=resid;
  id model price;
run; *ST202d03.sas;
```

Selected MODEL statement options:

DIST= specifies the response variable distribution.

SOLUTION requests a table of parameter estimates and standard errors for the model.

Partial PROC GLIMMIX Output

Model Information	
Data Set	STAT2.CARS
Response Variable	Price
Response Distribution	Lognormal
Link Function	Identity
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Residual

The Model Information table shows that the lognormal was selected as the response distribution for this model. PROC GLIMMIX uses restricted maximum likelihood as the default estimation method for normal or lognormal data.

Fit Statistics	
-2 Res Log Likelihood	27.80
AIC (smaller is better)	37.80
AICC (smaller is better)	38.64
BIC (smaller is better)	49.52
CAIC (smaller is better)	54.52
HQIC (smaller is better)	42.48
Pearson Chi-Square	4.09
Pearson Chi-Square / DF	0.05

The Fit Statistics table provides several information criteria statistics, including the AIC, AICC, and BIC. PROC GLIMMIX computes these statistics differently than PROC GLMSELECT does, so you should not make comparisons on the basis of these criteria between models, which were fit using different procedures.

Parameter Estimates						
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	
Intercept	2.1022	0.1054	77	19.95	<.0001	
s_Hwmpg	-0.04193	0.01038	77	-4.04	0.0001	
s_Hwmpg^2	0.001599	0.000694	77	2.30	0.0240	
Horsepower	0.004907	0.000784	77	6.26	<.0001	
Scale	0.05317	0.008569	.	.	.	.

The results indicate that all three predictors in the model are highly significant. Based on the parameter estimates, you can write the following regression equation:

$$\log(\text{Price}) = 2.1022 - 0.04193 * \text{Hwmpg} + 0.001599 * \text{Hwmpg}^2 + 0.004907 * \text{Horsepower}$$

The **Scale** parameter shown in the table is the estimate of the residual variance for the lognormal model.

Check for homogeneity of variances using the Spearman correlation coefficient on the log-scale data.

```
data check3;
  set out;
  abserror=abs(resid);
run;

proc corr data=check3 spearman nosimple;
  var abserror pred;
run;          *ST202d03.sas;
```

The CORR Procedure		
2 Variables:		abserror pred
<b>Spearman Correlation Coefficients, N = 81</b>		
Prob >  r  under H0: Rho=0		
	abserror	pred
<b>abserror</b>	1.00000 0.0812	0.19492 0.0812
<b>pred</b> Linear Predictor	0.19492 0.0812	1.00000

The Spearman correlation coefficient between the absolute value of the residuals and the predicted values on the log scale is 0.19492. The *p*-value of 0.0812 indicates that there is not enough evidence to reject the null hypothesis of constant variance.

In order to obtain predicted values on the original data scale, you must back-transform the predicted values produced by PROC GLIMMIX. This is done using a DATA step to compute the estimated means on the scale of the original data.

```
data _null_;
  set params;
  if Effect='Scale' then call symput('var',Estimate);
run;

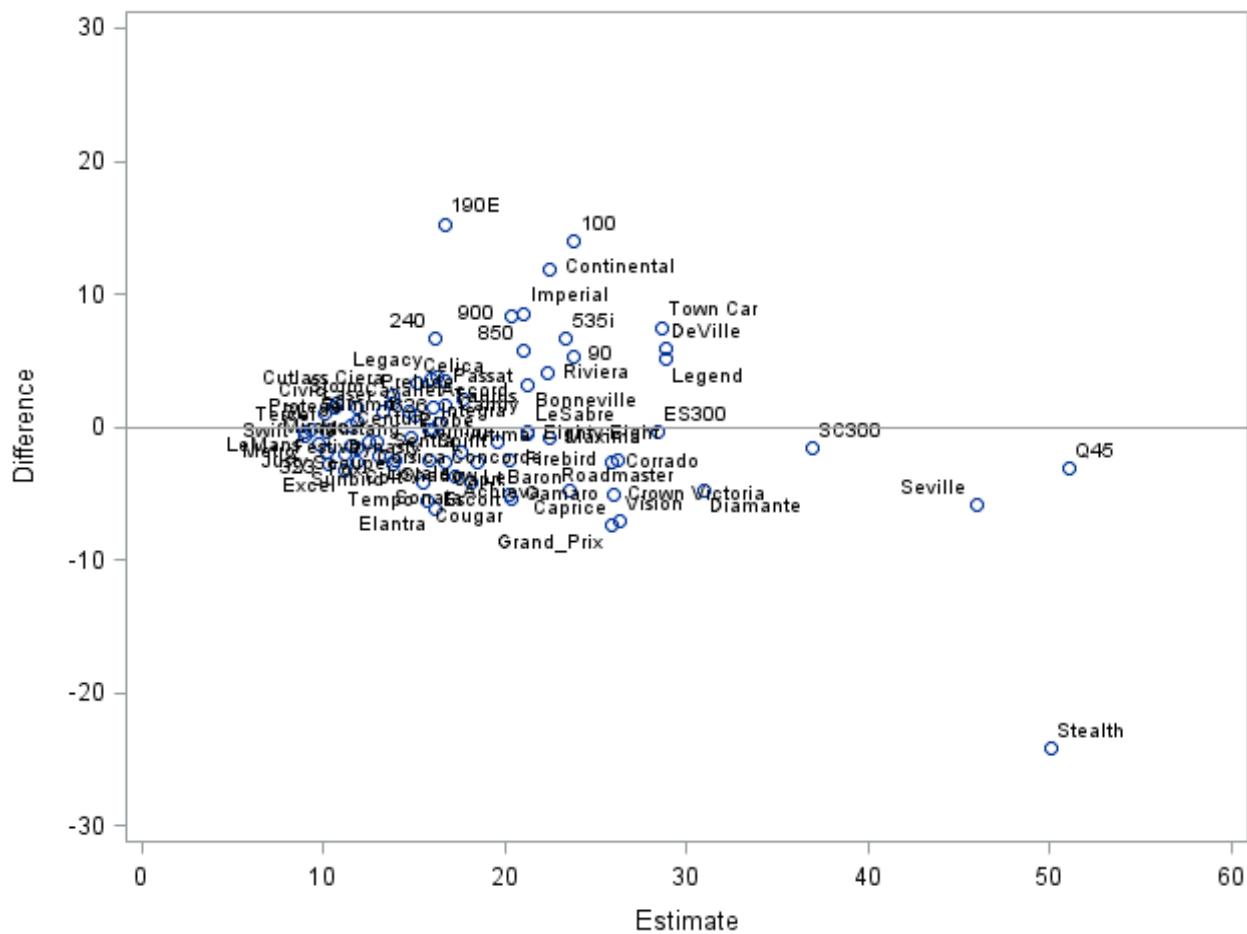
data back;
  set check3;
  Estimate = exp(pred + &var/2);
  Difference = Price-Estimate;
run;

proc sgplot data=back;
  scatter x=Estimate y=Difference / datalabel=model;
  xaxis min=0 max=60;
  yaxis min=-30 max=30;
  refline 0;
run;
```

Selected SCATTER statement option:

DATALABEL displays a label for each data point. If you specify a variable, then the values of that variable are used for the data labels. If you do not specify a variable, then the values of the response variable are used for the data labels.

## PROC SGSCATTER Output



The residuals for cars with predicted values above \$30,000 are all negative, and the residual for the Dodge Stealth is still large. Overall, the lognormal model seems to provide a better fit to the data than the original model.

Recall that the independent variables in this model were selected using ordinary least squares regression, assuming that the errors were normally distributed. When switching to a lognormal model, you should return to the modeling cycle and repeat the variable selection process.

An alternative to using a generalized linear model is to use heteroscedasticity-consistent standard errors.

You can request these with the HCC option in the MODEL statement in PROC REG using the &\_GLSMOD macro variable created from PROC GLMSELECT in ST202d01.

```
proc reg data=d_carfinal;
  model price = &_GLSMOD / hcc hccmethod=3;
run;
quit; *ST202d04.sas
```

Selected MODEL statement options:

- HCC requests heteroscedasticity-consistent standard errors of the parameter estimates. You can use the HCCMETHOD= option to specify the method that is used to compute the heteroscedasticity-consistent covariance matrix.
- HCCMETHOD=0, 1, 2, or 3 specifies the method that is used to obtain a heteroscedasticity-consistent covariance matrix for use with the ACOV, HCC, or WHITE option in the MODEL statement and for heteroscedasticity-consistent tests with the TEST statement. The default is HCCMETHOD=0. MacKinnon and White (1985) introduced three alternative heteroscedasticity-consistent covariance matrix estimators that are all asymptotically equivalent to the estimator HC0. These estimators (labeled HC1, HC2, and HC3) typically have better small sample behavior than HC0. Long and Ervin (2000) studied the performance of these estimators and recommend using the HC3 estimator if the sample size is less than 250.

PROC REG Output

The REG Procedure Model: MODEL1 Dependent Variable: Price					
Number of Observations Read		81			
Number of Observations Used		81			
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4448.69341	1482.89780	65.73	<.0001
Error	77	1737.20536	22.56111		
Corrected Total	80	6185.89877			
Root MSE		4.74985	R-Square	0.7192	
Dependent Mean		18.64321	Adj R-Sq	0.7082	
Coeff Var		25.47766			

The ANOVA table is the same as previously seen.

Variable	Label	DF	Parameter Estimates					Heteroscedasticity Consistent		
			Parameter Estimate	Standard Error	t Value	Pr >  t		Standard Error	t Value	Pr >  t
Intercept	Intercept	1	4.03949	2.17024	1.86	0.0665	3.04604	1.33	0.1887	
s_Hwympg	s_Hwympg	1	-0.80407	0.21378	-3.76	0.0003	0.21049	-3.82	0.0003	
s_Hwympg_2	s_Hwympg^2	1	0.04350	0.01430	3.04	0.0032	0.01424	3.05	0.0031	
Horsepower	Horsepower	1	0.09730	0.01614	6.03	<.0001	0.02360	4.12	<.0001	

HCC Approximation Method: HC3

The table for the parameter estimates now includes the heteroscedasticity-consistent standard errors and the results of the tests based on them. All three of the independent variables, **Hwympg**, **Hwympg<sup>^2</sup>**, and **Horsepower**, are significant.

The results of the tests using the consistent covariance estimates do not differ substantially from the tests that assume constant variance. However, in the presence of heteroscedasticity, you should select candidate models using the consistent variance estimates.



## Exercises

---

### 2. Fitting a Lognormal Model

- a. Use PROC GLIMMIX to fit a lognormal regression model. Use **Price** as the response variable and **Citympg**, **Citympg<sup>2</sup>**, **EngineSize**, **Horsepower**, **Horsepower<sup>2</sup>**, and **Weight** as the independent variables. Use an ODS OUTPUT statement to save the parameter estimates to a data set and an OUTPUT statement to create a data set containing the predicted values. Compare the parameter estimates from this model for **LogPrice** that you obtained in the previous exercise. Using the fit statistics produced by PROC GLIMMIX, evaluate the fit of this model compared to the lognormal model that was fit in class.
- b. Generate estimates of the mean of **Price** by back-transforming the predicted values from the lognormal regression model. Plot the difference between the observed and estimated versus the predicted values. Do these estimates appear to be better than those obtained from the model developed in class?

### Advanced

### 3. Identifying Outlying Values

In the plot that you created for the previous exercise, turn on the DATALABEL option to identify outlying values. You might choose the variable that you consider most relevant to identify these observations. Do you see any patterns?

## 2.3 Chapter Summary

---

After candidate models are identified, they should be evaluated for goodness-of-model fit, model assumptions, multicollinearity, and influential observations. The following table summarizes tools and statistics that are helpful for model diagnostics and remedial measures:

Model Fit and Assumptions	How to Diagnose	Suggested Remedial Measures
Model Fit	<ul style="list-style-type: none"> <li>Residual plots and the R-F plot</li> </ul>	<ul style="list-style-type: none"> <li>Re-specify the model.</li> </ul>
Normality	<ul style="list-style-type: none"> <li>Normality tests of residuals</li> <li>Normal Probability plot for residuals</li> <li>Histogram for residuals</li> </ul>	<ul style="list-style-type: none"> <li>Use procedures such as PROC LOGISTIC, PROC GENMOD, PROC GLIMMIX, or PROC NLIN to model data from the appropriate distributions.</li> <li>Transform the dependent variable.</li> </ul>
Constant variance	<ul style="list-style-type: none"> <li>Residual plot</li> <li>Summary statistics</li> <li>Spearman rank correlation coefficient</li> <li>Test produced by the SPEC option in the MODEL statement of PROC REG</li> </ul>	<ul style="list-style-type: none"> <li>Use PROC GENMOD, PROC GLIMMIX, or PROC MIXED to model the nonconstant variance.</li> <li>Transform the dependent variable.</li> <li>Add the HCC option to the MODEL statement in PROC REG to obtain tests based on heteroscedasticity-consistent standard errors.</li> <li>Fit weighted least squares models.</li> </ul>
Independent Observations	<ul style="list-style-type: none"> <li>Understand how the data were gathered. If they were gathered over time or represent data from a complex survey design or repeated measures, the data is not independent.</li> <li>A plot of residual versus the ordering component</li> <li>Durbin-Watson statistic for time series data</li> </ul>	<ul style="list-style-type: none"> <li>Use PROC AUTOREG or PROC ARIMA to model time series data.</li> <li>Use PROC MIXED or PROC GLIMMIX to model repeated measures data.</li> <li>Use the SURVEY procedures to model data gathered from a complex sample design.</li> </ul>
Multicollinearity	<ul style="list-style-type: none"> <li>Correlation statistics</li> <li>VIF, COLLIN, and COLLINOINT options in the MODEL statement from PROC REG</li> </ul>	<ul style="list-style-type: none"> <li>Center the variables that are involved in higher-ordered terms.</li> <li>Exclude redundant variables.</li> <li>Redefine the predictor variables.</li> <li>Use ridge regression or principal component regression.</li> </ul>
Influential observations	<ul style="list-style-type: none"> <li>RSTUDENT residuals, LEVERAGE, Cook's D, DFFITS, DFBETAS, COVRATIO statistics, and their plots</li> </ul>	<ul style="list-style-type: none"> <li>Perform a sensitivity analysis.</li> <li>Redefine the model.</li> <li>Delete the influential observations when appropriate.</li> </ul>

When model assumptions are violated, one possible solution is to fit a generalized linear model using a more appropriate probability distribution. To determine which distribution to use, you can do the following:

- use theoretical knowledge or previous research to select an appropriate distribution
- consider the rate at which the variance increases as the dependent variable increases

After the models are evaluated, model refinement might be necessary. If so, you should perform model diagnostics on the model for the transformed data. A final step is to evaluate the model's predictive capability with data that were not used to build the model.

## 2.4 Solutions

---

### Solutions to Exercises

#### 1. Performing Model Diagnostics on the Revised Model

- a. Use the appropriate options in the MODEL statement of PROC REG to evaluate multicollinearity, influential observations, and the constant variance assumption. Generate the necessary plots to check the assumptions of regression. Use the OUTPUT statement to save the residuals and calculate the Spearman rank correlation coefficient between the absolute value of the residuals and the predicted values to check the constant variance assumption. Is multicollinearity a problem for this model? Does this model meet the assumptions for linear regression?

```

ods output ParameterEstimates=param;
proc glmselect data=STAT2.cars4 outdesign(addinputvars)=d_carslog;
  effect p_city=polynomial(citympg / degree=2
    standardize(method=moments)=center);
  effect p_hp=polynomial(horsepower / degree=2
    standardize(method=moments)=center);
  model logprice = p_city enginesize p_hp weight / selection=none;
run;

proc reg data=d_carslog plots (label)=all;
  model logprice = & GLSMOD
    / vif collin collinoint influence spec partial;
  output out=check r=Residual p=Pred rstudent=rstudent h=leverage;
  id model;
run;
quit;

data check;
  set check;
  abserror=abs(residual);
run;

proc corr data=check spearman nosimple;
  var abserror pred;
run;                                *ST202s01.sas;

```

## Partial PROC REG Output

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	Intercept	1	2.02225	0.37759	5.36	<.0001	0
s_Citympg	s_Citympg	1	-0.03877	0.01269	-3.05	0.0031	9.84600
s_Citympg_2	s_Citympg^2	1	0.00197	0.00058731	3.35	0.0013	3.68546
EngineSize	EngineSize	1	-0.17068	0.06052	-2.82	0.0062	7.22881
s_Horsepower	s_Horsepower	1	0.00442	0.00123	3.58	0.0006	7.56525
s_Horsepower_2	s_Horsepower^2	1	-0.00001481	0.00000751	-1.97	0.0522	2.17676
Weight	Weight	1	0.00040871	0.00015203	2.69	0.0089	14.39378

The VIF for **Weight** is 14.39. This indicates moderate collinearity between **Weight** and one or more other variables in the model.

Collinearity Diagnostics									
Number	Eigenvalue	Condition Index	Proportion of Variation						
			Intercept	s_Citympg	s_Citympg_2	EngineSize	s_Horsepower	s_Horsepower_2	Weight
1	3.52397	1.00000	0.00025068	0.00021217	0.00308	0.00136	0.00102	0.01385	0.00017953
2	2.10685	1.29330	0.00003113	0.02045	0.02815	0.00002193	0.01841	0.00051501	9.091122E-7
3	0.94125	1.93492	0.00031176	0.00368	0.04111	0.00084716	0.03194	0.13926	0.00016684
4	0.32804	3.27760	0.00013715	0.00981	0.26124	0.00140	0.05853	0.43599	5.483234E-7
5	0.08149	6.57589	0.00133	0.56438	0.38326	0.00029441	0.48108	0.26639	0.00060632
6	0.01699	14.40302	0.04826	0.28778	0.26876	0.67853	0.04849	0.02861	0.00815
7	0.00142	49.76137	0.94968	0.11368	0.01439	0.31754	0.36054	0.11538	0.99090

The collinearity diagnostics table suggests that moderate multicollinearity might exist between **Weight** and the intercept. You might or might not want to remove the variable **Weight**, depending on the objectives of the study and other considerations.

Collinearity Diagnostics (intercept adjusted)									
Number	Eigenvalue	Condition Index	Proportion of Variation						
			s_Citympg	s_Citympg_2	EngineSize	s_Horsepower	s_Horsepower_2	Weight	
1	3.70756	1.00000	0.00612	0.00459	0.00813	0.00782	0.00349	0.00465	
2	1.39299	1.63144	0.00608	0.07396	0.00070578	0.00690	0.15059	0.00002773	
3	0.63975	2.40734	0.00167	0.14757	0.04033	0.00116	0.27162	0.00773	
4	0.14132	5.12198	0.00181	0.02510	0.34592	0.47190	0.37294	0.00578	
5	0.07131	7.21067	0.93818	0.74613	0.16736	0.08397	0.02964	0.00883	
6	0.04707	8.87522	0.04614	0.00264	0.43756	0.42826	0.17171	0.97298	

When the intercept is adjusted out of the model, there seems to be no apparent collinearity among the predictor variables.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
25	30.71	0.1989

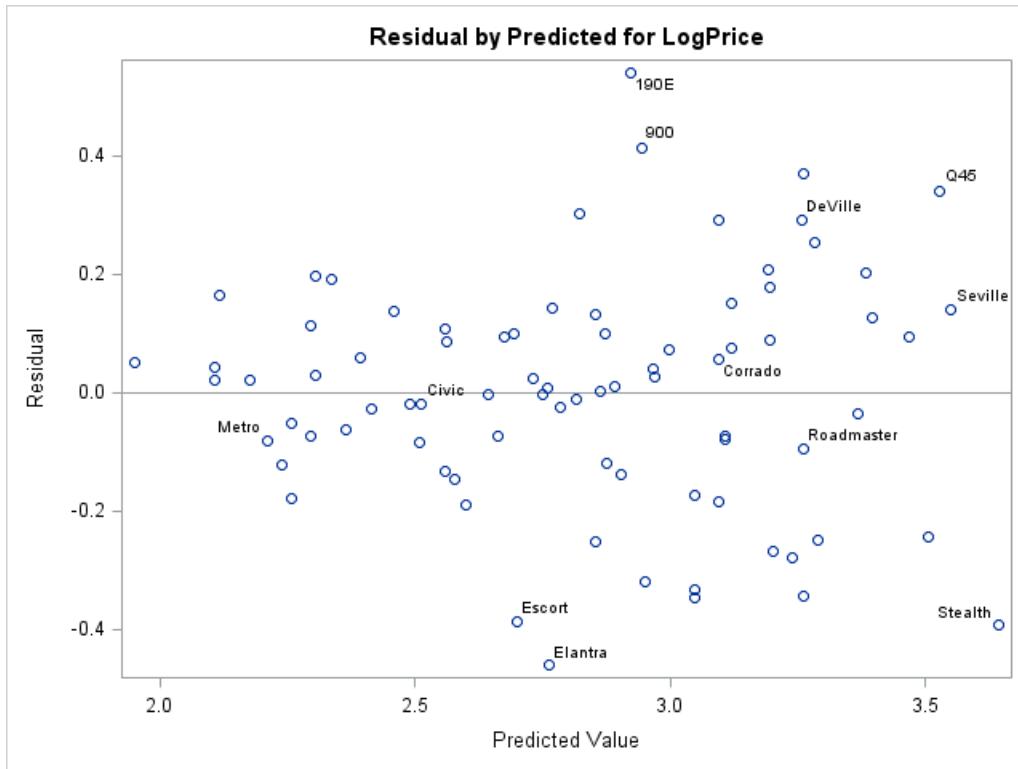
WARNING: The average covariance matrix for the SPEC test has been deemed singular which violates an assumption of the test. Use caution when interpreting the results of the test.

The SPEC test indicates that there is not enough evidence to reject the null hypothesis that the model is correctly specified, that the errors are independent of the predictor variables, and that the variances are constant. However, you should use caution when interpreting the results due to the warning message in the Log window.

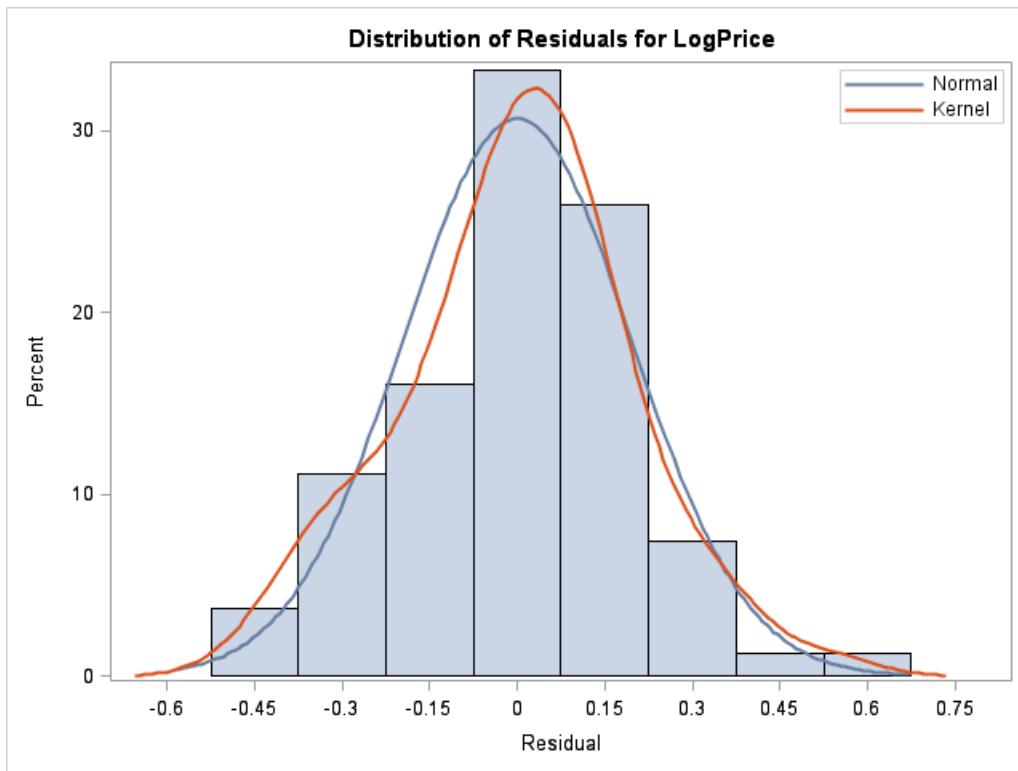
#### Partial CORR Output

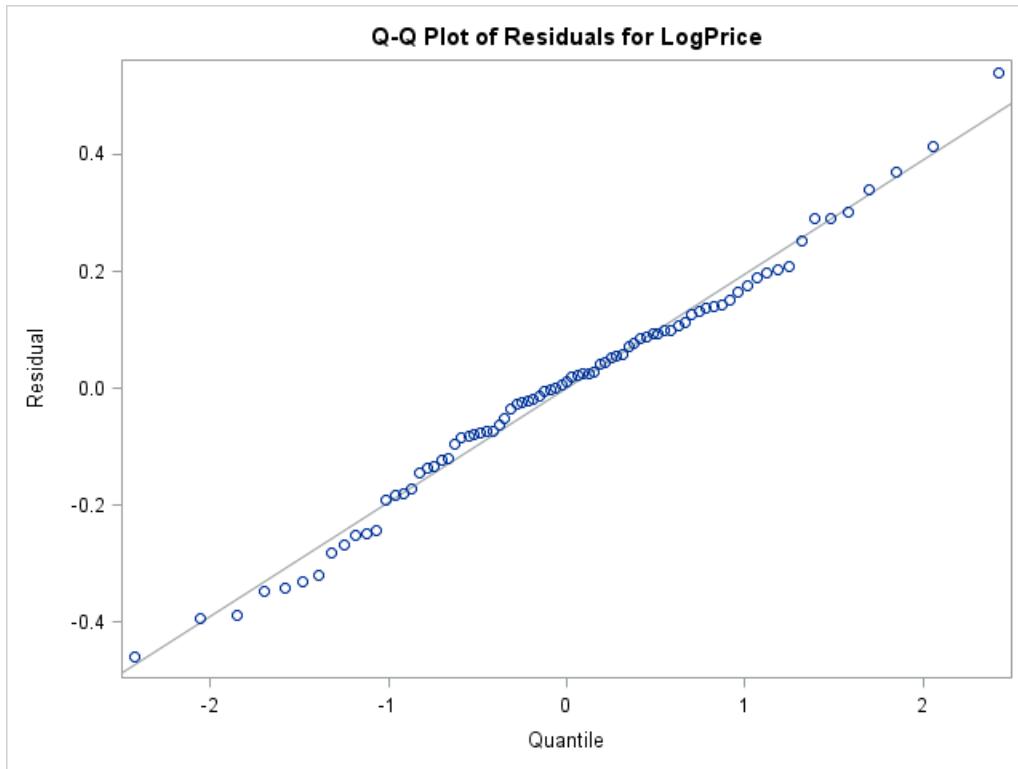
Spearman Correlation Coefficients, N = 81 Prob >  r  under H0: Rho=0		
	AbsError	Pred
<b>AbsError</b>	1.00000 0.0001	0.41820 0.0001
<b>Pred</b> Predicted Value of LogPrice	0.41820 0.0001	1.00000

The Spearman correlation coefficient is lower for this model than for the original model with **Price** as the dependent variable (0.41820 versus 0.60274). However, the significant *p*-value for this test (0.0001) indicates that the variance is not stabilized for this model, which contradicts the results of the SPEC test. However, recall that the results of the SPEC test are questionable because an assumption for the test was violated.



The residual plot looks better than the one for the model with **Price** as the outcome variable. However, the variance still appears to be smaller at the lower range of the predicted values than at the higher range. This corroborates the results of the Spearman correlation coefficient test.

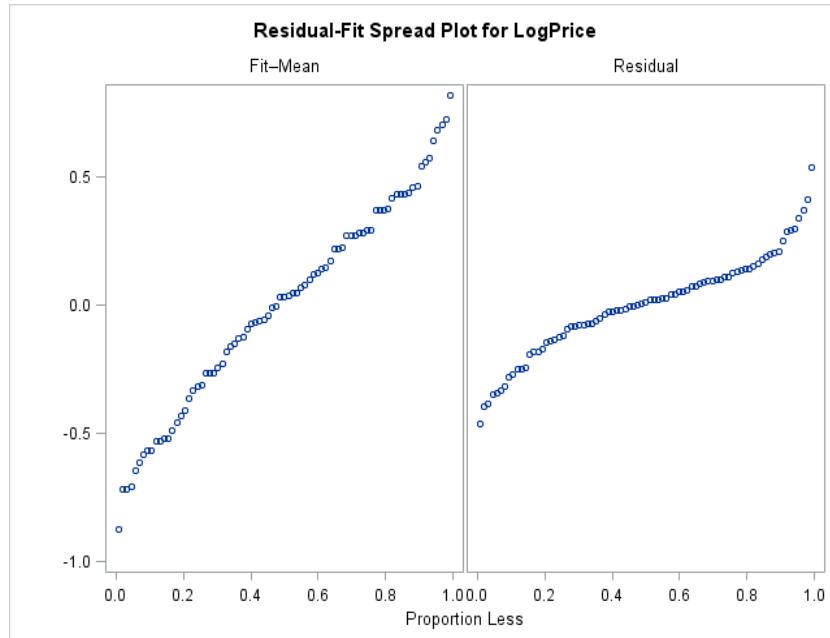




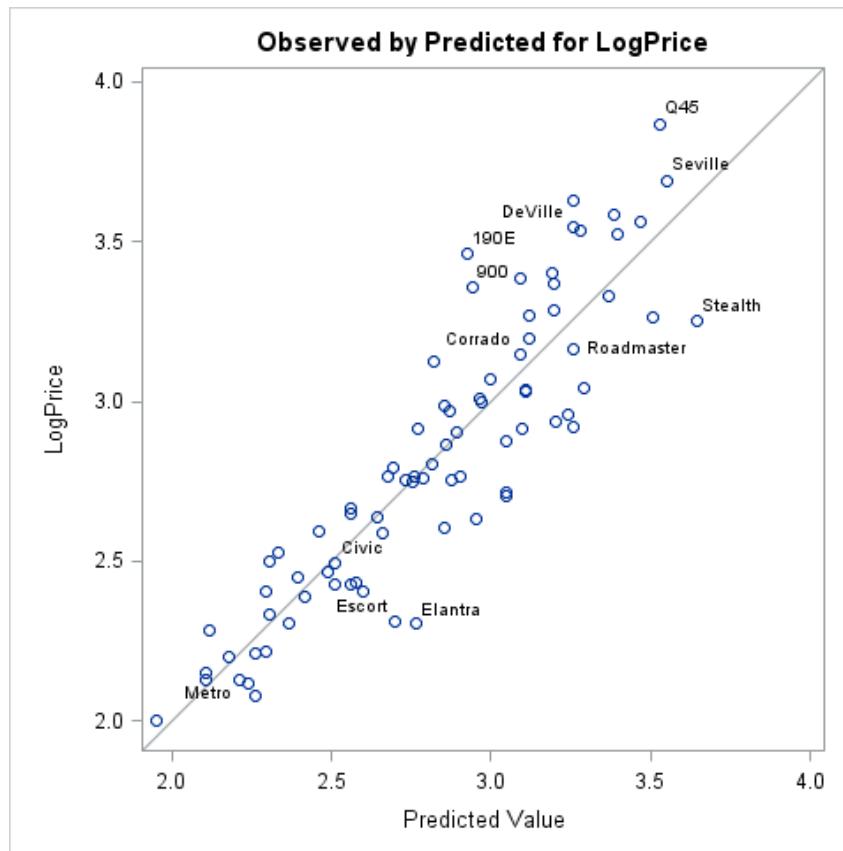
Neither the histogram of the residuals nor the normal quantile plot indicates any problems with the normality assumption of the error terms. This model seems to meet the assumptions of normality and independence for linear regression, but does not appear to meet the assumption of constant variance.

- b.** To assess the model fit, examine the R-F plot, the plot of the observed values versus the predicted value, and the plots of residuals versus the independent variables. What are your conclusions?

#### Partial PROC REG ODS Graphics Output

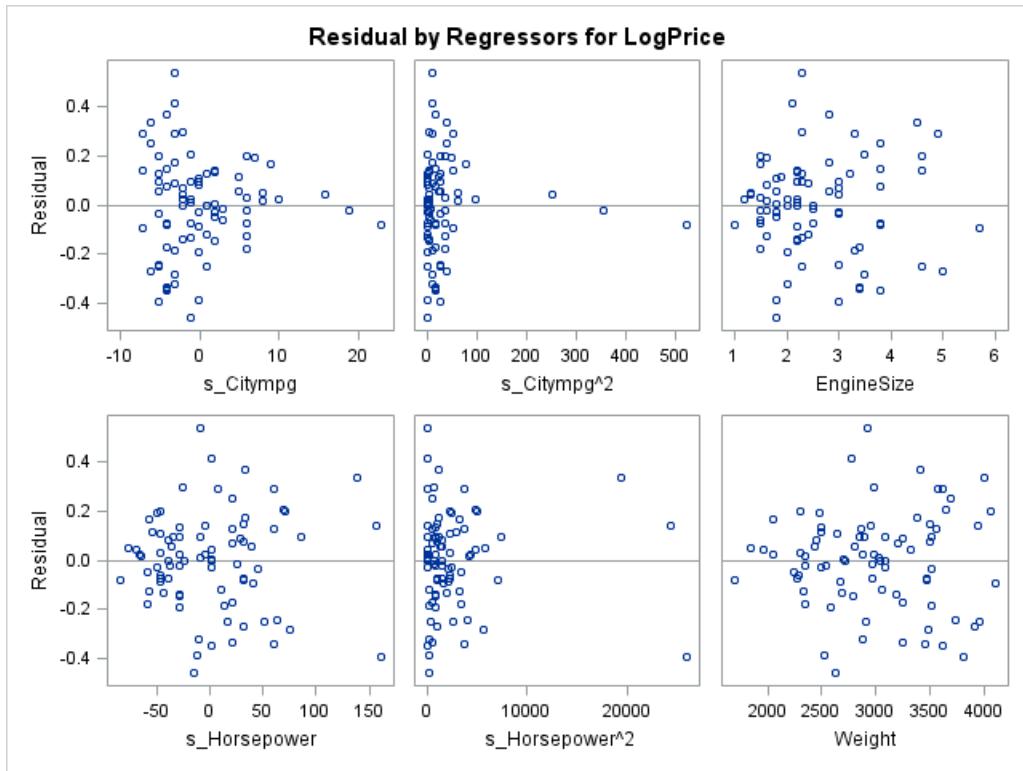


Because the spread of the residual plot is smaller than that of the fit-mean plot, the model explains most of the variation in the data.



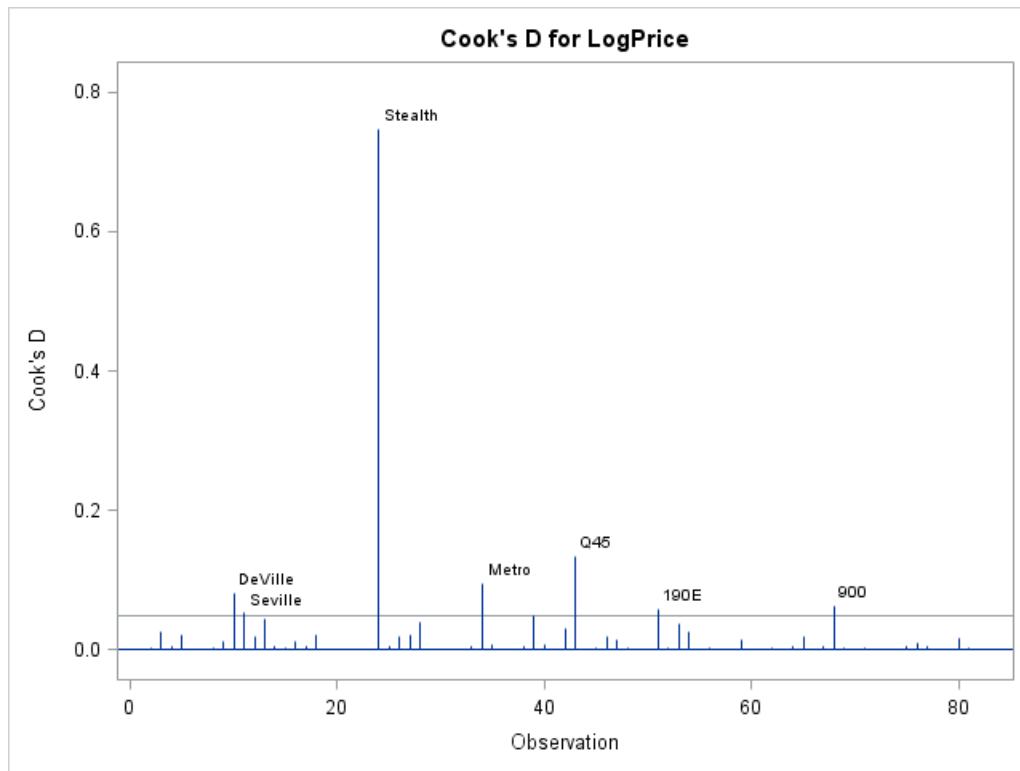
The plot of the observed values versus the predicted values has a fairly tight distribution around the 45-degree reference line. This indicates a good fit. The fit for observation #24 (the Dodge Stealth) seems to improve in this graph.

However, in the following graphs of the residuals versus the predictors, the model does not seem to fit the Dodge Stealth well. Consequently, the potential improvement seen here might be due solely to the change in scale from **Price** to **LogPrice**.

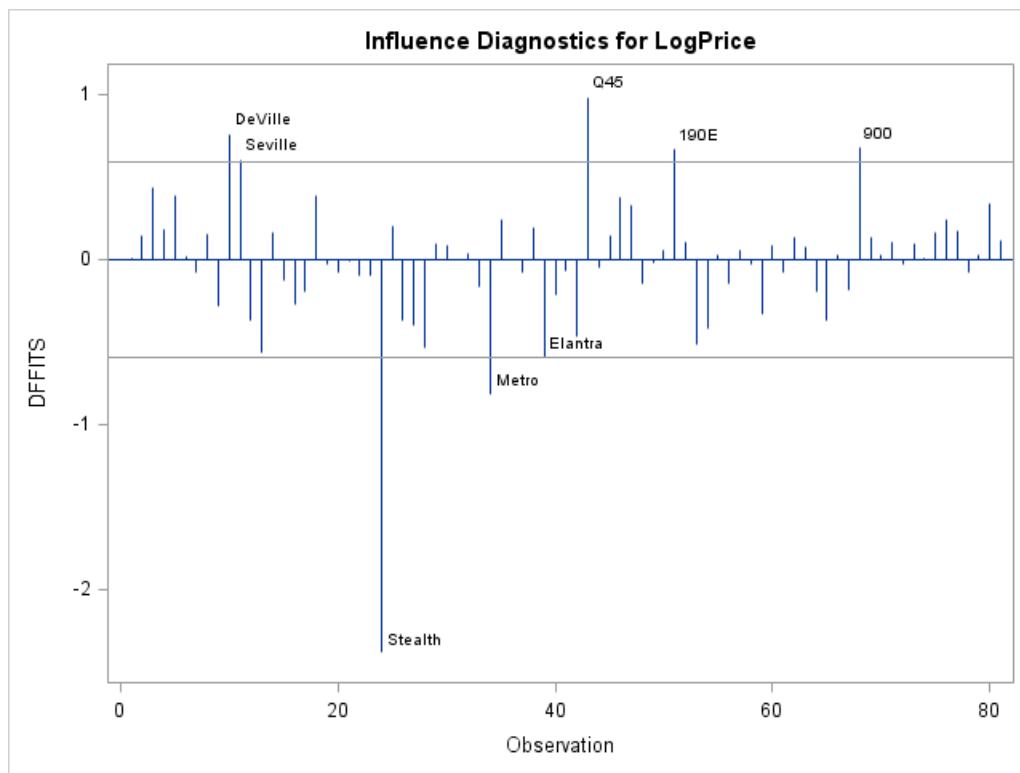


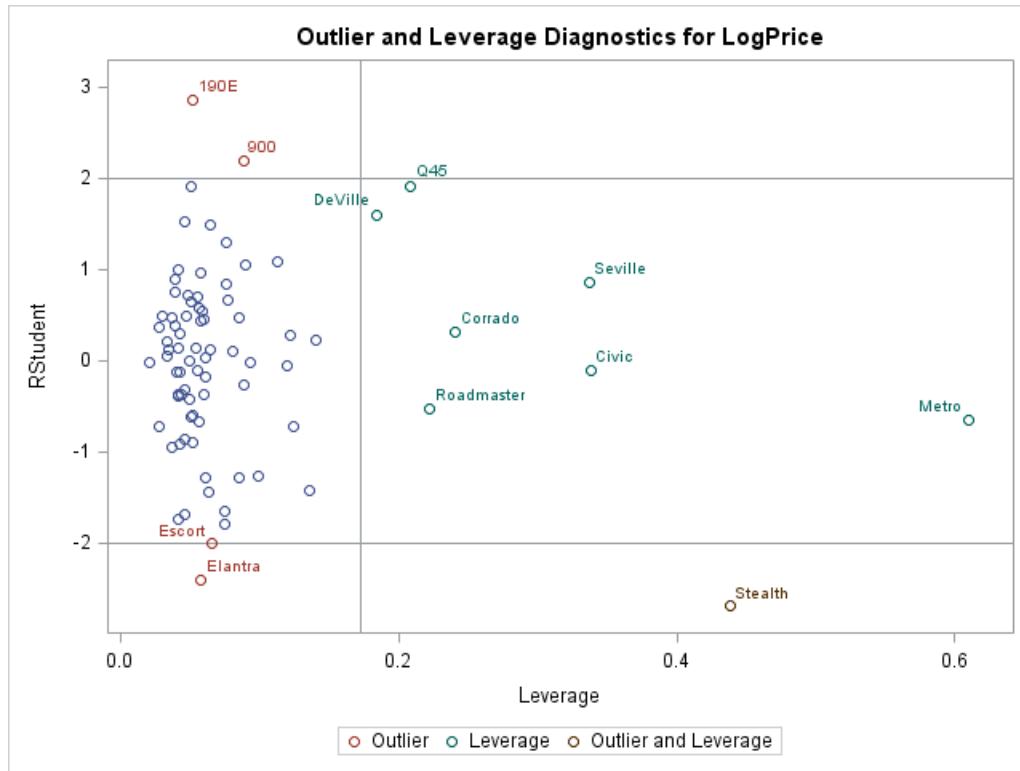
The residual plots for **Citympg** and **Citympg<sup>2</sup>** for this model look similar to the residual plots for **Hwympg** and **Hwympg<sup>2</sup>** for the model with **Price** as the dependent variable. Also seen is the linear pattern for the three data points with (centered) values of **Citympg** above 10 mpg. The residual plot for **EngineSize** shows no apparent patterns. Only three observations have values of **Horsepower** above 100, so they are prominent in the residual plots of **Horsepower** and **Horsepower<sup>2</sup>**. The residual plot for **Weight** exhibits no apparent pattern. The outlying point with the large negative residual in all the graphs is observation #24, the Dodge Stealth.

- c. Examine the plots of the Cook's D, DFFITS, and DFBETA statistics, the plot of RSTUDENT versus LEVERAGE, and the partial leverage plots to identify any outlying or influential observations. Output the potentially influential observations to a data set. What are your conclusions?

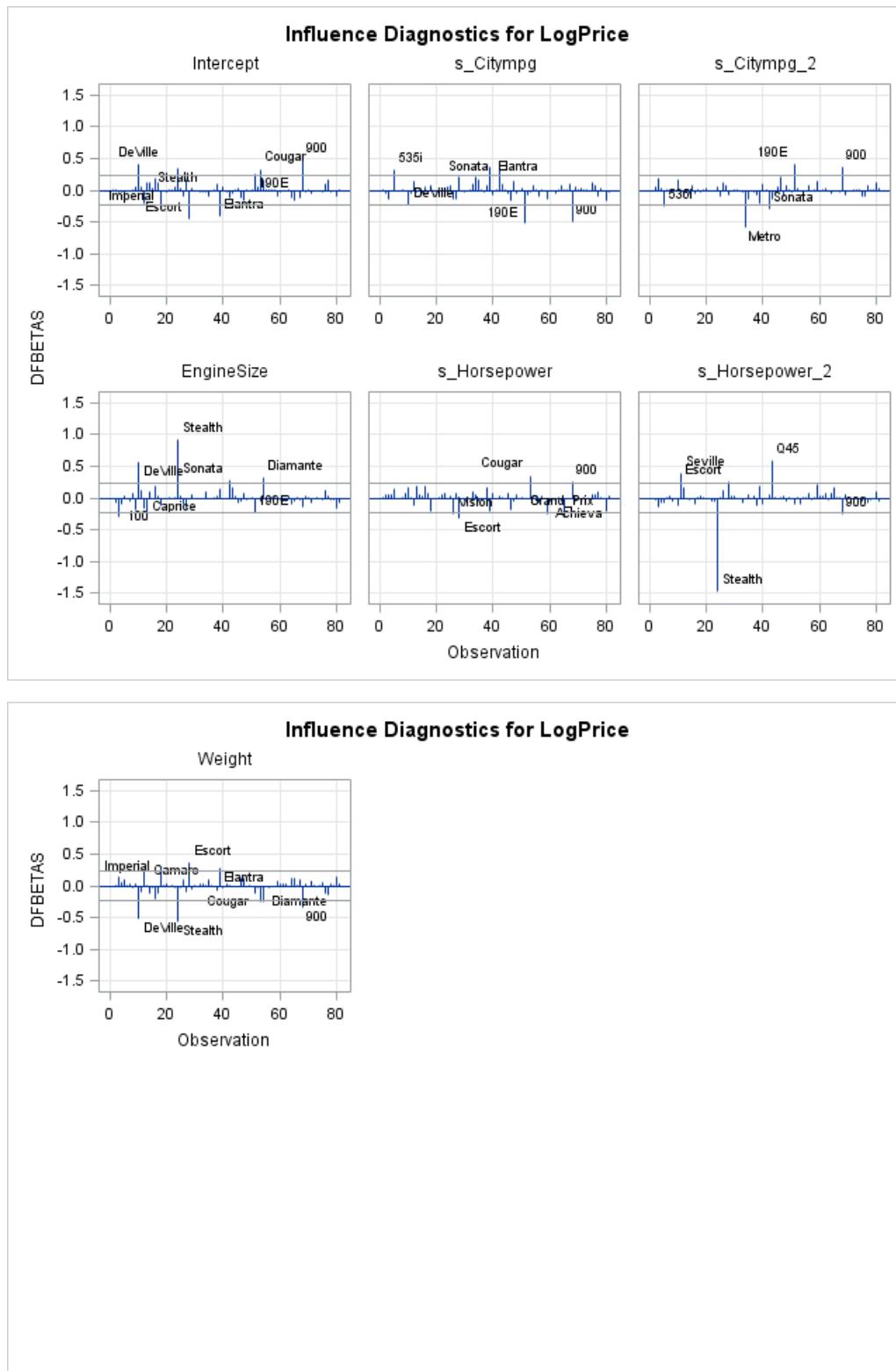


The plots of Cook's D and DFFITS for the model with **LogPrice** as the independent variable flag more influential observations than the model with **Price** as the outcome variable. The Dodge Stealth is flagged as highly influential by both measures.

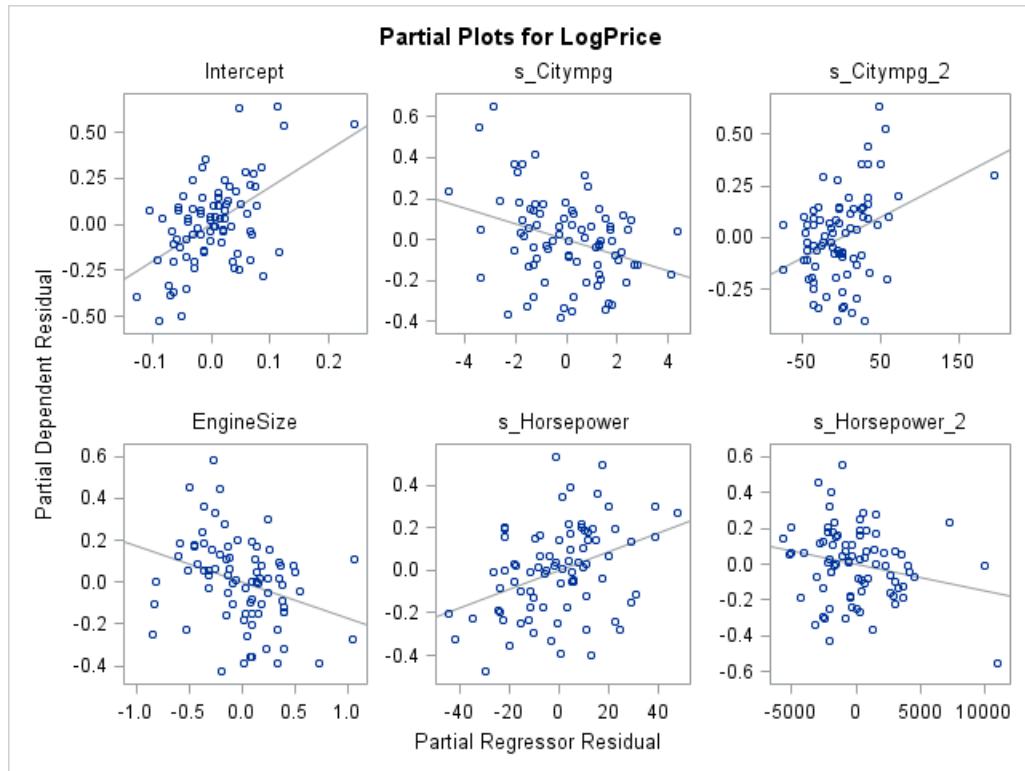




The RStudent versus Leverage plot for this model identifies observations that are outliers and have high leverage. The Dodge Stealth and Mercedes-Benz 190E are still flagged as outliers for this model. The Audi 100 and Lincoln Continental were outliers for the original model, but not here. The SAAB 900 is flagged as an outlier for this model, but not for the original one. For this model of **LogPrice**, two more points are potentially more influential than for the model with **Price** as the dependent variable.



The Dodge Stealth exhibits influence on the parameter estimates for **Intercept**, **EngineSize**, **Horsepower<sup>2</sup>**, and **Weight**. You can output these influential observations that were flagged by the influence statistics to a data set to further examine them.



Although these partial leverage plots for the model for **LogPrice** look better than the ones for **Price**, several points still appear to be influential or suffering from lack of fit.

```
%let numparms = 7; /* # of predictor variables + 1 */
%let numobs = 81; /* # of observations */
%let idvars = manufacturer model price; /* relevant identification
variable(s) */

data influence;
  set check;
  absrstud=abs(rstudent);
  if absrstud ge 2 then output;
  else if leverage ge (2*&numparms /&numobs) then output;
run;

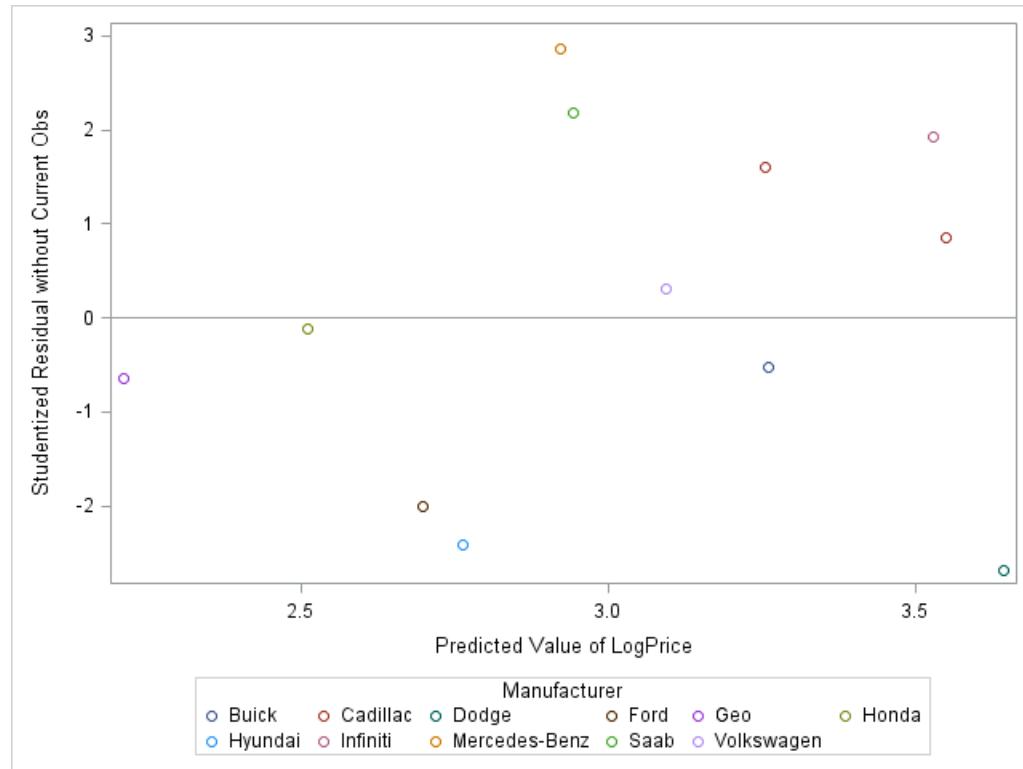
proc print data=influence;
  var &idvars;
run; *ST202s01.sas;
```

Obs	Manufacturer	Model	Price
1	Buick	Roadmaster	23.7
2	Cadillac	DeVille	34.7
3	Cadillac	Seville	40.1
4	Dodge	Stealth	25.8
5	Ford	Escort	10.1
6	Geo	Metro	8.4
7	Honda	Civic	12.1
8	Hyundai	Elantra	10.0
9	Infiniti	Q45	47.9
10	Mercedes-Benz	190E	31.9
11	Saab	900	28.7
12	Volkswagen	Corrado	23.3

Three more observations are either outlying or more influential for this model with **LogPrice** as the dependent variable than the previous model for **Price**.

- d. Use the output data set containing the influence statistics that you created in the previous exercise. Use PROC SG PLOT to create a scatter plot of the residuals versus the predicted values. Look in the SG PLOT documentation for how to use the GROUP option. Plot the residuals versus the predicted values. Use the variable **Manufacturer** as the grouping variable. Do you see any patterns?

```
proc sgplot data=influence;
  scatter y=rstudent x=pred / group=manufacturer;
  refline 0;
run; *ST202s01.sas;
```



No manufacturer seems to have more observations that are influential than any of the other manufacturers.

## 2. Fitting a Lognormal Model

- Use PROC GLIMMIX to fit a lognormal regression model. Use **Price** as the response variable and **Citympg**, **Citympg^2**, **EngineSize**, **Horsepower**, **Horsepower^2**, and **Weight** as the independent variables. Use an ODS OUTPUT statement to save the parameter estimates to a data set and an OUTPUT statement to create a data set containing the predicted values. Compare the parameter estimates from this model for **LogPrice** that you obtained in the previous exercise. Using the fit statistics produced by PROC GLIMMIX, evaluate the fit of this model compared to the lognormal model that was fit in class.

```
ods output ParameterEstimates=params;
proc glimmix data=STAT2.cars;
effect p_city=polynomial(citympg / degree=2
    standardize(method=moments)=center);
effect p_hp=polynomial(horsepower / degree=2
    standardize(method=moments)=center);
model price = p_city enginesize p_hp weight / dist=lognormal
    solution;
output out=out pred=pred;
id manufacturer model price citympg enginesize;
run; *ST202s02.sas;
```

## Partial PROC GLIMMIX Output

Parameter Estimates						
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	
Intercept	2.0222	0.3776	74	5.36	<.0001	
s_Citympg	-0.03877	0.01269	74	-3.05	0.0031	
s_Citympg^2	0.001968	0.000587	74	3.35	0.0013	
EngineSize	-0.1707	0.06052	74	-2.82	0.0062	
s_Horsepower	0.004417	0.001233	74	3.58	0.0006	
s_Horsepower^2	-0.00001	7.506E-6	74	-1.97	0.0522	
Weight	0.000409	0.000152	74	2.69	0.0089	
Scale	0.04119	0.006772	.	.	.	

The parameter estimates for the lognormal model are identical to those obtained using ordinary least squares regression with the response variable **LogPrice**.

Fit Statistics	
-2 Res Log Likelihood	47.61
AIC (smaller is better)	63.61
AICC (smaller is better)	65.83
BIC (smaller is better)	82.05
CAIC (smaller is better)	90.05
HQIC (smaller is better)	70.97
Pearson Chi-Square	3.05
Pearson Chi-Square / DF	0.04

The information criteria statistics are all larger for this model than for the previous model with **Hwympg**, **Hwympg^2**, and **Horsepower** as independent variables. The current model might be more complex than necessary.

- b. Generate estimates of the mean of **Price** by back-transforming the predicted values from the lognormal regression model. Plot the difference between the observed and estimated versus the predicted values. Do these estimates appear to be better than those obtained from the model developed in class?

```
data _null_;
  set params;
  if Effect='Scale' then call symput('var', Estimate);
run;
```

```

data out;
  set out;
  Estimate=exp(pred + &var/2);
  Difference = price - estimate;
run;

proc print data=out;
  var manufacturer model price estimate difference;
run;

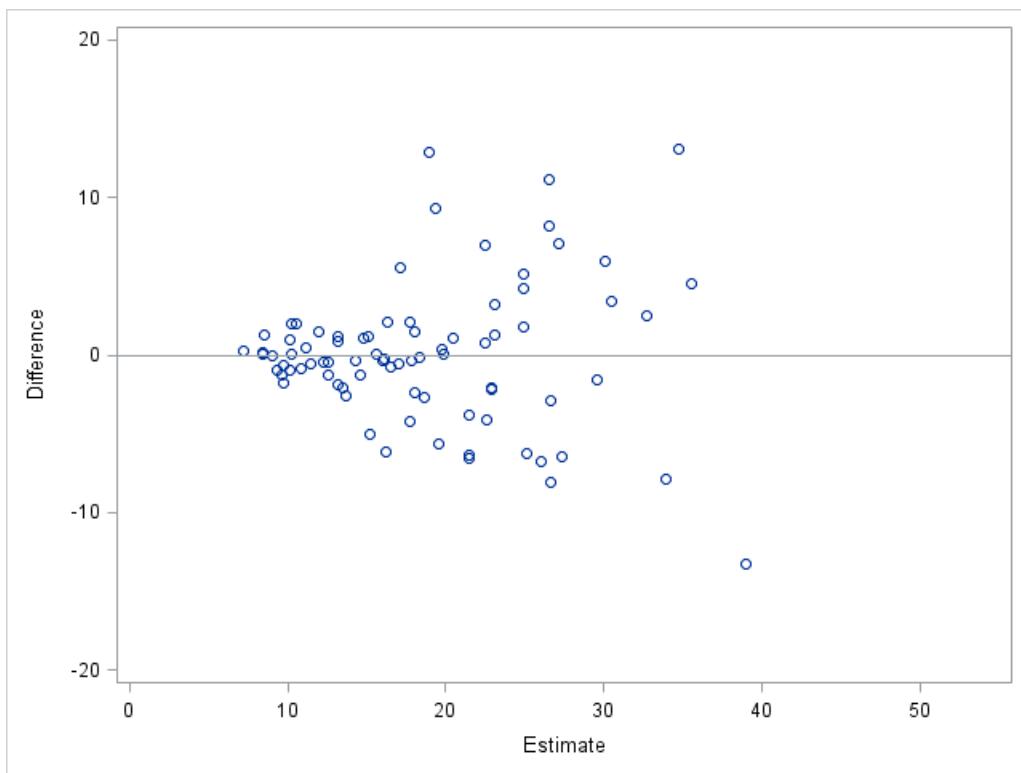
proc sgplot data=out;
  scatter y=difference x=estimate;
  xaxis min=0 max=55;
  yaxis min=-20 max=20;
  refline 0;
run;                                *ST202s02.sas;

```

Partial PROC PRINT Output

Obs	Manufacturer	Model	Citympg	Horsepower	EngineSize	Price	Estimate	Difference
1	Acura	Integra	25	140	1.8	15.9	16.1214	-0.2214
2	Acura	Legend	18	200	3.2	33.9	30.4785	3.4215
3	Audi	100	19	172	2.8	37.7	26.5819	11.1181
4	Audi	90	20	172	2.8	29.1	24.9006	4.1994
5	BMW	535i	22	208	3.5	30.0	24.8786	5.1214
6	Buick	Century	22	110	2.2	15.7	15.6498	0.0502
7	Buick	LeSabre	19	170	3.8	20.8	22.8552	-2.0552
8	Buick	Riviera	19	170	3.8	26.3	23.0899	3.2101
9	Buick	Roadmaster	16	180	5.7	23.7	26.5963	-2.8963
10	Cadillac	DeVille	16	200	4.9	34.7	26.5020	8.1980
11	Cadillac	Seville	16	295	4.6	40.1	35.5693	4.5307
12	Chevrolet	Camaro	19	160	3.4	15.1	21.4769	-6.3769

## PROC SGLOT Output

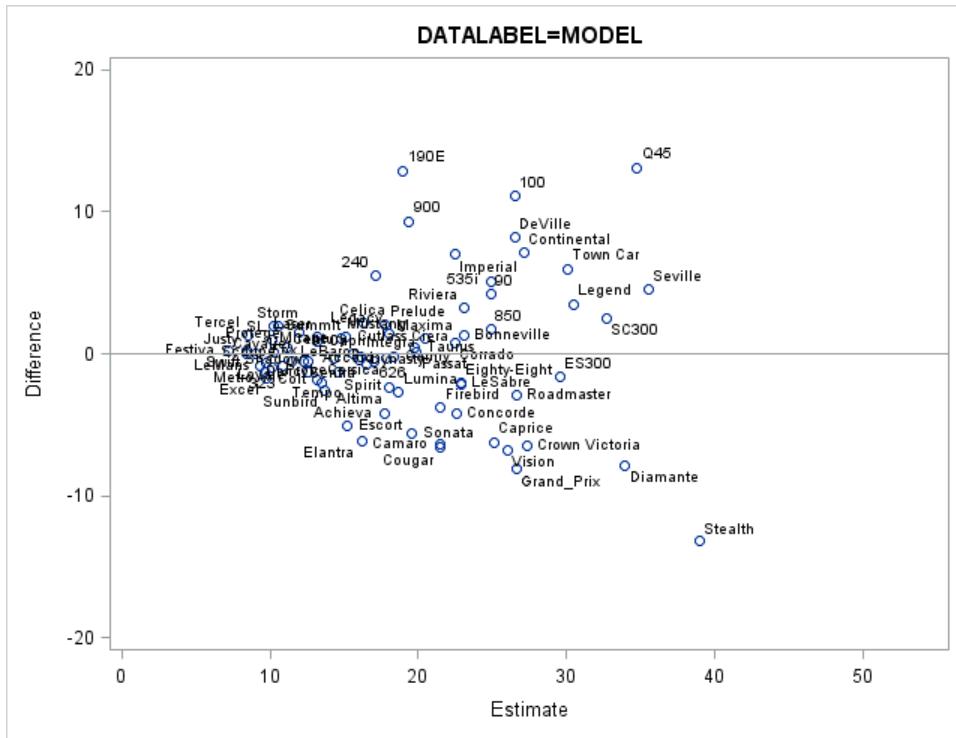


There do not appear to be any extreme values of **Difference** (in other words, residuals on the original scale) for this model. In that sense, it might be a better model than the one developed in class.

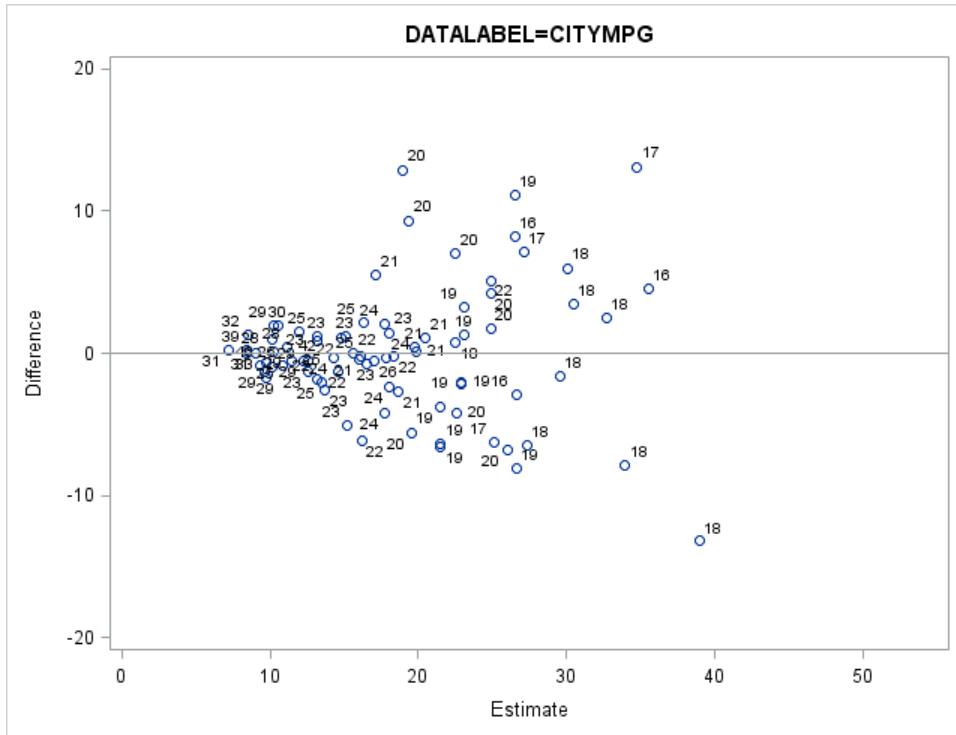
### 3. Identifying Outlying Values

In the plot that you created for the previous exercise, turn on the DATALABEL option to identify outlying values. You might choose the variable that you consider most relevant to identify these observations. Do you see any patterns?

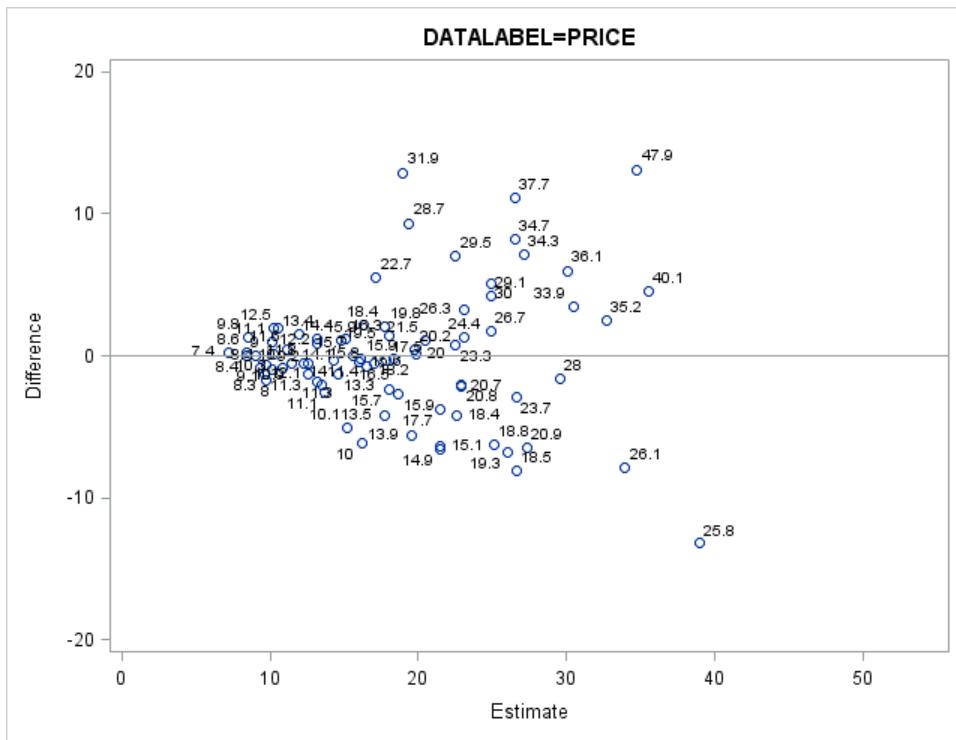
```
proc sgplot data=out;
  scatter y=difference x=estimate / datalabel=variable;
  xaxis min=0 max=55;
  yaxis min=-20 max=20;
  refline 0;
  title 'DATALABEL=VARIABLE';
run; *ST202s02.sas;
```



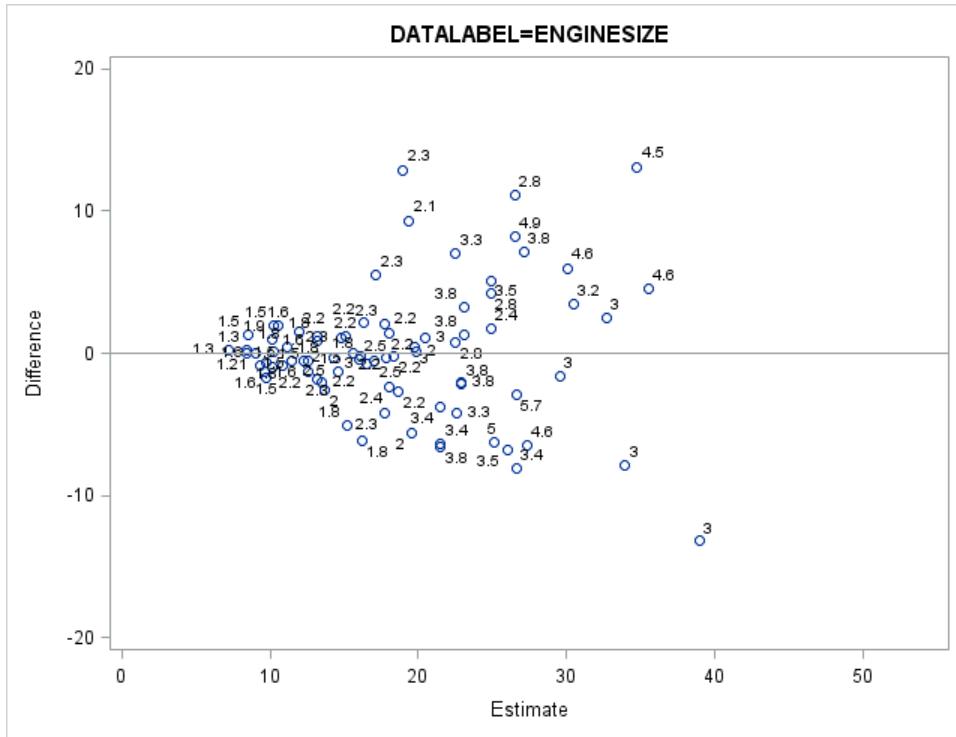
The points with the most extreme values for **Difference** appear to be in the category of luxury cars, such as the Mercedes-Benz 190E, the Infinity Q45, and the Dodge Stealth.



The pattern in this graph is that cars with lower values of **Citympg** have the most extreme residuals on the back-transformed scale.



From this graph it appears that no cars at the lowest price ranges have extreme values of **Difference**. This indicates that the model fits those cars the best.



None of the points with extreme values for **Difference** has unusual values for **EngineSize**. All of these graphs seem to indicate that the model fits poorly for cars in the higher price ranges, with large engine sizes, and in the lower range of fuel efficiency.

## Solutions to Student Activities (Polls/Quizzes)

### 2.01 Multiple Choice Poll – Correct Answer

Regression model diagnostics might include checking for which of the following?

- a. model fit
- b. model assumptions (independent normal errors with constant variance)
- c. multicollinearity
- d. influential observations
- e. a and b
- f. all of the above.

20

### 2.02 Multiple Choice Poll – Correct Answer

In the previous demonstration, which of the following did you discover?

- a. The residuals appear to be normally distributed.
- b. The residual variances might not be constant.
- c. There does not seem to be apparent multicollinearity.
- d. There can be some influential observations.
- e. all of the above
- f. none of the above

23

## 2.03 Multiple Choice Poll – Correct Answer

What did you find out about the model diagnostics in the previous exercise?

- a. The residuals are normally distributed.
- b. There seems to be apparent multicollinearity.
- c. The error variance might not be constant.
- d. a and c
- e. all of the above



# Chapter 3 Analysis of Variance

<b>3.1 ANOVA Review.....</b>	<b>3-3</b>
Demonstration: Two-Way Analysis of Variance .....	3-12
Exercises .....	3-22
<b>3.2 Postfitting Analyses.....</b>	<b>3-23</b>
Demonstration: The LSMESTIMATE Statement .....	3-34
Exercises .....	3-37
<b>3.3 Contrasts and Estimates (Self-Study) .....</b>	<b>3-38</b>
Demonstration: Contrasts and Estimates .....	3-50
Exercises (Self-Study) .....	3-53
<b>3.4 Evaluations of Model Assumptions and Remedial Measures.....</b>	<b>3-54</b>
Demonstration: Identifying Violations of ANOVA Assumptions .....	3-64
Demonstration: Accounting for Unequal Variances .....	3-74
Exercises .....	3-82
<b>3.5 Chapter Summary.....</b>	<b>3-83</b>
<b>3.6 Solutions .....</b>	<b>3-85</b>
Solutions to Exercises .....	3-85
Solutions to Student Activities (Polls/Quizzes) .....	3-97



## 3.1 ANOVA Review

---

### Objectives

- Describe the relationship between analysis of variance and linear regression.
- Fit a two-way ANOVA model and interpret the interaction.

3

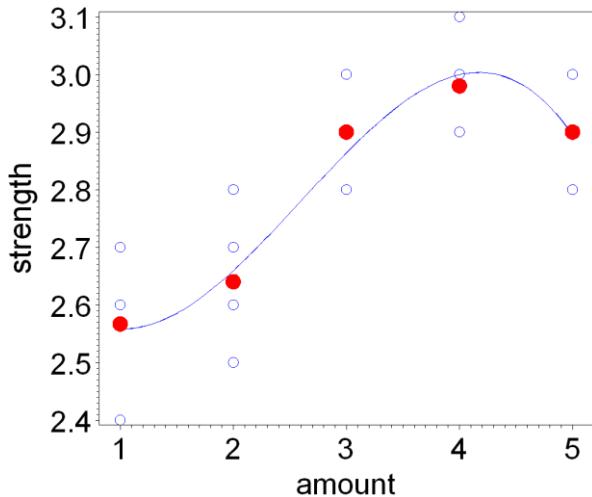
### 3.01 Poll

Both ANOVA models and regression models are general linear models that use OLS to obtain parameter estimates and standard errors.

- True
- False

4

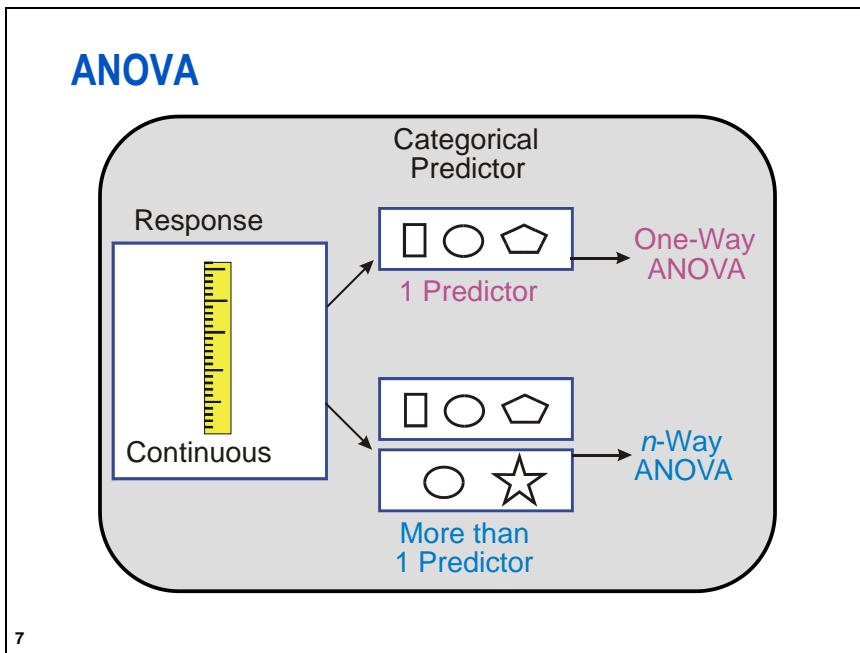
## Regression and ANOVA



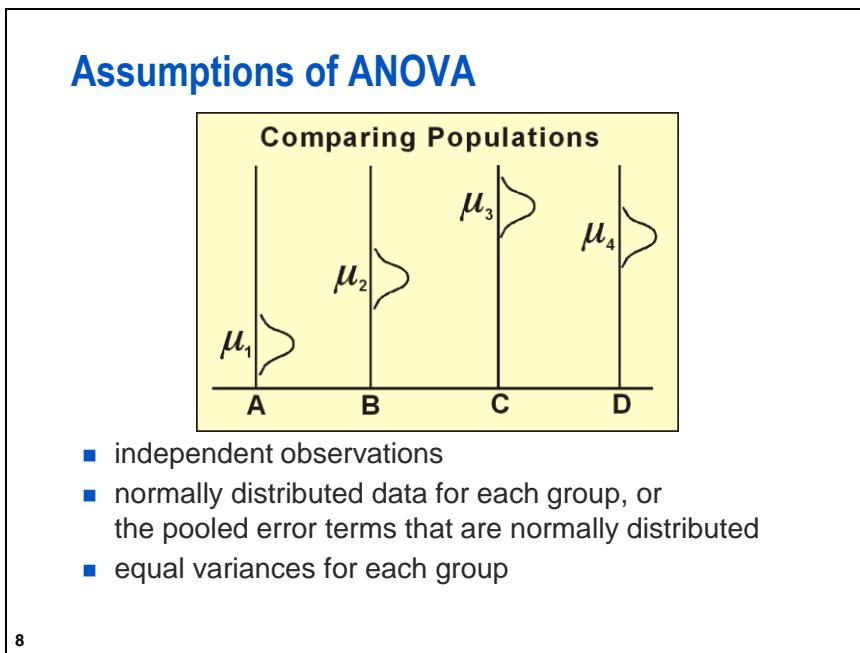
6

Recall the cubic model that you fit to the **STAT2.paper** data set. When a linear regression model is fit to the data, you are assuming that the mean of the dependent variable (**Strength**) at each value of the independent variable (**Amount**) follows a mathematical relationship. This relationship is estimated and presented as a line on the graph. You are interested in estimating the nature of the relationship (the slopes), whether this relationship is significant (slopes not equal to zero), and to what extent the variations in the data are explained by your model.

In analysis of variance, the mean of the dependent variable (**Strength**) at each value of the independent variable (**Amount**) is also computed. However, there is no mathematical relationship defined between these average values and the values of the independent variable. As a matter of fact, the independent variable might not be on an interval scale. The independent variables can be numeric and character variables and they can be on a nominal or ordinal scale. When this is the case, a regression model might not be appropriate because the mathematical relationship between the mean of the dependent variable and the independent variables might not be defined. The independent variables group your data into different groups. The question of interest is whether the group means are equal to each other.



Analysis of variance (ANOVA) is a statistical technique that is used to compare the means of two or more groups of observations or treatments. You should have a continuous dependent variable and one or more discrete independent variables (also called *predictor* or *explanatory variables*). They separate your data into several groups. You want to evaluate the effects of these categorical variables to see whether they explain the variations in the response variable. You can perform a one-way or an *n*-way ANOVA. The *n* can be replaced with the number of categorical predictor variables in your model.



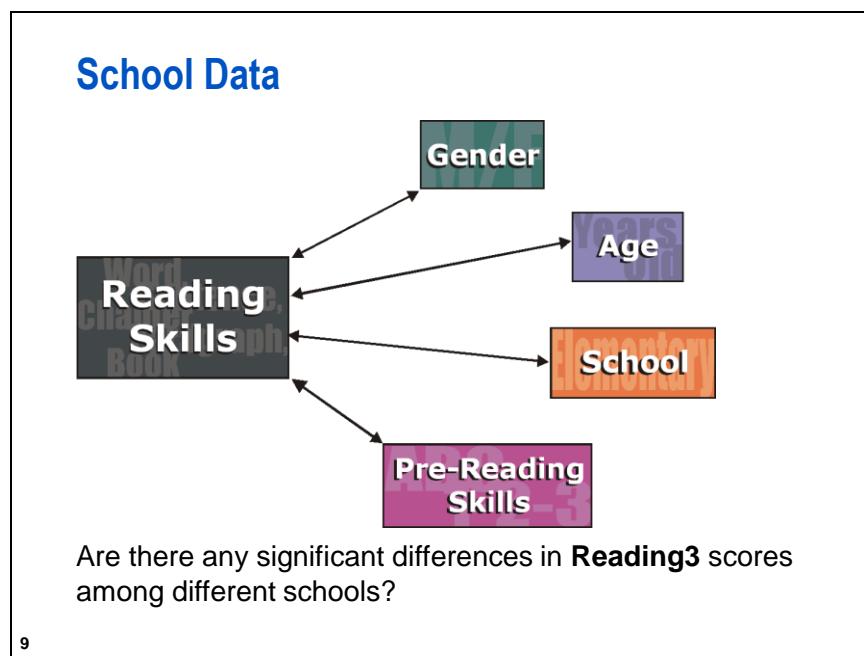
The assumption of independent observations means that no observations provide any information about any other observation that you collect. For example, measurements are **not** repeated on the same subject.

The assumption that the data for each group are approximately normal can be verified by examining plots of the data. In theory, the data for each group should be checked separately for normality. In practice, the data or the residuals as a whole are usually checked for normality. This assumption can be relaxed when the sample size is large enough.

The assumption of equal variances can be checked by looking at descriptive statistics and plots of the data and by conducting a test for equal variances.

If these assumptions are *not* valid, the probability of drawing incorrect conclusions from the analysis might be increased.

 Evaluation of model assumptions and remedial measures are discussed later.



9  
Data were collected by a school district to assess the reading skills progress of students in their first year of formal schooling. A random sample of students was selected from all the first-year students in the district. These are the variables in the **STAT2.school** data set:

<b>ID</b>	ID number of student	<b>Phonics2</b>	score on letter sound test in the winter
<b>Gender</b>	gender of student ( <i>F</i> , <i>M</i> )	<b>Words2</b>	score on word identification test in the winter
<b>Age</b>	student's age (rounded to nearest tenth of a year)	<b>Phonics3</b>	score on letter sound test in the spring
<b>School</b>	school student attends	<b>Words3</b>	score on word identification test in the spring
<b>Teacher</b>	name of student's teacher	<b>Reading2</b>	score on reading test in the winter
<b>Semesters</b>	number of semesters student attended in the district	<b>Fluency2</b>	score on reading fluency test in the winter
<b>Letters1</b>	score on letter identification test in the fall	<b>Reading3</b>	score on reading test in the spring
<b>Phonics1</b>	score on letter sound test in the fall	<b>Fluency3</b>	score on reading fluency test in the spring
<b>Words1</b>	score on word identification test in the fall		

To evaluate **School** and **Gender** effects, that is, to examine whether the average reading values are the same or different among different schools, genders, or combinations of schools and genders, you can conduct a two-way analysis of variance.

## Two-Way ANOVA Model

$$\text{Reading3} = \text{Base Level} + \text{School} + \text{Gender} + \frac{\text{School Unexplained}}{\text{Gender}} + \text{Variation}$$

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

10

$Y_{ijk}$  the observed **Reading** test score for **School  $i$** , **Gender  $j$** , and **Student  $k$**

$\mu$  the overall population mean of the response, **Reading3**

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

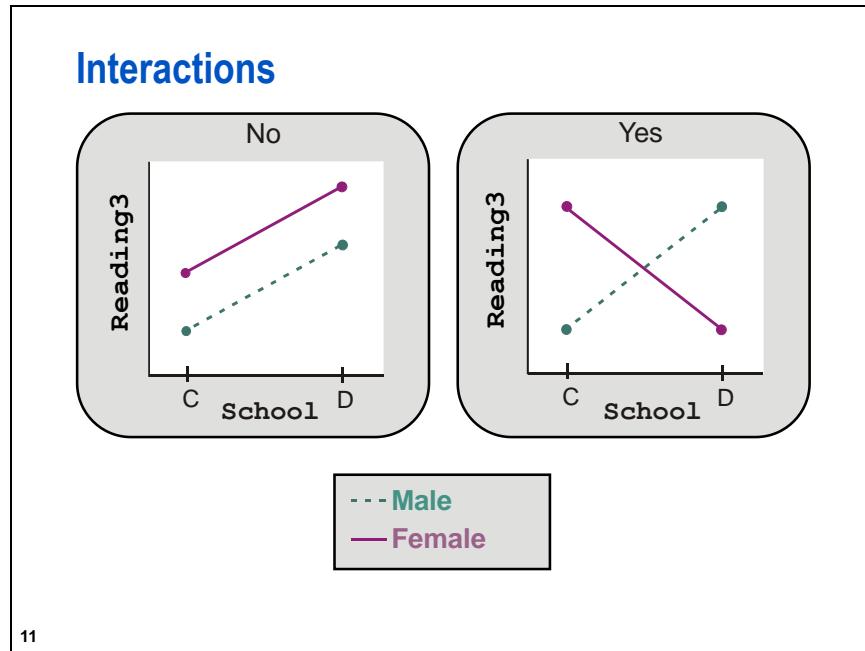
$\beta_j$  the effect of the  $j^{\text{th}}$  gender

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

$\varepsilon_{ijk}$  error term

In the model, the following conditions are assumed:

- The observations are independent.
- The data are normal for each treatment, or the pooled error terms are normal.
- The variances are approximately equal for each treatment group.



Interaction refers to the fact that the effect of one factor (**A**) on the response variable depends on the levels of the other factor (**B**). For example, when there is a significant interaction between **School** and **Gender**, the effect of **School** on **Reading3** is not the same for female students as it is for male students. In other words, the difference in average **Reading3** values between **School C** and **School D** for female students is not the same as the difference in average **Reading3** values between **School C** and **School D** for male students. Notice that the two lines do not need to cross for significant interaction to exist. When there is no interaction, the two lines should be close to parallel.

When you analyze an  $n$ -way ANOVA, the first consideration must be whether there is interaction between the factors. This is done by looking at the test for interaction on the ANOVA table.

If there is no interaction between the factors, then the tests for the individual factor effects might be considered in the table to determine the significance or nonsignificance of these factors.

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

## Interaction

- For observational studies, when the interactions are not significant, usually you can delete the interactions and analyze the main effects.
- When the interactions are significant, try to understand the interactions by examining the interaction plots.

12

For observational studies, when the interaction is not statistically significant, it is a common practice to delete the nonsignificant interaction from the model and then analyze the main effects. However, in some situations, experimenters like to leave the nonsignificant interactions in the model. These situations typically include when the data are from designed experiments, when the degrees of freedom for the residuals are not too small (for example, greater than 5), and when the  $F$  statistic for the nonsignificant interaction term is not too small (for example, greater than 2) (Nelder, Kutner, Wasserman, and Nachtsheim 1996).

When the interaction is statistically significant, you need to understand the nature of the interaction, usually by producing interaction plots, testing for simple effects, or examining pairwise comparisons.



When a predictor is continuous or ordinal, an ANOVA model yields algebraically the same results as treating the variable as continuous and using the appropriate number of polynomial terms. When you use the polynomial approach, your interactions are created using the individual polynomial terms. Thus, when you consider whether to remove an interaction, you might be able to remove some of the higher order interaction terms while retaining others. This flexibility might enable you to retain interaction terms using fewer degrees of freedom than would be required when treating the predictor as categorical.

## 3.02 Multiple Choice Poll

Which of the following are **false** for interpreting a significant interaction between **Gender** and **School**?

- The effect of **Gender** depends on the levels of **School**.
- The effect of **School** depends on the levels of **Gender**.
- The difference between levels of **Gender** is not the same across schools.
- The difference between levels of **School** is not the same across genders.
- The interaction is not intuitive to interpret, so you can ignore it.

13

## The GLM Procedure

General form of the GLM procedure:

```
PROC GLM options PLOTS (global-plot-options) =  

  (plot-request (specific-plot-options));  

CLASS variables;  

MODEL dependents=independents / options;  

MEANS effects / options;  

LSMEANS effects / options;  

STORE <OUT=item-store-name/LABEL=  

  'label' ;  

RUN;
```

15

The GLM procedure uses the method of least squares to fit general linear models. Among the statistical methods available in PROC GLM are regression, analysis of variance, analysis of covariance, multivariate analysis of variance, and partial correlation.

Selected GLM procedure statements:

**CLASS** specifies classification variables for the analysis.

**MODEL** specifies dependent and independent variables for the analysis.

MEANS	computes the arithmetic means and standard deviations of all continuous variables in the model (both dependent and independent) for each <i>effect</i> listed in the MEANS statement. You can specify only classification effects in the MEANS statement, that is, effects that contain only classification variables. Notice that the arithmetic means are not adjusted for other effects in the model. For adjusted means, use the LSMEANS statement.
LSMEANS	computes least squares means for each <i>effect</i> listed in the LSMEANS statement. You can specify only classification effects in the LSMEANS statement. In contrast to the MEANS statement, the LSMEANS statement performs multiple comparisons on interactions as well as main effects. You can also specify options to perform multiple comparisons. Least squares means are <i>predicted population margins</i> . That is, they estimate the marginal means over a balanced population. In a sense, least squares means are to unbalanced data as class and subclass arithmetic means are to balanced data.
STORE	requests that the procedure save the context and result of the statistical analysis for additional processing by PROC PLM.

#### Selected ODS GRAPHICS:

For a two-way ANOVA model, the GLM procedure produces an *interaction plot* of the response values, with the X axis representing one CLASS variable and the marker style representing the other. The predicted means are connected by lines.

For an LSMEANS statement with the PDIFF=ALL option, the GLM procedure produces a *diffogram*, which displays all pairwise LS-means differences and their significance.



## Two-Way Analysis of Variance

Consider the **STAT2.school** data set and the following variables in the data set:

**Reading3** score on reading test in the spring

**School** schools students attend

**Gender** gender of the students (*F, M*)

Recall that you conducted initial data explorations on **Reading3** previously. No particular concerns were noted about unusual data values or the distribution of the data. Now, use the GLM procedure to generate an interaction plot and the analysis of variance. The graph enables you to visually evaluate the interaction between the factors.

```
ods html style=statistical;
title "STAT2.SCHOOL DATA SET";
proc glm data=STAT2.school plots(unpack)=all;
  class school gender;
  model reading3 = school|gender;
run;
quit;                                         *ST203d01.sas;
```

Selected GLM statement options:

**PLOTS** (*global-plot-options*) = *plot-request (options)...* *plot-request (options)*

controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. For example:

**PLOTS= NONE**

**PLOTS=(DIAGNOSTICS RESIDUALS)**

PROC GLM Output

The GLM Procedure		
Class Level Information		
Class	Levels	Values
<b>School</b>	4	Cottonwood Dogwood Maple Pine
<b>Gender</b>	2	F M
<b>Number of Observations Read</b>		190
<b>Number of Observations Used</b>		179

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	7	35362.7215	5051.8174	3.50	0.0016
<b>Error</b>	171	246765.9489	1443.0757		
<b>Corrected Total</b>	178	282128.6704			

R-Square	Coeff Var	Root MSE	Reading3 Mean
0.125343	79.53939	37.98784	47.75978

The ANOVA table tests the hypothesis that the treatment group means are equal. The *p*-value given is 0.0016. Presuming an alpha equal to 0.05, you reject the null hypothesis and conclude that all treatment means are not equal. Which factors explain this difference?

The descriptive statistics indicate that the average test score for all observations is 47.76. The R square for this model is approximately 0.13. This indicates that this model accounts for approximately 13% of the variability in **Reading3** scores in the school district.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
<b>School</b>	3	14130.80515	4710.26838	3.26	0.0228
<b>Gender</b>	1	4026.54944	4026.54944	2.79	0.0967
<b>School*Gender</b>	3	17205.36689	5735.12230	3.97	0.0091

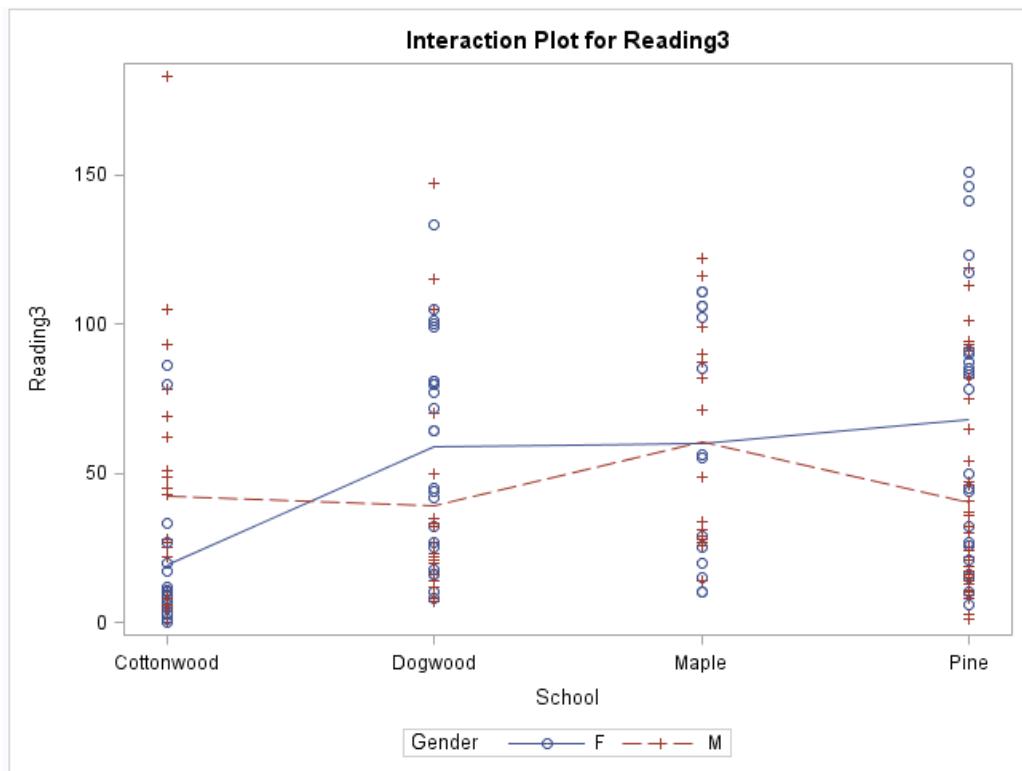
Source	DF	Type III SS	Mean Square	F Value	Pr > F
<b>School</b>	3	17905.24929	5968.41643	4.14	0.0073
<b>Gender</b>	1	1578.63006	1578.63006	1.09	0.2971
<b>School*Gender</b>	3	17205.36689	5735.12230	3.97	0.0091

There are actually four types of sums of squares available in PROC GLM. However, for one-way analysis of variance, all types of sums of squares are the same. For two-way analysis of variance, all types of sums of squares are the same *if* the data are balanced. *Balanced data* are data that have an equal sample size for all treatment combinations. The four types of sums of squares are generally not identical for unbalanced data. Type I and Type III sums of squares are produced by default in PROC GLM and Type III sums of squares are the most commonly used sums of squares for unbalanced data ANOVA.

The sums of squares are used to test the null hypothesis that the effect of the terms in the model is insignificant. You should consider the test for the interaction first. The *p*-value is 0.0091. Presuming an alpha of 0.05, you reject the null hypothesis. You have sufficient evidence to conclude that there is an interaction between **School** and **Gender**.

The interaction is best illustrated by an interaction plot, which is produced by default for a two-way ANOVA.

## Partial PROC GLM Output



The interaction is evident from the graph. Also evident is the large amount of variability in the scores around each mean. Boys and girls have the same **Reading3** scores at Maple, but the scores for boys and girls seem to differ at all of the other schools. The difference in average **Reading3** values for female students at Cottonwood and Dogwood (approximately -40) does not appear to be the same as the difference in average **Reading3** values for the male students at Cottonwood and Dogwood (approximately 4). Differences in **Reading3** values can also be found between other schools. Which of the differences are statistically significant?

Because of the interaction, the tests for the individual factor effects might be misleading due to masking of these effects by the interaction. Therefore, testing the means for the two factors separately might lead to erroneous conclusions.

The LSMEANS statement in PROC GLM can be used to perform multiple comparisons on interactions.

```
proc glm data=STAT2.school;
  class school gender;
  model reading3 = school|gender;
  lsmeans school*gender / pdiff adjust=tukey cl;
run;
quit; *ST203d02.sas;
```

### Selected LSMEANS statement options:

PDIFF=*difftype* requests that *p*-values for differences of the LS-means be produced. The optional *difftype* specifies which differences to display. Possible values for *difftype* are ALL, CONTROL, CONTROLL, and CONTROLU. The ALL value requests all pairwise differences, and it is the default. The CONTROL value requests the differences with a control that, by default, is the first level of each of the specified LS-mean effects.

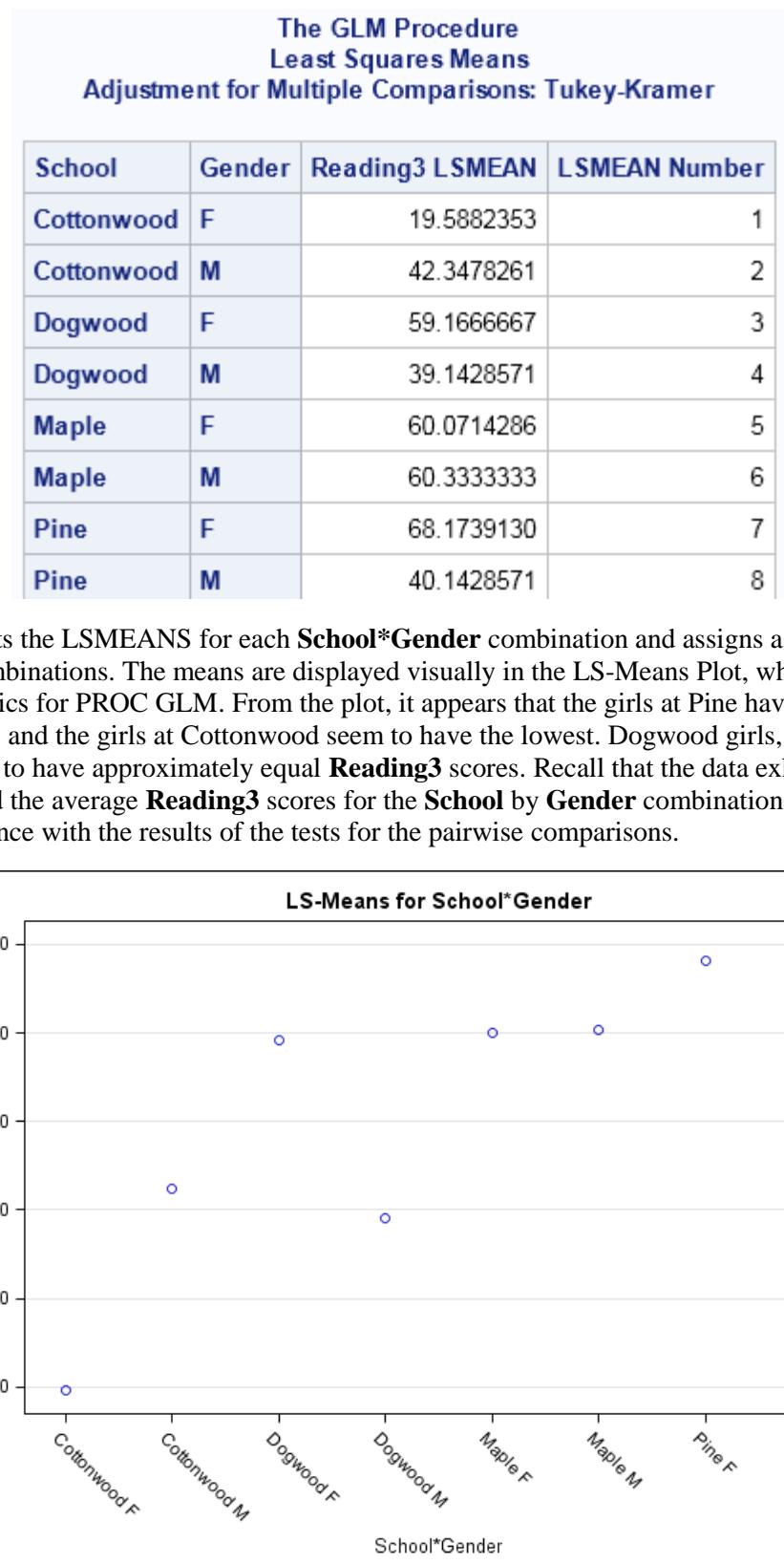
If you specify an LSMEANS statement with the PDIF option, the GLM procedure produces a plot appropriate for the type of LS-means comparison. For PDIF=ALL (which is the default if you specify only PDIF), the procedure produces a diffogram. The diffogram displays all pairwise LS-means differences and their significance. The display is also known as a *mean-mean scatter plot* (Hsu 1996). For PDIF=CONTROL, the procedure produces a display of each non-control LS-mean compared to the control LS-mean, with two-sided confidence intervals for the comparison. For PDIF=CONTROLL and PDIF=CONTROLU, a similar display is produced, but with one-sided confidence intervals. Finally, for the PDIF=ANOM option, the procedure produces an *analysis of means* plot, which compares each LS-mean to the average LS-mean.

ADJUST= specifies the multiple comparisons adjustment. If no *difftype* is specified, the default for the ADJUST= option is T (that is, no adjustment). For PDIF=ALL, ADJUST=TUKEY is the default. In all other instances, the default value for the ADJUST= option is DUNNETT.

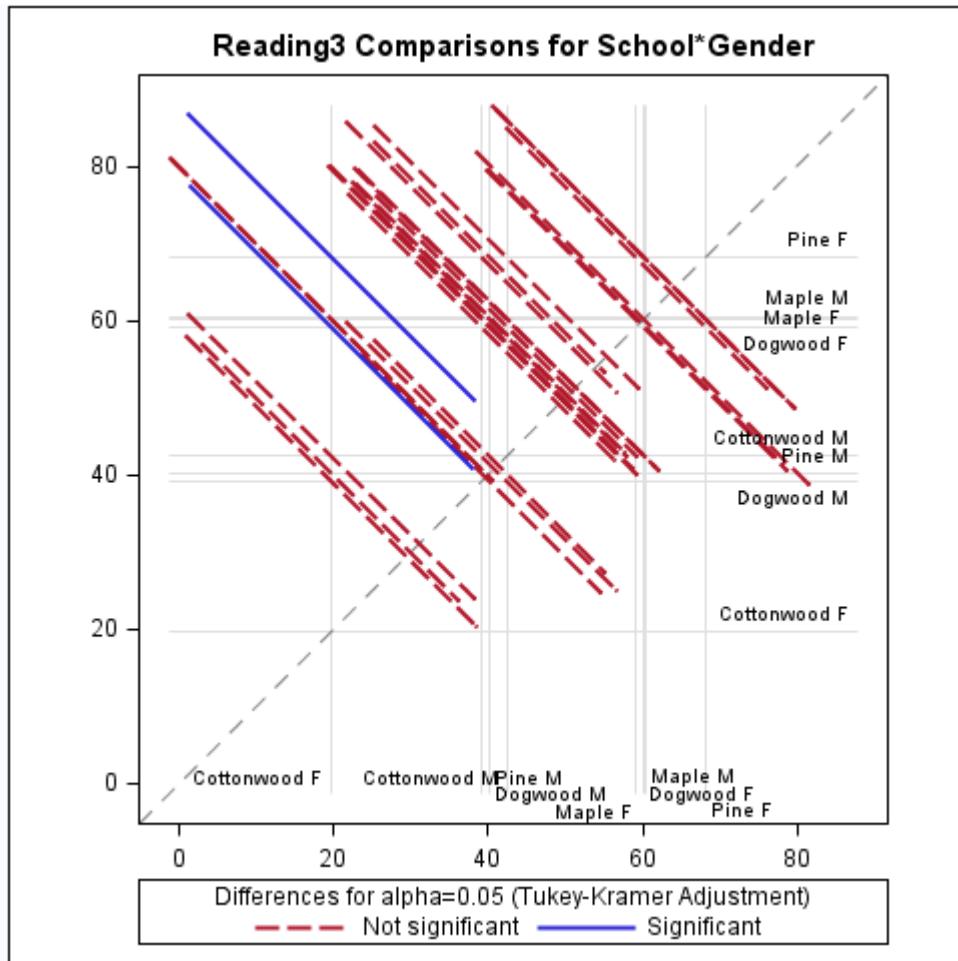
PDIFF= <i>difftype</i>	Default ADJUST=
Not specified	T (no adjustment)
ALL	TUKEY
CONTROL	DUNNETT
CONTROLL	
CONTROLU	

CL requests confidence limits for the individual LS-means. If you specify the PDIF option, confidence limits for differences between means are produced as well. You can control the confidence level with the ALPHA= option. Notice that, if you specify an ADJUST= option, the confidence limits for the differences are adjusted for multiple inference, but the confidence intervals for individual means are *not* adjusted.

## Partial PROC GLM Output



Because the PDIFF=ALL option was used in the LSMEANS statement, the ODS Graphics for PROC GLM includes a diffogram. This diffogram displays all pairwise LS-means differences and their significance. The display is also known as a *mean-mean scatter plot* or a *pairwise-difference plot*. Each line segment is centered at the intersection of two least squares means. The length of the line segments corresponds to the width of a 95% confidence interval for the difference between the two least squares means. The length of the segment is adjusted for the rotation. If a line segment crosses the dashed 45-degree line, the comparison between the two factor levels is not significant. Otherwise, it is significant. The horizontal and vertical axes of the plot are drawn in least squares means units, and the grid lines are placed at the values of the least squares means.



The background grid of the difference plot is drawn at the values of the least squares means for the eight **School\*Gender** combinations. These grid lines are used to find a particular comparison by intersection. Also, the labels of the grid lines indicate the ordering of the least squares means.

The diffogram visually displays the test results of the pairwise comparisons. It also confirms that at a Tukey-adjusted alpha level of 0.05, the **Reading3** scores for Cottonwood girls are significantly different from the scores for girls at Maple and Pine. No other comparisons are significant at this alpha level.

If you prefer, you can interpret the group differences through the tabular output.

<b>Least Squares Means for effect School*Gender</b> $\Pr >  t  \text{ for } H_0: LS\text{Mean}(i) = LS\text{Mean}(j)$ <b>Dependent Variable: Reading3</b>									
i/j	1	2	3	4	5	6	7	8	
<b>1</b>	0.5712	0.0265	0.7629	0.0687	0.0560	0.0024	0.5649		
<b>2</b>	0.5712		0.7972	1.0000	0.8668	0.8438	0.2967	1.0000	
<b>3</b>	0.0265	0.7972		0.6451	1.0000	1.0000	0.9922	0.5139	
<b>4</b>	0.7629	1.0000	0.6451		0.7517	0.7191	0.1893	1.0000	
<b>5</b>	0.0687	0.8668	1.0000	0.7517		1.0000	0.9984	0.6873	
<b>6</b>	0.0560	0.8438	1.0000	0.7191	1.0000		0.9985	0.6431	
<b>7</b>	0.0024	0.2967	0.9922	0.1893	0.9984	0.9985		0.0910	
<b>8</b>	0.5649	1.0000	0.5139	1.0000	0.6873	0.6431	0.0910		

School	Gender	Reading3 LSMEAN	95% Confidence Limits	
Cottonwood	F	19.588235	1.401585	37.774886
Cottonwood	M	42.347826	26.712273	57.983380
Dogwood	F	59.166667	43.860320	74.473014
Dogwood	M	39.142857	22.779684	55.506030
Maple	F	60.071429	40.030716	80.112141
Maple	M	60.333333	40.972166	79.694501
Pine	F	68.173913	52.538360	83.809467
Pine	M	40.142857	28.572346	51.713368

This table gives the  $p$ -values for a test of the null hypothesis that the group means are equal.

Presuming an alpha equal to 0.05, the following conclusions can be drawn:

- The average **Reading3** test score for female students at Cottonwood is significantly different from that of female students at Dogwood and female students at Pine.

Additional significant differences can be found if the alpha level is set to be 0.10.

- The average **Reading3** test score for female students at Cottonwood is significantly different from that of female students at Maple and male students at Maple.
- The average **Reading3** test score for female students at Pine is significantly different from that of male students at Pine.

The table also shows the 95% confidence limits for the mean of each group. Recall that these limits are *not* adjusted to control the experimentwise error rate.

Least Squares Means for Effect School*Gender			
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)
1	2	-22.759591	-60.046027 14.526845
1	3	-39.578431	-76.533252 -2.623610
1	4	-19.554622	-57.588179 18.478935
1	5	-40.483193	-82.556014 1.589627
1	6	-40.745098	-82.041685 0.551488
1	7	-48.585678	-85.872114 -11.299242
1	8	-20.554622	-54.065527 12.956283
2	3	-16.818841	-50.835264 17.197583
2	4	3.204969	-31.980394 38.390331
2	5	-17.723602	-57.240440 21.793235
2	6	-17.985507	-56.674864 20.703849
2	7	-25.826087	-60.202482 8.550308
2	8	2.204969	-28.034724 32.444662
3	4	20.023810	-14.809942 54.857561
3	5	-0.904762	-40.108855 38.299332
3	6	-1.166667	-39.536535 37.203202
3	7	-9.007246	-43.023670 25.009177
3	8	19.023810	-10.806033 48.853652
4	5	-20.928571	-61.151123 19.293980
4	6	-21.190476	-60.600367 18.219414
4	7	-29.031056	-64.216418 6.154307
4	8	-1.000000	-32.156254 30.156254
5	6	-0.261905	-43.582918 43.059109
5	7	-8.102484	-47.619322 31.414353
5	8	19.928571	-16.047572 55.904715
6	7	-7.840580	-46.529936 30.848777
6	8	20.190476	-14.874731 55.255683
7	8	28.031056	-2.208637 58.270748

This table gives the 95% confidence limits for the differences between the groups. These intervals *are* adjusted with Tukey's method to control the experimentwise error rate. The only two comparisons that are significantly different from zero are the comparisons of group 1 (Cottonwood girls) to group 3 (Dogwood girls) and to group 7 (Pine girls).

Comparing group 1 to group 3, the average **Reading3** score for Cottonwood girls is almost 40 points lower (-39.58) than the average score for Dogwood girls. The simultaneous confidence interval estimates the difference in the scores for girls at these two schools to be between -76.53 points and -2.62 points.

Comparing group 1 to group 7, the average **Reading3** score for Cottonwood girls is almost 50 points lower (-48.59) than the average score for Pine girls. The simultaneous confidence interval estimates the difference in the scores for girls at these two schools to be between -11.30 and -85.87 points.

When there is a significant interaction, interpreting main effects might not be appropriate. It might be of interest to test the effect of one factor within each level of the other factor. These are known as *tests for simple effects*.

For example, you might want to determine whether there are significant differences between female students and male students for each **School**, or whether there are significant differences among the schools for each **Gender**. Although some of these tests might be included with the PDIFF option, if you are interested only in this type of test, it is easier to use the SLICE= option.

```
proc glm data=STAT2.school;
  class school gender;
  model reading3 = school|gender;
  lsmeans school*gender / slice=gender slice=school;
run;
quit;                                         *ST203d03.sas;
```

Selected LSMEANS statement option:

SLICE= specifies effects within which to test for differences between interaction LS-mean effects. This can produce what are known as tests of simple effects.

Partial PROC GLM Output

### The GLM Procedure Least Squares Means

School*Gender Effect Sliced by Gender for Reading3					
Gender	DF	Sum of Squares	Mean Square	F Value	Pr > F
F	3	25876	8625.421606	5.98	0.0007
M	3	5152.566673	1717.522224	1.19	0.3151

The *p*-values for **Gender F** (0.0007) is significant at the alpha = 0.05 level. Therefore, there is a significant difference among the schools' average **Reading3** scores for female students. However, there is *not* sufficient evidence to conclude that there is a difference among different schools for male students.

School*Gender Effect Sliced by School for Reading3					
School	DF	Sum of Squares	Mean Square	F Value	Pr > F
Cottonwood	1	5063.439962	5063.439962	3.51	0.0628
Dogwood	1	4490.673016	4490.673016	3.11	0.0795
Maple	1	0.496716	0.496716	0.00	0.9852
Pine	1	11677	11677	8.09	0.0050

Presuming an alpha level of 0.05, the average **Reading3** values for Pine school are significantly different between female students and male students. At a slightly higher alpha level (0.08), you find that the average **Reading3** values are significantly different between female students and male students for Cottonwood School and Dogwood School, respectively.



## Exercises

---

### 1. Generating a Two-Way Analysis of Variance

A computer service center has four technicians who specialize in repairing three brands of computer disk drives for desktop computers. The service center wants to study the effects of the technician and brand of the disk drive on the service time. The data are stored in the **STAT2.disks** data set. These are the variables in the data set:

- Technician** name of the technician (*Angela, Bob, Justin, or Karen*)  
**Brand** brand of disk drive (1, 2, or 3)  
**Time** time for repair (in minutes)
- a. Generate a two-way analysis of variance with **Time** as the dependent variable and **Technician** and **Brand** as the independent variables. Include the interaction between the independent variables in your model. Presuming a level of significance of 0.05, is the overall *F* test significant in your model? Is there a significant interaction?
  - b. Examine the interaction plot. Does this graph verify the conclusion reached in the test for interaction? Why or why not?
  - c. Is it appropriate to examine the tests for the main effects shown in the PROC GLM output?
  - d. Use the LSMEANS statement with the SLICE= option to determine whether there are differences between the technicians for each brand of disk drive. Also examine the differences between the brands of disk drive for each technician. What are your conclusions?

### 3.03 Multiple Answer Poll

The significant interaction between **Brand** and **Technician** indicates which of the following?

- a. The differences in the average repair time between the technicians differ across different brands.
- b. The differences in the average repair time between the brands differ across different technicians.
- c. The significant interaction is not something that you should worry about.

19

## 3.2 Postfitting Analyses

---

### Objectives

- Estimate relationships of interest using the STORE statement in PROC GLM and the LSMESTIMATE statement PROC PLM.

22

## The LSMESTIMATE Statement



The LSMESTIMATE statement enables you to estimate any linear combination of the least square means and test custom hypotheses of interest.

25

The MEANS and LSMEANS statements of the GLM procedure only enable you to test pairwise differences. However, there are times when pairwise differences are not the only tests of interest in an analysis. For example, you might want to test to see whether the average of a number of levels of an effect is equal to another level. The LSMESTIMATE statement enables you to estimate and test any linear combination of cell means.

## 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.

26

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>;
  LSMEANS <model-effects> </ options>;
  LSMESTIMATE model-effect <'label'> values
    <divisor=n><, ...<'label'> values
    <divisor=n>> </ options>;
  SHOW options;
  WHERE expression;
RUN;
```

27

The PLM procedure performs postfitting statistical analyses for the contents of a SAS item store that were previously created with the STORE statement in some other SAS/STAT procedure. An item store is a special binary file format defined by SAS. It is used to store and restore information with a hierarchical structure.

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 postfitting tasks.

The use of item stores and PROC PLM enables you to separate common postprocessing 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:

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.

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.

### Example 1: Hypothesis of Interest

Is the average **Reading3** value for female students at Cottonwood and Dogwood the same as the average **Reading3** value for female students at Maple and Pine?

## Writing LSMESTIMATE Coefficients – Approach 1

1. Write the hypothesis of interest in terms of the cell means.
2. Compute the coefficients for the LSMESTIMATE statement.

29

You begin by writing the hypothesis of interest in terms of the cell means. From this, you obtain the coefficients for the LSMESTIMATE statement.

## Cell Means

School	Gender		$\mu_{..}$
	F	M	
Cottonwood	$\mu_{11}$	$\mu_{12}$	$\mu_{1..}$
Dogwood	$\mu_{21}$	$\mu_{22}$	$\mu_{2..}$
Maple	$\mu_{31}$	$\mu_{32}$	$\mu_{3..}$
Pine	$\mu_{41}$	$\mu_{42}$	$\mu_{4..}$
	$\mu_{..1}$	$\mu_{..2}$	$\mu_{...}$

30

An easy way to begin with a two-way ANOVA is to make a table that lists the two factors and their levels in the order in which SAS reads them. The order for the factors is determined by the order in which they are entered into the CLASS statement. Generally, the first factor should be used as the row variable and the second factor as the column variable. The order of the factor levels is alphanumeric.

The body of the table represents the specific levels of the **School\*Gender** interaction term. The entries of the column on the right side of the table are the sums of the rows and represent the main effects of the schools. The entries of the row at the bottom are the sums of the columns and represent the main effect of the genders.

-  It is important to note that if a factor is entered into SAS as a character variable, then 11 is before 2. However, if the factor is a numeric variable, then 2 comes before 11. When in doubt about the order of the factor levels, examine the Class Level Information table in the PROC GLM or PROC GLMSELECT output.

## Hypothesis in Terms of Cell Means

Is the average **Reading3** value for female students at Cottonwood and Dogwood the same as the average **Reading3** value for female students at Maple and Pine?

$$\frac{1}{2}(\mu_{11} + \mu_{21}) = \frac{1}{2}(\mu_{31} + \mu_{41})$$

31

## Compute the Coefficients for the LSMESTIMATE Statement

$$\frac{1}{2}(\mu_{11} + \mu_{21}) = \frac{1}{2}(\mu_{31} + \mu_{41})$$

$$\frac{1}{2}\mu_{11} + \frac{1}{2}\mu_{21} - \frac{1}{2}\mu_{31} - \frac{1}{2}\mu_{41} = 0$$

**Coefficients:**  
**School\*Gender** 0.5 0 0.5 0 -0.5 0 -0.5 0;

32

The hypothesis is simplified and rewritten with zero on one side of the equation, with the levels of each factor in the correct order. The coefficients are then available and the LSMESTIMATE statement can be included in your program.

The fractional coefficients can be written as decimals if they are decimals that do ***not*** repeat. In this example,  $1/2$  can be written as 0.5. However, accuracy and precision would be lost by writing  $1/3$  as 0.33. In such situations, you can multiply all coefficients by the common denominator to clear the fractions. The DIVISOR= option eliminates the need for fractions, which you see in the demonstration.

## Writing Coefficients – Approach 2

School	Gender		
	F	M	
Cottonwood			
Dogwood			
Maple			
Pine			

33

Another approach to writing the LSMESTIMATE statement is to begin with the two-way table based on the variables listed in the CLASS statement.

## Writing Coefficients – Approach 2

School	Gender		
	F	M	
Cottonwood	0.5		
Dogwood	0.5		
Maple	-0.5		
Pine	-0.5		

34

Express the hypothesis of interest in terms of coefficients for the corresponding cell means.

## Writing Coefficients – Approach 2

School	Gender		Sum
	F	M	
Cottonwood	0.5	0	
Dogwood	0.5	0	
Maple	-0.5	0	
Pine	-0.5	0	
<b>Sum</b>			

**Coefficients:**

**School\*Gender** 0.5 0 0.5 0 -0.5 0 -0.5 0;

36

Within the body of the chart, fill in zeros for any school by **Gender** combinations not involved in the hypothesis of interest. Write the coefficients for the LSMESTIMATE statement based on the filled-in table. For hypotheses on interaction terms, the LSMESTIMATE statement does not require coefficients for the main effects. For this reason, you are not obliged to fill in the margins of the table for this type of hypothesis.

## Example 2: Hypothesis of Interest

Is the **Reading3** value for Cottonwood the same as the **Reading3** value averaged across Dogwood, Maple, and Pine?

37



Although in the presence of significant interactions, the tests for main effects might be misleading, you are introduced to this example for illustrations of writing the LSMESTIMATE statement for main effects.

## Example 2 – Approach 1

$$\mu_{1\cdot} = \frac{1}{3}(\mu_{2\cdot} + \mu_{3\cdot} + \mu_{4\cdot})$$

$$\mu_{1\cdot} - \frac{1}{3}(\mu_{2\cdot} + \mu_{3\cdot} + \mu_{4\cdot}) = 0$$

If the test concerns only the marginal means (main effects), you need only to specify the coefficients associated with the main effects.

42

...

For a hypothesis involving main effects, the LSMESTIMATE statement does not require that you provide coefficients for the interaction term. It equally distributes the coefficients of the specified effect (**School**) to the levels of the higher-ordered effect (**School\*Gender** interaction) behind the scenes.

## Example 2 – Approach 1

$$\mu_{1\cdot} = \frac{1}{3}(\mu_{2\cdot} + \mu_{3\cdot} + \mu_{4\cdot})$$

$$\mu_{1\cdot} - \frac{1}{3}(\mu_{2\cdot} + \mu_{3\cdot} + \mu_{4\cdot}) = 0$$

If the test concerns only the marginal means (main effects), you need only to specify the coefficients associated with the main effects.

**Coefficients:**

**School 1** -0.333333 -0.333333 -0.333333;

or

**School 3** -1 -1 -1 / divisor=3;

43

For fractional coefficients that have repeating decimals, such as 1/3, accuracy and precision is lost when you write 1/3 as 0.33. In such situations, you must carry the decimal places to at least six digits or more, or multiply all coefficients by the common denominator to clear the fractions. If you multiply all coefficients by the common denominator in an LSMESTIMATE statement, you must use the DIVISOR= option to eliminate the need for fractions, but maintain the same magnitude of the difference that you are estimating.

## Example 2 – Writing Coefficients Approach 2

School	Gender		
	F	M	
Cottonwood			
Dogwood			
Maple			
Pine			

44

## Example 2 – Writing Coefficients Approach 2

School	Gender		Sum
	F	M	
Cottonwood			1
Dogwood			-0.333333
Maple			-0.333333
Pine			-0.333333
<b>Sum</b>			0

46

You need only provide the coefficients for the main effect of **School**. The LSMESTIMATE statement does not need coefficients for the interaction terms for a hypothesis involving only main effects.

## Example 2 – Writing Coefficients Approach 2

School	Gender		Sum
	F	M	
Cottonwood			1
Dogwood			-0.333333
Maple			-0.333333
Pine			-0.333333
<b>Sum</b>			0

### Coefficients:

**School 1** -0.333333 -0.333333 -0.333333;

or

**School 3** -1 -1 -1 / divisor=3;

47



## The LSMESTIMATE Statement

Generate an analysis of variance that tests and estimates the contrasts discussed in this section.

```
ods html close;
proc glm data=STAT2.school;
  class school gender;
  model reading3 = school|gender;
  store out=STAT2.schoolstore;
run;
quit;
ods html;
proc plm restore=STAT2.schoolstore;
  lsmeestimate school*gender 'Female Cottonwood&Dogwood vs. Female
                                Maple&Pine'
                    .5 0 .5 0 -.5 0 -.5 0 / elsm;
  lsmeestimate school 'Cottonwood vs. Dogwood, Maple and Pine'
                    1 -0.333333 -0.333333 -0.333333;
  Lsmeestimate school 'Cottonwood vs. Dogwood, Maple and Pine'
                    3 -1 -1 -1 / divisor=3;
run;                                              *ST203d04.sas;
```

The ODS statements before and after PROC GLM close and open the HTML destination respectively. Because PROC GLM is run with a STORE statement, you do not view the results now but use them for postfitting analyses. The STORE statement requests that PROC GLM save the context and results of the statistical analysis for use with PROC PLM.

Selected PROC PLM statement option:

**RESTORE=***item-store-specification*

specifies the source item store for processing. This option is required because, in contrast to SAS data sets, there is no default item store. An item-store-specification consists of a one- or two-level name as with SAS data sets. As with data sets, the default library association of an item store is with the Work library, and any stores created in this library are deleted when the SAS session concludes.

Selected PROC PLM statement:

**LSMESTIMATE**      provides a mechanism for obtaining custom hypothesis tests among least squares means.

Selected LSMESTIMATE statement options:

**DIVISOR=***number*    specifies a value by which to divide all coefficients so that fractional coefficients can be entered as integer numerators.

**ELSM**                requests that the K matrix coefficients be displayed. These are the coefficients that apply to the LS-means. This option is useful to ensure that you assigned the coefficients correctly to the LS-means.

## PROC PLM Output

The PLM Procedure		
Store Information		
<b>Item Store</b>		STAT2.SCHOOLSTORE
<b>Data Set Created From</b>		STAT2.SCHOOL
<b>Created By</b>		PROC GLM
<b>Date Created</b>		04SEP15:10:21:48
<b>Response Variable</b>		Reading3
<b>Class Variables</b>		School Gender
<b>Model Effects</b>		Intercept School Gender School*Gender

Class Level Information		
Class	Levels	Values
School	4	Cottonwood Dogwood Maple Pine
Gender	2	F M

The Store Information table and Class Level Information table are produced by default for the stored GLM model.

Least Squares Means Estimate Coefficients			
Effect	School	Gender	Row1
School*Gender	Cottonwood	F	0.5
School*Gender	Cottonwood	M	
School*Gender	Dogwood	F	0.5
School*Gender	Dogwood	M	
School*Gender	Maple	F	-0.5
School*Gender	Maple	M	
School*Gender	Pine	F	-0.5
School*Gender	Pine	M	

The output above is generated by the ELSM option in the LSMESTIMATE statement. This estimate is designed to compute the difference in female students' average test scores at Cottonwood and Dogwood, to the average of female student test scores at Maple and Pine. The output shows the order of the least squares means and enables you to check to see whether you created the intended estimate.

Least Squares Means Estimate						
Effect	Label	Estimate	Standard Error	DF	t Value	Pr >  t
School*Gender	Female Cottonwood&Dogwood vs. Female Maple&Pine	-24.7452	8.8152	171	-2.81	0.0056

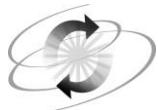
The LSMESTIMATE provides the estimate, its standard error and degrees of freedom, a *t*-statistic, and a *p*-value. The first LSMESTIMATE statement indicates that the average of the scores of the female students of Cottonwood and Dogwood are 24.75 points lower than the average of the scores of the female students at Maple and Pine. This difference is significantly different from 0.

Partial PROC PLM Output

Least Squares Means Estimate						
Effect	Label	Estimate	Standard Error	DF	t Value	Pr >  t
School	Cottonwood vs. Dogwood, Maple and Pine	-23.5371	6.9799	171	-3.37	0.0009

Least Squares Means Estimate						
Effect	Label	Estimate	Standard Error	DF	t Value	Pr >  t
School	Cottonwood vs. Dogwood, Maple and Pine	-23.5371	6.9799	171	-3.37	0.0009

The last two LSMESTIMATE statements produced identical results. The difference is in the syntax. One used the decimal coefficients, and the other one used the integer coefficients with the DIVISOR= option. The results show that the **Reading3** test scores for Cottonwood are 23.54 below the average **Reading3** test scores for Dogwood, Maple, and Pine. This difference is significantly different from zero (*p*-value=0.0009).



## Exercises

---

### 2. Writing LSMESTIMATE Statements

- a. Continue the analysis on the **STAT2.disks** data set and use an LSMESTIMATE statement to compare the average service time for *Bob* for **Brand 2** with the average service time for *Justin* for Brand 2. Use the ELSM option to verify that your coefficients are correct. Is the average service time significantly different for these two technician and brand combinations?
- b. Use an LSMESTIMATE statement to compute the difference between the lowest and highest average service times: *Angela* for **Brand 2** and *Karen* for **Brand 3**. Use the ELSM option to verify that your estimate coefficients are correct. What is the estimate of the difference between the two service times? Are they significantly different?

### 3.05 Multiple Choice Poll

Assume that the CLASS statement reads as follows:

CLASS SCHOOL GENDER

Which of the following statements is correct for comparing  
**Reading3** values for Male Cottonwood and Male Pine?

- a. LSMESTIMATE School\*Gender 1 0 0 -1;
- b. LSMESTIMATE School\*Gender 0 1 0 0 0 0 0 -1;
- c. LSMESTIMATE School 1 0 0 -1 Gender 0 1;

49

## 3.3 Contrasts and Estimates (Self-Study)

### Objectives

- Test hypotheses of interest using the CONTRAST statement.
- Estimate relationships of interest using the ESTIMATE statement.

53

## Contrasts



Contrasts enable you to perform any custom hypothesis test.

54

The MEANS and LSMEANS statements of the GLM procedure enable you only to test pairwise differences. However, there are times when pairwise differences are not the only tests of interest in an analysis. For example, you might want to test to see whether the average of a number of levels of an effect is equal to another level. The CONTRAST statement enables you to test any linear combination of cell means. The ESTIMATE statement enables you to estimate any linear combination, as well as conduct the test.

## CONTRAST and ESTIMATE Statements in PROC GLM

```
CONTRAST 'label' effect values / options;
ESTIMATE 'label' effect values / options;
```

- They are very similar in syntax.
- For one-degree-of-freedom tests, the ESTIMATE statement produces the difference and the significance of the difference. The CONTRAST statement produces only the significance of the difference.
- For more-than-one-degree-of-freedom tests, you can use only the CONTRAST statement.

55

- CONTRAST enables you to perform custom hypothesis tests by specifying an **L** vector or matrix for testing the univariate hypothesis  $\mathbf{L}\beta = 0$ . There is no limit to the number of CONTRAST statements that you can specify, but they must appear after the MODEL statement.
- ESTIMATE enables you to estimate linear functions of the parameters by multiplying the vector **L** by the parameter estimate vector **b**, which results in **Lb**. There is no limit to the number of ESTIMATE statements that you can specify, but they must appear after the MODEL statement.

In the CONTRAST and ESTIMATE statements:

- label* identifies the contrast (or estimate) on the output. A label is required for every specified contrast (or estimate). Labels must be enclosed in quotation marks.
- effect* identifies an effect that appears in the MODEL statement, or the INTERCEPT effect. The INTERCEPT effect can be used when an intercept is fitted in the model. You do not need to include all effects that are in the MODEL statement.
- values* are constants that are elements of the **L** vector associated with the effect.

### Example 1: Hypothesis of Interest

Is the average **Reading3** value for female students at Cottonwood and Dogwood the same as the average **Reading3** value for female students at Maple and Pine?

## Writing Contrasts – Approach 1

1. Write the hypothesis of interest in terms of the cell means.
2. Rewrite the hypothesis in terms of the model parameters.
3. Compute the coefficients for the CONTRAST statement.

57

If you begin by writing the hypothesis of interest in terms of the cell means, you reduce the possibility of attempting to test a combination of parameters that is not estimable. You can then rewrite the hypotheses in terms of the model parameters using the equation  $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ .

## Cell Means

School	Gender		
	F	M	
Cottonwood	$\mu_{11}$	$\mu_{12}$	$\mu_{1\cdot}$
Dogwood	$\mu_{21}$	$\mu_{22}$	$\mu_{2\cdot}$
Maple	$\mu_{31}$	$\mu_{32}$	$\mu_{3\cdot}$
Pine	$\mu_{41}$	$\mu_{42}$	$\mu_{4\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot \cdot}$

58

An easy way to begin with a two-way ANOVA is to make a table that lists the two factors and their levels in the order in which SAS reads them. The order for the factors is determined by the order in which they are entered into the CLASS statement. Generally, the first factor should be used as the row variable and the second factor as the column variable. The order of the factor levels is alphanumeric.



It is important to note that if a factor is entered into SAS as a character variable, then 11 is before 2. However, if the factor is a numeric variable, then 2 comes before 11. When in doubt about the order of the factor levels, examine the Class Level Information table in the PROC GLM or GLMSELECT output.

## Hypothesis in Terms of Cell Means

Is the average **Reading3** value for female students at Cottonwood and Dogwood the same as the average **Reading3** value for female students at Maple and Pine?

$$\frac{1}{2}(\mu_{11} + \mu_{21}) = \frac{1}{2}(\mu_{31} + \mu_{41})$$

59

## Model Parameters

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

implies

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

60

The two-way ANOVA model shown above implies that the cell means can be written as the sum of the overall mean, the effect of the  $i^{th}$  level of the first factor, the  $j^{th}$  level of the second factor, and the interaction between the two factors. The estimates of these parameters are used to estimate the cell means.

The hypothesis must be simplified and rewritten with zero on one side of the equation. The terms must be written in the order in which they appear in the model, with the levels of each factor in the correct order as well. The CONTRAST or ESTIMATE statement can then be included in your program.

### Rewrite the Hypothesis in Terms of the Model Parameters

$$\begin{aligned} \frac{1}{2}(\mu_{11} + \mu_{21}) - \frac{1}{2}(\mu_{31} + \mu_{41}) &= 0 \\ \frac{1}{2}(\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}) \\ - \frac{1}{2}(\mu + \alpha_3 + \beta_1 + (\alpha\beta)_{31} + \mu + \alpha_4 + \beta_1 + (\alpha\beta)_{41}) &= 0 \end{aligned}$$

61

### Compute the Coefficients for the CONTRAST Statement

$$\begin{aligned} 0.5\alpha_1 + 0.5\alpha_2 - 0.5\alpha_3 - 0.5\alpha_4 + 0\beta_1 + 0\beta_2 \\ + 0.5(\alpha\beta)_{11} + 0(\alpha\beta)_{12} + 0.5(\alpha\beta)_{21} + 0(\alpha\beta)_{22} \\ - 0.5(\alpha\beta)_{31} + 0(\alpha\beta)_{32} - 0.5(\alpha\beta)_{41} + 0(\alpha\beta)_{42} &= 0 \end{aligned}$$

#### Coefficients:

**School** 0.5 0.5 -0.5 -0.5

**School\*Gender** 0.5 0 0.5 0 -0.5 0 -0.5 0;

63

In this example, all of the  $\beta$  terms are canceled. Therefore, you do not need to include the factor **Gender** in the CONTRAST or ESTIMATE statement. The order for the levels for the interaction term **School\*Gender** is determined on a row-by-row basis. The fractional coefficients can be written as decimals if they are decimals that do **not** repeat. In this example,  $1/2$  can be written as  $0.5$ . However, accuracy and precision is lost by writing  $1/3$  as  $0.33$ . In such situations, you must carry the decimal places to at least six digits or more, or multiply all coefficients by the common denominator to clear the fractions. In an ESTIMATE statement, the DIVISOR= option eliminates the need for fractions as shown in the demonstration.

## Writing Contrasts – Approach 2

School	Gender		
	F	M	
Cottonwood			
Dogwood			
Maple			
Pine			

64

Another approach to writing the CONTRAST or ESTIMATE statement is to begin with a two-way table based on the variables listed in the CLASS statement.

## Writing Contrasts – Approach 2

School	Gender		Sum
	F	M	
Cottonwood	0.5		
Dogwood	0.5		
Maple	-0.5		
Pine	-0.5		

65

Next, you express the hypothesis of interest in terms of coefficients for the corresponding cell means.

## Writing Contrasts – Approach 2

School	Gender		Sum
	F	M	
Cottonwood	0.5	0	0.5
Dogwood	0.5	0	0.5
Maple	-0.5	0	-0.5
Pine	-0.5	0	-0.5
<b>Sum</b>	0	0	0

**Coefficients:**

**School** 0.5 0.5 -0.5 -0.5

**School\*Gender** 0.5 0 0.5 0 -0.5 0 -0.5 0;

67

Then you compute the sums of the coefficients for the corresponding columns and rows.  
Write the coefficients based on the completed table.

## Example 2: Hypothesis of Interest

Is the **Reading3** value for Cottonwood the same as the **Reading3** value averaged across Dogwood, Maple, and Pine?

68



Although in the presence of significant interactions, the tests for main effects might be misleading, this example illustrates writing the CONTRAST or ESTIMATE statement for main effects.

## Example 2 – Approach 1

$$\mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$$

$$\mu_1 - \frac{1}{3}(\mu_2 + \mu_3 + \mu_4) = 0$$

If the test concerns only the marginal means (main effects), you need only to specify the coefficients associated with the main effects. The coefficients for the interaction terms are assigned automatically.

$$\alpha_1 - \frac{1}{3}(\alpha_2 + \alpha_3 + \alpha_4) = 0$$

74

PROC GLM equally distributes the coefficients of the specified effect (**School**) to the levels of the higher-ordered effect (**School\*Gender** interaction). For this example, PROC GLM assumes that the coefficients for **School\*Gender** are as follows:

0.5 0.5 -0.166667 -0.166667 -0.166667 -0.166667 -0.166667

## Example 2 – Writing Contrasts Approach 1

$$\alpha_1 - \frac{1}{3}(\alpha_2 + \alpha_3 + \alpha_4) = 0$$

**Coefficients:**

**School 1** -0.333333 -0.333333 -0.333333;

or

**School 3** -1 -1 -1; (for CONTRAST statement)

**School 3** -1 -1 -1 / divisor=3; (for ESTIMATE statement)

75

For fractional coefficients that have repeating decimals, such as 1/3, accuracy and precision is lost by writing 1/3 as 0.33. In such situations, you must carry the decimal places to at least six digits or more, or multiply all coefficients by the common denominator to clear the fractions. If you multiply all coefficients by the common denominator in an ESTIMATE statement, you must use the DIVISOR= option to eliminate the need for fractions, but maintain the same magnitude of the difference that you are estimating.

## Example 2 – Writing Contrasts Approach 2

School	Gender		Sum
	F	M	
Cottonwood			1
Dogwood			-0.333333
Maple			-0.333333
Pine			-0.333333
<b>Sum</b>			0

78

## Example 2 – Writing Contrasts Approach 2

School	Gender		Sum
	F	M	
Cottonwood	0.5	0.5	1
Dogwood	-0.166667	-0.166667	-0.333333
Maple	-0.166667	-0.166667	-0.333333
Pine	-0.166667	-0.166667	-0.333333
<b>Sum</b>	0	0	0

**Coefficients:**

**School 1** -0.333333 -0.333333 -0.333333;

or

**School 3** -1 -1 -1; (for CONTRAST statement)

**School 3** -1 -1 -1 / divisor=3; (for ESTIMATE statement)

79

You can equally distribute the marginal coefficients to the corresponding cells in the same row, as shown above in gray, and specify the coefficients accordingly in the CONTRAST or ESTIMATE statement. Alternatively, knowing that omitting the coefficients for the interaction term PROC GLM automatically equally distributes the marginal coefficients to the corresponding higher-ordered terms (for example, **School\*Gender**), you can omit the interaction term and the associated coefficients from the CONTRAST or ESTIMATE statement.

## Example 3 – Hypothesis of Interest

Are the average **Reading3** values for Cottonwood, Maple, and Pine the same?

$$\mu_{1.} = \mu_{3.} = \mu_{4.}$$

$$\mu_{1.} - \mu_{3.} = 0, \quad \mu_{1.} - \mu_{4.} = 0$$

- Only the CONTRAST statement can be used.
- The coefficients are the following:  
**School 1** 0 -1 0,  
**School 1** 0 0 -1;

83

Whereas the previous examples were single degree-of-freedom hypotheses, this is a multiple degree-of-freedom hypothesis with two degrees of freedom. (One degree of freedom is required for each set of equal marks in the expression.) The expression  $\mu_1=\mu_3=\mu_4$  defines three sets of equalities. They are  $\mu_1=\mu_3$ ,  $\mu_1=\mu_4$ , and  $\mu_3=\mu_4$ . You can select any two of these three equalities to define the contrast.



## Contrasts and Estimates

Generate an analysis of variance that tests and estimates the contrasts discussed in this section.

```
ods html style=journal;
proc glm data=STAT2.school;
  class school gender;
  model reading3=school|gender;
  contrast 'Female Cottonwood&Dogwood vs. Female Maple&Pine'
    school .5 .5 -.5 -.5
    school*gender .5 0 .5 0 -.5 0 -.5 0;
  estimate 'Female Cottonwood&Dogwood vs. Female Maple&Pine'
    school .5 .5 -.5 -.5
    school*gender .5 0 .5 0 -.5 0 -.5 0;
  estimate 'Cottonwood vs. Dogwood, Maple and Pine'
    school 1 -0.333333 -0.333333 -0.333333 / e;
  estimate 'Cottonwood vs. Dogwood, Maple and Pine'
    school 3 -1 -1 -1 / divisor=3 ;
  contrast 'Cottonwood vs. Maple vs. Pine'
    school 1 0 -1 0,
    school 1 0 0 -1;
run;
quit;                                         *ST203d05.sas;
```

Selected CONTRAST statement option:

- E displays the entire **L** vector. This option is useful in confirming the ordering of parameters for specifying **L**.

Selected ESTIMATE statement option:

- DIVISOR=*number* specifies a value by which to divide all coefficients so that fractional coefficients can be entered as integer numerators.

## Partial PROC GLM Output

The GLM Procedure	
Coefficients for Estimate Cottonwood vs. Dogwood, Maple and Pine	
	Row 1
Intercept	0
School Cottonwood	1
School Dogwood	-0.333333
School Maple	-0.333333
School Pine	-0.333333
Gender F	0
Gender M	0
School*Gender Cottonwood F	0.5
School*Gender Cottonwood M	0.5
School*Gender Dogwood F	-0.1666665
School*Gender Dogwood M	-0.1666665
School*Gender Maple F	-0.1666665
School*Gender Maple M	-0.1666665
School*Gender Pine F	-0.1666665
School*Gender Pine M	-0.1666665

The output above is generated by the E option in the ESTIMATE statement. This estimate is designed to compute the difference in students' average test scores between Cottonwood and Dogwood, Maple, and Pine (averaged across all three schools). The output shows the order of the parameters and enables you to check to see whether you created the intended contrast or estimate. Notice that the coefficients for higher-ordered terms are assigned automatically by the procedure.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Female Cottonwood&Dogwood vs. Female Maple&Pine	1	11371.11146	11371.11146	7.88	0.0056
Cottonwood vs. Maple vs. Pine	2	17887.36484	8943.68242	6.20	0.0025

Parameter	Estimate	Standard Error	t Value	Pr >  t
Female Cottonwood&Dogwood vs. Female Maple&Pine	-24.7452198	8.81524564	-2.81	0.0056
Cottonwood vs. Dogwood, Maple and Pine	-23.5371449	6.97991584	-3.37	0.0009
Cottonwood vs. Dogwood, Maple and Pine	-23.5371453	6.97991637	-3.37	0.0009

The output from all the CONTRAST statements in the program is grouped; the same is true for the ESTIMATE statement. The CONTRAST statement generates an *F* test with the null hypothesis that the contrast is significantly different from zero. For example, the *p*-value of 0.0056 for the first contrast indicates that there is a significant difference in the average **Reading3** test scores between female students at Cottonwood and Dogwood and the female students at Maple and Pine. The *p*-value of 0.0025 for the second contrast indicates that there is a significant difference in the average **Reading3** test scores among Cottonwood, Maple, and Pine.

The output from all the ESTIMATE statements in the program is grouped together. The first ESTIMATE statement provides identical test results to the first CONTRAST statement. However, the magnitude of the difference (-24.75) is produced by the ESTIMATE statement. The last two ESTIMATE statements produced identical results. The difference is in the syntax. One used the decimal coefficients; the other one used the integer coefficients with the DIVISOR= option. The ESTIMATE statements show that the **Reading3** test scores for Cottonwood are 23.54 below the average **Reading3** test scores for Dogwood, Maple, and Pine, and this difference is significantly different from zero (*p*-value=0.0009).



## Exercises (Self-Study)

---

### 3. Writing CONTRAST and ESTIMATE Statements

- a. Continue the analysis on the **STAT2.disks** data set and use a CONTRAST statement to compare the average service time for *Bob* for **Brand 2** with the average service time for *Justin* for **Brand 2**. Use the E option to verify that your contrast coefficients are correct. Is the average service time significantly different for these two technician and brand combinations?
- b. Use an ESTIMATE statement to compute the difference between the lowest and highest average service times: *Angela* for **Brand 2** and *Karen* for **Brand 3**. Use the E option to verify that your estimate coefficients are correct. What is the estimate of the difference between the two service times? Are they significantly different?

### 3.06 Poll

When you write the CONTRAST or the ESTIMATE statement, the trailing zeros can be omitted, but the leading zeros or intermittent zeros must be specified.

- True
- False

87

## 3.4 Evaluations of Model Assumptions and Remedial Measures

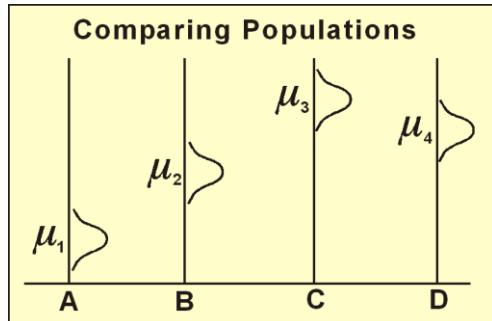
---

### Objectives

- Define Type I and Type II error rates.
- Evaluate ANOVA model assumptions using statistical techniques and graphs.
- Examine the consequences of assumption violations.

90

## Assumptions of ANOVA



- independent observations
- normally distributed data for each group, or the pooled error terms that are normally distributed
- equal variances for each group

91

The assumption of independent observations means that no observations provide any information about any other observations that you collect. For example, measurements are *not* repeated on the same subject.

The assumption that the data for each group are approximately normal can be verified by examining plots of the data. In theory, the data for each group should be checked separately for normality. In practice, residuals are usually checked for normality. Residuals are the differences between the observed and predicted values for each observation.

The assumption of equal variances can be checked by looking at descriptive statistics and plots of the data. A test for equal variances for one-way ANOVA can also be conducted.

If these assumptions are *not* valid, then the probability of drawing incorrect conclusions from the analysis might be increased.

 The assumption of normality can be relaxed when the sample size is large enough.

## Type I and Type II Error Rates

Decision	Reality	
	$H_0$ is true	$H_0$ is false
Fail to reject $H_0$	✓	Type II error ( $\beta$ )
Reject $H_0$	Type I error ( $\alpha$ )	✓ power=1- $\beta$

92

The Type I error rate, often denoted as  $\alpha$ , is the probability of wrongly rejecting the null hypothesis when  $H_0$  is true. The Type I error rate is also called the *significance level* of a test. A customary level for alpha for a hypothesis test is 0.05.

The Type II error rate, often denoted as  $\beta$ , is the probability of failing to reject the null hypothesis when  $H_0$  is false. The power of a statistical test is equal to  $1-\beta$ . This is the probability that you correctly reject the null hypothesis, or the probability of detecting a true effect.

You prefer that your tests have a low Type I error rate and a high power. However,  $\alpha$  and  $\beta$  cannot be determined independently of each other. They also depend on the sample size and the standard errors of the test of interest.

## Robustness of ANOVA

- ANOVA is robust against departures from normality, especially with large enough sample sizes.
- ANOVA is robust against unequal variances when sample sizes are equal.



93

See Miller (1997) for more information.

## Effects of Violations of the Assumptions of ANOVA: Independence

- Dependence
  - The Type I error rate might increase if the observations are positively correlated.



- The power might suffer if the observations are negatively correlated.



94

Violation of the independence assumption affects the accuracy of the standard errors and thus the results of significance tests. If the observations are positively correlated, then standard errors might be underestimated, which can lead to test statistics that are too large. Significance might be detected when it is not truly there, which results in an increase in Type I error rate. If the observations are negatively correlated, then the opposite situation arises. That is, standard errors are overestimated, test statistics might be too small, and true significance is not detected. Thus, the result for negatively correlated data is that power might suffer and the ability of the statistical test to detect a *true* difference is reduced.

## Effects of Violations of the Assumptions of ANOVA: Normality

- Nonnormality
  - The Type I error rate is not appreciably increased, but power might suffer.



95

ANOVA is robust against departures from normality, especially with a large enough sample size. This means that the probability of incorrectly rejecting the null hypothesis is not appreciably increased over the set value. However, power might suffer when the normality assumption is violated. This means that the probability of rejecting the null hypothesis when the null hypothesis should be rejected decreases. In other words, the ability of the statistical test to detect a *true* difference is reduced.

### Details

Lack of normality has very little effect on the significance level of the  $F$  test in the case of one-way ANOVA, and even less in  $n$ -way ANOVA. The asymptotic robustness of the  $F$  test follows from the multivariate central limit theorem.

## Effects of Violations of the Assumptions of ANOVA: Constant Variance

- Unequal Variances When Sample Sizes are Unequal

- » The Type I error rate might be increased if the smaller group has the larger variance.



- The power might be decreased if the larger group has the larger variance.



96

When sample sizes are equal, ANOVA is also robust against unequal variances. However, when sample sizes of the groups are unequal, the effect of unequal variances is more pronounced. If the variances of groups with a larger sample size are larger, the ANOVA loses power. If the variances of the groups with smaller sample sizes are larger, the probability of incorrectly rejecting the null hypothesis is increased.

More detailed information can be found from Miller (1997).

### Details

To see the effect of unequal variances with unequal sample sizes, examine the expected value of the  $F$  test:

$$E(F) = E\left(\frac{MS(treatment)}{MS(error)}\right) \approx \frac{\sum_{i=1}^I (N - n_i) \sigma_i^2}{\frac{N(I-1)}{\sum_{i=1}^I (n_i - 1) \sigma_i^2}}, \text{ where}$$

$N$  is the overall sample size,

$n_i$  is the sample size for group  $i$ ,

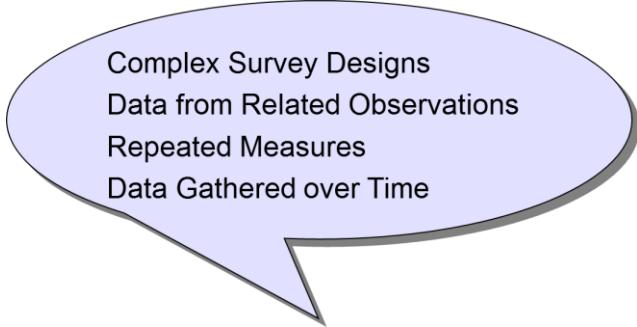
$I$  is the number of groups, and

$\sigma_i^2$  is the variance for group  $i$ .

In the case where the larger variance is associated with the smaller sample size, the numerator is larger than the case for equal variances and the denominator is smaller. This results in a lower  $p$ -value and inflated Type I error rates. In the case where the larger variance is associated with the larger sample size, the numerator is smaller and the denominator is larger, resulting in a lower  $p$ -value and loss of power. The results are approximately the same because the expected value of a ratio is not exactly equal to the ratio of the expected values.

## Evaluating Independence

- To evaluate independence, know the source of your data.



Complex Survey Designs  
Data from Related Observations  
Repeated Measures  
Data Gathered over Time

- Data from these types of studies are not independent.

97

Knowing how your data are generated and collected helps you evaluate the assumption of independence. Correlated observations can arise in data from a complex survey design, any type of clustered data, repeated measures on a given subject, or data gathered over time. Data can be correlated even when measurements are not taken on the same subjects. For example, members of the same household, littermates in animal studies, and colleagues in a company are examples of potentially non-independent data.

## Evaluating Normality

The following tools are useful for evaluating normality:

- normal probability plots
- histograms for residuals

Available via the ODS Graphics in PROC GLM

- descriptive statistics of residuals
- tests of normality for residuals

Available from PROC UNIVARIATE

98

Many formal tests of normality are available. The null hypothesis for these tests is that the data are normally distributed. The tests for normality are good tools, but they should not be used as the sole determining factor. The tests tend to be too liberal when the sample size is small and too conservative when the sample size is large.

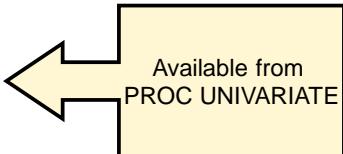
Normal probability plots of the residuals and histograms of the residuals are graphical methods to evaluate the normality of the data and are available as part of the ODS Graphics output for PROC GLM. Normal probability plots graph the distribution of the residuals against how the residuals would be distributed if they were normally distributed. If the residuals are normally distributed, the plot should be a straight line with a slope of one. Histograms can be constructed with normal curves superimposed on them. This enables you to visually compare the distribution of the residuals against normally distributed data with the same mean and variance.

Recall that the skewness and kurtosis statistics are equal to zero for normally distributed data. These statistics can be examined to assist in evaluating the normality of the data. You can output the residuals to a data set and obtain these statistics from PROC UNIVARIATE.

-  None of these tools should be used as the sole factor for determining the normality of the data. Each should be evaluated and a decision made by the analyst as to whether the data are distributed normally enough to conduct an analysis of variance.
-  Although the kurtosis statistic is usually equal to 3 for the normal distribution, it was rescaled in PROC UNIVARIATE to be 0 for normally distributed data.

## Evaluating the Homogeneity of Variances

The following tools are useful for evaluating the homogeneity of variances:

- statistical tests for one-way ANOVA  Available from PROC GLM
- residual plots
- descriptive statistics  Available from PROC UNIVARIATE

99

For a one-way analysis of variance, there are several formal statistical tests that were developed to evaluate the homogeneity of the variances. Among the tests that are available in PROC GLM are tests developed by Bartlett, Brown, and Forsythe, Levene, and O'Brien. It should be noted that Bartlett's test should be used only if the data are normally distributed, and the default Levene's test is considered to be the standard test.



Unless the group variances are extremely different or the number of groups is large, the usual ANOVA test is relatively robust when the groups are all about the same size. As Box (1953) notes, “To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!”

Plots of residuals versus predicted values can be used to visually examine whether the variances appear to be the same for different groups. Descriptive statistics can be calculated for each group and the variances compared to determine the relative magnitude of the variances in relation to each other.

## Details

Bartlett's test for homogeneity of variance (HOV) in a one-way ANOVA model (HOVTEST=BARTLETT in the MEANS statement) is a modification of the normal-theory likelihood ratio test. Although Bartlett's test has accurate Type I error rates and optimal power when the underlying distribution of the data is normal, it can be very inaccurate if that distribution is even slightly nonnormal (Box 1953). Therefore, Bartlett's test is not recommended for routine use.

All other HOV tests available in PROC GLM (Brown and Forsythe, Levene, and O'Brien) use the approach that leads to tests that are much more robust to the underlying distribution. This approach transforms the original values of the dependent variable to derive a *dispersion variable* and then performs analysis of variance on this variable. The significance level for the test of homogeneity of variance is the *p*-value for the ANOVA *F* test on the dispersion variable. The difference among the different tests depends on how the dispersion variable is derived.

Levene's test (Levene 1960) is widely considered to be the standard homogeneity-of-variance test (the HOVTEST=LEVENE option in the MEANS statement). O'Brien (1979) proposes a test (HOVTEST=OBRIEN) that is basically a modification of Levene's squared dispersion variable. Brown and Forsythe (1974) suggest using the absolute deviations from the group *medians* as the dispersion variable (HOVTEST=BF).

Simulation results (Conover et al. 1981, Olejnik and Algina 1987) show that, although all of these ANOVA-based tests are reasonably robust to the underlying distribution, the Brown-Forsythe test seems best at providing power to detect variance differences while protecting the Type I error probability. However, because the within-group medians are required for the Brown-Forsythe test, it can be resource intensive if there are very many groups or if some groups are very large.

### 3.07 Poll

You always examine plots, such as a normal probability plot, to evaluate normality because the normality test might be sensitive to sample size.

- True
- False

100



## Identifying Violations of ANOVA Assumptions

Recall the last model that you fit to the **STAT2.school** data set. Now, use the diagnostic plots that are available in PROC GLM to evaluate the ANOVA model assumptions. You can also output the residuals and use PROC UNIVARIATE to look at the descriptive statistics of the residuals.

```
proc glm data=STAT2.school plots(unpack)=diagnostics;
  class gender semesters school;
  model reading3 = gender semesters school gender*school;
  output out=check r=residuals p=predicted;
run;
quit;

goptions reset=all;
ods select moments BasicMeasures GoodnessOfFit;
proc univariate data=check;
  var residuals;
  histogram / normal;
run;                                         *ST203d06.sas;
```

Selected PROC GLM statement option:

**PLOTS** (*global-plot-options*) = (*plot-request (options)* ... *plot-request (options)*)

controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses from around the plot request. The available plots include the following:

**DIAGNOSTICS** <(*LABEL UNPACK*)> requests that a panel of regression diagnostics for the fit be displayed. The panel displays scatter plots of residuals, absolute residuals, studentized residuals, and observed responses by predicted values; studentized residuals by leverage; Cook's *D* by observation; a Q-Q plot of residuals; a residual histogram; and a residual-fit spread plot. The *LABEL* option displays labels on observations satisfying  $RSTUDENT > 2$ ,  $LEVERAGE > \frac{2p}{n}$ , and on the Cook's *D* plot,  $COOKSD > \frac{4}{n}$ , where  $n$  is the number of observations used in fitting the model, and  $p$  is the number of parameters in the model. The label is the first ID variable if the *ID* statement is specified. Otherwise, it is the observation number. The *UNPACK* option unpanels the diagnostic display and produces the series of individual plots that form the paneled display.

Selected PROC GLM statement:

<b>OUTPUT</b>	creates a new SAS data set that saves diagnostic measures calculated after fitting the model. All the variables in the original data set are included in the new data set, along with variables created in this statement. These new variables contain the values of a variety of diagnostic measures that are calculated for each observation in the data set.
---------------	---

The ODS SELECT statement specifies the selected tables to be displayed in the Output window.

## PROC UNIVARIATE Output

The UNIVARIATE Procedure Variable: residuals			
Moments			
<b>N</b>	179	<b>Sum Weights</b>	179
<b>Mean</b>	0	<b>Sum Observations</b>	0
<b>Std Deviation</b>	36.1366339	<b>Variance</b>	1305.85631
<b>Skewness</b>	0.78202265	<b>Kurtosis</b>	0.26740459
<b>Uncorrected SS</b>	232442.423	<b>Corrected SS</b>	232442.423
<b>Coeff Variation</b>	.	<b>Std Error Mean</b>	2.70097883

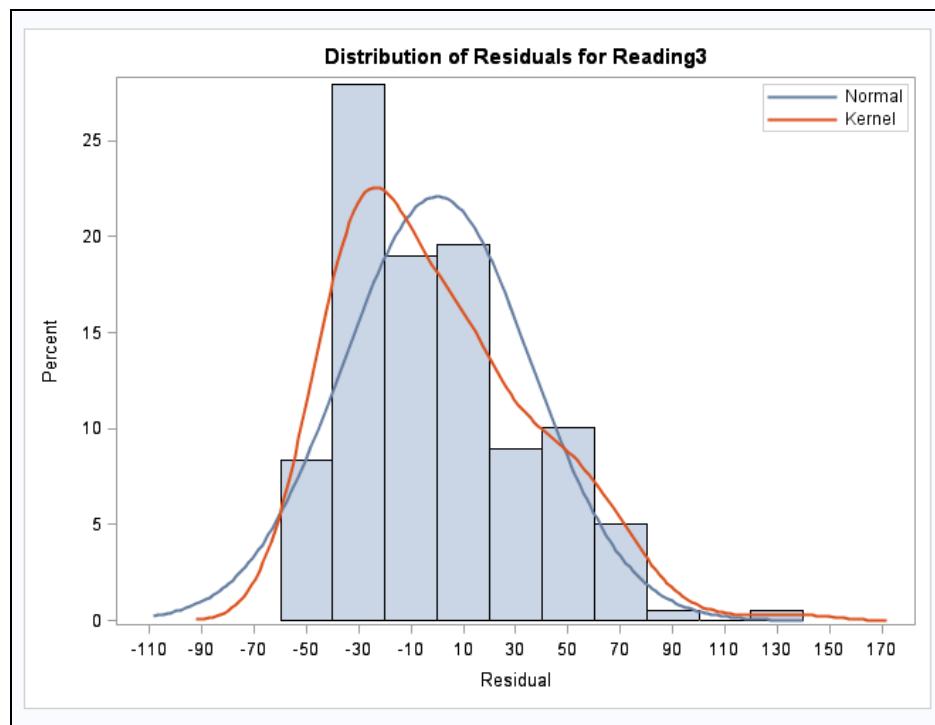
Basic Statistical Measures			
Location		Variability	
<b>Mean</b>	0.0000	<b>Std Deviation</b>	36.13663
<b>Median</b>	-8.4423	<b>Variance</b>	1306
<b>Mode</b>	-36.9996	<b>Range</b>	192.00301
		<b>Interquartile Range</b>	50.31893

The residuals are skewed to the right.

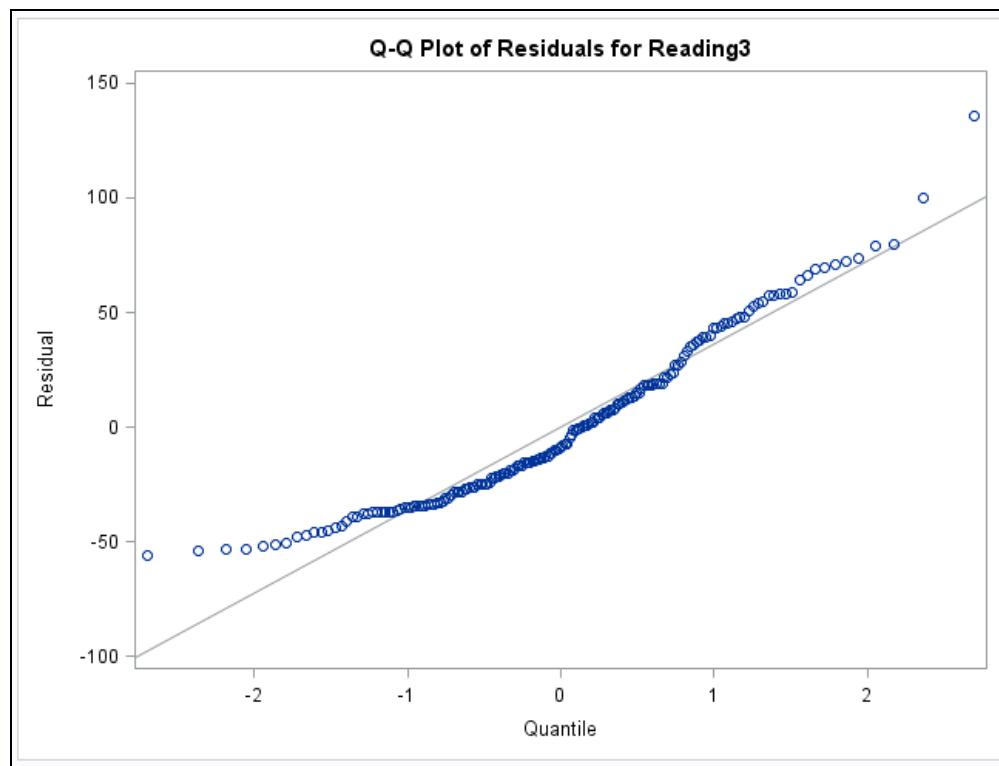
The UNIVARIATE Procedure Fitted Normal Distribution for residuals				
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
<b>Kolmogorov-Smirnov</b>	<b>D</b>	0.10497637	<b>Pr &gt; D</b>	<0.010
<b>Cramer-von Mises</b>	<b>W-Sq</b>	0.46280986	<b>Pr &gt; W-Sq</b>	<0.005
<b>Anderson-Darling</b>	<b>A-Sq</b>	2.80971012	<b>Pr &gt; A-Sq</b>	<0.005

The normality tests indicate that residuals are not normally distributed. However, the skewness and kurtosis statistics indicate that the departure from normality might be minor.

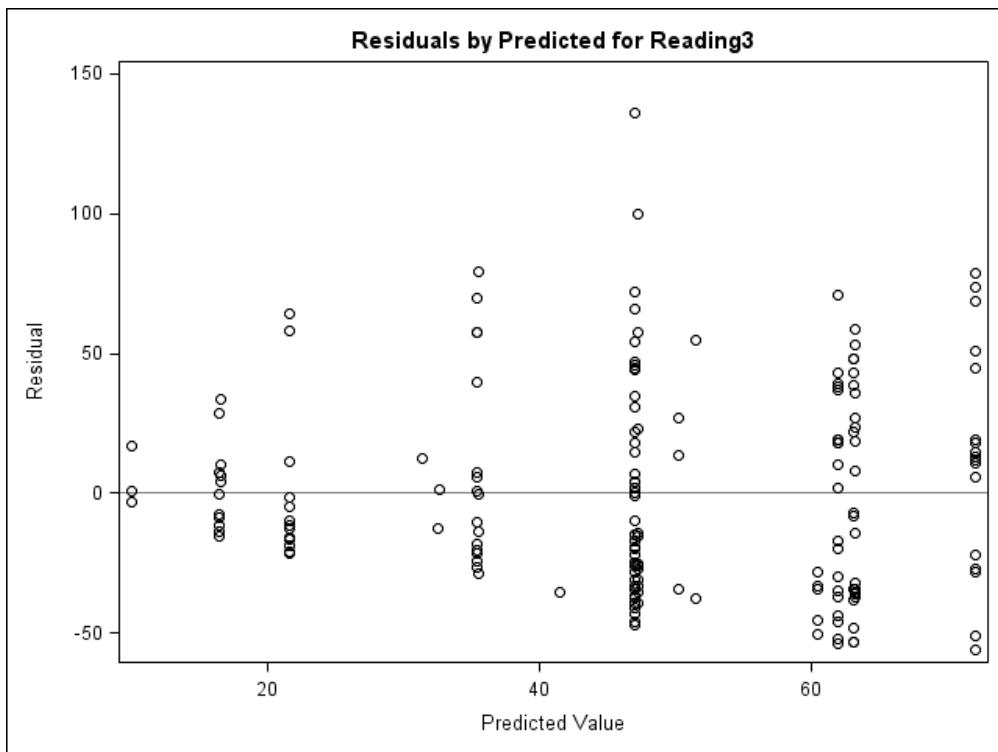
## PROC GLM Output



The histogram of the residuals indicates that they are skewed slightly to the right.



The normal probability plot also indicates that residuals are skewed slightly to the right.



The residual plot shows that the model seems to fit the data fairly well. The variances might not be the same for all groups. Notice that for some groups, the relatively large range of the residuals might be due to the fact that there are more observations in these groups.

The homogeneity of variances tests in PROC GLM are available only for one-way ANOVA. One approach to testing the homogeneity of variance assumption for two-way or higher-ordered ANOVA would be to create a new variable in a DATA step that is the actual treatment group for each observation. Essentially, this creates one factor with 24 levels ( $2 \times 4 \times 3$ ) for this example, and treats the analysis as a one-way ANOVA. This enables you to assess the equality of variances using the HOVTEST option in the MEANS statement in PROC GLM.

```

data school;
  set STAT2.school;
  group=compress(gender||school||semesters);
run;

ods select classlevels hovftest means;
proc glm data=school;
  class group;
  model reading3=group;
  means group / hovtest;
run;
quit; *ST203d06.sas;

```

Selected MEANS statement option:

HOVTEST performs the default test, which is Levene's squared residuals test for homogeneity (equality) of variances. The null hypothesis for this test is that the variances are equal. Other tests are available including Bartlett's test, O'Brien's test, and the Brown-Forsythe test.

## Partial PROC GLM Output

Class Level Information		
Class	Levels	Values
Group	23	FCottonwood4 FCottonwood6 FDogwood4 FDogwood6 FDogwood8 FMaple4 FMaple6 FMaple8 FPine4 FPine6 FPine8 MCottonwood4 MCottonwood6
		MCottonwood8 MDogwood4 MDogwood6 MDogwood8 MMaple4 MMaple6 MMaple8 MPine4 MPine6 MPine8

The reason that only 23 levels were created is that there is one empty cell for *FCottonwood8*.

## The GLM Procedure

Levene's Test for Homogeneity of Reading3 Variance					
ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	15	46196378	3079759	0.74	0.7420
Error	155	6.4598E8	4167588		

The null hypothesis for tests for homogeneity of variance is that the variances are equal. The alternative hypothesis is that the variances are not all equal. The *p*-value for Levene's test is 0.7420. Therefore, you do not reject the null hypothesis. The variances do not appear to be sufficiently unequal to cause concern for the validity of the ANOVA model under the equal variance assumption.

### The GLM Procedure

Level of Group	N	Reading3	
		Mean	Std Dev
FCottonwood4	3	15.000000	10.5830052
FCottonwood6	14	20.571429	27.8477020
FDogwood4	3	52.333333	32.1299445
FDogwood6	20	60.950000	37.3679781
FDogwood8	1	44.000000	.
FMaple4	1	106.000000	.
FMaple6	12	59.583333	41.5133239
FMaple8	1	20.000000	.
FPine4	5	22.000000	9.1378334
FPine6	17	85.411765	41.2008174
FPine8	1	6.000000	.
MCottonwood4	4	45.500000	42.0277686
MCottonwood6	17	43.647059	46.2722125
MCottonwood8	2	25.000000	28.2842712
MDogwood4	4	44.750000	48.2104760
MDogwood6	13	40.153846	41.9877882
MDogwood8	4	30.250000	13.4008706
MMaple4	1	14.000000	.
MMaple6	13	65.923077	35.9060383
MMaple8	1	34.000000	.
MPine4	9	43.888889	34.2689526
MPine6	27	45.555556	35.4480479
MPine8	6	10.166667	8.5654344

The MEANS statement produces the sample size, arithmetic means, and standard deviations of **Reading3** for each group. Because you have unbalanced data, you might want to examine the results from the LSMEANS statement for means comparisons.

## Dealing with Correlated Error Terms

When the **Independence Assumption** is violated, you can do the following:

-  use the MIXED, GLIMMIX, or GENMOD procedures (GEE) to model clustered data, including repeated measures
-  use the SAS SURVEY procedures to model data from complex survey designs

103

If the independence assumption is violated, other tools than PROC GLM must be used for the analysis of variance. These include the MIXED, GLIMMIX, or GENMOD procedures to model repeated measures. For data gathered from a complex survey design, you can use PROC SURVEYREG with a CLASS statement to perform the ANOVA.

## Dealing with Nonnormal Data

When the **Normality Assumption** is violated, you can do the following:

-  use the GENMOD or GLIMMIX procedure with the correct distribution and link functions
-  conduct a nonparametric analysis for one-way ANOVA
-  transform the response variable
-  rank the dependent variable and perform an ANOVA on ranks

104

If the assumption of normality is violated, transformation of the response variable often normalizes the data, and an analysis of variance can then be conducted on the transformed data. Transformations might also correct unequal variance problems.

When appropriate, you can use PROC GENMOD or PROC GLIMMIX with the appropriate distribution and link function to fit a generalized linear model to a nonnormal data.

When data are extremely skewed or there are extreme outliers, transformation of the data might not correct the problem. In this case, nonparametric ANOVA might be appropriate. The NPAR1WAY procedure can be used to fit a nonparametric one-way ANOVA. For two-way or higher-ordered nonparametric ANOVA, you might consider ranking your dependent variable and use PROC GLM to perform ANOVA on ranks (Iman 1988 and Iman 1982).

-  Detailed information about the NPAR1WAY procedure can be found in the SAS online documentation.

However, remember that ANOVA is robust to departures from normality, particularly with a large enough sample size.

## Dealing with Heteroscedasticity

When the Equal Variance Assumption is violated, you can do the following:

-  use the GENMOD, MIXED, or GLIMMIX procedures to model the nonconstant variance
-  transform the response variable
-  use a Welch's ANOVA for a one-way model
-  conduct a nonparametric ANOVA
-  transform the response variable

105

When variances are unequal, you can use the GENMOD, MIXED, or GLIMMIX procedures to model the nonconstant variances. Various approaches are available in these procedures and one, the GROUP=option in PROC GLIMMIX, is shown in a demonstration.

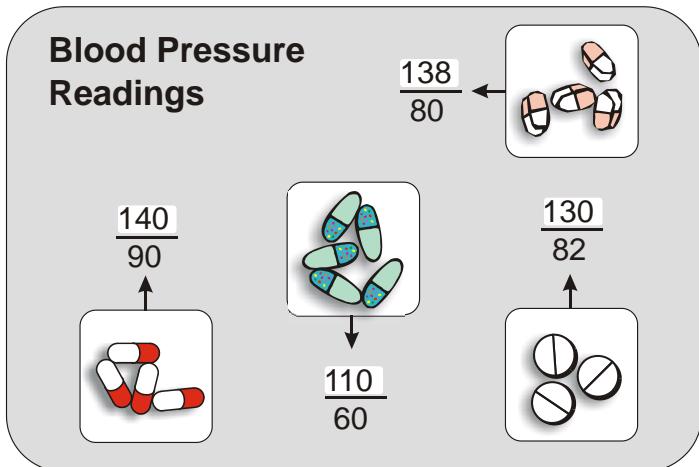
If the variances are unequal and it is a one-way ANOVA, then Welch's variance-weighted ANOVA can be used. It is specifically designed for unequal variance situations. It should be noted, however, that other GLM procedure statements such as the LSMEANS and CONTRAST statements are not valid for a Welch ANOVA. You might also use nonparametric ANOVA for this situation.

When variances are unequal, another approach is to transform the dependent variable to stabilize the variances.

-  ANOVA is robust to unequal variances when the sample sizes are equal.

## Accounting for Unequal Variances

### Example



106

Consider an experiment to evaluate the effect of four different drugs on blood pressure values. The drugs are administered to randomly selected subjects. The change in systolic blood pressure for each subject is recorded. You want to compare the average change in blood pressure for the different drugs. The data are stored in the **STAT2.pressure2** data set. These are the variables in the data set:

**BPChange** change in blood pressure values after the drug is administered to the subject

**Drug** the drugs included in the study (1, 2, 3, 4)

### The Model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

107

$Y_{ij}$  is the observed change in blood pressure for the  $j^{th}$  person on the  $i^{th}$  drug.

$\mu$  is the overall population mean of the response, **BPChange**.

$\alpha_i$  is the effect of the  $i^{th}$  **Drug**.

$\varepsilon_{ij}$  is the error term.

The assumption is  $\varepsilon \sim N(0, \sigma^2)$ .



## Accounting for Unequal Variances

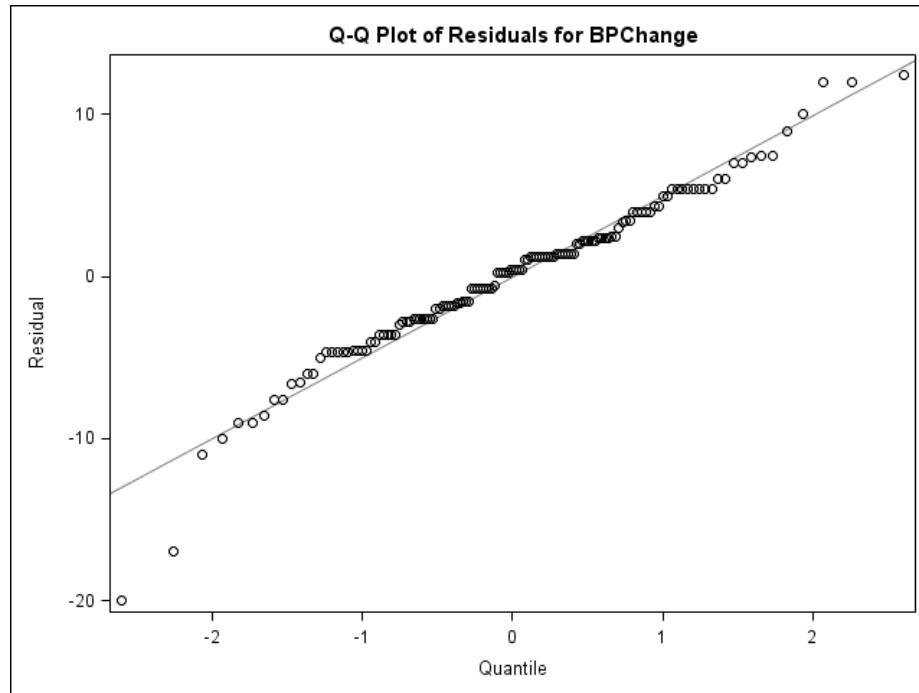
Use PROC GLM to generate a box plot of the data in the **STAT2.pressure2** data set. Examine the ANOVA to determine whether the drugs have different effects on the change in blood pressure values. Verify the assumptions of the ANOVA.

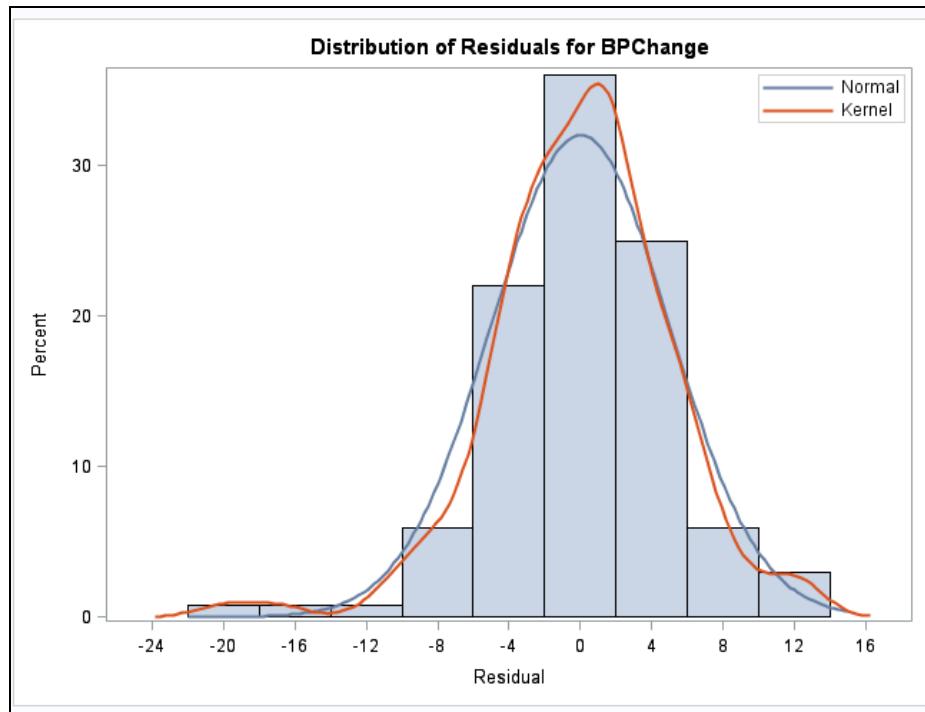
```
ods graphics / imagemap=on;
ods select modelanova overallanova QQPlot ResidualHistogram boxplot
      HOVFTest;
proc glm data=STAT2.pressure2 plots (unpack)=all;
  class drug;
  model bpchange=drug;
  means drug / hovtest;
  id drug;
run; quit; *ST203d07.sas;
```

Selected ODS Graphics statement option:

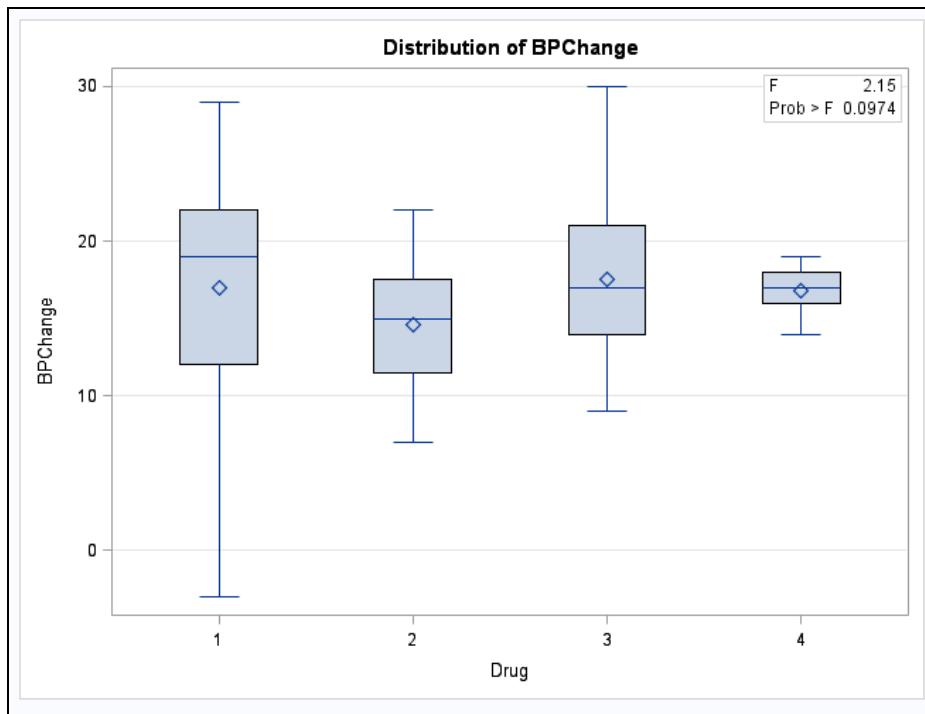
IMAGEMAP=OFF|ON specifies whether data tips are generated.

Partial PROC GLM Output





Based on the graphs, the residuals appear to be normally distributed.



The variances in **BPChange** do not seem to be constant across different types of drugs.

## Partial PROC GLM Output

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	163.639496	54.546499	2.15	0.0974
Error	132	3354.242857	25.410931		
Corrected Total	135	3517.882353			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Drug	3	163.6394958	54.5464986	2.15	0.0974

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Drug	3	163.6394958	54.5464986	2.15	0.0974

**Drug** is not significant ( $p$ -value=0.0974) in this model.

Levene's Test for Homogeneity of BPChange Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Drug	3	60380.8	20126.9	9.75	<.0001
Error	132	272616	2065.3		

The  $p$ -value for Levene's test is extremely small. Therefore, you reject the null hypothesis and conclude that the variances are not equal. This ANOVA assumption was violated.

In this case, you met the normal distribution assumption but not the equal variance assumption. If you attempt to transform the response variable, you might be able to stabilize the variances, but you might violate the assumption of normality in the process.

Analysis with the GLM procedure assumes independence, constant variance, and normality. Other SAS/STAT procedures (MIXED, GLIMMIX, GENMOD, and SURVEYREG) fit statistical models to data with correlations or nonconstant variability.

One approach is to adopt a model that allows for nonconstant variance. You can modify your assumption of constant variance to enable the variance to be estimated for each level of **Drug**.

## The Model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma_i^2)$$

109

The subscript on the variance term indicates that the variance is no longer constant, but varies across levels of **Drug**, which results in four variance estimates instead of one.

## Accounting for Unequal Variances

- The GLM procedure
  - models data that
    - are independent with constant variance
    - are from the normal distribution
  - has an option for Welch's ANOVA that can model nonconstant variance for one-way ANOVA only.
- The GLIMMIX and GENMOD procedures
  - model data that
    - might have correlations or nonconstant variance
    - are not necessarily from the normal distribution
  - have several tools for modeling nonconstant variance for a variety of scenarios.

110

PROC GLM performs statistical analyses for General Linear Models and assumes normality, independence, and constant variance. The procedure has only one tool for accounting for nonconstant variance: Welch's ANOVA for one-way analysis of variance. PROC GLIMMIX performs analyses for Generalized Linear Models and does not require the assumptions of normality, independence, or constant variance. The only assumption is that the response belongs to the exponential family of distributions, which includes the normal, gamma, binomial, Poisson, negative binomial, and several other distributions.

Although Welch's ANOVA is appropriate for this one-way ANOVA, PROC GLIMMIX provides a more general approach. The emphasis here is on using the GLIMMIX procedure to account for nonconstant variance where the normality assumption is met. (More information about PROC GLIMMIX is provided in a later chapter.)

## GLIMMIX Procedure

General form of the GLIMMIX procedure:

```
PROC GLIMMIX <options>;
  CLASS variables;
  MODEL response<(response options)>=<fixed-effects>
    </options>;
  RANDOM _RESIDUAL_ </options>;
  COVTEST '<label>' <test-specification> </options>;
RUN;
```

111

The PROC GLIMMIX and MODEL statements are required, and the MODEL statement must appear after the CLASS statement if a CLASS statement is included. The RANDOM and COVTEST statements can appear multiple times; the CLASS and MODEL statements can appear only once. Explanations of the following statements are provided in the context of the current use.

Selected GLIMMIX procedure statements:

**CLASS** names the classification variables to be used in the analysis. If the CLASS statement is used, it must appear before the MODEL statement.

**COVTEST** provides a mechanism to obtain statistical inferences for the covariance parameters. Significance tests are based on the ratio of (residual) likelihoods or pseudo-likelihoods. Confidence limits and bounds are computed as Wald or likelihood ratio limits. You can specify multiple COVTEST statements.

**MODEL** names the dependent variable and the fixed effects.

**RANDOM** The RANDOM \_RESIDUAL\_ statement defines **R**, the residual variance-covariance matrix.

Selected RANDOM statement option:

**GROUP** estimates covariance parameters by groups.

```
proc glimmix data=STAT2.pressure2;
  class drug;
  model bpchange=drug;
  random _residual_ / group=drug;
  covtest 'common variance' homogeneity;
run;                                *ST203d07.sas;
```

The RANDOM \_RESIDUAL\_ statement with the GROUP=DRUG option enables the residual variance to vary by levels of **Drug** to account for the possibility of nonconstant variance. The COVTEST statement with the HOMOGENEITY keyword tests whether the variances are constant across levels of **Drug**. The null hypothesis for this test is that the variances are constant with the alternative hypothesis that they are not constant across the levels of **Drug**.

The GLIMMIX Procedure	
Model Information	
<b>Data Set</b>	STAT2.PRESSURE2
<b>Response Variable</b>	BPChange
<b>Response Distribution</b>	Gaussian
<b>Link Function</b>	Identity
<b>Variance Function</b>	Default
<b>Variance Matrix</b>	Diagonal
<b>Estimation Technique</b>	Restricted Maximum Likelihood
<b>Degrees of Freedom Method</b>	Containment

The Model Information table provides basic information about the model including the data set and response variable. It indicates that the response is assumed to follow a Gaussian (that is, normal) distribution and that the identity link function is used. This means that the data are modeled on their original scale without any transformation. The default variance matrix for the normal distribution is diagonal. Thus, the off-diagonal elements, which are covariances, are equal to zero. This corresponds to the assumption of independent observations. The default estimation technique is Restricted Maximum Likelihood, which differs slightly from Maximum Likelihood used by PROC GLM. The Degrees of Freedom Method is Containment. The default settings are acceptable for the analysis. (Further information is available in the online documentation.)

Covariance Parameter Estimates			
Cov Parm	Group	Estimate	Standard Error
<b>Residual (VC)</b>	Drug 1	60.7879	14.9649
<b>Residual (VC)</b>	Drug 2	16.1774	4.1091
<b>Residual (VC)</b>	Drug 3	22.4874	5.4540
<b>Residual (VC)</b>	Drug 4	2.4168	0.5862

The previous analysis in PROC GLM assumed constant variance across the levels of **Drug**; it was estimated by mean squared error to be 25.41. This model does not assume constant variance, but estimates a separate variance for each level of **Drug**, as shown in the Covariance Parameter Estimates table. For drugs 1 through 4, the variances are estimated to be 60.79, 16.18, 22.49, and 2.42, respectively.

Tests of Covariance Parameters Based on the Restricted Likelihood					
Label	DF	-2 Res Log Like	ChiSq	Pr > ChiSq	Note
common variance	3	815.75	69.36	<.0001	DF

**DF:** P-value based on a chi-square with DF degrees of freedom.

The results of the COVTEST statement reject the null hypothesis of constant variance across the levels of **Drug** with a *p*-value less than 0.0001.

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Drug	3	132	3.23	0.0245

When you properly account for the unequal variances, the effect of **Drug** is now significant. Previously, when you assumed constant variance, it was not significant, with a *p*-value of 0.0974.

## Welch's ANOVA (Self-Study)

Welch's ANOVA (Welch 1951) was developed for a one-way ANOVA when the assumption of equal variances is not met. It uses weights for each group based on the variance for that group. It was shown that even in the case where the group variances are homogeneous, there is little loss of power with the use of Welch's ANOVA relative to the standard ANOVA *F* test (Kotz, Johnson, and Read eds. 1988). It is a variance-weighted ANOVA and is robust to the assumption of equal variances, but is valid only for one-way ANOVA. It is an option in the MEANS statement in PROC GLM.

```
/*WELCH'S ANOVA - Self-study*/
proc glm data=STAT2.pressure2;
  class drug;
  model bpchange=drug;
  means drug / welch;
run;
quit; *ST203d08.sas;
```

### The GLM Procedure

Welch's ANOVA for BPChange			
Source	DF	F Value	Pr > F
Drug	3.0000	3.16	0.0306
Error	62.2065		

With a *p*-value of 0.0306 for the overall F test for **Drug**, you reject the null hypothesis. As with the results from PROC GLIMMIX, there is sufficient evidence to conclude that there are differences among the four drugs. It seems that the unweighted ANOVA from PROC GLM seen previously lost power due to the unequal variances. Because the Welch's ANOVA is a weighted analysis, the unequal variances were accounted for in the results.



## Exercises

---

### 4. Evaluating Model Assumptions

- a. Request the DIAGNOSTICS plots from PROC GLM and examine the histogram and normal probability plot of the residuals from the two-way ANOVA model fit on the **STAT2.disks** data set. What do you conclude?
- b. Check the equal variance assumption by examining the residual plots generated in Exercise 3.a. Use a DATA step to create a single variable, **Group**, which includes both the level of **Technician** and the level of **Brand**.

Hint:

```
Group=compress(technician || brand);
```

Generate a one-way ANOVA with **Group** as the independent variable, and request a homogeneity of variance test. What do you conclude?

## 3.5 Chapter Summary

---

Analysis of variance (ANOVA) is a statistical technique used to compare the means of two or more groups of observations or treatments. For this type of problem, you have a continuous dependent variable and one or more discrete independent variables.

The null hypothesis for an ANOVA is that there is no treatment effect, or all of the population means are equal. The alternative hypothesis is that there are some treatment effects, or at least one population mean is different from at least one other population mean. For an ANOVA with more than two groups, after you determine that the population means are not all equal, you can determine which means are significantly different using multiple comparison tests.

When analyzing an  $n$ -way ANOVA, you first examine whether there is significant interaction between the factors. This is done by looking at the test for interaction on the ANOVA table. If there is no significant interaction between the factors, then you can delete the nonsignificant interaction and refit the model. If there is a significant interaction between the factors, then the tests for the individual factor effects (that is, tests for main effects) might be misleading due to masking of these effects by the interaction. In the case of a significant interaction, the LSMEANS statement in the GLM procedure can be used to perform multiple comparison tests to compare treatment means. You can also use the SLICE option to test the effect of one factor within each level of the other factor; this is known as a test for simple effects.

The STORE statement in PROC GLM enables you to store the results of your model for postfitting analyses using PROC PLM. The LSMESTIMATE statement in PROC PLM enables you to estimate and test any linear combination of the least square means. To write an LSMESTIMATE statement, you should first write the hypothesis of interest in terms of the cell means. From these, you can compute the coefficients for the statement. Alternatively, you can construct a two-way table and fill in the coefficients based on the hypothesis of interest.

When the sample sizes are unequal for treatment combinations, the data are unbalanced. In unbalanced data, it is important to use least squares means because the unequal number of observations is taken into account.

The assumptions of ANOVA are that the observations are independent, the data for each treatment are normally distributed, and all treatments have equal variances. Violations of these assumptions cause various problems with the analysis.

ANOVA is robust against departures from normality, especially with a large enough sample size. This means that the probability of incorrectly rejecting the null hypothesis is not appreciably increased over the set  $\alpha$ -level. However, power might suffer when the normality assumption is violated.

When sample sizes are equal, ANOVA is also robust against unequal variances. However, when sample sizes of the groups are unequal, the effect of unequal variances is more pronounced. If the variances of groups with a larger sample size are larger, the ANOVA loses power. If the variances of the groups with smaller sample sizes are larger, the probability of incorrectly rejecting the null hypothesis is increased.

The effect of dependence of the observations depends on the nature of the relationship. If observations are positively correlated, the probability of incorrectly rejecting the null hypothesis is increased. If the observations are negatively correlated, the power suffers.

The assumption of normality can be verified by using a combination of tests of normality, graphs, and descriptive statistics of the residuals. None of these tools should be used as the sole factor for determining the normality of the data. Each should be evaluated and a decision made by the analyst as to whether the data are distributed normally enough to conduct an analysis of variance.

The homogeneity of variances can be evaluated by conducting a formal statistical test for one-way ANOVA and by examining the residual plot.

If the assumptions of the ANOVA are violated, some alternatives are as follows:

- use the GENMOD or GLIMMIX procedure with the appropriate distribution and link function for lack of normality
- use the MIXED or GLIMMIX procedure to model the nonconstant variance or correlated errors
- use the SAS SURVEY procedures to model data from complex survey designs
- use a Welch's ANOVA in PROC GLM for a one-way ANOVA model with nonconstant variance
- conduct a nonparametric ANOVA
- transform the response variable for nonconstant variance or lack of normality

General form of the GLM procedure:

```
PROC GLM options PLOTS (global-plot-options)=  

                    (plot-request (specific-plot-options));  

CLASS variables;  

MODEL dependents=independents / options;  

LSMEANS effects / options;  

MEANS effects / options;  

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

RUN;
```

General form of the PLM procedure:

```
PROC PLM RESTORE=item-store-specification <options>;  

LSMESTIMATE <model-effects> <'label'> values <divisor=n> ...<'label'> values  

                    </options>;  

SHOW options;  

WHERE expression;  

RUN;
```

General form of GLIMMIX procedure:

```
PROC GLIMMIX <options>;  

CLASS variables;  

MODEL response<(response options)>=<fixed-effects>  

                    </options>;  

RANDOM _RESIDUAL_ </options>;  

COVTEST <'label'> <test-specification> </options>;  

RUN;
```

# 3.6 Solutions

## Solutions to Exercises

### 1. Generating a Two-Way Analysis of Variance

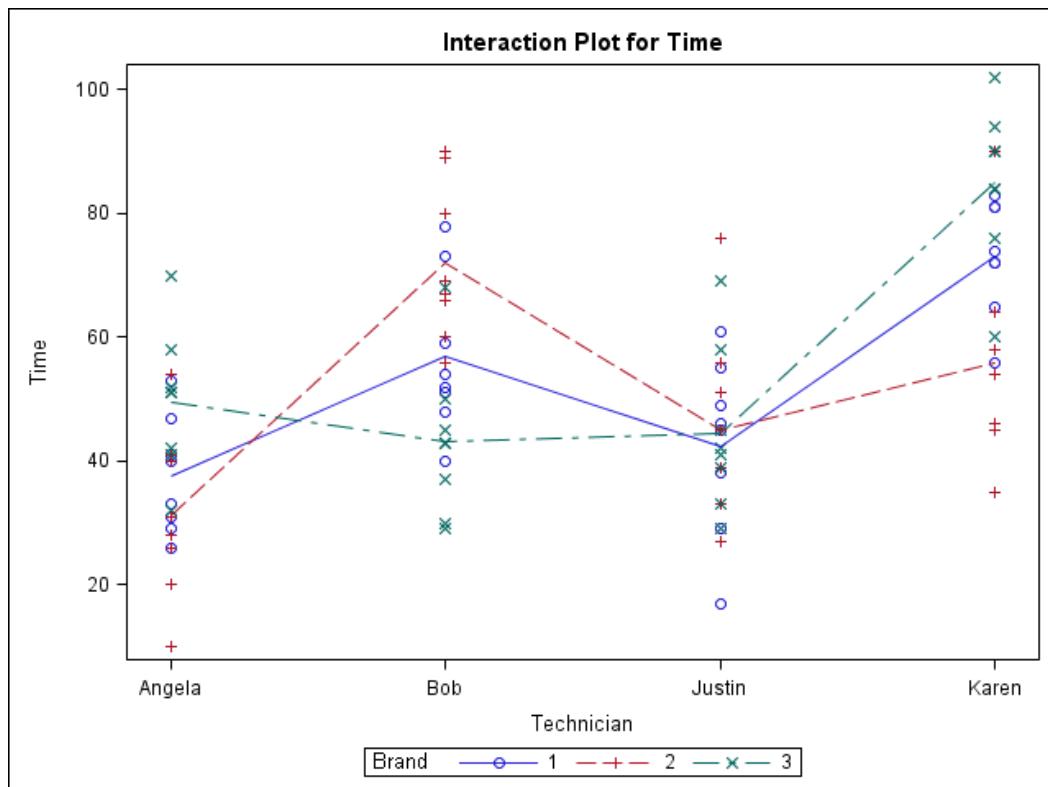
- a. Generate a two-way analysis of variance with **Time** as the dependent variable and **Technician** and **Brand** as the independent variables. Include the interaction between the independent variables in your model. Presuming a level of significance of 0.05, is the overall *F* test significant in your model? Is there a significant interaction?

```
ods html close; ods listing style=statistical;
proc glm data=STAT2.disks;
  class technician brand;
  model time=technician|brand;
run; quit; *ST203s01.sas;
```

The GLM Procedure										
Class Level Information										
Class	Levels	Values								
Technician	4	Angela Bob Justin Karen								
Brand	3	1 2 3								
Number of Observations Read			96							
Number of Observations Used			96							
The GLM Procedure										
Dependent Variable: Time										
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F					
Model	11	23048.45833	2095.31439	12.38	<.0001					
Error	84	14213.50000	169.20833							
Corrected Total	95	37261.95833								
R-Square Coeff Var Root MSE Time Mean										
0.618552 24.53377 13.00801 53.02083										
Source	DF	Type I SS	Mean Square	F Value	Pr > F					
Technician	3	14797.87500	4932.62500	29.15	<.0001					
Brand	2	343.14583	171.57292	1.01	0.3672					
Technician*Brand	6	7907.43750	1317.90625	7.79	<.0001					
Source	DF	Type III SS	Mean Square	F Value	Pr > F					
Technician	3	14797.87500	4932.62500	29.15	<.0001					
Brand	2	343.14583	171.57292	1.01	0.3672					
Technician*Brand	6	7907.43750	1317.90625	7.79	<.0001					

The  $p$ -value for the overall  $F$  test is less than 0.0001, which is less than the significance level of 0.05. Therefore, the test is significant and you can conclude that at least one treatment mean is different from one other treatment mean. The  $p$ -value for the test for interaction is also less than 0.0001. Therefore, the interaction is significant.

- b. Examine the interaction plot. Does this graph verify the conclusion reached in the test for interaction? Why or why not?



The graph verifies that there is an interaction between **Technician** and **Brand**. This is evident because the lines on the graph are not parallel. In particular, notice that *Bob* repairs **Brand 3** in the shortest amount of time, but **Brand 3** takes *Karen* the longest amount of time to repair.

- c. Is it appropriate to examine the tests for the main effects shown in the PROC GLM output?

The tests for the main effects in the PROC GLM output can be misleading in the presence of an interaction. Therefore, it is not appropriate to examine these tests. Instead, you should test for simple effects using the SLICE= option.

- d. Use the LSMEANS statement with the SLICE= option to determine whether there are differences between the technicians for each brand of disk drive. Also examine the differences between the brands of disk drive for each technician. What are your conclusions?

```
proc glm data=STAT2.disks;
  class technician brand;
  model time=technician|brand;
  lsmeans technician*brand / slice=brand slice=technician;
run;
quit;
ods listing close;
ods html; *ST203s01.sas;
```

## Partial PROC GLM Output

The GLM Procedure Least Squares Means					
Technician	Brand	Time	LSMEAN		
Angela	1	37.500000			
Angela	2	31.250000			
Angela	3	49.625000			
Bob	1	56.875000			
Bob	2	72.125000			
Bob	3	43.125000			
Justin	1	42.500000			
Justin	2	45.000000			
Justin	3	44.500000			
Karen	1	73.000000			
Karen	2	55.750000			
Karen	3	85.000000			

The GLM Procedure Least Squares Means					
Technician*Brand Effect Sliced by Brand for Time					
Brand	DF	Sum of Squares	Mean Square	F Value	Pr > F
1	3	6115.093750	2038.364583	12.05	<.0001
2	3	7159.093750	2386.364583	14.10	<.0001
3	3	9431.125000	3143.708333	18.58	<.0001

Based on the *p*-values shown here, the service time for at least one technician is different from one other technician for each brand of disk drive.

The GLM Procedure Least Squares Means					
Technician*Brand Effect Sliced by Technician for Time					
Technician	DF	Sum of Squares	Mean Square	F Value	Pr > F
Angela	2	1396.583333	698.291667	4.13	0.0195
Bob	2	3367.000000	1683.500000	9.95	0.0001
Justin	2	28.000000	14.000000	0.08	0.9207
Karen	2	3459.000000	1729.500000	10.22	0.0001

Within each of the technicians, except for *Justin*, the average service time for at least one brand of disk drive is different from the service times for the other brands.

## 2. Writing LSMESTIMATE Statements

- Continue the analysis on the **STAT2.disks** data set and use an LSMESTIMATE statement to compare the average service time for *Bob* for **Brand 2** with the average service time for *Justin* for **Brand 2**. Use the ELSM option to verify that your coefficients are correct. Is the average service time significantly different for these two technician and brand combinations?

Presuming that  $\alpha$  represents **Technician** and  $\beta$  represents **Brand**, the equation for this contrast in terms of cell mean is the following:

$$\mu_{22} - \mu_{32} = 0 \Rightarrow 0 * \mu_{11} + 0 * \mu_{12} + 0 * \mu_{13} + 0 * \mu_{21} + 1 * \mu_{22} + 0 * \mu_{23} + 0 * \mu_{31} - 1 * \mu_{32} + 0 * \mu_{33}$$

If you use approach 2, then consider the following two-way table:

<b>Technician</b>	<b>Brand</b>			<b>Sum</b>
	1	2	3	
Angela				
Bob		1		
Justin		-1		

<b>Technician</b>	<b>Brand</b>			<b>Sum</b>
	1	2	3	
Angela	0	0	0	
Bob	0	1	0	
Justin	0	-1	0	
Karen	0	0	0	
<b>Sum</b>				

```

title; ods listing close;
proc glm data=STAT2.disks;
  class technician brand;
  model time = technician|brand;
  store out=STAT2.diskstore;
run;
ods html;
proc plm restore=STAT2.diskstore;
  lsmeans technician*brand
    'Bob Brand 2 vs Justin Brand 2'
    0 0 0 0 1 0 0 -1 0 0 0 0 / elsm;
run; quit;
*ST203s02.sas;

```

Least Squares Means Estimate Coefficients			
Effect	Technician	Brand	Row1
Technician*Brand	Angela	1	
Technician*Brand	Angela	2	
Technician*Brand	Angela	3	
Technician*Brand	Bob	1	
Technician*Brand	Bob	2	1
Technician*Brand	Bob	3	
Technician*Brand	Justin	1	
Technician*Brand	Justin	2	-1
Technician*Brand	Justin	3	
Technician*Brand	Karen	1	
Technician*Brand	Karen	2	
Technician*Brand	Karen	3	

The output from the ELSM options confirms that you are comparing repair times for *Bob* on **Brand 2** to those for *Justin* on **Brand 2**.

Least Squares Means Estimate							
Effect	Label	Estimate	Standard Error	DF	t Value	Pr >  t	
Technician*Brand	Bob Brand 2 vs Justin Brand 2	27.1250	6.5040	84	4.17	<.0001	

The results indicate that *Bob* takes a little more than 27 minutes longer on average than *Justin* to repair **Brand 2**, and the difference is significantly different from 0.

- b. Use an LSMESTIMATE statement to compute the difference between the lowest and highest average service times: *Angela* for **Brand 2** and *Karen* for **Brand 3**. Use the ELSM option to verify that your estimate coefficients are correct. What is the estimate of the difference between the two service times? Are they significantly different?

```
proc plm restore=STAT2.diskstore;
  lsmeestimate technician*brand
    'Lowest (Angela Brand 2) vs. Highest (Karen Brand 3)'
    0 1 0 0 0 0 0 0 -1 / elsm;
run;
quit; *ST203s02.sas;
```

Least Squares Means Estimate Coefficients			
Effect	Technician	Brand	Row1
Technician*Brand	Angela	1	
Technician*Brand	Angela	2	1
Technician*Brand	Angela	3	
Technician*Brand	Bob	1	
Technician*Brand	Bob	2	
Technician*Brand	Bob	3	
Technician*Brand	Justin	1	
Technician*Brand	Justin	2	
Technician*Brand	Justin	3	
Technician*Brand	Karen	1	
Technician*Brand	Karen	2	
Technician*Brand	Karen	3	-1

Least Squares Means Estimate						
Effect	Label	Estimate	Standard Error	DF	t Value	Pr >  t
Technician*Brand	Lowest (Angela Brand 2) vs. Highest (Karen Brand 3)	-53.7500	6.5040	84	-8.26	<.0001

The estimate of the difference between *Angela* for **Brand 2** and *Karen* for **Brand 3** is -53.75 minutes. On average, *Angela* repairs **Brand 2** in almost an hour less than *Karen* repairs **Brand 3**.

### 3. Writing CONTRAST and ESTIMATE Statements (Self-Study)

- a. Use a CONTRAST statement to compare the average service time for *Bob* for **Brand 2** with the average service time for *Justin* for **Brand 2**. Presuming that  $\alpha$  represents **Technician** and  $\beta$  represents **Brand**, the equation for this contrast in terms of cell mean is the following:

$$\mu_{22} - \mu_{32} = 0$$

Because cell means are related to the model parameters by the following:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

you can rewrite the hypothesis in terms of model parameters as the following:

$$\alpha_2 - \alpha_3 + (\alpha\beta)_{22} - (\alpha\beta)_{32} = 0$$

or

$$\begin{aligned}
 & 0\alpha_1 + 1\alpha_2 - 1\alpha_3 + 0\alpha_4 \\
 & + 0(\alpha\beta)_{11} + 0(\alpha\beta)_{12} + 0(\alpha\beta)_{13} \\
 & + 0(\alpha\beta)_{21} + 1(\alpha\beta)_{22} + 0(\alpha\beta)_{23} \\
 & + 0(\alpha\beta)_{31} - 1(\alpha\beta)_{32} + 0(\alpha\beta)_{33}
 \end{aligned}$$

Notice that trailing zeros can be omitted in the CONTRAST and ESTIMATE statements.

If you use approach 2, then consider the following two-way table:

<b>Technician</b>	<b>Brand</b>			<b>Sum</b>
	1	2	3	
Angela				
Bob		1		
Justin		-1		
Karen				
<b>Sum</b>				

The resulting table is as follows:

<b>Technician</b>	<b>Brand</b>			<b>Sum</b>
	1	2	3	
Angela	0	0	0	0
Bob	0	1	0	1
Justin	0	-1	0	-1
Karen	0	0	0	0
<b>Sum</b>	0	0	0	0

```

ods html close; ods listing;
proc glm data=STAT2.disks;
  class technician brand;
  model time = technician|brand;
  contrast 'Bob brand 2 vs Justin brand 2'
    technician 0 1 -1 0
    technician*brand 0 0 0 0 1 0 0 -1 0 / E;
run;
quit;
*ST203s03.sas;

```

Notice that trailing zeros were omitted in this CONTRAST statement in the SAS code. Also, in the Coefficients table below, notice that PROC GLM filled in the trailing zeros and assigned the appropriate coefficients to the levels of the **Technician** by **Brand** interaction.

## Partial PROC GLM Output

Coefficients for Contrast Bob brand 2 vs Justin brand 2					
Row 1					
					Intercept 0
Technician	Angela	0			
Technician	Bob	1			
Technician	Justin	-1			
Technician	Karen	0			
Brand	1	0			
Brand	2	0			
Brand	3	0			
Technician*Brand	Angela 1	0			
Technician*Brand	Angela 2	0			
Technician*Brand	Angela 3	0			
Technician*Brand	Bob 1	0			
Technician*Brand	Bob 2	1			
Technician*Brand	Bob 3	0			
Technician*Brand	Justin 1	0			
Technician*Brand	Justin 2	-1			
Technician*Brand	Justin 3	0			
Technician*Brand	Karen 1	0			
Technician*Brand	Karen 2	0			
Technician*Brand	Karen 3	0			
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Bob brand 2 vs Justin brand 2	1	2943.062500	2943.062500	17.39	<.0001

The *p*-value of <0.0001 indicates that there is a significant difference between the average service time for Bob brand 2 and the average service time for Justin brand 2.

- b. Use an ESTIMATE statement to compute the difference between the lowest and highest average service times: *Angela* for **Brand 2** and *Karen* for **Brand 3**. Presuming that  $\alpha$  represents **Technician** and  $\beta$  represents **Brand**, the equation for this contrast in terms of cell mean is  $\mu_{12} - \mu_{43} = 0$ .

If you use approach 2, then consider the following two-way table:

Technician	Brand			Sum
	1	2	3	
Angela		1		
Bob				
Justin				
Karen			-1	
<b>Sum</b>				

The resulting table is as follows:

Technician	Brand			Sum
	1	2	3	
Angela	0	1	0	1
Bob	0	0	0	0
Justin	0	0	0	0
Karen	0	0	-1	-1
<b>Sum</b>	<b>0</b>	<b>1</b>	<b>-1</b>	<b>0</b>

```
proc glm data=STAT2.disks;
  class technician brand;
  model time=technician|brand;
  estimate 'Lowest (Angela Brand 2) vs. Highest (Karen Brand 3)'
    technician 1 0 0 -1
    brand 0 1 -1
    technician*brand 0 1 0   0 0 0   0 0 0   0 0 -1 / E;
run;
quit;
ods listing close; ods html;
*ST203s03.sas;
```

#### Partial PROC GLM Output

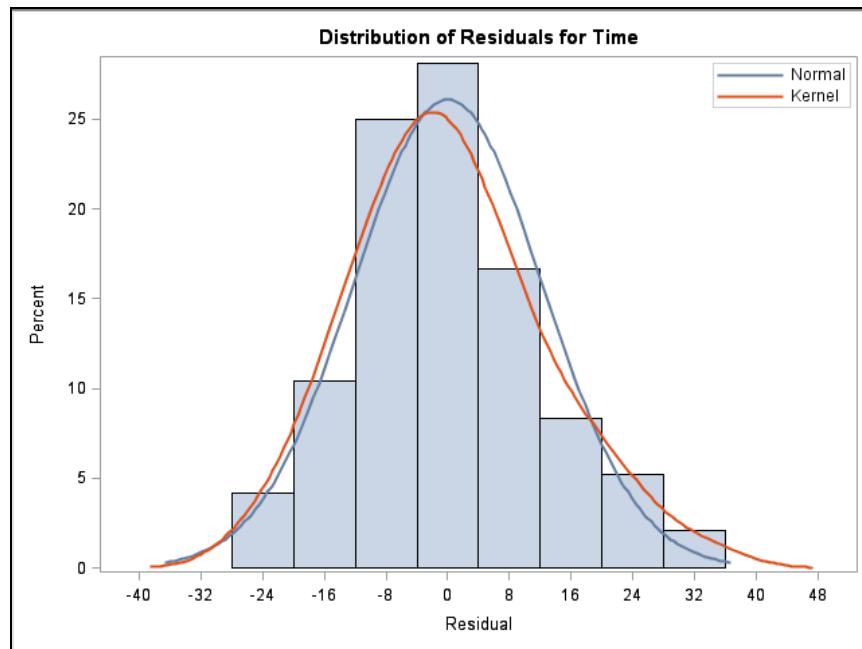
Coefficients for Estimate Lowest (Angela Brand 2) vs. Highest (Karen Brand 3)		
Row 1		
Intercept		0
Technician	Angela	1
Technician	Bob	0
Technician	Justin	0
Technician	Karen	-1
Brand	1	0
Brand	2	1
Brand	3	-1
Technician*Brand	Angela 1	0
Technician*Brand	Angela 2	1
Technician*Brand	Angela 3	0
Technician*Brand	Bob 1	0
Technician*Brand	Bob 2	0
Technician*Brand	Bob 3	0
Technician*Brand	Justin 1	0
Technician*Brand	Justin 2	0
Technician*Brand	Justin 3	0
Technician*Brand	Karen 1	0
Technician*Brand	Karen 2	0
Technician*Brand	Karen 3	-1
Standard		
Parameter	Estimate	Error
Lowest (Angela Brand 2) vs. Highest (Karen Brand 3)	-53.7500000	6.50400518
	t Value	Pr >  t
	-8.26	<.0001

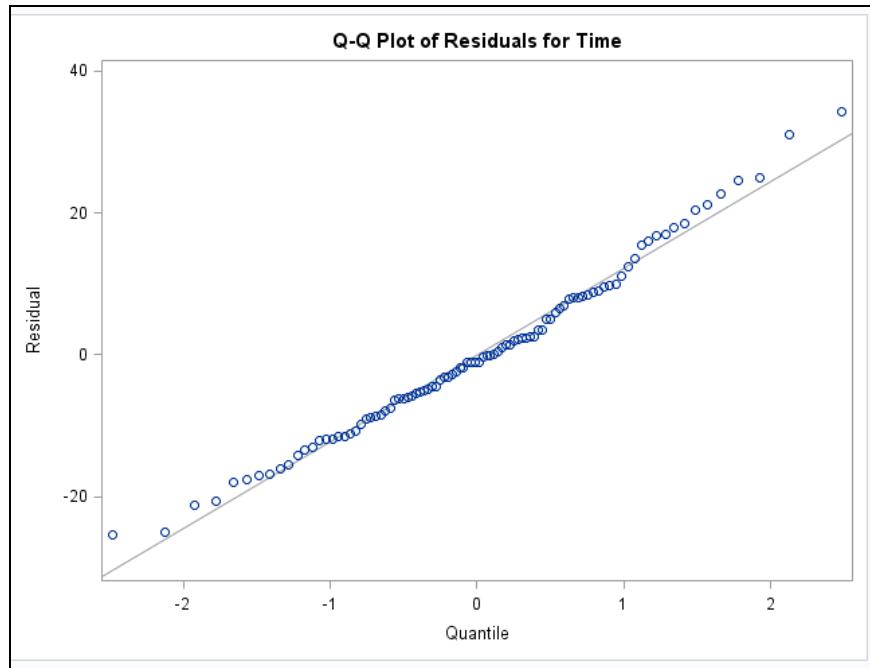
The estimate of the difference between *Angela* for **Brand 2** and *Karen* for **Brand 3** is -53.75 minutes. On average, *Angela* repairs **Brand 2** in almost an hour less than *Karen* repairs **Brand 3**.

#### 4. Evaluating Model Assumptions

- Request the DIAGNOSTICS plots from PROC GLM and examine the histogram and normal probability plot of the residuals. What do you conclude?

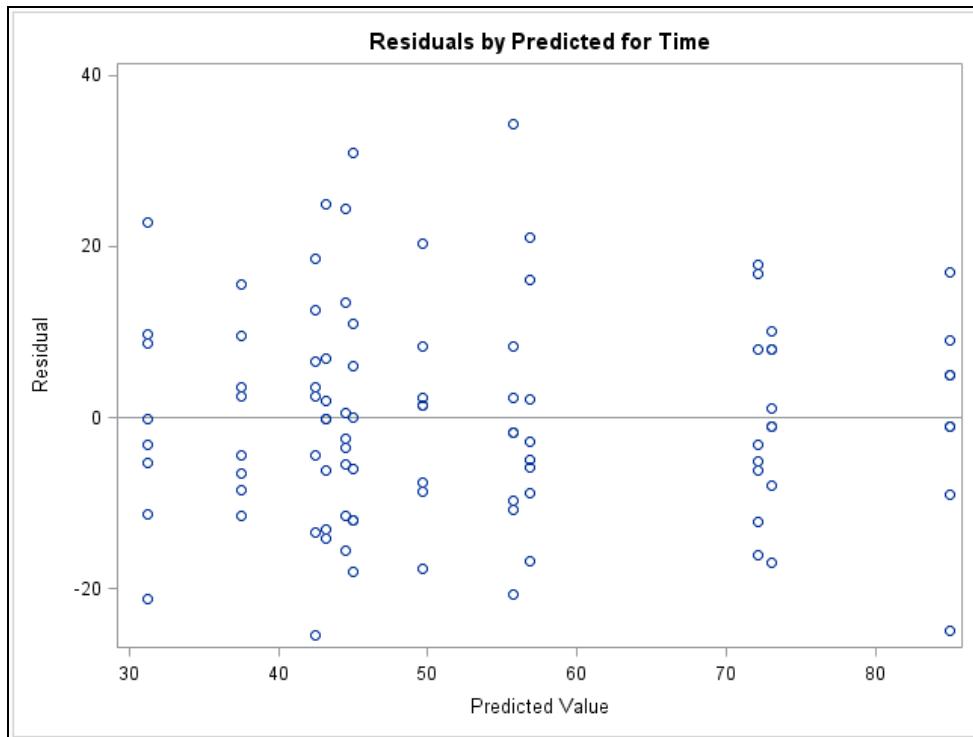
```
proc glm data=STAT2.disks plots(unpack)=diagnostics;
  class technician brand;
  model time=technician|brand;
run;
quit; *ST203s04.sas;
```





Neither the histogram nor the normal probability plot indicates any violation of the normality assumption.

- b.** Check the equal variance assumption by examining the residual plots generated in Exercise 3.a. Use a DATA step to create a single variable, **Group**, which includes both the level of **Technician** and the level of **Brand**. Generate a one-way ANOVA with **Group** as the independent variable, and request a homogeneity of variance test. What do you conclude?



The residual plot indicates that the model fits your data fairly well and the variances appear to be constant across different groups.

```
data disks;
  set STAT2.disks;
  Group=compress(technician||brand);
run;

proc glm data=disks;
  class group;
  model time=group;
  means group/hovtest;
run;
quit; *ST203s04.sas;
```

Partial PROC GLM Output

Levene's Test for Homogeneity of Time Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
group	11	208736	18976.0	0.40	0.9528
Error	84	3994307	47551.3		

The Levene's test for homogeneity of variance has a *p*-value equal to 0.9528. Therefore, you do not reject the null hypothesis of equality of variances.

Level of group	N	Time	
		Mean	Std Dev
Angela1	8	37.5000000	9.3808315
Angela2	8	31.2500000	13.6565003
Angela3	8	49.6250000	11.5750656
Bob1	8	56.8750000	12.7664460
Bob2	8	72.1250000	12.7999721
Bob3	8	43.1250000	12.4147090
Justin1	8	42.5000000	14.2026155
Justin2	8	45.0000000	15.8835045
Justin3	8	44.5000000	13.1148770
Karen1	8	73.0000000	9.1339242
Karen2	8	55.7500000	16.4468842
Karen3	8	85.0000000	12.6942057

The summary statistics produced by the MEANS statement show that the group size is equal (8) and the means and standard deviations for each **Technician/Brand** combination are shown.

## Solutions to Student Activities (Polls/Quizzes)

### 3.01 Poll – Correct Answer

Both ANOVA models and regression models are general linear models that use OLS to obtain parameter estimates and standard errors.

- True  
 False

5

### 3.02 Multiple Choice Poll – Correct Answer

Which of the following are **false** for interpreting a significant interaction between **Gender** and **School**?

- a. The effect of **Gender** depends on the levels of **School**.
- b. The effect of **School** depends on the levels of **Gender**.
- c. The difference between levels of **Gender** is not the same across schools.
- d. The difference between levels of **School** is not the same across genders.
- e. The interaction is not intuitive to interpret, so you can ignore it.

14

### 3.03 Multiple Answer Poll – Correct Answers

The significant interaction between **Brand** and **Technician** indicates which of the following?

- a. The differences in the average repair time between the technicians differ across different brands.
- b. The differences in the average repair time between the brands differ across different technicians.
- c. The significant interaction is not something that you should worry about.

20

### 3.05 Multiple Choice Poll – Correct Answer

Assume that the CLASS statement reads as follows:

CLASS SCHOOL GENDER

Which of the following statements is correct for comparing **Reading3** values for Male Cottonwood and Male Pine?

- a. LSMESTIMATE School\*Gender 1 0 0 -1;
- b. LSMESTIMATE School\*Gender 0 1 0 0 0 0 0 -1;
- c. LSMESTIMATE School 1 0 0 -1 Gender 0 1;

50

### 3.06 Poll – Correct Answer

When you write the CONTRAST or the ESTIMATE statement, the trailing zeros can be omitted, but the leading zeros or intermittent zeros must be specified.

- True  
 False

88

### 3.07 Poll – Correct Answer

You always examine plots, such as a normal probability plot, to evaluate normality because the normality test might be sensitive to sample size.

- True  
 False

101



# Chapter 4     Analysis of Covariance (ANCOVA)

<b>4.1</b>	<b>Introduction to Analysis of Covariance (ANCOVA) .....</b>	<b>4-3</b>
	Demonstration: Conducting an Analysis of Covariance Using PROC GLM .....	4-12
	Exercises .....	4-16
<b>4.2</b>	<b>Least Squares Means for ANCOVA Models.....</b>	<b>4-17</b>
	Demonstration: Least Squares Means and Multiple Comparison Tests .....	4-19
	Exercises .....	4-25
<b>4.3</b>	<b>Diagnostics and Remedial Measures for ANCOVA Models.....</b>	<b>4-26</b>
	Demonstration: Diagnostics and Remedial Measures for ANCOVA Models .....	4-28
	Exercises (Take-home) .....	4-34
<b>4.4</b>	<b>Chapter Summary.....</b>	<b>4-35</b>
<b>4.5</b>	<b>Solutions .....</b>	<b>4-36</b>
	Solutions to Exercises .....	4-36
	Solutions to Student Activities (Polls/Quizzes) .....	4-44



# 4.1 Introduction to Analysis of Covariance (ANCOVA)

## Objectives

- Perform an analysis of covariance using the GLM procedure.
- Interpret the parameter estimates from an analysis of covariance.
- Generate adjusted means for the CLASS variable in an analysis of covariance.

3

## Regression and ANOVA



Both PROC REG and PROC GLM use ordinary least squares to fit general linear models to your data with a continuous response variable.



Statements and options in PROC REG make the procedure easier to use for regression analysis where the independent variables are continuous.



Statements and options in PROC GLM make the procedure more convenient for analysis of variance where the independent variables are discrete.

4

Both PROC REG and PROC GLM are tools to fit general linear models  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , where the dependent variable is an independent continuous variable following a normal distribution with an expected value  $\mathbf{X}\boldsymbol{\beta}$  and a constant variance  $\sigma^2$ , given values of predictor variables. This assumption is often checked using the residuals from the fitted model. Both procedures use the ordinary least squares method to obtain parameter estimates and the standard errors.

PROC REG is more convenient to use for regression analysis when you have continuous predictor variables. The automatic model selection methods, the model diagnostics statistics, and other statements and options make it easier to use for regression analysis.

PROC GLM is more convenient to use for analysis of variance when you have classification variables and you want to evaluate the effects of these variables on the observed response variable. The CLASS statement makes it easier to use for ANOVA, and PROC GLM produces plots via ODS Graphics for ANCOVA models.

## What Happens If You Have Both Continuous and Discrete Independent Variables?

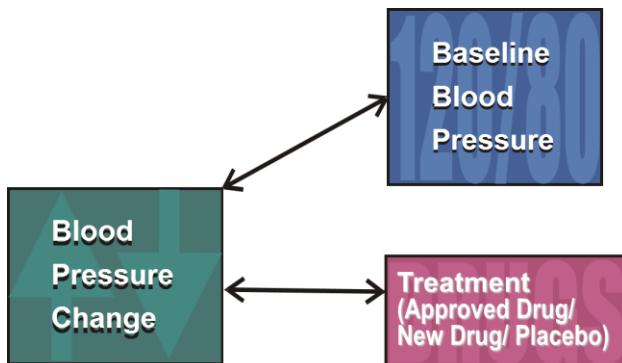
Analyze the data in PROC GLM

- Include the continuous variables as well as the discrete variables using PROC GLM.
- This is called an analysis of covariance (or ANCOVA) model.

5

If you have both continuous and discrete continuous variables, you can include the continuous as well as the discrete variables and use PROC GLM. You can also create indicator (dummy) variables for the discrete independent variables and run a regression analysis using PROC REG.

## Drug Test Example



6

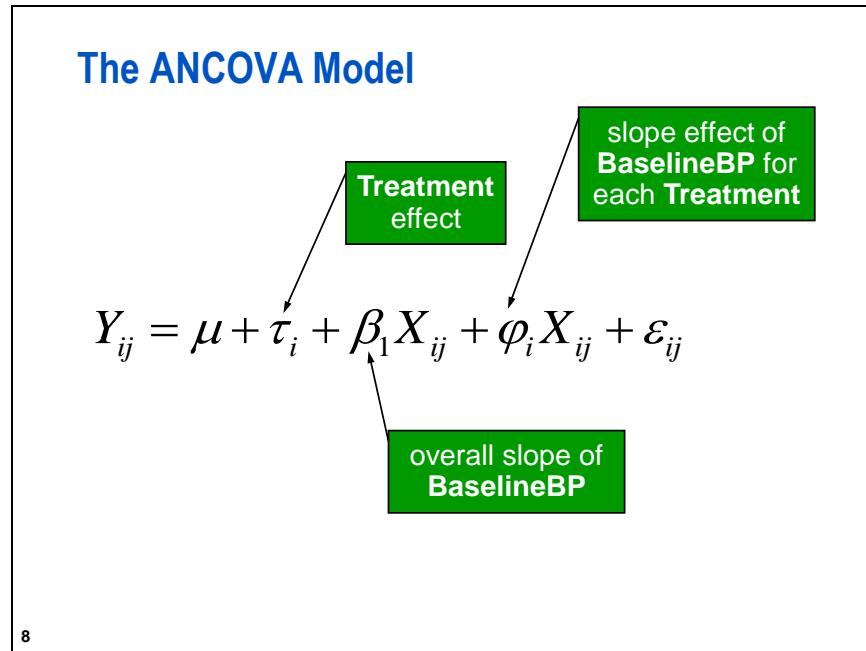
Data from a clinical trials study were collected to compare the effect of baseline blood pressure values on blood pressure changes after a certain treatment. The data are stored in the **STAT2.trials** data set. These are the variables in the data set:

<b>Subject</b>	subject identification
<b>Treatment</b>	drug treatment ( <b>Approved Drug</b> , <b>New Drug</b> , <b>Placebo</b> )
<b>Baselinebp</b>	baseline (starting) level of diastolic blood pressure
<b>BPChange</b>	change in diastolic blood pressure from the beginning of the study.

## The Data

Obs	Subject	Treatment	Baselinebp	BPChange
1	N1	New Drug	100.3	-20.4
2	N2	New Drug	91.6	-5.7
3	N3	New Drug	90.0	-4.3
4	N4	New Drug	93.7	-7.9
5	N5	New Drug	95.4	-10.1
6	N6	New Drug	99.6	-16.4
7	N7	New Drug	97.9	-15.0
8	N8	New Drug	96.0	-11.1
9	N9	New Drug	96.9	-10.8
10	N10	New Drug	95.4	-5.4
11	N11	New Drug	93.9	-8.7
12	N12	New Drug	89.7	-2.8
13	N13	New Drug	94.6	-10.3
14	N14	New Drug	98.0	-13.2
15	N15	New Drug	93.4	-9.9
16	N16	New Drug	93.4	-9.5
17	N17	New Drug	101.7	-18.3
18	N18	New Drug	101.6	-18.8
:				

7



The analysis of covariance (ANCOVA) model can be written as shown above, where

$Y_{ij}$   $j^{\text{th}}$  observed response value of **BPChange** in the  $i^{\text{th}}$  **Treatment** group

$\mu$  overall intercept

$\tau_i$  the effect of the  $i^{\text{th}}$  **Treatment** on the intercept

$\beta$  overall slope

$\phi_i$  the effect of the  $i^{\text{th}}$  **Treatment** on the slope of **BaselineBP**

$X_{ij}$   $j^{\text{th}}$  value of the continuous variable or covariate, **BaselineBP**, in the  $i^{\text{th}}$  **Treatment** group

$\varepsilon_{ij}$  random error.

In an ANCOVA model, the discrete independent variables are usually called *classification, class, or grouping variables*. The continuous independent variables are usually called *covariates*.

## The Model in PROC GLM

$$Y_{ij} = \mu + \tau_i + \beta_1 X_{ij} + \varphi_i X_{ij} + \varepsilon_{ij}$$

```
class Treatment;
model BPChange=Treatment BaselineBP
Treatment*BaselineBP;
```

9

The CLASS statement in PROC GLM creates columns in the design matrix,  $\mathbf{X}$ , for each classification variable. The number of design columns is the same as the number of levels for a given CLASS variable. The value of each design column is either 0 or 1 across all observations. You list interaction terms directly in the MODEL statement in GLM. The design columns are created for interaction terms involving variables listed in the CLASS statement.

## PROC GLM Creates Indicator Variables

For the treatment effect: **Approved Drug**, **New Drug**, or **Placebo**

$$D_{Approved} = \begin{cases} 1 & \text{if Approved Drug} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{New} = \begin{cases} 1 & \text{if New Drug} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{Placebo} = \begin{cases} 1 & \text{if Placebo} \\ 0 & \text{otherwise} \end{cases}$$

10

continued...

PROC GLM creates numeric indicator variables for variables listed in the CLASS statement. For the **Treatment** variable, three indicator variables are created—one for each level of treatment: **Approved Drug**, **New Drug**, and **Placebo**.

## The Model in PROC GLM

```
class Treatment;
model BPChange=Treatment BaselineBP
Treatment*BaselineBP;
```

$$Y_{ij} = \mu + \tau_i + \beta_1 X_{ij} + \varphi_i X_{ij} + \varepsilon_{ij}$$

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} -20.4 \\ -5.7 \\ \vdots \\ 0.2 \\ \vdots \\ -1.8 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{32} \\ \vdots \\ \varepsilon_{93} \end{bmatrix}$$

11

continued...

For this example, the **Y** matrix contains the observed values of blood pressure change for the 93 observations and the **ε** matrix contains the random error terms for the observations.

## The Model in PROC GLM

8 design columns

$$\mathbf{X} = \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 & 100.3 & 100.3 & 0 & 0 \\ 1 & \vdots & \vdots & \vdots & 91.6 & 91.6 & \vdots & \vdots \\ \vdots & & & & \vdots & \vdots & & \vdots \\ 0 & 1 & & & 92.8 & 0 & 92.8 & 0 \\ 1 & \vdots & \vdots & \vdots & & & \vdots & \\ 1 & & & & & & \vdots & \\ 0 & 1 & & & 93.4 & & 0 & 93.4 \\ 1 & \vdots & \vdots & \vdots & \vdots & & \vdots & \\ 1 & 0 & 0 & 1 & 99.5 & 0 & 0 & 99.5 \end{bmatrix}}_{93 \times 8} \quad \mathbf{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \beta \\ \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{bmatrix}_{8 \times 1}$$

12

The **X** matrix consists of eight columns:

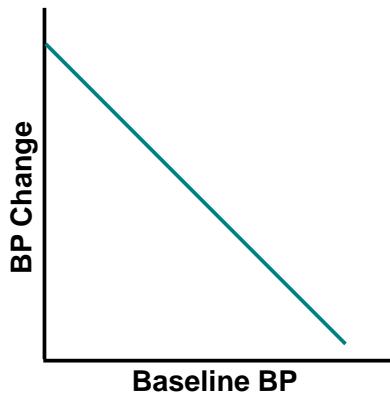
- a column of 1s corresponding to the intercept
- three columns of 0s or 1s corresponding to the three levels of the class variable **Treatment** (**Approved Drug**, **New Drug**, and **Placebo**)
- a column corresponding to the value of **BaselineBP**
- three columns corresponding to the interaction terms between the three levels of the class variable **Treatment** and **BaselineBP**

The eight parameters are intercept, (the effect of) **Treatment 1 (Approved Drug)**, (the effect of) **Treatment 2 (New Drug)**, (the effect of) **Treatment 3 (Placebo)**, (the slope of) **BaselineBP**, the slope effect for **Approved Drug**, the slope effect of **New Drug**, and the slope effect of **Placebo**.

The first four columns in the **X** matrix are linearly dependent. This means that PROC GLM fits an over-parameterized model fit. The last four columns are also linearly dependent.

- ☞ There are an infinite number of ways to code class variables with dummy variables, but only a few are easily interpretable. Many statistics are invariant to different parameterizations of the model containing class variables.
- ☞ You can also create indicator (dummy) variables for the discrete independent variables and run a regression analysis using PROC REG. This is shown in a later section.

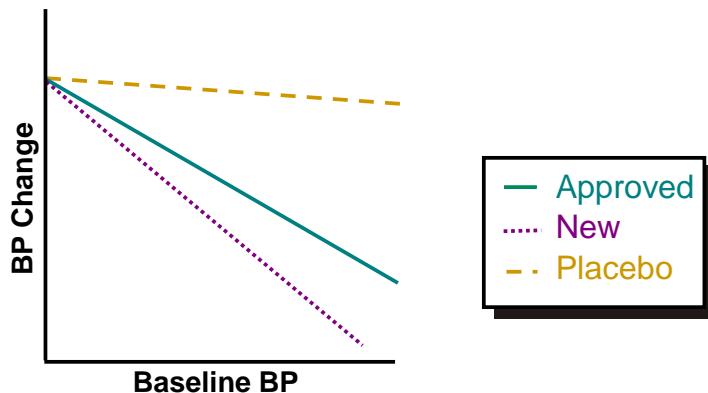
## A Model with Equal Slopes and Intercepts



13

The simplest situation is that both the slopes and intercepts for the three treatments would be the same. If this is the case, the change in blood pressure is the same for the treatments when the baseline blood pressure is zero. The change in blood pressure increases or decreases at the same rate for the treatments at different baseline blood pressure values. This indicates that the type of drug is not significant in determining the change in blood pressure.

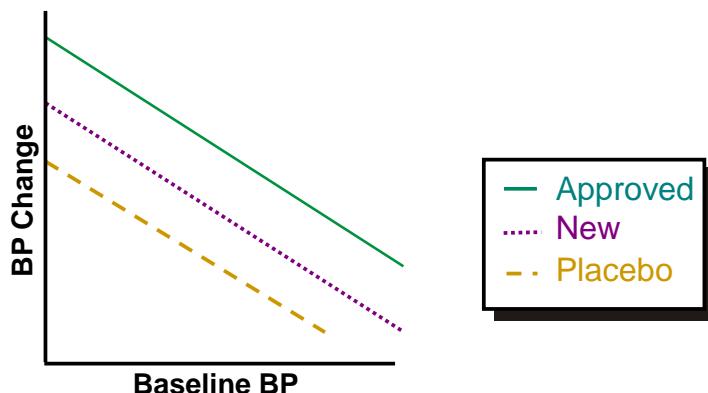
### An Equal-Intercepts but Unequal Slopes Model



14

This graph is a depiction of a situation where the grouping variable (in this example, **Treatment**) has three levels. In this situation, the three levels of the grouping variable could have the same intercept, but different slopes. If this is the case, the change in blood pressure is the same for the treatments when the baseline blood pressure is zero. However, the change in blood pressure increases or decreases at a different rate for at least two of the treatments with a change in baseline blood pressure.

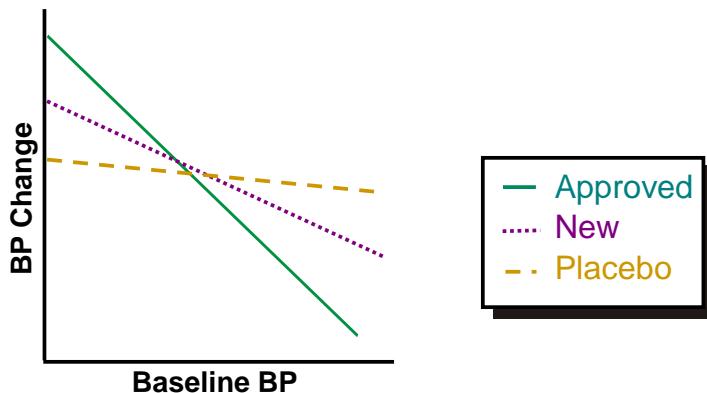
### An Equal-Slopes but Unequal Intercepts Model



15

A third possibility is that the intercepts are different, but the slopes are the same for the treatments. In this example, the change in blood pressure is different for at least two of the treatments when the baseline blood pressure is zero. However, the change in blood pressure increases or decreases at the same rate for the three treatments with a change in baseline blood pressure.

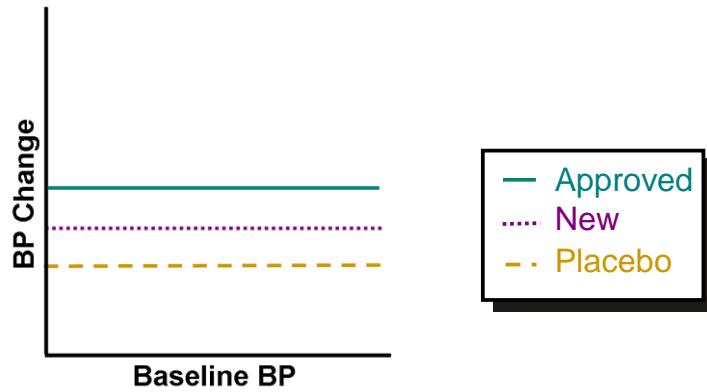
## Unequal Slopes and Intercepts



16

Another possibility is that both the slopes and intercepts are different for at least two of the treatments.

## Equal Slopes ANOVA Model



17

A special case of the Equal Slopes Model is where all of the slopes are zero. In this case, the continuous covariate makes no contribution to the model and an analysis of variance should be used.



## Conducting an Analysis of Covariance Using PROC GLM

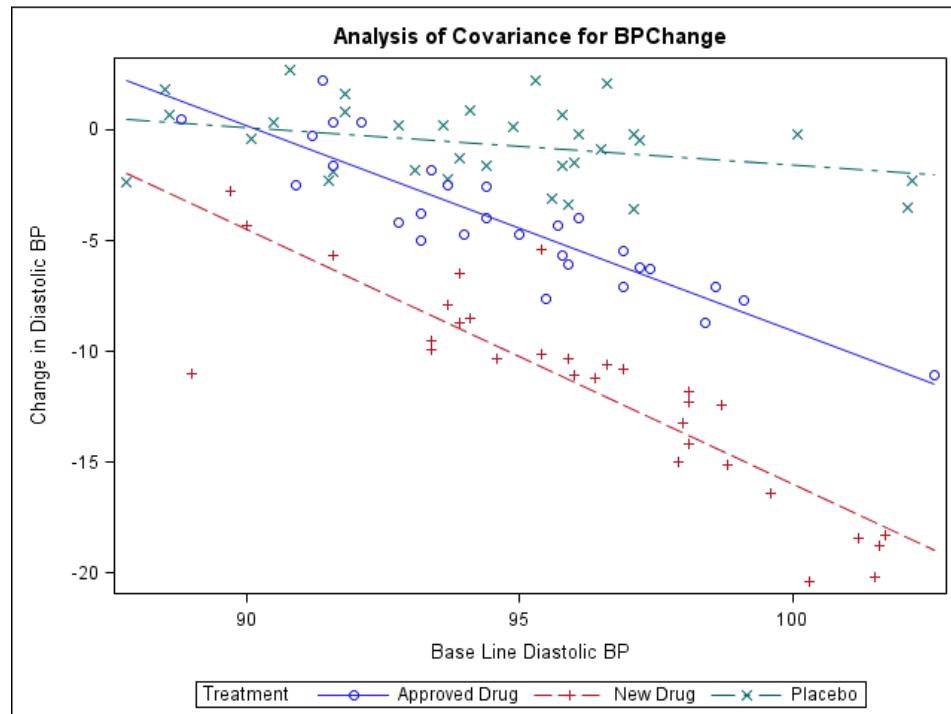
The **STAT2.trials** data set contains the following variables:

<b>Subject</b>	subject identification
<b>Treatment</b>	drug treatment ( <b>New Drug</b> , <b>Approved Drug</b> , or <b>Placebo</b> )
<b>BaselineBP</b>	baseline (starting) level of diastolic blood pressure
<b>BPChange</b>	change in diastolic blood pressure from the beginning of the study.

You are interested in determining the effect of treatment and baseline blood pressure on the change in blood pressure. Begin by examining the data with a scatter plot of **BPChange** versus **BaselineBP**.

If you specify an analysis of covariance model, with one or two CLASS variables and one continuous variable, the GLM procedure produces an analysis of covariance plot of the response values versus the covariate values, with lines representing the fitted relationship within each classification level.

```
ods html style=listing;
proc glm data=STAT2.trials;
  class treatment;
  model bpchange = treatment baselinebp treatment*baselinebp /
    solution;
  title 'Analysis of Covariance';
run;                                *ST204d01.sas;
```



It seems that **Approved Drug** and **New Drug** have the same slope. It appears that **Placebo** has a different slope than the other two treatments. It is difficult to determine whether the intercepts for **Approved Drug** and **New Drug** are different from each other in the graph.

**The GLM Procedure**

Class Level Information		
Class	Levels	Values
Treatment	3	Approved Drug New Drug Placebo

Number of Observations Read	93
Number of Observations Used	93

---

**The GLM Procedure**

**Dependent Variable: BPChange Change in Diastolic BP**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2713.007382	542.601476	166.96	<.0001
Error	87	282.744446	3.249936		
Corrected Total	92	2995.751828			

R-Square	Coeff Var	Root MSE	BPChange Mean
0.905618	-33.29821	1.802758	-5.413978

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Treatment	2	2004.713803	1002.356901	308.42	<.0001
BaselineBP	1	510.649344	510.649344	157.13	<.0001
BaselineBP*Treatment	2	197.644236	98.822118	30.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Treatment	2	163.2624078	81.6312039	25.12	<.0001
BaselineBP	1	541.6044650	541.6044650	166.65	<.0001
BaselineBP*Treatment	2	197.6442355	98.8221178	30.41	<.0001

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	15.16097099	B	8.59030873	1.76 0.0811
Treatment Approved Drug	68.37420774	B	13.80649413	4.95 <.0001
Treatment New Drug	83.73827468	B	12.49532810	6.70 <.0001
Treatment Placebo	0.00000000	B	.	.
BaselineBP	-0.16733979	B	0.09100548	-1.84 0.0694
BaselineBP*Treatment Approved Drug	-0.75860830	B	0.14588398	-5.20 <.0001
BaselineBP*Treatment New Drug	-0.98130328	B	0.13099717	-7.49 <.0001
BaselineBP*Treatment Placebo	0.00000000	B	.	.

**Note:** The  $X'X$  matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The  $F$  statistic for the **BaselineBP\*Treatment** interaction term tests the null hypothesis that the slopes of **BaselineBP** are equal across all treatments. The  $p$ -value is less than 0.0001. Therefore, you reject the null hypothesis and conclude that the slopes for the treatments are not all equal to each other. The unequal slopes model is needed for this data set.

The parameter estimates can be used to write the regression equations for each treatment.

The intercept term estimates the intercept for the last level of the grouping variable, in this case **Placebo**.

- The parameter estimate for the **Treatment Approved Drug** is estimating the difference between the intercept of **Approved Drug** and the intercept of the last group, **Placebo**. Likewise, the parameter estimate for the **Treatment New Drug** is estimating the difference between the intercept of **New Drug** and the intercept of **Placebo**.
- The parameter estimate for the **Treatment Placebo** is zero because it estimates the difference between the intercept of **Placebo** and the intercept of the last group, **Placebo**.
- The parameter estimate for **BaselineBP** is the estimate of the slope for the last level of the grouping variable, **Placebo**.
- The parameter estimate for approved drug **BaselineBP\*Treatment** for **Approved Drug** estimates the difference in the slope of **Approved Drug** and the slope of the last group, **Placebo**.
- The parameter estimate for **BaselineBP\*Treatment** for **New Drug** estimates the difference in the slope of **New Drug** and the slope of the last group, **Placebo**.
- The parameter estimate for **BaselineBP\*Treatment** for **Placebo** is zero because it estimates the difference in the slope of **Placebo** and itself.

The three regression models are as follows:

- for the approved drug:

$$\begin{aligned} \text{BPChange} &= (15.1610 + 68.3742) + (-0.1673 - 0.7586) * \text{BaselineBP} \\ &= 83.5352 - 0.9259 * \text{BaselineBP} \end{aligned}$$

- for the new drug:

$$\begin{aligned} \text{BPChange} &= (15.1610 + 83.7383) + (-0.1673 - 0.9813) * \text{BaselineBP} \\ &= 98.8993 - 1.1486 * \text{BaselineBP} \end{aligned}$$

- for the placebo:

$$\text{BPChange} = 15.1610 - 0.1673 * \text{BaselineBP}$$

## 4.01 Multiple Choice Poll

In the previous demonstration, the parameter estimate for **BaselineBP** is -0.1673. This is the slope that corresponds to

- a. Approved Drug
- b. New Drug
- c. Placebo



## Exercises

---

### 1. Generating an ANCOVA Model

- a. You are interested in determining the effect of **Words1** and **Gender** on **Reading3** scores. Generate an ANCOVA plot from PROC GLM and perform an analysis of covariance. Use the **STAT2.school** data set. Use the test for the interaction term to determine whether both slopes are equal.
- b. Generate the most appropriate model and use the SOLUTION option. Use the parameter estimates to write the regression equation for each gender.

## 4.02 Quiz

How do you interpret the significant term  
**Words1ByGenderF** for this model?

23

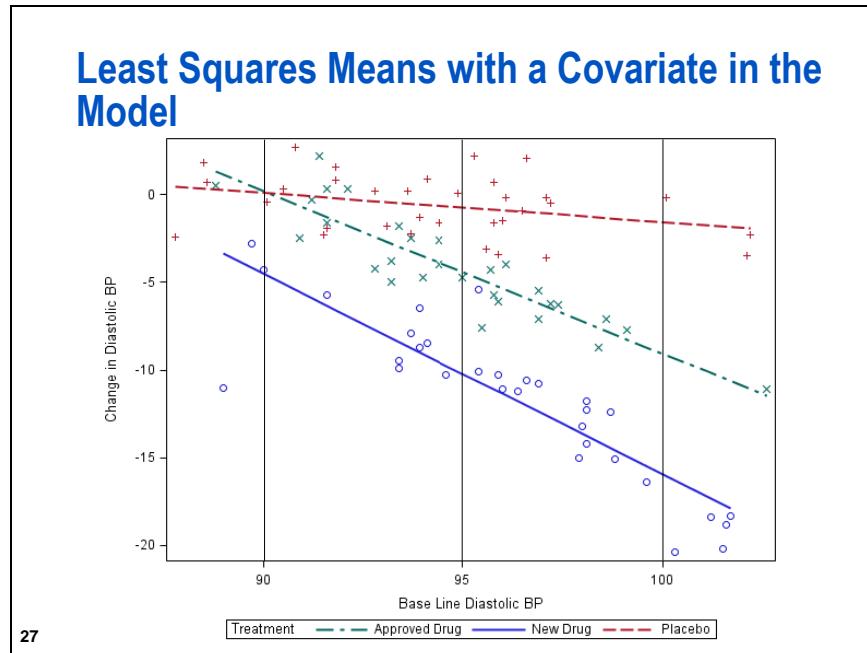
# 4.2 Least Squares Means for ANCOVA Models

---

## Objectives

- Estimate least squares means for an ANCOVA model.
- Perform multiple comparison tests using the LSMEANS statement in PROC GLM.

26



In analysis of covariance, your main interest is generally to compare group means. As discussed in a previous chapter, you can use the LSMEANS statement in PROC GLM to generate least squares means or adjusted means. Looking at the graph above, the question arises as to where along the lines, you should compute and compare these adjusted means.

If you compare the group means of **Placebo** versus **Approved Drug** at a **BaselineBP** measure of 90, you get different results than if you make the comparisons at a **BaselineBP** measure of 100. By default, the LSMEANS statement gives the estimate of the average dependent variable for each group at the mean value of the covariate. You can also specify other values of the covariate in the program. (Making comparisons at the mean value of the covariate eliminates inconsistent results. This action for ANCOVA models is equivalent to centering the continuous covariates for regression models with indicator variables.) In addition to obtaining the least squares means, you can also use multiple comparison tests to determine which group means are different, just as you can in an ANOVA.



## Least Squares Means and Multiple Comparison Tests

The mean of **BaselineBP** is approximately 95. Therefore, when least squares means are requested for **Treatment**, they are calculated at this mean value for the covariate **BaselineBP**. Because the slopes for the three treatments are not equal, it might be important to calculate least squares means at other values of **BaselineBP**. Suppose you are also interested in comparing the three treatments when **BaselineBP** values are 90 and 100. You can use the AT option in the LSMEANS statement. ODS Graphics can produce mean plots and diffograms for the LSMEANS statements.

```
proc glm data=STAT2.trials;
  class treatment;
  model bpchange = treatment|baselinebp;
  lsmeans treatment / pdiff adjust=tukey;
  lsmeans treatment / at baselinebp=90 pdiff adjust=tukey;
  lsmeans treatment / at baselinebp=100 pdiff adjust=tukey;
  title 'Least Squares Means for ANCOVA Model';
run;
quit; *ST204d02.sas;
```

Selected LSMEANS statement option:

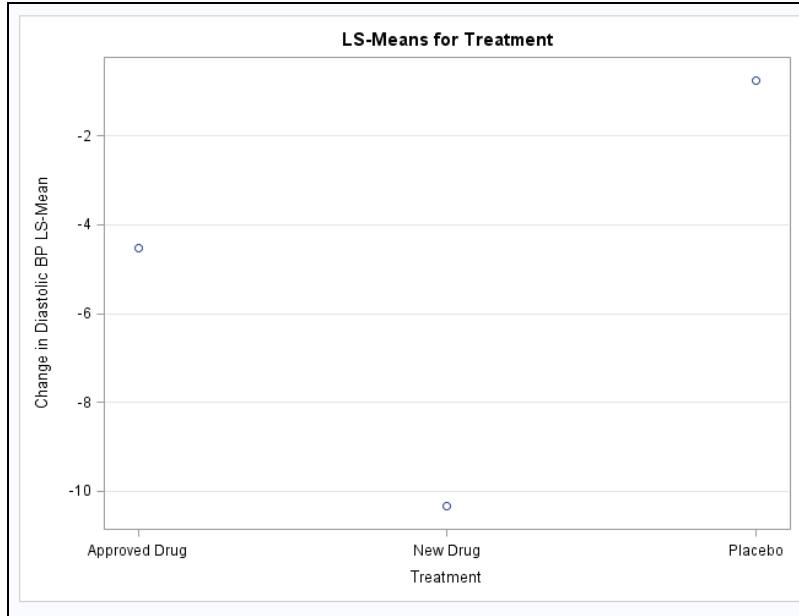
AT *variable*=*value* enables you to modify the values of the covariates used in computing LS-means. By default, all covariate effects are set equal to their mean values for computation of standard LS-means. The AT option enables you to set the covariates to any values that you consider interesting.

Partial PROC GLM Output

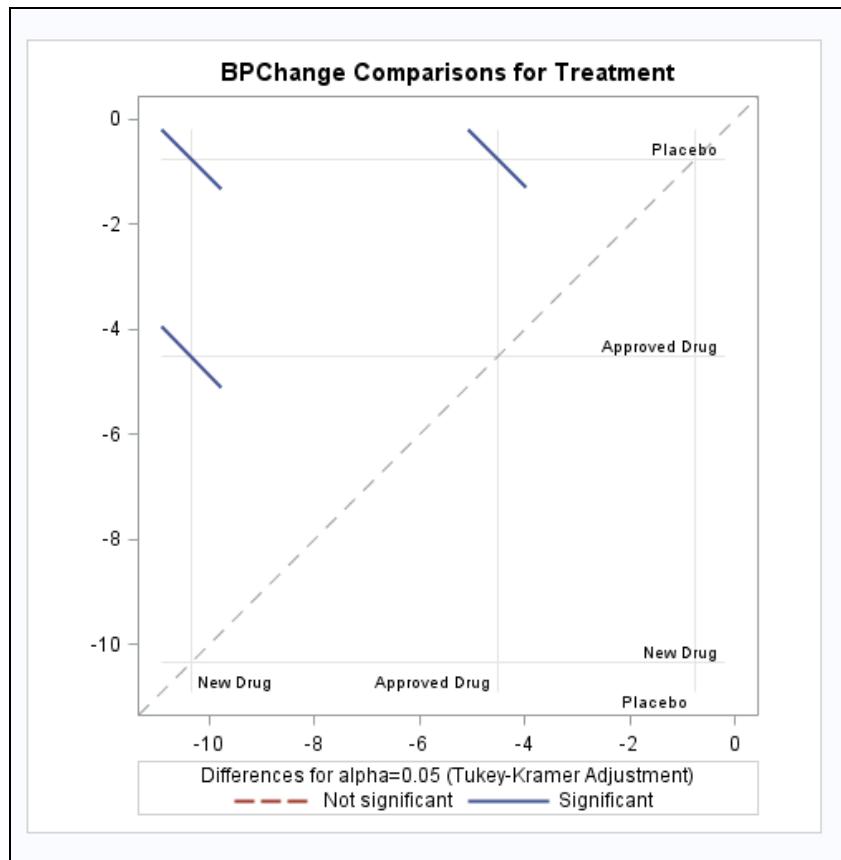
The GLM Procedure Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer		
Treatment	BPChange LSMEAN	LSMEAN Number
Approved Drug	-4.5214897	1
New Drug	-10.3354757	2
Placebo	-0.7528634	3

Least Squares Means for effect Treatment Pr >  t  for H0: LSMean(i)=LSMean(j) Dependent Variable: BPChange				
i/j	1	2	3	
1		<.0001	<.0001	
2	<.0001		<.0001	
3	<.0001	<.0001		

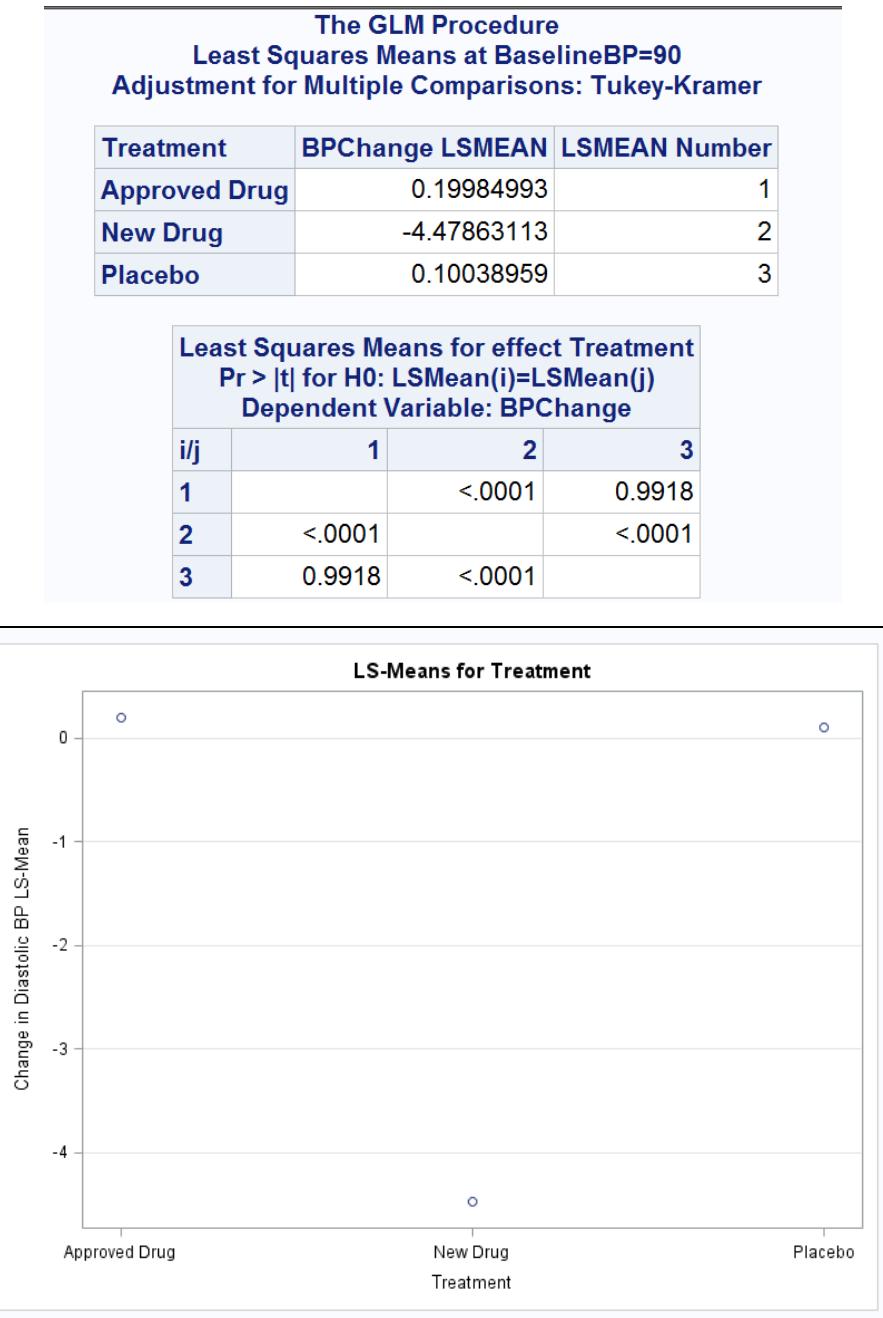


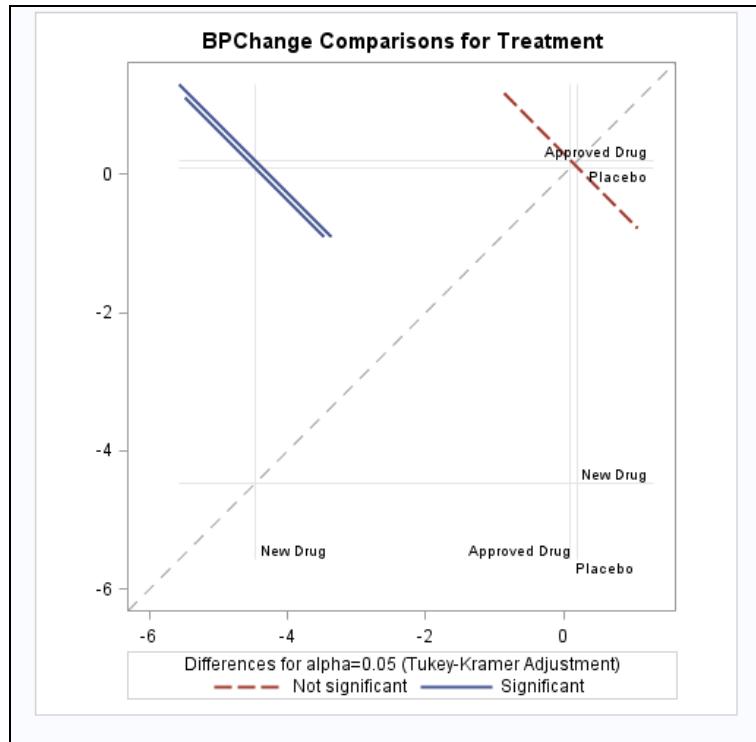
The means for the groups at the mean **BaselineBP** are shown above. By default, these are calculated at the mean values of the covariate **BaselineBP**, which is approximately 95.



The diffogram provides a graphical test of significance for mean comparisons. Any line segments that cross the diagonal reference line are displayed as red dashed lines. They indicate that the least squares means for these groups are not significantly different from each other at alpha = 0.05. Lines that do not cross the reference line represent groups that have least squares means that are significantly different from one another.

At the mean value of **BaselineBP** (approximately 95), the least squares means for all treatment groups are significantly different from each other.





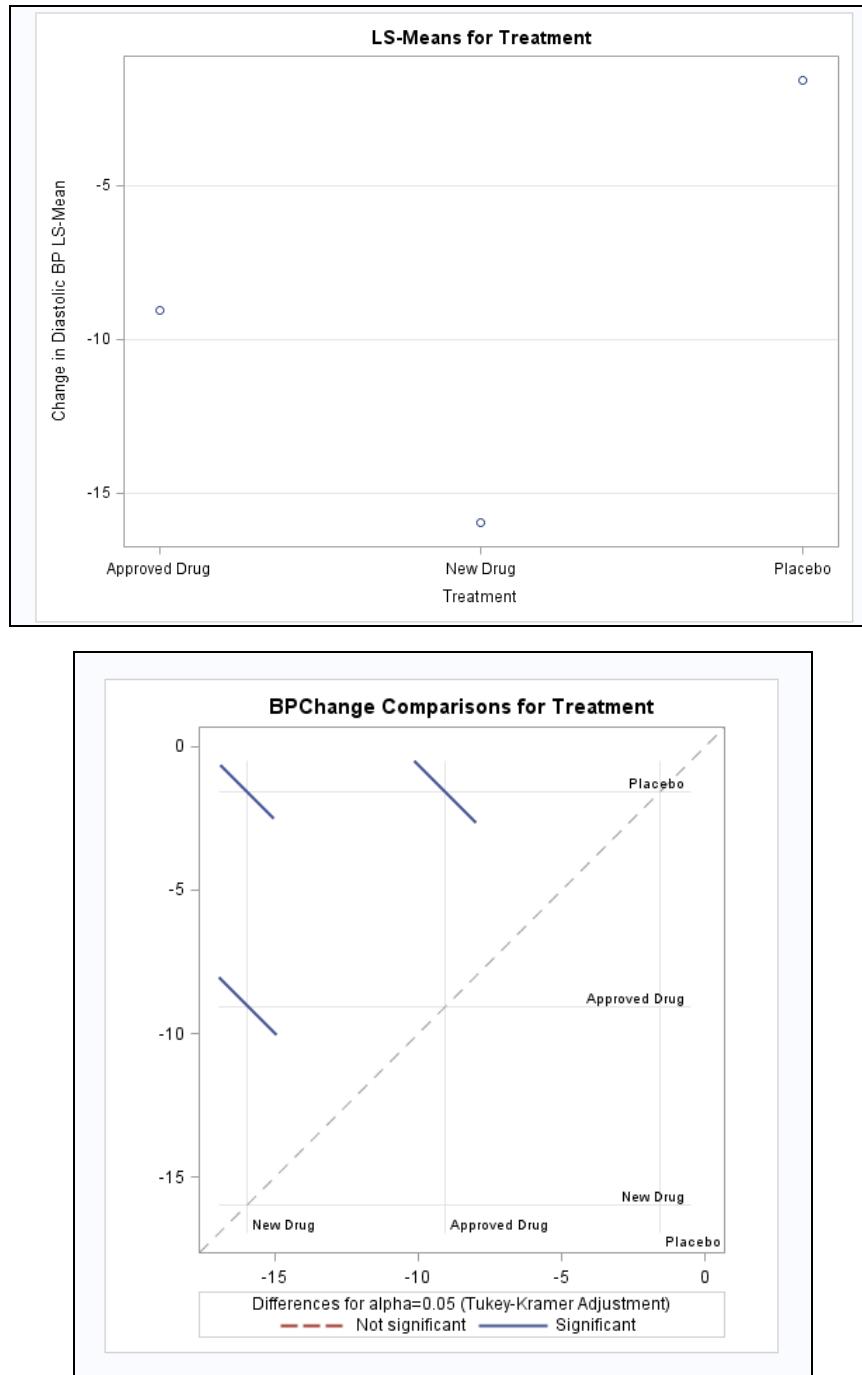
When **BaselineBP** is equal to 90, **Approved Drug** is not significantly different from **Placebo**, as indicated by the *p*-value of 0.9918 and the red dashed line on the diffogram. **Approved Drug** and **New Drug** are found to be statistically different from one another.

**The GLM Procedure**  
**Least Squares Means at BaselineBP=100**  
**Adjustment for Multiple Comparisons: Tukey-Kramer**

Treatment	BPChange LSMEAN	LSMEAN Number
Approved Drug	-9.0596311	1
New Drug	-15.9650619	2
Placebo	-1.5730083	3

**Least Squares Means for effect Treatment**  
**Pr > |t| for H0: LSMean(i)=LSMean(j)**  
**Dependent Variable: BPChange**

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



When **BaselineBP** is equal to 100, all treatments are found to be statistically different from each other.

You should verify the assumptions of any model.

### 4.03 Poll

In ANCOVA models, the least squares means for a CLASS variable is adjusted for the covariate.

- True
- False



## Exercises

---

### 2. Least Squares Means for an ANCOVA Model

Use the LSMEANS statement to determine whether there is a difference between the genders when **Words1** equals the average value, 40, or 60. Use the ADJUST=TUKEY option. Examine the means plots and the diffograms.

## 4.3 Diagnostics and Remedial Measures for ANCOVA Models

### Objectives

- Describe and demonstrate diagnostic tools for ANCOVA models.
- Define remedial measures for ANCOVA models.

33

### Diagnostic Tools for ANCOVA Models

#### PROC GLM

- ◆ Diagnostic plots
- ◆ Residual plots

#### PROC REG

- ◆ Multicollinearity diagnostics
- ◆ All plots are available in GLM plus plots for DFBETAS and DFFITS.
- ◆ Indicator variables must be created by PROC GLMSELECT.

34

PROC GLM has options that request a panel of summary diagnostic plots. This includes scatter plots of residuals, absolute residuals, studentized residuals, and observed responses by predicted values; studentized residuals by leverage; Cook's D by observation; a Q-Q plot of residuals; a residual histogram; and a residual-fit spread plot.

PROC REG has all the diagnostics available in PROC GLM, plus additional ones. These include multicollinearity diagnostics and plots of the DFBETA and DFFITS statistics. PROC REG has no class statement to create the indicator variable needed for the discrete variables in an ANCOVA model. However, PROC GLMSELECT has an option to create and output the design variables, including indicator variables for discrete variables. You create these variables in PROC GLMSELECT and then use them in PROC REG to take advantage of the additional diagnostics.

-  Because all the diagnostics available in the GLM procedure are also available in the REG procedure, you can complete all of the diagnostics in PROC REG.

## The GLMSELECT Procedure

General form of the GLMSELECT procedure:

```
PROC GLMSELECT outdesign=data-set-name <options>;  
  CLASS variables;  
  MODEL dependents=independents / options;  
RUN;
```



## Diagnostics and Remedial Measures for ANCOVA Models

Verify the model assumptions and check for multicollinearity and influential observations using PROC GLMSELECT and PROC REG.

```
ods select none;
proc glmselect data=STAT2.trials outdesign=design;
  class treatment;
  model bpchange = treatment|baselinebp / selection=none;
run;
%put macro variable _glsmod=&_glsmod; *ST204d03.sas;
```

Because the MODEL statement is exactly the same as the one used in PROC GLM, the OUTDESIGN statement creates a data set called **work.design**. The data set contains the variables that you need for PROC REG. In addition, a macro variable (**\_glsmod**) that contains the independent variables is created. Because you are interested only in creating the output data set, the ODS SELECT NONE statement turns off all printed output. The SELECTION=NONE option in the MODEL statement turns off the default setting of automatic model selection.

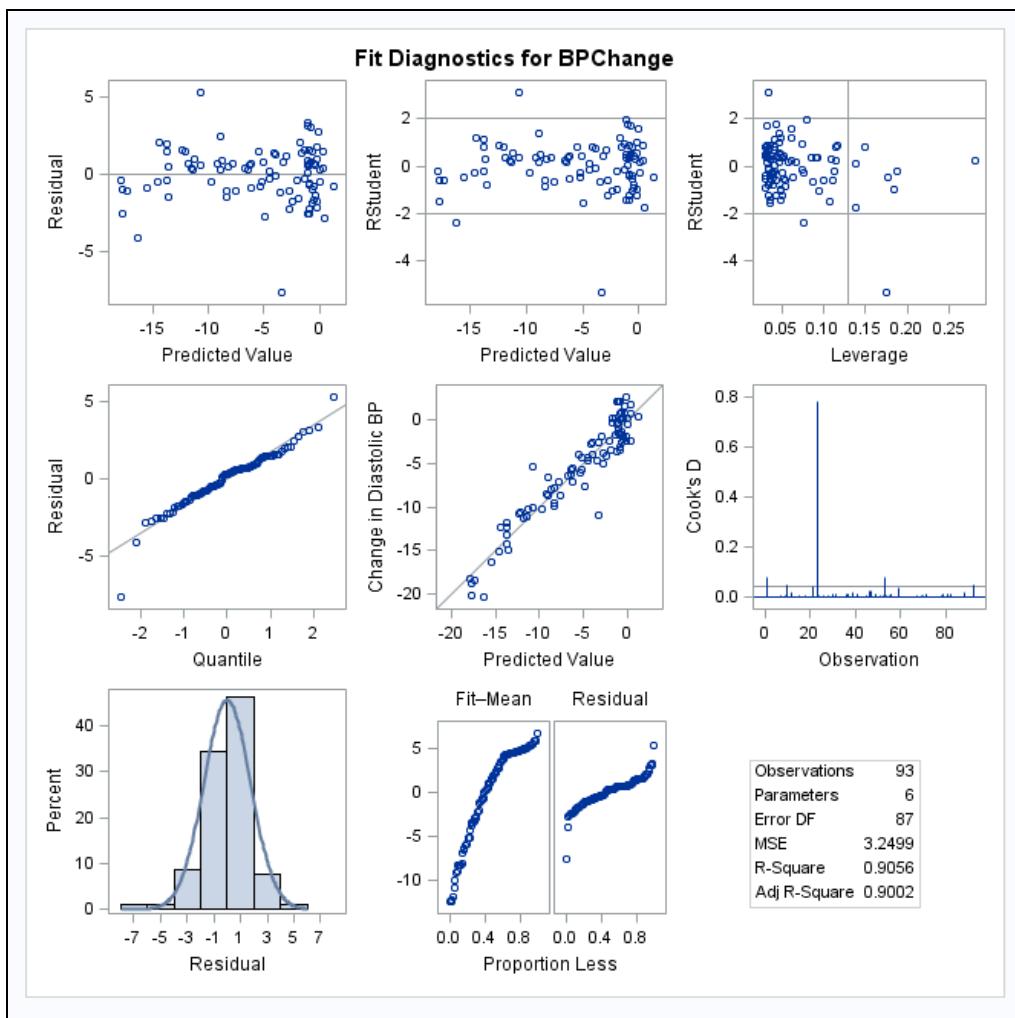
### Partial Log

```
macro variable _glsmod=Treatment_Approved_Drug Treatment_New_Drug Treatment_Placebo BaselineBP
BaselineBP_Treatment_Approved_Dr BaselineBP_Treatment_New_Drug
BaselineBP_Treatment_Placebo
```

Using a %PUT statement, you can see that the variables contained in the macro variable **\_glsmod** are the independent variables listed in the MODEL statement.

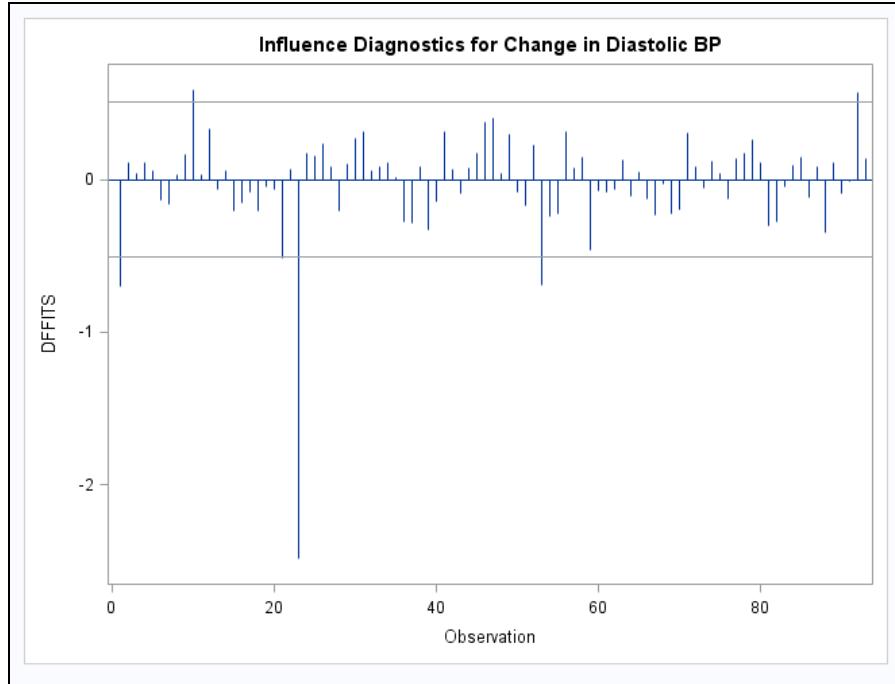
```
ods select ParameterEstimates DiagnosticsPanel DFFITSPlot
      DFBETASPanel;
proc reg data=design plots=(dfbetas dffits);
  model bpchange=&_glsmod / vif influence;
  title 'Check Collinearity on ANCOVA Model';
run;
quit; *ST204d03.sas;
```

By using the **\_GLSMOD** macro variable, you can avoid entering all the variable names in the MODEL statement. Alternatively, you can copy the results from the %PUT statement found in the SAS log in to the MODEL statement.

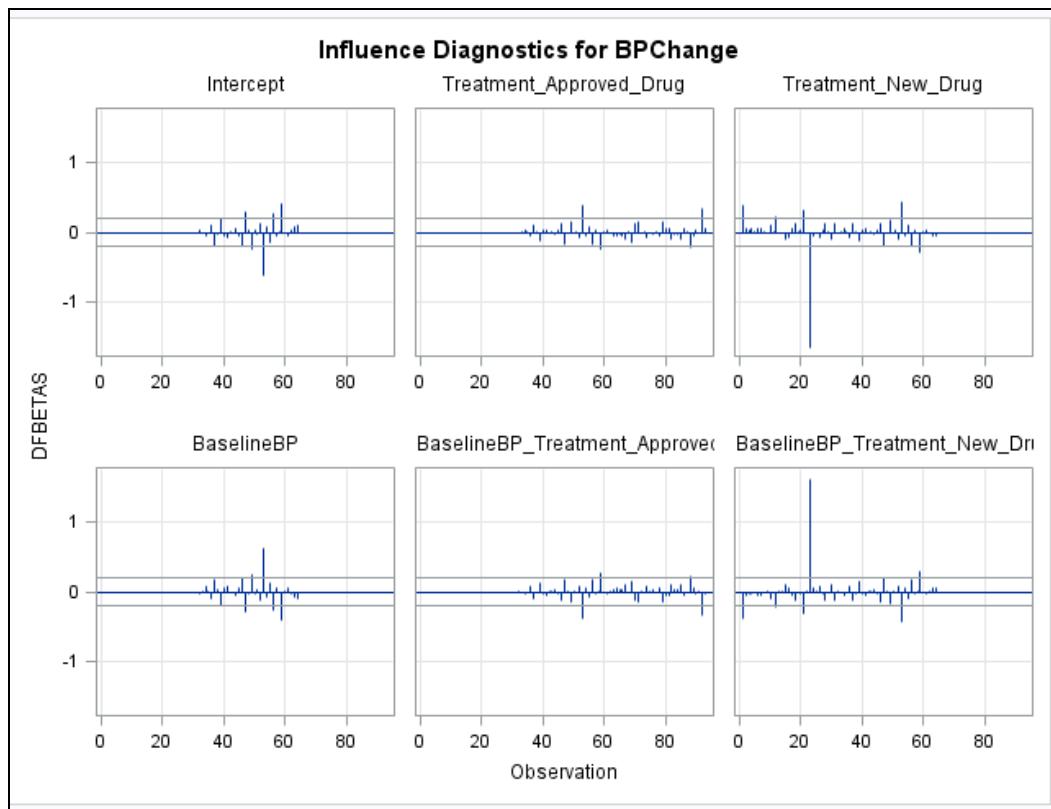


Examine the residual plots to check the assumptions of constant variance and normality. In the Residual and RStudent plots, the residuals appear to be a random scatter around a zero reference line and display no heteroscedasticity. The quantile plot and histogram of the residuals indicate no problems with the normality assumption.

The plot of the observed values versus the predicted values indicate a good fit for the model. The Fit-Mean residual plot indicates the model accounts for a good deal of the variability in the change in diastolic blood pressure measurements, as does the adjusted R square of 0.9002.



The plots produced by the INFLUENCE option indicate that several observations might be exerting influence on the model and should be investigated.



The large variance inflation factors indicate that multicollinearity is a problem for this model.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	Intercept	B	15.16097	8.59031	1.76	0.0811	0
Treatment_Approved_Drug	Treatment Approved Drug	B	68.37421	13.80649	4.95	<.0001	1170.54204
Treatment_New_Drug	Treatment New Drug	B	83.73827	12.49533	6.70	<.0001	992.86667
Treatment_Placebo	Treatment Placebo	0	0	.	.	.	.
BaselineBP	BaselineBP	B	-0.16734	0.09101	-1.84	0.0694	2.73171
BaselineBP_Treatment_Approved_Dr	BaselineBP*Treatment Approved Drug	B	-0.75861	0.14588	-5.20	<.0001	1174.93867
BaselineBP_Treatment_New_Drug	BaselineBP*Treatment New Drug	B	-0.98130	0.13100	-7.49	<.0001	1012.69304
BaselineBP_Treatment_Placebo	BaselineBP*Treatment Placebo	0	0	.	.	.	.

Multicollinearity is frequently present for ANCOVA. If it occurs, centering the continuous covariate is one remedial measure that you can take. Use the STDIZE procedure to center the continuous covariate, **BaselineBP**. Then repeat the diagnostic process for the centered variable.

 The estimates for this model match those obtained from PROC GLM. This is because the two models were parameterized the same way. Recall that the design variables used in PROC REG were created in PROC GLMSELECT. The GLMSELECT procedure created exactly the same design columns that PROC GLM did for the categorical variables, resulting in the same model.

```
proc stdize data=STAT2.trials method=mean
            out=trials2c (rename=(baselinebp=baselinebpc));
  var baselinebp;
run;
ods select none;
proc glmselect data=trials2c outdesign=design2c;
  class treatment;
  model bpchange = treatment|baselinebpc / selection=none;
title 'Check Collinearity on Centered ANCOVA Model';
run;

ods select ParameterEstimates DiagnosticsPanel DFFITSPlot
      DFBETASPanel;
proc reg data=design2c;
  model bpchange=&_glsmod / vif influence;
run;
quit; *ST204d03.sas;
```

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	Intercept	B	-0.75286	0.32152	-2.34	0.0215	0
Treatment_Approved_Drug	Treatment Approved Drug	B	-3.76863	0.46584	-8.09	<.0001	1.33260
Treatment_New_Drug	Treatment New Drug	B	-9.58261	0.46884	-20.44	<.0001	1.39780
Treatment_Placebo	Treatment Placebo	0	0	.	.	.	.
baselinebpc	baselinebpc	B	-0.16734	0.09101	-1.84	0.0694	2.73171
baselinebp_Treatment_Approved_Dr	baselinebp*Treatment Approved Drug	B	-0.75861	0.14588	-5.20	<.0001	1.65282
baselinebp_Treatment_New_Drug	baselinebp*Treatment New Drug	B	-0.98130	0.13100	-7.49	<.0001	2.07541
baselinebp_Treatment_Placebo	baselinebp*Treatment Placebo	0	0	.	.	.	.

The parameter estimates can be used to write the regression equations for each treatment.

The three regression models are as follows:

- for the approved drug:  

$$\text{BPChange} = (-0.7529 - 3.7663) + (-0.1673 - 0.7586) * \text{BaselineBP}$$

$$= -4.5192 - 0.9259 * \text{BaselineBP}$$
- for the new drug:  

$$\text{BPChange} = (-0.7529 - 9.5826) + (-0.1673 - 0.9813) * \text{BaselineBP}$$

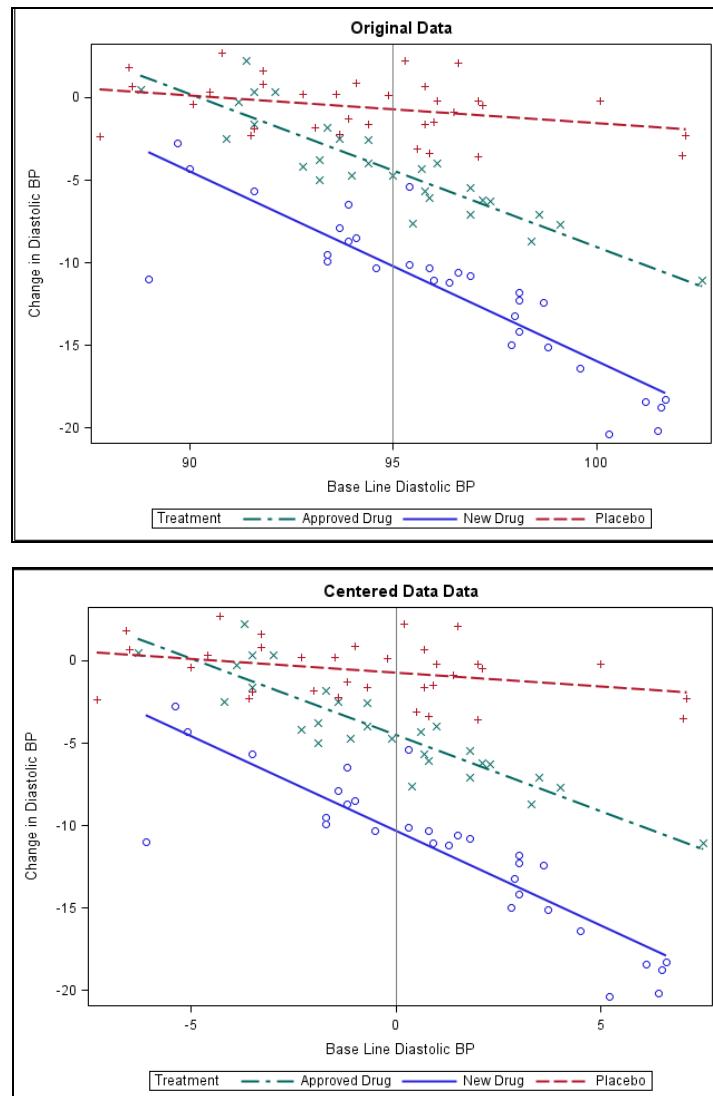
$$= 10.3355 - 1.1486 * \text{BaselineBP}$$
- for the placebo:  

$$\text{BPChange} = -0.7529 - 0.1673 * \text{BaselineBP}.$$

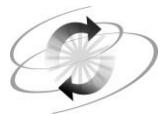
Only the estimated intercepts changed. The slope parameters are unchanged. To see why this is so, look at plots of the original and centered data with the regression lines overlaid.

```
ods html style=listing;
proc sgplot data=STAT2.trials;
  reg y=bpchange x=baselinebp / group=treatment;
  *xaxis values=(0 to 125 by 25);
  refline 95 / axis=x;
title 'Original Data';
run;

proc sgplot data=trials2c;
  reg y=bpchange x=baselinebpc / group=treatment;
  refline 0 / axis=x;
title 'Centered Data';
run;                                         *ST204d03.sas;
```



You can see that the only effect of centering **BaselineBP** is to shift the cloud of data points so that it is centered around 0 instead of (approximately) 95. The intercepts for the three equations change, but the slopes do not.



## **Exercises (Take-home)**

---

### **3. Model Diagnostics**

Check the model for multicollinearity. If multicollinearity seems to be a problem, center **Words1** and refit the model with the centered variable. Compare the results to the previous model.

## 4.4 Chapter Summary

---

Analysis of covariance is used when the response variable is continuous and some of the predictor variables are continuous while other predictor variables are discrete. In the simplest of these situations there is one continuous and one discrete predictor variable. In that case, the possible relationships between the variables are the following:

- Both the slopes and intercepts for the two groups are the same. In this case, a linear regression model is appropriate, because the discrete predictor variable is not significant.
- The intercepts are the same, but the slopes are different.
- The intercepts are different, but the slopes are the same.
- Both the slopes and the intercepts are different.
- The intercepts are different, and the slopes are zero. In this case, an ANOVA model is appropriate because the continuous predictor variable is not significant.

You use the GLM procedure to conduct an analysis of covariance. Via ODS Graphics, an ANCOVA plot is created for a model with a continuous covariate and one or two classification variables. The CLASS statement in PROC GLM creates indicator variables for the discrete variables listed in the statement. You can also specify interaction terms directly in the MODEL statement. The test for the interaction term in an ANCOVA model tests whether the slopes are equal for all groups. If the slopes are not different, remove the interaction term and fit a common slope model for all groups. If the slopes are different, use the unequal slopes model. Use the SOLUTION option in the MODEL statement to obtain parameter estimates when there is a CLASS variable specified in the model. You can use the LSMEANS statement to compute and compare the least squares means for each group at a specified value of the covariate. Via ODS Graphics, you get a mean plot and a diffogram for the comparisons of the least square means.

You can do model diagnostics using the diagnostic plots in PROC GLM. However, some diagnostics that are not available in PROC GLM are available in the REG procedure. Both of these procedures are used for fitting general linear models using the same approach,, that is, ordinary least squares. For this reason, you can use PROC REG to perform diagnostics on your ANCOVA model. However, the REG procedure requires numeric variables only. You can use PROC GLMSELECT to create design variables for all of your independent variables, including indicator variables for discrete variables. The procedure creates a macro variable `_glsmod` that you can use in PROC REG.

# 4.5 Solutions

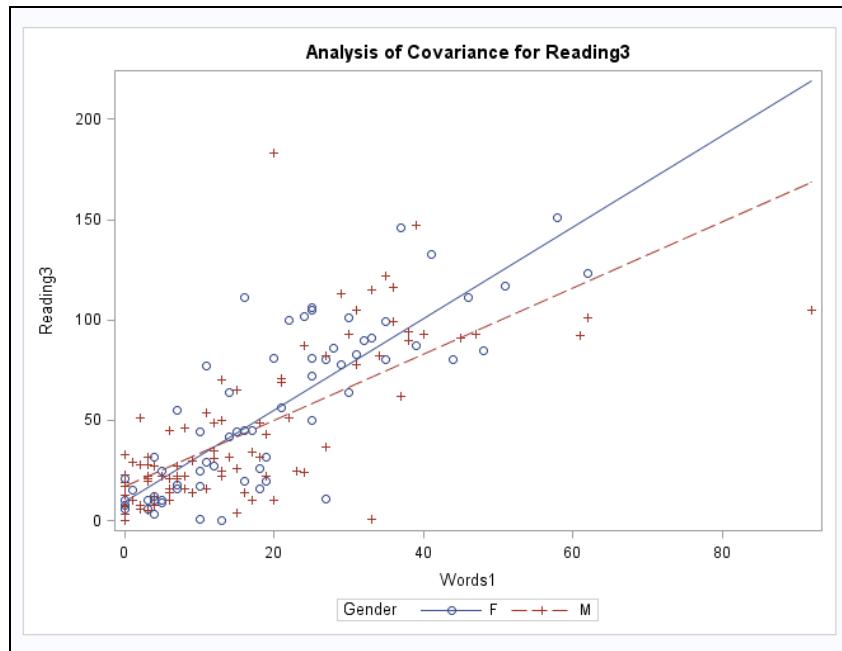
---

## Solutions to Exercises

### 1. Generating an Analysis of Covariance

- a. You are interested in determining the effect of **Words1** and **Gender** on **Reading3** scores. Generate an ANCOVA plot from PROC GLM and perform an analysis of covariance. Use the **STAT2.school** data set. Use the test for the interaction term to determine whether both slopes are equal.

```
proc glm data=STAT2.school;
  class gender;
  model reading3=gender words1 gender*words1;
run;                      *ST204s01.sas;
quit;
```



The ANCOVA plot suggests that the slopes might be unequal, but the intercepts do not seem to be significantly different.

The GLM Procedure						
Dependent Variable: Reading3						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	3	144567.4193	48189.1398	74.19	<.0001	
Error	150	97428.6911	649.5246			
Corrected Total	153	241996.1104				
R-Square	Coeff Var	Root MSE	Reading3 Mean			
0.597396	51.75800	25.48577	49.24026			
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
Gender	1	4246.1640	4246.1640	6.54	0.0116	
Words1	1	136779.4255	136779.4255	210.58	<.0001	
Words1*Gender	1	3541.8299	3541.8299	5.45	0.0209	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
Gender	1	838.9809	838.9809	1.29	0.2576	
Words1	1	138976.2850	138976.2850	213.97	<.0001	
Words1*Gender	1	3541.8299	3541.8299	5.45	0.0209	

The interaction term tests whether the slope parameters are equal. Presuming an alpha equal to 0.05, you reject the null hypothesis. There is sufficient evidence to conclude that the slope parameters for the two groups are **not** equal.

- b. Generate the most appropriate model and use the SOLUTION option. Use the parameter estimates to write the regression equation for each gender.

```
proc glm data=STAT2.school;
  class gender;
  model reading3=gender|words1 / solution;
run;                                         *ST204s01.sas;
quit;
```

Partial PROC GLM Output

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	17.17307433	B	3.84718936	4.46 <.0001
Gender F	-7.45084258	B	6.55582187	-1.14 0.2576
Gender M	0.00000000	B	.	.
Words1	1.65070369	B	0.16394259	10.07 <.0001
Words1*Gender F	0.62715889	B	0.26857250	2.34 0.0209
Words1*Gender M	0.00000000	B	.	.

**Note:** The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The equation for **Gender=M** is the following:

$$\text{Reading3}=17.17307+1.65070*\text{Words1}$$

The equation for **Gender=F** is the following:

$$\text{Reading3} = (17.17307 - 7.45084) + (1.65070 + 0.62716) * \text{Words1} = 9.72223 + 2.27786 * \text{Words1}$$

## 2. Least Squares Means for an ANCOVA Model

Use the LSMEANS statement to determine whether there is a difference between the genders when **Words1** equals the average value, 40, or 60. Use the ADJUST=TUKEY option. Examine the means plots and the diffograms.

```
ods html close; ods html style=journal;
proc glm data=STAT2.school;
  class gender;
  model reading3=gender|words1;
  lsmeans gender / pdiff adjust=tukey;
  lsmeans gender / at words1=40 pdiff adjust=tukey;
  lsmeans gender / at words1=60 pdiff adjust=tukey;
title 'L-S Means';
run;
quit;                                *ST204s02.sas;
```

Partial PROC GLM Output

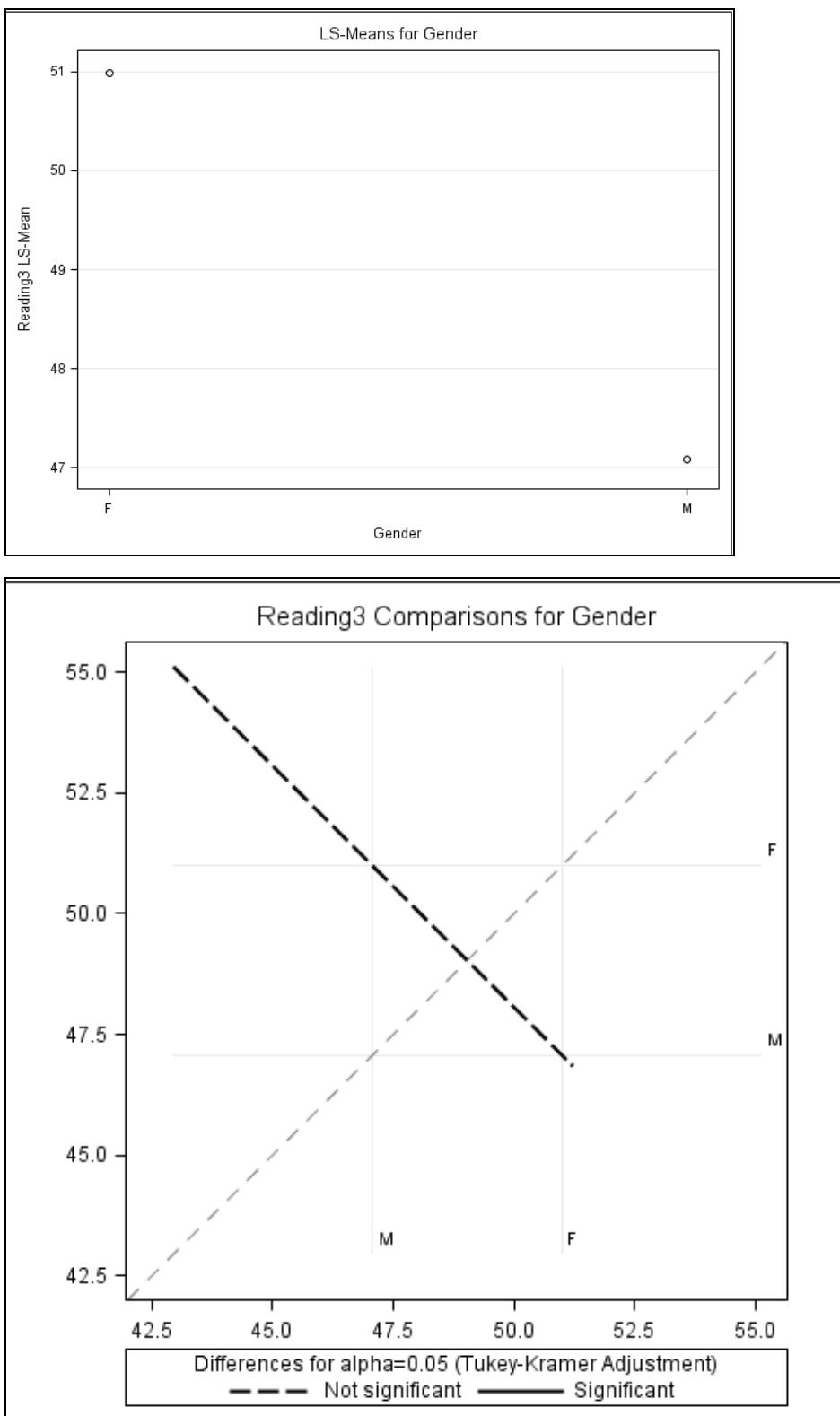
Least squares means are calculated with **Words1** at the mean score of 18.12.

### L-S Means

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

<i>H0:LSMean1=LSMean2</i>		
<i>Gender</i>	<i>Reading3 LSMEAN</i>	<i>Pr &gt;  t </i>
<i>F</i>	50.9900018	0.3515
<i>M</i>	47.0786801	

The *p*-value shown is 0.3515. Therefore, you do not reject the null hypothesis. At the average value of **Words1** (18.12), there is **not** sufficient evidence to conclude that there is a difference in the average **Reading3** scores for boys and girls.



Notice that the line segment crosses the diagonal reference line in the diffogram. This indicates no significant difference between the two least squares means when **Words1** equals its average value of 18.12.

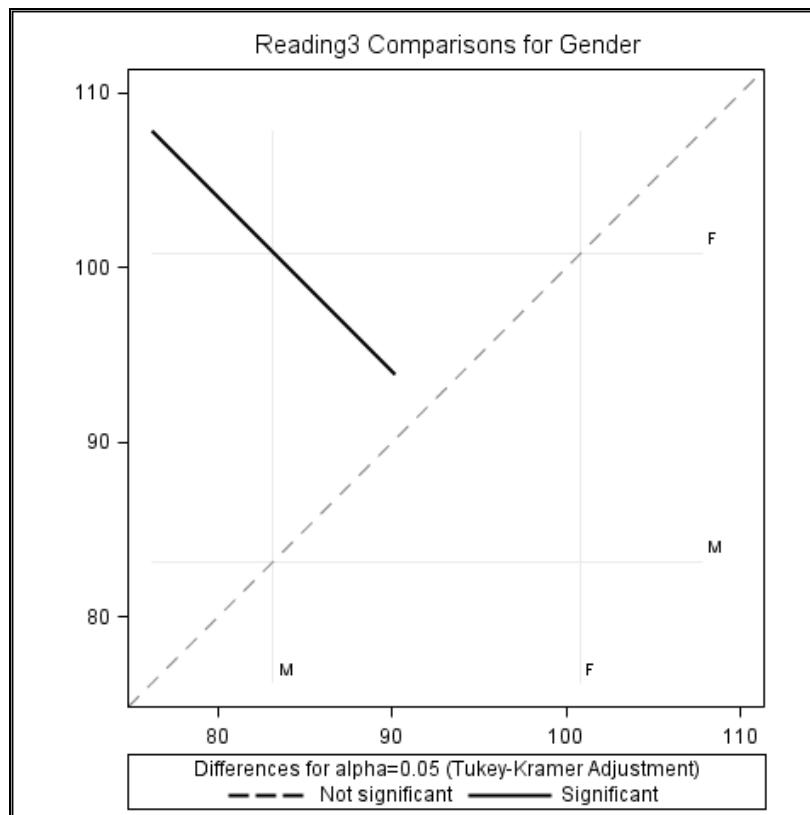
For the second LSMEANS statement, least squares means are calculated with **Words1** at a value of 40.

### L-S Means

**The GLM Procedure**  
**Least Squares Means at Words1=40**  
**Adjustment for Multiple Comparisons: Tukey-Kramer**

Gender	$H_0: LSMean1 = LSMean2$	
	Reading3 LSMEAN	Pr >  t
F	100.836735	0.0136
M	83.201222	

The  $p$ -value is 0.0136. Presuming a level of significance of 0.05, you reject the null hypothesis. When **Words1**=40, there **is** sufficient evidence to conclude that there is a difference in the average **Reading3** scores for boys and girls. Notice that the line segment in the diffogram does not cross the diagonal reference line.



Notice that the line segment does not cross the diagonal reference line. This indicates that when **Words1** equals 40, **Reading3** scores for boys and girls do differ significantly.

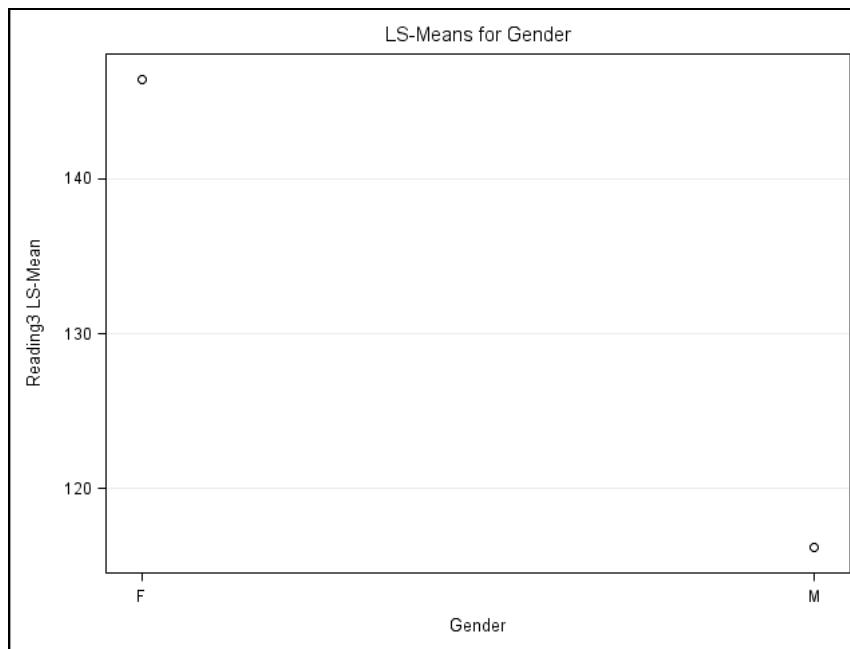
For the third LSMEANS statement, least squares means are calculated with **Words1** at a value of 60.

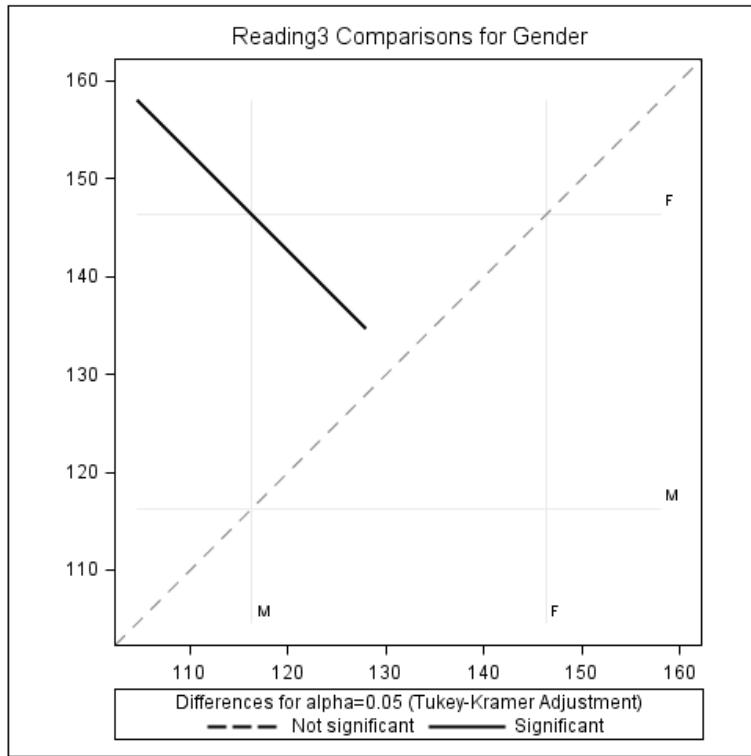
**L-S Means**

**The GLM Procedure**  
**Least Squares Means at Words1=60**  
**Adjustment for Multiple Comparisons: Tukey-Kramer**

		$H_0: LS\text{Mean}1 = LS\text{Mean}2$
Gender	Reading3 LSMEAN	$Pr >  t $
F	146.393986	0.0117
M	116.215296	

The  $p$ -value shown is 0.0117. Presuming a level of significance of 0.05, you reject the null hypothesis. When **Words1**=60, there **is** sufficient evidence to conclude that there is a difference in the average **Reading3** scores for boys and girls.





The diffogram illustrates that when **Words1** equals 60, **Reading3** scores for boys and girls are significantly different.

- The TUKEY adjustment might not be necessary because you were comparing only two groups in each LSMEANS statement. The adjusted tests are the same as the unadjusted tests for this situation. It was used for practice purpose.

### 3. Model Diagnostics

Check the model for multicollinearity. If multicollinearity appears to be a problem, center **Words1** and refit the model with the centered variable. Compare the results to the previous model.

```
ods select none;
proc glmselect data=STAT2.school outdesign=design;
  class gender;
  model reading3=gender|words1 / selection=none;
run;
%put macro variable _glsmod=&_glsmod;

ods select ParameterEstimates;
proc reg data=design;
  model reading3=&_glsmod / vif;
title 'Check Collinearity on ANCOVA Model';
run; quit; *ST204s03.sas;
```

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation	
Intercept	Intercept	B	17.17307	3.84719	4.46	<.0001	0	
Gender_F	Gender F	B	-7.45084	6.55582	-1.14	0.2576	2.48566	
Gender_M	Gender M	0	0	.	.	.	.	
Words1	Words1	B	1.65070	0.16394	10.07	<.0001	1.61124	
Words1_Gender_F	Words1*Gender F	B	0.62716	0.26857	2.34	0.0209	3.27030	
Words1_Gender_M	Words1*Gender M	0	0	.	.	.	.	

Multicollinearity does not appear to be a problem with this model on the original scale of the data. All of the variance inflation factors are below 10. You can center **Words1**, because this gives the intercept the interpretation of the average **Reading3** score for boys with the average **Words1** score. The current interpretation of the intercept is the average **Reading3** score for boys with the **Words1** score of 0.

## Solutions to Student Activities (Polls/Quizzes)

### 4.01 Multiple Choice Poll – Correct Answer

In the previous demonstration, the parameter estimate for **BaselineBP** is -0.1673. This is the slope that corresponds to

- a. Approved Drug
- b. New Drug
- c. Placebo

20

### 4.02 Quiz – Correct Answer

How do you interpret the significant term **Words1ByGenderF** for this model?

The slopes for Words1 are different between male students and female students.

24

### 4.03 Poll – Correct Answer

In ANCOVA models, the least squares means for a CLASS variable is adjusted for the covariate.

- True
- False



# Chapter 5 Introduction to Generalized Linear Models

<b>5.1 Introduction to Generalized Linear Models.....</b>	<b>5-3</b>
<b>5.2 Introduction to Poisson Regression and Negative Binomial Regression .....</b>	<b>5-8</b>
Demonstration: Fitting a Poisson Regression Model for Count Data .....	5-16
Demonstration: Modeling Overdispersion By Using the Negative Binomial Distribution.....	5-28
Exercises .....	5-37
Demonstration: Fitting a Poisson Regression Model for Rate Data .....	5-44
<b>5.3 Introduction to Gamma Regression .....</b>	<b>5-47</b>
Demonstration: Fitting a Gamma Regression Model.....	5-51
Exercises .....	5-63
<b>5.4 Chapter Summary.....</b>	<b>5-64</b>
<b>5.5 Solutions .....</b>	<b>5-66</b>
Solutions to Exercises .....	5-66
Solutions to Student Activities (Polls/Quizzes) .....	5-79



# 5.1 Introduction to Generalized Linear Models

---

## Objectives

- Define generalized linear models.
- Identify examples of generalized linear models.

3

## General Linear Models

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

$$\varepsilon_i \sim i.i.d. N(0, \sigma^2)$$

$$\begin{array}{l} \xrightarrow{\hspace{1cm}} E(y_i) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \\ \text{var}(y_i) = \sigma^2 \end{array}$$

5

$y_i$  is the  $i^{th}$  observed value for the response variable.

$x_{1i} \dots x_{ki}$  is the  $i^{th}$  value for the independent variables  $x_1 \dots x_k$ .

$\beta_1 \dots \beta_k$  are the regression coefficients for the corresponding independent variables.

$\varepsilon_i$  is the  $i^{th}$  value of random errors.

Assume that the random errors are independently and identically distributed following a normal distribution with a mean of 0 and a variance of  $\sigma^2$ . For general linear models, you have  $E(\varepsilon_i) = 0$  and  $Var(\varepsilon_i) = \sigma^2$ . It follows that  $E(y) = \mathbf{X}\beta$  and  $Var(y) = \sigma^2 I_n$ , where  $I_n$  is an  $n \times n$  identity matrix.

## Generalized Linear Models

$$g(E(y_i)) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} = \mathbf{X}\beta$$

- The distribution of the observations can come from the exponential family of distributions.
- The variance of the response variable can be expressed as a function of its mean.
- $\mathbf{X}\beta$  is fit to a function of  $E(y)$  (called a link function) suggested by the distribution of the observations:

$$g(E(y)) = g(\mu) = \mathbf{X}\beta$$

**Link function**

6

Generalized linear models were defined by Nelder and Wedderburn (1972) and expanded upon by McCullagh and Nelder (1989). A generalized linear model extends general linear models in three ways.

1. The distribution of the observations can come from the family of exponential distributions; no assumption of normality is required. This includes distributions such as the normal, gamma, Poisson, binomial, and negative binomial distributions.
2. The variance of the response variable can be specified as a function of its mean. (In the case of the normal distribution, the relationship can be expressed as  $\sigma^2 = \mu^2 \cdot \sigma^2$ . See below for details.)
3. The link function  $g$  is introduced to fit the linear model. This link function must be monotonic, but does not have to be the identity function, as is the case for general linear models.

### Details

From the SAS online documentation, the general form for a density or probability function of the exponential family can be expressed as  $f(y | \theta) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\}$ , where  $\theta$  is the natural location parameter, and  $\phi$  is the dispersion parameter.

For this family of distributions, the variance of  $y$ ,  $Var(y)$ , can be expressed as a function of the mean of  $y$  as  $Var(y) = V(\mu)a(\phi)$ , where  $V(\mu)$  denotes the variance function, and  $a(\phi)$  denotes a function of the dispersion parameter. For the normal distribution,  $V(\mu)$  is the identity function and  $a(\phi) = \sigma^2$ , so the relationship between the mean and variance can be expressed as  $Var(Y) = V(\mu)a(\phi) = \mu^0 \cdot \sigma^2$ .

An alternative parameterization for the general form for a density or probability mass function for an exponential family distribution is  $f(y|\theta) = h(y) \cdot c(\theta) \cdot e^{\sum t_i(y_i)(W_i(\theta))}$ . When written in this form, the canonical link can be identified as  $W_i(\theta)$ . (Casella and Berger 1990)

### Logit Link Function for Binary Response

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

7

The *logit transformation* is the log of the odds, which is the ratio of the probability of the outcome to the probability of no outcome. To create a linear model, the logit transformation is applied to the probability. Unlike a probability, the logit is unbounded because transforming the probability to odds removes the upper bound. Taking the natural logarithm of the odds removes the lower bound. The model (also called the *logistic regression model*) is now linear because the logit is linear in its parameters. Furthermore, the model gives estimated probabilities that are between 0 and 1.

#### Details

To identify the canonical link for binary data, start with the mass function of a Bernoulli (binary) random variable and rearrange to the following format (shown previously):

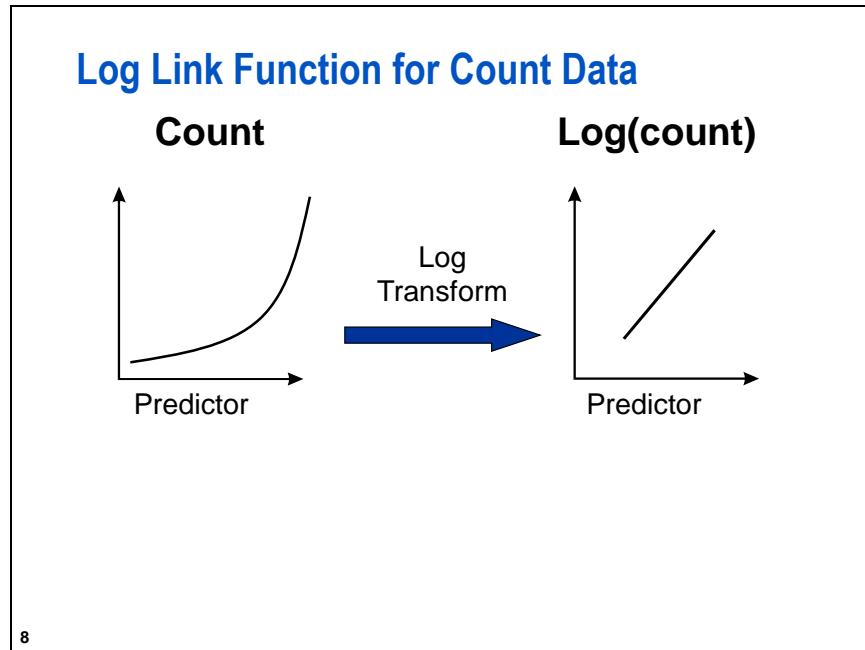
$$\begin{aligned}
 f(y|\theta) &= h(y) \cdot c(\theta) \cdot e^{\sum t_i(y_i)(W_i(\theta))} \\
 f(y|p) &= p^y \cdot (1-p)^{(1-y)} = p^y \cdot (1-p)^1 \cdot (1-p)^{-y} \\
 &= \frac{p^y \cdot (1-p)}{(1-p)^y} = (1-p) \cdot I_y \cdot \left(\frac{p}{1-p}\right)^y = (1-p) \cdot I_y \cdot e^{\log\left(\frac{p}{1-p}\right)y} \\
 &= (1-p) \cdot I_y \cdot e^{y \cdot \log\left(\frac{p}{1-p}\right)} \quad \text{where } I_y = \begin{cases} 1 & \text{for } y=0,1 \\ 0 & \text{otherwise} \end{cases}
 \end{aligned}$$

In this format you can see that the canonical link is  $\log\left(\frac{p}{1-p}\right)$ .

In PROC GENMOD, the location parameter is  $p$ , and the scale parameter is 1.



The notation **log** refers to the natural logarithm, which is the logarithm to the base e.



8

The log link function is often applied to count data that have nonnegative integer values. The log transformation removes the lower bound and creates a linear model.

#### Details

The probability mass function for Poisson distribution is given by the following:

$$f(y | \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad y = 0, 1, 2, \dots$$

In this expression,  $y$  is a nonnegative integer value and  $\lambda$  is the expected value of  $Y$ . It can be shown that  $\text{Var}(y) = \lambda$  and the scale parameter is 1.

To identify the canonical link, rearrange it to the following format:  $f(y | \lambda) = h(y) \cdot c(\lambda) \cdot e^{t(y)(W(\lambda))}$

$$f(y | \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} = \frac{e^{-\lambda} \cdot e^{\log(\lambda^y)}}{y!} = e^{-\lambda} \cdot \frac{1}{y!} \cdot e^{y(\log(\lambda))}.$$

The last term indicates that the log is the canonical link.

In PROC GENMOD, the location parameter is  $\lambda$  and the scale parameter is 1.

Examples of Generalized Linear Models					
Model	Response	Distribution	Mean	Variance	Canonical Link
Linear Regression	Continuous	Normal	$\mu$	$\sigma^2$	identity $\mu$
Logistic regression	Dichotomous	Binomial	$\pi$	$\pi(1 - \pi)/n$	logit $\log[\pi/(1-\pi)]$
Poisson Regression	Count	Poisson	$\lambda$	$\lambda$	log $\log(\lambda)$
Gamma Regression	Continuous	Gamma	$\mu$	$\mu^2/\nu$	*inverse $1/\mu$

<sup>9</sup> \* Models often use the LOG link in practice.

For the gamma distribution, the canonical link function (the default link function in PROC GENMOD) is the inverse. However, when practical or theoretical limitations indicate, modelers often use the log link function to constrain the unlinked expected values to be positive.

### Details

The probability density for the normal distribution is  $f(y | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(y-\mu)^2}{(2\sigma^2)}}$ ,

where  $\mu$  and  $\sigma$  are the location and scale parameters, respectively, in PROC GENMOD.

The probability density for the gamma distribution is  $f(y | \nu, \mu) = \frac{1}{\Gamma(\nu) \cdot y} \cdot \left(\frac{y\nu}{\mu}\right)^\nu \cdot e^{-\left(\frac{y\nu}{\mu}\right)}$ ,

where  $\mu$  and  $\nu$  are the location and scale parameters, respectively, in PROC GENMOD.

## 5.02 Multiple Choice Poll

Which of the following is *not* true?

- a. Logistic regression, Poisson regression, and gamma regression are all examples of generalized linear models.
- b. In generalized linear models, the mean and variance are unrelated.
- c. Canonical link functions are commonly used link functions for many exponential family distributions.
- d. Although the canonical link function for the gamma distribution is the inverse, modelers often use the logarithm link function.

11

## 5.2 Introduction to Poisson Regression and Negative Binomial Regression

---

### Objectives

- Define Poisson regression models.
- Define overdispersion in Poisson regression models.
- Use the GENMOD procedure to fit a Poisson regression model and correct for overdispersion.
- Define Poisson regression for rate data.

14

## Poisson Regression

Poisson regression

- is one type of generalized linear model
- assumes that the response variable follows a Poisson distribution that is conditional on the values of the predictor variables
- can be used to model the number of occurrences of an event of interest or the rate of occurrence of an event of interest as a function of some predictor variables
- is most appropriate for counts of rare events.



15

Poisson regression is often used to analyze count data, where the response variable has nonnegative integer values (0, 1, 2, 3, and so on). Poisson regression can also be used to model the rate or incidence of an event. This type of outcome is widely seen in the medical sciences, biological sciences, social sciences, agriculture, engineering, and business.

The assumption in Poisson regression is that the conditional distribution of the response variable follows a Poisson distribution. This distribution is the benchmark distribution for count data in much the same way that the normal distribution is the benchmark for continuous data. Although ordinary least squares regression can be used to analyze count data, Poisson regression has the advantage of being precisely customized to the discrete, often skewed distribution of count data.

In addition to being skewed, the sample distribution should have a fairly small mean if Poisson regression is the method of choice. The mean should certainly be below 10, preferably below 5, and ideally close to 1 (Zar 1996).

- ✍ The gamma distribution or the lognormal distribution might be more appropriate for highly skewed data with large mean values.
- ✍ Many times, count data have too many zeros to use the standard Poisson model and a zero-inflated Poisson model must be used. This is discussed in the Fitting Poisson Regression Models Using the GENMOD Procedure course.

## Poisson Regression Outcome Variables

Examples include the following:

- number of ear infections in infants
- number of equipment failures
- colony counts for bacteria or viruses
- counts of a rare disease in a population
- number of fatal crashes at an intersection
- homicide rates in a given state
- rate of insurance claims
- number of infected areas per unit volume of a tree
- response rates to a marketing campaign



16

When outcomes occur over time, space, or some other index of size, it might be more relevant to model the **rate** of occurrence rather than the counts. If the size measures vary across observations, modeling the rates provides the necessary standardization to ensure that the outcomes are comparable. Poisson regression for rates is discussed later in this chapter.

## Poisson versus Normal Distribution

Poisson distribution

- is skewed to the right for rare events
- is for nonnegative integer values
- has only one parameter (the mean)
- has a variance that is equal to the mean.

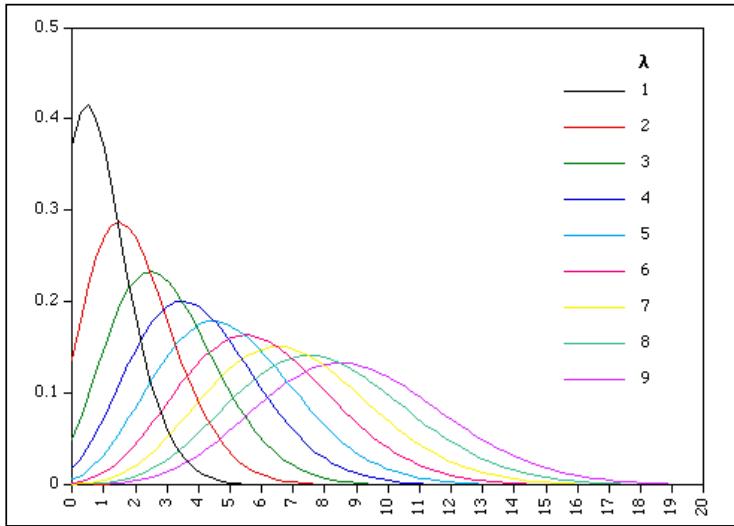
Normal distribution

- is symmetric
- can be for negative as well as positive real values
- has two unrelated parameters (mean and variance).

17

The Poisson distribution is fully defined by one parameter, the mean ( $\lambda$ ), which must be positive. An unusual property of the Poisson distribution is that the mean and variance are equal. This can be a serious limitation because count observations often exhibit variability exceeding that predicted by the Poisson distribution. This leads to overdispersion, which is addressed in a later section.

## Poisson Distributions with Different Means



18

For rare events, the Poisson distribution is different from the normal distribution. As the mean increases, however, the Poisson distribution approximates the normal distribution.

- ✍ The Poisson distribution is a discrete distribution, which might be more conventionally represented by a bar chart. The graph shown above represents a continuous approximation to the bar chart for the Poisson distribution.

## Poisson Regression Model

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

$$\mu = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}$$

$$= e^{\beta_0} \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \cdots e^{\beta_k X_k}$$

19

Because the Poisson mean is required to be positive, an additive model for the mean is unsatisfactory. Linear predictors can become negative for certain parameter combinations. Poisson regression models use a log link function that relates the expected value of the response variable to the linear predictor. This ensures that the mean remains positive for all linear predictors, and hence positive for all parameter and covariate combinations. The fitted values are now the exponentiation of the linear predictor.

The method of maximum likelihood is used to estimate the parameters of Poisson regression models. This method finds the parameter estimates that are most likely to occur given the data. These parameter estimates maximize the likelihood function, which expresses the probability of the observed data as a function of the unknown parameters.

## Poisson Regression Parameter Estimates

$$e^{\hat{\beta}} = \text{multiplicative effect on } \hat{\mu} \text{ for a one-unit change in } X$$

**Example 1, if**

$$e^{\hat{\beta}_1} = 1.20, \text{ then a one-unit increase in } X_1 \text{ yields a 20\% increase in the estimated mean.}$$

**Example 2, if**

$$e^{\hat{\beta}_2} = 0.80, \text{ then a one-unit increase in } X_2 \text{ yields a 20\% decrease in the estimated mean.}$$

20

Because the Poisson model uses a log link function, the parameter estimates represent the expected change in the log scale. If you calculate  $100(e^{\hat{\beta}} - 1)$ , you obtain the percent change in the expected number of events with each one-unit increase in the predictor variable.

For example, if  $e^{\hat{\beta}_1} = 1.20$ , then a one-unit increase in  $X_1$  yields a 20% increase in the estimated mean.

If  $e^{\hat{\beta}_2} = 0.80$ , then a one-unit increase in  $X_2$  yields a 20% decrease in the estimated mean.

## 5.03 Multiple Choice Poll

Why can you not use OLS regression for a count of rare events with a skewed distribution?

- a. OLS regression assumes normal distribution of errors.
- b. OLS regression assumes constant variance.
- c. OLS regression can produce both positive and negative predicted values.
- d. all the above

21

## Female Horseshoe Crab Example



Number of Satellites

- Color
- Spine condition
- Carapace width in centimeters
- Weight in grams

24

The data come from a study that was conducted on the mating habits of female horseshoe crabs. The population of horseshoe crabs is monitored because they provide a critical food source for migrating birds. Each year, at the end of May and during June, hundreds of thousands of horseshoe crabs emerge from Delaware Bay to lay and fertilize their eggs.

Each female horseshoe crab had a male crab resident in her nest. The study investigated factors affecting whether the female horseshoe crab had any other males, called *satellites*, residing nearby. The response variable for each female horseshoe crab is her number of satellites. The data are stored in **STAT2.crab**.

## The Data

Width	Weight	Color	Spine	Satellites
28.3	3.05	2	3	8
22.5	1.55	3	3	0
26.0	2.30	1	1	9
24.8	2.10	3	3	0
26.0	2.60	3	3	4
23.8	2.10	2	3	0
26.5	2.35	1	1	0
24.7	1.90	3	2	0
23.7	1.95	2	1	0
25.6	2.15	3	3	0
24.3	2.15	3	3	0
25.8	2.65	2	3	0
28.2	3.05	2	3	11
21.0	1.85	4	2	0
26.0	2.30	2	1	14
27.1	2.95	1	1	8
...				

**Color:** 1=Light Medium 2=Medium 3=Dark Medium 4=Dark

**Spine:** 1=Both Good 2=One Worn or Broken 3=Both Worn or Broken

25

These are the variables in the **STAT2.crab** data set:

- Satellites** the number of satellites or male horseshoe crabs residing nearby
- Color** female horseshoe crab's color (*Light Medium, Medium, Dark Medium, Dark*)
- Spine** spine condition (*Both Good, One Worn/Broken, Both Worn/Broken*)
- Width** carapace width in centimeters
- Weight** weight in kilograms

 The data are from an example in Agresti (1996).

## The GENMOD Procedure

General form of the GENMOD procedure:

```
PROC GENMOD options PLOTS=requests;  

CLASS variables;  

MODEL response=effects / options;  

ESTIMATE 'label'effect values / options;  

RUN;
```

26

The GENMOD procedure fits generalized linear models with a number of built-in link functions and probability distributions. The available link functions are the identity, log, logit, probit, power, cumulative complementary log-log, cumulative logit, cumulative probit, and complementary log-log. PROC GENMOD also allows user-defined link functions. The available probability distributions are binomial, gamma, inverse Gaussian, multinomial, negative binomial, normal, and Poisson. You can also specify a user-defined probability distribution.

Selected PROC GENMOD statement options:

PLOTS (*global-plot-options*) = (*plot-request (options)*... *plot-request (options)*)

Here are some examples:

- PLOTS=ALL
- PLOTS=PREDICTED
- PLOTS=(PREDICTED RESCHI)
- PLOTS(UNPACK)=DFBETA

The default plots with the *plots=all* option in the PROC GENMOD statement include diagnostic plots, DFBETA plots, standardized DFBETA plots, and plots of the residuals versus the observation number.

Selected GENMOD procedure statements:

- |          |  |
|----------|--|
| CLASS    | specifies the classification variables to be used in the analysis. If the CLASS statement is used, it must appear before the MODEL statement.  |
| MODEL    | specifies the response variable and the predictor variables.   |
| ESTIMATE | provides a means for obtaining a test for a specified hypothesis concerning the model parameters. It can be used to produce the ratio of the expected number of events for subjects in one category compared to another category along with the 95% confidence limits. |



## Fitting a Poisson Regression Model for Count Data

---

First, use the SGLOT procedure to examine a histogram of the data. Superimpose a normal distribution and a nonparametric curve fit to the data to see whether the normal distribution is a good fit. Further explore the data by looking at summary statistics from PROC UNIVARIATE.

```

title;
proc sgplot data=STAT2.crab;
  histogram satellites;
  density satellites;
  density satellites / type=kernel;
run;

ods select moments goodnessoffit;
proc univariate data=STAT2.crab;
  var satellites;
  histogram / normal;
run;                                         *ST205d01.sas;

```

Selected SGLOT procedure statement:

DENSITY      creates a density curve that shows the distribution of values for a numeric variable.

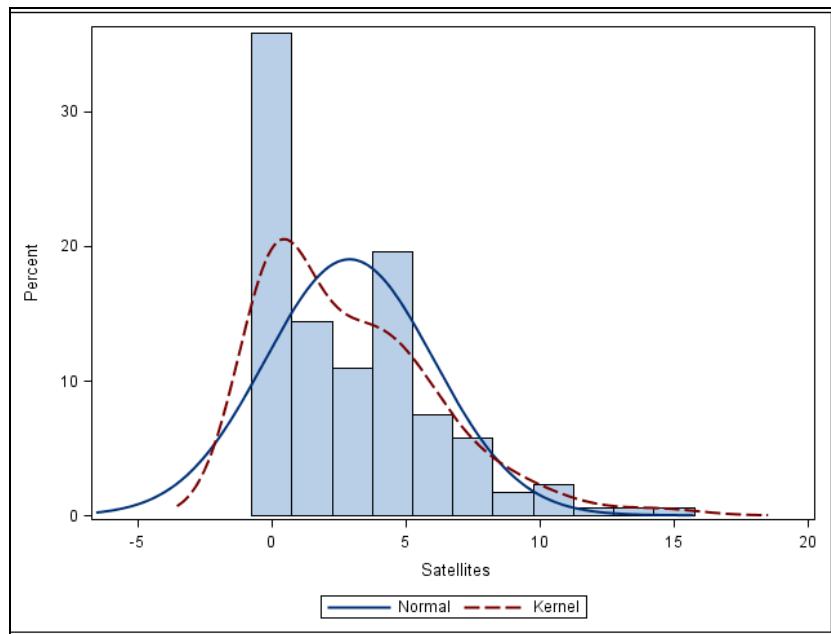
Selected DENSITY statement option:

TYPE      specifies the type of distribution curve that is used for the density plot. Specify one of the following keywords:

NORMAL = (*normal-options*) specifies a normal density estimate, with a mean and a standard deviation. The normal density is the default.

KERNEL = (*kernel-opts*) specifies a nonparametric kernel density estimate.

## PROC SGLOT Output



The histogram shows that the data are skewed to the right. The kernel and normal density superimposed on the histogram indicate that the variable **Satellites** does not follow a normal distribution.

## Partial PROC UNIVARIATE Output

The UNIVARIATE Procedure Variable: Satellites			
Moments			
N	173	Sum Weights	173
Mean	2.91907514	Sum Observations	505
Std Deviation	3.14833571	Variance	9.91201774
Skewness	1.14523732	Kurtosis	1.26072382
Uncorrected SS	3179	Corrected SS	1704.86705
Coeff Variation	107.853877	Std Error Mean	0.23936353
Basic Statistical Measures			
Location		Variability	
Mean	2.919075	Std Deviation	3.14834
Median	2.000000	Variance	9.91202
Mode	0.000000	Range	15.00000
		Interquartile Range	5.00000

The skewness statistic (1.145) is greater than 0. The mean value (2.92) is greater than the median (2.0). These indicate a possible skewed-to-the-right distribution. The average for **Satellites** is relatively small, and the variance (9.91) is much bigger than the mean (2.92).

Now, use the GENMOD procedure to fit a Poisson regression model.

```
proc genmod data=STAT2.crab;
  class color spine;
  model satellites=width weight color spine
    / dist=poi link=log type3;
  title 'Poisson Model';
run; *ST205d01;
```

Selected MODEL statement options:

DIST= specifies the built-in probability distribution to use in the model. If you specify the DIST= option and you omit a user-defined link function, a default link function is chosen as displayed in the following table:

DIST=	Distribution	Default Link Function
BINOMIAL   BIN   B	binomial	logit
GAMMA   GAM  G	gamma	inverse ( power(-1) )
IGAUSSIAN   IG	inverse Gaussian	inverse squared ( power(-2) )
MULTINOMIAL   MULT	multinomial	cumulative logit
NEGBIN   NB	negative binomial	log
NORMAL   NOR   N	normal	identity
POISSON   POI   P	Poisson	log

LINK= specifies the link function to be used in the model. The keywords and their associated built-in link functions are as follows:

LINK=	Link Function
CUMCLL   CCLL	cumulative complementary log-log
CUMLOGIT   CLOGIT	cumulative logit
CUMPROBIT   CPROBIT	cumulative probit
CLOGLOG   CLL	complementary log-log
IDENTITY   ID	identity
LOG	log
LOGIT	logit
PROBIT	probit
POWER( <i>number</i> )   POW( <i>number</i> )	power with $\lambda = \text{number}$

 If you specify no distribution and no link function, then the GENMOD procedure defaults to the normal distribution with the identity link function.

TYPE3 requests that statistics for Type 3 contrasts be computed for each effect specified in the MODEL statement. The default analysis is to compute likelihood ratio statistics for the contrasts or score statistics for GEE (Generalized Estimating Equations).

#### Partial PROC GENMOD Output

The GENMOD Procedure		
Model Information		
Data Set	STAT2.CRAB	
Distribution	Poisson	
Link Function	Log	
Dependent Variable	Satellites	
Number of Observations Read 173		
Number of Observations Used 173		
Class Level Information		
Class	Levels	Values
Color	4	1 2 3 4
Spine	3	1 2 3

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	165	549.5856	3.3308
Scaled Deviance	165	549.5856	3.3308
Pearson Chi-Square	165	533.8165	3.2353
Scaled Pearson X2	165	533.8165	3.2353
Log Likelihood		77.5928	
Full Log Likelihood		-452.4416	
AIC (smaller is better)		920.8833	
AICC (smaller is better)		921.7613	
BIC (smaller is better)		946.1096	

The Criteria For Assessing Goodness Of Fit table provides statistics for testing the goodness of fit of the model. The measures are the deviance and the Pearson chi-squared statistic. The values of these statistics divided by the squared scale parameter (that is, the dispersion parameter) are called scaled deviance and scaled Pearson chi-squared. Because the scale parameter by definition is 1 for Poisson regression, the statistics (original and scaled) are equal.

The Value/DF values are computed by dividing the goodness-of-fit statistics by the degrees of freedom. (The degrees of freedom for the Deviance and Pearson Chi-Square are equal to the number of observations minus the number of regression parameters estimated.) These values for the scaled deviance or the scaled Pearson chi-square are useful for assessing the goodness of model fit. Values close to 1 indicate good model fit. The Value/DF column in the table has 3.3308 for scaled deviance and 3.2353 for scaled Pearson chi-square. They are not close to 1. This might indicate overdispersed data, which can occur frequently in Poisson regression, and occasionally in logistic regression. Overdispersion does not affect the parameter estimates, but it causes the estimates of the standard error of the parameter estimates to be underestimated. More detailed discussion is provided later in the course.

Other fit statistics include the Akaike information criterion (AIC), the corrected Akaike information criterion (AICC), and the Bayesian information criterion (BIC). Each is a measure of goodness of model fit that balances model fit against model simplicity. These criteria are useful in selecting among models, with smaller values representing better model fit.

- ✍ The scale parameter can be estimated from your data. For the Poisson distribution, you divide the Pearson chi-square statistic (or the deviance statistic) by the degrees of freedom (which is indicated by the Value/DF column), and then take the square root.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq	
Intercept	1	-0.8054	0.9443	-2.6562	1.0454	0.73	0.3937	
Width	1	0.0167	0.0489	-0.0791	0.1126	0.12	0.7321	
Weight	1	0.4965	0.1663	0.1706	0.8223	8.92	0.0028	
Color	1	1	0.5309	0.2269	0.0861	0.9756	5.47	0.0193
Color	2	1	0.2660	0.1650	-0.0574	0.5895	2.60	0.1070
Color	3	1	0.0172	0.1809	-0.3375	0.3718	0.01	0.9245
Color	4	0	0.0000	0.0000	0.0000	0.0000	.	.
Spine	1	1	-0.0873	0.1199	-0.3223	0.1478	0.53	0.4667
Spine	2	1	-0.2377	0.1980	-0.6258	0.1505	1.44	0.2301
Spine	3	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

**Note:** The scale parameter was held fixed.

The Analysis Of Maximum Likelihood Parameter Estimates table provides the parameter estimates and the *p*-values for testing whether the estimates are different from zero. The LR Statistics for Type 3 Analysis gives the tests of significance for each of the parameters. However, because the data exhibit overdispersion, these results might not be reliable.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Width	1	0.12	0.7324
Weight	1	9.04	0.0026
Color	3	9.24	0.0263
Spine	2	1.79	0.4076

The likelihood ratio test for the predictor variables indicates a significant **Weight** effect and a significant **Color** effect. You can compare these tables to those obtained after correcting for the overdispersion.

## Overdispersion

- Poisson regression models assume that the variance is equal to the mean.
- Count data often exhibit variability that exceeds the mean.
- Overdispersion leads to underestimates of the standard errors of parameter estimates.
- Overdispersion results in overestimates of the test statistic and liberal  $p$ -values.

29

Overdispersion occurs when the observed variance is larger than the nominal variance for a particular distribution. In the Poisson distribution, it is assumed that the mean equals the variance. However, when you model count data, the variances are usually much higher than the means. When the model for the mean is correct but the variance does not follow what is defined by the Poisson distribution, the maximum likelihood estimates of the model parameters are still consistent, but the standard errors are incorrect. In fact, overdispersion leads to underestimates of the standard errors and overestimates of the test statistics, which increase the Type I error rate. Underdispersion can also occur where the standard errors are overestimated and the test statistics are underestimated, which increase the Type II error rate.

Overdispersion is not a problem in ordinary least squares (OLS) regression because the normal distribution has a separate parameter, the *variance*, to describe variability. OLS regression is generally not appropriate for count data because it assumes constant variances, which might lead to incorrect inferences for count data.

## 5.05 Poll

Is overdispersion a problem in ordinary least squares regression?

- Yes
- No

30

## Causes of Overdispersion

- subject heterogeneity due to an under-specified model
- outliers in the data
- positive correlation between the responses in clustered data

32

Poisson regression models assume that the response variable has a Poisson distribution conditional on the values of the predictor variables. If some relevant predictor variables are not in the model, then the unexplained heterogeneity among the subjects causes greater variation in the response than the Poisson predicts. If the variance equals the mean when all the relevant predictor variables are controlled for, it exceeds the mean when relevant predictor variables are not controlled for (Agresti 1996).

Therefore, overdispersion occurs when you have an under-specified model, and the variability between subjects is not being adequately accounted for. Because there is no random error term in a Poisson regression model, there is no way to account for the extra variability caused by the omitted important predictor variables. Therefore, assuming a Poisson distribution for a count variable is often too simplistic because most models are missing relevant predictor variables.

Overdispersion can also occur due to clustering in the population. Examples of naturally occurring clusters are families, households, litters, colonies, and neighborhoods. If your sample includes a large number of observations within a cluster, the positive correlation between the responses might cause overdispersion (McCullagh and Nelder 1989).

## Correcting for Overdispersion



- Make sure that you do not have erroneous data.
- Recheck your model to include all important variables. Re-specify your model if necessary.

After these checks are completed, you can do one of the following:

- Use the negative binomial distribution to model the overdispersion (DIST=NEGBIN option in the MODEL statement in PROC GENMOD).
- or
- Apply a multiplicative adjustment factor to adjust the standard errors accordingly (PSCALE or DSCALE option in the MODEL statement in PROC GENMOD).

33

One way to deal with overdispersion is to introduce a more flexible distribution than the Poisson distribution in the model. A related distribution for count data that permits the variance to exceed the mean is the negative binomial distribution. The binomial distribution counts the number of success in a fixed number of Bernoulli trials. The ***negative*** binomial distribution counts the number of Bernoulli trials that are required to obtain a set number of successes.

Another way of accounting for overdispersion is to fit a Poisson model and use the PSCALE or DSCALE option in the MODEL statement to adjust the standard errors to account for overdispersion. This fixes the scale parameter at the value of 1 in the estimation procedure. After the parameter estimates are determined, the exponential family dispersion parameter is assumed to be given by Pearson's chi-square statistic divided by the degrees of freedom. All statistics such as standard errors and likelihood ratio statistics are adjusted appropriately. For the binomial and Poisson distributions, which have no free scale parameter, this can be used to specify an *overdispersed* model.

## Negative Binomial Distribution

The negative binomial distribution

- is the distribution for count data that permits the variance to exceed the mean
- enables the model to have greater flexibility in modeling the relationship between the mean and the variance of the response variable than the Poisson model has.

34

The negative binomial distribution can be used to describe the distribution arising from an experiment consisting of a sequence of independent Bernoulli trials. The probability of success for each trial,  $p$ , is constant across the experiment. The experiment continues until a fixed number of successes are achieved.

This distribution is a generalization of the Poisson distribution for count data. If the distribution of  $Y$  is Poisson, given the mean at a fixed setting of the predictors, and the mean itself follows a gamma distribution, then it follows that the marginal distribution for the response variable  $Y$  is a negative binomial. Unlike the Poisson distribution, the negative binomial distribution provides a way to model subject heterogeneity and account for overdispersion.

The negative binomial distribution is appropriate for aggregated events. For example, the distribution of an organism in space might be a negative binomial. Organisms tend to aggregate, either because aggregation enhances survival, or because individuals of the same species favor the same habitat or environmental conditions. For example, humans aggregate in towns and cities, bacteria aggregate in colonies, birds aggregate in flocks, cows aggregate in herds, and fish aggregate in schools. The relative abundance of species in ecological communities results in overdispersion when you use a Poisson distribution. You can use the negative binomial distribution.

## Negative Binomial Model

Response Variable	Distribution	Link Function	Variance Function
Count	Negative Binomial	Natural Log	$\mu + k\mu^2$

35

The relationship between the variance and the mean for a negative binomial distribution has a dispersion parameter that must be estimated or set to a fixed value. The dispersion parameter,  $k$ , enables the variance to exceed the mean and enables the negative binomial distribution to account for overdispersion.

### Details

The probability mass function for the negative binomial distribution is as follows:

$$f(y) = \frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)} * \frac{(ku)^y}{(1+ku)^{y+1/k}} \text{ for } y = 0, 1, 2, 3\dots$$

## Dispersion Parameter $k$

- The dispersion parameter  $k$  is not allowed to vary over observations.
- The limiting case when the parameter  $k$  is equal to 0 corresponds to a Poisson regression model.
- When the parameter is greater than 0, overdispersion is evident and the standard errors increase. The fitted values are similar, but the larger standard errors reflect the overdispersion that is uncaptured with the Poisson model.

36

If the model for the mean is correctly specified, the parameter estimates from the negative binomial model are consistent (that is, as the sample size increases, the probability that the parameter estimate approaches the true value increases), even if the true distribution is not negative binomial. If overdispersion is evident, then the standard errors of the parameter estimates in the Poisson model are underestimated. It is always useful to examine the dispersion parameter to see how much greater than 0 it is. If the dispersion parameter is much greater than 0, then the negative binomial model is more appropriate than the Poisson model and the inferences from the negative binomial model are more accurate.



## Modeling Overdispersion By Using the Negative Binomial Distribution

Use the GENMOD procedure to fit a negative binomial model to the **STAT2.crab** data set in order to account for overdispersion.

```
ods html style=listing;
proc genmod data=STAT2.crab;
  class color spine;
  model satellites=width weight color spine
    / dist=negbin link=log type3 ;
  title 'Negative Binomial to Account for Overdispersion';
run; *ST205d02.sas;
```

Partial PROC GENMOD Output

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
<b>Deviance</b>	165	196.5122	1.1910
<b>Scaled Deviance</b>	165	196.5122	1.1910
<b>Pearson Chi-Square</b>	165	154.2446	0.9348
<b>Scaled Pearson X2</b>	165	154.2446	0.9348
<b>Log Likelihood</b>		157.3742	
<b>Full Log Likelihood</b>		-372.6602	
<b>AIC (smaller is better)</b>		763.3204	
<b>AICC (smaller is better)</b>		764.4247	
<b>BIC (smaller is better)</b>		791.7000	

Algorithm converged.

Because the extra dispersion parameter in the negative binomial distribution models the overdispersion, the Pearson Chi-Square / DF value is relatively close to 1, as expected. For the negative binomial distribution, the Deviance and Scaled Deviance is the same unless you specify SCALE=D. The Person Chi-Square and Scaled Pearson X2 are the same unless you specify SCALE=P. Doing so causes the scale parameter, phi, to be estimated using the Pearson (SCALE=P) or deviance statistic (SCALE=D).

- For a fixed value of the dispersion parameter, the scaled deviance is defined to be twice the difference between the maximum achievable log likelihood and the log likelihood at the maximum likelihood estimates of the regression parameters.

Comparing the fit statistics for this negative binomial model to the fit statistics for the Poisson model, you can see that the negative binomial model fits the data better.

MODELS			
FIT STATISTICS		Original Model	Negative Binomial Model
AIC (smaller is better)		920.883	763.3204
AICC (smaller is better)		921.7613	764.4247
BIC (smaller is better)		946.1096	791.7000

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq	
Intercept	1	-0.8136	2.0928	-4.9154 3.2882	0.15	0.6975	
Width	1	-0.0024	0.1108	-0.2197 0.2148	0.00	0.9825	
Weight	1	0.7009	0.4038	-0.0906 1.4923	3.01	0.0826	
Color	1	0.5790	0.4782	-0.3582 1.5162	1.47	0.2260	
Color	2	0.2582	0.3053	-0.3401 0.8566	0.72	0.3976	
Color	3	0.2582	0.3053	-0.3401 0.8566	0.72	0.3976	
Color	4	0.0000	0.0000	0.0000 0.0000	.	.	
Spine	1	-0.0428	0.2469	-0.5267 0.4412	0.03	0.8625	
Spine	2	-0.2855	0.3683	-1.0074 0.4364	0.60	0.4383	
Spine	3	0.0000	0.0000	0.0000 0.0000	.	.	
Dispersion	1	1.0363	0.1891	0.7247 1.4819			

**Note:** The negative binomial dispersion parameter was estimated by maximum likelihood.

The dispersion parameter is estimated at 1.0363 and is significantly different from zero (as indicated by the confidence interval). Recall that for the negative binomial distribution, the limiting value of 0 for the dispersion parameter corresponds to a Poisson regression model. This implies that overdispersion is evident if a Poisson model were used. Thus, the standard errors from the negative binomial model are more appropriate than those from the Poisson model.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Width	1	0.00	0.9825
Weight	1	3.02	0.0823
Color	3	2.76	0.4297
Spine	2	0.59	0.7452

With the negative binomial model, none of the effects is significant at an alpha level of 0.05. If you want to reduce your model, you can remove the nonsignificant factors one at a time, starting with the least significant one (**Width**). Reducing your model in this way, the final model has only one significant term, **Weight**.

Refit the model with **Weight** as the only predictor variable. Turn on the influence diagnostics with the **DIAGNOSTICS** option and request all available plots. Request the **HTML** format for the output with tooltips turned on to identify potentially influential or outlying observations. Use the **ANALYSIS** style.

```
ods graphics/ reset=all imagemap=on;
ods html style=analysis;
proc genmod data=STAT2.crab plots(unpack)=all;
  model satellites= weight / dist=negbin link=log type3
    diagnostics;
title2 'Reduced Model'; run;      *ST205d02.sas;
```

Selected PROC GENMOD statement options:

Selected PLOTS= *requests*:

**ALL** produces all available plots.

**STDRESCHI** (*options*)

plots standardized Pearson residuals. The STDRESCHI plot request has the following options:

**INDEX** plots as a function of observation number.

**XBETA** plots as a function of linear predictor.

If you do not specify an option, standardized Pearson residuals are plotted as a function of observation number.

Selected MODEL statement option:

**DIAGNOSTICS|INFLUENCE** requests that case deletion diagnostic statistics be displayed.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
<b>Deviance</b>	171	196.1605	1.1471
<b>Scaled Deviance</b>	171	196.1605	1.1471
<b>Pearson Chi-Square</b>	171	147.9590	0.8653
<b>Scaled Pearson X2</b>	171	147.9590	0.8653
<b>Log Likelihood</b>		155.7126	
<b>Full Log Likelihood</b>		-374.3219	
<b>AIC (smaller is better)</b>		754.6437	
<b>AICC (smaller is better)</b>		754.7857	
<b>BIC (smaller is better)</b>		764.1036	

Algorithm converged.

The AICC (754.79) and BIC (764.10) fit statistics for the reduced model are slightly improved when compared to the full model (AICC=764.43, BIC=791.70).

Analysis Of Maximum Likelihood Parameter Estimates						
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8647	0.4495	-1.7456 0.0163	3.70	0.0544
Weight	1	0.7603	0.1769	0.4136 1.1069	18.48	<.0001
Dispersion	1	1.0740	0.1935	0.7545 1.5288		

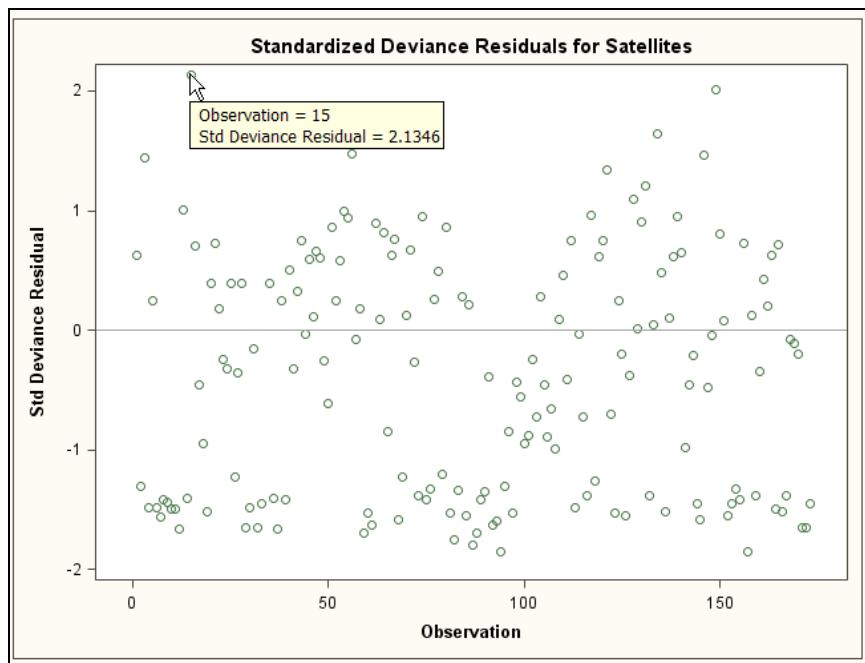
**Note:** The negative binomial dispersion parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Weight	1	18.77	<.0001

When you use the negative binomial distribution to account for overdispersion, the only factor that is significant in predicting the number of satellites is **Weight**. When you compare this to the previous Poisson model that exhibited overdispersion, **Color** was incorrectly identified as a significant factor. Perhaps this indicates that the Poisson model suffered from a Type I error.

The plots created by ODS Graphics with the DIAGNOSTICS option in the MODEL statement include plots of the standardized deviance residuals, Cook's D statistic, and standardized DFBETA plots for the intercept and **Weight**.

#### PROC GENMOD Output

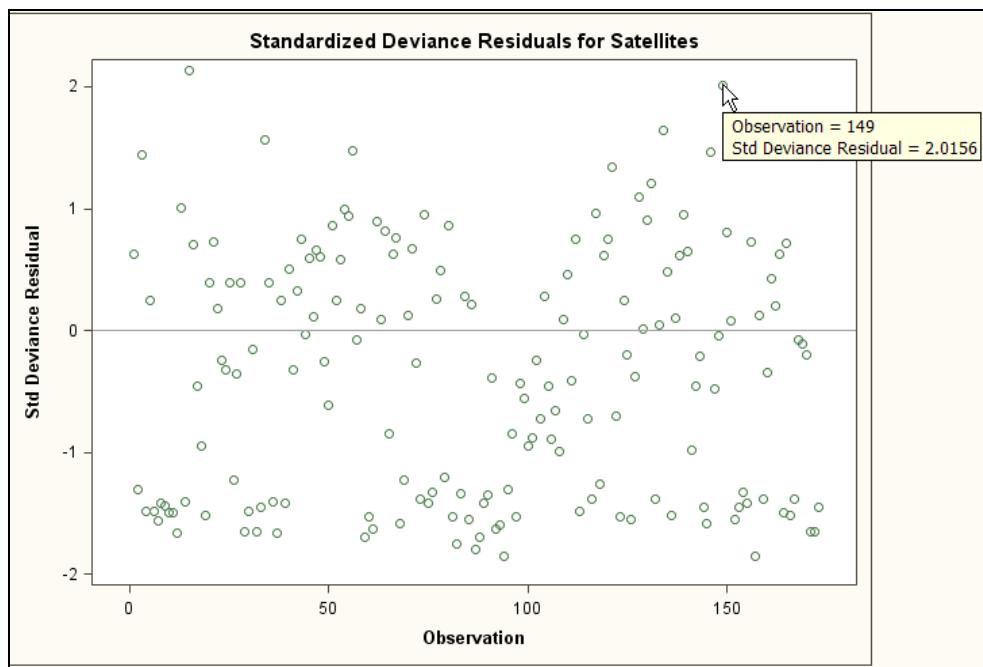


The deviance residuals can take unusual patterns, but unlike standard residuals from a general linear model, they are not used to validate model assumptions. Instead these plots are used to visually inspect for influential observations and outliers, looking for points with large deviance values that are separated from other points (Allison 2010).

Standardized residuals higher than 2 or less than -2 should be examined. The residual for observation #15 falls outside this range. Use the SAS Explorer window to open the **STAT2.crab** data set.

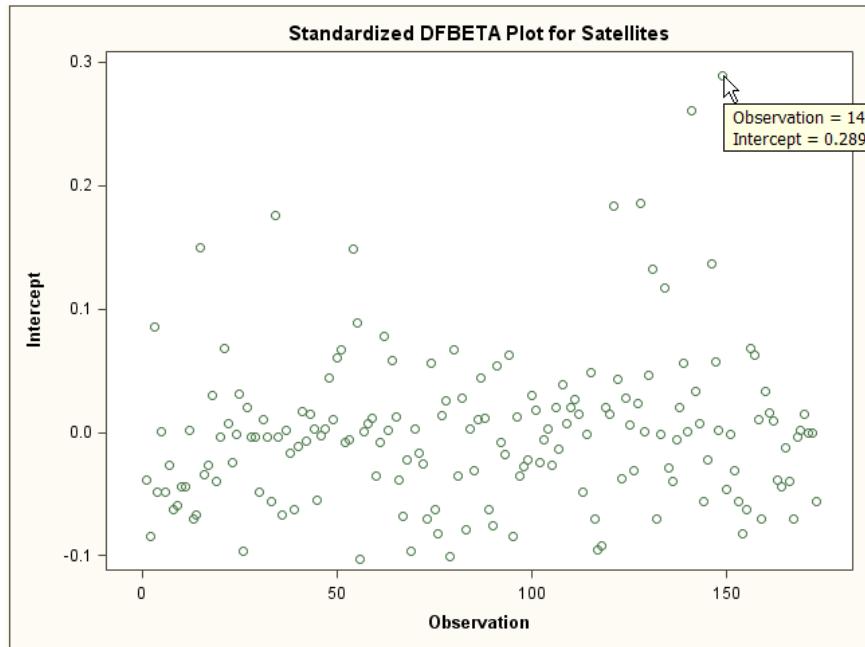
	Color	Spine	Width	Satellites	Weight
13		2	3	28.2	11
14		4	2	21	0
15		2	1	26	14
16		1	1	27.1	8
17		2	3	25.2	1
18		2	3	29	1
19		4	3	24.7	0
20		2	3	27.4	5

Because **Weight** is the only variable in this model, look there for unusual values. This female weighs 2.3 kilograms, close to the average of 2.4 kilograms. What is unusual about this crab is that she has 14 satellites outside her nest. Recall from the data exploration that the mean value of **Satellites** was 2.9. The model does not fit this unusual observation well.

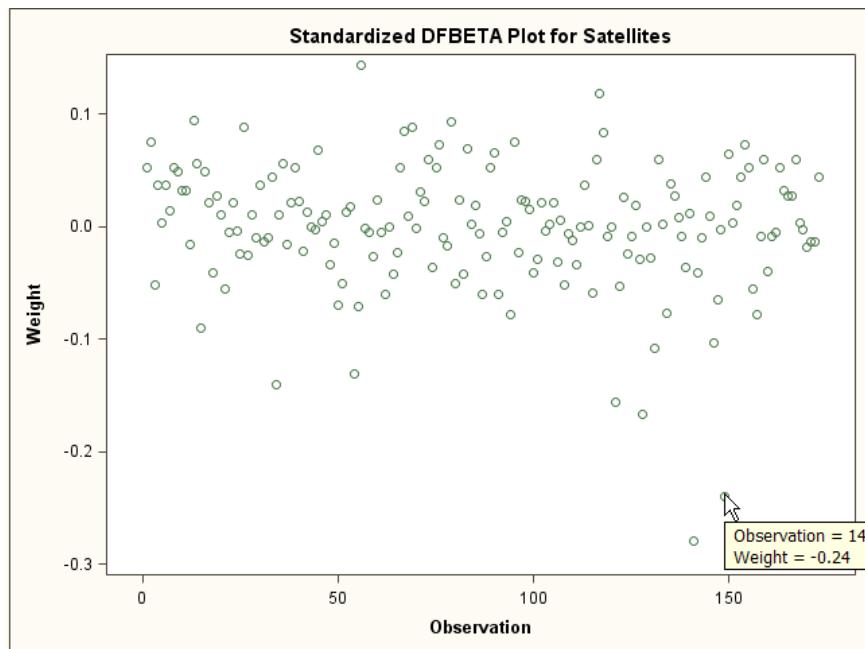


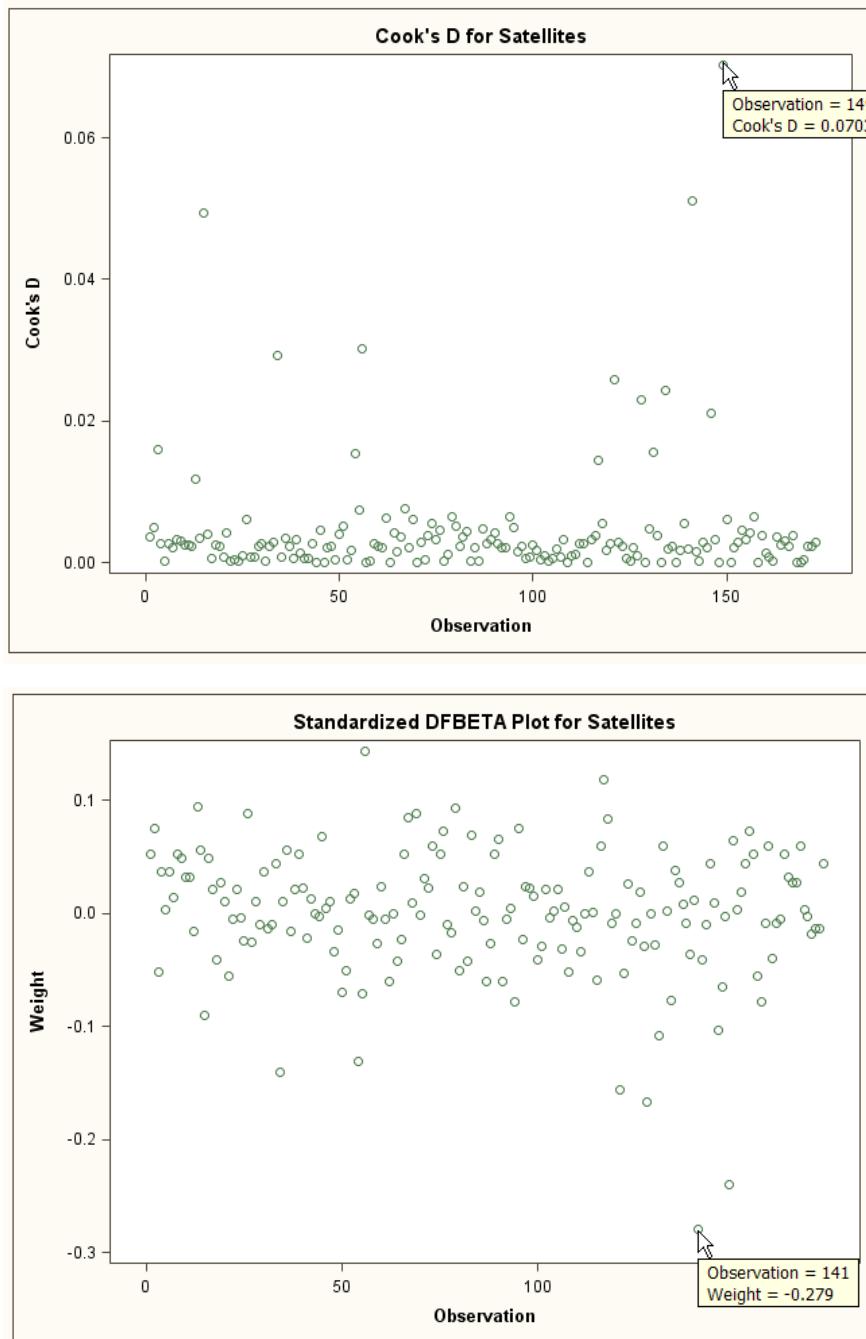
The other data point with a standardized deviance residual more than 2 is observation 149. She weighs 1.9 kilograms, which is below average. She had 10 satellites outside her nest. That is more than the model predicts, given her weight.

	Color	Spine	Width	Satellites	Weight
149	3	3	24	10	1.9

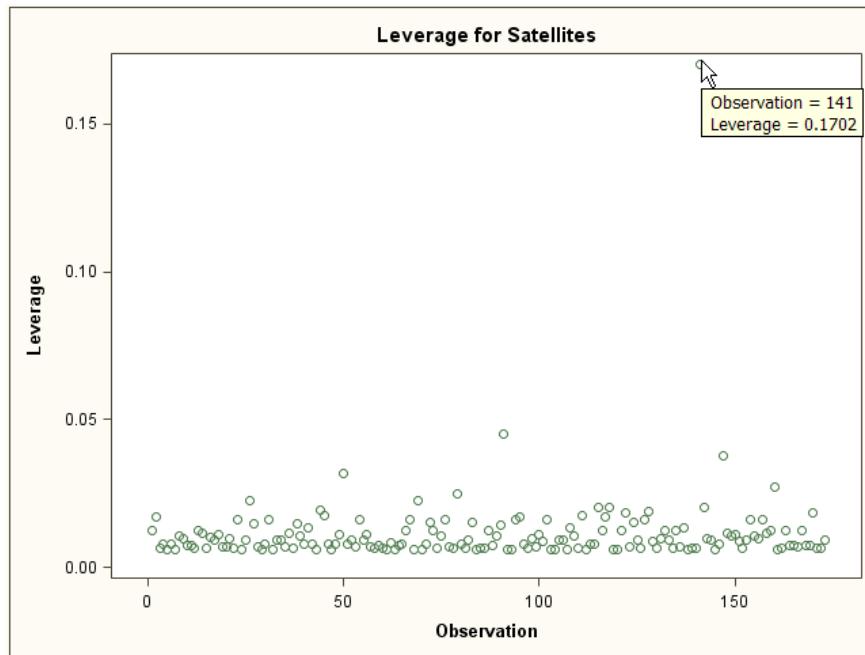


This same observation exhibits a positive influence on the parameter estimate for the intercept and a negative influence on the parameter estimate for **Weight**. This data point is also flagged as influential by the plot of the Cook's D statistic. (Notice that the values shown below the observation numbers in the graphs are the values of the statistic plotted on the vertical axis.)





Observation 141 appears to be exhibiting negative influence on the parameter estimate for **Weight** and seems to be potentially influential in the leverage plot.



This data point represents a very large crab weighing 5.2 kilograms with 7 satellites.

	Color	Spine	Width	Satellites	Weight
141	2	1	33.5	7	5.2
142	2	3	30.5	3	3.325
143	3	3	29	3	2.925
144	2	1	24.3	0	2
145	2	3	25.8	0	2.4
146	4	3	25	8	2.1

The three observations flagged in these plots, plus any others that appear to be influential, should be investigated to make sure that they are not data entry errors.

## 5.06 Multiple Answer Poll

Failing to correct for overdispersion results in which of the following? (Choose all that apply.)

- a. underestimated parameter estimates
- b. overestimated parameter estimates
- c. underestimated standard errors for parameter estimates
- d. overestimated test statistics, and therefore, a too small  $p$ -value



## Exercises

---

### 1. Fitting a Poisson Regression Model Using the GENMOD Procedure

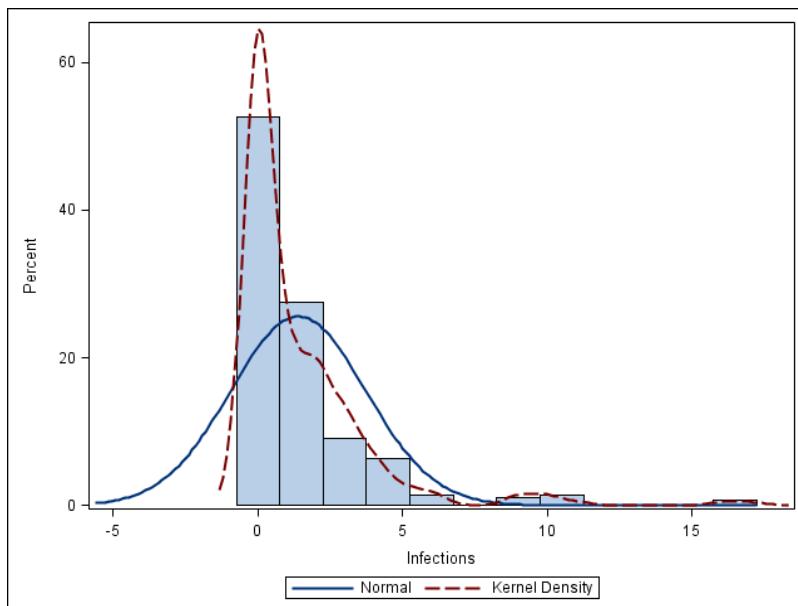
A survey was undertaken to examine which factors are related to ear infections among swimmers. The response variable is the number of self-diagnosed ear infections reported by the participant. The data are stored in the **STAT2.earinfection** data set.

These are the variables in the data set:

<b>Infections</b>	number of self-diagnosed ear infections
<b>Swimmer</b>	swimmer's perception of whether he or she is a frequent ocean swimmer or occasional ocean swimmer ( <i>Frequent</i> =frequent, <i>Occasional</i> =occasional)
<b>Location</b>	swimmer's typical swimming location ( <i>NonBeach</i> =not a beach swimmer, <i>Beach</i> =usually a beach swimmer)
<b>Age</b>	age in years
<b>Gender</b>	gender of swimmer ( <i>Male</i> =male, <i>Female</i> =female).

 The data were obtained with permission from the OZDATA website. This website is a collection of data sets and is maintained in Australia.

- a. First, examine the histogram shown below of the dependent variable, **Infections**. Notice that a normal distribution and a nonparametric curve are superimposed on the histogram to help evaluate whether the normal distribution is a good fit.



- b. Further explore the data by looking at summary statistics from PROC UNIVARIATE. Use an ODS SELECT statement to look at the Moments table and the Goodness of Fit table. How would you describe the distribution of this variable?

- c. Use the GENMOD procedure to fit a Poisson regression model to the data. Which factors seem to be significant? Is overdispersion a problem for this model?
- d. Use the GENMOD procedure to fit a negative binomial regression model to the data. Does this model account for the overdispersion? What factors are now significant?

## Advanced

- e. Look up the PSCALE option in the SAS online documentation. Use the PSCALE option in the MODEL statement to adjust for the possible overdispersion. What factors are now significant? How do you back-transform the model to obtain the model for the average count of ear infections for female occasional beach swimmers?

## 5.07 Quiz

In the Poisson model, the parameter estimate for swimmer **Freq** is **-0.6086**. How do you interpret this value?

Notice that **Exp(-0.6086) = 0.544**.

41

## Poisson Regression for Rates

- When events occur over time, space, or some other index of exposure, it is more relevant to model the rate at which they occur rather than the number of events.
- Rates provide the necessary standardization to make the outcomes comparable.
- You use the OFFSET= option in the MODEL statement in PROC GENMOD.

43

Poisson regression can be used to analyze the rate or incidence of an event. Rates are simply counts divided by some measure of exposure (time, space, population, and so on.). For example, in the female horseshoe crab study, the number of satellites was collected for each nesting female horseshoe crab. Analyzing discrete counts is appropriate. However, if the number of satellites is counted over different sizes of the area, then some type of standardization is needed. Dividing the number of satellites in a certain area by the size of the area enables the outcomes to be comparable.

Another example of standardization is the incidence of certain diseases by county. Because more at-risk subjects result in more occurrences of the disease, you need to adjust for the number of subjects at risk in each county. Simply modeling the number of occurrences of disease by county would be misleading if the number at risk in each county varies.

## Examples of Poisson Regression for Rates

- how crime rates are related to the city's unemployment rate
- how melanoma incidence rates are related to demographic variables
- how the rate of loan defaults is related to region of the country
- how response rates to marketing campaigns relate to known characteristics of the recipients

44

Modeling the rate of an event is common in many fields of study. For example, an epidemiologist might want to examine how the county's cancer rate is related to certain demographic variables and health policy programs in a county. Furthermore, a business analyst might want to examine how the rate of insurance claims is related to demographic variables and the number of years since the last claim on the policy or how the rate of loan defaults is related to a region of the country.

## Poisson Model for Rates

$$\log\left(\frac{\mu}{T}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

$$\log(\mu) = \underbrace{\log(T)}_{\text{OFFSET = Variable}} + \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

$$\mu = T * e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

45

For Poisson regression models for rates, the log of the incidence (where  $T$  is a measure of exposure) is modeled as a linear function of the explanatory variables. Rearranging the terms in the model, it can be shown that the log of the mean can be modeled as a linear function of the explanatory variables and the log of the measure of exposure. The log of the measure of exposure is called the *offset variable*.

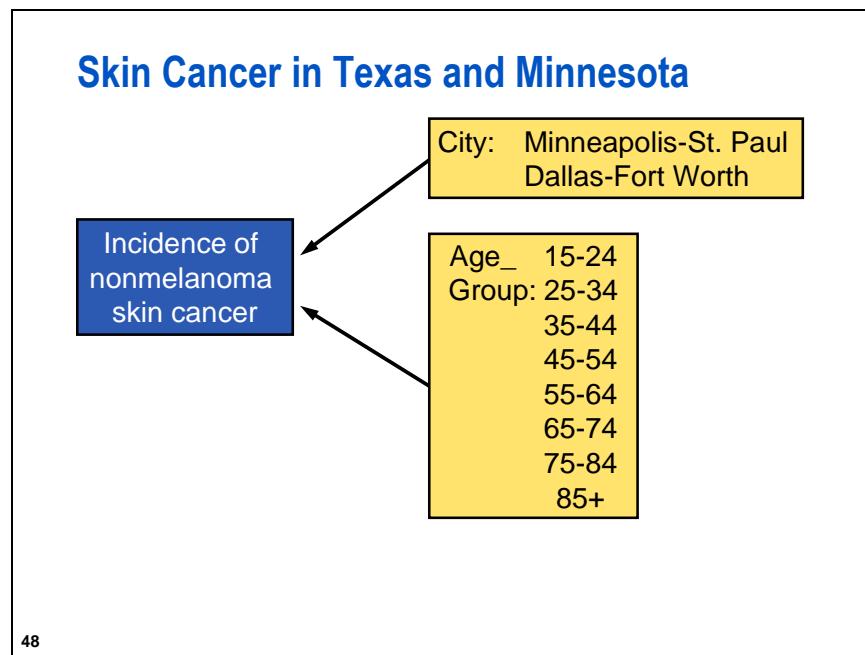
-  If the measure of exposure is the same for every subject, then the offset variable can be absorbed into the intercept.

If you exponentiate both sides of the model expression, you obtain the expected number of events. Notice that the expected number of events is proportional to the index of exposure times the marginal effects of the explanatory variables. The scale of the index of exposure does not affect the parameter estimates for the explanatory variables. For example, expressing the homicide rate by 1,000 people or by 100,000 people does not matter because the differences in exposure scales are reflected in the different values of the intercept.

## 5.08 Multiple Choice Poll

You want to model the rate of car insurance claims by geographic zone. The offset variable is which of the following?

- a. the number of claims
- b. the area of the geographic zone
- c. the number of insured in each geographic zone
- d. the population in the geographic zone



48

A study was conducted to examine the incidence of non-melanoma skin cancer among women in Minneapolis-St. Paul, Minnesota, and Dallas-Fort Worth, Texas. The investigators expect sun exposure to be greater in Texas than in Minnesota. The data are stored in a data set named **STAT2.skin**.

**The Data**

obs	Cases	City	Age	Population	Log_Pop
1	1	MSP	15-24	172675	12.0592
2	16	MSP	25-34	123065	11.7205
3	30	MSP	35-44	96216	11.4744
4	71	MSP	45-54	92051	11.4301
5	102	MSP	55-64	72159	11.1866
6	130	MSP	65-74	54722	10.9100
7	133	MSP	75-84	32185	10.3793
8	40	MSP	85+	8328	9.0274
9	4	DFW	15-24	181343	12.1081
10	38	DFW	25-34	146207	11.8928
11	119	DFW	35-44	121374	11.7066
12	221	DFW	45-54	111353	11.6205
13	259	DFW	55-64	83004	11.3266
14	310	DFW	65-74	55932	10.9319
15	295	DFW	75-84	36518	10.5056
16	65	DFW	85+	7583	8.9337

49

The number of cases is summarized for each combination of **City** by **Age**. These are the variables in the **STAT2.skin** data set:

<b>Cases</b>	number of non-melanoma skin cancer cases
<b>City</b>	city of residence (coded as <i>MSP</i> for Minneapolis-St.. Paul and <i>DFW</i> for Dallas-Fort Worth)
<b>Age</b>	age group
<b>Population</b>	number of people at risk
<b>Log_Pop</b>	log of the population

-  The data were obtained with permission from the website <http://www.statsci.org/data>. This website is a collection of data sets and is maintained in Australia.



## Fitting a Poisson Regression Model for Rate Data

Fit a Poisson regression model for the skin cancer data set **STAT2.skin** in PROC GENMOD.

```
proc genmod data=STAT2.skin;
  class city age;
  model cases= city age / offset=log_pop dist=poi link=log type3;
  title 'Poisson Regression Model for Skin Cancer Rates';
run; *ST205d03.sas;
```

Selected MODEL statement option:

**OFFSET** specifies a variable in the input data set to be used as an offset variable. This variable cannot be a classification variable, and it cannot be the response variable or one of the other explanatory variables.

Partial PROC GENMOD Output

### Poisson Regression Model for Skin Cancer Rates

#### The GENMOD Procedure

Model Information	
Data Set	STAT2.SKIN
Distribution	Poisson
Link Function	Log
Dependent Variable	Cases
Offset Variable	Log_Pop

Number of Observations Read	16
Number of Observations Used	16

Class Level Information		
Class	Levels	Values
City	2	DFW MSP
Age	8	15-24 25-34 35-44 45-54 55-64 65-74 75-84 85+

The Model Information table shows that the offset variable is the log of the population.

- ✍ If you used reference cell coding in the CLASS statement (PARAM=REF), you would see a different presentation of the Class Level Information table. The default coding is GLM coding.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	7	7.4808	1.0687
Scaled Deviance	7	7.4808	1.0687
Pearson Chi-Square	7	7.3536	1.0505
Scaled Pearson X2	7	7.3536	1.0505
Log Likelihood		7585.8374	
Full Log Likelihood		-50.9961	
AIC (smaller is better)		119.9922	
AICC (smaller is better)		149.9922	
BIC (smaller is better)		126.9455	

The Criteria For Assessing Goodness Of Fit table shows that the model fits your data reasonably well because the Value/DF value is close to 1 for the scaled deviance and scaled Pearson chi-square statistics.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
City	1	268.36	<.0001
Age	7	2678.94	<.0001

The LR Statistics For Type 3 Analysis table shows that the variables **City** and **Age** are significantly related to the outcome. In other words, there is a significant difference in skin cancer rates between Minneapolis-St.. Paul and Dallas-Fort Worth, and there is at least one significant difference in skin cancer rates between the age groups.

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.4869	0.1036	-5.6900 -5.2839	2805.29	<.0001
City	DFW	1	0.8091	0.0518	0.7077 0.9106	244.40	<.0001
City	MSP	0	0.0000	0.0000	0.0000 0.0000	.	.
Age	15-24	1	-6.1743	0.4577	-7.0715 -5.2772	181.95	<.0001
Age	25-34	1	-3.5443	0.1675	-3.8725 -3.2160	447.83	<.0001
Age	35-44	1	-2.3272	0.1275	-2.5770 -2.0773	333.33	<.0001
Age	45-54	1	-1.5793	0.1138	-1.8024 -1.3562	192.48	<.0001
Age	55-64	1	-1.0872	0.1109	-1.3045 -0.8698	96.09	<.0001
Age	65-74	1	-0.5289	0.1086	-0.7418 -0.3160	23.71	<.0001
Age	75-84	1	-0.0997	0.1089	-0.3132 0.1138	0.84	0.3602
Age	85+	0	0.0000	0.0000	0.0000 0.0000	.	.
Scale		0	1.0000	0.0000	1.0000 1.0000		

The Analysis Of Maximum Likelihood Parameter Estimates table shows that there is a significant difference in skin cancer rates between Dallas-Fort Worth and Minneapolis-St.. Paul. The positive coefficient for Dallas-Fort Worth indicates that Dallas-Fort Worth has a higher skin cancer rate. There are also significant differences between the age groups 15-24 and 85+, 25-34 and 85+, 35-44 and 85+, 45-54 and 85+, 55-64 and 85+, and 65-74 and 85+. The only group that shows no significant difference in skin cancer rates when compared to the 85+ age group is the age group 75-84.

-  You can write CONTRAST or ESTIMATE statements (or both) with the EXP option to obtain meaningful contrasts of interests. More information can be found in the Fitting Poisson Regression Models Using the GENMOD Procedure Live Web class.

## 5.3 Introduction to Gamma Regression

### Objectives

- Define the gamma distribution.
- Use the GLIMMIX procedure to fit a gamma regression model.

53

### The GLIMMIX Procedure

In addition to PROC GENMOD, you can also use PROC GLIMMIX to fit a generalized linear model for nonnormal responses:

- logistic regression models for binary outcomes
- Poisson regression models for counts of rare events
- gamma regression model for continuous, skewed, positive values

54

In addition to PROC GENMOD, PROC GLIMMIX is also a modeling procedure that is used to fit generalized linear models. In a previous chapter, you used the GLIMMIX procedure to account for nonconstant variances for a normal distribution. Fitting generalized linear models for nonnormal responses, such as logistic regression for binary outcomes, Poisson or negative binomial regression for counts, and gamma regression for continuous skewed, positive values are additional applications for PROC GLIMMIX.

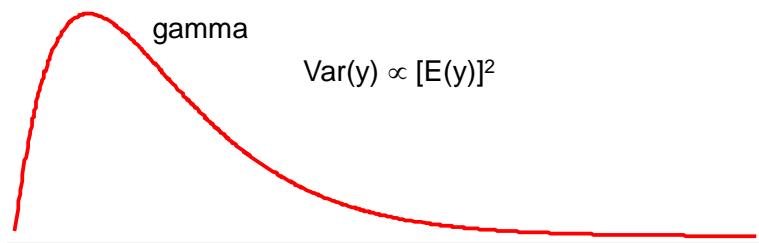


You can also use PROC GLIMMIX to fit mixed models. (An introduction to this topic is available in a later chapter.)

## Gamma Distribution

A gamma distribution

- is a skewed distribution for positive values
- has a variance that is proportional to the squared mean
- has lighter tails than a lognormal distribution.



55

Sometimes a normal distribution curve is not a good description of the data because they are not symmetrically distributed with respect to their values. In such cases, nonsymmetrical distribution functions can be used to describe the data, such as the gamma distribution.

### Details

The probability density function for a gamma distribution is the following:

$$f(y) = \frac{\left(\frac{y-\mu}{\beta}\right)^{\alpha-1} \exp\left(-\frac{y-\mu}{\beta}\right)}{\beta \Gamma(\alpha)}, \quad y \geq \mu; \quad \alpha, \beta > 0$$

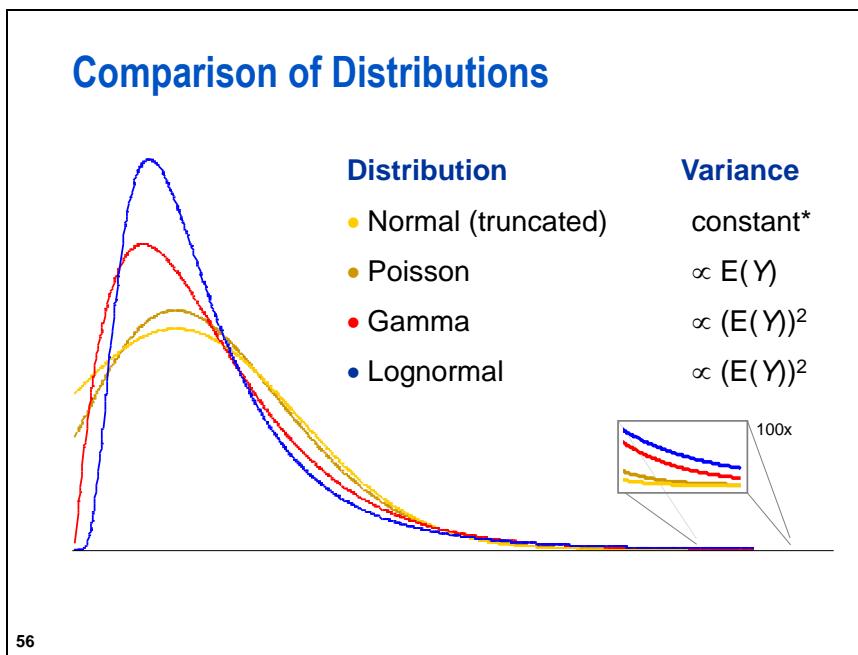
where  $\alpha$  is the shape parameter,  $\mu$  is the location parameter,  $\beta$  is the scale parameter, and  $\Gamma$  is the gamma function:  $\Gamma(\gamma) = \int_0^\infty t^{\gamma-1} e^{-t} dt$ .

The case where  $\mu=0$  and  $\beta=1$  is called standardized gamma distribution. The equation for the standard gamma distribution reduces to  $f(y) = \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}$ .

It can be shown that for the gamma distribution, the variance is proportional to the square of the mean. This information can be used to model positive continuous variables that exhibit this relationship between the variances and the means.



The gamma distribution is used to model variables with positive values. Zero and negative values are not allowed in the gamma distribution. In fact, PROC GLIMMIX excludes zeros and negative values for gamma distribution models.



For many monetary-related models, the residual variance increases with the predicted value. Some of the commonly used distributions for this type of data are Poisson, gamma, and lognormal distributions.

As mentioned earlier, Poisson regression is useful for modeling the count or rate of rare events. When the expected value increases, the Poisson distribution approaches the normal distribution.

Both the gamma and lognormal distributions are appropriate for continuous positive values whose variance increases in proportion to the square of the mean. Unlike the Poisson distribution, their skewness is independent of the expected value of the dependent variable. Therefore, for skewed distributions with relatively large means, gamma or lognormal distribution might be better choices.

The gamma regression model can be fit using the GLIMMIX procedure. The lognormal regression model can also be fit using PROC GLIMMIX.

One additional consideration for the distribution is the tail behavior. Although all of the distributions in the plot have the same expected value and variance, they have increasingly heavy tails. A few extreme outliers might indicate a lognormal distribution, whereas the absence of such might imply a gamma or less extreme distribution.

## GLIMMIX Procedure

General form of the GLIMMIX procedure:

```
PROC GLIMMIX <options>;
  CLASS variables;
  MODEL response<(response options)>=<fixed-effects>
    < DIST=keyword LINK=keyword options>;
RUN;
```

57

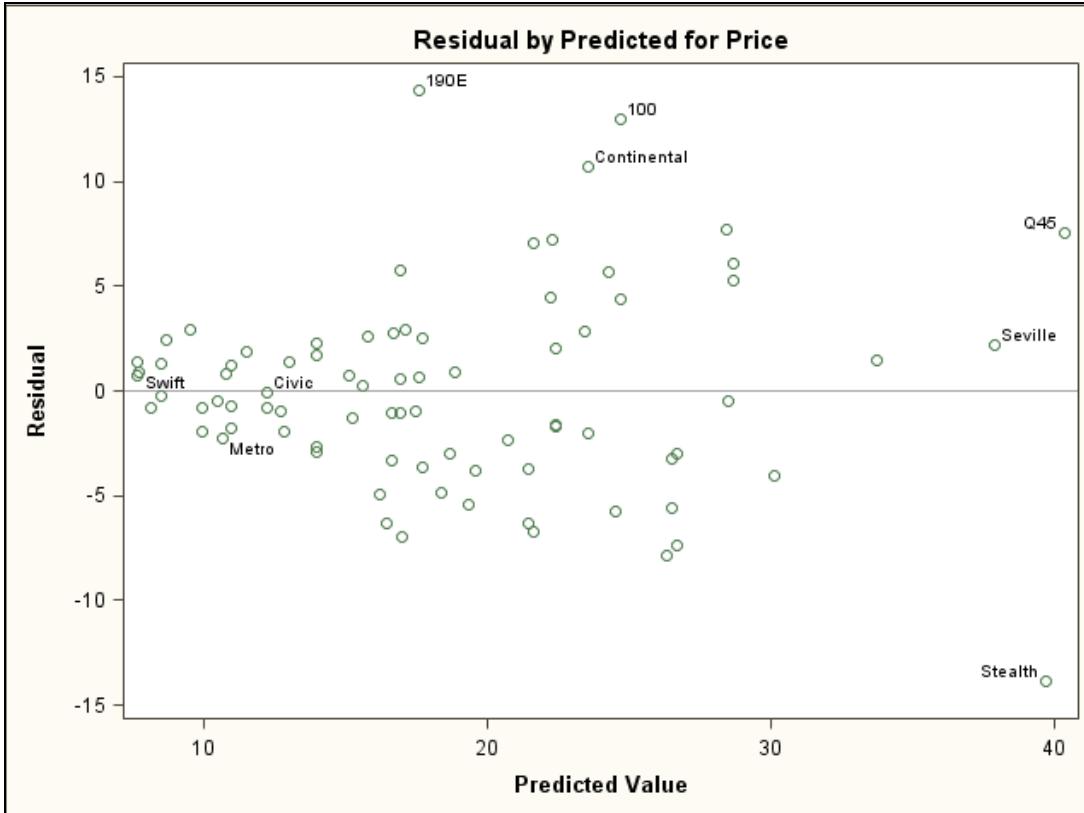
The DIST=*keyword* and LINK=*keyword* options in the MODEL statement identify the distribution that you want to model and the link that you want to use, respectively. For gamma regression, use the GAMMA distribution and log link keywords.



## Fitting a Gamma Regression Model

Recall the **STAT2.cars2** data set that you analyzed previously. You suspected that the variance might not be constant. Nonconstant variance violates one of the linear regression assumptions.

Previously Shown PROC REG Output



First, use the UNIVARIATE procedure to examine whether a gamma distribution describes the distribution of **Price** adequately.

```
ods select 'Panel 1' 'Parameter Estimates' 'Goodness of Fit';
proc univariate data=STAT2.cars2;
  var price;
  histogram /gamma(alpha=est sigma=est theta=est color=blue w=2)
              vaxis=0 to 14 by 2 midpoints=8 to 50 by 2;
  title 'Testing Gamma Distributions';
run; *ST205d04.sas;
```

Selected HISTOGRAM statement option:

**GAMMA(*gamma-options*)** fits gamma distribution with shape parameter  $\alpha$ , scale parameter  $\sigma$ , and threshold parameter  $\theta$ .

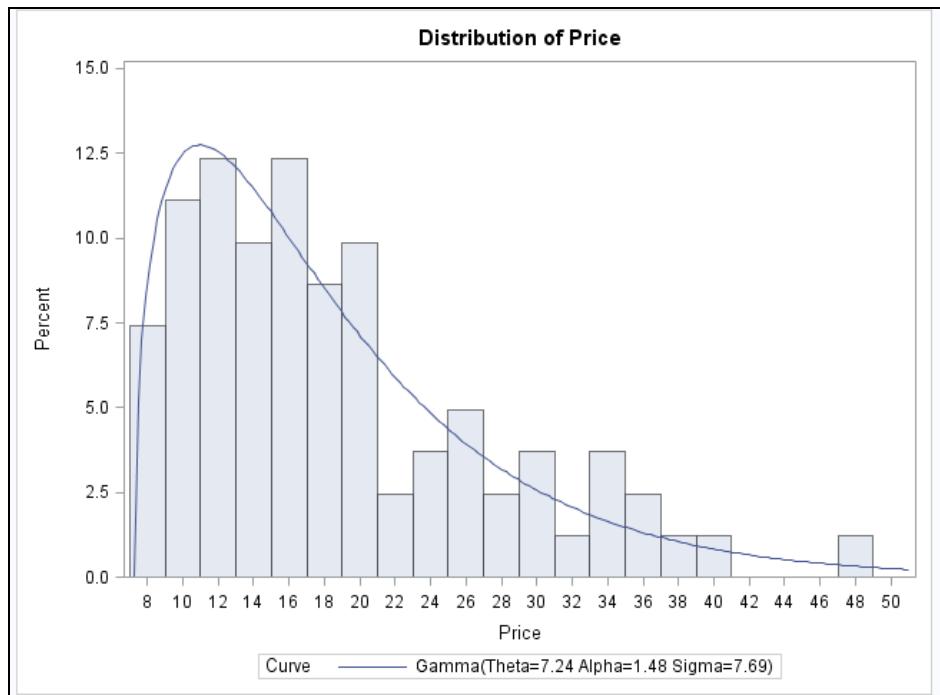
## Partial PROC UNIVARIATE Output

The UNIVARIATE Procedure Fitted Gamma Distribution for Price		
Parameters for Gamma Distribution		
Parameter	Symbol	Estimate
Threshold	Theta	7.243819
Scale	Sigma	7.686662
Shape	Alpha	1.483009
Mean		18.64321
Std Dev		9.36073

Goodness-of-Fit Tests for Gamma Distribution				
Test	Statistic	p Value		
Kolmogorov-Smirnov D	0.04958256	Pr > D	>0.500	
Cramer-von Mises W-Sq	0.02321245	Pr > W-Sq	>0.500	
Anderson-Darling A-Sq	0.17088853	Pr > A-Sq	>0.500	

The nonsignificant  $p$ -values for the tests for the gamma distribution indicate that you do not have enough evidence to reject the null hypothesis that the variable **Price** follows a gamma distribution.



The histogram looks consistent with the test.

Use the GLIMMIX procedure to fit a linear model to the **STAT2.cars2** data set that you created previously. Although the canonical link function for the gamma distribution is the inverse function, use the log link function because it is the most commonly used link function for gamma-distributed variables.

To identify the outlying observations, use the OUTPUT statement to create a data set containing the standardized residuals and predicted values. Add the ID option in the MODEL statement to include the desired variables in the output data set. Print observations with standardized residuals higher than 2 or less than -2.

```
proc glimmix data=STAT2.cars2 plots=studentpanel(unpack);
  model price=hwympg hwympg2 horsepower / dist=gamma link=log solution;
  id model hwympg hwympg2 horsepower price;
  output out=check1 student
    pred(ilink)= Pred stderr(ilink)=Stderr lcl(ilink)=LCL
    ucl(ilink)=UCL
    pred(noilink)= XB stderr(noilink)=StderrXB lcl(noilink)=LCLXB
    ucl(noilink)=UCLXB;
  title 'Cars Data Set - Gamma Distribution with Log Link';
run;
```

Selected MODEL statement options:

**DIST=***keyword*

specifies the built-in (conditional) probability distribution of the data. If you specify the DIST= option and you do not specify a user-defined link function, a default link function is chosen according to the following table:

Distribution	Default Link Function
beta	logit
binary	logit
binomial	logit
exponential	log
gamma	log
normal	identity
geometric	log
inverse Gaussian	inverse squared (power(-2))
lognormal	identity
multinomial	cumulative logit
negative binomial	log
Poisson	log
<i>t</i>	identity
multivariate	varied

If you do not specify a distribution, the GLIMMIX procedure defaults to the normal distribution for continuous response variables and to the multinomial distribution for classification or character variables, unless the events or trial syntax is used in the MODEL statement. If you choose the events or trial syntax, the GLIMMIX procedure defaults to the binomial distribution.

**LINK=***keyword*

specifies the link function in the generalized linear mixed model.

Selected OUTPUT statement options:

- keyword* (ILINK) requests that the statistics be on the original scale of the data.
- keyword*(NILINK) requests that the statistics be on the linked scale.

## Partial PROC GLIMMIX Output

The GLIMMIX Procedure	
<b>Model Information</b>	
Data Set	STAT2.CARS2
Response Variable	Price
Response Distribution	Gamma
Link Function	Identity
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Maximum Likelihood
Degrees of Freedom Method	Residual
Number of Observations Read	81
Number of Observations Used	81
<b>Dimensions</b>	
Covariance Parameters	1
Columns in X	4
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	81
<b>Fit Statistics</b>	
-2 Log Likelihood	447.25
AIC (smaller is better)	457.25
AICC (smaller is better)	458.05
BIC (smaller is better)	469.22
CAIC (smaller is better)	474.22
HQIC (smaller is better)	462.05
Pearson Chi-Square	4.37
Pearson Chi-Square / DF	0.06

The AIC statistic for this model is 457.25 and the BIC is 469.22.

Parameter Estimates						
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	
Intercept	2.1190	0.1088	77	19.48	<.0001	
Hwympg	-0.04333	0.01021	77	-4.24	<.0001	
Hwympg2	0.001598	0.000681	77	2.35	0.0215	
Horsepower	0.004970	0.000807	77	6.16	<.0001	
Scale	0.05074	0.007907	.	.	.	.

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Hwympg	1	77	18.01	<.0001
Hwympg2	1	77	5.51	0.0215
Horsepower	1	77	37.90	<.0001

All the parameter estimates are significant at the alpha level of 0.05. The model is as follows:

$$\log(E(\text{Price})) = 2.1190 - 0.0433*\text{Hwympg} + 0.0016*\text{Hwympg2} + 0.0050*\text{Horsepower}$$

or

$$E(\text{Price}) = e^{2.1190 - 0.0433*\text{Hwympg} + 0.0016*\text{Hwympg2} + 0.0050*\text{Horsepower}}$$

```
proc print data=check1 (obs=5);
  var Model Hwympg Hwympg2 Horsepower Price pred stderr lcl ucl xb
       stderrxb lclxb uclxb student;
  title2 'Predicted Values';
run; *ST205d04.sas;
```

Obs	Model	Hwympg	Hwympg2	Horsepower	Price	Pred	Stderr	LCL	UCL
1	Integra	0.96296	0.9273	140	15.9	16.0315	0.56619	14.9428	17.1995
2	Legend	-5.03704	25.3717	200	33.9	29.1315	1.28877	26.6751	31.8142
3	100	-4.03704	16.2977	172	37.7	23.9226	0.83774	22.3112	25.6503
4	90	-4.03704	16.2977	172	29.1	23.9226	0.83774	22.3112	25.6503
5	535i	-0.03704	0.0014	208	30.0	23.4386	1.62507	20.4161	26.9085

Obs	Model	XB	StderrXB	LCLXB	UCLXB	Pearson	Student
1	Integra	2.77456	0.035317	2.70423	2.84488	-0.03641	-0.03687
2	Legend	3.37182	0.044240	3.28373	3.45991	0.72665	0.74108
3	100	3.17482	0.035019	3.10509	3.24455	2.55665	2.58812
4	90	3.17482	0.035019	3.10509	3.24455	0.96077	0.97259
5	535i	3.15438	0.069333	3.01632	3.29244	1.24273	1.30614

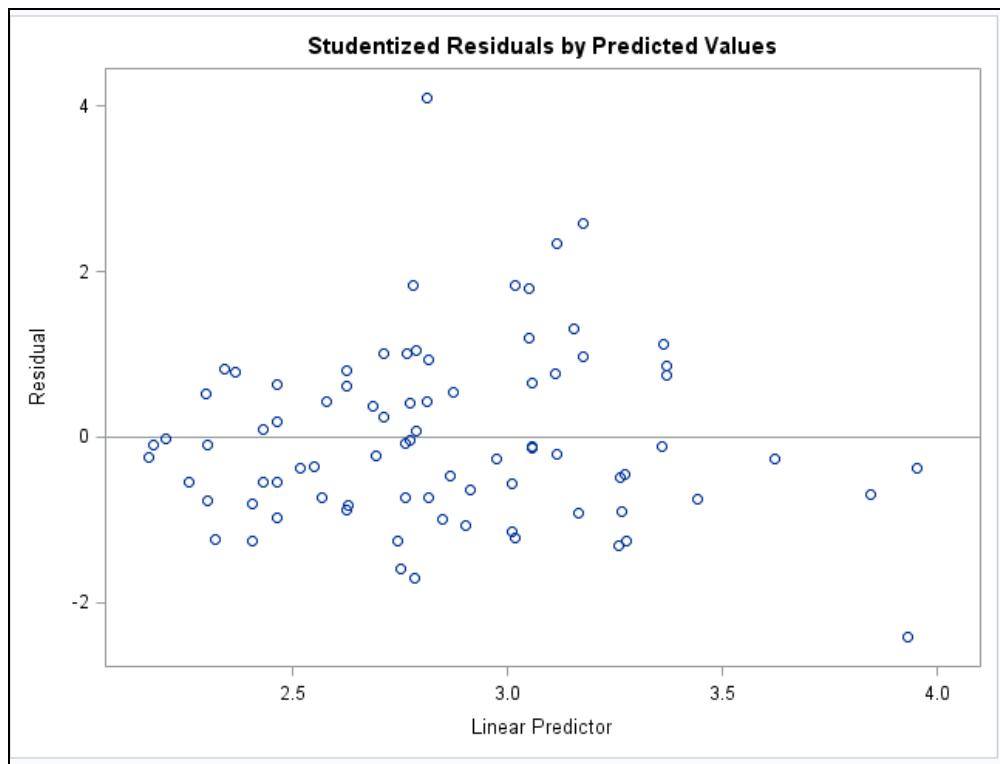
The output data set contains these variables:

- **Obs:** the observation number.
- **Model:** the model of the car.
- **Price:** the value of the response variable.
- **Hwympg, Hwympg2, Horsepower:** the values of the predictor variables.
- **Pred:** predicted mean,  $\hat{\mu} = g^{-1}(\eta)$ , where  $\eta = \mathbf{x}_i^T \hat{\beta}$  is the linear predictor and  $g$  is the link function.  
If there is an offset, it is included in  $\mathbf{x}_i^T \hat{\beta}$ .
- **Stderr:** standard error of the mean. The GLIMMIX procedure obtains standard errors on the scale of the mean by the delta method. If the link is a nonlinear function of the linear predictor, these standard errors are only approximate. For example:

$$\text{Var}[g^{-1}(\hat{\eta}_m)] \approx \left( \frac{\partial g^{-1}(t)}{\partial(t)}|_{\hat{\eta}_m} \right)^2 \text{Var}[\hat{\eta}_m]$$

- **LCL:** lower confidence limit of the predicted value of the mean. Confidence limits on the scale of the data are usually computed by applying the inverse link function to the confidence limits on the linked scale. The resulting limits on the data scale have the same coverage probability as the limits on the linked scale, but they are possibly asymmetric. The confidence coefficient is specified with the ALPHA= option.
- **UCL:** upper confidence limit of the predicted value of the mean.
- **XB:** estimate of the linear predictor  $\mathbf{x}_i^T \hat{\beta}$ . If there is an offset, it is included in  $\mathbf{x}_i^T \hat{\beta}$ .
- **StderrXB:** standard error of the linear predictor  $\mathbf{x}_i^T \hat{\beta}$ .
- **LCLXB:** lower confidence limit of the linear predictor.
- **UCLXB:** upper confidence limit of the linear predictor.
- **Student:** studentized residuals.

## PROC GLIMMIX Graphics Output



The Studentized residuals are plotted against the linear predictor. Several possible outliers might need further investigation, and variance of the residuals does not appear to be constant. In addition, relatively large predictions seem to have negative residuals, indicating that the predicted values are too large for high-priced cars. Based only on the values of **Hwympg** and **Horsepower**, your model is likely to overestimate the prices for these cars.

```
proc print data=check1 (obs=5);
  where student ge 2 | student le -2;
  var Model Hwympg Hwympg2 Horsepower Price pred stderr lcl ucl xb
    stderrxb lclxb uclxb student;
  title2 'Outlying Student Residuals';
run; *ST205d04.sas;
```

### Cars Data Set - Gamma Distribution with Log Link Outlying Student Residuals

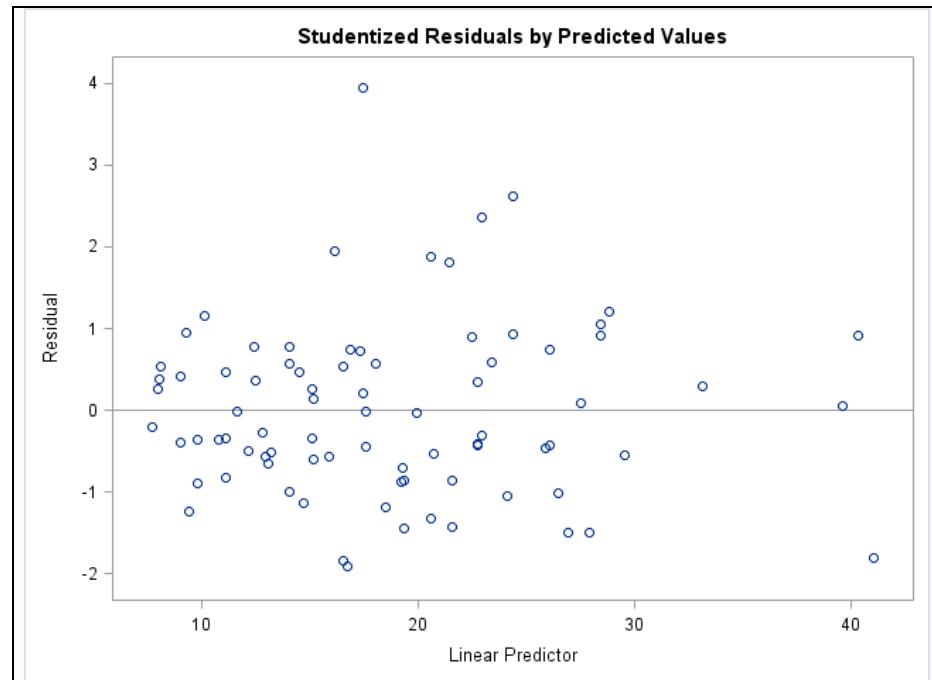
Obs	Model	Student	Price	Pred	Hwympg	Horsepower
3	100	2.58812	37.7	23.9226	-4.03704	172
24	Stealth	-2.42299	25.8	50.9005	-6.03704	300
46	Continental	2.34825	34.3	22.5375	-4.03704	160
51	190E	4.10448	31.9	16.6391	-1.03704	130

The car with the largest residual is the Mercedes-Benz 190E with a standardized residual of 4.104. The other cars with residuals greater than 2 or less than -2 are the Audi 100, Dodge Stealth, and Lincoln Continental. The model does not fit these four luxury cars well.

If you believe that the expected value of **Price** is linearly related to the predictor variables, but the variance of the residuals is not constant (for example, if you think the variance of the residuals is proportional to the mean squared), then you can use the gamma distribution with the identity link to fit a model to your data.

```
proc glimmix data=STAT2.cars2 plots=studentpanel (unpack);
  model price=hwympg hwympg2 horsepower / dist=gamma link=id solution;
  id model hwympg hwympg2 horsepower price;
  output out=check2 student=Student pred(ilink)=Pred;
  title 'Cars Data Set - Gamma Distribution with Identity Link';
run;

proc print data=check2;
  where student ge 2 | student le -2;
  var model student price pred hwympg horsepower;
title2 'Outlying Student Residuals';
run;
title1;
title2; *ST205d04.sas;
```



The residuals seem to be a random scatter about the zero reference line for this model and the variance seems to be more stable. This model has both positive and negative residuals in the high price ranges, whereas the gamma model with the log link had only negative residuals in the highest price ranges. In addition, this model has only three potentially outlying residuals, so it seems that this gamma model with the identity link might fit the data better than the gamma model with the log link.

**Cars Data Set - Gamma Distribution with Identity Link  
Outlying Student Residuals**

Obs	Model	Student	Price	Pred	Hwympg	Horsepower
3	100	2.61262	37.7	24.3606	-4.03704	172
46	Continental	2.36042	34.3	22.9487	-4.03704	160
51	190E	3.95081	31.9	17.4204	-1.03704	130

Partial PROC GLIMMIX Output

Fit Statistics	
<b>-2 Log Likelihood</b>	437.19
<b>AIC (smaller is better)</b>	447.19
<b>AICC (smaller is better)</b>	447.99
<b>BIC (smaller is better)</b>	459.17
<b>CAIC (smaller is better)</b>	464.17
<b>HQIC (smaller is better)</b>	452.00
<b>Pearson Chi-Square</b>	3.83
<b>Pearson Chi-Square / DF</b>	0.05

Another indication that the model seems to fit the data better than the previous model is that the AIC and BIC statistics are smaller for this model (447.19 versus 457.25 for AIC and 459.17 versus 469.24 for BIC).

Parameter Estimates						
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	
<b>Intercept</b>	1.5568	1.8223	77	0.85	0.3956	
<b>Hwympg</b>	-0.5171	0.1560	77	-3.32	0.0014	
<b>Hwympg2</b>	0.02937	0.009106	77	3.23	0.0018	
<b>Horsepower</b>	0.1177	0.01483	77	7.93	<.0001	
<b>Scale</b>	0.04491	0.007004	.	.	.	

The estimated equation is  $E(\text{Price}) = 1.5568 - 0.5171 * \text{Hwympg} + 0.0294 * \text{Hwympg2} + 0.1177 * \text{Horsepower}$ . Compared with the OLS regression model that you fit in the previous chapter,  $E(\text{Price}) = 4.0395 - 0.8041 * \text{Hwympg} + 0.0435 * \text{Hwympg2} + 0.0973 * \text{Horsepower}$ , the gamma regression model with an identity link provides different parameter estimates. However, this model accounts for heteroscedasticity.

- ✍ You should complete the entire modeling cycle to choose the best candidate model for the gamma regression with either the log or identity link.

### 5.09 Poll

A gamma regression model can be useful for positive values with large means that are skewed to the right.

- True
- False

59

### Summary of Approaches

- Problem:
    - nonconstant variance
  - Approaches:
    - ▶ Fit a gamma regression model with the log link function.
    - ▶ Fit a gamma regression model with the identity link function.
- 
- PROBLEM for OLS**

61

## Comparison of Results

Model	MSE	R Square	Adjusted R Square
OLS regression model	22.56	0.719	0.708
gamma regression with log link	28.88	0.641	0.627
gamma regression with identity link	23.14	0.712	0.701

62

The ordinary least squares regression model using the original dependent variable **Price** seems to fit your data well. However, the constant variance assumption is violated. Therefore, the standard errors for the parameter estimates are compromised. The other two models attempted to deal with the heteroscedasticity presented in the **STAT2.cars2** data set. The goodness of the model fit can be evaluated by the mean squared error, R square, or adjusted R square. These statistics should be computed on the original scale rather than the transformed scale to make them comparable. The general formulas for MSE, R square, and adjusted R square are the following:

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p}$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$adj. R^2 = 1 - \frac{SSE / df_E}{SST / df_T} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)}$$

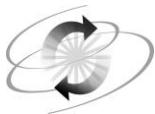
You can write a SAS program to compute these statistics for each model that you fit. The results are summarized above. The sample SAS program is provided in the appendix.

The gamma regression model is useful when the variance changes in proportion to the square of the mean. If the identity link function is used, then the parameter estimates are on the original scale and no back transformation is necessary. If other link functions are used (such as the log link), then the parameter estimates are on the linked scale. Applying the inverse link function provides the unbiased estimate of the mean on the original scale, but the parameter estimates are still on the transformed scale.

## 5.10 Poll

If you use PROC GLIMMIX with a log-link function, the predicted value is on the log scale. You need to request the predicted values with the ILINK option to obtain unbiased estimates of the means on the original scale.

- True
- False



## Exercises

---

### 2. Conducting a Linear Regression Analysis

A soft drink bottler is analyzing vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. The service activities include stocking the machine with beverage products and minor maintenance as required. The industrial engineer responsible for the study suggested that the two most important variables affecting the delivery time are the number of cases of product stocked and the distance walked by the route driver. The engineer collected 24 observations on delivery time (minutes), number of cases, and distance walked (feet). The data are found in the **STAT2.softdrinks** data set (based on data from Montgomery and Peck 1992).

These are the variables in the data set:

<b>Time</b>	delivery time in minutes
<b>Cases</b>	number of cases delivered
<b>Distance</b>	distance walked in feet
<b>LogTime</b>	logarithm of the delivery time

- Print the data and then create a histogram of the data on the original scale by using **Time** in the HISTOGRAM statement in PROC SGPlot. Overlay the kernel density and a normal density. Do the times seem to be normally distributed?
- Create a histogram of the data on the log-transformed scale by using **LogTime** in the HISTOGRAM statement in PROC SGPlot. Overlay the kernel density and a normal density. What do you conclude about the distribution of the log-transformed times?
- Use PROC GLIMMIX to model **Time** as a function of **Cases**, **Cases** squared, **Distance**, and **Distance** squared. In the MODEL statement, use the options DIST=GAMMA and LINK=LOG.
- Eliminate any terms that are not significant to obtain a final “candidate” model.
  - Use the PLOTS=STUDENTPANEL (UNPACK) option in the PROC GLIMMIX statement. This creates a plot of the studentized residuals versus the linear predictor.
  - Use the OUTPUT OUT=**gamma\_predicted** statement to save the predicted values in the data set **gamma\_predicted**.
  - Use the HTML output destination. Enable IMAGEMAP using the ODS GRAPHICS statement if you want to be able to identify particular observations in any of these graphs.
  - After looking at the plots of the studentized residuals, do you think the model fits well?
- Use the **gamma\_predicted** data set and PROC SGPlot with the REG statement to plot the predicted times against the observed times. Use the DATALABEL=**Distance** option in the REG statement.

## 5.4 Chapter Summary

---

Generalized linear models extend general linear models in three ways.

- The distribution of the random component can come from the family of exponential distributions.
- The variance of the response value is a specified function of its mean.
- The link function does not have to be the identity function.

The GENMOD and GLIMMIX procedures can be used to fit generalized linear models, including general linear models, logistic regression models, and Poisson regression models. They enable the user to specify the assumed probability distribution of the response variable. Based on the chosen distribution, a default link function is assigned. However, you can specify that a different link function be used.

One type of generalized linear model is a Poisson regression. The underlying assumption of Poisson regression is that the response variable,  $\mathbf{Y}$ , is distributed as a Poisson random variable. As the mean of the distribution becomes larger, the distribution can be better approximated by a normal distribution.

The Poisson distribution is very useful in modeling rare events. Any situation that involves a count of events in which large counts would be rare is a candidate for this type of analysis. In general, the distribution of these counts is skewed to the right, usually with a high number of zero occurrences.

One potential problem when you fit a generalized linear model is overdispersion. Overdispersion can be caused by misspecifying the probability distribution of the response variable or by heterogeneity among the sampled units. One solution for this problem is to use the PSCALE option in PROC GENMOD to adjust the standard errors and likelihood ratio statistics and other related statistics based on the estimated scale parameter. Another solution is to model the counts using the negative binomial distribution; it has an extra parameter in its variance function that accounts for overdispersion.

Poisson regression can be used to analyze the rate or incidence of an event. Rates are simply counts divided by some measure of exposure (time, space, population, and so on.). The log of the measure of exposure is called the offset variable and is modeled using the OFFSET= option in the MODEL statement in PROC GENMOD. The resulting model is the log of expected counts per exposure.

The GENMOD or GLIMMIX procedure can also be used to fit a gamma regression model. If the variable has continuous positive values, is highly skewed to the right, and has variances that are proportional to the squared mean, a gamma regression model might be appropriate. In gamma regression models, the log link function is commonly used rather than the canonical link function, which is the inverse function.

Sometimes an identity link function can fit your data well if the dependent variable itself is linearly related to the predictors. The gamma regression model might be an alternative way of addressing nonconstant variances as opposed to the common approach of transforming the dependent variable. If the variance and the mean exhibit a different relationship, then a different distribution can be chosen to model this relationship.

General form of the GENMOD and GLIMMIX procedures for modeling generalized linear models:

```
PROC GENMOD options PLOTS=requests;  
  CLASS variables;  
  MODEL response=effects / options;  
  ESTIMATE 'label' effect values / options;  
RUN;
```

```
PROC GLIMMIX <options> PLOTS=requests;  
  CLASS variables;  
  MODEL response <(response options)> =  
        <fixed-effects>  
        <DIST=keyword LINK=keyword options>;  
RUN;
```

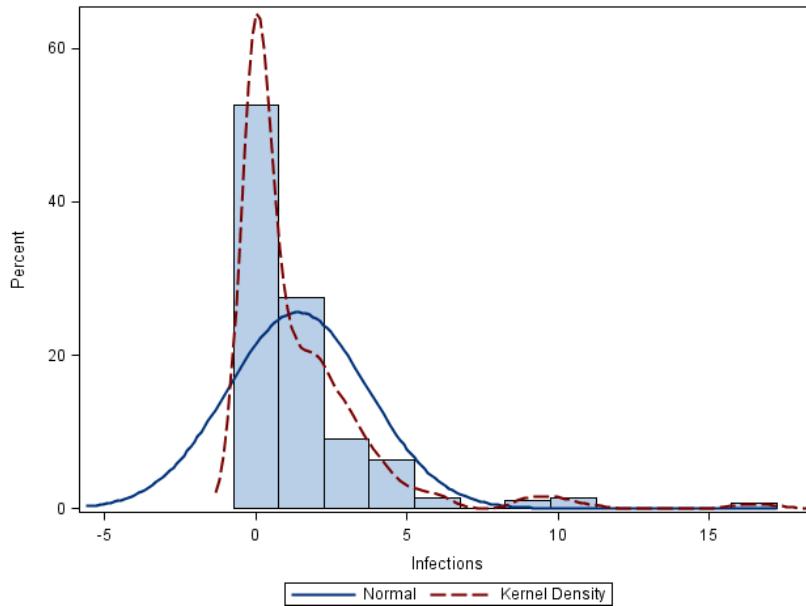
## 5.5 Solutions

### Solutions to Exercises

#### 1. Fitting a Poisson Regression Model Using the GENMOD Procedure

- a. First, examine the histogram shown below of the dependent variable, **Infections**; note that a normal distribution and a nonparametric curve have been superimposed on the histogram to help evaluate whether the normal distribution is a good fit. (The code to produce the histogram is shown below.)

```
title;
proc sgplot data=STAT2.earinfection;
  histogram infections;
  density infections;
  density infections / type=kernel;
run;                                              ST205s01.sas;
```



The histogram with the normal and kernel densities superimposed indicate that the normal distribution is not a good fit for the data.

- b. Further explore the data by looking at summary statistics from PROC UNIVARIATE. Use an ODS SELECT statement to look at the Moments table and the Goodness of Fit table. What do you think about the distribution of this variable?

```
ods select moments basicmeasures goodnessoffit;
proc univariate data=STAT2.earinfection normal;
  var infections;
run;                                              *ST205s01.sas;
```

Moments			
N	287	Sum Weights	287
Mean	1.38675958	Sum Observations	398
Std Deviation	2.33854124	Variance	5.46877513
Skewness	3.20185866	Kurtosis	14.180533
Uncorrected SS	2116	Corrected SS	1564.06969
Coeff Variation	168.633501	Std Error Mean	0.13803972

Basic Statistical Measures			
Location		Variability	
Mean	1.386760	Std Deviation	2.33854
Median	0.000000	Variance	5.46878
Mode	0.000000	Range	17.00000
		Interquartile Range	2.00000

The mean number of ear infections is about 1.39 and is low enough that the Poisson distribution would be appropriate for the data. However, the variance is about 5.47, much higher than the mean, so overdispersion might be a problem for the data.

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.2765899	Pr > D	<0.010
Cramer-von Mises	W-Sq	5.6097318	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	30.8888909	Pr > A-Sq	<0.005

The tests for normality have low *p*-values. This indicates that there is enough evidence to reject the assumption of normality for this data set.

- c. Use the GENMOD procedure to fit a Poisson regression model to the data. Which factors seem to be significant? Is overdispersion a problem for this model?

```
proc genmod data=STAT2.earinfection;
  class swimmer location gender;
  model infections=swimmer location age gender
    / dist=poisson link=log type3;
run; *ST205s01.sas;
```

Model Information	
Data Set	STAT2.EARINFECTION
Distribution	Poisson
Link Function	Log
Dependent Variable	Infections

Number of Observations Read	287
Number of Observations Used	287

Class Level Information		
Class	Levels	Values
Swimmer	2	Freq Occas
Location	2	Beach NonBeach
Gender	2	Female Male

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	282	760.0060	2.6951
Scaled Deviance	282	760.0060	2.6951
Pearson Chi-Square	282	963.5838	3.4170
Scaled Pearson X2	282	963.5838	3.4170
Log Likelihood		-235.6148	
Full Log Likelihood		-566.2004	
AIC (smaller is better)		1142.4008	
AICC (smaller is better)		1142.6143	
BIC (smaller is better)		1160.6982	

Algorithm converged.

The scaled deviance and scaled Pearson chi-square for the Value/DF column is 2.6951 and 3.4170 respectively. These values are not very close to 1, which indicates possible overdispersion.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3292	0.2517	0.8360 1.8225	27.90	<.0001
Swimmer	Freq	1	-0.6086	0.1050	-0.8145 -0.4028	33.59	<.0001
Swimmer	Occas	0	0.0000	0.0000	0.0000 0.0000	.	.
Location	Beach	1	-0.4896	0.1048	-0.6951 -0.2841	21.81	<.0001
Location	NonBeach	0	0.0000	0.0000	0.0000 0.0000	.	.
Age		1	-0.0261	0.0122	-0.0500 -0.0021	4.55	0.0330
Gender	Female	1	0.0294	0.1092	-0.1846 0.2433	0.07	0.7878
Gender	Male	0	0.0000	0.0000	0.0000 0.0000	.	.
Scale		0	1.0000	0.0000	1.0000 1.0000		

**Note:** The scale parameter was held fixed.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Swimmer	1	35.16	<.0001
Location	1	22.35	<.0001
Age	1	4.64	0.0312
Gender	1	0.07	0.7881

The significant factors that affect the average value of the number of **Infections** include **Swimmer**, **Location**, and **Age**. The variable **Gender** is not significant. However, your data might have overdispersion, which means that the standard errors might be underestimated. You should adjust for the overdispersion before deciding which factors are significant in this model.

- d. Use the GENMOD procedure to fit a negative binomial regression model to the data. Does this model account for the overdispersion? What factors are now significant?

```
proc genmod data=STAT2.earinfection;
  class swimmer location gender;
  model infections=swimmer location age gender
    / dist=negbin link=log type3;
run;                                *ST205s01.sas;
```

Model Information	
Data Set	STAT2.EARINFECTION
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	Infections

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	282	268.6917	0.9528
Scaled Deviance	282	268.6917	0.9528
Pearson Chi-Square	282	284.3559	1.0084
Scaled Pearson X2	282	284.3559	1.0084
Log Likelihood		-115.3547	
Full Log Likelihood		-445.9403	
AIC (smaller is better)		903.8807	
AICC (smaller is better)		904.1807	
BIC (smaller is better)		925.8376	

Algorithm converged.

The goodness-of-fit statistics indicate that the negative binomial model is a good fit to the data. The chi-square value divided by the degrees of freedom is close to 1.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.4874	0.4811	0.5445 2.4303	9.56	0.0020
Swimmer	Freq	1	-0.6177	0.1911	-0.9922 -0.2431	10.45	0.0012
Swimmer	Occas	0	0.0000	0.0000	0.0000 0.0000	.	.
Location	Beach	1	-0.4888	0.1976	-0.8761 -0.1016	6.12	0.0133
Location	NonBeach	0	0.0000	0.0000	0.0000 0.0000	.	.
Age		1	-0.0346	0.0218	-0.0774 0.0082	2.51	0.1132
Gender	Female	1	0.0888	0.2081	-0.3190 0.4966	0.18	0.6695
Gender	Male	0	0.0000	0.0000	0.0000 0.0000	.	.
Dispersion		1	1.7570	0.2740	1.2942 2.3851		

**Note:** The negative binomial dispersion parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Swimmer	1	10.16	0.0014
Location	1	6.02	0.0141
Age	1	2.49	0.1147
Gender	1	0.18	0.6689

Now, only **Swimmer** and **Location** are significant. **Age** is no longer significant in the negative binomial model.

- e. Use the PSCALE option in the MODEL statement to adjust for the possible overdispersion. What factors are now significant? How do you back-transform the model to obtain the model for the average count of ear infections for female occasional beach swimmers?

```
proc genmod data=STAT2.earinfection;
  class swimmer location gender;
  model infections=swimmer location age gender
    / dist=poisson link=log type3 pscale;
run;
*ST205s01.sas;
```

## Partial PROC GENMOD Output

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
<b>Deviance</b>	282	760.0060	2.6951
<b>Scaled Deviance</b>	282	222.4214	0.7887
<b>Pearson Chi-Square</b>	282	963.5838	3.4170
<b>Scaled Pearson X2</b>	282	282.0000	1.0000
<b>Log Likelihood</b>		-68.9544	
<b>Full Log Likelihood</b>		-566.2004	
<b>AIC (smaller is better)</b>		1142.4008	
<b>AICC (smaller is better)</b>		1142.6143	
<b>BIC (smaller is better)</b>		1160.6982	

Both the scaled Pearson chi-square (1.00) and the scaled deviance (0.7887) are now close to 1, indicating that overdispersion is no longer a problem for this model.

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
<b>Intercept</b>		1	1.3292	0.4652	0.4175 2.2410	8.16	0.0043
<b>Swimmer</b>	Freq	1	-0.6086	0.1941	-0.9891 -0.2281	9.83	0.0017
<b>Swimmer</b>	Occas	0	0.0000	0.0000	0.0000 0.0000	.	.
<b>Location</b>	Beach	1	-0.4896	0.1938	-0.8694 -0.1098	6.38	0.0115
<b>Location</b>	NonBeach	0	0.0000	0.0000	0.0000 0.0000	.	.
<b>Age</b>		1	-0.0261	0.0226	-0.0703 0.0182	1.33	0.2487
<b>Gender</b>	Female	1	0.0294	0.2018	-0.3661 0.4249	0.02	0.8842
<b>Gender</b>	Male	0	0.0000	0.0000	0.0000 0.0000	.	.
<b>Scale</b>		0	1.8485	0.0000	1.8485 1.8485		

**Note:** The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
<b>Swimmer</b>	1	282	10.29	0.0015	10.29	0.0013
<b>Location</b>	1	282	6.54	0.0111	6.54	0.0105
<b>Age</b>	1	282	1.36	0.2448	1.36	0.2438
<b>Gender</b>	1	282	0.02	0.8845	0.02	0.8844

Now, the significant variables are **Swimmer** and **Location**. The variable **Age** is no longer significant after adjusting for overdispersion.

The model for female occasional beach swimmers is  
 $\log(E(\text{Infections})) = 1.3292 - 0.4896 - 0.0261 * \text{Age} + 0.0294 = 0.869 - 0.0261 * \text{Age}$ .

It follows that the average number of ear infections for female occasional beach swimmers is  
 $E(\text{infections}) = e^{0.869 - 0.0261 \cdot \text{Age}}$ .

## 2. Conducting a Linear Regression Analysis

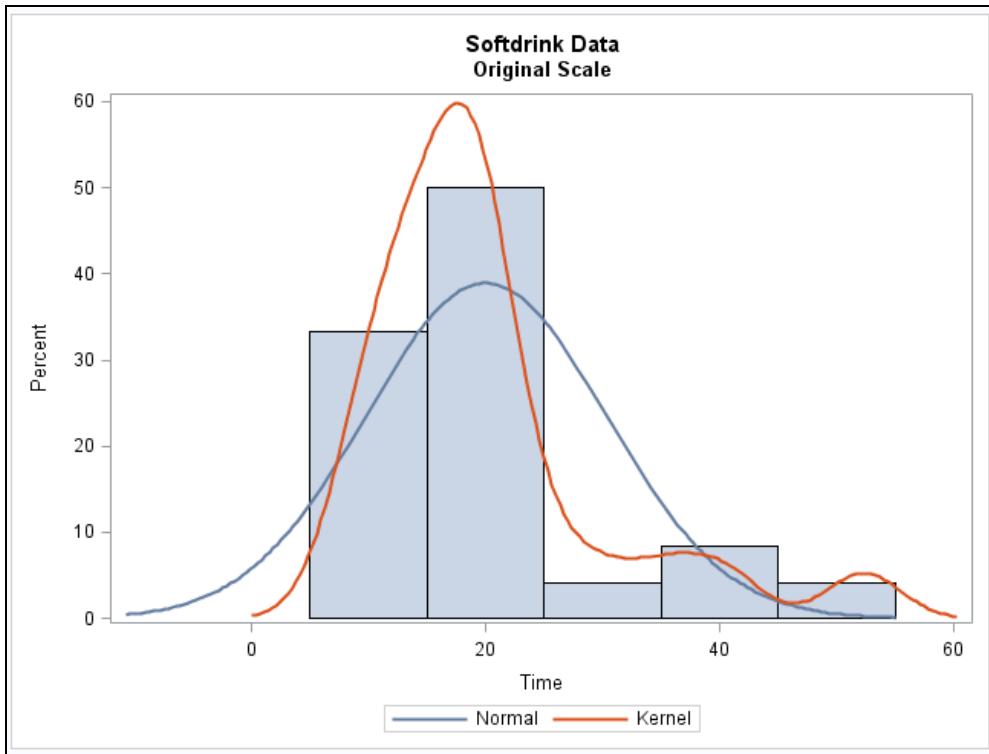
- a. Print the data and then create a histogram of the data on the original scale by using **Time** in the HISTOGRAM statement in PROC SGPlot. Overlay the kernel density and a normal density. Do the times seem to be distributed normally?

```
title;
proc print data=STAT2.softdrinks;
  title 'Softdrink Data';
run;

proc sgplot data=STAT2.softdrinks;
  histogram time;
  density time;
  density time / type=kernel;
  title1 'Softdrink Data';
  title2 'Original Scale';
run;                                         *ST205s02.sas;
```

**Softdrink Data**

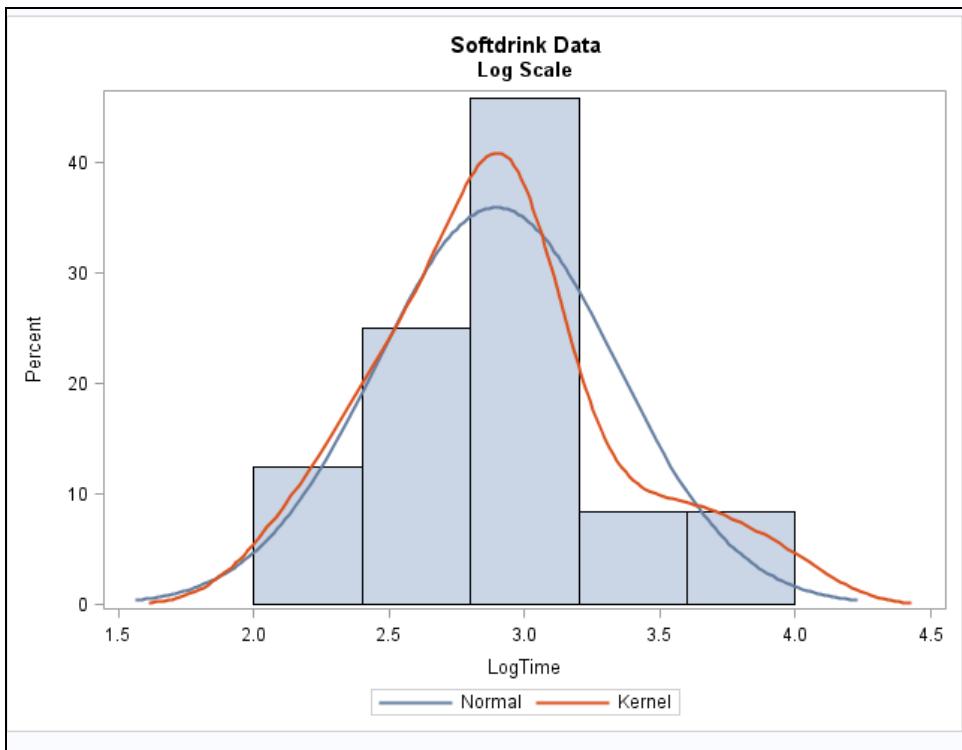
Obs	Time	Cases	Distance	LogTime
1	16.68	7	560	2.81421
2	11.50	3	220	2.44235
3	12.03	3	340	2.48740
4	14.88	4	80	2.70002
5	13.75	6	150	2.62104
6	18.11	7	330	2.89646
7	8.00	2	110	2.07944
8	17.83	7	210	2.88088
9	21.50	5	605	3.06805
10	40.33	16	688	3.69710
11	21.00	10	215	3.04452
12	13.50	4	255	2.60269
13	19.75	6	462	2.98315
14	24.00	9	448	3.17805
15	29.00	10	776	3.36730
16	15.35	6	200	2.73112
17	19.00	7	132	2.94444
18	9.50	3	36	2.25129
19	35.10	17	770	3.55820
20	17.90	10	140	2.88480
21	52.32	26	810	3.95738
22	18.75	9	450	2.93119
23	19.83	8	635	2.98720
24	10.75	4	150	2.37491



The data do not appear to be normally distributed. Because it is skewed to the right and has only positive values, the gamma distribution might be a good fit.

- b. Create a histogram of the data on the log-transformed scale by using **LogTime** in the HISTOGRAM statement in PROC SGPlot. Overlay the kernel density and a normal density. What do you conclude about the distribution of the log-transformed times?

```
proc sgplot data=STAT2.softdrinks;
  histogram logtime;
  density logtime;
  density logtime / type=kernel;
  title2 'Log Scale';
run;                                         *ST205s02.sas;
```



The log-transformed data seem to more closely resemble the normal distribution than the original data. Conduct a gamma analysis and compare how close the predicted values of **Time** are to the observed values.

- c. Use PROC GLIMMIX to model **Time** as a function of **Cases**, **Cases** squared, **Distance**, and **Distance** squared. In the MODEL statement, use the options DIST=GAMMA and LINK=LOG.

```
proc glimmix data=STAT2.softdrinks;
  model time= cases cases*cases distance distance*distance / solution
    dist=gamma link=log;
  title1 'Softdrink Data - Gamma Regression with Log Link';
  run; *ST205s02.sas;
```

Parameter Estimates						
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	
Intercept	1.9994	0.1041	19	19.21	<.0001	
Cases	0.1095	0.01818	19	6.02	<.0001	
Cases*Cases	-0.00202	0.000659	19	-3.06	0.0064	
Distance	0.000725	0.000531	19	1.37	0.1880	
Distance*Distance	-2.26E-7	0	19	-Infty	<.0001	
Scale	0.01498	0.004315	.	.	.	

The **distance\*distance** term has an estimate of 0 and a standard error of 0, so model reduction begins when you remove this term.

- For the gamma distribution, the scale parameter reported in PROC GLIMMIX is the reciprocal of the scale parameter reported in PROC GENMOD.

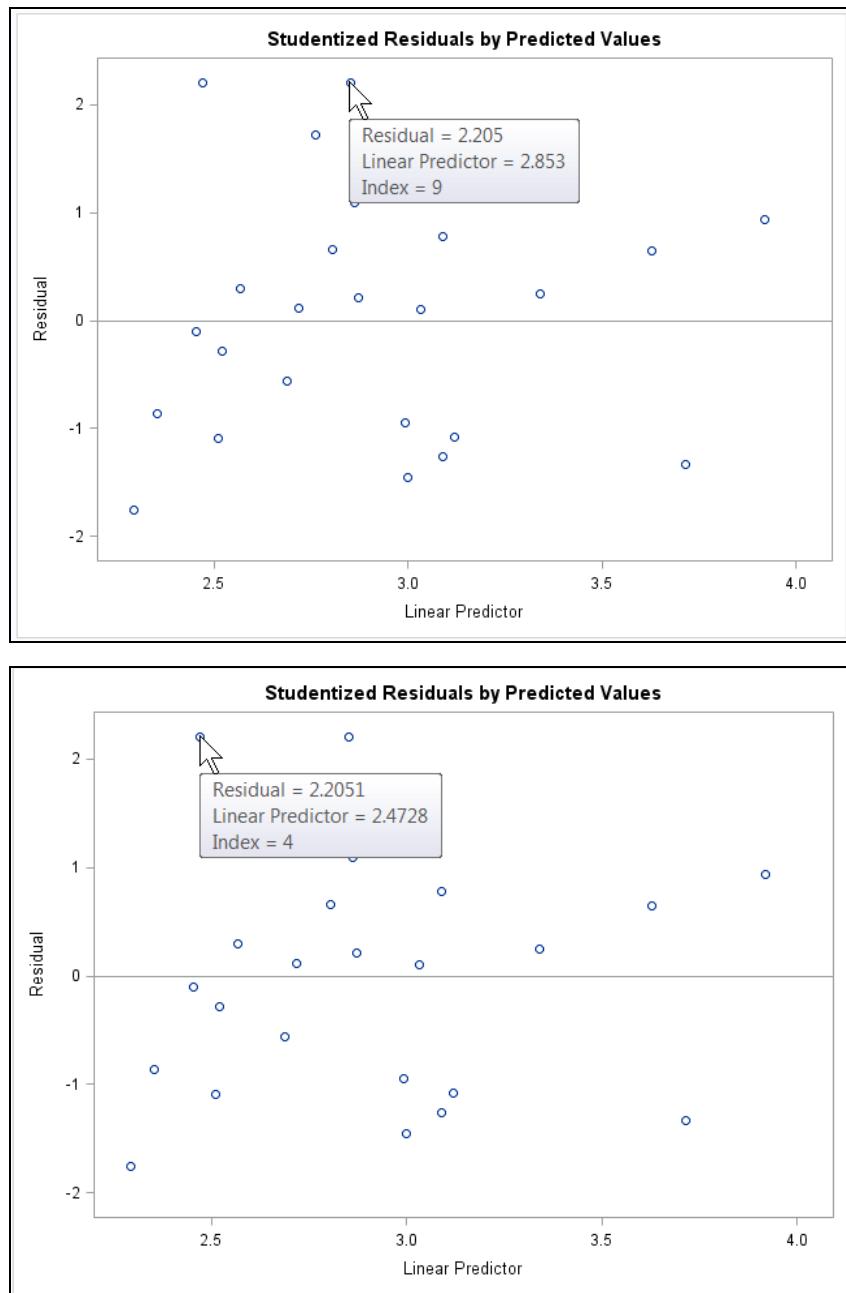
- d. Eliminate any terms that are not significant to obtain a final candidate model.
- Use the PLOTS=STUDENTPANEL (UNPACK) option in the PROC GLIMMIX statement. This creates a plot of the studentized residuals versus the linear predictor.
  - Use the OUTPUT OUT=**gamma\_predicted** statement to save the predicted values in the data set **gamma\_predicted**.
  - Use the HTML output destination. Enable IMAGEMAP using the ODS GRAPHICS statement if you want to be able to identify particular observations in any of these graphs.
  - After looking at the plots of the standardized residuals, do you think the model fits well?

```
*remove distance*distance;
ods graphics / imagemap=on;
proc glimmix data=STAT2.softdrinks plots=studentpanel(unpack);
  model time= cases cases*cases distance / dist=gamma link=log
    solution;
  output out=gamma_predicted pred(ilink)=pred;
title1 'Softdrink Data - Final Model';
run; *ST205s02.sas;
```

Parameter Estimates						
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	
Intercept	2.0211	0.08300	20	24.35	<.0001	
Cases	0.1103	0.01808	20	6.10	<.0001	
Cases*Cases	-0.00209	0.000623	20	-3.36	0.0031	
Distance	0.000550	0.000145	20	3.80	0.0011	
Scale	0.01506	0.004336	-	-	-	

After you remove **distance\*distance**, all terms in the model have reasonable estimates and are significant.

## PROC GLIMMIX ODS Graphics Output



The residuals seem to be a random scatter about zero with no apparent patterns. There are two data points with studentized residuals above 2 that you might want to investigate.

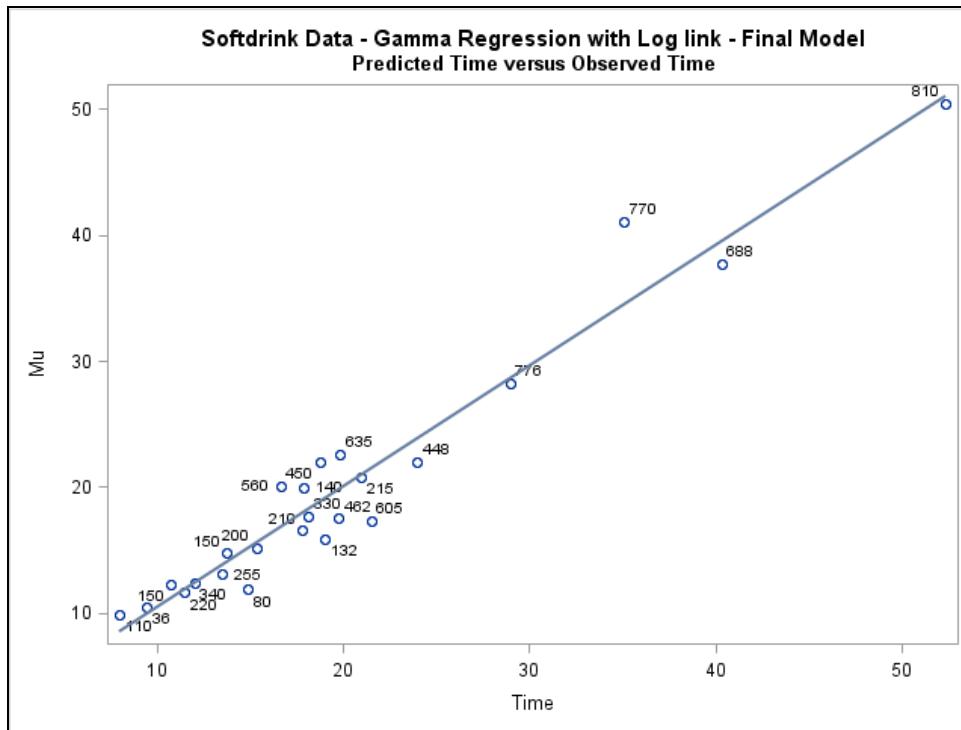
- e. Use the **gamma\_predicted** data set and PROC SGPLOT with the REG statement to plot the predicted times against the observed times. Use the DATALABEL=Distance option in the REG statement.

```

title 'Softdrink Data - Gamma Regression with Log link - Final Model';
proc sgplot data=gamma_predicted noautolegend;
  reg y=pred x=time / datalabel=distance;
  title2 'Predicted Time versus Observed Time';
run; *ST205s02.sas;

```

PROC SGPILOT Output



The predicted values for **Time** seem to align closely to the observed values, except for one data point, where the technician walked 770 feet to service the account. The model predicts the time needed to service the account for this observation.

## Solutions to Student Activities (Polls/Quizzes)

### 5.02 Multiple Choice Poll – Correct Answer

Which of the following is **not** true?

- a. Logistic regression, Poisson regression, and gamma regression are all examples of generalized linear models.
- b.** In generalized linear models, the mean and variance are unrelated.
- c. Canonical link functions are commonly used link functions for many exponential family distributions.
- d. Although the canonical link function for the gamma distribution is the inverse, modelers often use the logarithm link function.

12

### 5.03 Multiple Choice Poll – Correct Answer

Why can you not use OLS regression for a count of rare events with a skewed distribution?

- a. OLS regression assumes normal distribution of errors.
- b. OLS regression assumes constant variance.
- c. OLS regression can produce both positive and negative predicted values.
- d.** all the above

22

## 5.05 Poll – Correct Answer

Is overdispersion a problem in ordinary least squares regression?

- Yes
- No

31

## 5.06 Multiple Answer Poll – Correct Answers

Failing to correct for overdispersion results in which of the following? (Choose all that apply.)

- a. underestimated parameter estimates
- b. overestimated parameter estimates
- c. underestimated standard errors for parameter estimates
- d. overestimated test statistics, and therefore, a too small  $p$ -value

39

## 5.07 Quiz – Correct Answer

In the Poisson model, the parameter estimate for swimmer **Freq** is **-0.6086**. How do you interpret this value?

Notice that **Exp(-0.6086) = 0.544**.

**Comparing frequent swimmers and occasional swimmers, the log of the expected number of ear infections decreases by 0.6086 for frequent swimmers. In other words, the expected number of ear infections for frequent swimmers is 54.4% of the number of ear infections for occasional swimmers**

42

## 5.08 Multiple Choice Poll – Correct Answer

You want to model the rate of car insurance claims by geographic zone. The offset variable is which of the following?

- a. the number of claims
- b. the area of the geographic zone
- c. the number of insured in each geographic zone
- d. the population in the geographic zone

47

### 5.09 Poll – Correct Answer

A gamma regression model can be useful for positive values with large means that are skewed to the right.

- True
- False

60

### 5.10 Poll – Correct Answer

If you use PROC GLIMMIX with a log-link function, the predicted value is on the log scale. You need to request the predicted values with the ILINK option to obtain unbiased estimates of the means on the original scale.

- True
- False

65

# Chapter 6     Introduction to Linear Mixed Models

<b>6.1</b>	<b>Defining Linear Mixed Models .....</b>	<b>6-3</b>
<b>6.2</b>	<b>Using the GLIMMIX Procedure.....</b>	<b>6-13</b>
	Demonstration: Fitting a Mixed Model Using PROC GLIMMIX .....	6-21
	Exercises .....	6-27
<b>6.3</b>	<b>Solutions .....</b>	<b>6-29</b>
	Solutions to Exercises .....	6-29
	Solutions to Student Activities (Polls/Quizzes).....	6-34



## 6.1 Defining Linear Mixed Models

---

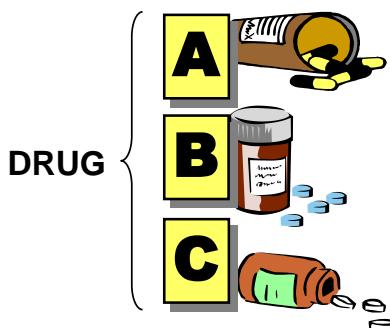
### Objectives

- Define fixed and random effects.
- List linear mixed model assumptions
- Describe PROC GLIMMIX syntax and estimation methods.

3

### Fixed Effects

- The levels of interest result from deliberate choice, not from sampling a distribution.
- Inferences are to be made only to those levels included in the study.



4

*Fixed effects* are those factors whose levels are selected deliberately to evaluate the differences. All levels of interest are in your data set. The researcher is interested in comparing the effects of the factors on the response variable only for those levels included in the study.

For example, in a drug study, you want to compare the effect of three drugs (A, B, and C). You are interested in the comparison of only these three drugs. You know what they are before conducting the experiment.

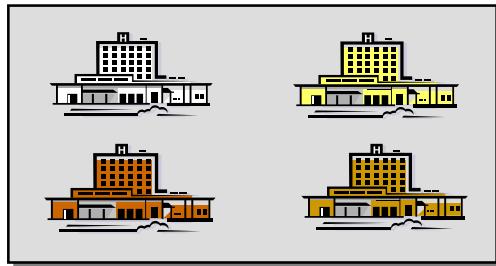
The variable **Drug** is a fixed effect. Other examples of fixed effects that represent all the levels of interest for the study might include **Gender**, **Treatment**, **Method**, and **Brand**.

A model containing only fixed effects is called a *fixed effects model*.

## Random Effects

- Levels represent a sample from a population with a probability distribution.
- Inferences about the fixed effects apply to all levels of random effects in the population, not only the subset of levels included in the study.

### CLINICS



5

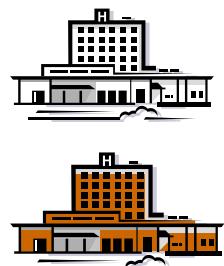
In some situations, a factor might have a large number of levels and the researcher or data analyst selects a subset of the levels to be included in the study. They represent a sample (although often an imperfect sample) from a population with a probability distribution. The inference about fixed effects from the data analysis applies to all population levels of random effects and not only the subset of levels included in the study. Effects like these are *random effects*. For example, in the same drug study, four clinics are randomly selected from a population of clinics in a region. The researcher wants to make an inference for the drug effects across the population of clinics, not only the ones included in the study. Then **Clinic** is a random effect.

Models in which all effects are random are called *random effects models*. Variances associated with random effects are known as *variance components*.

## Mixed Models

Models in which some factors are fixed effects and other factors are random effects are called ***mixed models***.

### CLINIC – random

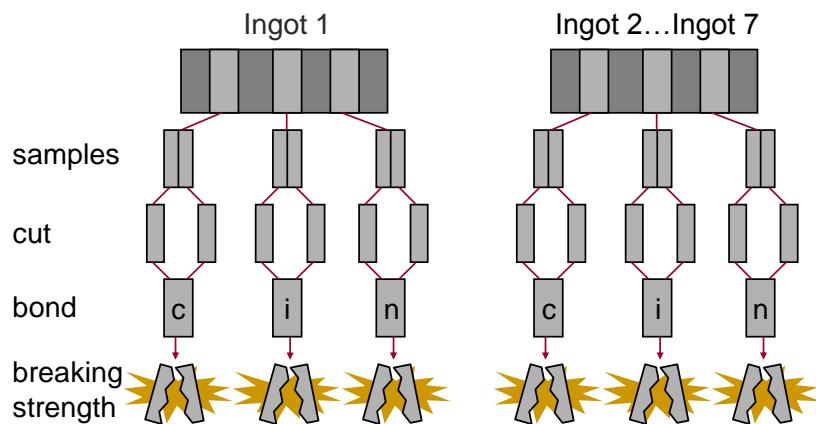


### DRUG – fixed



6

## Ingot Example



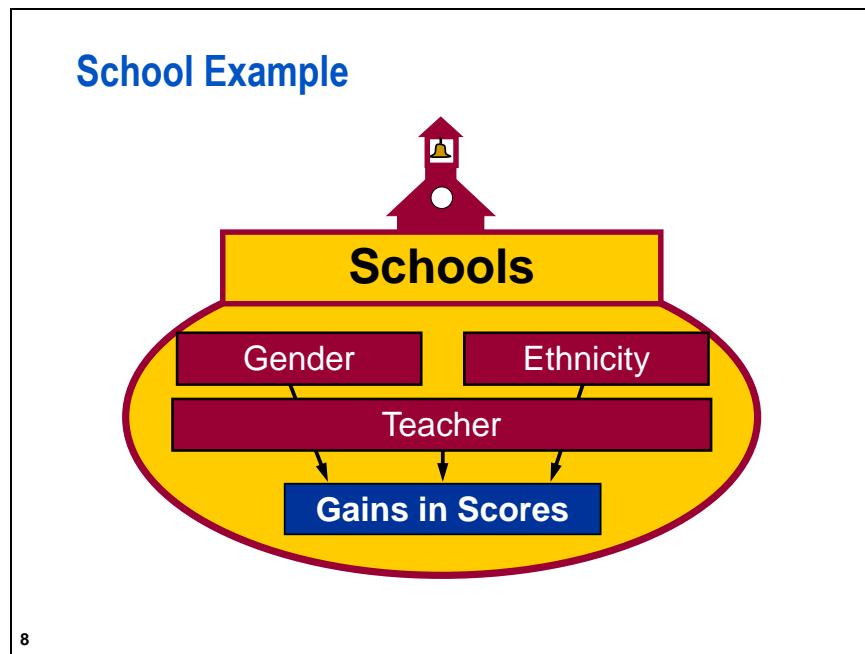
7

An engineer wants to test the strength of three metals used as bonding agents. Seven ingots made of a composition material are randomly selected from a population of ingots and are used for this strength test. A sample of material is taken from each ingot, and a bond is formed using one of the metals. The amount of pressure required to break the bond is then recorded.

This design consists of the following effects:

**Metal** a treatment effect. This is a ***fixed*** effect because only three metals (**Copper**, **Iron**, and **Nickel**) are used in the study, and the engineer is interested in making an inference about only these three metals.

**Ingots** a blocking effect. This is a **random** effect because the seven ingots are randomly selected from a population of ingots, and the inference about the treatment means is made over the entire population of ingots.



Gains in scores on a standardized test were recorded for 1,515 fourth-grade students in all schools in a district. **Gain** is defined as the score at the end of the year minus the score at the beginning of the year. The students' genders and ethnicities, as well as the identification numbers of the students' teachers, were also recorded. The primary objective was to evaluate and compare the schools in the gain scores. A secondary objective was to assess the effects of gender and ethnicity. This is a data set from an observational study.

The data consist of the following effects:

- School** is a **fixed** effect because only the schools included in the study are of interest.
- Ethnicity** is a **fixed** effect.
- Gender** is a **fixed** effect.
- Teacher** is a **random** effect because the teachers represent a sample of the population of teachers who could teach at the schools.

- ✍ If the effect level can reasonably be assumed to represent a probability distribution, then the effect is random. Blocks, laboratories, clinics, study centers, workers, teachers, sires, and so on, typically (but not always) represent a sample (although often an imperfect sample) of a population with a probability distribution (normal).
- ✍ The distinction between a fixed effect and a random effect might not always be clear in some situations. Some statisticians think that even if the levels of a factor were not randomly selected, the effect can still be considered random as long as the effects on the outcome are of a stochastic nature (Shabenberger, O. and Pierce, F. J. 2002).

## 6.01 Quiz

Three growing methods were studied to evaluate which one is more effective for growing grass. Five varieties of grass seeds were randomly selected from a large population of varieties. Each method was applied to each of the varieties. Dry yield is measured at the end of the season. Is variety fixed or random?

9

## Setup for the Poll

A health-care provider wants to study the variation of the cost of a medical procedure in two types of U.S. health-care facilities (clinics or hospitals). The three factors of interest are **State**, **City**, and **Type**.

- Five states are randomly selected within the U.S.
- Four cities are randomly selected within each state.
- Three of each type of facility (hospital or clinic) are selected within each city.

11

## 6.02 Poll

The variables **State** and **City** are both random effects and **Type** is a fixed effect.

- True
- False

12

## General Linear Models

fixed effects only model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mixed model

14

In the fixed-effects-only model that you typically fit using the GLM, GLMSELECT, or REG procedure, you have the following:

- Y** is the vector of observed response data values.
- X** is the known design matrix based on the model specification.
- $\beta$**  is the vector of unknown fixed-effect parameters.
- $\varepsilon$**  is the vector of random errors.

You assume that  $\boldsymbol{\varepsilon}$  is a vector of independent random errors that are normally distributed with a mean of 0 and a variance of  $\sigma^2$ .

In a linear mixed model that you fit using PROC MIXED, you have the following:

- $\mathbf{Y}$  is the vector of observed response data values.
- $\mathbf{X}$  is the known design matrix for the fixed effects.
- $\boldsymbol{\beta}$  is the vector of unknown fixed-effect parameters.
- $\mathbf{Z}$  is the known design matrix for the random effects.
- $\boldsymbol{\gamma}$  is the vector of unknown random effects parameters.
- $\boldsymbol{\varepsilon}$  is the vector of random errors.

The mixed model extends the general linear model by including random effects in the model. It also allows for correlated errors. The general linear model that you fit in previous chapters is a special case of the linear mixed model that you fit using PROC GLIMMIX.

## Linear Mixed Model Assumptions

- Random effects and random errors are normally distributed with mean zero and covariance matrices  $\mathbf{G}$  and  $\mathbf{R}$ , respectively.
- Random effects and random errors are independent of each other.
- The means (expected values) of the responses are linearly related to the predictor variables (*linear* in terms of fixed-effects parameters).

15

Because normal data can be modeled entirely in terms of the means and variances or covariances, the two sets of parameters in a mixed linear model actually specify the complete probability distribution of the data. The parameters of the mean model are referred to as *fixed-effects parameters*, and the parameters of the variance-covariance model are referred to as *covariance parameters*.

## GLIMMIX Procedure

General form of the GLIMMIX procedure:

```
PROC GLIMMIX options;  

  CLASS variables;  

  COVTEST 'label' test-specification /options;  

  LSMESTIMATE fixed-effect 'label' values /options;  

  MODEL response=fixed-effects / options;  

  RANDOM random-effects /options;  

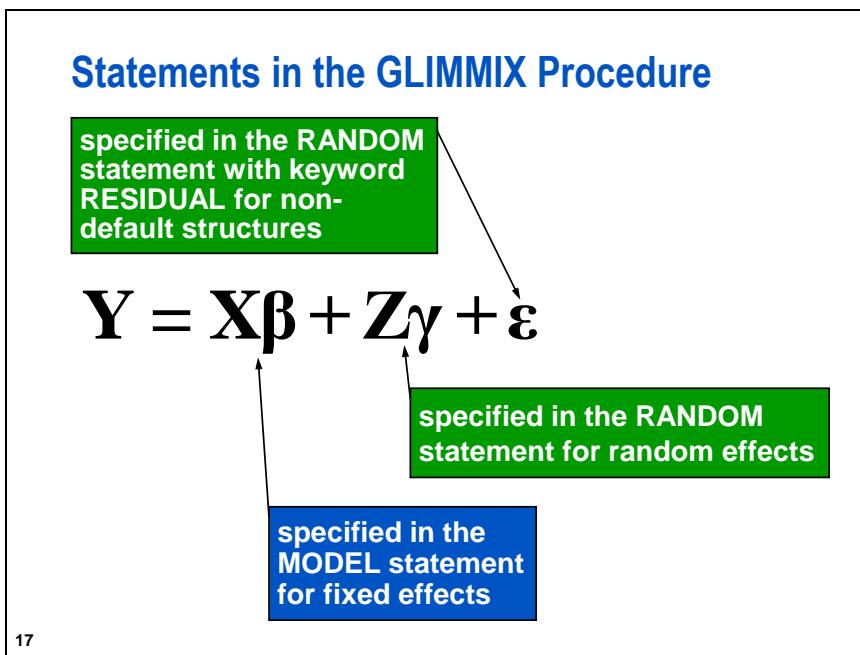
  RANDOM _RESIDUAL_ / options;  

RUN;
```

16

Selected GLIMMIX procedure statements:

- CLASS** names the classification variables to be used in the analysis. If the CLASS statement is used, it must appear before the MODEL statement. Classification variables can be either character or numeric.
- COVTEST** provides a mechanism to obtain statistical inferences for the covariance parameters. Significance tests are based on the ratio of (residual) likelihoods or pseudo-likelihoods. Confidence limits and bounds are computed as Wald or likelihood ratio limits. You can specify multiple COVTEST statements.
- LSMESTIMATE** requests custom hypothesis tests among the least squares means.
- MODEL** is a required statement that names a single dependent variable and the fixed effects. You do not specify random effects in the MODEL statement.
- RANDOM** defines the Z matrix of the mixed model, the random effects in the  $\gamma$  vector, the structure of G, and the structure of R. The random effects can be classification or continuous effects, and multiple RANDOM statements are possible. The RANDOM \_RESIDUAL\_ statement indicates a residual-type (R-side) random component that defines the R matrix.



In the GLIMMIX procedure, you use the MODEL statement to specify the fixed effects, the RANDOM statement to specify the random effects, and the RANDOM statement with the RESIDUAL keyword to specify the variance-covariance structure of the errors that is not the default  $\sigma^2 I_n$ . You should be careful when you specify more than one RANDOM statement in PROC GLIMMIX. In some cases, one statement with the specified variance-covariance structure captures all the variations in your data and is sufficient.

The GLIMMIX procedure distinguishes two types of random effects. Depending on whether the variance of the random effect is contained in  $G$  or in  $R$ , these are referred to as *G-side* and *R-side* random effects. R-side effects are also named *residual effects*. Simply, if a random effect is an element of  $G$ , it is a G-side effect. Otherwise, it is an R-side effect. Models fit with the GLIMMIX procedure can have none, one, or more of each type of effect.

- ✍ In the GLIMMIX procedure, all random effects are specified through the RANDOM statement. Various statistical analyses using PROC GLIMMIX are shown in the Statistical Analysis with the GLIMMIX Procedure course.

## The GLIMMIX Procedure

The GLIMMIX procedure

- can be used for analyzing linear mixed models and repeated measures data
- uses either the restricted maximum likelihood or maximum likelihood estimation method to estimate variance and covariance parameters for normally distributed data.
- uses the generalized least squares (GLS) method to estimate fixed-effect parameters and standard errors for linear mixed models.

18

PROC GLIMMIX was used earlier to fit a generalized linear model for Poisson outcomes. In addition, you can use PROC GLIMMIX to fit a linear model with or without random effects. This procedure is also commonly used for analyzing repeated measures data because of its large selection of variance-covariance matrices.

For linear mixed models, PROC GLIMMIX uses the restricted maximum likelihood estimation method to estimate variance and covariance parameters. It uses the generalized least squares estimation method for fixed effect parameter estimates and standard errors. This method takes into account the variance-covariance matrices for the random effects and residuals and therefore is more appropriate for linear mixed models than ordinary least squares method.

### DETAILS

For a linear mixed model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ , assuming that the random effect  $\boldsymbol{\gamma}$  and the residuals  $\boldsymbol{\varepsilon}$  are independently and normally distributed with the following:

$$E\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \mathbf{0} \text{ and } Var\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

It can be shown that for the observed response variable  $\mathbf{Y}$ , you have the following:

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \text{ and } Var(\mathbf{Y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} = \mathbf{V}$$

PROC GLIMMIX enables you to specify various covariance structures for both the **G** and **R** matrices. The default structure models a different variance component for each random effect.

The generalized least squares (GLS) estimates take into account the covariance matrices **G** and **R**. When you use this estimation method, it can be shown that the parameter estimates and variance are computed as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}, \text{ and } Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$$

The ordinary least squares (OLS) solution for a fixed effect model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$  is given by

$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ , and the standard errors are computed based on  $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . It can be seen that OLS is a special case of the GLS solution with  $\mathbf{V} = \sigma^2 I_n$ . The variance of the OLS solution is also a special case of the variance of the GLS solution.

### 6.03 Multiple Choice Poll

Which of the following are **true** about linear mixed models and the GLIMMIX procedure?

- Linear mixed models are a special case of general linear models that you fit using PROC GLM, PROC GLMSELECT, or PROC REG.
- Linear mixed models can handle normal and nonnormal responses.
- PROC GLIMMIX has a RANDOM statement to model random effects.
- PROC GLIMMIX can be used only to model random effects, and should not be used to model repeated measures data.

19

## 6.2 Using the GLIMMIX Procedure

### Objectives

- Explain crossed effects and nested effects.
- Use the GLIMMIX procedure to fit a linear mixed model.
- Explain the output from PROC GLIMMIX.
- Recognize the difference between treating a factor as random and treating it as fixed.

23

## Crossed Classifications

SCHOOL⇒ GENDER↓	Cottonwood	Dogwood
F	79, 84, 45, 72, .....	87, 93, 73, 96, ...
M	69, 98, 72, 88, ....	76, 82, 84, 89, ....

24

Data can be organized into two types of classification patterns, crossed or nested. If two factors are crossed, then observations are collected for each combination of each level of the two factors.

## Nested Classifications

SCHOOL ⇒	Cottonwood		Dogwood	
CLASS ⇒	1	2	1	2
Obs ⇒	46, 56, ...	54, 58, ...	59, 57, ...	69, 77, ...

25

For nested classifications, the samples are typically taken in several stages:

1. selection of main units (school)
2. selection of subunits from each main unit (class)
3. selection of sub-subunits from subunits, and so on (student)

Notice that in the example above, the two classes labeled "1" are not the same classes. One is from *Cottonwood* and the other is from *Dogwood*. You can renumber the classes from *Dogwood* in many other ways and it should not affect the analysis and results.

## Nested Classifications

- Sampling units are typically classified in a hierarchical manner.
- The classification factors at each stage are typically considered random effects.
- In some cases, factors at the first stage of sampling might be considered fixed.

26

## 6.04 Poll

To evaluate the productivity of two machines, five operators were chosen to operate on the machines. Each operator operated on each of the two machines three times. Operators and machines are crossed.

- True
- False

27

## 6.05 Poll

If, for each machine, five operators were selected and operated only on that particular machine, then operators are nested with machines.

- True
- False

29

## 6.06 Multiple Choice Poll

Five lots were chosen for the study. Four wafers were selected from each lot. The thickness of dioxide layers was measured from each wafer.

- a. Wafers are nested within lots.
- b. Wafers are crossed with lots.
- c. I do not know how wafers are related to lots.

31

## Teacher Example

Materials	A					B	C	D
Teacher	1	2	3	4	5	1...5	1...5	1...5
Student	1...6	1...6	1...6	1...6	1...6	1...6 for each teacher		

Y – represents scores on a standardized test.

33

A large school district is studying four sets of instructional materials and their relationship to test scores on a standardized test. Twenty teachers are randomly selected to use these materials in their classes. Each set of materials is used by five teachers in classes consisting of six randomly selected students. The data are stored in the **STAT2.scores** data set. These are the variables in the data set:

- Material** type of instructional material (*A, B, C, or D*)  
**Teacher** teachers selected for the study (1 to 5 for each type of material)  
**Student** student ID (1 to 6 for each teacher)  
**Score** score of student on standardized test.

The data have a nested classification because teachers are nested within materials. **Material** is considered a fixed effect because only four materials (*A, B, C, or D*) are used in the study and you are interested only in making inference about these four materials. **Teacher** is considered a random effect because they are randomly selected from a population of teachers. Teachers are nested within materials.

The purpose of the study is to accomplish the following:

- estimate and compare the treatment (**Material**) means over the entire population of teachers
- account for the variability in the response variable (**Score**) due to the **Teacher** variance

 In this example, the values for teacher (1 to 5) and student (1 to 6) are arbitrary. They can be coded with any other values, if the teachers within a material have distinct values, as do the students within a teacher.

## The Data

Obs	Material	Teacher	Student	Score
1	A	1	1	92
2	A	1	2	93
3	A	1	3	88
4	A	1	4	91
5	A	1	5	95
6	A	1	6	89
7	A	2	1	94
8	A	2	2	86
9	A	2	3	91
10	A	2	4	98
11	A	2	5	82
12	A	2	6	95
13	A	3	1	91
14	A	3	2	99
15	A	3	3	92
16	A	3	4	100
17	A	3	5	92
18	A	3	6	88
19	A	4	1	91
20	A	4	2	99
...				

34

## The Model

Materials fixed effects

$$y_{ijk} = \mu + \alpha_i + b(\alpha)_{ij} + \varepsilon_{ijk}$$

Teacher(Materials)  
random effects

random error

$$b(\alpha)_{ij} \sim N(0, \sigma_t^2)$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

36

$y_{ijk}$  test score for the  $i^{\text{th}}$  material, the  $j^{\text{th}}$  teacher nested within the  $i^{\text{th}}$  material, and the  $k^{\text{th}}$  student within the  $j^{\text{th}}$  teacher for the  $i^{\text{th}}$  material,  $i=1, 2, 3, 4$ ,  $j=1$  to 5 and  $k=1$  to 6.

$\mu$  overall mean.

$\alpha_i$  fixed effect associated with the  $i^{\text{th}}$  material (treatment).

$b(\alpha)_{ij}$  random effect associated with the  $j^{\text{th}}$  teacher nested within the  $i^{\text{th}}$  material,  $b(\alpha)_{ij} \sim \text{i.i.d. } N(0, \sigma_t^2)$ .

These random effects  $b(\alpha)_{ij}$ s are assumed to be independently and normally distributed with mean zero and variance  $\sigma_t^2$ . The variance  $\sigma_t^2$  is the parameter to be estimated in the mixed model for this effect.

$\varepsilon_{ijk}$  experimental error associated with students  $\varepsilon_{ijk} \sim$  i.i.d.  $N(0, \sigma^2)$ . The random errors are assumed to follow a normal distribution with mean zero and variance  $\sigma^2$ . The variance  $\sigma^2$  is the parameter to be estimated in the mixed model for random error.

The effects  $b(\alpha)_{ij}$  and  $\varepsilon_{ijk}$  are assumed to be independent random variables. Therefore,

- $E(y_{ijk}) = \mu + \alpha_i$  is the mean test score for  $i^{\text{th}}$  material averaged across all teachers and students in the population.
- $\text{var}(y_{ijk}) = \sigma^2 + \sigma_t^2$ : The variance of an observation is the sum of the variances due to teachers and students (random errors).

## Hypotheses in a Mixed Model

For fixed effects:  $H_0: \alpha_i = 0$

For random effects:  $H_0: \sigma_t^2 = 0$

37

In mixed model analysis, the hypotheses about the fixed effects remain the same as those in the fixed-effects model: whether there are significant treatment effects. The hypotheses about the random effects are whether the variance components associated with the random effects equal zero. In other words, whether there are significant variations due to these random variables. Often, inferences about random effects are of little interest. The primary role of random effects is to model sources of variation so that the fixed effects can be more accurately estimated and tested.

Before performing an analysis of variance, you should conduct an initial data exploration. You can use PROC SG PANEL to produce side-by-side box plots. The SG PANEL procedure creates a panel of graph cells for the values of one or more classification variables. For example, a data set contains continuous variable A and categorical variable B, and you want to compare the box plots of A for each value of B. Then, you can use the SG PANEL procedure to create these plots. The SG PANEL procedure creates a layout for you automatically and splits the panel into multiple graphs if necessary.

The SG PANEL procedure can create a wide variety of plot types, and overlay multiple plots together in each graph cell in the panel. It can also produce two different types of layout.

## The SGPROC Procedure

General form of the SGPROC procedure:

```
PROC SGPROC options;
  PANELBY variable(s) / option(s);
  HBOX response-variable / option(s);
  HISTOGRAM response-variable / option(s);
  REG X= numeric-variable Y= numeric-variable
        / option(s);
  SCATTER X= variable Y= variable / option(s);
  VBOX response-variable / option(s);
RUN;
```

39

Selected SGPROC procedure statements

**PANELBY** is a required statement and specifies one or more classification variables for the panel, the layout type, and other options for the panel.

**HISTOGRAM** creates a histogram that displays the frequency distribution of a numeric variable.

**HBOX** creates a horizontal box plot that shows the distribution of your data.

**REG** creates a fitted regression line or curve.

**SCATTER** creates a scatter plot.

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

Horizontal and vertical box plots display the distribution of data by using a rectangular box and whiskers. Whiskers are lines that indicate a data range outside of the box.

Selected PANELBY statement options:

**LAYOUT** specifies the type of layout that is used for the panel. Select one of the following values:

**LATTICE** when you specify two classification variables, arranges the cells so that the values of the first variable are columns and the values of the second variable are rows. You can use LATTICE only when you specify exactly two classification variables.

**PANEL** arranges the cells in rows and columns. The headings for each cell are placed at the top of the cell. This is the default.



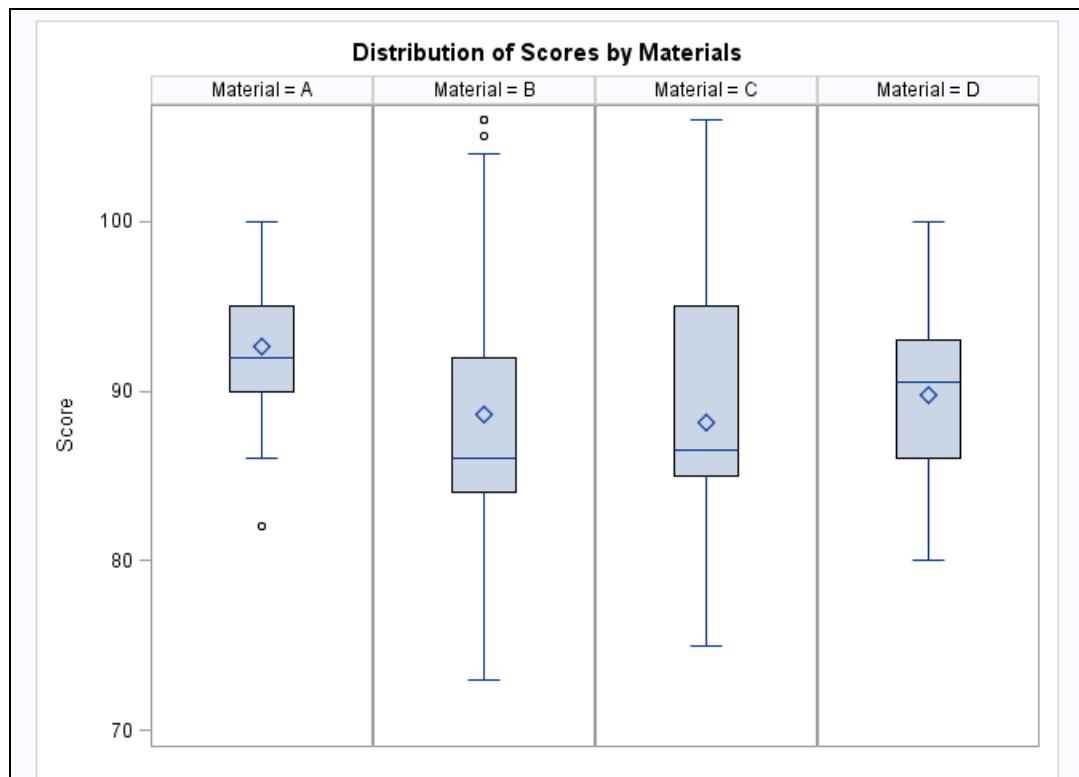
## Fitting a Mixed Model Using PROC GLIMMIX

These are the variables in the **STAT2.scores** data set:

<b>Material</b>	type of instructional materials ( <b>A</b> , <b>B</b> , <b>C</b> , or <b>D</b> )
<b>Teacher</b>	teachers selected for the study (5 for each <b>Material</b> )
<b>Student</b>	student ID (1 to 6 for each <b>Teacher</b> )
<b>Score</b>	score of student on standardized test.

Before performing an analysis of variance, you should conduct an initial data exploration. This can be accomplished, in part, by looking at side-by-side box plots.

```
title1 "Distribution of Scores by Materials";
proc sgpanel data=STAT2.scores;
  panelby material / columns=4;
  vbox score;
run; *ST206d01.sas;
```



There appear to be some differences among the four treatment means, although whether the difference is due to the material effect or random errors remains to be seen. Material A seems to be less variable than materials *B* and *C*.

Suppose you want to determine whether there is a significant difference in the mean test scores using four different teaching materials. You can use the GLIMMIX procedure to analyze the **STAT2.scores** data, and treat **Teacher(Material)** as a random effect.

Use the COVTEST statement with the GLM keyword to test your model against a null model of complete independence. All G-side covariance parameters are eliminated and the R-side covariance structure is reduced to a diagonal structure.

```
proc glimmix data=STAT2.scores;
  class material teacher;
  model score=material;
  random teacher(material);
  covtest 'Test Need for Random Effect' glm;
run; *ST206d02.sas;
```

All classification variables (fixed and random effects) are listed in the CLASS statement, but only the fixed effect (**Material**) is listed in the MODEL statement. The random effect (**Teacher(Material)**) is specified in the RANDOM statement.

#### PROC GLIMMIX Output

Model Information	
Data Set	STAT2.SCORES
Response Variable	Score
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Containment

The Model Information table displays basic information about the fitted model, such as the link and variance functions, the distribution of the response, and the data set. The default estimation technique for the normal distribution is Restricted Maximum Likelihood. The row in this table labeled *Degrees of Freedom Method* lists the method used for estimating the denominator degrees of freedom for the fixed effect. Five possibilities for this row are Containment, Between-Within, Residual, Satterthwaite, Kenward-Roger, and none. Containment is the default method when the RANDOM statement is specified. You can use the DDFM= option in the MODEL statement to specify other methods.

Class Level Information		
Class	Levels	Values
Material	4	A B C D
Teacher	5	1 2 3 4 5

The Class Level Information table lists the levels of every variable specified in the CLASS statement. You should check this information to make sure that the data are correct. You can adjust the order of the

CLASS variable levels with the ORDER= option in the PROC GLIMMIX statement. The default order is alphanumeric.

<b>Number of Observations Read</b>	120
<b>Number of Observations Used</b>	120

<b>Dimensions</b>	
<b>G-side Cov. Parameters</b>	1
<b>R-side Cov. Parameters</b>	1
<b>Columns in X</b>	5
<b>Columns in Z</b>	20
<b>Subjects (Blocks in V)</b>	1
<b>Max Obs per Subject</b>	120

The Number of Observations table shows the total number of observations in the data set and how many are used in fitting the model. All 120 observations on the data set were used for the analysis. The Dimensions table lists the sizes of relevant matrices. This table can be useful in determining CPU time and memory requirements. The G-side covariance parameter corresponds to  $\sigma^2$ ; the R-side covariance parameter is  $\sigma^2$ ; the five columns in the **X** matrix correspond to the intercept and four design columns for the classification variable **Material**; the 20 columns in the **Z** matrix correspond to the 20 teachers.

<b>Optimization Information</b>	
<b>Optimization Technique</b>	Dual Quasi-Newton
<b>Parameters in Optimization</b>	1
<b>Lower Boundaries</b>	1
<b>Upper Boundaries</b>	0
<b>Fixed Effects</b>	Profiled
<b>Residual Variance</b>	Profiled
<b>Starting From</b>	Data

<b>Iteration History</b>					
<b>Iteration</b>	<b>Restarts</b>	<b>Evaluations</b>	<b>Objective Function</b>	<b>Change</b>	<b>Max Gradient</b>
0	0	4	714.2814353	.	3.24E-14

Convergence criterion (ABSGCONV=0.00001) satisfied.

The optimization is performed using a Dual Quasi-Newton algorithm, and the rows of this table describe the iterations that this algorithm takes in order to minimize the objective function. Other algorithms are available by using the NLOPTIONS statement. The Iteration History table describes the optimization of the residual/restricted log likelihood function.

Fit Statistics	
-2 Res Log Likelihood	714.28
AIC (smaller is better)	718.28
AICC (smaller is better)	718.39
BIC (smaller is better)	720.27
CAIC (smaller is better)	722.27
HQIC (smaller is better)	718.67
Generalized Chi-Square	2001.58
Gener. Chi-Square / DF	17.25

The Fit Statistics table provides information for goodness of model fit. All information criteria consider both the fit of the model and the model complexity. The model with more parameters receives a larger penalty. Bayesian Information Criterion (BIC) tends to produce a larger penalty than both Akaike Information Criterion (AIC) and finite-sample corrected Akaike Information Criterion (AICC) for the same model. For all information criteria, smaller values indicate a better model.

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
Teacher(Material)	34.6596	13.2770
Residual	17.2550	2.4402

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Material	3	16	0.54	0.6647

The variance component estimates are  $\hat{\sigma}_t^2 = 34.6596$  and  $\hat{\sigma}^2 = 17.2550$ .

The Type 3 Tests of Fixed Effects table contains the  $F$  test for **Material** ( $F=0.54$ ), with a  $p$ -value of 0.6647. Therefore, there is not enough evidence to conclude that the average test scores for **Material** across all teachers are statistically significantly different at a 5% significance level.

 PROC GLIMMIX does not compute the sums of squares as PROC GLM does. Therefore, you do not see Sums of Squares in the Test of Fixed Effects table. The  $F$  statistic for the fixed effect is

$$\text{computed as } F = \frac{\hat{\beta}' L' (L Q^{-1} L')^{-1} L \hat{\beta}}{\text{rank}(L Q L)}, \text{ where}$$

$\hat{\beta}$  is the vector of the fixed-effect estimates.

$L$  is the coefficient matrix used to test for the fixed effects.

$Q$  depends on the estimation method and options

For a linear mixed model,  $Q = (X\hat{V}^{-1}X)$ . This statistic accounts for all variance components in the model as indicated by  $\hat{V}$  in the equation.

Tests of Covariance Parameters Based on the Restricted Likelihood						
Label	DF	-2 Res Log Like	ChiSq	Pr > ChiSq	Note	
Test Need for Random Effect	1	786.77	72.49	<.0001	MI	

#### MI: P-value based on a mixture of chi-squares.

The null hypothesis for the COVTEST statement is that a model that assumes independence (that is, no RANDOM statement) fits as well as the current model with the RANDOM statement. The low  $p$ -value (<.0001) provides evidence to reject the null hypothesis and conclude that the RANDOM statement is needed. The note indicates that the  $p$ -value is computed based on a mixture of chi-squares.

What would happen if you incorrectly specified Teacher(Material) as a fixed effect?

```
title 'Random Effect is Incorrectly Specified as Fixed Effect';
proc glimmix data=STAT2.scores;
  class material teacher;
  model score=material teacher(material);
  output out=checkvar variance=ResidualVariance;
run;

proc print data=checkvar (obs=1);
  var ResidualVariance;
run;                                         *ST206d03.sas;
```

The OUTPUT statement requests that the residual variance estimate be output to the **checkvar** data set, which is then printed.

PROC PRINT Output

#### Random Effect is Incorrectly Specified as Fixed Effect

Obs	ResidualVariance
1	17.255

The estimate for the Residual variance is the same as the one obtained from the previous model. This is because you have balanced data.

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Material	3	100	6.99	0.0003
Teacher(Material)	16	100	13.05	<.0001

Now, the  $F$  value for the **Material** effect is 6.99 with the denominator degrees of freedom 100. The  $p$ -value is 0.0003. You would incorrectly conclude that there is a significant difference in the average test scores among the four teaching materials.

## 6.08 Poll

Incorrectly specifying a random effect as a fixed effect can affect the inference about the treatment effects.

- True
- False



## Exercises

---

### 1. Analyzing Nested Data Using the MIXED Procedure

Recall that data were collected by a school district to assess the reading skill progress of students in their first year of formal schooling. The data are stored in the **STAT2.school** data set. In a previous chapter, you used ANOVA models to evaluate the significance of some factors in explaining the difference in the average **Reading3** test scores. There is another variable, **Teacher**, in the data set. Teachers are nested within schools. Factors of interest are **School** and **Gender**.

- a. Generate PROC PRINT output for the **STAT2.school** data set. Use PROC FREQ to examine whether the data are balanced or not for **School** and **Gender** combinations.
- b. Use PROC SGPANEL to generate side-by-side box plots of **Reading3** versus **School** and **Gender** combinations. (Hint: Put both **School** and **Gender** in the PANELBY statement.) Use a WHERE statement to make sure that the graph uses only the data values for observations in which all of **Reading3**, **Teacher**, **School**, and **Gender** are present. Why do you see graphs only for three out of four schools? Are there any groups that seem to be different from other groups?
- c. Are teachers crossed or nested within schools? Should you consider **Teacher** as a fixed effect or a random effect?
- d. Use the GLIMMIX procedure to determine whether there is a difference in the mean **Reading3** test scores among **School**, **Gender**, and **School\*Gender**. What are the estimates of the variance components? What do you conclude about the fixed effects?
- e. Use an LSMESTIMATE statement to estimate the difference in **Reading3** test scores between *Cottonwood* girls and all other students. What do you conclude?
- f. What happens if you incorrectly specify the random effect as a fixed effect?

## 6.09 Multiple Choice Poll

Which of the following is *true*?

- Specifying **Teacher(School)** as a random effect enables you to estimate the variance in **Reading3** scores among the teachers.
- Specifying **Teacher(School)** as a random effect enables you to apply your conclusions about the material effect over a population of teachers.
- both a and b
- none of the above

45

## Course Summary

Dependent Variables	Independent Variables			
	Continuous	Discrete	Continuous and Discrete	If Random Effects Are Present
Continuous	Linear Regression Models (GLMSELECT, REG)	ANOVA Models (GLM)	Linear Models (GLM)	Linear Mixed Models (GLIMMIX)
Discrete	Generalized Linear Models (GENMOD or GLIMMIX)	Generalized Linear Models (GENMOD or GLIMMIX)	Generalized Linear Models (GENMOD or GLIMMIX)	Generalized Linear Mixed Models (GLIMMIX)

47

# 6.3 Solutions

---

## Solutions to Exercises

### 1. Analyzing Nested Data Using the MIXED Procedure

- a. Generate listing output for the STAT2.school data set.

```
ods html close;
ods listing;
options formdlim="_" ;

proc print data=STAT2.school(obs=25) ;
run; *ST206s01.sas;
```

Partial PROC PRINT Output

L	P	P	P	R	F	F	R		S	e
e	h	h	h	e	l	l	e		T	m
t	o	W	o	W	a	u	u	a	G	
t	n	o	n	o	d	e	e	d	e	
e	i	r	i	r	i	n	n	i	n	
O	r	c	d	c	d	c	c	n	d	A
b	s	s	s	s	s	g	y	g	e	g
s	1	1	1	2	2	3	3	3	r	e
				2	2	3	3	r	D	1
1	15	36	2	74	46	47	77	9	7	29
2	8	7	1	12	23	52	38	1	0	8
3	3	1	0	35	1	10	51	0	0	0
4	39	13	13	40	16	63	37	7	9	28
5	55	17	22	46	43	47	67	18	12	50
6	36	17	16	50	37	38	38	12	11	39
7	24	22	10	58	45	49	54	9	8	35
8	55	26	28	53	51	51	66	46	55	85
9	55	19	19	57	46	56	55	22	22	59
10	59	19	92	57	81	41	119	79	82	119
11	50	29	37	56	55	55	74	26	26	66
12	24	9	7	48	42	31	49	10	10	35
13	33	28	31	51	61	40	96	46	57	84
14	44	41	21	52	57	58	61	50	52	61
15	41	24	12	56	44	48	70	13	9	61
16	52	25	35	58	67	51	74	39	45	98
17	24	4	7	40	44	47	48	7	5	48
18	34	6	2	54	44	54	69	11	7	60
19	18	6	6	28	26	46	76	7	5	66
20	51	18	27	71	83	67	92	19	17	23
21	.	.	.	.	32	29	.	.	11	9
22	.	.	.	48	45	55	53	11	9	30
23	.	.	.	44	43	.	.	19	18	.
24	.	.	.	45	55	41	27	8	7	26
25	.	.	.	59	39	55	56	13	11	45

Use PROC FREQ to examine whether the data are balanced or not for **School** and **Gender** combinations.

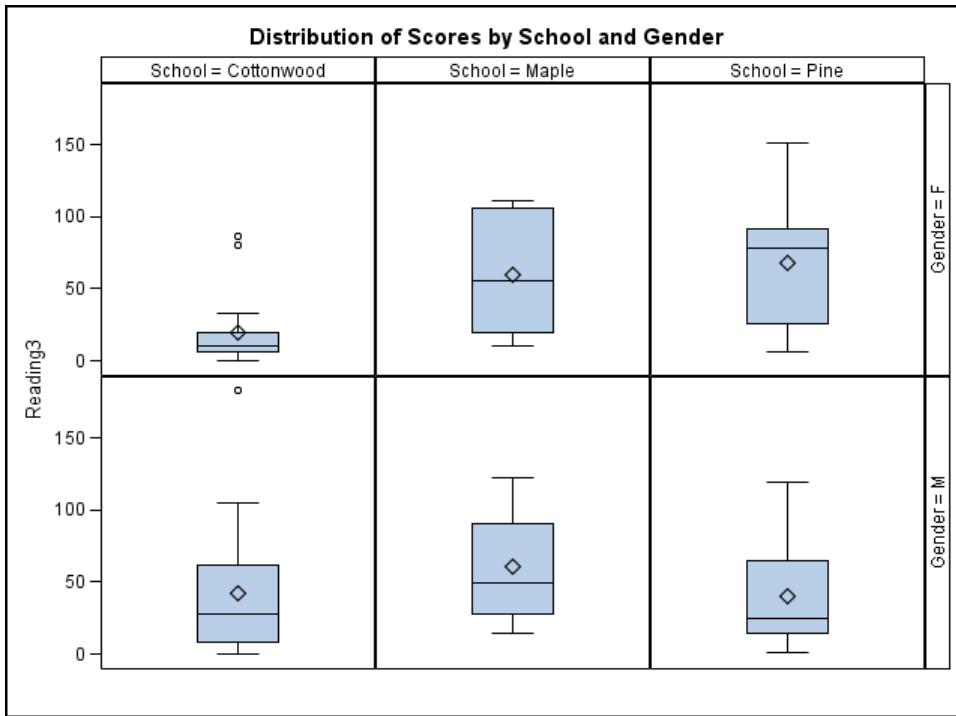
```
proc freq data=STAT2.school;
  tables school*gender / nocol norow nopercnt;
run; *ST206s01.sas;
```

The FREQ Procedure			
Table of School by Gender			
School	Gender		
Frequency	F	M	Total
Cottonwood	18	25	43
Dogwood	26	23	49
Maple	14	17	31
Pine	25	42	67
Total	83	107	190

The data are unbalanced.

- b. Use PROC SGpanel to create paneled box plots of **Reading3** versus **School** and **Gender** combinations.

```
title1 "Distribution of Scores by School and Gender";
proc sgpanel data=STAT2.school;
  where reading3 is not missing
    and teacher is not missing
    and school is not missing
    and gender is not missing;
  panelby school gender / layout=lattice;
  vbox reading3;
run; *ST206s01.sas;
```



You do not see data from *Dogwood* because all values for **Teacher** are missing for this school. Data from *Dogwood* are not included in the analysis.

The test scores for female students at *Cottonwood* seem to be different from other groups.

- c. Teachers are nested within schools. **Teacher(School)** should be considered a random effect.
- d. Use the GLIMMIX procedure to determine whether there is a difference in the mean **Reading3** test scores among **School**, **Gender**, and **School\*Gender**.

```
title;
proc glimmix data=STAT2.school;
  class school gender teacher;
  model reading3=school gender school*gender;
  random teacher(school);
run;                                *ST206s01.sas;
```

Partial PROC MIXED Output

Class Level Information		
Class	Levels	Values
School	3	Cottonwood Maple Pine
Gender	2	F M
Teacher	8	Miss Apple Miss Jones Miss Peters Mr. Johnson Mr. Rogers Mrs. King Mrs. Scott Ms. Chapman
Number of Observations Read		190
Number of Observations Used		134

Dimensions			
G-side Cov. Parameters	1		
R-side Cov. Parameters	1		
Columns in X	12		
Columns in Z	8		
Subjects (Blocks in V)	1		
Max Obs per Subject	134		
Covariance Parameter Estimates			
Cov Parm	Estimate	Standard Error	
Teacher(School)	51.9739	111.86	
Residual	1445.28	185.48	

The estimated teacher variance is 51.97 and the estimated residual variance is 1445.28.

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
School	2	5	3.87	0.0964
Gender	1	123	0.07	0.7853
School*Gender	2	123	5.16	0.0070

The *F* value for **School\*Gender** is 5.16 and the *p*-value is 0.0070. Averaged across all teachers, there is a significant effect of **School\*Gender** on the average **Reading3** test scores. Interpreting main effects in the presence of a significant interaction might be misleading. Instead, simple effects can be explored by using the LSMEANS statement with the SLICE option. However, for comparison sake, notice that the *p*-values for both the **School** and **Gender** effects are higher than 0.05.

- e. Use an LSMESTIMATE statement to estimate the difference in **Reading3** test scores between *Cottonwood* girls and all other students.

```
proc glimmix data=STAT2.school;
  class school gender teacher;
  model reading3=school gender school*gender;
  random teacher(school);
  lsmeans school*gender 'Cottonwood Girls vs. All Others'
    5 -1 -1 -1 -1 -1 / divisor=5 elsm;
run; *ST206s01.sas;
```

Because *Dogwood* is missing the value of **Teacher** for all observations, only six of the eight **School** by **Gender** by **Teacher** combinations have estimates. Thus, you need only six coefficients in the LSMESTIMATE statement. The ELSM option shows you how the coefficients are applied to the least squares means, after the DIVISOR=5 option is applied.

Least Squares Means Estimate Coefficients			
Effect	School	Gender	Row1
School*Gender	Cottonwood	F	1
School*Gender	Cottonwood	M	-0.2
School*Gender	Maple	F	-0.2
School*Gender	Maple	M	-0.2
School*Gender	Pine	F	-0.2
School*Gender	Pine	M	-0.2

## Partial PROC GLIMMIX Output

Least Squares Means Estimates						
Effect	Label	Standard				
		Estimate	Error	DF	t Value	Pr >  t
School*Gender	Cottonwood Girls vs. All Others	-34.0590	10.9071	123	-3.12	0.0022

The average **Reading3** test score for *Cottonwood* girls is 34.06 below the average scores for all other students. The standard error for the mean difference is 10.91. The *p*-value for the difference is 0.0022. Presuming an alpha level of 0.05, you would reject the null hypothesis and conclude that, on average, *Cottonwood* girls score significantly lower than all other students.

- f. What happens if you incorrectly specify the random effect as a fixed effect?

```
proc glimmix data=STAT2.school;
  class school gender teacher;
  model reading3=school gender school*gender teacher(school);
  output out=checkvar variance=ResidualVariance;
run;

proc print data=checkvar(obs=1);
  var ResidualVariance;
title 'Check Residual Variance';
run; title; *ST206s01.sas;
```

## PROC PRINT Output

Check Residual Variance	
Obs	Residual Variance
1	1437.06

The estimated residual variance is 1437.06. This is slightly different from the estimate from the previous model. Recall that you do not have balanced data.

## Partial PROC GLIMMIX Output

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
School	2	123	5.20	0.0068
Gender	1	123	0.20	0.6564
School*Gender	2	123	4.36	0.0148
Teacher(School)	5	123	1.65	0.1518

The *F* value for **School\*Gender** is 4.36 and the *p*-value is 0.0148. There is a significant effect of **School\*Gender** on the average **Reading3** test scores at a significance level of 0.05.

Interpreting the main effects of **School** and **Gender** in the presence of the significant interaction might be misleading. However, for comparison sake, notice that the *F* value for **School** is 5.20, the denominator degree of freedom for **School** is 123 (compared with 5 from the previous model), and the *p*-value is 0.0068. The *F* values and *p*-values for the effects are different from the ones obtained from the previous model. Specifying a random effect incorrectly as a fixed effect jeopardizes your conclusions about the treatment effects.

## Solutions to Student Activities (Polls/Quizzes)

### 6.01 Quiz – Correct Answer

Three growing methods were studied to evaluate which one is more effective for growing grass. Five varieties of grass seeds were randomly selected from a large population of varieties. Each method was applied to each of the varieties. Dry yield is measured at the end of the season. Is variety fixed or random?

Random

10

### 6.02 Poll – Correct Answer

The variables **State** and **City** are both random effects and **Type** is a fixed effect.

- True
- False

13

## 6.03 Multiple Choice Poll – Correct Answer

Which of the following are *true* about linear mixed models and the GLIMMIX procedure?

- a. Linear mixed models are a special case of general linear models that you fit using PROC GLM, PROC GLMSELECT, or PROC REG.
- b. Linear mixed models can handle normal and nonnormal responses.
- c. **PROC GLIMMIX has a RANDOM statement to model random effects.**
- d. PROC GLIMMIX can be used only to model random effects, and should not be used to model repeated measures data.

20

## 6.04 Poll – Correct Answer

To evaluate the productivity of two machines, five operators were chosen to operate on the machines. Each operator operated on each of the two machines three times. Operators and machines are crossed.

- True
- False

28

## 6.05 Poll – Correct Answer

If, for each machine, five operators were selected and operated only on that particular machine, then operators are nested with machines.

- True
- False

30

## 6.06 Multiple Choice Poll – Correct Answer

Five lots were chosen for the study. Four wafers were selected from each lot. The thickness of dioxide layers was measured from each wafer.

- a. Wafers are nested within lots.
- b. Wafers are crossed with lots.
- c. I do not know how wafers are related to lots.

32

## 6.08 Poll – Correct Answer

Incorrectly specifying a random effect as a fixed effect can affect the inference about the treatment effects.

- True
- False

42

## 6.09 Multiple Choice Poll – Correct Answer

Which of the following is **true**?

- a. Specifying **Teacher(School)** as a random effect enables you to estimate the variance in **Reading3** scores among the teachers.
- b. Specifying **Teacher(School)** as a random effect enables you to apply your conclusions about the material effect over a population of teachers.
- c. both a and b
- d. none of the above

46



# 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, Inc.
- Akaike, H. 1969. "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21, 243–247.
- Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle," *Proceedings of the 2<sup>nd</sup> International Symposium on Information Theory*. Editors: Petrov and Csaki. 267-281.
- Allison, P. D. 1991. *Logistic Regression Using the SAS® System, Theory and Application*, Cary, NC: SAS Institute Inc.
- Allison, P. D. 2012. *Logistic Regression Using SAS, Theory and Application Second Edition*, Cary, NC: SAS Institute Inc.
- Bates, D. M. and Watts, D. G. 1988. *Nonlinear Regression Analysis and Its Applications*, New York: John Wiley & Sons, Inc.
- Bowerman, B. L., O'Connell, R. T., and Dickey, D. A. 1986. *Linear Statistical Models, an Applied Approach*, Boston, MA: Duxbury Press.
- Box, G. E. 1953. "Non-normality and Tests on Variance," *Biometrika*, 40, 318 - 335.
- Box, G. E. 1954. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, II. Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification," *Annals of Mathematical Statistics*, 25, 484 - 498.
- Box, G. E. P. and Cox, D. R. 1964. "An Analysis of Transformations," *Journal of the Royal Statistics Society*, B-26, 211-252.
- Brocklebank, J. C. and Dickey, D. A. 2003. *SAS System for Forecasting Time Series*, Second Edition, Cary, NC: SAS Institute Inc.
- Brown, M. B. and Forsythe, A. B. 1974. "Robust Tests for Equality of Variances," *Journal of the American Statistical Association*, 69, 364 - 367.
- Carroll, R. J. and Ruppert, D. 1988. *Transformation and Weighting in Regression*, New York: Chapman and Hall.
- Casella, G. and Berger, R.. 1990. *Statistical Inference*, California: Wadsworth, Inc.
- Cochran, W. G. and Cox, G. M. 1957. *Experimental Designs*, New York: John Wiley & Sons, Inc.
- Conover, W. J., Johnson, M. E., and Johnson, M. M. 1981. "A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data," *Technometrics*, 23, 351 - 361.
- Craven, P. and Wahba, G. 1979. "Smoothing noisy data with spline functions," *Numer. Math.*, 31, 377-403.
- D'Agostino, R. B. and Stephens, M. A. 1996. *Goodness-of-fit Techniques*, New York: Marcel Dekker, Inc.
- Draper, N. R. and Smith, H. 1981. *Applied Regression Analysis*, Second Edition, New York: John Wiley & Sons, Inc.

- Efron, B., Hastie, T. J., Johnstone, I. M., and Tibshirani, R. 2004. "Least Angle Regression with Discussion." *Annals of Statistics*, 32, 407–499.
- Freund, R. J. and Littell, R. C. 2000. *SAS System for Regression, 3rd Edition*, Cary, NC: SAS Institute, Inc.
- Freund, R. J. and Wilson, W. J. 1998. *Regression Analysis*, San Diego: Academic Press.
- Fuller, W. A. 1978. *Introduction to Statistical Time Series*, New York: John Wiley & Sons, Inc.
- Griepentrog, G. L., Ryan, J. M., and Smith, L. D. 1982. "Linear Transformation of Polynomial Regression Models," *The American Statistician*, Vol. 36, No. 3, Part 1.
- Harrell, F. E. 2001. *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis*, New York: Springer-Verlag.
- Hastie and Tibshirani. 1990. *Generalized Additive Models*, New York: Chapman and Hall.
- Hocking, R. R. 1984. *The Analysis of Linear Models*, Monterey, CA: Brooks-Cole Publishing Co.
- Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. 1998. "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion," *Journal of the Royal Statistical Society B* 60, 271-293.
- Hurvich, C. M. and Tsai, C.-L. 1989. "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.
- Iman, R. 1988. "The Analysis of Complete Blocks Using Methods Based on Ranks," *Proceedings of the SAS Users Group International Conference, Volume 13*, 970-978, Cary, NC: SAS Institute, Inc.
- Iman, R. 1982. "Some Aspects of the Rank Transform in Analysis of Variance Problems," *Proceedings of the SAS Users Group International Conference, Volume 7*, 676-680, Cary, NC: SAS Institute, Inc.
- John, P. W. M. 1971. *Statistical Design and Analysis of Experiments*, New York: The Macmillan Company.
- Judge, G. G., Griffiths, W. E., Hill, R. C., and Lee, T.-C. 1980. *The Theory and Practice of Econometrics*, New York: John Wiley & Sons.
- Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. 1988. *Applied Regression Analysis and Other Multivariable Methods*, Boston, MA: PWS-Kent Publishing Company.
- Kotz, Samuel, Johnson, N. L., and Read, C. B., eds. 1988. *Encyclopedia of Statistical Sciences*, New York: John Wiley & Sons.
- Levene, H. 1960. "Robust Tests for the Equality of Variance," in I. Olkin, ed., "Contributions to Probability and Statistics," 278--292, Palo Alto, CA: Stanford University Press.
- Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. 1996. *SAS® System for Mixed Models*, Cary, NC: SAS Institute Inc.
- Littell, R. C., Stroup, W. W., and Freund, R. J. 2002. *SAS® System for Linear Models*, Cary, NC: SAS Institute Inc.
- 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.

- McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models*, Second Edition, London: Chapman and Hall.
- Myers, R. H. 1988. *Response Surface Methodology*, Virginia Polytechnic and State University.
- Myers, R. H. 1990. *Classical and Modern Regression with Applications*, Second Edition, Boston: PWS and Kent Publishing Company, Inc.
- Miller, D. M. 1984. "Reducing Transformation Bias in Curve Fitting," *The American Statistician*, 38, 2, 124-126.
- Miller, R. G., Jr. 1997. *Beyond ANOVA Basics of Applied Statistics*, Boca Raton, FL: Chapman & Hall/CRC.
- Milliken, G. A. and Johnson, D. E. 1992. *Analysis of Messy Data, Volume 1: Designed Experiments*, New York: Chapman & Hall.
- Montgomery, D. C. and Peck, E. A. 1982. *Introduction to Linear Regression Analysis*, New York: John Wiley & Sons, Inc.
- Nelder, J. A. and Wedderburn, R. W. M. 1972. "Generalized Linear Models," *Journal of the Royal Statistical Society A*, 135, 370-384.
- Neter, J., Kutner, M. H., Wasserman, W., and Nachtsheim, C. J. 1996. *Applied Linear Statistical Models*, Fourth Edition, New York: WCB McGraw Hill.
- Neter, J., Wasserman, W., and Kutner, M. 1990. *Applied Linear Statistical Models*, Third Edition, Richard D. Irwin Inc.
- O'Brien, R. G. 1979. "A General ANOVA Method for Robust Tests of Additive Models for Variances," *Journal of the American Statistical Association*, 74, 877 - 880.
- Olejnik, S. F. and Algina, J. 1987. "Type I Error Rates and Power Estimates of Selected Parametric and Non-parametric Tests of Scale," *Journal of Educational Statistics*, 12, 45 - 61.
- Peixoto, J. L. 1990. "A Property of Well-Formulated Polynomial regression Models," *The American Statistician*, Vol. 44, No. 1.
- Peixoto, J. L. 1987. "Hierarchical Variable Selection in Polynomial Regression Models," *The American Statistician*, Vol. 41, No. 4.
- Ratkowsky, D. 1990. *Handbook of Nonlinear Regression Models*, Marcel Dekker: New York and Basel.
- Rawlings, J. O., Pantula, S. G., and Dickey, D. A. 1998. *Applied Regression Analysis: A Research Tool*, Second Edition, New York: Springer.
- SAS Institute Inc. 2009. *JMP® 8: Statistics and Graphics Guide*, 2<sup>nd</sup> Edition. Cary, NC: SAS Institute, Inc.
- SAS Institute Inc. 2011. *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute, Inc.
- Schabenberger, O. and Pierce, F. J. 2002. *Contemporary Statistical Models for the Plant and Soil Sciences*, Boca Raton, FL: CRC Press LLC.
- Searle, S. R. 1987. *Linear Models for Unbalanced Data*, New York: John Wiley & Sons, Inc.

- Seber, G. A. F. and Wild, C. J. 1989. *Nonlinear Regression*, New York: John Wiley & Sons, Inc.
- Stokes, M. E., Davis, C. S., and Koch, G. G. 2000. *Categorical Data Analysis Using the SAS® System, 2nd Edition*, Cary, NC: SAS Institute Inc.
- Thode, Henry C. Jr. 2002. *Testing for Normality*, New York: Marcel Dekker, Inc.
- Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
- Welch, B. L. 1951. “On the Comparison of Several Mean Values: An Alternative Approach,” *Biometrika*, 38, 330-336.
- White, H. 1980. “A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity,” *Econometrics*, 48, 817 - 838.
- Zar, J. H. 1996. *Biostatistical Analysis, Third Edition*, New York: Prentice-Hall, Inc.
- Zou, H. and Hastie, T. 2005. “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

# Appendix B A Brief Review of Matrix Algebra

<b>B.1 Introducing Matrices.....</b>	<b>B-3</b>
<b>B.2 Basic Operations .....</b>	<b>B-3</b>
<b>B.3 Some Special Matrices .....</b>	<b>B-6</b>
<b>B.4 Determinant.....</b>	<b>B-10</b>
<b>B.5 Inverse Matrices.....</b>	<b>B-11</b>
<b>B.6 Linear Dependence and Rank.....</b>	<b>B-12</b>
<b>B.7 Generalized Inverse.....</b>	<b>B-14</b>
<b>B.8 Eigenvalues and Eigenvectors .....</b>	<b>B-15</b>
<b>B.9 Cholesky Root.....</b>	<b>B-17</b>



## B.1 Introducing Matrices

---

A *matrix* is a rectangular or square array of values arranged in rows and columns. An  $m \times n$  matrix  $\mathbf{A}$  has  $m$  rows and  $n$  columns, and has a general form of the following:

$$\mathbf{A}_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}_{m \times n}$$

The element  $a_{ij}$  denotes the element in the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column in matrix  $\mathbf{A}$ .

Example:

$$\mathbf{A} = \begin{bmatrix} 3 & 5 & 8 \\ 1 & 2 & 2 \end{bmatrix} \text{ is a } 2 \times 3 \text{ matrix; } \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 3 & 6 \\ 7 & 2 \\ 2 & 9 \end{bmatrix} \text{ is a } 4 \times 2 \text{ matrix.}$$

A *vector* is a matrix with only one column (a *column vector*) or only one row (a *row vector*). For example,

$$\mathbf{x} = \begin{bmatrix} 3 \\ -2 \\ 1 \\ 5 \end{bmatrix} \text{ is a column vector,}$$

and  $\mathbf{x}' = [3 \ -2 \ 1 \ 5]$  and  $\mathbf{y}' = [4 \ 7 \ -5]$  are row vectors.

A single number such as 2.4 or -6 is called a *scalar*. The elements of a matrix are usually scalars, although a matrix can be expressed as a matrix of smaller matrices.

## B.2 Basic Operations

---

### Transpose

The *transpose* of a matrix  $\mathbf{A}$  is the matrix whose columns are rows of  $\mathbf{A}$  (and therefore, whose rows are columns of  $\mathbf{A}$ ), with the order retained, from first to last. The transpose of  $\mathbf{A}$  is denoted by  $\mathbf{A}'$ .

Example:

$$\text{Let } \mathbf{A} = \begin{bmatrix} 3 & 2 & -6 \\ 7 & 1 & 2 \end{bmatrix}, \text{ then } \mathbf{A}' = \begin{bmatrix} 3 & 7 \\ 2 & 1 \\ -6 & 2 \end{bmatrix}.$$

You see that if  $\mathbf{A}$  is  $2 \times 3$ , then  $\mathbf{A}'$  is  $3 \times 2$ . In general, if  $\mathbf{A}$  is  $m \times n$ , then  $\mathbf{A}'$  is  $n \times m$ , and  $a'_{ij} = a_{ji}$ .

The transpose of a row vector is a column vector.

### Partitioned Matrices

The matrix  $\mathbf{A}$  can be written as a matrix of matrices:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

This specification of  $\mathbf{A}$  is called a *partitioning* of  $\mathbf{A}$ , and the matrices  $\mathbf{A}_{11}$ ,  $\mathbf{A}_{12}$ ,  $\mathbf{A}_{21}$ , and  $\mathbf{A}_{22}$  are said to be submatrices of  $\mathbf{A}$ .  $\mathbf{A}$  is called a *partitioned matrix*.

Example:

$$\text{Let } \mathbf{A} = \begin{bmatrix} 1 & 6 & 9 & 4 & 5 & 8 \\ 5 & 4 & 7 & 2 & 0 & 2 \\ 9 & 2 & 8 & 1 & 7 & 1 \\ 9 & 1 & 7 & 6 & 2 & 3 \\ 2 & 5 & 4 & 8 & 1 & 7 \end{bmatrix}.$$

Each of the arrays of numbers in the four sections of  $\mathbf{A}$  engendered by the dashed lines is a matrix:

$$\mathbf{A}_{11} = \begin{bmatrix} 1 & 6 & 9 & 4 \\ 5 & 4 & 7 & 2 \\ 9 & 2 & 8 & 1 \end{bmatrix} \quad \mathbf{A}_{12} = \begin{bmatrix} 5 & 8 \\ 0 & 2 \\ 7 & 1 \end{bmatrix}$$

$$\mathbf{A}_{21} = \begin{bmatrix} 9 & 1 & 7 & 6 \\ 2 & 5 & 4 & 8 \end{bmatrix} \quad \mathbf{A}_{22} = \begin{bmatrix} 2 & 3 \\ 1 & 7 \end{bmatrix}$$

### Trace

The sum of the diagonal elements of a square matrix is called the *trace* of the matrix, that is,

$$\text{tr}(\mathbf{A}) = a_{11} + a_{22} + \dots + a_{nn} = \sum_{i=1}^n a_{ii}.$$

Example:

$$\text{Let } \mathbf{A} = \begin{bmatrix} 1 & 5 & 9 \\ -3 & 2 & 8 \\ 4 & 7 & 6 \end{bmatrix}. \text{ Then, } \text{tr}(\mathbf{A}) = 1 + 2 + 6 = 9.$$

### Addition and Subtraction

Matrices of the same size are added or subtracted by adding or subtracting corresponding elements.

Example:

Let  $\mathbf{A} = \begin{bmatrix} 2 & 4 & 8 \\ 9 & -1 & 5 \end{bmatrix}$ , and  $\mathbf{B} = \begin{bmatrix} 1 & 5 & 3 \\ 2 & 3 & -7 \end{bmatrix}$ . Then,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2+1 & 4+5 & 8+3 \\ 9+2 & (-1)+3 & 5+(-7) \end{bmatrix} = \begin{bmatrix} 3 & 9 & 11 \\ 11 & 2 & -2 \end{bmatrix} \text{ and}$$

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} 2-1 & 4-5 & 8-3 \\ 9-2 & -1-3 & 5-(-7) \end{bmatrix} = \begin{bmatrix} -1 & -1 & 5 \\ 7 & -4 & 12 \end{bmatrix}.$$

### Multiplication

The *inner product* of two vectors  $\mathbf{a}' = [a_1 \ a_2 \ \dots \ a_n]$  and  $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$  is defined as

$$\mathbf{a}'\mathbf{x} = a_1x_1 + a_2x_2 + \dots + a_nx_n = \sum_{i=1}^n a_i x_i.$$

Example:

Let  $\mathbf{a}' = [3 \ 1 \ 10]$  and  $\mathbf{x} = \begin{bmatrix} 5 \\ 10 \\ 3 \end{bmatrix}$ . Then, the inner product  $\mathbf{a}'\mathbf{x} = 3(5) + 1(10) + 10(3) = 55$ .

The product  $\mathbf{AB}$  of two matrices  $\mathbf{A}$  and  $\mathbf{B}$  is defined and therefore exists only if the number of columns in  $\mathbf{A}$  equals the number of rows in  $\mathbf{B}$ .  $\mathbf{AB}$  has the same number of rows as  $\mathbf{A}$  and the same number of columns as  $\mathbf{B}$ . The  $ij^{\text{th}}$  element of  $\mathbf{AB}$  is the inner product of the  $i^{\text{th}}$  row of  $\mathbf{A}$  and  $j^{\text{th}}$  column of  $\mathbf{B}$ .

Example:

Let  $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 4 & 2 \end{bmatrix}$  and  $\mathbf{B} = \begin{bmatrix} 0 & 6 & 1 & 5 \\ 2 & 1 & -2 & 3 \\ 4 & 1 & 2 & 5 \end{bmatrix}$ . Then,

$\mathbf{AB} =$

$$\begin{bmatrix} 1(0)+2(2)+3(4) & 1(6)+2(1)+3(1) & 1(1)+2(-2)+3(2) & 1(5)+2(3)+3(5) \\ -1(0)+4(2)+2(4) & -1(6)+4(1)+2(1) & -1(1)+4(-2)+2(2) & -1(5)+4(3)+2(5) \end{bmatrix} = \begin{bmatrix} 16 & 11 & 3 & 26 \\ 16 & 0 & -5 & 17 \end{bmatrix}.$$

### Direct or Kronecker Product $\otimes$

Suppose  $\mathbf{A}$  is  $m \times n$  and  $\mathbf{B}$  is  $p \times q$ . Then the *direct* or *Kronecker product*  $\mathbf{A} \otimes \mathbf{B}$  is of size  $mp \times nq$  and is most easily described as the partitioned matrix:

$$\mathbf{A}_{m \times n} \otimes \mathbf{B}_{p \times q} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \dots & \dots & \dots & \dots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

Example:

Let  $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}$  and  $\mathbf{B} = \begin{bmatrix} 1 & -1 & 1 & 2 \\ 3 & 2 & 0 & 1 \\ -1 & 0 & 2 & 3 \end{bmatrix}$ . Then,

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} 1 & -1 & 1 & 2 & 2 & -2 & 2 & 4 \\ 3 & 2 & 0 & 1 & 6 & 4 & 0 & 2 \\ -1 & 0 & 2 & 3 & -2 & 0 & 4 & 6 \\ -1 & 1 & -1 & -2 & 3 & -3 & 3 & 6 \\ -3 & -2 & 0 & -1 & 9 & 6 & 0 & 3 \\ 1 & 0 & -2 & -3 & -3 & 0 & 6 & 9 \end{bmatrix}.$$

## B.3 Some Special Matrices

---

A *square* matrix is a matrix whose number of columns equals the number of rows. The elements on the diagonal,  $a_{11}, a_{22}, \dots, a_{nn}$ , are referred to as the *diagonal elements* or *diagonal* of the matrix. Elements of a square matrix other than the diagonal elements are called *off-diagonal* or *nondiagonal* elements.

A *diagonal* matrix is a square matrix having zero for all its nondiagonal elements.

Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$
 is a diagonal matrix.

A *triangular* matrix is a square matrix with all elements above (or below) the diagonal being zero.

Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 6 \\ 0 & -2 & 9 \\ 0 & 0 & 7 \end{bmatrix}$$
 and  $\mathbf{B} = \begin{bmatrix} 3 & 0 & 0 \\ -7 & 2 & 0 \\ 1 & -2 & 5 \end{bmatrix}$  are triangular matrices.

**A** is an *upper triangular matrix* and **B** is a *lower triangular matrix*.

A diagonal matrix having all diagonal elements equal to unity is called an *identity matrix*, or sometimes a *unit matrix*. It is usually denoted by the letter **I**. For example,

$$\mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ is an identity matrix of order 4.}$$

For any matrix **A** of order  $m \times n$ ,  $\mathbf{I}_m \mathbf{A}_{m \times n} = \mathbf{A}_{m \times n} \mathbf{I}_n = \mathbf{A}_{m \times n}$ .

A *symmetric matrix* is a square matrix when it equals its transpose.

Example:

$$\text{Let } \mathbf{A} = \begin{bmatrix} 1 & -2 & 3 \\ -2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}. \text{ Then, } \mathbf{A}' = \begin{bmatrix} 1 & -2 & 3 \\ -2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix} = \mathbf{A}.$$

The matrix **A** is symmetric. In symmetric matrices, the area above the diagonal is a mirror image of the area below the diagonal.

Vectors, whose every element is unity, are called *summing vectors* because they can be used to express a sum of numbers in matrix notation as an inner product.

Example:

**1'** = [1 1 1 1] is a summing vector of order 4. For **x'** = [2 4 -3 8],

$$\mathbf{1}' \mathbf{x} = [1 \ 1 \ 1 \ 1] \begin{bmatrix} 2 \\ 4 \\ -3 \\ 8 \end{bmatrix} = 2 + 4 - 3 + 8 = 11 = \mathbf{x}' \mathbf{1}.$$

It follows that  $\mathbf{1}'_n \mathbf{1}_n = n$  and  $\mathbf{1}_m \mathbf{1}'_n = \mathbf{J}_{m \times n}$ , a matrix having all elements unity.

A square **J** matrix is denoted by **J**<sub>n</sub>. **J**<sub>n</sub> = **1**<sub>n</sub>**1'**<sub>n</sub> and **J**<sub>n</sub><sup>2</sup> = n**J**<sub>n</sub>. A useful variant of **J**<sub>n</sub> is

$$\bar{\mathbf{J}}_n = \frac{1}{n} \mathbf{J}_n, \quad \bar{\mathbf{J}}_n^2 = \bar{\mathbf{J}}_n.$$

Therefore,  $\bar{\mathbf{J}}_n$  is an *idempotent matrix*, which satisfies  $\mathbf{K}^2 = \mathbf{K}$ .

**C**<sub>n</sub> = **I** -  $\bar{\mathbf{J}}_n$  = **I** -  $\frac{1}{n} \mathbf{J}_n$  is known as a *centering matrix*. It follows that

**C** = **C'** = **C**<sup>2</sup>, **C****1** = 0, and **CJ** = **J****C** = 0.

An *orthogonal matrix* **A** is a matrix having the property **AA'** = **I** = **A'A**.

Example:

$$\text{Let } \mathbf{A} = \begin{bmatrix} 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} & 0 \\ 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & -3/\sqrt{12} \end{bmatrix}.$$

You can verify that  $\mathbf{AA}' = \mathbf{I}$ . Therefore,  $\mathbf{A}$  is an orthogonal matrix. A matrix in the form shown above is called a *Helmert* matrix.

A *quadratic form* is the product of a row vector  $\mathbf{x}'$ , a matrix  $\mathbf{A}$ , and the column vector  $\mathbf{x}$ , that is,  $\mathbf{x}'\mathbf{Ax}$ . This is a quadratic function of the  $x$ s. Notice that to result in the same quadratic function of  $x$ s, you can use many different matrices. Each matrix has the same diagonal elements, and the sum of each pair of symmetrically placed off-diagonal elements  $a_{ij}$  and  $a_{ji}$  is the same.

For any particular quadratic form, there is a unique *symmetric matrix*  $\mathbf{A}$  for which the quadratic form can be expressed as  $\mathbf{x}'\mathbf{Ax}$ :

$$\mathbf{x}'\mathbf{Ax} = \sum_{i=1}^n x_i^2 a_{ii} + 2 \sum_{j=i+1}^n \sum_{i=1}^n x_i x_j a_{ij}$$

Example:

$$\mathbf{x}'\mathbf{Ax} = \mathbf{x}' \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \mathbf{x}$$

$$= a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + 2(a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3).$$

$$\text{If } \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 1 \\ 3 & 1 & 4 \end{bmatrix}, \text{ then}$$

$$\mathbf{x}'\mathbf{Ax} = [x_1 \ x_2 \ x_3] \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 1 \\ 3 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 + 5x_2^2 + 4x_3^2 + 4x_1x_2 + 6x_1x_3 + 2x_2x_3.$$

If  $\mathbf{x}'\mathbf{Ax} > 0$  for all  $\mathbf{x}$  other than  $\mathbf{x} = 0$ , then  $\mathbf{x}'\mathbf{Ax}$  is a *positive definite* quadratic form, and  $\mathbf{A} = \mathbf{A}'$  is correspondingly a *positive definite* (p.d.) matrix.

Example:

Let  $\mathbf{A} = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 5 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ . Then,

$$\begin{aligned}\mathbf{x}'\mathbf{Ax} &= [x_1 \ x_2 \ x_3] \begin{bmatrix} 2 & 2 & 1 \\ 2 & 5 & 1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= 2x_1^2 + 5x_2^2 + 2x_3^2 + 4x_1x_2 + 2x_1x_3 + 2x_2x_3 \\ &= (x_1 + 2x_2)^2 + (x_1 + x_3)^2 + (x_2 + x_3)^2 > 0, \text{ other than } \mathbf{x} = 0.\end{aligned}$$

The matrix  $\mathbf{A}$  is a positive definite matrix.

If  $\mathbf{x}'\mathbf{Ax} \geq 0$  for all  $\mathbf{x}$  and  $\mathbf{x}'\mathbf{Ax} = 0$  for some  $\mathbf{x} \neq 0$ , then  $\mathbf{x}'\mathbf{Ax}$  is a *positive semidefinite* quadratic form, and  $\mathbf{A} = \mathbf{A}'$  is correspondingly a *positive semidefinite (p.s.d.) matrix*. The two classes of matrices taken together, positive definite and positive semidefinite, are called *nonnegative definite (n.n.d.)*.

Example:

Let  $\mathbf{A} = \begin{bmatrix} 37 & -2 & -24 \\ -2 & 13 & -3 \\ -24 & -3 & 17 \end{bmatrix}$ . Then,

$$\begin{aligned}\mathbf{x}'\mathbf{Ax} &= [x_1 \ x_2 \ x_3] \begin{bmatrix} 37 & -2 & -24 \\ -2 & 13 & -3 \\ -24 & -3 & 17 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= 37x_1^2 + 13x_2^2 + 17x_3^2 - 4x_1x_2 - 48x_2x_3 - 6x_2x_3 \\ &= (x_1 - 2x_2)^2 + (6x_1 - 4x_3)^2 + (3x_2 - x_3)^2.\end{aligned}$$

This is zero for  $\mathbf{x}' = [2 \ 1 \ 3]$  and for any scalar multiple thereof, as well as for  $\mathbf{x} = 0$ . Then, the matrix  $\mathbf{A}$  is *positive semidefinite*.

Example:

$$\mathbf{x}'\mathbf{Cx} = \mathbf{x}'(\mathbf{I} - \bar{\mathbf{J}})\mathbf{x} = \sum_{i=1}^n (x_i - \bar{x})^2$$

is a positive semidefinite quadratic form because it is positive, except for being zero when all the  $x_i$ 's are equal. Its matrix,  $\mathbf{I} - \bar{\mathbf{J}}$ , which is idempotent, is also p.s.d., as are all symmetric idempotent matrices (except  $\mathbf{I}$ , which is the only p.d. idempotent matrix).

## B.4 Determinant

---

The *determinant* of a square matrix of order  $n$  (that is,  $\mathbf{A} = \{a_{ij}\}, i, j = 1, 2, \dots, n$ ) is the sum of all possible products of  $n$  elements of  $\mathbf{A}$  such that

1. each product has one and only one element from every row and column of  $\mathbf{A}$
2. the sign of a product being  $(-1)^p$  for  $p = \sum_{i=1}^n n_i$ , where by writing
  - a. the product with its  $i$  subscripts in natural order  $a_{1j_1} a_{2j_2} \cdots a_{ij_i} \cdots a_{nj_n}$
  - b. the  $j$  subscripts  $j_i, i = 1, 2, \dots, n$ , being the first  $n$  integers in some order  
 $n_i$  is defined as the number of  $j$ s less than  $j_i$  that follow  $j_i$  in this order.

Therefore, the determinant of a matrix  $\mathbf{A}$ , denoted by  $|\mathbf{A}|$ , is a polynomial of the elements of a square matrix. It is a scalar. It is the sum of certain products of the elements of the matrix from which it is derived, each product being multiplied by +1 or -1 according to certain rules.

Example:

$$|\mathbf{A}| = \begin{vmatrix} 3 & 7 \\ 2 & 6 \end{vmatrix} = 3(6) - 7(2) = 4$$

$$|\mathbf{A}| = \begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{vmatrix} = 1(+1) \begin{vmatrix} 5 & 6 \\ 8 & 10 \end{vmatrix} + 2(-1) \begin{vmatrix} 4 & 6 \\ 7 & 10 \end{vmatrix} + 3(+1) \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix}$$

$$= 1(50 - 48) - 2(40 - 42) + 3(32 - 35) = -3$$

The determinant that multiplies each element of the chosen row (in this case, the first row) is the determinant derived from  $|\mathbf{A}|$  by crossing out the row and column containing the element concerned.

For example, the first element, 1, is multiplied by the determinant  $\begin{vmatrix} 5 & 6 \\ 8 & 10 \end{vmatrix}$ , which is obtained from  $|\mathbf{A}|$

through crossing out the first row and column. Determinants obtained in this way are called *minors* of  $|\mathbf{A}|$ ,

that is,  $\begin{vmatrix} 5 & 6 \\ 8 & 10 \end{vmatrix}$  is the minor of the element 1 in  $|\mathbf{A}|$ , and  $\begin{vmatrix} 4 & 6 \\ 7 & 10 \end{vmatrix}$  is the minor of element 2.

The (+1) and (-1) factors in the expansion are decided on according to the following rule:

If  $\mathbf{A}$  is written in the form  $\mathbf{A} = \{a_{ij}\}$ , the product of  $a_{ij}$  and its minor in the expansion of determinant  $|\mathbf{A}|$  is multiplied by  $(-1)^{i+j}$ .

Therefore, because the element 1 in the example is the element  $a_{11}$ , its product with its minor is multiplied by  $(-1)^{1+1} = +1$ . For element 2, which is  $a_{12}$ , its product with its minor is multiplied by  $(-1)^{1+2} = -1$ .

Denote the minor of the element  $a_{ij}$  by  $|\mathbf{M}_{ij}|$ , where  $\mathbf{M}_{ij}$  is a submatrix of  $\mathbf{A}$  obtained by deleting the  $i$ th row and the  $j$ th column. The determinant of an  $n$ -order matrix is obtained by *expansion by the elements of a row (or column)*, or *expansion by minors*:

$$|\mathbf{A}| = \sum_{j=1}^n a_{ij} (-1)^{i+j} |\mathbf{M}_{ij}| \text{ for any row } i, \text{ or}$$

$$|\mathbf{A}| = \sum_{i=1}^n a_{ij} (-1)^{i+j} |\mathbf{M}_{ij}| \text{ for any column } j.$$

The signed minor  $(-1)^{i+j} |\mathbf{M}_{ij}|$  is called a *cofactor*:  $c_{ij} = (-1)^{i+j} |\mathbf{M}_{ij}|$ .

This expansion is used recurrently when  $n$  is large, that is, each  $|\mathbf{M}_{ij}|$  is expanded by the same procedure.

This method of evaluation requires lengthy computing for determinants of order exceeding 3 to 4. Fortunately, easier methods exist, but they are based on this expansion-by-minors method.

## B.5 Inverse Matrices

---

The *inverse* of a square matrix  $\mathbf{A}$  is a matrix whose product with  $\mathbf{A}$  is the identity matrix  $\mathbf{I}$ . The inverse matrix is denoted by  $\mathbf{A}^{-1}$ . The concept of “dividing” by  $\mathbf{A}$  in matrix algebra is replaced by the concept of multiplying by the inverse matrix  $\mathbf{A}^{-1}$ .

An inverse matrix  $\mathbf{A}^{-1}$  should have the following properties:

- $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$
- $\mathbf{A}^{-1}$  unique for given  $\mathbf{A}$

An *adjugate* (or *adjoint*) of matrix  $\mathbf{A}$ , denoted by  $\text{adj } \mathbf{A}$ , is obtained by replacing the elements in  $\mathbf{A}$  by their cofactors and then transposing it.

The inverse of matrix  $\mathbf{A}$ ,  $\mathbf{A}^{-1}$ , can be described as the adjugate of  $\mathbf{A}$  multiplied by the scalar  $1/|\mathbf{A}|$ :

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \text{adj } \mathbf{A}$$

where  $\text{adj } \mathbf{A}$  is the adjugate (or adjoint) matrix of  $\mathbf{A}$ . Therefore,

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \left[ \begin{array}{c} \mathbf{A} \text{ with every element} \\ \text{replaced by its cofactor} \end{array} \right]^{\text{transposed}}$$

Example:

Let  $\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 3 & 9 \end{bmatrix}$ . The determinant of  $\mathbf{A}$  is  $|\mathbf{A}| = \begin{vmatrix} 2 & 5 \\ 3 & 9 \end{vmatrix} = 18 - 15 = 3$ .

- The cofactor for  $a_{11} = 2$  is  $(-1)^{1+1} |9| = 9$ .
- The cofactor for  $a_{12} = 5$  is  $(-1)^{1+2} |3| = -3$ .
- The cofactor for  $a_{21} = 3$  is  $(-1)^{2+1} |5| = -5$ .
- The cofactor for  $a_{22} = 9$  is  $(-1)^{2+2} |2| = 2$ .

The adjugate matrix is  $\begin{bmatrix} 9 & -5 \\ -3 & 2 \end{bmatrix}$ , so the inverse is the following:

$$\mathbf{A}^{-1} = \frac{1}{3} \begin{bmatrix} 9 & -5 \\ -3 & 2 \end{bmatrix}$$

Conditions for existence of the inverse are as follows:

1.  $\mathbf{A}^{-1}$  can exist only when  $\mathbf{A}$  is square.
2.  $\mathbf{A}^{-1}$  does exist only if  $|\mathbf{A}|$  is nonzero.

A square matrix is said to be *singular* when its determinant is zero and *nonsingular* when its determinant is not zero.

Several computing procedures for inverting matrices are based on solving linear equations by successive elimination and backward substitution. The basic idea is as follows:

To solve the equation  $\mathbf{Ax} = \mathbf{b}$ , write the following matrix  $[\mathbf{A} \ \mathbf{I}]$ , where  $\mathbf{I}$  is the identity matrix of the same order as  $\mathbf{A}$ . Perform row operations to this matrix to make the left submatrix become  $\mathbf{I}$ , whereupon the right submatrix is  $\mathbf{A}^{-1}$ . Thus, starting with  $[\mathbf{A} \ \mathbf{I}]$ , you now have  $[\mathbf{I} \ \mathbf{A}^{-1}]$ , providing  $\mathbf{A}^{-1}$  exists.

 Detailed information about row operations can be found in any matrix algebra textbook.

## B.6 Linear Dependence and Rank

---

Define  $\mathbf{X}$  as the matrix having columns  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ , and  $\mathbf{a}$  as the vector of *as*:

$\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_n]$  and  $\mathbf{a}' = [a_1 \ a_2 \ \dots \ a_n]$ .

Then, the *linear combination* of the set of  $n$  vectors is the following:

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_n\mathbf{x}_n = \sum_{i=1}^n a_i\mathbf{x}_i = \mathbf{X}\mathbf{a}$$

You see that

- $\mathbf{X}\mathbf{a}$  is a column vector, a linear combination of the columns of  $\mathbf{X}$ .
- $\mathbf{b}'\mathbf{X}$  is a row vector, a linear combination of the rows of  $\mathbf{X}$ .
- $\mathbf{AB}$  is a matrix. Its rows are linear combinations of the rows of  $\mathbf{B}$ , and its columns are linear combinations of the columns of  $\mathbf{A}$ .

The vector  $\mathbf{X}\mathbf{a}$  is sometimes called the *linear transformation* of the vector  $\mathbf{a}$  to the vector  $\mathbf{X}\mathbf{a}$ , and  $\mathbf{X}$  is the matrix of the transformation.

Example:

Let  $\mathbf{X} = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 4 & 1 \\ -1 & 3 & 5 \\ 6 & 7 & 5 \end{bmatrix}$  and  $\mathbf{a}' = [a_1 \ a_2 \ a_3]$ . Then,

$$\mathbf{X}\mathbf{a} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3] \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 4 & 1 \\ -1 & 3 & 5 \\ 6 & 7 & 5 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} a_1 - 2a_2 + 0a_3 \\ 0a_1 + 4a_2 + a_3 \\ -a_1 + 3a_2 + 5a_3 \\ 6a_1 + 7a_2 + 5a_3 \end{bmatrix}$$

is a vector for any scalar values  $a_1$ ,  $a_2$ , and  $a_3$ .

If there exists a vector  $\mathbf{a} \neq 0$ , such that  $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_n\mathbf{x}_n = \mathbf{0}$ , then provided none of  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  is null, those vectors are said to be *linearly dependent vectors*.

An alternative statement of the definition is the following:

If  $\mathbf{X}\mathbf{a} = \mathbf{0}$  for some nonnull  $\mathbf{a}$ , then the columns of  $\mathbf{X}$  are *linearly dependent vectors*, provided none is null.

If  $\mathbf{a} = \mathbf{0}$  is the only vector for which  $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_n\mathbf{x}_n = \mathbf{0}$ , then provided none of  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  is null, those vectors (the columns of  $\mathbf{X}$ ) are said to be *linearly independent vectors*.

The *rank* of a matrix is the number of linearly independent rows (and columns) in the matrix. The rank of  $\mathbf{A}$  is denoted by  $r_{\mathbf{A}}$  or  $r(\mathbf{A})$ .

If  $r(\mathbf{A}_{n \times n}) = n$ , then  $\mathbf{A}$  is *nonsingular*, that is,  $\mathbf{A}^{-1}$  exists.

If  $r(\mathbf{A}_{n \times n}) < n$ , then  $\mathbf{A}$  is *singular* and  $\mathbf{A}^{-1}$  does not exist.

If  $r(\mathbf{A}_{p \times q}) = p < q$ , then  $\mathbf{A}$  has *full row rank*, that is, its rank equals its number of rows.

If  $r(\mathbf{A}_{p \times q}) = q < p$ , then  $\mathbf{A}$  has *full column rank*, that is, its rank equals its number of columns.

When  $r(\mathbf{A}_{n \times n}) = n$ ,  $\mathbf{A}$  has *full rank*, that is, its rank equals its order, it is nonsingular, its inverse exists, and it is called *invertible*.

Because it is easier to work with rank than determinants, you often determine the existence of  $\mathbf{A}^{-1}$  by ascertaining whether  $r(\mathbf{A}) < n$  or  $r(\mathbf{A}) = n$  rather than by ascertaining whether  $|\mathbf{A}|$  is zero or not.

Example:

Consider the following vectors:

$$\mathbf{x}_1 = \begin{bmatrix} 3 \\ -6 \\ 9 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 5 \\ -5 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{x}_4 = \begin{bmatrix} -6 \\ 12 \\ -18 \end{bmatrix} \quad \mathbf{x}_5 = \begin{bmatrix} 2 \\ -3 \\ 3 \end{bmatrix}$$

$$\text{Because } 2\mathbf{x}_1 + \mathbf{x}_4 = \begin{bmatrix} 6 \\ -12 \\ 18 \end{bmatrix} + \begin{bmatrix} -6 \\ 12 \\ -18 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \mathbf{0}, \text{ that is,}$$

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_4 = \mathbf{0} \text{ for } \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \text{ which is nonnull,}$$

$\mathbf{x}_1$  and  $\mathbf{x}_4$  are *linearly dependent* vectors. So are  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$  because

$$2\mathbf{x}_1 + 3\mathbf{x}_2 - 3\mathbf{x}_3 = \begin{bmatrix} 6 \\ -12 \\ 18 \end{bmatrix} + \begin{bmatrix} 0 \\ 15 \\ -15 \end{bmatrix} + \begin{bmatrix} -6 \\ -3 \\ -3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \mathbf{0}$$

However, consider

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 = \begin{bmatrix} 3a_1 \\ -6a_1 \\ 9a_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 5a_2 \\ -5a_2 \end{bmatrix} = \begin{bmatrix} 3a_1 \\ -6a_1 + 5a_2 \\ 9a_1 - 5a_2 \end{bmatrix}$$

There are no values  $a_1$  and  $a_2$ , which makes it a null vector other than  $a_1 = 0 = a_2$ . Therefore,  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are *linearly independent* vectors.

The rank of the matrix  $\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \mathbf{x}_4 \quad \mathbf{x}_5] = \begin{bmatrix} 3 & 0 & 2 & -6 & 2 \\ -6 & 5 & 1 & 12 & -3 \\ 9 & -5 & 1 & -18 & 3 \end{bmatrix}$  is 3.

The calculation of ranks can be performed by row operations, which is discussed in many matrix algebra textbooks.

 The linear dependence or linear independence of vectors is a characteristic pertaining to a set of vectors of the same order. It is not a characteristic of individual vectors.

Any nonnull matrix  $\mathbf{A}$  of rank  $r$  is equivalent to

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{C},$$

where  $\mathbf{I}_r$  is the identity matrix of order  $r$ , and the null submatrices are of appropriate order to make  $\mathbf{C}$  the same order as  $\mathbf{A}$ . For  $\mathbf{A}$  of order  $m \times n$ ,  $\mathbf{P}$  and  $\mathbf{Q}$  are nonsingular matrices of order  $m$  and  $n$ , respectively, being products of elementary operators.

The matrix  $\mathbf{C}$  is called the *equivalent canonical form* or the *canonical form under equivalence*. It always exists, and can be used to determine the rank of  $\mathbf{A}$ .

## B.7 Generalized Inverse

The generalized inverse plays an important role in understanding the solutions to linear equations  $\mathbf{Ax} = \mathbf{y}$ , when  $\mathbf{A}$  has no inverse (but has generalized inverse).

A *generalized inverse* of a matrix  $\mathbf{A}$  is any matrix  $\mathbf{G}$  such that  $\mathbf{AGA} = \mathbf{A}$ . An alternative symbol for  $\mathbf{G}$  is  $\mathbf{A}^-$ .

When  $\mathbf{A}$  is not full rank, as occurs in many general linear models with  $\mathbf{A} = \mathbf{X}'\mathbf{X}$ , an infinite number of generalized inverses exist. Several ways of obtaining  $\mathbf{G}$  exist. The approach to obtaining a generalized inverse used by SAS is to partition a singular matrix into several sets of matrices.

Example:

Consider an  $\mathbf{X}'\mathbf{X}$  matrix of rank  $r$  that can be partitioned as

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where  $\mathbf{A}_{11}$  is  $r \times r$  and of rank  $r$ . Then  $\mathbf{A}_{11}^{-1}$  exists, and a generalized inverse of  $\mathbf{X}'\mathbf{X}$  is

$$(\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Because the generalized inverse is not unique, the solutions to normal equations in general linear models,  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{y}$ , are not unique either. However, a class of linear functions ( $\mathbf{L}\mathbf{b}$ ) called *estimable functions* exists, and  $\mathbf{L}\mathbf{b}$  and its variance are invariant through all possible generalized inverses. In other words, the linear combination  $\mathbf{L}\mathbf{b}$  is unique and is an unbiased estimate of  $\mathbf{L}\beta$ .

## B.8 Eigenvalues and Eigenvectors

---

Consider equations

$$\mathbf{Ax} = \lambda\mathbf{x} \text{ or } (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

for  $\mathbf{x}$  (a vector) and  $\lambda$  (a scalar), and solve for  $\mathbf{x}$ .

If the matrix  $\mathbf{A} - \lambda\mathbf{I}$  is nonsingular, the unique solution to these equations is  $\mathbf{x} = 0$ . You only get a nontrivial solution when  $|\mathbf{A} - \lambda\mathbf{I}| = 0$ . If you expand this determinant, the condition becomes a polynomial equation in  $\lambda$  of degree  $p$ . This is called the *characteristic equation* of  $\mathbf{A}$ . Its  $p$  roots (which might be real or complex, simple or multiple) are called *eigenvalues* (or proper values, characteristic values, or latent roots) of  $\mathbf{A}$ . If  $\lambda$  is an eigenvalue, a nonzero vector  $\mathbf{x}$  satisfying  $\mathbf{Ax} = \lambda\mathbf{x}$  is called an *eigenvector* (or proper vector, characteristic vector, or latent vector) corresponding to  $\lambda$ . It is often convenient to normalize each eigenvector to have a squared length of 1.

Example:

The matrix  $\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 9 & 1 \end{bmatrix}$  has a characteristic equation:

$$\begin{bmatrix} 1 & 4 \\ 9 & 1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0; \text{ that is, } \begin{vmatrix} 1-\lambda & 4 \\ 9 & 1-\lambda \end{vmatrix} = 0.$$

$$(1-\lambda)^2 - 36 = 0, \text{ or } \lambda = -5 \text{ or } 7$$

It can be seen that

$$\begin{bmatrix} 1 & 4 \\ 9 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -3 \end{bmatrix} = -5 \begin{bmatrix} 2 \\ -3 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 4 \\ 9 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = 7 \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

Therefore,  $\begin{bmatrix} 2 \\ -3 \end{bmatrix}$  is an eigenvector corresponding to the eigenvalue  $-5$

and  $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$  is an eigenvector corresponding to the eigenvalue  $7$ .

### Some of the Basic Properties of Eigenvalues and Eigenvectors

For  $\mathbf{Ax} = \lambda\mathbf{x}$ ,

1.  $\mathbf{A}^k\mathbf{x} = \lambda^k\mathbf{x}$  and  $\mathbf{A}^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}$ , when  $\mathbf{A}$  is nonsingular.
2.  $c\mathbf{Ax} = c\lambda\mathbf{x}$  for any scalar  $c$ .
3.  $f(\mathbf{A})\mathbf{x} = f(\lambda)\mathbf{x}$  for any polynomial function  $f(\mathbf{A})$ .
4.  $\sum_{i=1}^n \lambda_i = \text{tr}(\mathbf{A})$  and  $\prod_{i=1}^n \lambda_i = |\mathbf{A}|$ , that is, the sum of eigenvalues of a matrix equals its trace, and their product equals its determinant.
5. If  $\mathbf{A}$  is symmetric, then
  - eigenvalues of matrix  $\mathbf{A}$  are all real
  - $\mathbf{A}$  is diagonable
  - eigenvectors are orthogonal to each other
  - the rank of  $\mathbf{A}$  equals the number of nonzero eigenvalues
  - positive definite matrices have eigenvalues all greater than zero and vice versa.

The matrix  $\mathbf{A}$  is *diagonable* when a nonsingular matrix  $\mathbf{X}$  exists, and consists of the  $n$  eigenvectors, that is,  $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_n]$ , where all  $n$  eigenvectors are linearly independent, such that

$$\mathbf{X}^{-1}\mathbf{AX} = \mathbf{D} = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}.$$

For symmetric matrix  $\mathbf{A}$ , because the eigenvectors are orthogonal to each other, the above equation becomes the following:

$$\mathbf{X}'\mathbf{AX} = \mathbf{D} = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$$

It follows that  $\mathbf{A}$  can be written as  $\mathbf{X}\mathbf{DX}'$ , or

$$\mathbf{A} = \lambda_1\mathbf{x}_1\mathbf{x}_1' + \lambda_2\mathbf{x}_2\mathbf{x}_2' + \dots + \lambda_n\mathbf{x}_n\mathbf{x}_n'.$$

This is called the *spectral decomposition* of the matrix  $\mathbf{A}$ .

When  $\mathbf{A}$  is nonsingular, the spectral decomposition of  $\mathbf{A}^{-1}$  is

$$\mathbf{A}^{-1} = \lambda_1^{-1}\mathbf{x}_1\mathbf{x}_1' + \lambda_2^{-1}\mathbf{x}_2\mathbf{x}_2' + \dots + \lambda_n^{-1}\mathbf{x}_n\mathbf{x}_n'$$

When  $\mathbf{A}$  is singular, a generalized inverse of  $\mathbf{A}$  can be obtained from its spectral decomposition in exactly the same way, by simply omitting the terms for which  $\lambda_i = 0$ .

## B.9 Cholesky Root

---

If  $\mathbf{A}$  is positive definite of size  $n \times n$ , you can find an upper triangular matrix  $\mathbf{U}$  such that  $\mathbf{A} = \mathbf{U}'\mathbf{U}$ , so that  $\mathbf{U}$  is a type of square root of  $\mathbf{A}$ , also referred to as the *Cholesky root* of the matrix  $\mathbf{A}$ .

The rule for the construction of  $\mathbf{U}$  is simply the following:

$$(\text{column } i \text{ of } \mathbf{U}) \times (\text{column } j \text{ of } \mathbf{U}) = a_{ij},$$

that is,

$$\begin{aligned} u_{11}^2 &= a_{11}, \\ u_{11}u_{12} &= a_{12}, \\ u_{11}u_{13} &= a_{13}, \\ &\dots \\ u_{12}^2 + u_{22}^2 &= a_{22}, \\ u_{12}u_{13} + u_{22}u_{23} &= a_{23}, \\ &\dots \\ u_{13}^2 + u_{23}^2 + u_{33}^2 &= a_{33}, \end{aligned}$$

and so on.

The procedure for forming  $\mathbf{U}$  (called the *square-root* or *Cholesky* procedure) provides a basis for an excellent numerical method for solving simultaneous linear equations and inverting matrices. This procedure can also be applied when  $\mathbf{A}$  is only positive semidefinite. If, when a zero  $u_{ii}$  occurs, you simply set all subsequent elements in the same row of  $\mathbf{U}$  to zero, the relation  $\mathbf{A} = \mathbf{U}'\mathbf{U}$  remains and the matrix  $\mathbf{U}^{-1}(\mathbf{U}')^{-1}$  is a generalized inverse of  $\mathbf{A}$ .



# Appendix C Review of Simple Linear Regression and One-Way ANOVA

<b>C.1 Univariate Analysis.....</b>	<b>C-3</b>
Demonstration: Exploring Data .....	C-7
<b>C.2 Simple Linear Regression.....</b>	<b>C-13</b>
Demonstration: Simple Linear Regression .....	C-19
<b>C.3 One-Way ANOVA Review.....</b>	<b>C-24</b>
Demonstration: One-Way ANOVA .....	C-31
Demonstration: Multiple Comparison Tests .....	C-36
<b>C.4 Chapter Summary.....</b>	<b>C-38</b>



## C.1 Univariate Analysis

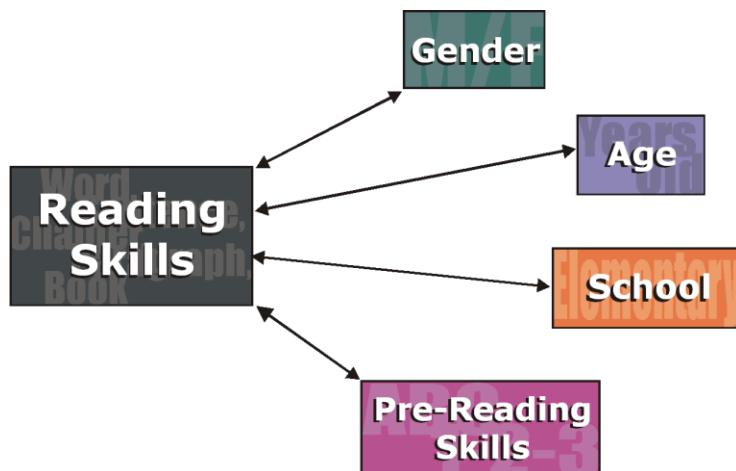
---

### Objectives

- Use the UNIVARIATE and SGANEL procedures to explore data.

3

### School Data



4

Data were collected by a school district to assess the reading skill progress of students in their first year of formal schooling. These are the variables in the **STAT2.school** data set:

<b>ID</b>	ID number of student
<b>Gender</b>	gender of student ( <i>M, F</i> )

<b>Age</b>	student's age (rounded to nearest tenth of a year)
<b>School</b>	school student attends
<b>Teacher</b>	name of student's teacher
<b>Semesters</b>	number of semesters student has attended in the district
<b>Letters1</b>	score on letter identification test in the fall
<b>Phonics1</b>	score on letter sound test in the fall
<b>Words1</b>	score on word identification test in the fall
<b>Phonics2</b>	score on letter sound test in the winter
<b>Words2</b>	score on word identification test in the winter
<b>Phonics3</b>	score on letter sound test in the spring
<b>Words3</b>	score on word identification test in the spring
<b>Reading2</b>	score on reading test in the winter
<b>Fluency2</b>	score on reading fluency test in the winter
<b>Reading3</b>	score on reading test in the spring
<b>Fluency3</b>	score on reading fluency test in the spring.

## Exploring Data

- Screen for unusual data values.
- Inspect the spread and shape of continuous variables.
- Characterize the location and dispersion.
- Draw preliminary conclusions.

5

Before using any inferential statistics, you should explore and describe your data. This helps ensure that your data are as error free as possible. At the same time, it enables you to find any unique aspects of your data and any extreme values that might affect your analysis.

Some useful tools for exploring your data are descriptive statistics, histograms, and box plots. In addition, you might want to determine whether your data are normally distributed.

## The UNIVARIATE Procedure

General form of the UNIVARIATE procedure:

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

6

The UNIVARIATE procedure generates descriptive statistics and provides details about the distribution of the variables. In addition, graphs of your data can be generated.

Selected UNIVARIATE procedure statements:

**VAR** specifies numeric variables to analyze and the order that they appear in the results. If no VAR statement is used, all numeric variables in the data set are analyzed.

**HISTOGRAM** generates a high-resolution histogram of the variables in the VAR statement.

**PROBPLOT** creates a high-resolution probability plot, which compares ordered variable values with the percentiles of a specified theoretical distribution. The default theoretical distribution is a normal distribution with the mean and standard deviation of your data.

## The SG PANEL Procedure

General form of the SG PANEL procedure:

```
PROC SG PANEL options;  
  PANELBY variable(s) / option(s);  
  VBOX response-variable / option(s);  
RUN;
```

7

Box plots provide information about the distribution of data. The plot divides the data into quartiles, with the box representing the middle 50% of the data.

Selected SG PANEL procedure statements:

PANELBY specifies one or more classification variables for the panel, the layout type, and other options for the panel.

VBOX creates a vertical box plot that shows the distribution of your data. Vertical box plots display the distribution of data by using a rectangular box and whiskers. Whiskers are lines that indicate a data range outside of the box.

Selected PANELBY statement option:

COLUMNS specifies the number of columns in the panel. By default, the number of columns is determined automatically based on the number of classifier values and the layout type. The SG PANEL procedure automatically splits the panel into multiple graphs (pages) as needed when your panel contains a large number of cells. You can control the number of cells in each graph by using the COLUMNS= and the ROWS= options.



## Exploring Data

---

Print the first 25 observations of the data set **STAT2.school**.

```
ods html close;
ods listing;
proc print data=STAT2.school (obs=25);
run; *ST20Cd01.sas;
```

PROC PRINT Output

	L	P	P	P	R	F	F	R		S
	e	h	h	h	e	l	l	e		e
	t	o	W	o	W	a	u	u	G	m
	t	n	o	n	o	d	e	e	d	e
	e	i	r	i	r	i	n	n	i	s
O	r	c	d	c	d	n	c	c	n	t
b	s	s	s	s	s	g	y	y	g	e
s	1	1	1	2	2	3	3	2	3	r
1	15	36	2	74	46	47	77	9	7	29
2	8	7	1	12	23	52	38	1	0	8
3	3	1	0	35	1	10	51	0	0	0
4	39	13	13	40	16	63	37	7	9	28
5	55	17	22	46	43	47	67	18	12	50
6	36	17	16	50	37	38	38	12	11	39
7	24	22	10	58	45	49	54	9	8	35
8	55	26	28	53	51	51	66	46	55	85
9	55	19	19	57	46	56	55	22	22	59
10	59	19	92	57	81	41	119	79	82	119
11	50	29	37	56	55	55	74	26	26	66
12	24	9	7	48	42	31	49	10	10	35
13	33	28	31	51	61	40	96	46	57	84
14	44	41	21	52	57	58	61	50	52	61
15	41	24	12	56	44	48	70	13	9	61
16	52	25	35	58	67	51	74	39	45	98
17	24	4	7	40	44	47	48	7	5	48
18	34	6	2	54	44	54	69	11	7	60
19	18	6	6	28	26	46	76	7	5	66
20	51	18	27	71	83	67	92	19	17	23
21	.	.	.	.	32	29	.	.	11	9
22	.	.	.	48	45	55	53	11	9	30
23	.	.	.	44	43	.	.	19	18	.
24	.	.	.	45	55	41	27	8	7	26
25	.	.	.	59	39	55	56	13	11	45
										33
										F
										7.3
										140
										Cottonwood
										Miss Jones
										6

You are interested in examining the scores on the spring reading tests (**Reading3**). You can begin by using the UNIVARIATE procedure to examine the distribution of the scores.

```
proc univariate data=STAT2.school;
var reading3;
histogram / normal;
probplot / normal(mu=est sigma=est);
id school;
run; *ST20Cd01.sas;
```

Selected HISTOGRAM statement option:

**NORMAL** superimposes a normal distribution curve on the histogram and requests tests for normality.

Selected PROBPLOT statement option:

**NORMAL** creates a normal probability plot. This is the default if you omit a distribution option. The normal-options MU=EST and SIGMA=EST add a line corresponding to estimated values of the mean and the standard deviation. Agreement between the reference line and the point pattern indicates that the normal distribution with estimated sample parameters is a good fit.

Examine the results in the Output window.

The UNIVARIATE Procedure			
Variable: Reading3			
Moments			
N	179	Sum Weights	179
Mean	47.7597765	Sum Observations	8549
Std Deviation	39.8119647	Variance	1584.99253
Skewness	0.87139671	Kurtosis	-0.1206008
Uncorrected SS	690427	Corrected SS	282128.67
Coeff Variation	83.358775	Std Error Mean	2.97568595

The Moments table lists some of the descriptive statistics for the variable **Reading3**. The sample mean is approximately 48. The coefficient of variation is approximately 83%. This indicates that the standard deviation is large compared to the mean. The skewness statistic is positive. This means that the data are skewed to the right.

Basic Statistical Measures			
Location		Variability	
Mean	47.75978	Std Deviation	39.81196
Median	32.00000	Variance	1585
Mode	10.00000	Range	183.00000
		Interquartile Range	66.00000
Tests for Location: Mu0=0			
Test	-Statistic-	-----	p Value-----
Student's t	t 16.05001	Pr >  t	<.0001
Sign	M 88.5	Pr >=  M	<.0001
Signed Rank	S 7876.5	Pr >=  S	<.0001

In the basic statistical measures table, the median is given as 32. This is much lower than the mean. It is also an indication that the data are skewed to the right.

Quantiles (Definition 5)		
Quantile	Estimate	
100% Max	183	
99%	151	
95%	119	
90%	105	
75% Q3	81	
50% Median	32	
25% Q1	15	
10%	8	
5%	5	
1%	0	
0% Min	0	

Extreme Observations					
-----Lowest-----			-----Highest-----		
Value	School	Obs	Value	School	Obs
0	Cottonwood	37	141	Pine	169
0	Cottonwood	3	146	Pine	148
1	Pine	163	147	Dogwood	59
1	Cottonwood	31	151	Pine	180
1	Cottonwood	29	183	Cottonwood	35

Missing Values					
-----Percent Of-----					
Missing Value	Count	All Obs	Missing Obs		
.	11	5.79	100.00		

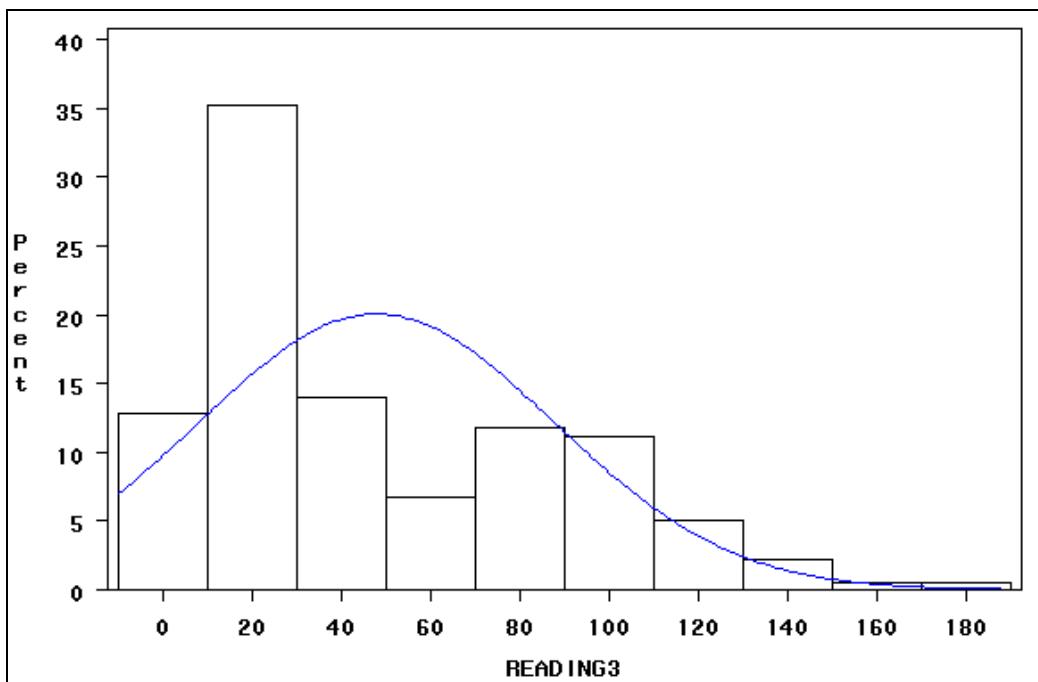
The lowest test score is zero and the highest is 183. The schools associated with these extreme values are also listed because of the **ID School;** statement. There are 11 missing values in this data set for the variable **Reading3**.

Fitted Distribution for Reading3				
Parameters for Normal Distribution				
Parameter	Symbol	Estimate		
Mean	Mu	47.75978		
Std Dev	Sigma	39.81196		
Goodness-of-Fit Tests for Normal Distribution				
Test	---Statistic---		-----p Value-----	
Kolmogorov-Smirnov	D	0.17530962	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.27496460	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	7.03638617	Pr > A-Sq	<0.005
Quantiles for Normal Distribution				
-----Quantile-----				
Percent	Observed	Estimated		
1.0	0.000	-44.85670		
5.0	5.000	-17.72508		
10.0	8.000	-3.26131		
25.0	15.000	20.90701		
50.0	32.000	47.75978		
75.0	81.000	74.61254		
90.0	105.000	98.78086		
95.0	119.000	113.24463		
99.0	151.000	140.37626		

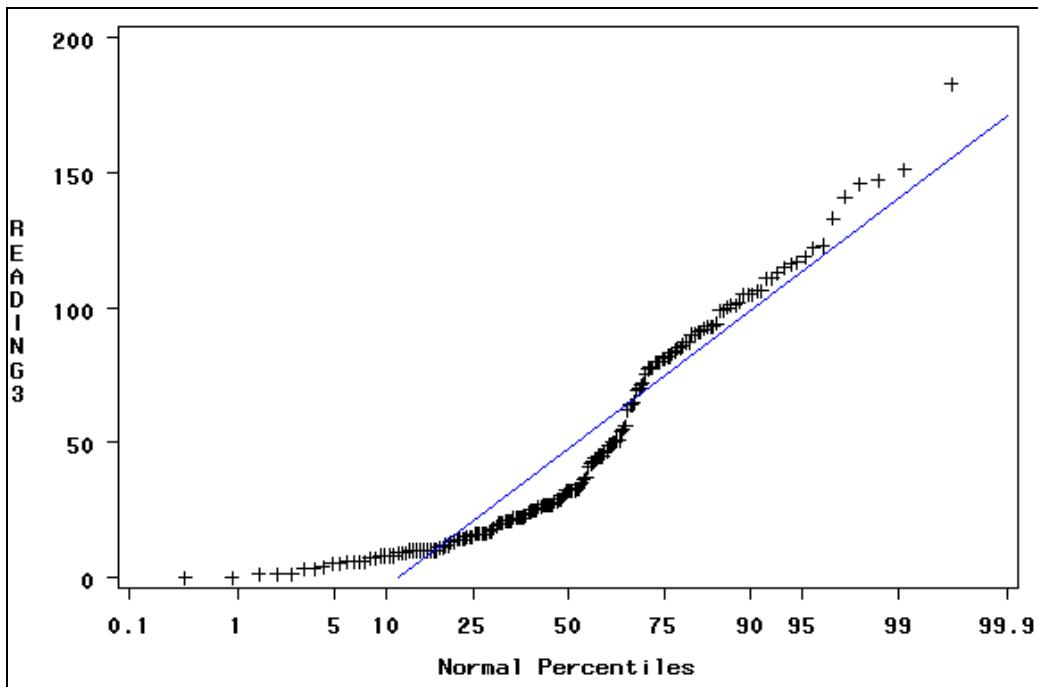
There are three tests for normality given in the output. The null hypothesis for these tests is that the data are normally distributed. Presuming an alpha equal to 0.05, the *p*-values for all of these tests are less than alpha. Therefore, you reject the null hypothesis and conclude that the data are not normally distributed.

- ✍ All the normality tests depend on the sample size. As the sample size becomes larger, increasingly smaller departures from normality can be detected. A small sample size likely yields a less powerful test and you might want to use a higher alpha value. In general, it is recommended that you also examine some graphs, such as a normal probability plot, to evaluate the normality assumption.

Examine the histogram and probability plot.



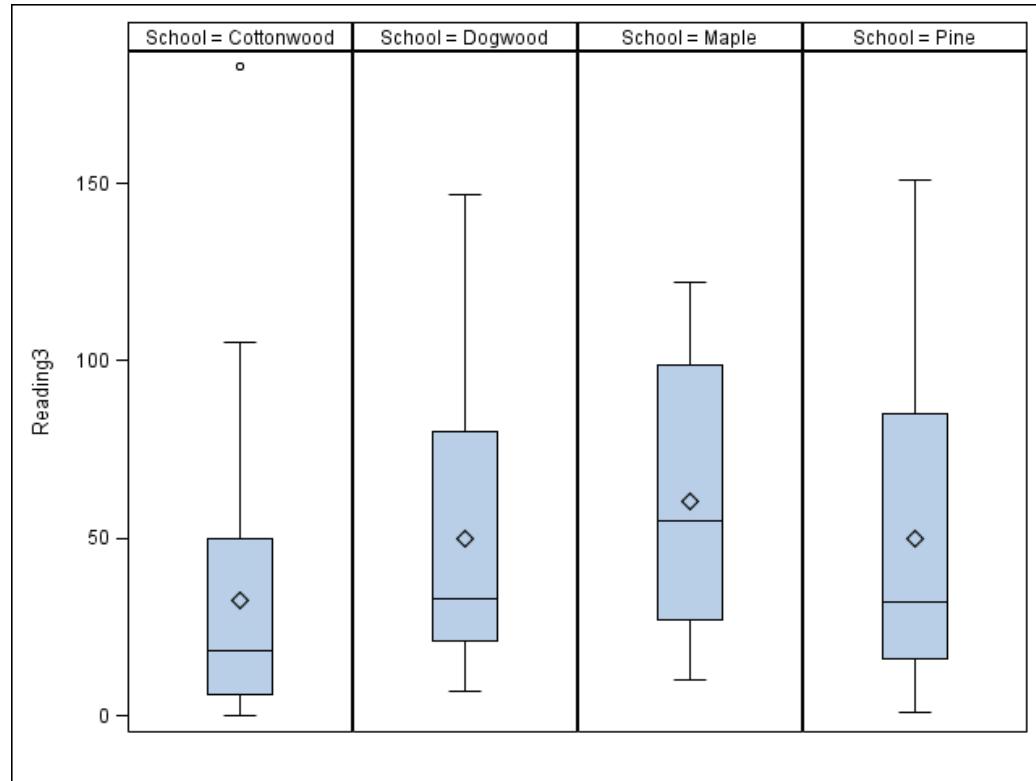
The normal distribution graph is superimposed on the histogram. The graph shows that the data are not normally distributed. Instead, it is skewed to the right.



If the data are normally distributed, the normal probability plot should be a straight line close to the reference line. The curvature evident on this plot is another indication that the data are not normally distributed. Instead, it is skewed to the right.

You can generate high-resolution side-by-side box plots with PROC SGPMANL. For example, suppose you want to compare the reading scores for the different schools in the district.

```
proc sgpanel data=STAT2.school;
  panelby school / columns=4;
  vbox reading3;
run; *ST20Cd01.sas;
```



The box plot shows that in general the distribution is skewed to the right (skewed to the top in box plots). Cottonwood School has one unusual observation, which should be verified for correctness. The data for all of the schools are skewed to the right. There might be differences in the means between the schools.

## C.2 Simple Linear Regression

### Objectives

- Explain the concepts of simple linear regression.
- Fit a simple linear regression model using the REG procedure.
- List linear regression model assumptions.

10

### Linear Regression Analysis

The objectives of linear regression are to

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

11

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

## Ordinary Least Squares (OLS) Estimates

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

residual

The line fitted by least squares is the one that makes the sum of squared residuals as small as possible.

12

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

$Y$  response variable

$X$  predictor variable

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

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

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

The parameters  $\beta_0$  and  $\beta_1$  are estimated using the ordinary least squares (OLS) method. OLS solutions have good statistical properties such as normal, unbiased, and minimum variance.

### Details

Let the sum of squares of deviations from the true line be

$$S = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

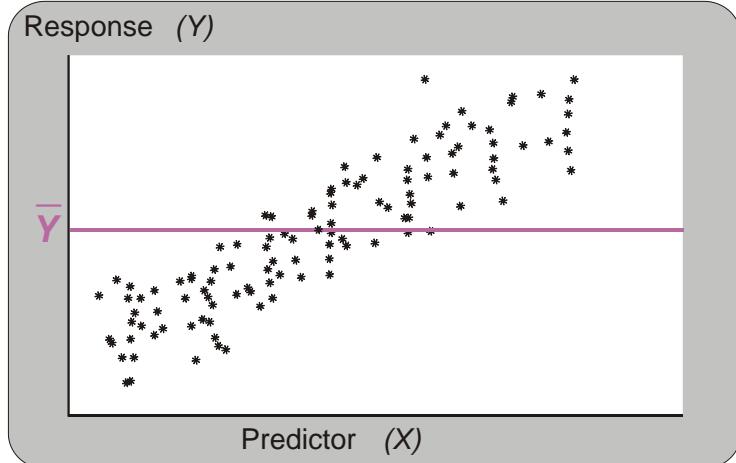
You shall choose the estimates for  $\beta_0$  and  $\beta_1$  such that the sum of squared residuals  $S$  is minimized. You can determine the estimates by taking the first derivative with respect to  $\beta_0$  and  $\beta_1$ , respectively, and then setting the results to zero. It follows that

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sum X_i^2 - n \bar{X}^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$



The general form of the solutions in matrix notation is  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ .

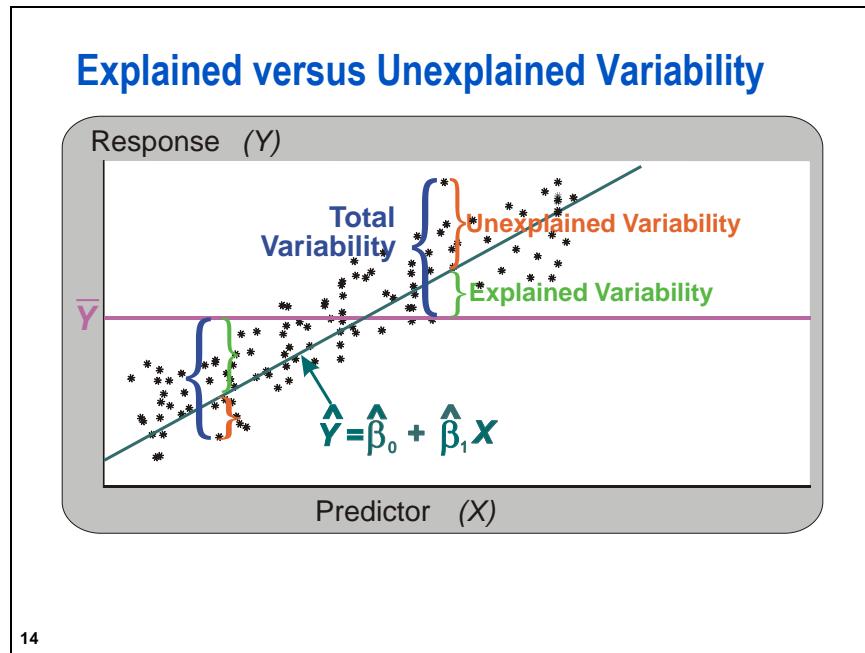
## The Significance of Linear Regression Model



13

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

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



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

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

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

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

It can be shown that Total variability = Explained variability + Unexplained variability, or  $SST = SSR + SSE$ . The model is significant when the explained variability (SSR) is relatively large compared to the unexplained variability (SSE). The coefficient of determination ( $R^2$ ) that measures the percent of variation in the data explained by the model, is computed as  $1 - SSE/SST$ , or  $SSR/SST$ .

## The Simple Linear Regression Model Hypothesis Test

**Null Hypothesis:  $H_0: \beta_1 = 0$**

The simple linear regression model does not fit the data better than the mean model.

**Alternative Hypothesis:  $H_1: \beta_1 \neq 0$**

The simple linear regression model does fit the data better than the mean model.

15

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

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

### Details

The significance of the model can be tested using the  $F$  statistic shown in the ANOVA table below.

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Squares (MS=SS/df)	$F$ value	$p$ -value
Due to regression	1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	MSR=SSR/dfR	MSR/MSE	If $< \alpha$ (predefined, for example, 0.05), then significant model
About regression (residual)	$n-2$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	MSE=SSE/dfE		
Total, corrected for mean $\bar{Y}$	$n-1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$			

## The REG Procedure

General form of the REG procedure:

```
PROC REG options PLOTS (global-plot-options) =  

   (plot-request (specific-plot-options));  

MODEL dependents=regressors / options;  

PLOT yvariable*xvariable =symbol / options;  

RUN;
```

16

Selected REG procedure statements:

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

**PLOT** displays scatter plots with Y variable on the vertical axis and X variable on the horizontal axis.

 PROC REG can be used interactively. After you specify a model with a MODEL statement and run PROC REG with a RUN statement, a variety of statements can be executed without re-invoking PROC REG. Notice that the MODEL statement can be repeated. This is an important difference from the GLM procedure, which allows only one MODEL statement.

If you use PROC REG interactively, you can end the REG procedure with a DATA step, another PROC, an ENDSAS statement, or with a QUIT statement. Additional RUN statements do not end PROC REG but tell the procedure to execute additional statements. Using PROC REG interactively enables you to fit a model, perform diagnostics, and then refit the model, and perform diagnostics on the refitted model.



## Simple Linear Regression

---

The school district is interested in predicting the spring reading scores (**Reading3**) based on tests taken by students in the fall. Use the REG procedure to generate a regression with **Reading3** as the response variable and **Words1** as the predictor variable. In order to plot the baseline model, it might be useful to compute the average value of **Reading3** first.

```
proc means data=STAT2.school mean;
  var reading3;
  where words1 is not missing;
run;                                              *ST20Cd02.sas;
```

The WHERE statement is used to ensure that the mean is computed for the same data set used to fit the linear regression model.

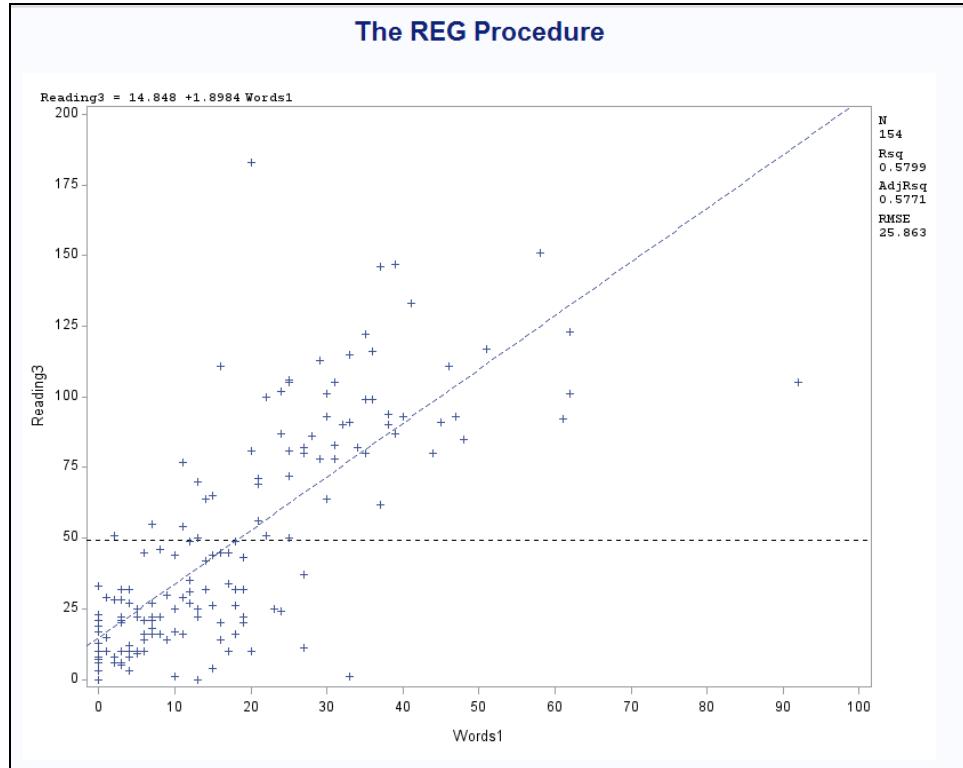
PROC MEANS Output

The MEANS Procedure	
Analysis Variable : Reading3	
	Mean
	49.2402597

```
proc reg data=STAT2.school;
  model reading3 = words1;
  plot reading3*words1 / vref=49.24;
run;
quit;                                              *ST20Cd02.sas;
```

Selected PLOT statement option:

VREF= specifies reference lines perpendicular to the vertical axis. The vertical reference line specified here is at the value of the mean of the variable **Reading3**. This enables you to visually compare the baseline model with the regression model.



The plot shows a scatter plot of the points with the regression line and mean line superimposed on it. The regression line appears to fit the data better than the baseline model. In addition, the regression equation and some of the model statistics are shown.

## PROC REG Output

The REG Procedure Model: MODEL1 Dependent Variable: Reading3					
<b>Number of Observations Read</b>					190
<b>Number of Observations Used</b>					154
<b>Number of Observations with Missing Values</b>					36
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	140325	140325	209.79	<.0001
Error	152	101671	668.88715		
Corrected Total	153	241996			
Root MSE 25.86285 R-Square 0.5799 Dependent Mean 49.24026 Adj R-Sq 0.5771 Coeff Var 52.52379					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	14.84765	3.15938	4.70	<.0001
Words1	1	1.89837	0.13107	14.48	<.0001

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

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

- The model DF is the number of parameters in the model minus one. In this case, there are two parameters to be estimated, the intercept and the slope.
- The corrected total DF is the sample size minus one.

The sums of squares provide information about the sources of variability in the response variable.

- The model sum of squares is associated with the variability in **Reading3** that is explained by the variable **Words1**.
- The error sum of squares is associated with the variability in **Reading3** that is not explained by the variable **Words1**.
- The corrected total sum of squares is associated with the overall variability in **Reading3**. The sum of the model and error sums of squares is equal to the corrected total sum of squares.

The mean squares are calculated by dividing the sums of squares by their corresponding degrees of freedom.

The *F* value is calculated by dividing the mean square for the model by the mean square for error.

The *p*-value tests the hypothesis that the slope for the model is equal to zero. It is the probability of obtaining an *F* value this large or larger if the null hypothesis is true. In this case, presuming an alpha of 0.05, the *p*-value is less than alpha. You reject the null hypothesis and conclude that you have sufficient evidence that the slope of the regression line is not equal to zero. Therefore, this model fits the data better than the baseline model.

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

Root MSE	The root mean square error is an estimate of the standard deviation of the response variable at each value of the explanatory variable. It is the square root of the MSE.
Dependent Mean	The overall mean of the response variable.
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"> <li>• calculated as <math>\left( \frac{\text{RootMSE}}{\bar{Y}} \right) * 100</math></li> <li>• a unitless measure, so it can be used to compare data that have different units of measurement or different magnitudes.</li> </ul>
R-Square	The coefficient of determination is usually referred to as the “R square.” This value is <ul style="list-style-type: none"> <li>• between 0 and 1.</li> <li>• the proportion of variability observed in the data explained by the regression line. In this example, the value is 0.5799, which means that the regression line explains 58% of the total variation in the response values.</li> </ul>
Adj R-Sq	The adjusted R square is the R square adjusted for the number of parameters in the model. This statistic is useful in multiple regression and is discussed in detail in a later section.
 The Parameter Estimates table defines the predictive 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 parameter 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 parameters by their corresponding standard errors.
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 explanatory variable is different from 0. Therefore, you can conclude that the predictor variable explains a significant portion of the variability in the response variable.

The Parameter Estimates table gives the estimates of the intercept and slope. In this case, the regression equation can be written as **Reading3** = 14.84765 + 1.89837 \* **Words1**. The slope of **Words1** indicates that for every one-unit increase in the fall word test score (**Words1**), the spring reading test score (**Reading3**) increases by 1.89837. You must be careful when you attempt to interpret the intercept. Recall that the intercept is the value of Y when X is equal to zero. The linear relationship exists over the range of Xs in the data set, but there is no guarantee, or evidence, that the same linear relationship exists outside of the range of the X variable. Therefore, in many cases, the value of the intercept has no physical meaning. In this example, the value of the intercept does have meaning because you do have values of **Words1** at zero.

### Bonus Program

The following program uses a macro variable to automate the reference of the mean value in the graph:

```
options symbolgen;
proc sql noprint;
  select avg(reading3)
    into :mean
   from STAT2.school
  where words1 is not missing;
quit;

proc reg data=STAT2.school;
  model reading3 = words1;
  plot reading3*words1 / vref=&mean;
run;
quit;                                ST20Cd02.sas;
```

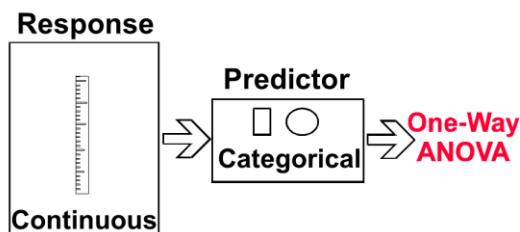
## C.3 One-Way ANOVA Review

### Objectives

- Identify differences in population means with ANOVA.
- Determine which population means are different using multiple comparison methods.

19

### One-Way ANOVA

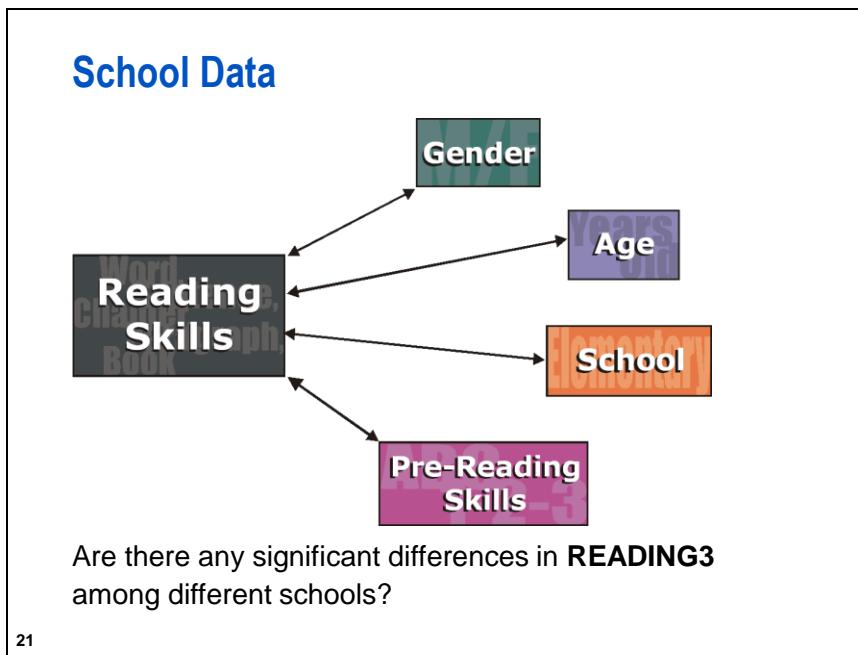


Are there any differences in the population means?

20

Analysis of variance (ANOVA) is a statistical technique used to compare the means of two or more groups of observations, or treatments. For one-way ANOVA, you have the following:

- continuous dependent, or response, variable
- discrete independent variable, also called a *predictor* or *explanatory* variable, that separates your data into several groups



Data were collected by a school district to assess the reading skill progress of students in their first year of formal schooling. These are the variables in the **STAT2.school** data set:

<b>ID</b>	ID number of student
<b>Gender</b>	gender of student (F, M)
<b>Age</b>	student's age (rounded to nearest tenth of a year)
<b>School</b>	school student attends
<b>Teacher</b>	name of student's teacher
<b>Semesters</b>	number of semesters student has attended in the district
<b>Letters1</b>	score on letter identification test in the fall
<b>Phonics1</b>	score on letter sound test in the fall
<b>Words1</b>	score on word identification test in the fall
<b>Phonics2</b>	score on letter sound test in the winter
<b>Words2</b>	score on word identification test in the winter
<b>Phonics3</b>	score on letter sound test in the spring
<b>Words3</b>	score on word identification test in the spring
<b>Reading2</b>	score on reading test in the winter
<b>Fluency2</b>	score on reading fluency test in the winter
<b>Reading3</b>	score on reading test in the spring
<b>Fluency3</b>	score on reading fluency test in the spring.

To evaluate the **School** effect, that is, to examine whether the average readings values are the same or different among different schools, you can conduct a one-way analysis of variance.

## The One-Way ANOVA Model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$E(y_{ij}) = \mu + \alpha_i$$

$$\text{var}(y_{ij}) = \sigma^2$$

22

$y_{ij}$  the  $j^{\text{th}}$  value of **Reading3** for the  $i^{\text{th}}$  school.

$\mu$  the overall mean of the response **Reading3**.

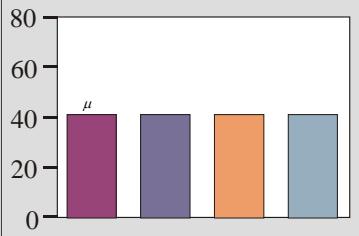
$\alpha_i$  the  $i^{\text{th}}$  school effect. It is estimated by taking the difference between the mean value of **Reading3** for the  $i^{\text{th}}$  school and the overall mean  $\mu$ .

$\varepsilon_{ij}$  error term, or residual.

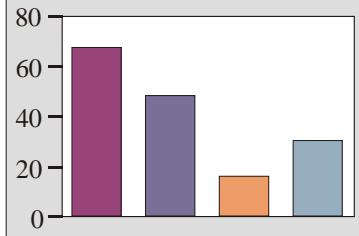
It is assumed that the residuals are independently and normally distributed with a mean of zero and a constant variance of  $\sigma^2$ . It follows that the expected (average) value of **Reading3** for school  $i$  is  $\mu + \alpha_i$ , and the variance of **Reading3** for each school is  $\sigma^2$ .

## The ANOVA Hypothesis

$$H_0: \alpha_i=0 \text{ for all } i$$



$$H_1: \text{At least one } \alpha_i \neq 0$$



23

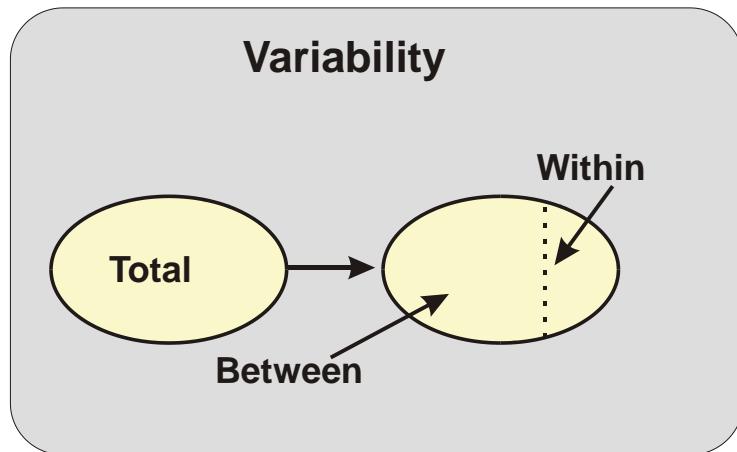
Differences between sample means are usually present. The objective is to determine whether these differences are significant. In other words, is the difference more than what might be expected to occur by chance? If the difference is more than what might be expected to occur by chance, you have sufficient evidence to conclude that there are significant differences between the population means.

The hypotheses for ANOVA are as follows:

$H_0: \alpha_i=0$  for all  $i$ s, or there is no treatment effect. In other words, all of the group means are equal.

$H_1: \text{at least one } \alpha_i \neq 0$ , or there is a treatment effect. In other words, **at least** one group mean is different from at least one other group mean.

## Partitioning Variability in ANOVA



24

As its name implies, analysis of variance analyzes the variances of the data to determine whether there is a difference between the group means. The total variation is partitioned into two parts: the between-group variation and the within-group variation.

Between Group Variation	the sum of the squared differences between the mean for each group and the overall mean, $\sum n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2$ . This is the variability explained by the independent variable. It is the model sum of squares.
Within Group Variation	the sum of the squared differences between each observed value and the mean for its group, $\sum \sum (y_{ij} - \bar{y}_{i\cdot})^2$ . This is the variability not explained by the independent variable. It is the error sum of squares.
Total Variation	the sum of the squared differences between each observed value and the overall mean, $\sum \sum (y_{ij} - \bar{y}_{..})^2$ . This is the overall variability in the response variable and is equal to the between-group variation plus the within-group variation. It is the corrected total sum of squares.

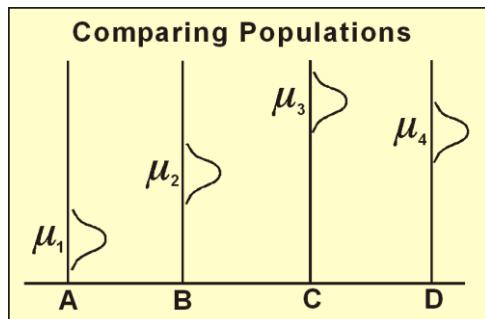
It can be shown that Total variation = between-group variation + within-group variation, or  $SST=SSTR+SSE$ . The model is significant when the between-group variation (SSTR) is relatively large compared to the within-group variation (SSE).

## Details

The significance of the model can be tested using the  $F$  statistic shown in the ANOVA table below.

	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Squares (MS=SS/df)	F value	p-value
Between treatments	$r-1$	$\sum n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2$	$MSTR=SSTR/(r-1)$	$MSR/MSE$	If $< \alpha$ (predefined, for example, 0.05), then significant model
Residuals (within treatments)	$n-r$	$\sum \sum (y_{ij} - \bar{y}_{i\cdot})^2$	$MSE=SSE/(n-r)$		
Total, corrected for mean $\bar{Y}$	$n-1$	$\sum \sum (y_{ij} - \bar{y}_{..})^2$			

## Assumptions of ANOVA



- independent observations
- normally distributed data for each group, or the pooled error terms are normally distributed
- equal variances for each group

25

The assumption of independent observations means no observations provide any information about any other observation that you collect. For example, measurements are *not* repeated on the same subject.

The assumption that the data for each group is approximately normal can be verified by examining plots of the data. In theory, the data for each group should be checked separately for normality. In practice, the data or the residuals as a whole are usually checked for normality. This assumption can be relaxed when the sample size is large enough.

The assumption of equal variances can be checked by looking at descriptive statistics and plots of the data and by conducting a test for equal variances.

If these assumptions are *not* valid, the probability of drawing incorrect conclusions from the analysis might be increased.

## The GLM Procedure

General form of the GLM procedure:

```
PROC GLM options PLOTS (global-plot-options) =  

                    (plot-request (specific-plot-options));  

CLASS variables;  

MODEL dependents=independents / options;  

LSMEANS effects / options;  

RUN;
```

26

The GLM procedure uses the method of least squares to fit general linear models. Among the statistical methods available in PROC GLM are regression, analysis of variance, analysis of covariance, multivariate analysis of variance, and partial correlation.

Selected GLM procedure statements:

**CLASS** specifies classification variables for the analysis.

**MODEL** specifies dependent and independent variables for the analysis.

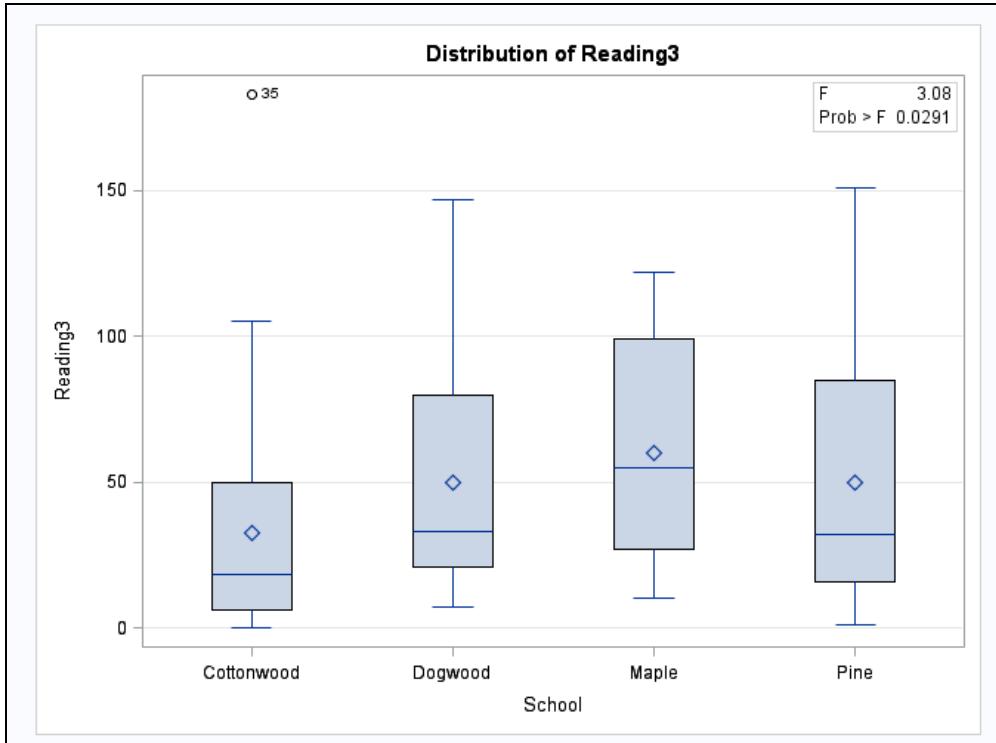
**LSMEANS** computes least squares means for each *effect* listed in the LSMEANS statement. You might specify only classification effects in the LSMEANS statement, that is, effects that contain only classification variables. You might also specify options to perform multiple comparisons. In contrast to the MEANS statement, the LSMEANS statement performs multiple comparisons on interactions as well as main effects.

Least squares means are *predicted population margins*. That is, they estimate the marginal means over a balanced population. In a sense, least squares means are to unbalanced data as class and subclass arithmetic means are to balanced data.



## One-Way ANOVA

Recall that in a previous section, box plots of the variable **Reading3** by **School** were created.



Are the differences among the four schools in the average **Reading3** test scores significant? Use the GLM procedure to conduct a formal statistical test.

```
proc glm data=STAT2.school;
  class school;
  model reading3=school;
run; *ST20Cd03.sas;
```

PROC GLM Output

The GLM Procedure		
Class Level Information		
Class	Levels	Values
School	4	Cottonwood Dogwood Maple Pine
Number of Observations Read		190
Number of Observations Used		179

The Class Level Information table specifies the class variable, the total number of levels and the values of the class variable. It also lists total number of observations in the data and how many observations have been used in the analysis.

### The GLM Procedure

#### Dependent Variable: Reading3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14130.8051	4710.2684	3.08	0.0291
Error	175	267997.8652	1531.4164		
Corrected Total	178	282128.6704			

R-Square	Coeff Var	Root MSE	Reading3 Mean
0.050086	81.93781	39.13332	47.75978

Source	DF	Type I SS	Mean Square	F Value	Pr > F
School	3	14130.80515	4710.26838	3.08	0.0291

Source	DF	Type III SS	Mean Square	F Value	Pr > F
School	3	14130.80515	4710.26838	3.08	0.0291

The Analysis of Variance table shows the partitioning of the variability.

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

- Model DF is the number of treatment groups minus one.
- Error DF is the number of observations used in the analysis minus the number of treatment groups.
- Corrected Total DF is the number of observations used in the analysis minus one.

Mean Squares are calculated by dividing the sums of squares by the corresponding degrees of freedom.

- Mean square for error (MSE) is an estimate of  $\sigma^2$ , the constant variance assumed for all treatments.
- If all treatment population means are equal, the mean square for the model (also called mean square treatment, or MSTR) is also an estimate of  $\sigma^2$ .
- If not all treatment population means are equal, MSTR estimates  $\sigma^2$  plus a positive constant.

The *F* value is MSTR divided by MSE. If the *F* statistic is significantly larger than one, it supports rejecting the null hypothesis, concluding that all treatment means are not equal.

The *F* value and corresponding *p*-value are reported in the analysis of variance table. Presuming an alpha equal to 0.05, the *p*-value is less than alpha. Therefore, you reject the null hypothesis and conclude that there are differences in the population means among different schools.

The coefficient of determination, or R square, is a measure of the proportion of variability explained by the independent variables in the analysis. This statistic is calculated as SSTR divided by SST.

The value of R square is between 0 and 1. The value is

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

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

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

The **Reading3** mean is the mean of all of the data values in the variable **Reading3** without regard to the treatment.

For a one-way analysis of variance, the information about the class variable in the model is an exact duplicate of the model line in the analysis of variance table. The significant *p*-value indicates that **School** is a significant factor in explaining variations in the **Reading3** values. Therefore, if there is at least one group mean different from one other group mean, you want to conduct multiple comparison tests to determine which group means differ.

## Multiple Comparison Tests

Not all population means are equal, but which ones are

**different?**

28

After you determined that the population means for all treatments are not equal, you want to take this further. Determine which means are significantly different. *Multiple comparisons*, also known as *post-hoc tests*, enable you to make this determination.

## Multiple Comparison Methods

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

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

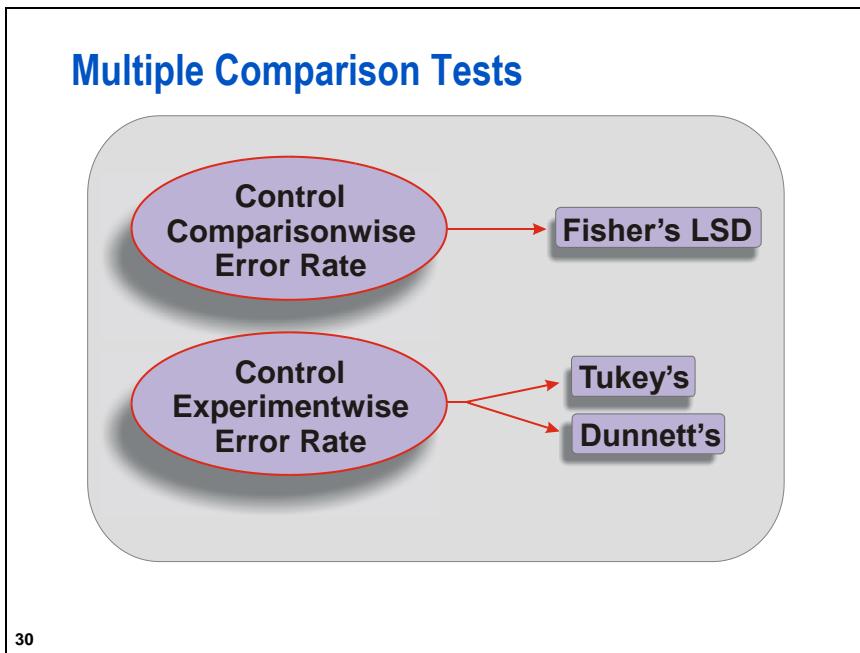
29

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

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

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

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



There are many multiple comparison methods discussed in statistical literature. Among the tests available in SAS are the following:

- Fisher's LSD      to control the comparisonwise error rate
- Tukey's            to control the experimentwise error rate when comparing all pairs of means
- Dunnett's          to control the experimentwise error rate when comparing each mean with the mean of a control group.

SAS generates a variety of multiple comparison tests in addition to those mentioned here. For information about the multiple comparisons tests, refer to the SAS online documentation.



## Multiple Comparison Tests

---

Recall that you determined through analysis of variance that the four schools did not all have the same average **Reading3** test score. However, during data exploration, there were no clear differences noted between the schools. At this point, you want to determine which scores are different among the schools.

There are four schools, so there are a total of six pairwise comparisons that can be made. Because you are interested in conducting all of these comparisons and controlling the experimentwise error rate, Tukey's test is appropriate.

```
proc glm data=STAT2.school;
  class school;
  model reading3=school;
  lsmeans school / pdiff adjust=tukey;
run;                                *ST20Cd04.sas;
```

Selected LSMEANS statement option:

- PDIFF= requests that  $p$ -values for differences of the least squares means be produced. The optional *difftype* specifies which differences to display. Possible values for *difftype* are ALL, CONTROL, CONTROLL, and CONTROLU. The ALL value requests all pairwise differences, and it is the default. The CONTROL value requests the differences with a control that, by default, is the first level of each of the specified least-squares-mean effects.
- ADJUST= requests a multiple comparison adjustment for the  $p$ -values and confidence limits for the differences of LS-means. The ADJUST= option modifies the results of the TDIFF and PDIFF options. Thus, if you omit the TDIFF or PDIFF option, then the ADJUST= option has no effect. By default, PROC GLM analyzes all pairwise differences unless you specify ADJUST=DUNNETT. In that case, PROC GLM analyzes all differences with a control level. The default is ADJUST=T, which really signifies no adjustment for multiple comparisons. ADJUST=TUKEY performs Tukey's studentized range test and it controls experimentwise error rate.

## Partial PROC GLM Output

Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer			
School	Reading3 LSMEAN	LSMEAN Number	
Cottonwood	32.6750000		1
Dogwood	49.8222222		2
Maple	60.2068966		3
Pine	50.0615385		4

Least Squares Means for effect School Pr >  t  for H0: LSMean(i)=LSMean(j) Dependent Variable: Reading3				
i/j	1	2	3	4
1		0.1859	0.0227	0.1243
2	0.1859		0.6812	1.0000
3	0.0227	0.6812		0.6522
4	0.1243	1.0000	0.6522	

The output shows the least squares means in **Reading3** for each school and the *p*-values for the differences between each pair of schools.

There is a statistically significant difference between *Cottonwood* and *Maple*. No other means are found to be statistically different at the 0.05 level of significance.

## C.4 Chapter Summary

---

Before you conduct any type of inferential analysis, it is important to explore and describe your data. Descriptive statistics, histograms, and box plots are some useful tools for exploring your data.

The UNIVARIATE procedure generates descriptive statistics and graphs for numeric variables. The SGPALEL procedure generates side-by-side box plots that enable you to compare the distribution of a variable for two or more groups.

Linear regression is used to assess the significance of the continuous predictor variables when explaining the variability or behavior of a continuous response variable. Linear regression is also used to predict the values of the response variable given the values of the predictor variables.

When you perform a simple linear regression, the null hypothesis is that the simple linear regression does *not* fit the data better than the baseline model. The alternative hypothesis is that the simple linear regression model does fit the data better than the baseline model.

The assumptions for linear regression are

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

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

The null hypothesis for an ANOVA is that there is no treatment effect, or all of the population means are equal. The alternative hypothesis is that there are some treatment effects, or at least one population mean is different from at least one other population mean. For an ANOVA with more than two groups, after you determined that the population means are not all equal, you can determine which means are significantly different using multiple comparison tests.

General form of the UNIVARIATE procedure:

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

General form of the SGPALEL procedure:

```
PROC SGPALEL options;
  PANELBY variable(s) / option(s);
  VBOX response-variable / option(s);
RUN;
```

General form of the REG procedure:

```
PROC REG options;  
  MODEL dependents=regressors / options;  
  PLOT yvariable*xvariable / options;  
RUN;
```

General form of the GLM procedure:

```
PROC GLM options;  
  CLASS variables;  
  MODEL dependents=independents / options;  
  LSMEANS effects / options;  
RUN;
```



# Appendix D Additional Topics

<b>D.1 Nonlinear Regression.....</b>	<b>D-3</b>
Demonstration: Fitting a Nonlinear Regression Model .....	D-15
<b>D.2 Local Regression.....</b>	<b>D-29</b>
Demonstration: Fitting a Local Regression Model .....	D-39
<b>D.3 Modeling Data with Autocorrelation .....</b>	<b>D-51</b>
Demonstration: Detecting Autocorrelation .....	D-56
Demonstration: Modeling Autocorrelation .....	D-62
<b>D.4 Transforming the Dependent Variable as a Remedial Measure .....</b>	<b>D-67</b>
Demonstration: Transformations Based on Relationship between the Variance and Mean.....	D-71
Demonstration: Box-Cox Transformation to Stabilize Nonconstant Variances .....	D-74
Demonstration: Back-Transformation of the Model .....	D-83
Demonstration: Transforming Variables as a Remedial Measure for Departures from Normality .....	D-88
Exercises .....	D-96
<b>D.5 Weighted Least Squares .....</b>	<b>D-98</b>
Demonstration: Weighted Least Squares Using PROC REG.....	D-100
<b>D.6 Evaluating the Importance of Parameters.....</b>	<b>D-105</b>
Demonstration: Evaluating the Importance of Parameters .....	D-106
<b>D.7 A Sample SAS Program for Comparing Model Fit.....</b>	<b>D-107</b>
<b>D.8 Incorrectly Treating Random Effects as Fixed.....</b>	<b>D-109</b>
Demonstration: Comparing Expected Mean Squares for Fixed and Random Effects.....	D-110
<b>D.9 Solutions .....</b>	<b>D-114</b>

Solutions to Exercises .....	D-114
Solutions to Polls and Quizzes .....	D-124

## D.1 Nonlinear Regression

---

### Objectives

- Compare linear and nonlinear regression models.
- Describe different estimation techniques used in nonlinear regression.
- List nonlinear regression model assumptions.
- Identify starting values for parameter estimates.
- Use the NLIN procedure to analyze data.
- Check the assumptions of the nonlinear model.

3

### Linear versus Nonlinear Regression Models

A linear regression model is linear in the parameters. That is, there is only one parameter in each term of the model and each parameter is a multiplicative constant on the independent variables of that term.

A nonlinear model is nonlinear in the parameters.

4

## Examples of Linear Models

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 \ln(X) + \varepsilon$$

$$Y = e^{\beta_0 + \beta_1 X} \varepsilon$$

5

The models shown have a single parameter in each term and the parameters are multiplicative constants of the independent variable.

The last listed model might seem to be nonlinear. However, as shown below, by taking the natural log of each side of the equation, the model can be linearized. Therefore, it is considered to be a linear model.

$$Y = e^{\beta_0 + \beta_1 X} \varepsilon$$

$$\ln(Y) = \ln(e^{\beta_0 + \beta_1 X} \varepsilon)$$

$$\ln(Y) = \ln(e^{\beta_0 + \beta_1 X}) + \ln(\varepsilon)$$

$$\ln(Y) = \beta_0 + \beta_1 X + \ln(\varepsilon)$$

## Examples of Nonlinear Models

$$Y = e^{\beta_0 + \beta_1 X} + \varepsilon$$

$$Y = \beta_1 e^{-\beta_2 X} + \varepsilon$$

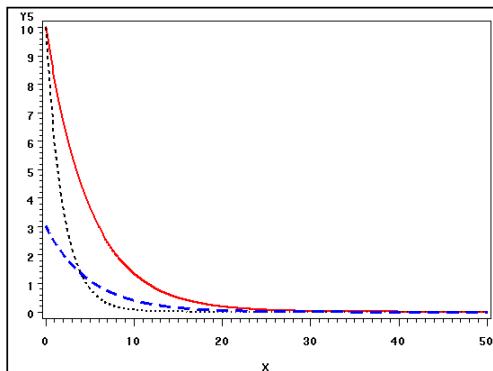
$$Y = \beta_1 + (\beta_2 - \beta_1) e^{-\beta_3 X} + \varepsilon$$

$$Y = \beta_1 X^{\beta_2} + \varepsilon$$

6

Nonlinear models most often occur as a result of some known physical phenomenon, such as exponential growth or decay. All of the models shown here are nonlinear in the parameters and cannot be linearized by a transformation of the equation.

## Examples of Nonlinear Models



$$Y = \beta_1 X^{\beta_2} + \varepsilon$$

7

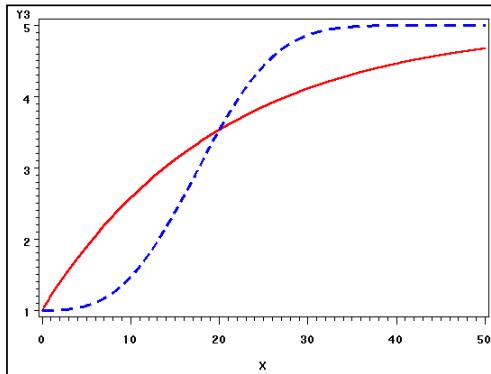
Three examples of the same basic model are shown here. In this model, the parameters have the following geometric interpretation:

$\beta_1$  the  $y$ -intercept

$\beta_2$  the rate parameter

For the model shown here as a solid line,  $\beta_1$  is equal to 10 and  $\beta_2$  is equal to  $-0.2$ . For the model shown as the line with the shorter dashes,  $\beta_1$  is equal to 10 and  $\beta_2$  is equal to  $-0.5$ . For the model shown as the line with longer dashes,  $\beta_1$  is equal to 3 and  $\beta_2$  is equal to  $-0.2$ .

## Examples of Nonlinear Models



$$Y = \beta_1 + (\beta_2 - \beta_1)e^{-(\beta_3 X)^{\beta_4}} + \varepsilon$$

8

This model is a Weibull model.

In this model, the parameters have the following geometric interpretation:

$\beta_1$  the upper horizontal asymptote.

$\beta_2$  the  $y$ -intercept.

$\beta_3$  the rate parameter.

$\beta_4$  the shape parameter. If the shape parameter is greater than one, the curve is sigmoidal.

In the graphs shown here, the parameter values for the solid curve are as follows:

- $\beta_1$  is equal to 5.
- $\beta_2$  is equal to 1.
- $\beta_3$  is equal to 0.05.
- $\beta_4$  is equal to 1.

The parameter values for the dashed curve are as follows:

- $\beta_1$  is equal to 5.
- $\beta_2$  is equal to 1.
- $\beta_3$  is equal to 0.2.
- $\beta_4$  is equal to 3.

Other examples of nonlinear models are summarized below (*JMP Statistics and Graphics Guide*, 2008).

Data Reference	Formula	Model
Myers (1988), p.310	$\frac{\theta_1 \times x}{\theta_2 + x}$	Michaelis-Menten
Draper and Smith (1981), p.522, L	$\theta_1 \times [1 - \exp(-\theta_2 x)]$	
Draper and Smith (1981), p.475	$\theta_1 + (0.49 - \theta_1) \times \exp[-\theta_2 \times (x - 8)]$	
Draper and Smith (1981), p.519, H	$\exp\left\{-\theta_1 \times x_1 \times \exp\left[-\theta_2 \left(\frac{1}{x_2} - \frac{1}{620}\right)\right]\right\}$	
Draper and Smith (1981), p.521, K	$\theta_1 x^{\theta_2}$	
Draper and Smith (1981), p.524, N	$\theta_1 \times [1 - \theta_2 \times \exp(\theta_3 x)]$	
Draper and Smith (1981), p.524, N	$\theta_1 - \ln[1 + \theta_2 \times \exp(-\theta_3 x)]$	Log-logistic
Draper and Smith (1981), p.524, P	$\theta_1 + \frac{\theta_2}{x^{\theta_3}}$	
Draper and Smith (1981), p.524, P	$\ln[\theta_1 \times \exp(-\theta_2 x) + (1 - \theta_1) \times \exp(-\theta_3 x)]$	
Bates and Watts (1988), p. 310	$\theta_1 + \theta_2 \times \exp(\theta_3 x)$	Asymptotic Regression
Bates and Watts (1988), p. 310	$\frac{\theta_1}{1 + \theta_2 \times \exp(\theta_3 x)}$	Logistic
Bates and Watts (1988), p. 310	$\theta_1 \times \exp[-\exp(\theta_2 - \theta_3 - x)]$	Gompertz Growth
Bates and Watts (1988), p. 271	$\frac{\theta_1 \theta_3 \left(x_2 - \frac{x_3}{1.632}\right)}{1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3}$	
Bates and Watts (1988), p. 310	$\frac{\theta_2 \theta_3 + \theta_1 x^{\theta_4}}{\theta_3 + x^{\theta_4}}$	Morgan-Mercer-Flodin
Bates and Watts (1988), p. 310	$\frac{\theta_1}{\left[1 + \theta_2 \times \exp(-\theta_3 x)\right] \left(\frac{1}{\theta_4}\right)}$	Richards Growth
Bates and Watts (1988), p. 274	$\frac{\theta_1}{\theta_2 + x_1} + \theta_3 x_2 + \theta_4 x_2^2 + \theta_5 x_2^3 + (\theta_6 + \theta_7 x_2^2) \times x_2 \times \exp\left(\frac{-x_1}{\theta_8 + \theta_9 x_2^2}\right)$	

## Model Specification

In order for each nonlinear model to be analyzed, you must specify the following:

- the model equation (the MODEL statement)
- names and starting values of the parameters to be estimated (the PARMs statement)

9

Specification of the starting values for the parameters is an important part of the process. Incorrect specification can lead to nonconvergence of the model, or convergence to a local minimum rather than a global minimum. Starting values might be determined based on knowledge of the physical phenomenon of the model or by using a linear regression.

## Estimation Techniques for Nonlinear Regression

- Steepest-descent or gradient (METHOD=GRADIENT)
  - can be fast at the beginning but slow to converge.
- Newton (METHOD=NEWTON)
  - can be computationally expensive.
- Gauss-Newton (METHOD=GAUSS)
  - is the default method.
- Marquardt (METHOD=MARQUARDT)
  - can be an alternative method when the default method does not work.

10

Because nonlinear models cannot be solved explicitly, iterative numerical methods must be used to estimate the parameters. The NLIN procedure provides five different estimation methods. All of the estimation techniques are designed to reduce the residual sum of squares of the model and attempt to find the minimum value. Each method begins with a starting value of the parameters provided by the modeler.

The iterative procedure stops when the sum of squared errors (SSE) converges. Other convergence criteria are possible in the NLIN procedure.

The steepest-descent method moves in the direction that provides the most rapid decrease in the residual sum of squares. Although this method might seem to be very fast at the outset, it can be very slow to converge.

The Newton method regresses the residuals onto a function of the first and second partial derivatives of the model with respect to the parameters until the SSE converges.

The Gauss-Newton and Marquardt methods regress the residuals onto the partial derivatives of the model until the estimates converge. The Gauss-Newton method performs reasonably well under most circumstances and is the default method.

The Marquardt method is a compromise between the Gauss-Newton and steepest-descent methods. It is equivalent to performing a series of ridge regressions and is useful when the parameter estimates are highly correlated or the objective function is not well approximated by a quadratic.

 Beginning in SAS®9, METHOD=DUD is no longer supported. PROC NLIN uses METHOD=GAUSS if you specify METHOD=DUD. In the earlier releases, the DUD (*Doesn't Use Derivatives*) method is similar to the Gauss-Newton method, except that the derivatives are estimated from the history of iterations rather than supplied.

## Potential Lack of Convergence of Nonlinear Estimates

Convergence might not be obtained under certain conditions. These might include the following:

- incorrect specification of the model
- poor initial starting values
- over-defined models
- inappropriate estimation methods
- insufficient data

11

Incorrect model specification is when the data do not exhibit the same relationship as the one given in the model.

As mentioned earlier, poor initial starting values can cause lack of convergence or convergence to a local minimum.

A model that is over defined is one that has more parameters than are necessary for the relationship. If two or more parameters are very close in value, they could and should be represented by a single parameter.

Although the default estimation method (GAUSS) works well in many cases, it might not yield convergence in other cases. Different estimation methods might work better.

Finally, there might be insufficient data to generate a model. In this case, additional data must be collected.

More information can be found in Rawlings, J.O., Pantula, S.G., and Dickey, D.A. (1998).

## Nonlinear Regression Assumptions

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i$$

- $\varepsilon_i$  is normally and independently distributed with a mean of zero and a constant variance of  $\sigma^2$ .
- The sample size  $n$  is reasonably large. The adequacy of the sample size can be assessed by
  - quick convergence
  - a small value for Hougaard's measure of skewness
  - confirmed close-to-linear distribution of regression parameter estimates using bootstrap.

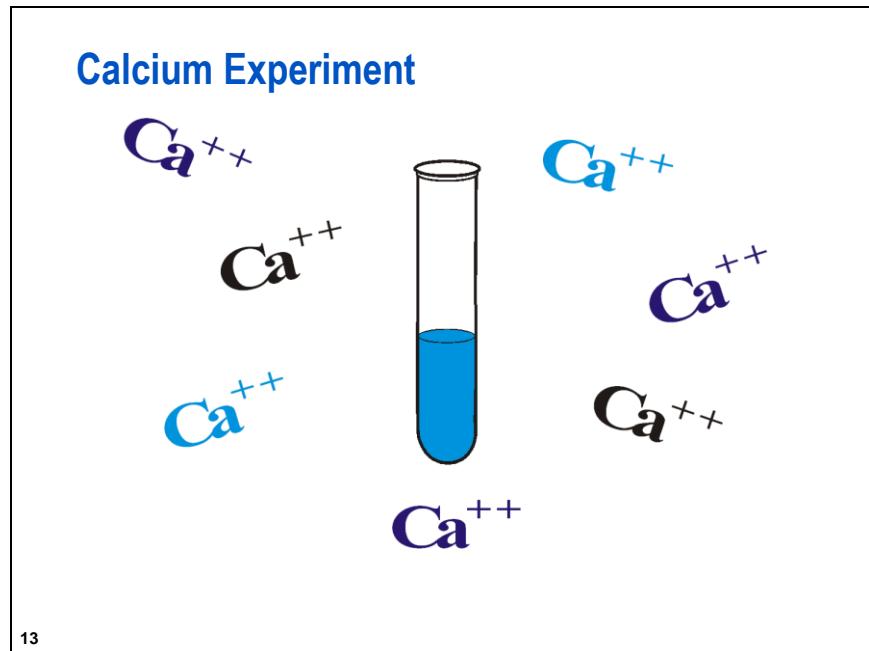
12

Most properties of linear least squares regressions apply only approximately or asymptotically for nonlinear least squares regression models. Inferences about the regression parameters in nonlinear regression are usually based on large-sample theory. This theory tells you that the least squares (and maximum likelihood) estimators for nonlinear regression models with independent normal error terms and constant variance, when the sample size is large, are approximately normally distributed and almost unbiased, and have almost minimum variance. This large-sample theory also applies when the error terms are not normally distributed (Neter, Kutner, Nachtsheim, and Wasserman 1996).

How do you know whether your sample size is large enough? Unfortunately, no simple rule exists that tells you when it is appropriate to use the large-sample theory to make inference about the parameter estimates and when it is not appropriate. For some nonlinear regression models, the sample size can be quite small for the large-sample approximation to be good. For other nonlinear regression models, the sample size might need to be quite large. However, a number of guidelines were developed to assess the appropriateness of using the large-sample theory.

1. Quick convergence of the iterative procedure in finding the estimates of the nonlinear regression parameters is often an indication that the asymptotic properties of the regression parameters are applicable. Slow convergence suggests caution and consideration of other guidelines before large-sample inference is used.
2. Several measures were developed for providing guidance about the appropriateness of the use of the inference. Hougaard's measure of skewness is available in PROC NLIN. An indication of little skewness ( $<0.1$ ) supports the approximate normality of the sampling distributions of the regression parameters and consequently the applicability of the inference provided by the procedure.

3. Bootstrap sampling provides a direct means of examining whether the sampling distributions of the nonlinear regression parameter estimates are approximately normal, whether the variances of the sampling distributions are near the variances for the linear approximated model, and whether the bias in each parameter estimates is fairly small. If so, the sampling behavior of the nonlinear regression estimates is said to be close-to-linear and the inference might be appropriate to use.



An experiment was conducted by Howard Grimes, Washington State University, to determine the amount of radioactive calcium in cells. In particular, the relationship between the amount of time the cells were in a suspension and the amount of radioactive calcium in the cells were of interest.

The researchers believed that the response would follow a Weibull growth model:

$$y_i = a + (b - a)e^{-(cx)^d} + \varepsilon_i$$

, which is clearly a nonlinear model.

## The Data

Suspension	Time	Calcium
1	0.45	0.34170
2	0.45	-0.00438
3	0.45	0.82531
4	1.30	1.77967
5	1.30	0.95384
6	1.30	0.64080
7	2.40	1.75136
8	2.40	1.27497
9	2.40	1.17332
10	4.00	3.12273
11	4.00	2.60958
12	4.00	2.57429
13	6.10	3.17881
14	6.10	3.00782
15	6.10	2.67061
16	8.05	3.05959
17	8.05	3.94321
18	8.05	3.43726

and so on;

14

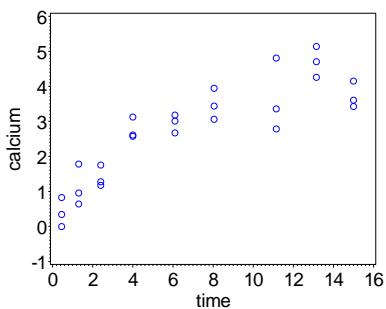
The data are stored in the **STAT2.calcium** data set. These are the variables in the data set:

**Suspension** the suspension identification number

**Time** time in hours since the cells were put in the suspension

**Calcium** the amount of radioactive calcium in the cells

## Calcium versus Time



$$y_i = a + (b - a)e^{-(cx_i)^d} + \varepsilon_i$$

Annotations for the equation:

- y-intercept**: Points to the term  $a$ .
- shape**: Points to the term  $(cx_i)^d$ .
- rate**: Points to the term  $e^{-}$ .
- upper horizontal asymptote**: Points to the term  $b$ .

15

In order to analyze the data set **STAT2.calcium**, you must first specify the model.

The researchers in this case have scientific reason to believe that the model should be

$y_i = a + (b - a)e^{-(cx_i)^d} + \varepsilon_i$ . Therefore, four parameters ( $a$ ,  $b$ ,  $c$ , and  $d$ ) must be estimated. The independent variable, **Time**, is  $x$ , and the dependent variable, **Calcium Concentration**, is  $y$ .

After the model is specified, the next step is to determine the starting values of the parameters to be estimated. This task can be simplified by examining the terms in the model. The distribution is as follows:

$$y_i = a + (b - a)e^{-(cx_i)^d} + \varepsilon_i$$

where

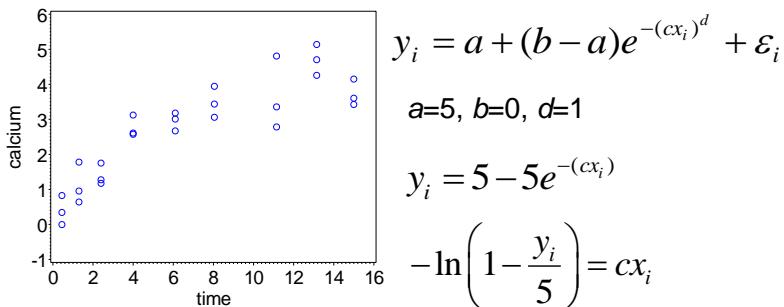
$a$  = the upper horizontal asymptote

$b$  = the  $y$ -intercept

$c$  = rate parameter

$d$  = shape parameter. (If  $d$  is greater than one, then it implies a sigmoidal curve.)

### Calcium versus Time – Starting Values



16

From the scatter plot, you can see that the  $y$ -intercept is close to zero and that the horizontal asymptote is close to 5. Therefore, a good starting value for  $a$  is 5, and a good starting value for  $b$  is 0. The data do not appear to be sigmoidal. Therefore, a good starting value for  $d$  is 1.

Based on these starting values for  $a$ ,  $b$ , and  $d$ , the REG procedure can be used to obtain a starting value for  $c$ . Using the starting values for  $a$ ,  $b$ , and  $d$ , the model reduces to  $y_i = 5 - 5e^{-(cx_i)}$ . Taking the natural logarithm of both sides simplifies this to  $-\ln(1 - \frac{y_i}{5}) = cx_i$ .

## The NLIN Procedure

General form of the NLIN procedure:

```
PROC NLIN options;  

  PARAMETERS parameter=values...;  

    program statements;  

  MODEL dependent=expression;  

  OUTPUT OUT=SAS-data-set keyword=names;  

RUN;
```

17

Selected NLIN procedure statements:

**PARAMETERS** identifies a parameter to be estimated, both in subsequent procedure statements and in the output. Several parameter names and values can appear. The parameter names must all be valid SAS names and must not duplicate the names of any variables in the data set to which the NLIN procedure is applied. *Values* specify the possible starting values of the parameter.

**MODEL** defines the prediction equation by declaring the dependent variable and defining an expression that evaluates predicted values. The expression can include parameter names, variables in the data set, and variables created by program statements in the NLIN procedure. Most operators or functions that can be used in a DATA step can also be used in the MODEL statement. The computational methods assume that the model is a continuous function and a smooth function of the parameters.

**OUTPUT** specifies an output data set to contain statistics calculated for each observation.

PROC NLIN supports many statements that are similar to SAS programming statements used in a DATA step. However, there are some differences in capabilities. For more information, refer to PROC NLIN in the SAS online documentation.



## Fitting a Nonlinear Regression Model

As discussed previously, a reasonable set of starting values is  $a=5$ ,  $b=0$ , and  $d=1$ . The starting value for  $c$  can be obtained by fitting a linear model to  $-\ln(1 - \frac{y_i}{5}) = cx_i$ .

First, use a DATA step to transform the response variable **calcium**, and then use PROC REG to estimate the parameter  $c$ .

```
data calcium;
  set STAT2.calcium;
  trans=-log((1-(calcium/5)));
run;

proc reg data=calcium;
  model trans=time / noint;
run;                                *ST20Dd01.sas;
```

Selected MODEL statement option:

**Noint** suppresses the intercept term that is otherwise included in the model.

Partial PROC REG Output

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Time	1	0.13679	0.01244	10.99	<.0001

The estimate for  $c$  from the regression is 0.13679. Therefore, a good starting value for  $c$  in your nonlinear regression is 0.14.

Now that starting values for all of the parameters are determined, you can do the nonlinear regression.

```
proc nlin data=STAT2.calcium hougaard;
  parms a=5 b=0 c=.14 d=1;
  model calcium=a+(b-a)*exp(-(c*time)**d);
  output out=check p=est;
run;

proc sgplot data=check;
  scatter y=calcium x=time;
  series y=est x=time / lineattrs=(color=blue pattern=1);
run;
quit;                                *ST20Dd01.sas;
```

Selected PROC NLIN statement options:

**HOUGAARD** adds Hougaard's measure of skewness to the parameter estimation table.

**METHOD=** specifies the iterative method that PROC NLIN uses. The available methods are the GAUSS, MARQUARDT, NEWTON, and GRADIENT methods. If you omit the METHOD= option, METHOD=GAUSS is used.

Selected NLIN procedure statement:

**BOUNDS** restrains the parameter estimates within specified bounds. Each bound contains a list of parameters, an inequality comparison operator, and a value.

### PROC NLIN Output

**The NLIN Procedure**  
**Dependent Variable Calcium**  
**Method: Gauss-Newton**

Iterative Phase					
Iter	a	b	c	d	Sum of Squares
0	5.0000	0	0.1400	1.0000	8.6107
1	4.4450	0.0415	0.1733	1.0271	7.9200
2	4.2097	0.0813	0.2051	1.0623	7.5274
3	4.2435	0.0816	0.2101	1.0642	7.4587
4	4.2470	0.0798	0.2102	1.0627	7.4585
5	4.2469	0.0800	0.2102	1.0628	7.4585

NOTE: Convergence criterion met.

The default method is the Gauss-Newton method. There were five iterations after which the convergence criterion was met. PROC NLIN determines convergence using the relative offset measure of Bates and Watts. When this measure is less than  $10^{-5}$ , convergence is declared.

Estimation Summary	
<b>Method</b>	Gauss-Newton
<b>Iterations</b>	5
<b>Subiterations</b>	1
<b>Average Subiterations</b>	0.2
<b>R</b>	7.557E-6
<b>PPC(b)</b>	0.000228
<b>RPC(b)</b>	0.001696
<b>Object</b>	3.975E-9
<b>Objective</b>	7.458546
<b>Observations Read</b>	27
<b>Observations Used</b>	27
<b>Observations Missing</b>	0

The estimation summary displays a summary of the procedure including several convergence measures: R, PPC, RPC, and Object.

- The *R measure* is the relative offset convergence measure of Bates and Watts.
- The *PPC (Prospective Parameter Change) measure* indicates the parameter that has the largest PPC values of all of the parameters (in this case parameter *B*). It indicates the relative amount that the parameter estimate would change if an additional iteration were performed.
- The *RPC (Retrospective Parameter Change) value* indicates the relative amount that the parameter estimate of *b* changed in the last iteration.
- The *Object measure* indicates the relative amount the objective function value changed in the last iteration.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	3	45.7751	15.2584	47.05	<.0001
Error	23	7.4585	0.3243		
Corrected Total	26	53.2336			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits		Skewness
			Lower Limit	Upper Limit	
a	4.2469	0.5458	3.1178	5.3759	3.9654
b	0.0800	0.6745	-1.3154	1.4753	-2.2177
c	0.2102	0.0542	0.0980	0.3223	-0.4485
d	1.0628	0.4867	0.0560	2.0696	-0.1045

The null hypothesis is that all the parameter estimates equal zero. The significant *p*-value (<0.0001) indicates that at least one parameter estimate is significantly different from zero.

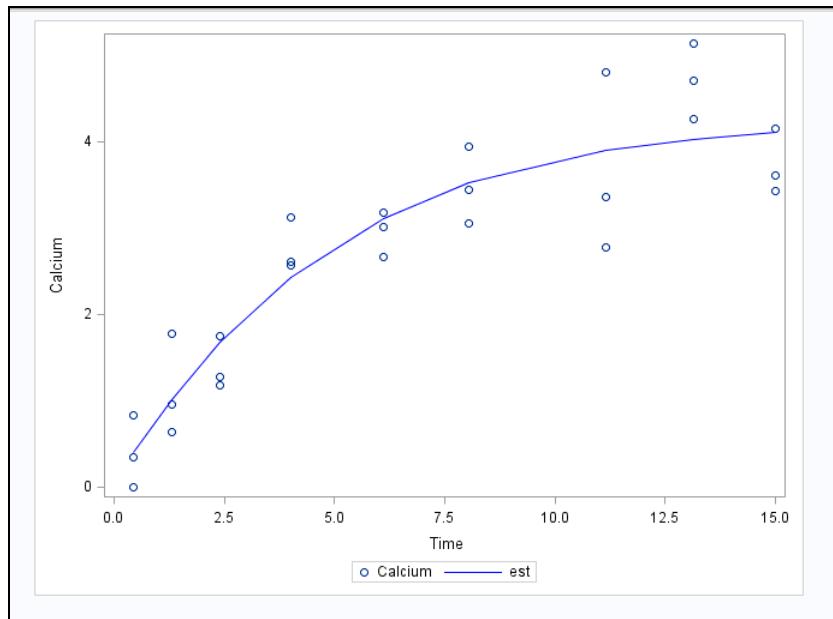
The parameter estimates and their asymptotic standard errors and 95% confidence intervals are also displayed. Hougaard's measure of skewness can be used to assess whether a parameter is close to linear or whether it contains considerable nonlinearity. A "close-to-linear" nonlinear regression model, first described by Ratkowsky (1990), is a model that produces parameters having properties similar to those produced by a linear regression model. That is, the least squares estimates of the parameters are close to being unbiased, normally distributed, minimum variance estimators. If the absolute value of the Hougaard's measure of skewness is

- less than .1, then the estimator is very close to linear in behavior
- between .1 and .25, then the estimator is reasonably close to linear
- between .25 and 1, then the skewness is very apparent
- greater than one, then the nonlinear behavior is considerable.

If the estimator exhibits nonlinear behavior, there is bias in the reported standard error and confidence limits that render them invalid. With the exception of parameter *d*, the skewness measures are too large to consider the estimates linear, and the confidence intervals should not be used. This problem can often be remedied with reparameterization of the model form.

Approximate Correlation Matrix				
	a	b	c	d
a	1.000000	-0.5746506	-0.6536171	-0.8084012
b	-0.5746506	1.0000000	-0.1582127	0.8659924
c	-0.6536171	-0.1582127	1.0000000	0.2051860
d	-0.8084012	0.8659924	0.2051860	1.0000000

The correlation matrix is a measure of the correlation between the parameters. If any correlations are above 0.98 or 0.99, the parameters are excessively correlated. This can be an indication that the model is over-parameterized and the model should be simplified. There do not appear to be any excessively high correlations in this case.



The curve appears to fit the data reasonably well, although at two times, all of the observed points are above the curve. The data seem to indicate that the calcium begins to fall after some amount of time. This indication in the data should be discussed with subject-matter experts. If this is indeed a physical phenomenon, then the model should be revised to allow for this decline.

When performing nonlinear regression, you do not have to fit each parameter in the equation. Instead, you can fix one or more of the parameters to constant values. Recall that the parameter estimate for  $d$  was very close to one. In addition, the approximate 95% confidence interval included one. Because  $d$  is an exponent in the model, if it is equal to one, the parameter is not necessary. Therefore,  $d$  should be removed from the model, which makes it a simple exponential growth model. It is also appropriate to adjust the starting values for the remaining parameters to be the estimates from the previous model.

```
proc nlin data=STAT2.calcium hougaard;
  parms a=4.25 b=0.08 c=.21;
  model calcium=a+(b-a)*exp(-(c*time));
run; *ST20Dd01.sas;
```

Partial PROC NLIN Output

The NLIN Procedure				
Dependent Variable Calcium				
Method: Gauss-Newton				
Iterative Phase				
Iter	a	b	c	Sum of Squares
0	4.2500	0.0800	0.2100	7.5120
1	4.3091	-0.00080	0.2085	7.4645
2	4.3089	-0.00098	0.2086	7.4645

Only two iterations were required to reach convergence in this case.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	45.7691	22.8846	73.58	<.0001
Error	24	7.4645	0.3110		
Corrected Total	26	53.2336			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits		Skewness
			a	b	
a	4.3089	0.3492	3.5882	5.0296	0.9790
b	-0.00098	0.3477	-0.7186	0.7166	-0.1738
c	0.2086	0.0567	0.0916	0.3256	0.3524

Approximate Correlation Matrix			
	a	b	c
a	1.0000000	0.4659636	-0.8712185
b	0.4659636	1.0000000	-0.7059815
c	-0.8712185	-0.7059815	1.0000000

The model is significant and the mean square error decreased slightly. Hougaard's skewness statistics still indicate that the parameter estimates for parameters *a* and *c* behave nonlinearly. This makes the approximate confidence intervals for these parameters unreliable. The correlation matrix does not indicate any problems with collinearity of the parameters.

As a final step in developing this model, *b* is not significantly different from zero. Therefore, it can be removed from the model, and again adjust the initial parameter estimates.

```
proc nlin data=STAT2.calcium hougaard;
  parms a=4.31 c=.21;
  model calcium=a+(-a)*exp(-(c*time));
run; *ST20Dd01.sas;
```

Partial PROC NLIN Output

**The NLIN Procedure**  
**Dependent Variable Calcium**  
**Method: Gauss-Newton**

Iterative Phase			
Iter	a	c	Sum of Squares
0	4.3100	0.2100	7.4665
1	4.3094	0.2085	7.4645
2	4.3094	0.2085	7.4645

NOTE: Convergence criterion met.

Again, only two iterations were required to reach convergence in this case.

**Note:** An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	240.8	120.4	403.22	<.0001
Error	25	7.4645	0.2986		
Uncorrected Total	27	248.3			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits		Skewness
			a	c	
a	4.3094	0.3029	3.6855	4.9332	0.6047
c	0.2085	0.0393	0.1275	0.2895	0.3816

Approximate Correlation Matrix		
	a	c
a	1.0000000	-0.8654763
c	-0.8654763	1.0000000

The model is significant and the mean square error decreased slightly. Hougaard's skewness statistics still indicate that the parameter estimates for the parameters behave nonlinearly. This makes the approximate confidence intervals unreliable. The correlation matrix does not indicate any problems with collinearity of the parameters.

Recall that in a previous chapter, several possible transformations for variables were discussed. These same transformations can be used on the parameters for this model to attempt to make the parameters close to linear. The appropriate transformation can be determined through a process of trial and error, observing the skewness statistics. For example, if after a transformation, a formerly positive skewness statistic becomes negative, the transformation that is used is too extreme and a less extreme transformation should be chosen.

For this model, the first transformation chosen was the natural logarithm of the parameters. This transformation reduced the skewness of the parameter *a* somewhat and made the skewness of the parameter *c* negative. After several iterations, the chosen transformations were the inverse-squared transformation for *a* and the cube-root transformation for *c*. The program to accomplish this is illustrated below.

```
proc nlin data=STAT2.calcium hougaard;
  parms a=4.31 c=.21;
  model calcium=a+(-a)*exp(-(c*time));
  output out=ac parms=A C;
run;

data _null_;
  set ac(obs=1);
  call symput('ap',1/a**2);
  call symput('cp',c**(1/3));
run;
```

```

ods output ParameterEstimates=est;
proc nlin data=STAT2.calcium hougaard;
  parms ap=&ap cp=&cp;
  a=1/sqrt(ap);
  c=cp**3;
  model calcium=a+(-a)*exp(-(c*time));
  output out=check r=residual p=predicted;
run;                                *ST20Dd01.sas;

```

First, rerun the procedure to create an output data set that contains the parameter estimates. Then, use a DATA step to transform these parameter estimates, create macro variables **ap** and **cp**, and use them as the initial values for the new (transformed) parameters *ap* and *cp* in the PARMS statement in PROC NLIN. The NLIN procedure uses program statements to transform the parameters to the original ones. In this way, the MODEL statement itself does not change in syntax, although the parameters are now *ap* and *cp*, which are related to the original parameters *a* and *c* through the corresponding transformations. The OUTPUT statement is used to save the residuals and predicted values for a model assumptions check. The ODS OUTPUT statement is used to save the parameter estimates for *ap* and *cp* to a data set for further processing.

**The NLIN Procedure**  
**Dependent Variable Calcium**  
**Method: Gauss-Newton**

Iterative Phase			
Iter	ap	cp	Sum of Squares
0	0.0538	0.5930	7.4645

NOTE: Convergence criterion met.

Estimation Summary	
Method	Gauss-Newton
Iterations	0
R	1.35E-6
PPC(ap)	8.267E-7
RPC	.
Object	.
Objective	7.464514
Observations Read	27
Observations Used	27
Observations Missing	0

You provided the solutions as the starting values. Therefore, no iterations are needed.

**Note:** An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	240.8	120.4	403.22	<.0001
Error	25	7.4645	0.2986		
Uncorrected Total	27	248.3			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits		Skewness
ap	0.0538	0.00757	0.0383	0.0694	0.0280
cp	0.5930	0.0373	0.5162	0.6697	0.00438

Approximate Correlation Matrix		
	ap	cp
ap	1.0000000	0.8654763
cp	0.8654763	1.0000000

The mean square error is exactly the same as the previous model. This is because this model is the same as the previous model with different parameterizations.

The Hougaard skewness statistics for the reparameterized model indicate that the parameters are very close to linear. Therefore, the confidence intervals are appropriate. Because the parameter estimates had a particular meaning in the original model, it might be desirable to obtain the estimates and the confidence intervals for those original parameters. This can be done by back transforming the estimates and the confidence intervals.

```

data est;
  set est;
  if parameter='ap' then do;
    Orig_Parameter='a';
    Orig_Estimate=1/sqrt(estimate);
    Orig_LowerCL=1/sqrt(upperCL);
    Orig_UpperCL=1/sqrt(lowerCL);
    output;
  end;

  if parameter='cp' then do;
    Orig_Parameter='c';
    Orig_Estimate=estimate**3;
    Orig_LowerCL=lowerCL**3;
    Orig_UpperCL=upperCL**3;
    output;
  end;
run;

proc print data=est;
  var orig_parameter orig_estimate orig_lowerCL orig_upperCL;
run;                                         *ST20Dd01.sas;

```

## PROC PRINT Output

Obs	Orig_Parameter	Orig_Estimate	Orig_LowerCL	Orig_UpperCL
1	a	4.30936	3.79485	5.11264
2	c	0.20848	0.13753	0.30041

The parameter estimates are the same as the ones obtained from the original model, but the confidence intervals are different. The 95% confidence intervals from the original model were [3.6855, 4.9332] for  $a$  and [0.1275, 0.2895] for  $c$ . They are both symmetric about the parameter estimates, which might not be appropriate because of the values of skewness statistics. The confidence intervals shown above ([3.79485, 5.11264] for  $a$  and [0.13753, 0.30041] for  $c$ ) are based on transformed parameters. They are asymmetric about the estimates and are more appropriate than the original ones because they are based on parameters that are close to linear. Refer to Seber, G.A.F., and Wild, C.J. (1989) for more discussions about nonlinear regression models.

The fitted model for the **calcium** data is predicted calcium =  $4.31 - 4.31 \times e^{(-0.21 \times \text{time})}$ .

You might want to evaluate the nonlinear regression assumptions by

- plotting the residuals versus the predicted values
- plotting the data points with the nonlinear model superimposed on the plot
- generating a histogram and normal probability plot of the residuals to check for normality
- testing the residuals for normality.

```

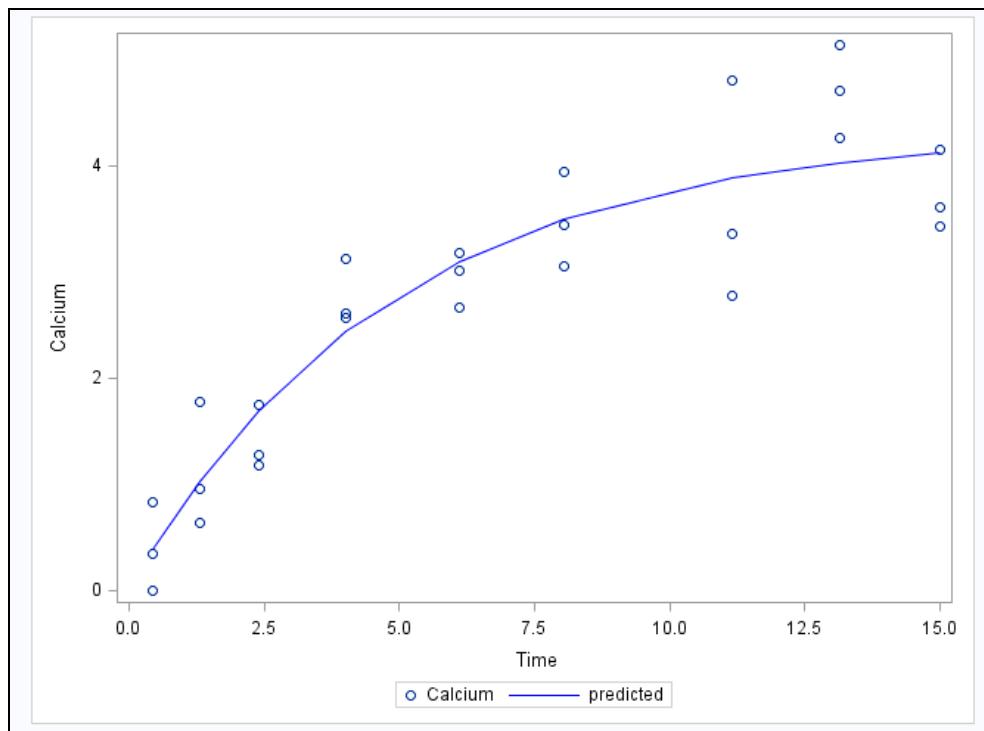
proc sgplot data=check;
  scatter y=calcium x=time;
  series y=predicted x=time / lineattrs=(color=blue pattern=1);
run;
quit;                                *ST20Dd01.sas;

proc sgplot data=check;
  scatter y=residual x=predicted;
  refline 0;
run;                                    *ST20Dd01.sas;

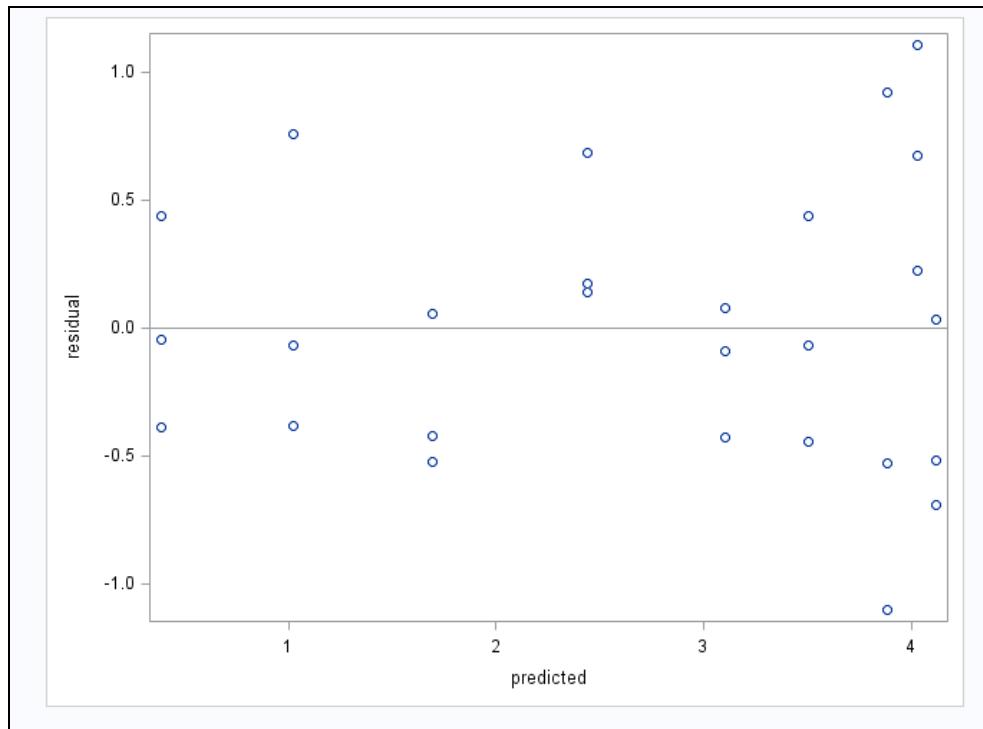
proc univariate data=check;
  var residual;
  histogram / normal;
  probplot / normal(mu=est sigma=est);
run;                                    *ST20Dd01.sas;

```

## PROC SGLOT Output



The curve is very similar to the first model that you fit to the data. It appears to fit the data reasonably well, although at two times, all of the observed points are above the curve. Again, the data seem to indicate that the calcium begins to fall after some amount of time. This indication in the data should be discussed with subject-matter experts. If this is indeed a physical phenomenon, then the model should be revised to allow for this decline.



The graph of the residuals versus the predicted values might or might not indicate that the variances are equal. You might suspect that more variability is associated with the larger predicted values. However, the violation of equal variance assumption, if any, appears to be minor.

#### Partial PROC UNIVARIATE Output

##### The UNIVARIATE Procedure Variable: residual

Moments			
N	27	Sum Weights	27
Mean	-0.0000942	Sum Observations	-0.0025434
Std Deviation	0.53581405	Variance	0.28709669
Skewness	0.24447358	Kurtosis	-0.3554961
Uncorrected SS	7.46451428	Corrected SS	7.46451404
Coeff Variation	-568802.71	Std Error Mean	0.10311746

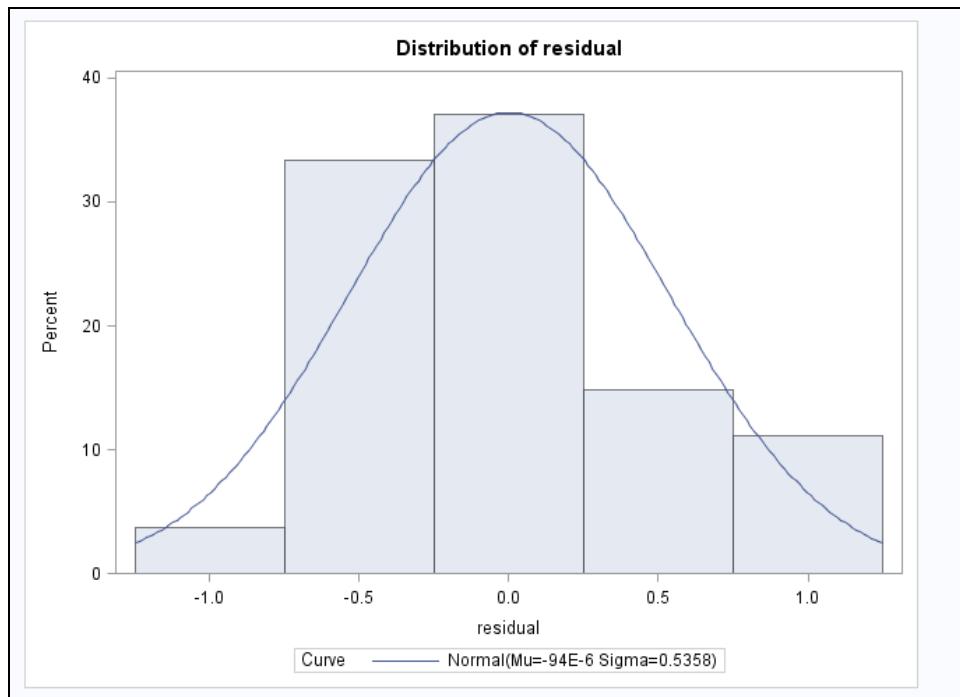
The skewness and kurtosis are close to zero. The residuals are skewed slightly to the right and the tails are slightly light compared with a normally distributed data.

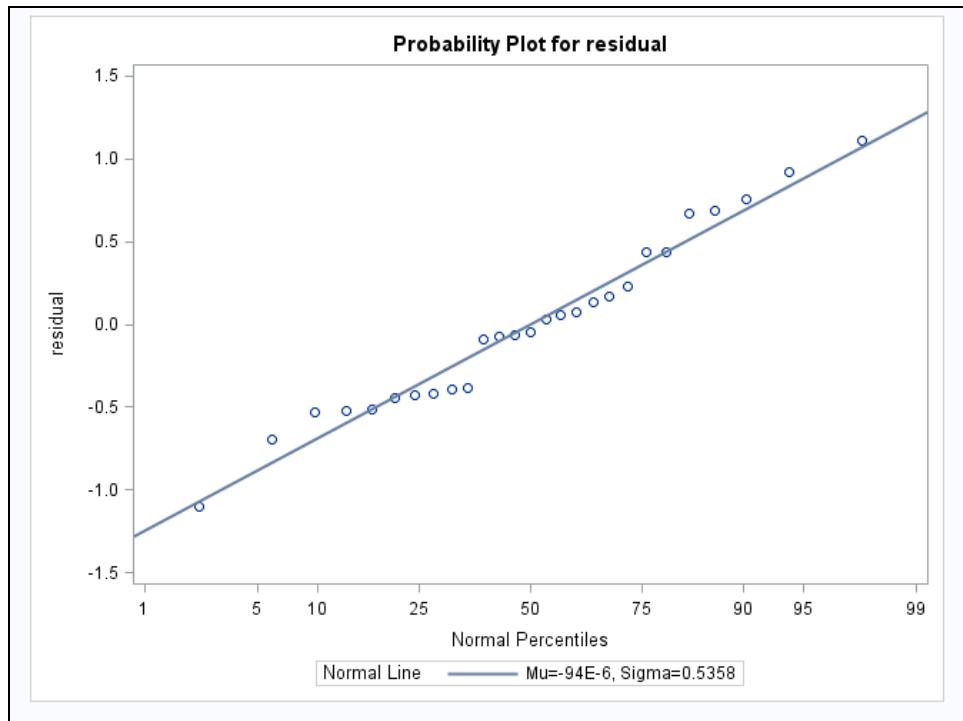
**The UNIVARIATE Procedure**  
**Fitted Normal Distribution for residual**

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	-0.00009
Std Dev	Sigma	0.535814

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic	p Value		
Kolmogorov-Smirnov D	0.13251318	Pr > D	>0.150	
Cramer-von Mises W-Sq	0.05813583	Pr > W-Sq	>0.250	
Anderson-Darling A-Sq	0.36179391	Pr > A-Sq	>0.250	

The tests for normality indicate that the residuals are normally distributed.





The histogram and normal probability plot do not indicate any serious violations of the normality assumption.

All indications are that this model meets the assumptions of the analysis. You can make the inference about the parameter estimates based on the procedure output.

## Summary

A linear regression model is linear in the parameters. That is, there is only one parameter in each term of the model and each parameter is a multiplicative constant on the independent variables of that term. A nonlinear model is nonlinear in the parameters.

For each nonlinear model, you must specify the model equation and starting values of the parameters to be estimated. Determination of the model equation is generally done based on subject-matter knowledge of the behavior of the variables. Starting values of the parameter estimates can sometimes be determined based on knowledge of the physical phenomenon of the model or through examining the data. Incorrect model specification or poor initial starting values can sometimes lead to nonconvergence.

Because nonlinear models cannot be solved explicitly, iterative numerical methods must be used to estimate the parameters. There are several estimation techniques. Each is designed to reduce the residual sum of squares of the model and attempts to find the minimum value. The default method in the NLIN procedure is the Gauss-Newton method.

As with all models, after a model is generated, it is important to check for violations of the assumptions, such as nonnormality of the residuals or nonconstant variance. For nonlinear models, it is also important to check the parameters to see whether they are close to linear. If they exhibit nonlinear behavior, there is bias in the reported standard errors and confidence limits. This problem can often be remedied with reparameterization of the model form.

General form of the NLIN procedure:

```
PROC NLIN options;  
  PARAMETERS parameter=values...;  
  program statements;  
  MODEL dependent=expression;  
  OUTPUT OUT=SAS-data-set keyword=names...;  
RUN;
```

## D.2 Local Regression

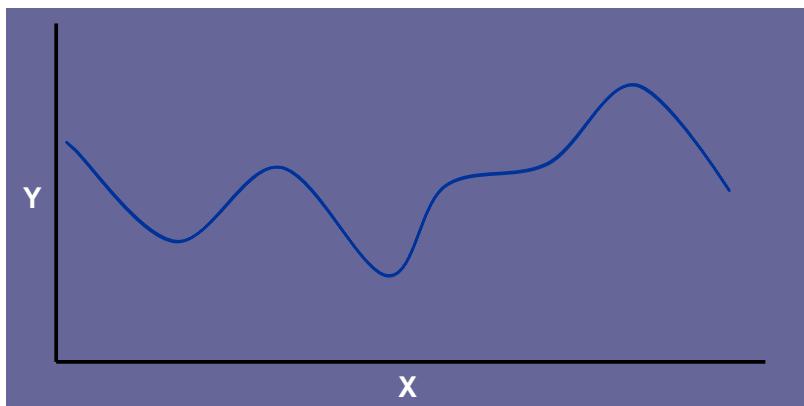
### Objectives

- Explain the need for local regression.
- Describe the process of fitting a local regression.
- Explain the process of kd tree construction in choosing the local regression fit points.
- Use the LOESS procedure to fit a local regression model.

20

### Idea of Local Regression

It is difficult to find an appropriate parametric form for complicated curves.

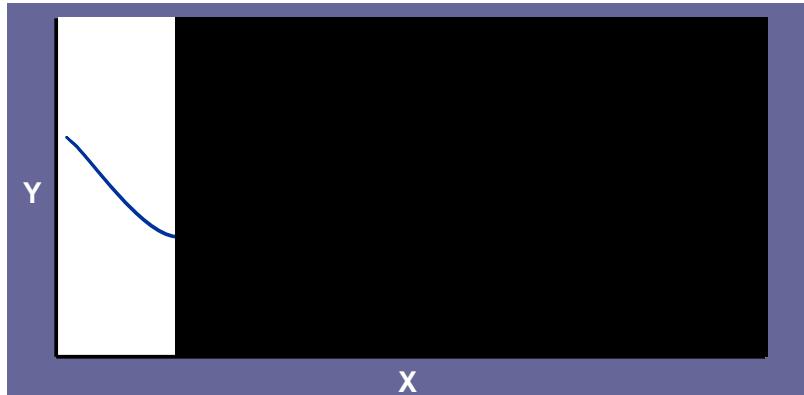


21

At times, the relationship between the response variable and the variable of interest is sufficiently complicated that it cannot be estimated by any obvious equation.

## Idea of Local Regression

Locally such curves can be well approximated by low-degree polynomials.

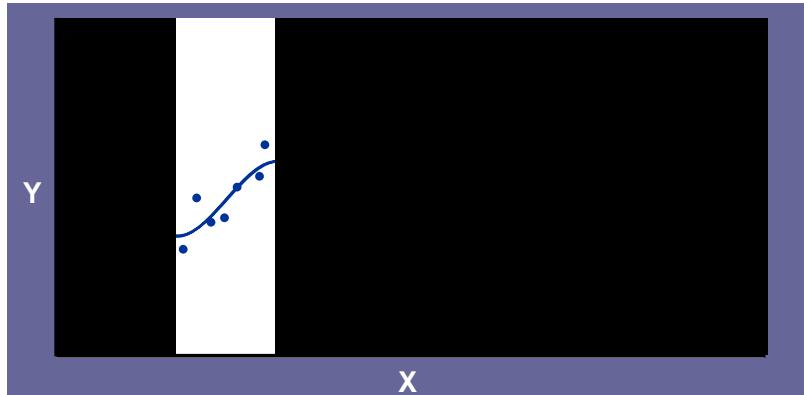


23

The idea of local regression is that near any chosen value of X, the regression function can be locally approximated by the value of a function in some parametric class, such as a linear or quadratic function.

## Idea of Local Regression

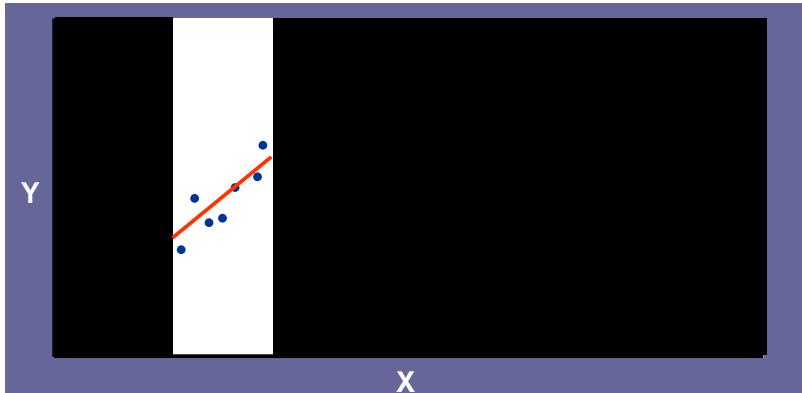
Consider this earlier window, which shows the true curve and some data points.



30

## Idea of Local Regression

Find a local fit by linear regression.



32

The local approximation is found by fitting a regression surface to the data points within a chosen neighborhood.

## How Do You Define the Neighborhood?

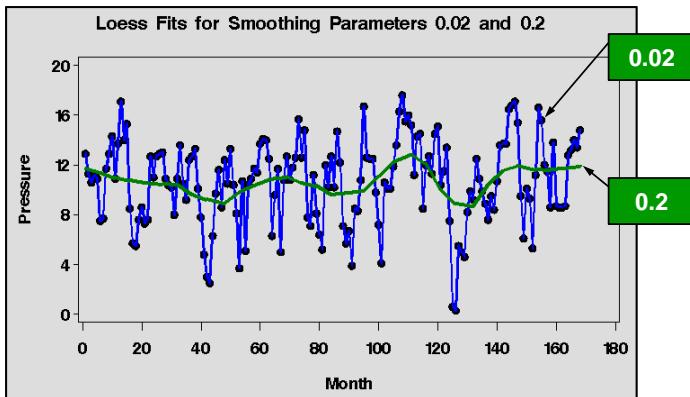
The smoothing parameter

- defines the fraction of data that is considered in the local neighborhood.
- is always between 0 and 1.
- controls the smoothness of the curve. As the value decreases, the model becomes less smooth and more noisy. As the value increases, the model becomes more smooth and less informative.

33

In local regression, the radius of each local neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. This fraction of the data, called the *smoothing parameter*, controls the smoothness of the estimated surface. If the smoothing parameter is too small, the curve is over fit and too noisy. If the smoothing parameter is too large, the curve might miss some important features of the relationship.

## Over and Under Smoothed LOESS Fits



34

## Smoothing Parameter Selection Methods

The smoothing parameter value is selected to yield a minimum in one of two types of criteria.

- a criterion that incorporates both the tightness of the fit and model complexity
  - generalized cross validation (GCV)
  - bias corrected Akaike information criteria (AICC)
  - bias corrected Akaike information criteria (AICC1)
- a criterion that is based on degrees of freedom
  - DF1(target)
  - DF2(target)
  - DF3(target)

35

Smoothing parameter values are selected by minimizing some criterion. There are two types of criteria to be minimized in PROC LOESS. One type incorporates both the tightness of the fit and the model complexity. There is a penalty designed to decrease with increasing smoothness of fit. Included among these methods are generalized cross validation (GCV) and two bias-corrected Akaike information criteria (AICC and AICC1). It was shown that the bias corrected Akaike information criteria avoid the tendency to undersmooth that often occurs when using the classical Akaike information criterion or generalized cross validation. AICC and AICC1 differ in how the bias is corrected. Generally speaking, AICC1 is much more computationally expensive than AICC. AICC can be regarded as an approximation of AICC1 and generally performs well in all circumstances.

A second class of methods seeks to set an approximate measure of model degrees of freedom to a specified target value. These methods are useful for making meaningful comparisons between loess fits and other nonparametric and parametric fits. The approximate model degrees of freedom in a nonparametric fit is a number that is analogous to the number of free parameters in a parametric model. There are three commonly used measures of model degrees of freedom in nonparametric models. These criteria are Trace ( $L$ ), Trace ( $L^T L$ ), and 2 Trace  $L$  - Trace ( $L^T L$ ). A discussion of their properties can be found in Hastie and Tibshirani (1990). You invoke these methods by specifying the `SELECT=DFCriterion(target)` option in the MODEL statement, where `DFCriterion` is one of DF1, DF2, or DF3.

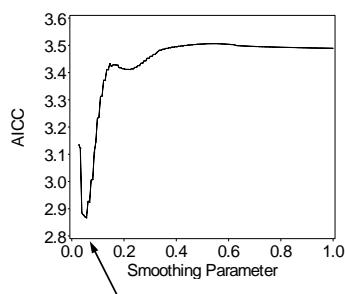
## Smoothing Parameter Selection

In the MODEL statement in PROC LOESS, use the

- `SMOOTH=value-list` option to list smoothing parameters to choose from
- `SELECT=` option to specify the automatic smoothing parameter selection method:
  - AICC, GCV, and DF1 are more computationally efficient.
  - AICC is the default.
  - DF1, DF2, and DF3 are useful for making meaningful comparisons between loess fits and other nonparametric and parametric fits.

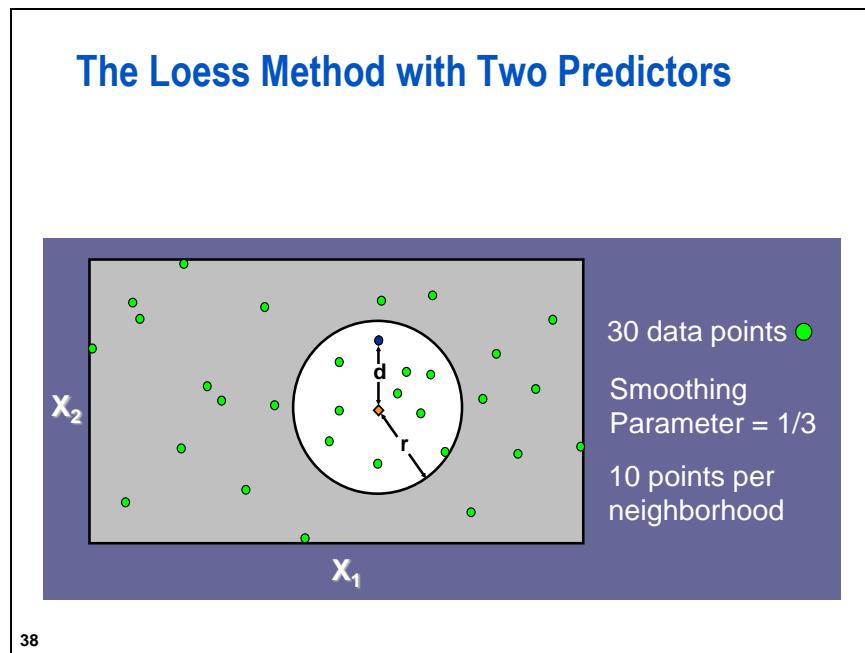
36

## AICC Statistic



37

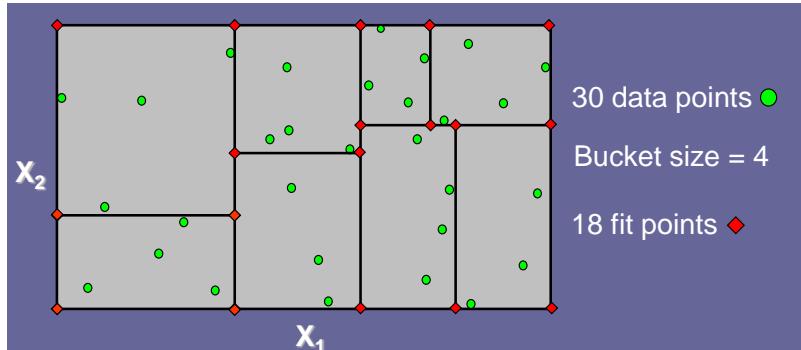
When evaluating the bias-corrected Akaike information criteria, you should choose the smoothing parameter that yields the smallest AICC statistic from among those evaluated smoothing parameters.



The steps in local regression are as follows:

1. Choose a smoothing parameter, which determines the fraction of the data in each local neighborhood.
2. Choose a point at which you want a loess fit. This point does not need to be an original data point.
3. Find the smallest circle about the fit that encloses the desired number of data points. This circle defines the local neighborhood.
4. Determine a local weight for each point in the local neighborhood. One weighting scheme often used is the tri-cube local weight:  $\left(1 - \frac{d^3}{r^3}\right)^3$ .
5. Locally fit a surface using weighted linear or quadratic least squares.
6. Evaluate this regression surface at the fit point to get the loess prediction at this point.
7. Repeat this procedure for every point at which you want a loess fit.

## kd Tree Construction



53

The method used by SAS to determine the fit points for a local regression is a *kd tree*. This method encompasses the following steps:

1. Select the bucket size. The default is (number of points \* smoothing parameter) / 5.
2. Begin with the smallest cell enclosing the data. In the case of a model with one independent variable, the cell is a line segment. In the case of a model with two independent variables, the cell is a rectangle.
3. Select the direction of the longest cell edge as the split coordinate.
4. Divide the cell in two at the median of the split coordinate of the data in the cell.
5. Repeat this process for each child cell until all cells have no more than the number of points specified in the bucket size.
6. The corners of the cells of the final kd tree are used as the fit points.
7. The fit at any other point is found by interpolation from the fitted values at the vertices of the enclosing kd tree cell.

## The Local Regression Model

$$y_i = f(x_i) + \varepsilon_i$$

where  $\varepsilon_i$ s are independent random errors with mean zero.

You can obtain the confidence limits of the predictions if one of the following conditions is met:

- The residuals  $\varepsilon_i$ s are normally distributed with constant variance, or  $\varepsilon_i$  is heteroscedastic, but  $a_i\varepsilon_i$  has constant variance where  $a_i$ s are a priori weights that are specified in the WEIGHT statement in PROC LOESS.
- The error distribution is symmetric.

75

If you denote the  $i^{\text{th}}$  measurement of the response by  $y_i$  and the corresponding measurement of predictors by  $x_i$ , then

$$y_i = f(x_i) + \varepsilon_i$$

where  $f$  is the regression function and the  $\varepsilon_i$  values are independent random errors with mean zero. If the errors are normally distributed with constant variance, then you can obtain confidence intervals for the predictions from PROC LOESS. You can also obtain confidence limits in the case where  $\varepsilon_i$  is heteroscedastic, but  $a_i\varepsilon_i$  has constant variance and  $a_i$  are a priori weights that are specified using the WEIGHT statement of PROC LOESS. You can do inference in the case in which the error distribution is symmetric by using iterative reweighting.

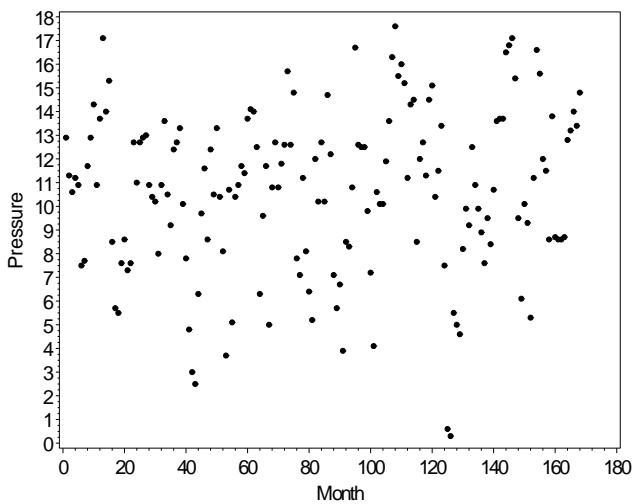
## ENSO Example



76

The ENSO data consist of monthly averaged atmospheric pressure differences between Easter Island and Darwin, Australia, for a period of 168 months. The scientists are studying these differences over time.

## Pressure versus Month



77

Data are stored in the **STAT2.enso** data set. These are the variables in the data set:

- Pressure** monthly averaged atmospheric pressure differences between two locations  
**Month** months since the beginning of the study

From the scatter plot of the data, it is not obvious how to define a parametric relationship between **Pressure** and **Month**. Local regression might be useful.

Another modeling approach is to use procedures provided in SAS/ETS to model the time series data.

## The LOESS Procedure

General form of the LOESS procedure:

```
PROC LOESS;  
  MODEL dependents=regressors / options;  
  RUN;
```

78

The LOESS procedure implements a nonparametric method for estimating regression surfaces. PROC LOESS enables great flexibility because no assumptions about the parametric form of the regression surface are needed. You can use PROC LOESS for situations in which you do not know a suitable parametric form of the regression surface. Furthermore, it is suitable when there are outliers in the data and a robust fitting method is necessary.

Selected PROC LOESS statement:

**MODEL**      names the dependent variables and the independent variables. Variables specified in the MODEL statement must be numeric variables.



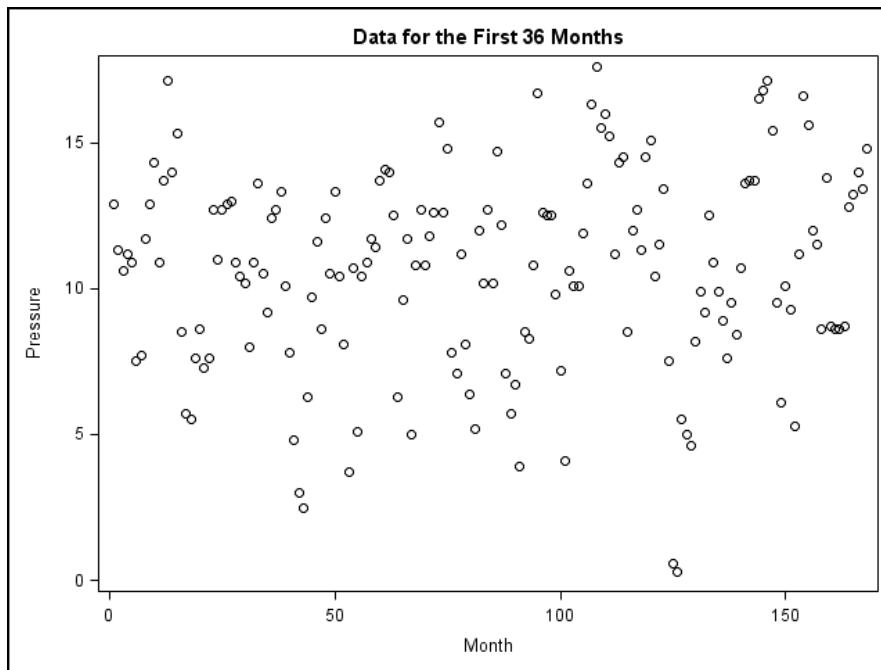
## Fitting a Local Regression Model

Before generating a model for any data, you should conduct an initial data exploration. In part, this can be accomplished by looking at a scatter plot of the data.

The program is saved in **ST20Dd02.sas** and provided with the course data.

```
proc sgplot data=STAT2.enso;
    scatter y=pressure x=month;
    title 'Data for the First 36 Months';
run;                                *ST20Dd02.sas;
```

PROC SGPlot Output



There seem to be some patterns present in the data. Now, use the LOESS procedure to fit a local regression to the data.

```
title;
proc loess data=STAT2.enso plots=all;
    model Pressure = Month;
run;                                *ST20Dd02.sas;
```

In order to fit a local regression, one of the first steps is to determine an appropriate smoothing parameter. You can use the SMOOTH= option in the MODEL statement to specify a list of smoothing parameter values to choose from. You can also use the SELECT= option in the MODEL statement to specify the automatic smoothing parameter selection method to use by PROC LOESS.

If you omit both the SMOOTH= and SELECT= options and you have one dependent variable, PROC LOESS uses the default automatic smoothing parameter selection technique, that is, AICC using the golden section search. *Golden section search* is a bisection method of locating a minimum. Sometimes it can result in a local minimum rather than a global minimum.

The PLOTS=ALL option in the PROC LOESS statement requests that diagnostic and fit plots be displayed for the model.

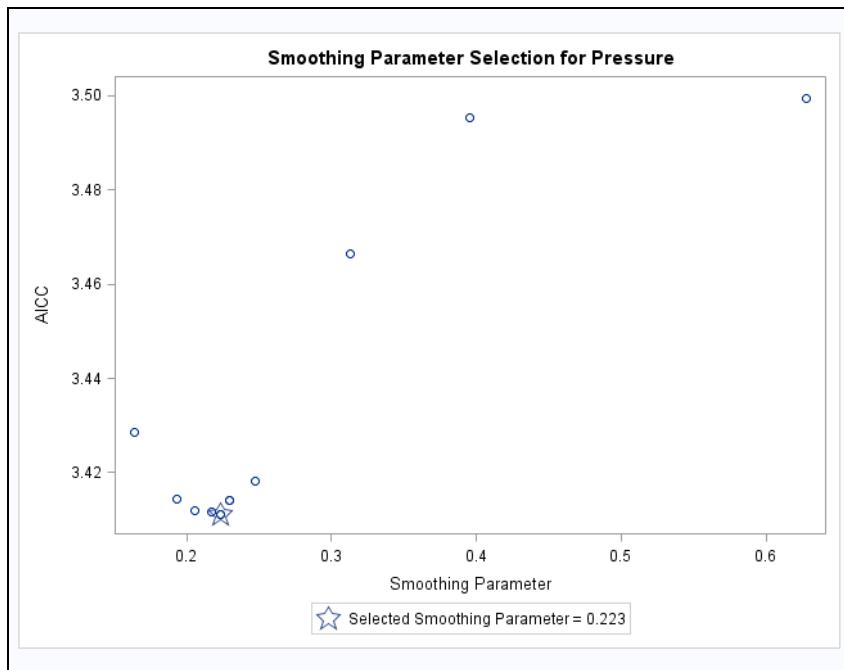
#### Partial PROC LOESS Output

The LOESS Procedure	
Independent Variable Scaling	
Scaling applied: None	
Statistic	Month
Minimum Value	1.00000
Maximum Value	168.00000

This table summarizes the information about the independent variables.

Optimal Smoothing Criterion	
AICC	Smoothing Parameter
3.41105	0.22321

The smoothing parameter is selected to be 0.223. Notice that the default AICC criterion was used to select the smoothing parameter. The technique used was golden section search, which might result in a local minimum.

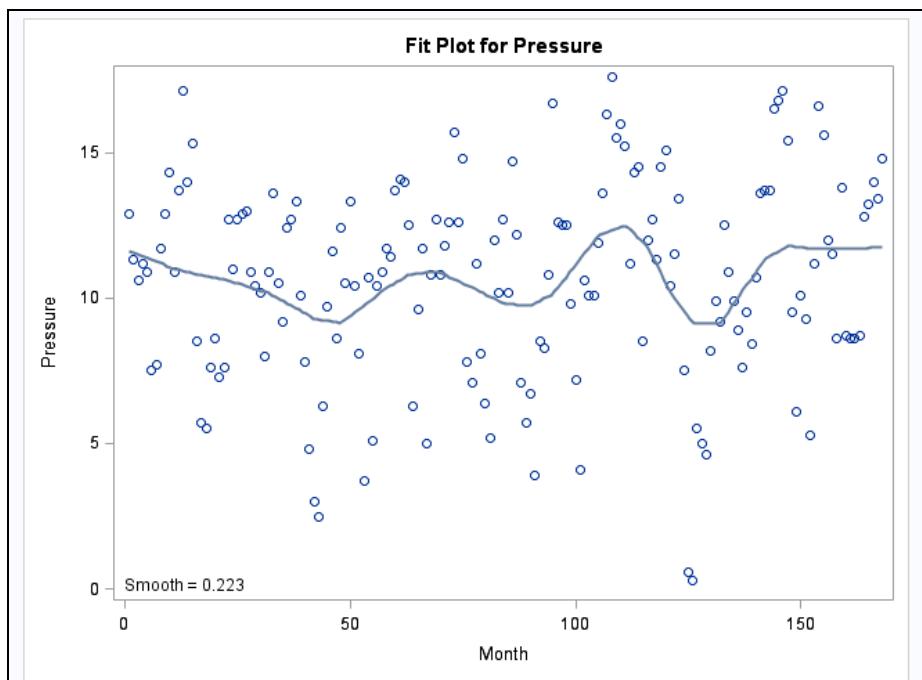


The LOESS Procedure  
 Selected Smoothing Parameter: 0.223  
 Dependent Variable: Pressure

Fit Summary	
Fit Method	kd Tree
Blending	Linear
Number of Observations	168
Number of Fitting Points	33
kd Tree Bucket Size	7
Degree of Local Polynomials	1
Smoothing Parameter	0.22321
Points in Local Neighborhood	37
Residual Sum of Squares	1654.27725
Trace[L]	8.74180
GCV	0.06522
AICC	3.41105

This table provides a summary of model fit.

#### PROC LOESS ODS Graphics Output



The plot of the model fit indicates possible oversmoothing. Recall that if you have one dependent variable, PROC LOESS uses the golden section search technique by default to find a local minimum of the AICC criterion. You can use the SELECT= option with the GLOBAL modifier in parentheses to specify that a global minimum be found.

```
title;
proc loess data=STAT2.enso plots(unpack) =all ;
  model Pressure=Month / r select=aicc(global)
    details(tree fitpoints statout);
run;
*ST20Dd02.sas;
```

Selected MODEL statement options:

**R** specifies that residuals are to be included in the Output Statistics table.

**SELECT=** specifies that automatic smoothing parameter selection be done using the named criterion or DFCriterion. Valid values for the criterion are as follows:

- AICC specifies the AICC criterion, one type of bias-corrected Akaike information criteria.
- AICCI specifies the AICCI criterion, another type of bias-corrected Akaike information criteria.
- GCV specifies the generalized cross validation criterion.
- DF1 specifies a criterion that is based on  $\text{Trace}(L)$ .
- DF2 specifies a criterion that is based on  $\text{Trace}(L^T L)$ .
- DF3 specifies a criterion that is based on  $2\text{Trace}(L) - \text{Trace}(L^T L)$ .

You can specify the following modifiers in parentheses after the specified criterion to alter the behavior of the SELECT= option:

**GLOBAL** specifies that a global minimum be found within the range of examined smoothing parameter values. This modifier has no effect if you also specify the SMOOTH= option in the MODEL statement.

**STEPS** specifies that all models evaluated in the selection process be displayed.

**RANGE(lower,upper)**

specifies that only smoothing parameter values greater than or equal to *lower* and less than or equal to *upper* be examined.

**DETAILS** selects which tables to display, where *tables* is one or more of kdTree (or TREE), PredAtVertices (or FITPOINTS), and OutputStatistics (or STATOUT). A specification of kdTree outputs the kd tree structure. PredAtVertices outputs fitted values and coordinates of the kd tree vertices where the local least squares fitting is done. OutputStatistics outputs the predicted values and other requested statistics at the points in the input data set. The kdTree and PredAtVertices specifications are ignored if the DIRECT option is specified in the MODEL statement. Specifying the DETAILS option with no qualifying list outputs all tables.

Other options in the MODEL statement might include the following:

**SMOOTH=** specifies a list of positive smoothing parameter values. A separate fit is obtained for each specified smoothing value.

**DEGREE=** sets the degree of the local polynomials to use for each local regression. The valid values are 1 for local linear fitting or 2 for local quadratic fitting. The default value is 1.

DIRECT specifies that local least squares fits are to be done at every point in the input data set. When the DIRECT option is not specified, a computationally faster method is used. This faster method performs local fitting at vertices of a kd tree decomposition of the predictor space followed by blending of the local polynomials to obtain a regression surface.

Notice that AICC1 is much more computationally expensive than AICC and GCV, especially when combined with the GLOBAL modifier. AICC can be considered as an approximation of AICC1 and generally performs well in all circumstances.

Partial PROC LOESS Output

Optimal Smoothing Criterion	
AICC	Smoothing Parameter
2.86660	0.05655

When you use the SELECT=AICC(GLOBAL) option, the smoothing parameter was found to be 0.05655, which is much smaller than the previous 0.22321. This indicates that the golden section search used in the previous example resulted in a local minimum rather than the global minimum. This is evident from one of the previous slides showing how the AICC statistic changes as the smoothing parameter values change.

The LOESS Procedure Selected Smoothing Parameter: 0.057 kd Tree				
Node	Left Child	Right Child	Split Direction	Split Value
0	1	2	Month	84.00000
1	3	4	Month	42.00000
2	169	170	Month	126.00000
3	5	6	Month	21.00000
4	87	88	Month	63.00000
5	7	8	Month	11.00000
6	47	48	Month	32.00000
7	9	10	Month	6.00000
8	29	30	Month	16.00000
9	11	12	Month	3.00000
10	21	22	Month	9.00000
11	13	14	Month	2.00000
12	17	18	Month	5.00000
13	15	16	Month	1.00000
14			TERMINAL	

The request for tree details in the MODEL statement displays the kd tree construction that is used to determine fit points for the model.

**Selected Smoothing Parameter: 0.057**  
**Dependent Variable: Pressure**

<b>Predicted Values at kd Tree Vertices</b>		
<b>Vertex Number</b>	<b>Month</b>	<b>Predicted Value</b>
<b>1</b>	1.00000	12.52410
<b>2</b>	2.00000	11.84500
<b>3</b>	3.00000	11.14819
<b>4</b>	4.00000	10.43739
<b>5</b>	5.00000	9.80596
<b>6</b>	6.00000	9.73752
<b>7</b>	7.00000	10.11769
<b>8</b>	8.00000	10.82563
<b>9</b>	9.00000	11.70848
<b>10</b>	10.00000	12.69091
<b>11</b>	11.00000	13.52426
<b>12</b>	12.00000	13.98697
<b>13</b>	13.00000	14.13017

The request for fit-points details in the MODEL statement displays the predicted values at each of the vertices of the kd tree.

<b>Output Statistics</b>				
<b>Obs</b>	<b>Month</b>	<b>Pressure</b>	<b>Predicted Pressure</b>	<b>Residual</b>
<b>1</b>	1.00000	12.90000	12.52410	0.37590
<b>2</b>	2.00000	11.30000	11.84500	-0.54500
<b>3</b>	3.00000	10.60000	11.14819	-0.54819
<b>4</b>	4.00000	11.20000	10.43739	0.76261
<b>5</b>	5.00000	10.90000	9.80596	1.09404
<b>6</b>	6.00000	7.50000	9.73752	-2.23752
<b>7</b>	7.00000	7.70000	10.11769	-2.41769
<b>8</b>	8.00000	11.70000	10.82563	0.87437
<b>9</b>	9.00000	12.90000	11.70848	1.19152
<b>10</b>	10.00000	14.30000	12.69091	1.60909
<b>11</b>	11.00000	10.90000	13.52426	-2.62426
<b>12</b>	12.00000	13.70000	13.98697	-0.28697
<b>13</b>	13.00000	17.10000	14.13017	2.96983
<b>14</b>	14.00000	14.00000	13.59185	0.40815
<b>15</b>	15.00000	15.30000	12.02974	3.27026

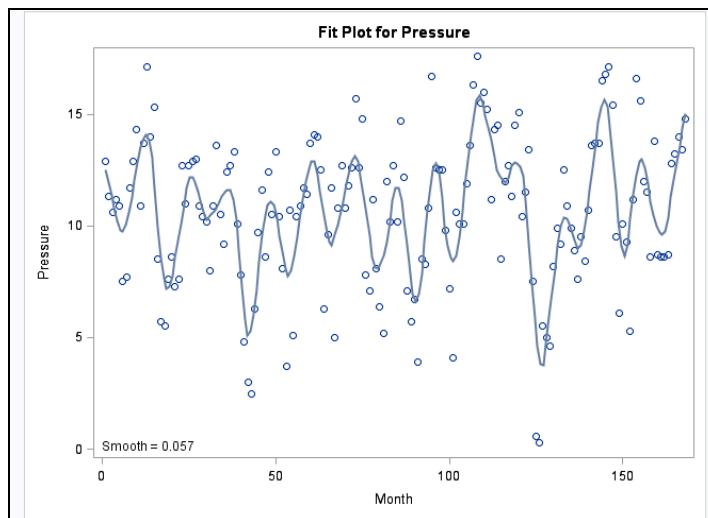
The request for output statistics in the MODEL statement displays the predicted values and residuals for each point in the data set.

**The LOESS Procedure**  
**Selected Smoothing Parameter: 0.057**  
**Dependent Variable: Pressure**

Fit Summary	
<b>Fit Method</b>	kd Tree
<b>Blending</b>	Linear
<b>Number of Observations</b>	168
<b>Number of Fitting Points</b>	168
<b>kd Tree Bucket Size</b>	1
<b>Degree of Local Polynomials</b>	1
<b>Smoothing Parameter</b>	0.05655
<b>Points in Local Neighborhood</b>	9
<b>Residual Sum of Squares</b>	604.00712
<b>Trace[L]</b>	36.89173
<b>GCV</b>	0.03514
<b>AICC</b>	2.86660

A summary of model fit is one of the default outputs by PROC LOESS. Notice that in this example, the fit points are at every observation, which is not necessarily the case for other data sets.

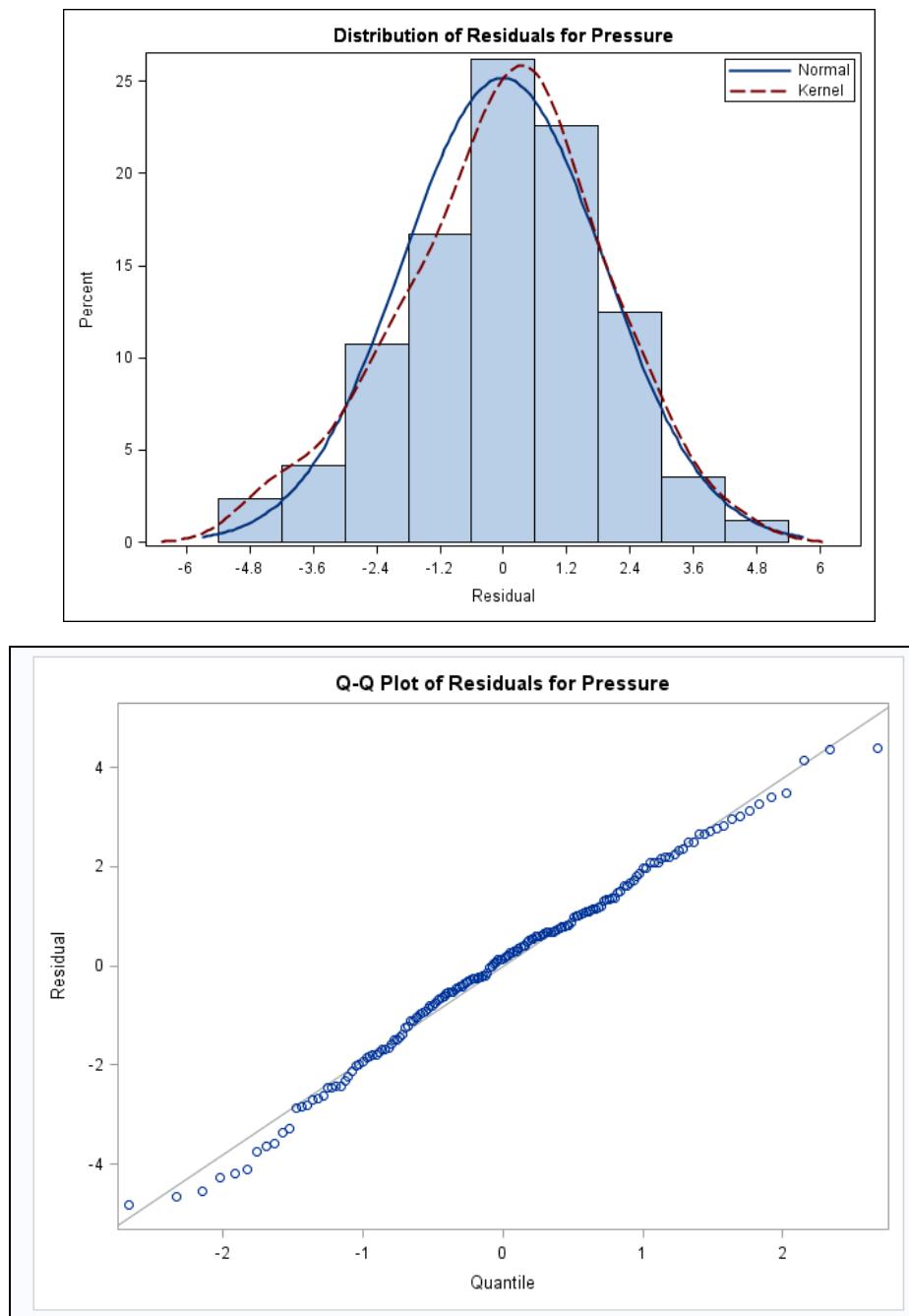
#### PROC LOESS ODS Graphics Output

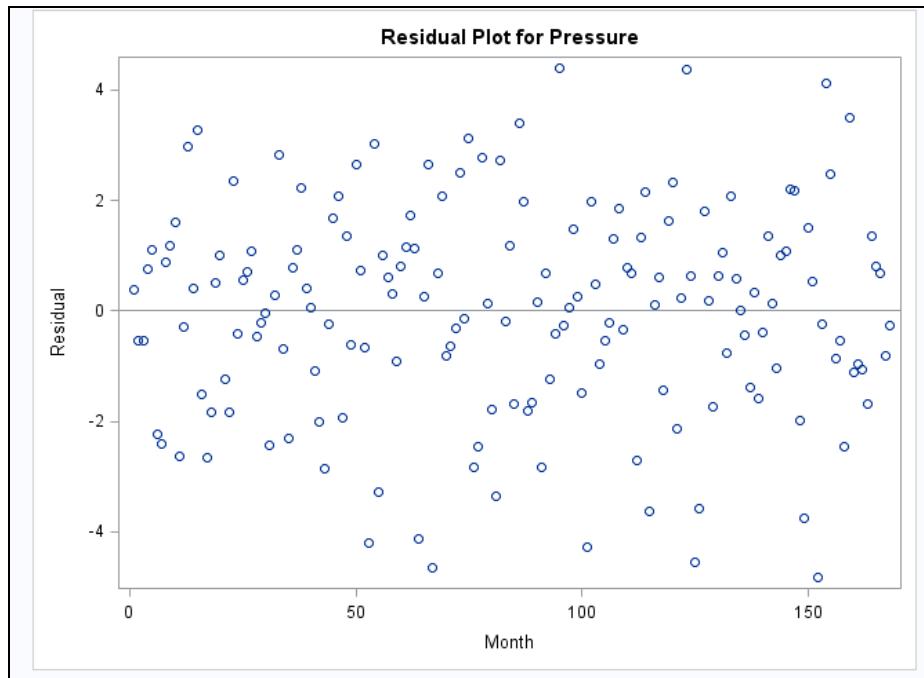


The model appears to fit the data well.

It is often desirable to obtain not only the predictions but also the confidence limits about the predictions. In computing confidence limits, PROC LOESS assumes that the error distribution is normal. If the error distribution is symmetric but not normal, then you can still compute asymptotically valid confidence limits using PROC LOESS by iterative reweighting. This is requested using the ITERATIONS= option in the MODEL statement. Now, you might want to evaluate the distribution of the random errors.

#### PROC LOESS ODS Graphics Output





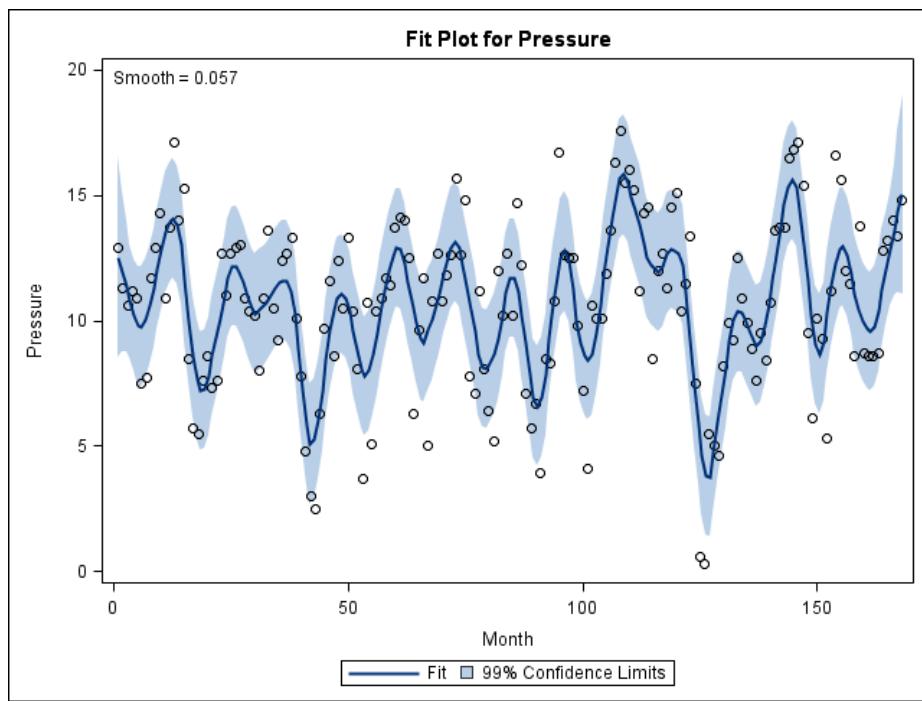
Based on the graphs of the residuals, the residuals appear to be normally distributed with constant variance. Therefore, confidence limits requested by using the CLM option in the MODEL statement are valid.

```
proc loess data=STAT2.enso plots=fitplot;
  model pressure = month / smooth=0.05655 clm alpha=0.01;
run;
*ST20Dd02.sas;
```

Selected MODEL statement options:

- CLM** requests that confidence limits on the mean predicted value be added to the Output Statistics table. By default, 95% limits are computed. The use of this option implicitly selects the MODEL option DFMETHOD=EXACT if the DFMETHOD= option was not explicitly used.
- ALPHA=** sets the significance level used for the construction of confidence intervals for the current MODEL statement.

## PROC LOESS ODS Graphics Output



By using the CLM option in the MODEL statement, you obtain confidence intervals on the fit plot produced by PROC LOESS.

Another feature of the LOESS procedure is the ability to score other data sets. The SCORE statement enables you to produce predicted values for any data set that includes all of the independent variables specified in the MODEL statement. You can also specify the CLM option in the SCORE statement to obtain confidence intervals for the predicted values. The SCORE statement produces predicted values for values of the independent variable that are not included in the original data set, provided that the values are within the range of data values in the original data set. Scoring is done by first finding the cell in the kd tree that contains the specified point and then interpolating the scored value from the predicted values at the vertices of this cell.

For example, suppose you are interested in generating predicted values when **Month** is equal to a multiple of 10. The values of interest can be entered into a SAS data set and the SCORE statement can be used to generate these predicted values.

```

data test;
  do Month=10 to 160 by 10;
  output;
  end;
run;
proc loess data=STAT2.enso;
  model pressure = month / smooth=0.05655;
  score data=test / print;
run;                                *ST20Dd02.sas;

```

Selected LOESS procedure statement:

**SCORE**      The fitted loess model is used to score the data in the specified SAS data set. This data set must contain all the regressor variables specified in the MODEL statement. If no data set is named in the SCORE statement, the data set named in the PROC LOESS statement is scored. You use the PRINT option in the SCORE statement to request that the Score Results table be displayed.

Partial PROC LOESS Output

**The LOESS Procedure**  
**Smoothing Parameter: 0.057**

Score Results		
Obs	Month	Predicted Pressure
1	10.00000	12.69091
2	20.00000	7.58389
3	30.00000	10.24924
4	40.00000	7.74043
5	50.00000	10.64859
6	60.00000	12.89636
7	70.00000	11.61962
8	80.00000	8.19405
9	90.00000	6.53191
10	100.00000	8.69673
11	110.00000	15.21142
12	120.00000	12.76831
13	130.00000	7.55758
14	140.00000	11.07955

## Summary

Local regression fits a curve to data when the relationship between the response and predictor variables cannot be estimated by any obvious parametric equation. Locally, such curves can be approximated by low degree polynomials. This local approximation is found by fitting a regression surface to the data points within a chosen neighborhood.

The local neighborhood is chosen so that it contains a specified percentage of the data points. This fraction of the data is called the smoothing parameter and controls the smoothness of the estimated surface. If the smoothing parameter is too small, the curve is over fit and is too noisy. If the smoothing parameter is too large, the curve might miss some important features of the relationship.

Several smoothing parameter selection methods are available in PROC LOESS. They all select the smoothing parameter value to yield a minimum of an optimization criterion. One of the criteria is the bias-corrected Akaike information criterion (AICC), which is the default method. Others include GCV, AICC1, DF1, DF2, and DF3.

The steps in local regression are

1. Choose a smoothing parameter, which determines the fraction of the data in each local neighborhood.
2. Choose a point at which you want a loess fit. This point does not have to be an original data point. The method used by the LOESS procedure to determine the fit points for a local regression is a kd tree.
3. Find the smallest circle about the fit that encloses the desired number of data points. This circle defines the local neighborhood.
4. Determine a local weight for each point in the local neighborhood. One weighting scheme that is often used is the tri-cube local weight:  $\left(1 - \frac{d^3}{r^3}\right)^3$ .
5. Locally fit a surface using weighted linear or quadratic least squares.
6. Evaluate this regression surface at the fit point to get the loess prediction at this point.
7. Repeat this procedure for every point at which you want a loess fit.

General form of the LOESS procedure:

```
PROC LOESS;
  MODEL dependents=regressors / options;
RUN;
```

## D.3 Modeling Data with Autocorrelation

### Objectives

- Introduce the concept of autocorrelation.
- Demonstrate methods to detect autocorrelation.
- Introduce the Autoregressive Error Model.
- Model autocorrelation using the AUTOREG procedure.

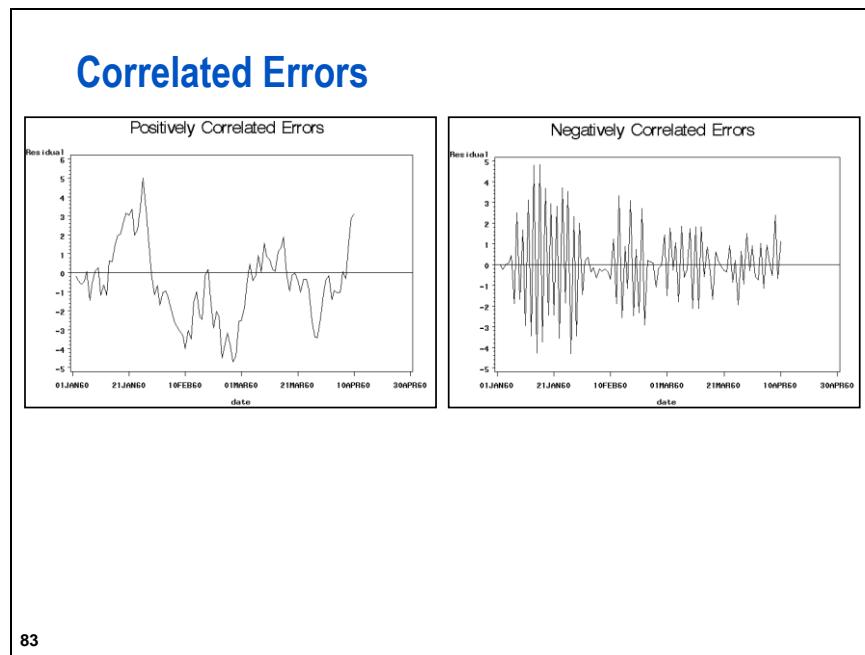
81

### Correlated Error Terms

- Correlated errors can arise from
  - time series data
  - repeated measures on a given subject
  - nested or hierarchical data
  - complex survey design data.
- Correlated errors can affect the standard errors of the parameters, confidence intervals, and tests of significance.
- Tools other than PROC REG are needed to model data with correlated error terms.

82

One of the assumptions in linear regression is independent errors. Error terms are correlated if the values of the errors depend on the other values of errors. They often arise from time series data or repeated measures data. Correlated errors can affect the standard errors of the parameter estimates, and therefore, can affect the confidence intervals and the significance tests for the parameters.



When data are collected over time, correlations between error terms are referred to as *autocorrelation*. Two types of autocorrelation are positive autocorrelation and negative autocorrelation.

For *positive autocorrelation*,

- values tend to be followed by values with the same sign
- estimates of variance ignoring positive correlation are usually too small.

For *negative autocorrelation*,

- values tend to alternate between positive and negative values.

Estimates of variance ignoring positive correlation are usually too large.

## Durbin-Watson $d$ Statistic

- The Durbin-Watson  $d$  statistic can be used to test for autocorrelation in time series data.
- Small  $d$  (near 0) indicates positively correlated residuals.
- Large  $d$  (near 4) indicates negatively correlated residuals.
- A  $d$  statistic that is approximately 2 indicates random residuals.
- A  $d$  statistic can be obtained by the DWPROB option in the MODEL statement in PROC REG.

84

### Details

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \approx 2 \left[ 1 - \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \right] = 2(1 - r)$$

where  $d$  is the Durbin-Watson statistic,  $t$  represents the time order in which the data were collected,  $e_t$  is the residual at time  $t$ , and  $r$  is the first-order autocorrelation.

## First-Order Autocorrelation

The first-order autocorrelation

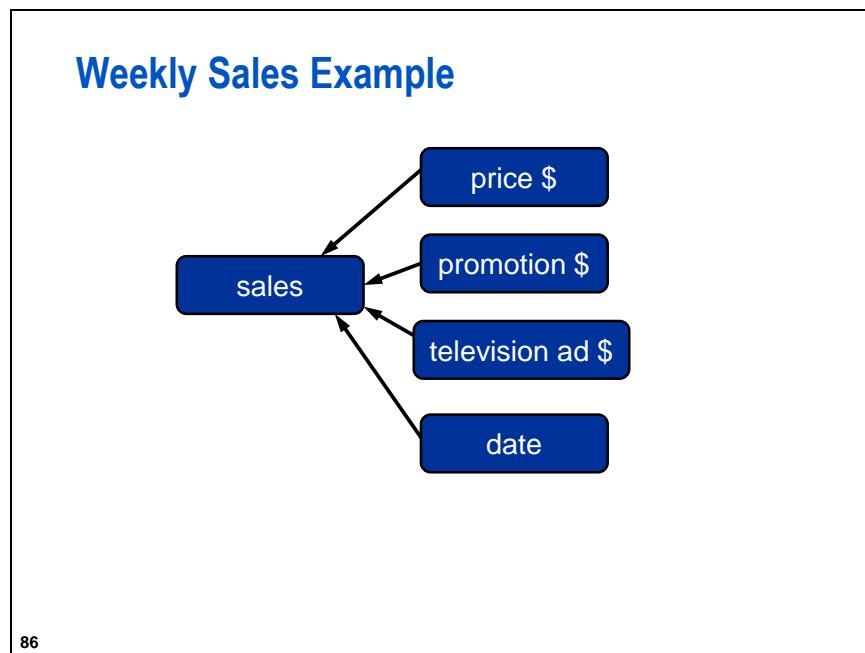
- measures the correlation between the residual values and the lagged residual values
- is approximately related to the Durbin-Watson statistic as  $r = 1 - d/2$
- close to 1 or -1 indicates a high degree of autocorrelation
- close to 0 might or might not indicate independent error terms
- can be obtained by the DWPROB option in the MODEL statement in PROC REG.

85

**Details**

$$r = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \approx 1 - \frac{d}{2}$$

where  $r$  is the first-order autocorrelation,  $t$  represents the time order in which the data were collected,  $e_t$  is the residual at time  $t$ , and  $d$  is the Durbin-Watson statistic.



Weekly sales data were collected over the past three years for a national retail store. The goal of the study is to evaluate whether the total sales are affected by the average unit price and the expenses associated with TV commercials and non-TV promotion programs. Data are stored in the **STAT2.sales** data set.

## The Data

Obs	Date	Sales	Price	Promotion	TVAd
1	05/30/2000	9859	133.499	0.84936	17.49
2	06/06/2000	7645	145.677	0.00000	25.25
3	06/13/2000	7968	133.127	0.91753	2.52
4	06/20/2000	6851	142.618	0.00000	6.81
5	06/27/2000	6568	143.918	0.00000	0.00
6	07/04/2000	6077	142.137	0.12669	0.00
7	07/11/2000	6929	134.819	3.08270	0.95
8	07/18/2000	6431	141.102	0.79101	0.00
9	07/25/2000	6970	138.818	0.90677	0.00
10	08/01/2000	9402	129.557	2.30656	0.00
11	08/08/2000	7993	133.895	1.49199	0.00
12	08/15/2000	8403	135.231	1.02416	0.00
13	08/22/2000	8568	133.015	1.21113	0.00
14	08/29/2000	9221	135.658	0.76901	0.00
15	09/05/2000	10785	130.207	1.56996	23.02

and so on;

87

These are the variables in the **STAT2.sales** data set:

- Date** the day of the week
- Sales** total weekly sales in terms of the number of transactions
- Price** average unit price in dollars for the same type of merchandise
- Promotion** cost in hundreds of dollars associated with promotional programs
- TVAd** cost in thousands of dollars associated with television commercials

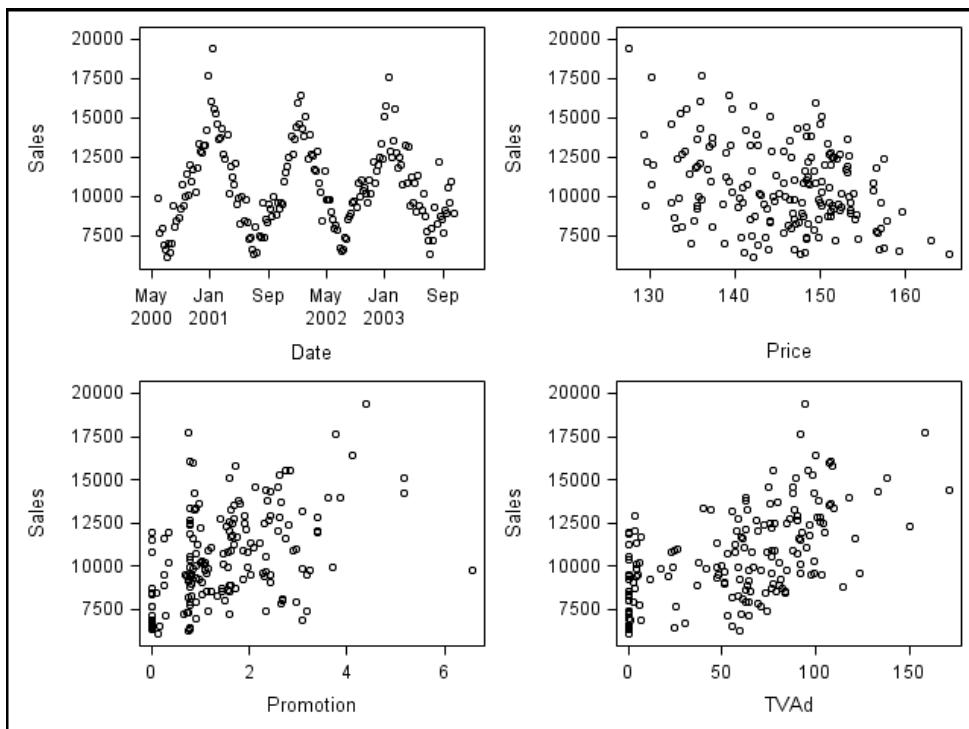


## Detecting Autocorrelation

First, produce scatter plots to examine how **Sales** is related to other variables in the data.

```
proc sgscatter data=STAT2.sales;
  plot sales*(date price promotion tvad);
  format date mmddyy10.;
run;
quit; *ST20Dd03.sas;
```

PROC SGSCATTER Output



Although there does not seem to be a linear relationship between **Sales** and **Date**, the graph indicates a cyclical trend. The data seem to be positively correlated. There seems to be a negative relationship between **Sales** and **Price**, a positive relationship between **Sales** and **Promotion**, and a positive relationship between **Sales** and **TVAd**.

Now, test for autocorrelation using PROC REG.

```
proc reg data=STAT2.sales plots(unpack)=all;
  model sales= price promotion tvad / dwprob;
  plot r.*obs. ;
  title 'Using PROC REG for Autocorrelated Data';
run;
quit; *ST20Dd03.sas;
```

Selected MODEL statement option:

DWPROB calculates a Durbin-Watson statistic  $d$  and a  $p$ -value to test whether the errors have first-order autocorrelation. It is not necessary to specify the DW option if the DWPROB option is specified. (This test is appropriate only for time series data.) The sample autocorrelation of the residuals is also produced.

Partial PROC REG Output

### Using PROC REG for Autocorrelated Data

The REG Procedure  
Model: MODEL1  
Dependent Variable: Sales

Number of Observations Read	177
Number of Observations Used	177

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	708233430	236077810	79.25	<.0001
Error	173	515335219	2978816		
Corrected Total	176	1223568649			

Root MSE	1725.92477	R-Square	0.5788
Dependent Mean	10604	Adj R-Sq	0.5715
Coeff Var	16.27579		

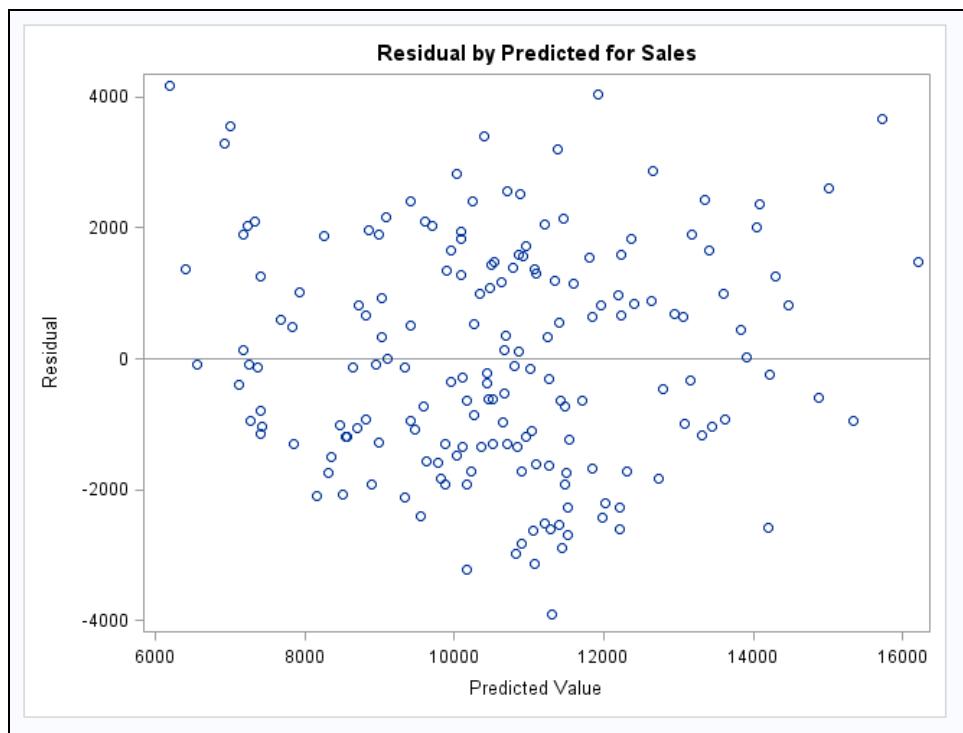
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	29996	2925.41408	10.25	<.0001
Price	1	-153.79136	19.90505	-7.73	<.0001
Promotion	1	277.61828	133.73533	2.08	0.0394
TVAd	1	43.76232	3.56681	12.27	<.0001

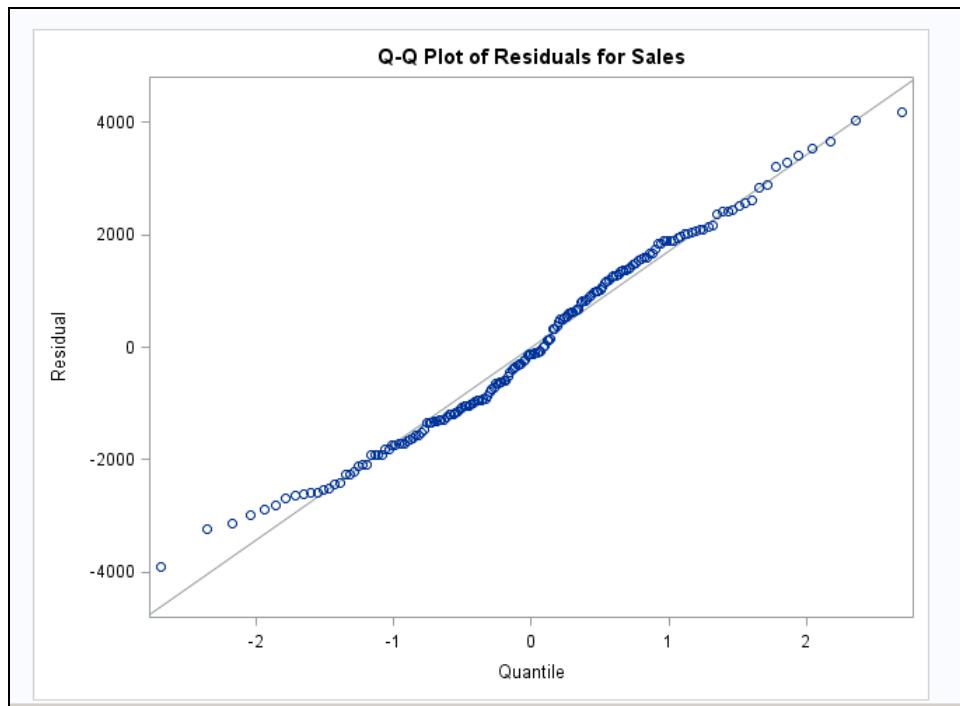
All the parameters are significant at  $\alpha=0.05$ .

The REG Procedure	
Model: MODEL1	
Dependent Variable: Sales	
Durbin-Watson D	0.628
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	177
1st Order Autocorrelation	0.679

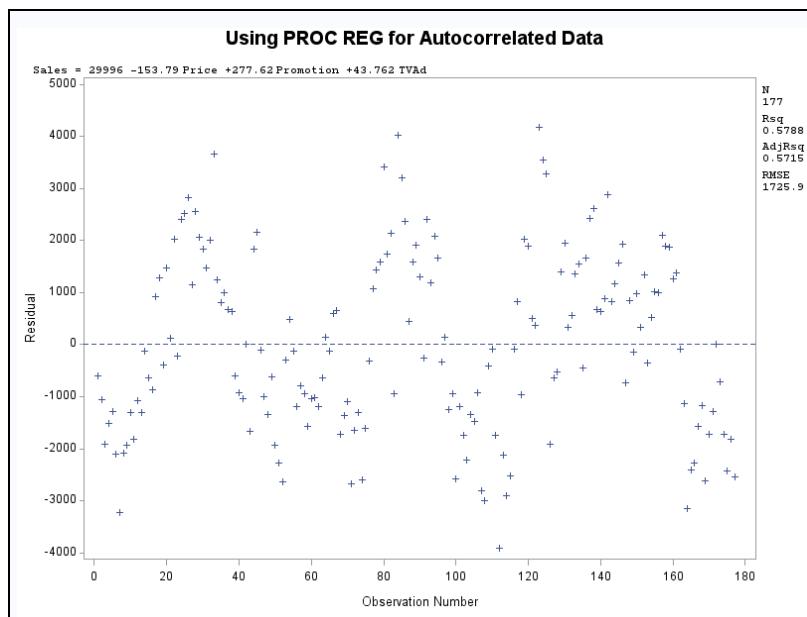
The first-order autocorrelation is 0.679, which measures the correlation between the value for **Sales** at time  $t$  and the previous time point  $t-1$ . The Durbin-Watson statistic  $d$  is 0.628. The small  $p$ -value ( $<.0001$ ) labeled Pr < DW indicates that there is a positive autocorrelation among the residuals.

#### PROC REG ODS Graphics Output





The residual plot and the normal quantile plot do not reveal any apparent departure from normality and the constant variance assumptions, nor do they indicate a poor model fit.



The plot of residuals versus the observation number (reflects the date order) shows a strong positive autocorrelation.

## Autoregressive Error Model in the AUTOREG Procedure

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + v_t$$

$$v_t = -\rho_1 v_{t-1} - \cdots - \rho_m v_{t-m} + \varepsilon_t$$

$$\varepsilon_t \sim i.i.d. N(0, \sigma^2)$$

89

Violation of the independent errors assumption has three important consequences for ordinary regression.

1. Statistical tests of the significance of the parameters and the confidence limits for the predicted values are not correct.
2. The estimates of the regression coefficients are not as efficient as they would be if the autocorrelation were taken into account.
3. The ordinary regression residuals contain information that can be used to improve the prediction of future values because they are not independent.

The AUTOREG procedure solves this problem by augmenting the regression model with an autoregressive model for the random error, thereby accounting for the autocorrelation of the errors. Instead of the usual regression model, the autoregressive error model is used.

The notation  $\varepsilon_t \sim i.i.d. N(0, \sigma^2)$  indicates that each  $\varepsilon_t$  is normally and independently distributed with mean 0 and variance  $\sigma^2$ .

By simultaneously estimating the regression coefficients and the autoregressive error model parameters, the AUTOREG procedure corrects the regression estimates for autocorrelation. Thus, this type of regression analysis is often called *autoregressive error correction* or *serial correlation correction*.

-  More sophisticated approaches to time series regression are available in PROC ARIMA in SAS/ETS software.

## The AUTOREG Procedure

General form of the AUTOREG procedure:

```
PROC AUTOREG options PLOTS (global-plot-options)
   =(plot-request (specific-plot-options));
MODEL dependent=regressors / options;
OUTPUT OUT=SAS-data-set options;
RUN;
```

90

The AUTOREG procedure available in SAS/ETS software estimates and forecasts linear regression models for time series data when the errors are autocorrelated or heteroscedastic. The autoregressive error model is used to correct for autocorrelation. The generalized autoregressive conditional heteroscedasticity (GARCH) model and its variants are used to model and correct for heteroscedasticity.

Selected AUTOREG procedure statements:

**MODEL** specifies the dependent variable and independent regressor variables for the regression model. If no independent variables are specified in the MODEL statement, only the mean is fitted. (This is a way to obtain autocorrelations of a series.)

**OUTPUT** creates an output SAS data set.

More detailed information about using PROC AUTOREG and other procedures in SAS/ETS to analyze time series data can be found in the SAS online documentation.

 For the first-order autoregressive model, you can use the MIXED procedure in SAS/STAT software.



## Modeling Autocorrelation

Now, use the AUTOREG procedure to model the autocorrelation. The appropriate order of autocorrelation can be determined by trial and error, or by fitting stepwise autoregression models. Because the first three orders of autocorrelation are significant by trial and error, you specify a model with third-order autocorrelation.

```
proc autoreg data=STAT2.sales plots(unpack)=all;
  model sales=price promotion tvad / nlag=3 method=ml dwprob;
  title 'Using PROC AUTOREG for Autocorrelated Data';
run; *ST20Dd03.sas;
```

Selected MODEL statement options:

**NLAG=** specifies the order of the autoregressive error process or the subset of autoregressive error lags to be fitted. If the NLAG= option is not specified, PROC AUTOREG does not fit an autoregressive model.

**METHOD** requests the type of estimates to be computed. Valid values are ML (maximum likelihood), ULS (unconditional least squares), YW (Yule-Walker), and ITYW (iterative Yule-Walker). If the GARCH= or LAGDEP option is specified, the default is METHOD=ML. Otherwise, the default is METHOD=YW.

**DWPROB** produces the Durbin-Watson statistic  $d$  and the associated  $p$ -values. The first-order autocorrelation is also produced.

The NLAG=3 option was specified in the MODEL statement because the fourth lag when specifying NLAG=4 was not significant.

Partial PROC AUTOREG Output

### Using PROC AUTOREG for Autocorrelated Data

#### The AUTOREG Procedure

Ordinary Least Squares Estimates			
SSE	515335219	DFE	173
MSE	2978816	Root MSE	1726
SBC	3157.50842	AIC	3144.80382
MAE	1446.41854	AICC	3145.03638
MAPE	14.3576609	HQC	3149.95631
Durbin-Watson	0.6279	Regress R-Square	0.5788
		Total R-Square	0.5788

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	0.6279	<.0001	1.0000

**Note:** Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

Parameter Estimates						
Variable	DF	Estimate	Standard Error	t Value	Approx Pr >  t	
Intercept	1	29996	2925	10.25	<.0001	
Price	1	-153.7914	19.9051	-7.73	<.0001	
Promotion	1	277.6183	133.7353	2.08	0.0394	
TVAd	1	43.7623	3.5668	12.27	<.0001	

The Ordinary Least Squares Estimates table provides identical results as those from PROC REG.

The Total R-square statistic is computed as  $R^2_{\text{tot}} = 1 - [\text{SSE}/\text{SST}]$ , where SST is the sum of squares for the original response variable corrected for the mean and SSE is the final error sum of squares. The Total R square is a measure of how well the next value can be predicted using the structural part of the model and the past values of the residuals.

The Regress R-square statistic is computed as  $R^2_{\text{reg}} = 1 - [\text{TSSE}/\text{TSST}]$ , where TSST is the total sum of squares of the transformed response variable corrected for the transformed intercept, and TSSE is the error sum of squares for this transformed regression problem. The Regress R square is a measure of the fit of the structural part of the model after transforming for the autocorrelation and is the R square for the transformed regression. The regression R square and the total R square should be the same when there is no autocorrelation correction (OLS regression).

Estimates of Autocorrelations																							
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1
0	2911498	1.000000												*****	*****	*****	*****	*****	*****	*****	*****	*****	*****
1	1978097	0.679408												*****	*****	*****	*****	*****	*****	*****	*****	*****	*****
2	1624497	0.557959												*****	*****	*****	*****	*****	*****	*****	*****	*****	*****
3	1304448	0.448033												*****	*****	*****	*****	*****	*****	*****	*****	*****	*****

Preliminary MSE 1516058

The first-order autocorrelation is 0.679. The second-order autocorrelation is 0.558, and the third-order autocorrelation is 0.448.

## Using PROC AUTOREG for Autocorrelated Data

### The AUTOREG Procedure

Maximum Likelihood Estimates			
SSE	102340865	DFE	170
MSE	602005	Root MSE	775.88987
SBC	2889.30478	AIC	2867.07173
MAE	576.764293	AICC	2867.73445
MAPE	5.48893412	HQC	2876.08859
Log Likelihood	-1426.5359	Regress R-Square	0.6483
Durbin-Watson	1.9317	Total R-Square	0.9164
		Observations	177

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	1.9317	0.3242	0.6758

The *p*-value for testing positive autocorrelation is  $\text{Pr} < \text{DW} = 0.3242$ . The *p*-value for testing negative autocorrelation is 0.6758. After you account for the third-order autoregressive error process, the residuals seem to be uncorrelated.

Notice that the total R square is 0.9164, which is much higher than the R square from the OLS regression (0.5788). This total R-square value measures the variations in the data explained by the structural part of the model ( $X\beta$ ) plus the autoregressive error part of the model. The high value indicates a good model fit. The regression R square (0.6483) measures the structural part of the model after transforming for the autocorrelation, and is close to the R-square value from the OLS regression.

Parameter Estimates					
Variable	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	32069	1975	16.23	<.0001
Price	1	-152.1254	11.6443	-13.06	<.0001
Promotion	1	102.6632	51.9434	1.98	0.0497
TVAd	1	3.1971	2.2757	1.40	0.1619
AR1	1	-0.7623	0.0756	-10.08	<.0001
AR2	1	-0.3782	0.0926	-4.09	<.0001
AR3	1	0.1948	0.0770	2.53	0.0124

Autoregressive parameters assumed given					
Variable	DF	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	1	32069	1970	16.28	<.0001
Price	1	-152.1254	11.5848	-13.13	<.0001
Promotion	1	102.6632	51.5181	1.99	0.0479
TVAd	1	3.1971	2.2716	1.41	0.1611

The tables provide parameter estimates and standard errors for the regression coefficients and the autoregressive coefficients. The standard errors for the regression coefficients from the second table are computed, based on the assumption that the estimated autoregressive parameters are the true values. This produces smaller standard errors compared with the first table, which does not make this assumption. Therefore, the variation in the autoregressive parameter estimates is accounted for. These two tables should have similar results.

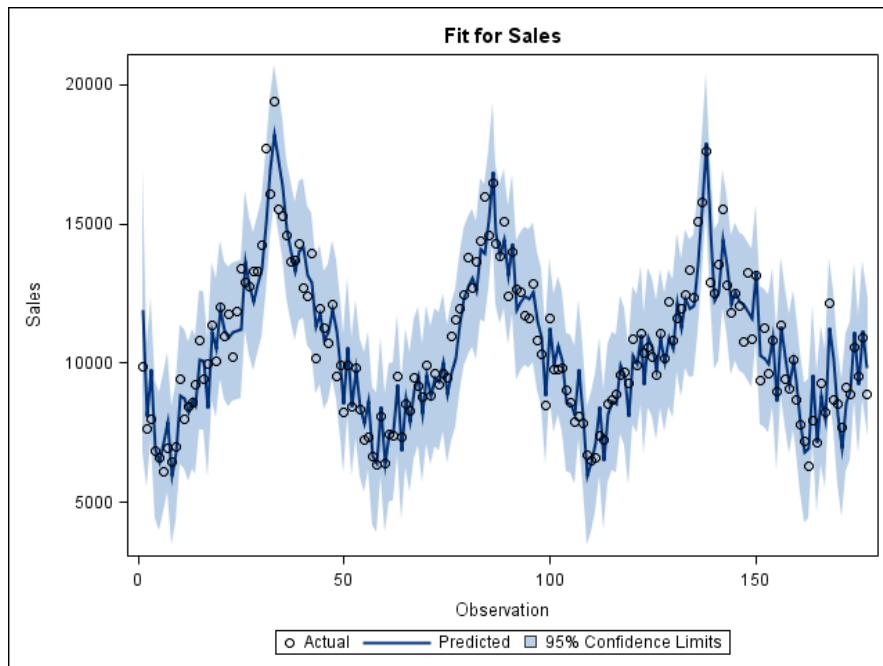
After you account for the autoregressive error process, **TVAd** is **not** significant. Remember that failing to account for positive correlations results in underestimated standard errors and therefore inflates the Type I error rate. With a correct model accounting for the autoregressive error process, the valid inferences are obtained for the predictor variables.

The estimated model is as follows:

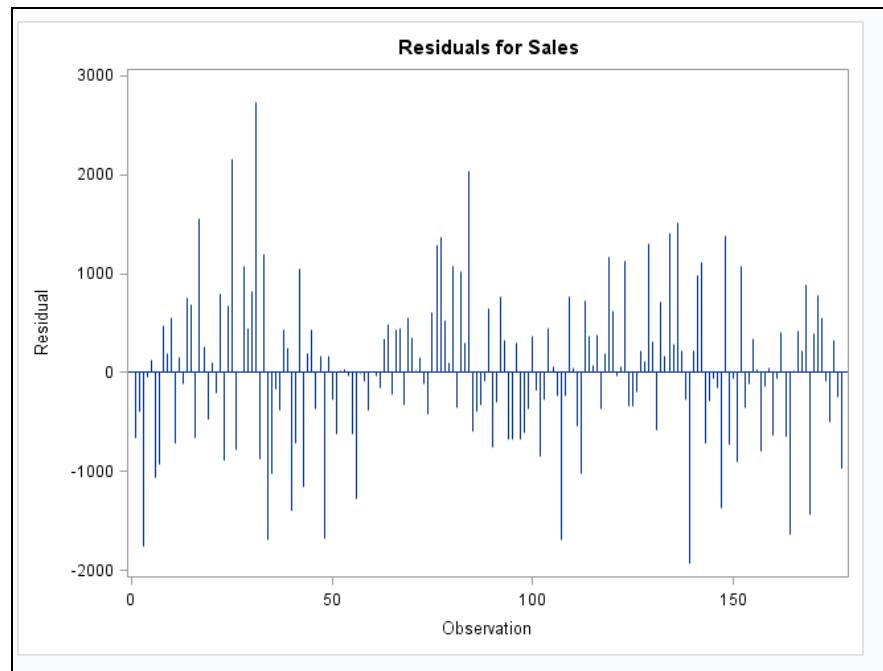
$$\begin{aligned} \text{Sales}_t &= 32069 - 152.1254 * \text{Price} + 102.6632 * \text{Promotion} + 3.1971 * \text{TVAd} + v_t \\ v_t &= 0.7623 * v_{t-1} + 0.3782 * v_{t-2} - 0.1948 * v_{t-3} + e_t \\ \text{estimated var}(e_t) &= 602005 \end{aligned}$$

-  For illustration purpose, the variable **TVAd** is kept in the model. You might want to rerun the model without **TVAd** and write your estimated model accordingly.

## PROC AUTOREG ODS Graphics Output



The model appears to fit the data well and the residuals do not appear to be correlated.



More detailed information about using PROC AUTOREG and other procedures in SAS/ETS, such as PROC ARIMA for more sophisticated approaches to analyze time series data, can be found in the SAS online documentation or Brocklebank and Dickey (2003).

## D.4 Transforming the Dependent Variable as a Remedial Measure

### Objectives

- Discuss transforming response variables as a remedial measure when model assumptions are violated.
- Demonstrate transformations to stabilize variances and/or correct for nonnormality.
- Use PROC TRANSREG to automatically select an appropriate Box-Cox transformation.

93

### Transforming the Dependent Variable

- Transforming the dependent variable is one of the common approaches to deal with nonnormal data, nonconstant variances, or both.
- To determine the appropriate transformation,
  - use theoretical knowledge or previous research
  - use trial and error
  - consider the rate at which the variance increases as the dependent variable increases.
  - use PROC TRANSREG for Box-Cox transformations.

94

To determine which transformation to use, you can do the following:

- use theoretical knowledge or previous research to select an appropriate transformation
- try several transformations and select a transformation that seems to stabilize the variance
- consider the rate at which the variance increases as the dependent variable increases
- use PROC TRANSREG to determine the appropriate Box-Cox transformation

Box-Cox transformations are commonly used as a remedial measure when the data are not normally distributed or suffer from nonconstant variance. This family of transformations affects both the normality and constant variance of the data, so both assumptions must be assessed after a transformation is made.

In addition, when you transform the dependent variable, it is *not* uncommon to observe a change in the relationship with one or more of the independent variables. In practice, after you transform the dependent variable, you should consider starting the process of variable selection again. This was explored in previous exercises.

## Box-Cox Transformation

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \text{Log}(y) & (\lambda = 0) \end{cases}$$

$\lambda$  is the power parameter and can be determined using maximum likelihood criterion.

95

Box and Cox present a method for determining the appropriate power transformation for the dependent variable to stabilize variances or to correct for nonnormality. The Box-Cox (1964) transformation of the positive dependent variable  $y$  is controlled by the parameter  $\lambda$ . Transformations linearly related to square root, inverse, quadratic, cubic, and so on, are all special cases of Box-Cox transformations. The limit as  $\lambda$  approaches 0 is the log transformation.

The BOXCOX option in PROC TRANSREG can be used to perform a Box-Cox transformation of the dependent variable. You can specify a list of power parameters using the LAMBDA= transformation option. By default, LAMBDA=-3 TO 3 BY 0.25. The procedure chooses the optimal power parameter using a maximum likelihood criterion (Draper and Smith 1981, pp. 225-226). You can specify the PARAMETER=c transformation option when you want to shift the values of  $y$ , usually to avoid negatives.

## Details

There are a number of transformations common in statistical modeling. One way to determine which transformation is most appropriate is to evaluate the relationship of the variance to the mean. If the variance is proportional to  $\mu^{2\lambda}$ , then the appropriate transformation is  $y^{1-\lambda}$ , with the exception of when  $\lambda=1$ . When  $\lambda=1$ , the appropriate transformation is  $\ln(y)$ .

### Examples of Box-Cox Transformations

If  $\sigma^2 \propto \mu^{2\lambda}$  (or if  $\sigma \propto \mu^\lambda$ ), then  $y' = y^{1-\lambda}$  ( $\lambda \neq 1$ ).

When  $\lambda = 1$ ,  $y' = \log(y)$ .

$\lambda$	0	$\frac{1}{2}$	$\frac{2}{3}$	1	$\frac{3}{2}$	2	3
$\sigma \propto$	$\mu^0$	$\mu^{\frac{1}{2}}$	$\mu^{\frac{2}{3}}$	$\mu^1$	$\mu^{\frac{3}{2}}$	$\mu^2$	$\mu^3$
$y'$	$y$	$\sqrt{y}$	$y^{\frac{1}{3}}$	$\log(y)$	$\frac{1}{\sqrt{y}}$	$\frac{1}{y}$	$\frac{1}{y^2}$

weaker ← → stronger

96

The transformations shown above are arranged in the order of the amount of curvature that they induce. For example, if the logarithm transformation does not seem to stabilize the variance enough, you might try the more severe transformations to the right of the log transformation. If the log transformation seems to over-correct the heteroscedasticity, you might try the transformations to the left of the log transformation, which are less severe.

- ✍ The parameterization of Box-Cox transformations in PROC TRANSREG differs from that used in the table shown above.

## D.01 Multiple Choice Poll

You can use the following approach to determine the appropriate transformation of the dependent variable:

- a. theoretical knowledge or past studies
- b. how variances and means are related
- c. PROC TRANSREG
- d. trial and error
- e. c and d
- f. all of the above



## Transformations Based on Relationship between the Variance and Mean

---

As an example of how to determine the appropriate transformation for a data set by examining the relationships between the variance and mean (as shown in the previous table), do the following for the `sasuer.cars` data set:

1. Use PROC RANKS to create 20 groups based on **Price** and then calculate the group means and standard deviations.
2. Use a DATA step to calculate the ratio of the standard deviation divided by powers of the mean for each of the 20 groups that you created.
3. From the chart on the previous page, use the relationship for the identity (no) transformation, square root transformation, log transformation and inverse transformation.
4. Use SQL to calculate the quotient of the maximum and minimum ratio for each transformation. The transformation with the smallest quotient seems to stabilize the variance best.

```

proc rank data=STAT2.cars out=partitioned groups=20;
  var price;
  ranks group;
run;

proc means data=partitioned nway noprint;
  class group;
  var price;
  output out=check_transform mean=Mean var=var n=N;
run;

data check_transform;
  set check_transform;
  ID_Transform=var;
  Sqrt_Transform=var/mean;
  Log_Transform=var/(mean**2);
  Inverse_Transform=var/(mean**4);
run;

proc sql;
  select max(ID_transform)/min(ID_transform) as IDQuotient,
         max(sqrt_transform)/min(sqrt_transform) as SqrtQuotient,
         max(log_transform)/min(log_transform) as LogQuotient,
         max(inverse_transform)/min(inverse_transform) as
           InverseQuotient
  from check_transform;
quit;                                *ST2Dd04.sas

```

**SQL Output**

IDQuotient	SqrtQuotient	LogQuotient	Inverse Quotient
104.6327	41.06413	37.00948	300.6351

By examining the quotients, you conclude that the log transformation stabilizes the variance best.

## The TRANSREG Procedure

General form of the TRANSREG procedure:

```
PROC TRANSREG options PLOTS (global-plot-options) =  

  (plot-request (options) ... plot-request (options));  

MODEL transform (dependents / t-options)  

  = transform (independents / t-options / a-options);  

RUN;
```

100

PROC TRANSREG (transformation regression) fits linear models (potentially with spline and other nonlinear transformations). It can be used to code experimental designs before their use in other analyses.

Selected PROC TRANSREG statement option:

PLOTS        controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request.

Selected TRANSREG procedure statements:

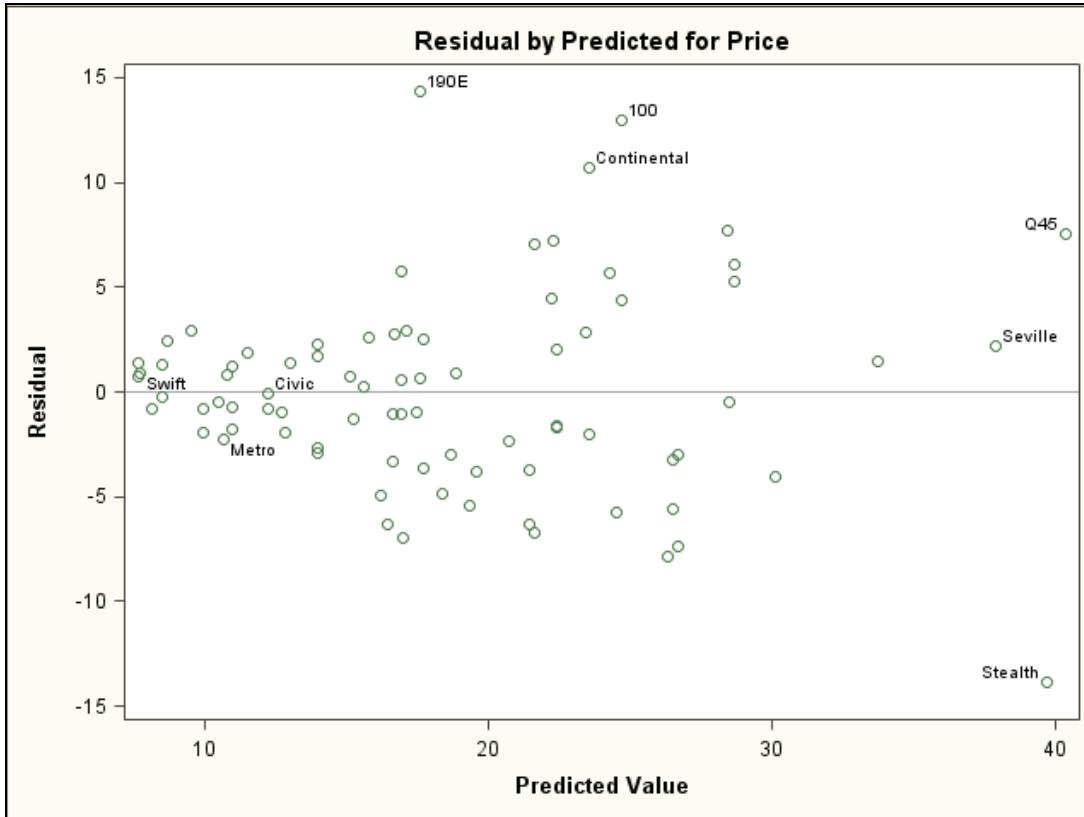
MODEL        specifies the dependent and independent variables and specifies the transformation to apply to each variable. Only one MODEL statement can appear in the TRANSREG procedure. The *t*-options are transformation options, and the *a*-options are the algorithm options. The *t*-options provide details for the transformation; these depend on the *transform* chosen. The *t*-options are listed after a slash in the parentheses that enclose the variable list (either *dependents* or *independents*). The *a*-options control the algorithm used, details of iteration, details of how the intercept and dummy variables are generated, and displayed output details. The *a*-options are listed after the entire model specification (the *dependents*, *independents*, transformations, and *t*-options) and after a slash.

Many *t*-options and *a*-options are available in PROC TRANSREG. The BOXCOX transformation can be applied only to the dependent variable. Refer to the SAS 9 online documentation for more detailed information.



## Box-Cox Transformation to Stabilize Nonconstant Variances

Recall the graph of the residuals versus the predicted values from the model for the **STAT2.cars2** data.



The variability of the residuals seems to increase as the predicted value increases. The Spearman rank correlation coefficient that you computed earlier indicates a significant correlation between the absolute value of the residuals and the predicted values (0.603). Use PROC TRANSREG to determine the appropriate Box-Cox transformation to stabilize the variance.

```

ods html close;
ods rtf file="BoxPlots.rtf" style=journal;
ods rtf select BoxCoxFPlot BoxCoxLogLikePlot RMSEPlot;

proc transreg data=STAT2.cars2 ss2 test cl nomiss
  plots=boxcox (rmse unpack);
  model boxcox(price / convenient)=identity(hwympg hwympg2
    horsepower);
run;
quit;

ods rtf close;                                *ST20Dd05.sas;

```

Selected ODS RTF statement options:

- FILE= specifies the RTF file or SAS catalog to which to write. This file remains open until you do one of the following actions: (1) close the RTF destination with ODS RTF CLOSE or ODS\_ALL\_CLOSE; (2) specify a different file to which to write.
- PATH= specifies the location of an external file or a SAS catalog for all markup files.
- STYLE= provides formatting information for specific visual aspects of your SAS output. The appearance of tables and graphs is coordinated within a particular style. For tables, this information typically includes a list of font definitions and a list of colors. Each font definition specifies a family, size, weight, and style. Colors are associated with common areas of output, including titles, footnotes, BY groups, table headers, and table cells. For graphs, styles also control the appearance of graph elements including lines, markers, fonts, and colors. ODS styles also include elements specific to statistical graphics, such as the style of fitted lines, confidence bands, and prediction limits. For more information about styles, see *SAS® Output Delivery System: User’s Guide*. Each output destination has a default style, which can be changed to any of the following styles:
- HTMLBLUE – all-color style whose dominant colors are shades of blue with sans-serif fonts.
  - DEFAULT – color style whose dominant colors are gray, blue, and white, with bold sans-serif fonts.
  - STATISTICAL – color style whose dominant colors are blue, creamy gray, and white, with sans-serif fonts.
  - LISTING (default in LISTING) – color style, similar to DEFAULT, but with a white background.
  - JOURNAL – black-and-white style with filled areas, and with sans-serif fonts.
  - JOURNAL2 – black-and-white style, similar to JOURNAL, but with empty areas.
  - RTF (default in RTF) – color style whose dominant colors are blue, white, and black, with Times Roman fonts.
  - ANALYSIS – color style, similar to STATISTICAL, whose dominant color is tan.

## Selected PROC TRANSREG statement options:

- SS2 produces a regression table based on Type II sums of squares. Tests of the contribution of each transformation to the overall model are displayed and output to the OUTTEST= data set when you specify the OUTTEST= option. When you specify the SS2 option, the TEST option is implied.
- TEST generates an ANOVA table. PROC TRANSREG tests the null hypothesis that the vector of scoring coefficients for all of the transformations is zero.
- CL requests confidence limits on the parameter estimates in the displayed output.
- NOMISS excludes all observations with missing values from the analysis, but does not exclude them from the OUT= data set. If you omit the NOMISS *a-option*, PROC TRANSREG simultaneously computes the optimal transformations of the nonmissing values and estimates the missing values that minimize squared error. The NOMISS option is recommended for the Box-Cox transformation.



These options can also be specified in the MODEL statement rather than in the PROC TRANSREG statement.

## PLOTS=BOXCOX

requests a display of the results of the Box-Cox transformation. These results are displayed by default when there is a Box-Cox transformation. The BOXCOX plot request has the following options:

RMSE plots the root mean square error as a function of lambda.

T plots *t* statistics rather than *F* statistics.

UNPACK plots the *t* or *F* and log-likelihood plots in separate panels.

## Selected MODEL statement options:

- BOXCOX finds a Box-Cox transformation of the specified variables. The BOXCOX transformation can be used only with dependent variables that are numeric and typically continuous.
- IDENTITY specifies variables that are not changed by the iterations. It is used for variables when no transformation and no missing data estimation are desired.

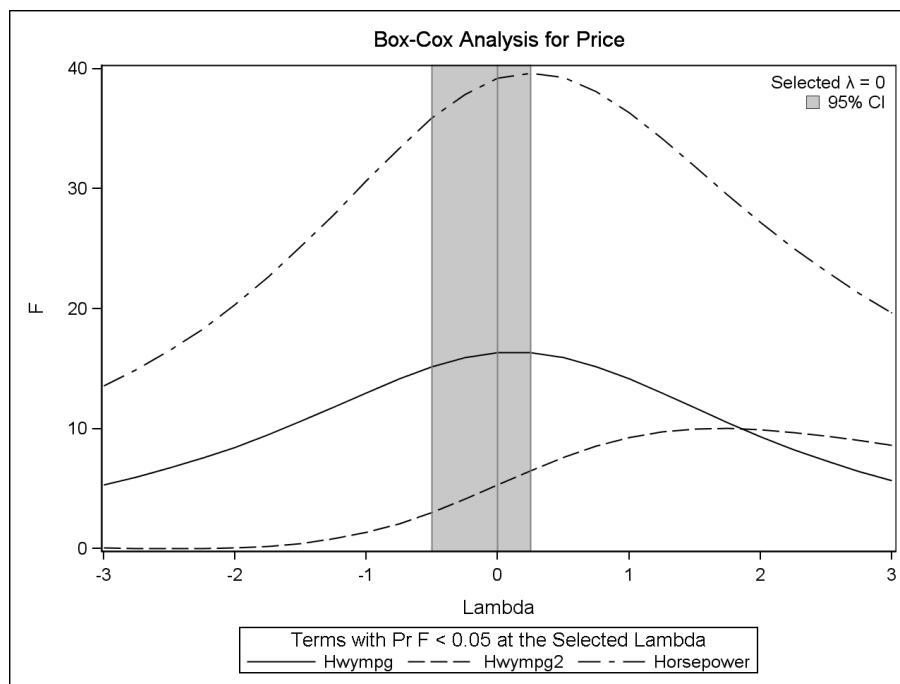
Selected BOXCOX *t*-options:

- LAMBDA= specifies a list of Box-Cox power parameters. The default is LAMBDA = -3 TO 3 BY 0.25. PROC TRANSREG tries each power parameter in the list and selects the best one.

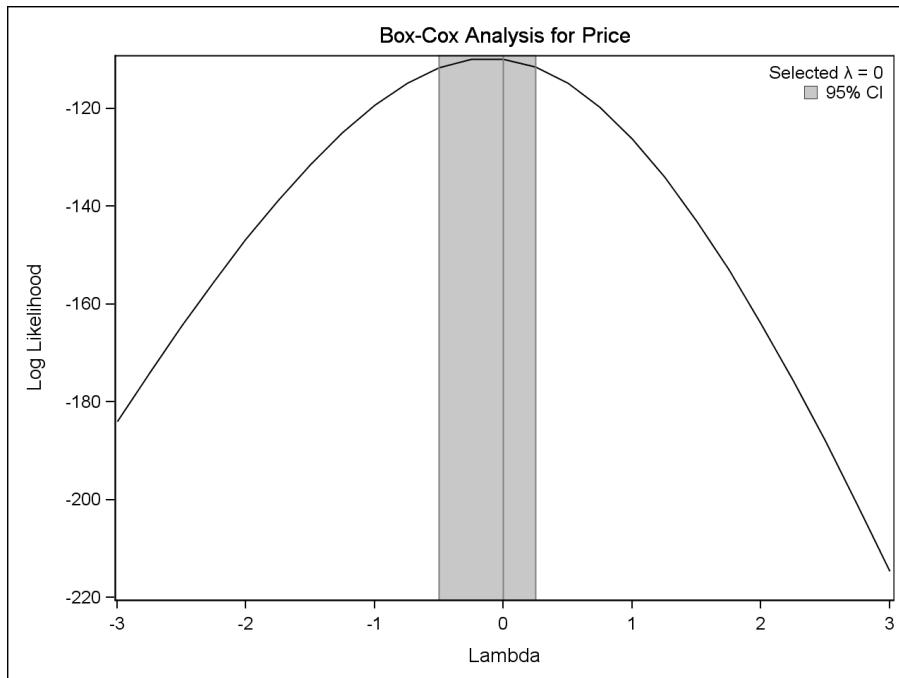
**CONVENIENT** specifies that a power parameter from the convenient lambda list is to be used for the final transformation instead of the LAMBDA= value, if a convenient lambda value is in the confidence interval. When the confidence interval for the power parameter includes one of the values in this list, PROC TRANSREG reports it and can use the convenient power parameter instead of the more optimal power parameter. The default convenient lambda list is 1.0, 0.0, 0.5, -1.0, -0.5, 2.0, -2.0, 3.0, -and 3.0. By default, a linear transformation is preferred over log, square root, inverse, inverse square root, quadratic, inverse quadratic, cubic, and inverse cubic. You can specify a custom convenient lambda list with the CLL= option.

If you specify the CONVENIENT *t*-option, then PROC TRANSREG uses the first convenient power parameter in the list that is in the confidence interval. For example, if the optimal power parameter is 0.25 and 0.0 is in the confidence interval but not 1.0, then the convenient power parameter is 0.0.

**CLL=** specifies a user-defined Box-Cox convenient lambda list. The default is CLL=1.0 0.0 0.5 -1.0 -0.5 2.0 -2.0 3.0 -3.0.



The graph shown above plots the *F* statistic against lambda. The *F* statistic for **Hwmpg** is the highest for lambda=0. The *F* statistic for **Horsepower** peaks at a lambda slightly above 0, but the statistic for **Hwmpg2** peaks at a higher value.



Maximizing the log-likelihood function is the basis for determining the best lambda for a Box-Cox transformation. This plot of the log-likelihood function versus lambda shows that the function is maximized at  $\lambda=0$ . This indicates that a log transformation of the dependent variable is the appropriate transformation.



The plot of the root mean square (RMSE) versus lambda indicates that RMSE is minimized at lambda=0.

To obtain transformation information for Box-Cox in tabular rather than graphical output, run the program with the PLOTS=NONE option in the PROC TRANSREG statement.

```

ods listing;
proc transreg data=STAT2.cars2 ss2 test cl nomiss plots=none;
  model boxcox(price / convenient)=identity(hwympg hwympg2
                                              horsepower);
run;
quit;                                         *ST20Dd05.sas;

```

The TRANSREG Procedure  
Box-Cox Transformation Information for Price

Lambda	R-Square	Log Like
-3.00	0.55	-183.964
-2.75	0.57	-174.052
-2.50	0.59	-164.519
-2.25	0.61	-155.417
-2.00	0.64	-146.811
-1.75	0.66	-138.780
-1.50	0.68	-131.429
-1.25	0.70	-124.882
-1.00	0.71	-119.289
-0.75	0.73	-114.822
-0.50	0.74	-111.659 *
-0.25	0.74	-109.973 *
0.00 +	0.75	-109.898 <
0.25	0.75	-111.514 *
0.50	0.74	-114.825
0.75	0.73	-119.763
1.00	0.72	-126.207
1.25	0.70	-134.002
1.50	0.68	-142.984
1.75	0.66	-153.002
2.00	0.64	-163.922
2.25	0.62	-175.631
2.50	0.60	-188.037
2.75	0.57	-201.065
3.00	0.55	-214.654

< - Best Lambda  
\* - 95% Confidence Interval  
+ - Convenient Lambda

For this model, the best and also the most convenient value for lambda is 0, which corresponds to a logarithm transformation.

The TRANSREG Procedure					
Dependent Variable BoxCox(Price)					
Number of Observations Read					81
Number of Observations Used					81
The TRANSREG Procedure Hypothesis Tests for BoxCox(Price)					
Univariate ANOVA Table Based on the Usual Degrees of Freedom					
Source	DF	Sum of Squares	Mean Square	F Value	Liberal p
Model	3	12.08755	4.029184	75.78	>= <.0001
Error	77	4.09424	0.053172		
Corrected Total	80	16.18179			
The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.					
Root MSE		0.23059	R-Square	0.7470	
Dependent Mean		2.82388	Adj R-Sq	0.7371	
Coeff Var		8.16573	Lambda	0.0000	

The ANOVA table is consistent with the output from the previous PROC REG model that you fit with the transformed variable **LogPrice** as the dependent variable. The note “The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal” points out that there were data-driven dependent variable transformations. A model was fit to the results without any adjustments for the data-driven nature of the transformations. The chances of falsely rejecting the null hypothesis  $H_0$  are increased.

The TRANSREG Procedure							
Univariate Regression Table Based on the Usual Degrees of Freedom							
Type II							
Variable	DF	Coefficient	95% Confidence Limits	Sum of Squares	Mean Square	F Value	Liberal p
Intercept	1	2.1022267	1.8924313	2.3120221	21.1691	21.1691	398.13 >= <.0001
Identity(Hwympg)	1	-0.0419279	-0.0625938	-0.0212621	0.8678	0.8678	16.32 >= 0.0001
Identity(Hwympg2)	1	0.0015985	0.0002162	0.0029808	0.2820	0.2820	5.30 >= 0.0240
Identity(Horsepower)	1	0.0049065	0.0033461	0.0064670	2.0845	2.0845	39.20 >= <.0001
The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.							

The regression coefficient estimates are consistent with the results from the previous PROC REG run with **LogPrice** as the dependent variable. After you transformed the data, check to see whether the variances were stabilized using the Spearman correlation coefficient.

- ✍ Different approaches of determining the appropriate transformations might not always produce consistent results. The Box-Cox transformation might be a good tool to use as an initial step.
- ✍ After the appropriate transformation is determined, continue with the modeling cycle by finding a new candidate model. Then, proceed through the cycle until all diagnostics are completed and the model is finalized.

## D.02 Quiz

Is the statement below true or false? Explain, your answer.

Although there are several methods to determine the appropriate transformation to stabilize variances and correct for nonnormality, the Box-Cox transformation in PROC TRANSREG provides an automated way of doing so.

102

The answer is True. PROC TRANSREG does automate the selection of the appropriate transformation to stabilize variances and correct for nonnormality.

Have the students change their status to indicate True or False. Have someone come over the phone to explain the answer.

## Interpretation of the Transformed Model

- A reverse transformation applied to the transformed model produces an unbiased estimate of the **median** of the original data rather than the **mean**.
- Sometimes it is desirable, or even necessary, to have an estimate of the mean.
- In order to estimate the mean of the original data, a back transformation to the original scale is necessary and an adjustment is required to minimize the bias.

104

Presuming the transformed model meets the assumptions of regression analysis, when transforming back to the original scale of the data, the resulting model is an unbiased estimate of the median rather than the mean. In some cases, an estimate of the median is appropriate. In other cases, an estimate of the mean is necessary.

In order to obtain an estimate of the mean of the original data, an adjustment factor is necessary. This adjustment factor depends on the transformation that was applied to the data.

## Low-Biased Estimates of the Mean: Logarithmic Transformation

Transformed Model:  $\ln(Y) = \beta_0 + \beta_1 X + \varepsilon$

Detransformed Model with Adjustment Factor:

$$E(Y) \approx e^{\hat{\beta}_0 + \hat{\beta}_1 X + \frac{\hat{\sigma}^2}{2}}$$

105

### Details

Transforming data back to the original scale requires an adjustment factor to obtain a low-biased estimate of the mean. Shown below are the back transformations with the correct adjustment factors for data transformed with the log, square-root, inverse square root, and inverse transformations. More general results for negative fractional power ( $-1/N$ ) transformations can be found in Miller (1984).

Transformation	Transformed Model	Back Transformation with Adjustment Factor
Logarithmic	$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon$	$\hat{E}(Y) = e^{\hat{\beta}_0 + \hat{\beta}_1 X + \frac{\hat{\sigma}^2}{2}}$
Square Root	$\sqrt{Y} = \beta_0 + \beta_1 X + \varepsilon$	$\hat{E}(Y) = (\hat{\beta}_0 + \hat{\beta}_1 X)^2 + \hat{\sigma}^2$
Inverse Transformation	$\frac{1}{Y} = \beta_0 + \beta_1 X + \varepsilon$	$\hat{E}(Y) = \frac{1}{\hat{\beta}_0 + \hat{\beta}_1 X} \left( 1 + \frac{\hat{\sigma}^2}{(\hat{\beta}_0 + \hat{\beta}_1 X)^2} \right)$
Inverse Square Root	$\frac{1}{\sqrt{Y}} = \beta_0 + \beta_1 X + \varepsilon$	$\hat{E}(Y) = \frac{1}{(\hat{\beta}_0 + \hat{\beta}_1 X)^2 + \hat{\sigma}^2} \times \left\{ 1 + \frac{2\hat{\sigma}^4 + 4(\hat{\beta}_0 + \hat{\beta}_1 X)^2 \hat{\sigma}^2}{[(\hat{\beta}_0 + \hat{\beta}_1 X)^2 + \hat{\sigma}^2]^2} \right\}$



## Back-Transformation of the Model

In order to obtain estimates of the mean for the original data, you must transform the model back to the original scale, using the low-bias adjustment factor.

First, rerun the regression and create an output data set with the residuals, the mean squared error, and the predicted values. Check for homogeneity of variances with the Spearman correlation coefficient on the log-transformed data. Then, use a DATA step to compute the low-bias adjusted mean values on the original scale, and plot the residuals versus the predicted values on the original scale.

```

data STAT2.logcars2;
  set STAT2.cars2;
  LogPrice=log(price);
run;

proc reg data=STAT2.logcars2;
  model logprice = hwympg hwympg2 horsepower;
  ods output ANOVA=ANOVATable;
  output out=out p=pred r=resid;
run;
quit;

data out;
  set out;
  abserror=abs(resid);
run;

proc corr data=out spearman nosimple;
  var abserror pred;
run;                                *ST20Dd06.sas;

```

### The CORR Procedure

Spearman Correlation Coefficients, N = 81  
Prob > |r| under H0: Rho=0

	abserror	pred
abserror	1.00000	0.19492 0.0812
pred	0.19492	1.00000
	Predicted Value of LogPrice	0.0812

The Spearman correlation coefficient between the absolute value of the residuals and the predicted values on the log-transformed data is 0.19492. The *p*-value of 0.0812 indicates that not enough evidence exists to reject the hypothesis of homogeneity of variances on the log-transformed scale of the data.

```

data _null_;
  set ANOVATable;
  if source='Error' then call symput('var', MS);
run;

```

```

data out;
  set out;
  Estimate = exp(pred + &var/2);
  Difference = price - estimate;
run;

proc print data=out;
  var manufacturer model hwympg horsepower price estimate difference;
run;

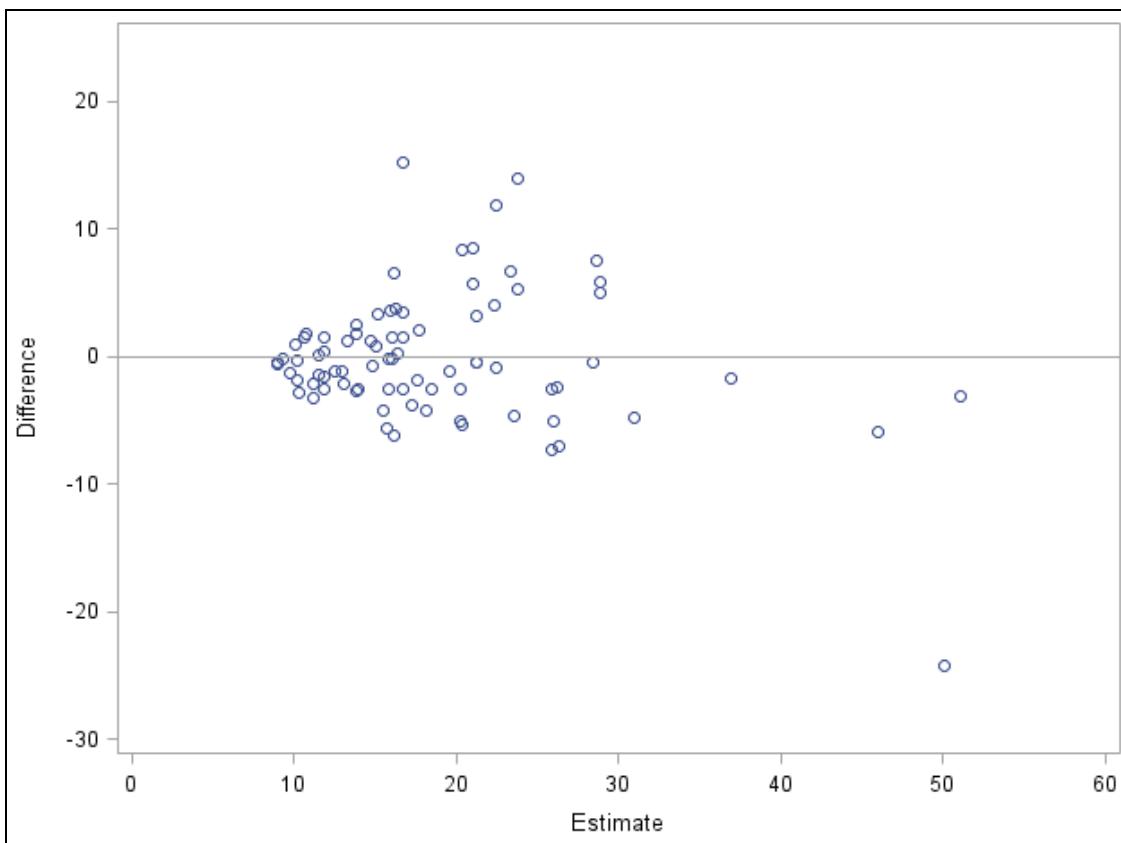
proc sgplot data=out;
  scatter x=estimate y=difference;
  xaxis min=0 max=60;
  yaxis min=-30 max=25;
  refline 0;
run;
quit;
*ST20Dd06.sas;

```

Partial PROC PRINT Output

Obs	Manufacturer	Model	Hwympg	Horsepower	Price	Estimate	Difference
1	Acura	Integra	0.9630	140	15.9	16.0680	-0.1680
2	Acura	Legend	-5.0370	200	33.9	28.8430	5.0570
3	Audi	100	-4.0370	172	37.7	23.7611	13.9389
4	Audi	90	-4.0370	172	29.1	23.7611	5.3389
5	BMW	535i	-0.0370	208	30.0	23.3576	6.6424
6	Buick	Century	0.9630	110	15.7	13.8687	1.8313
7	Buick	LeSabre	-2.0370	170	20.8	21.2204	-0.4204
8	Buick	Riviera	-3.0370	170	26.3	22.3093	3.9907
9	Buick	Roadmaster	-5.0370	180	23.7	26.1470	-2.4470
10	Cadillac	DeVille	-5.0370	200	34.7	28.8430	5.8570
11	Cadillac	Seville	-5.0370	295	40.1	45.9699	-5.8699
12	Chevrolet	Camaro	-2.0370	160	15.1	20.2043	-5.1043
...							
24	Dodge	Stealth	-6.0370	300	25.8	50.0063	-24.2063
25	Eagle	Summit	2.9630	92	12.2	11.8226	0.3774
26	Eagle	Vision	-2.0370	214	19.3	26.3337	-7.0337
27	Ford	Crown Victoria	-4.0370	190	20.9	25.9551	-5.0551
28	Ford	Escort	-0.0370	127	10.1	15.6974	-5.5974
29	Ford	Festiva	2.9630	63	7.4	10.2546	-2.8546
30	Ford	Mustang	-1.0370	105	15.9	14.7199	1.1801
...							
51	Mercedes-Benz	190E	-1.0370	130	31.9	16.6408	15.2592
...							
79	Volkswagen	Passat	-0.0370	134	20.0	16.2459	3.7541
80	Volvo	240	-2.0370	114	22.7	16.1222	6.5778
81	Volvo	850	-2.0370	168	26.7	21.0132	5.6868

## PROC SGLOT Output



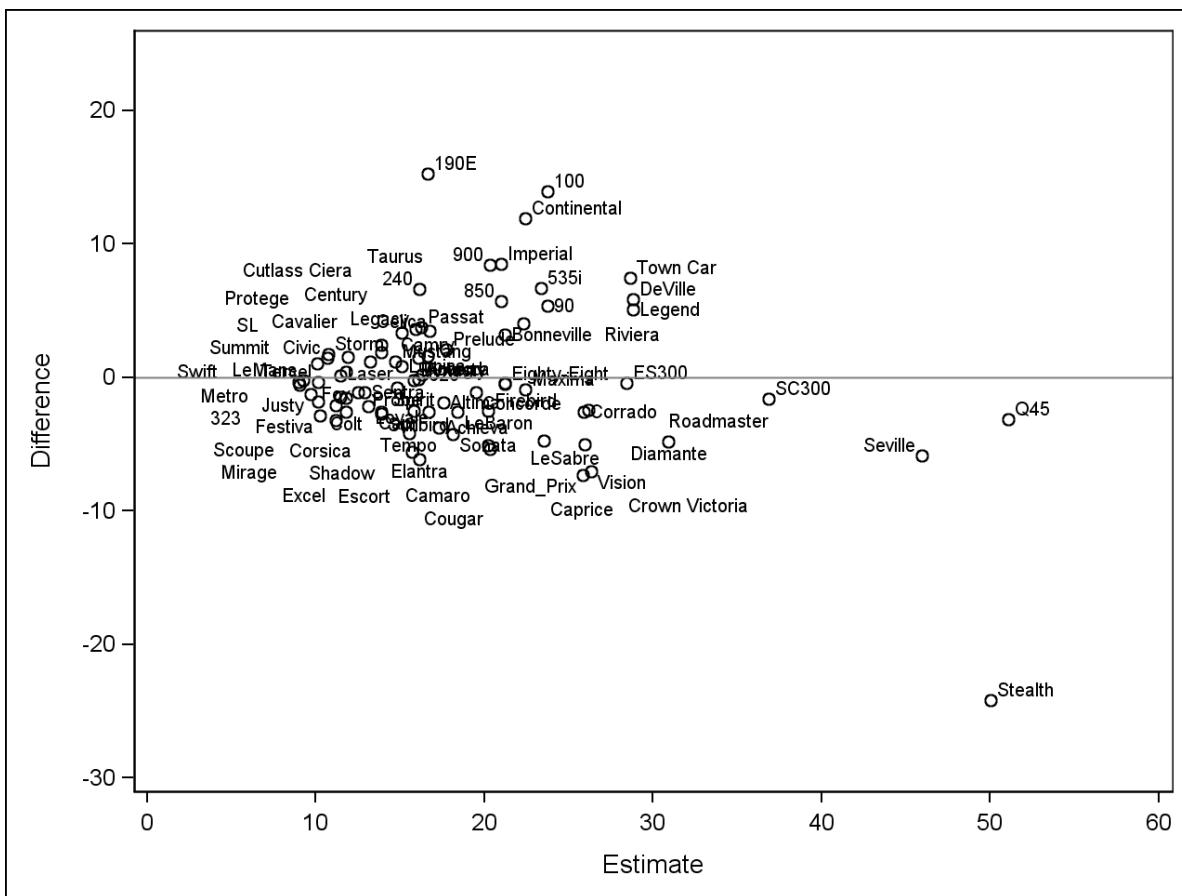
There is at least one observation that seems to have a large difference (in absolute value) between the observed **Price** and the estimated value based on the model and the back-transformation. Rerunning the program with the **DATALABEL** option enables you to see which observations might be suffering from lack of fit.

```
proc sgplot data=out;
  scatter x=estimate y=difference /datalabel=model;
  xaxis min=0 max=60;
  yaxis min=-30 max=25;
  refline 0;
run;
quit; *ST20Dd06.sas;
```

Selected SCATTER statement option:

**DATALABEL** displays a label for each data point. If you specify a variable, then the values of that variable are used for the data labels. If you do not specify a variable, then the values of the Y variable are used for the data labels.

## PROC SGLOT Output



The residuals for predicted values above \$30,000 have negative residuals. The Dodge Stealth still has a large negative residual for this model.

Remember that the independent variables in this model were selected before transforming the response variable. You should return to the modeling cycle and select the appropriate variables for the transformed dependent variable. Then perform model diagnostics (as was done in the exercises).

## D.03 Poll

When you transform the dependent variable to meet the model assumptions, you can always detransform the predicted value to obtain the unbiased estimates of the means on the original scale.

- True
- False

107

Recall that in the previous demonstrations for this course, you found that residuals from the ANOVA model for **STAT2.school** data set did not seem to be normal. The variances seemed to be equal.

Transformation of the **Reading3** scores might correct the normality problem, but you must take care to ensure that it does not create an unequal variance problem. Recall that ANOVA is robust to departures from normality, especially with a large enough sample size. Because the departure from normality was not extreme, a transformation might not be needed.

However, if you decide that a transformation is appropriate, you can use PROC TRANSREG to select the appropriate Box-Cox transformation to normalize the data. Box Cox transformations are defined only for positive values of the dependent variable, so the PARAMETER= option must be used to add a specified amount to each value of **Reading3** to eliminate any values of zero.



## Transforming Variables as a Remedial Measure for Departures from Normality

```

title;
proc transreg data=STAT2.school nomiss;
  model boxcox(reading3/parameter=0.0000001)=class(gender
    semesters school gender*school);
run;
quit; *ST20Dd07.sas;

```

Selected MODEL statement options:

- CLASS** expands the variables to a set of coded or dummy variables. PROC TRANSREG uses the values of the formatted variables to determine class membership. The specification class (x1 x2) fits a simple main-effects model. Class (x1 | x2) fits a main-effects and interactions model. Variables specified with the CLASS expansion can be either character or numeric; numeric variables should be discrete.
- PARAMETER=** specifies the transformation parameter. The **PARAMETER=t-option** is available for the BOXCOX, EXP, LOG, POWER, SMOOTH, SSPLINE, and PBSPLINE transformations. For BOXCOX, the parameter is the value to add to each value of the variable before a Box-Cox transformation, because Box-Cox transformations can be performed only on a dependent variable with nonnegative values.

The TRANSREG Procedure  
Box-Cox Transformation Information for Reading3

Lambda	R-Square	Log Like
-3.00	0.04	-10367.6
-2.75	0.04	-9517.4
-2.50	0.04	-8668.8
-2.25	0.04	-7821.9
-2.00	0.04	-6977.3
-1.75	0.04	-6135.5
-1.50	0.04	-5297.3
-1.25	0.04	-4464.3
-1.00	0.04	-3638.5
-0.75	0.04	-2824.3
-0.50	0.04	-2031.1
-0.25	0.05	-1287.9
0.00	0.12	-720.7
0.25	0.22	-585.0 <
0.50	0.21	-590.3
0.75	0.19	-614.2
1.00	0.18	-646.8
1.25	0.16	-685.6
1.50	0.15	-729.5
1.75	0.14	-777.8
2.00	0.13	-829.8
2.25	0.12	-885.2
2.50	0.11	-943.5
2.75	0.10	-1004.6
3.00	0.09	-1068.0

< - Best Lambda

PROC TRANSREG selected 0.25 as the best lambda. Because the transformation corresponding to this lambda might be difficult to interpret, the program below makes logarithm and square root transformations in addition to the transformation that was selected. The program creates descriptive statistics and plots of the residuals for each transformation.

 Recall that the Box-Cox transformation is  $y' = \frac{y^\lambda - 1}{\lambda} \quad (\lambda \neq 0)$ .  
 $\log(y) \quad (\lambda = 0)$

```
data school_t;
  set STAT2.school;
  IDReading3=reading3;
  BoxCoxReading3=(reading3**.25 - 1) / .25;
  LogReading3=log(reading3);
  SqrtReading3=sqrt(reading3);
run;                                *ST20Dd07.sas;
```

```
%macro transform (trans);
title "&trans.Residuals";
ods select ResidualHistogram QQPlot ResidualByPredicted;
proc glm data=school_t plots(unpack)=diagnostics;
  class gender semesters school;
  model &trans.reading3 = gender semesters school gender*school;
  output out=&trans.check r=&trans._Residuals;
run;
quit;

ods select TestsForNormality;
ods output moments=&trans.moments;
proc univariate data=&trans.check normal;
  var &trans._residuals;
  histogram / normal;
run;
%mend transform;

%transform (ID);
%transform (BoxCox);
%transform (Log);
%transform (Sqrt);
title;

data compare (keep=varname Skewness Kurtosis);
  length varname $20;
  set idmoments boxcoxmoments logmoments sqrtmoments;
  Skewness=cvalue1;
  Kurtosis=cvalue2;
  where label1= 'Skewness';
run;

proc print data=compare;
title 'Compare Skewness and Kurtosis for the Transformations';
run;
title;                                *ST20Dd07.sas;
```

## PROC PRINT Output

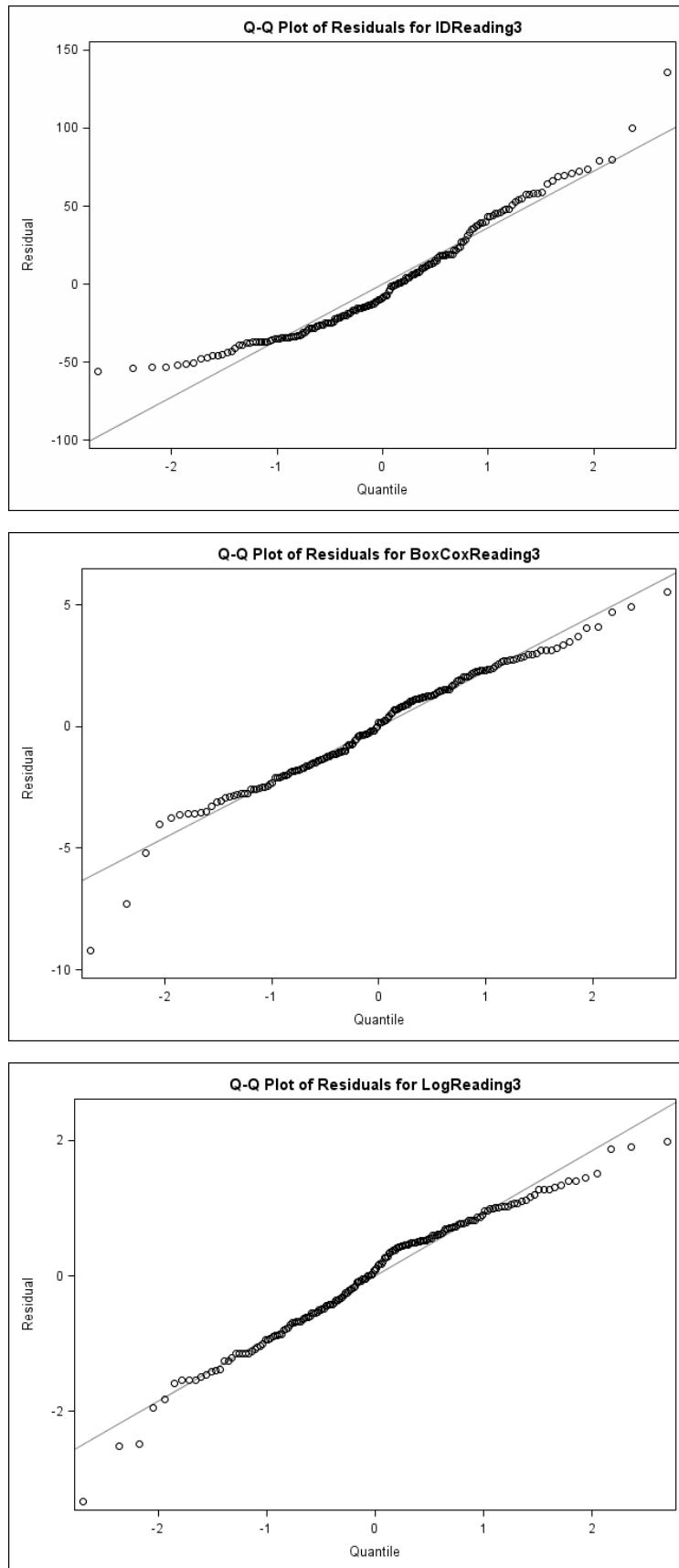
Compare Skewness and Kurtosis for the Transformations				
	Obs	varname	Skewness	Kurtosis
1		ID_Residuals	0.78202265	0.26740459
2		BoxCox_Residuals	-0.4301994	0.82945211
3		Log_Residuals	-0.4850826	0.20954248
4		Sqrt_Residuals	0.24914819	-0.6535832

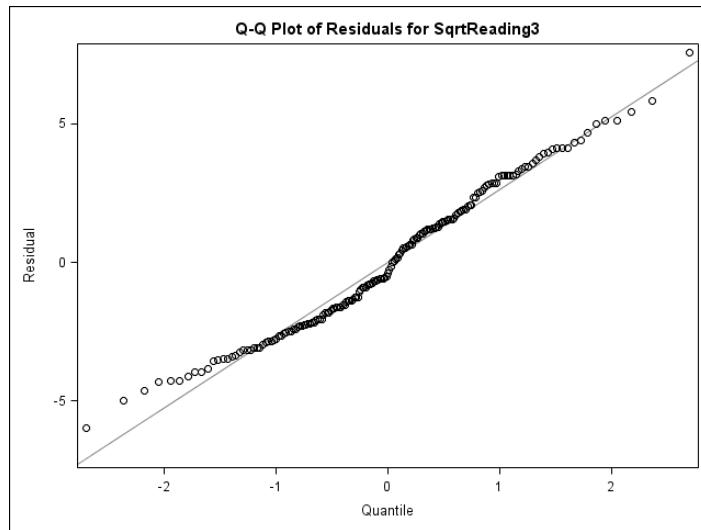
Comparing the descriptive statistics for the residuals from the model for the untransformed data (**IDResiduals**) to the residuals from the models for the transformed data indicates that all of the transformations reduced the absolute value of the skewness statistic. The log transformation also reduced the absolute value of the kurtosis statistic. (As was noted previously, the skewness and kurtosis for the original model indicate only slight departures from normality.)

## Selected PROC UNIVARIATE Output

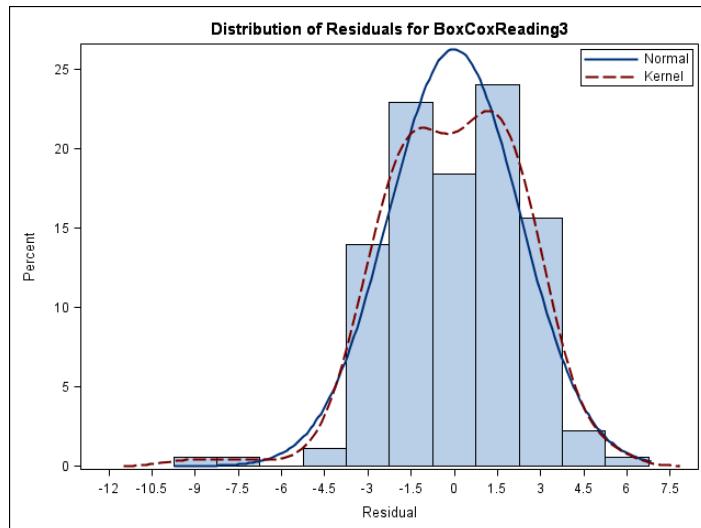
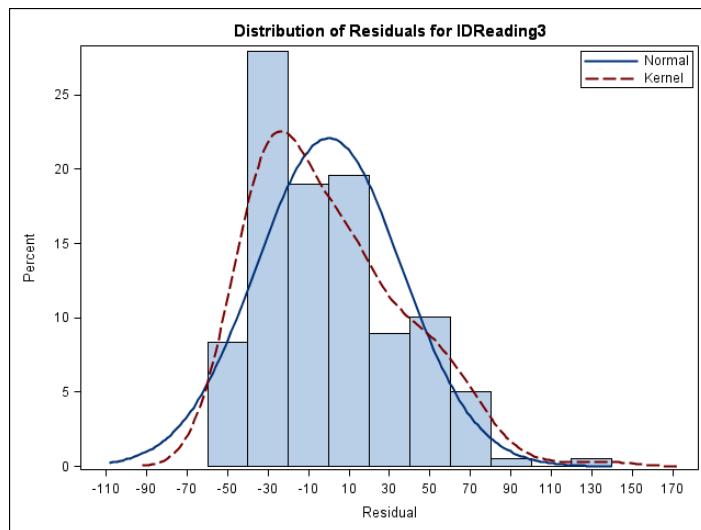
The UNIVARIATE Procedure				
Variable: ID_Residuals				
Tests for Normality				
Test	--Statistic--	-----	p Value-----	
Shapiro-Wilk	W	0.945331	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.104976	Pr > D	<0.0100
Cramér-von Mises	W-Sq	0.46281	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.80971	Pr > A-Sq	<0.0050
Variable: BoxCox_Residuals				
Tests for Normality				
Test	--Statistic--	-----	p Value-----	
Shapiro-Wilk	W	0.977746	Pr < W	0.0058
Kolmogorov-Smirnov	D	0.067715	Pr > D	0.0439
Cramér-von Mises	W-Sq	0.115535	Pr > W-Sq	0.0731
Anderson-Darling	A-Sq	0.755435	Pr > A-Sq	0.0486
Variable: Log_Residuals				
Tests for Normality				
Test	--Statistic--	-----	p Value-----	
Shapiro-Wilk	W	0.977654	Pr < W	0.0060
Kolmogorov-Smirnov	D	0.097165	Pr > D	<0.0100
Cramér-von Mises	W-Sq	0.208907	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.153318	Pr > A-Sq	0.0050
Variable: Sqrt_Residuals				
Tests for Normality				
Test	--Statistic--	-----	p Value-----	
Shapiro-Wilk	W	0.980745	Pr < W	0.0141
Kolmogorov-Smirnov	D	0.079566	Pr > D	<0.0100
Cramér-von Mises	W-Sq	0.224171	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.267907	Pr > A-Sq	<0.0050

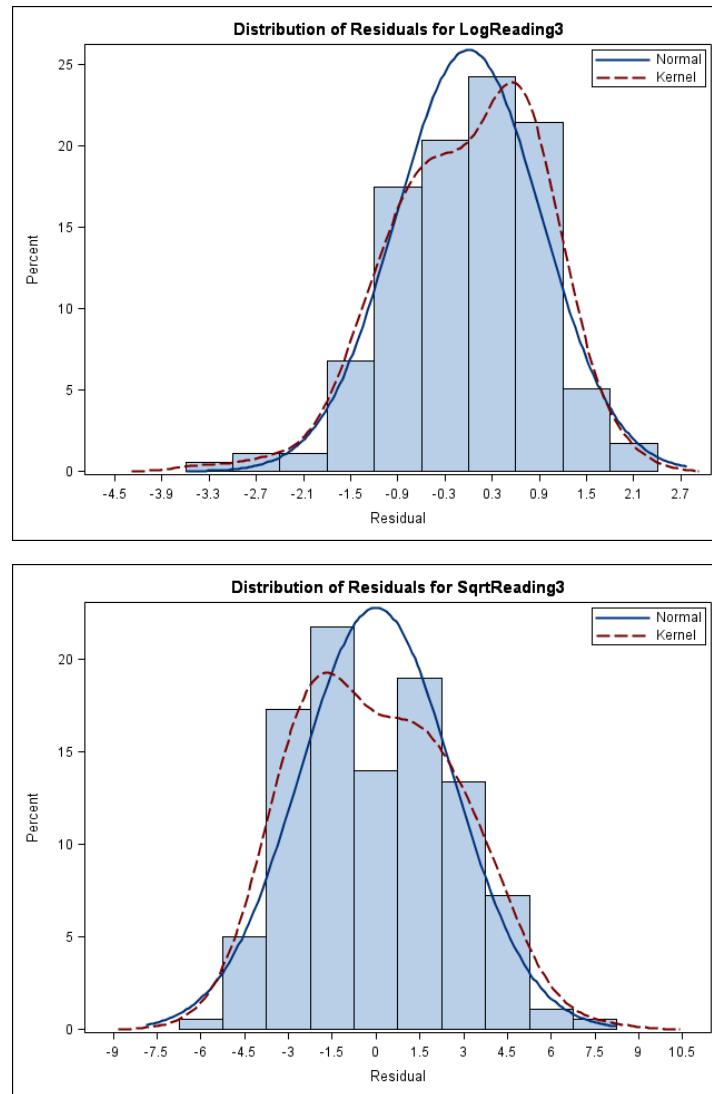
The results of the tests for normality should be considered in conjunction with the graphs, but the test for normality of the residuals from the model using the Box-Cox transformation has the highest *p*-values for the tests. The Cramér-von Mises test for this transformation has a *p*-value of 0.0731. This indicates that there is not enough evidence to reject the assumption of normality. The Kolmogorov-Smirnov and Anderson-Darling tests have *p*-values just below 0.05.



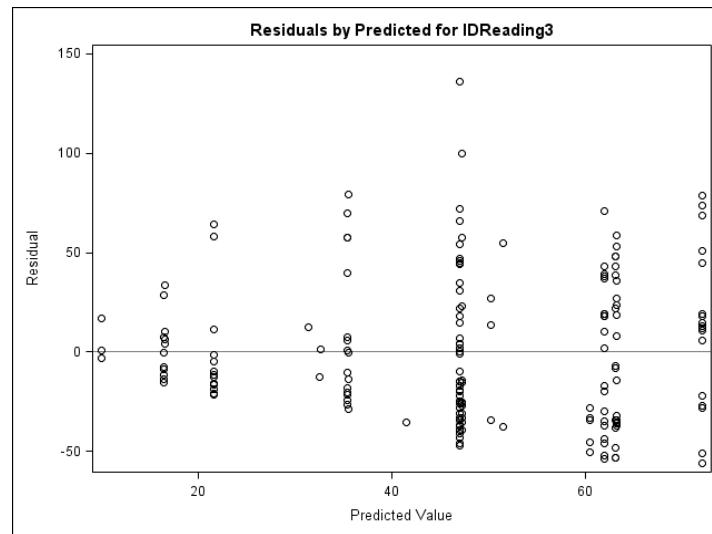


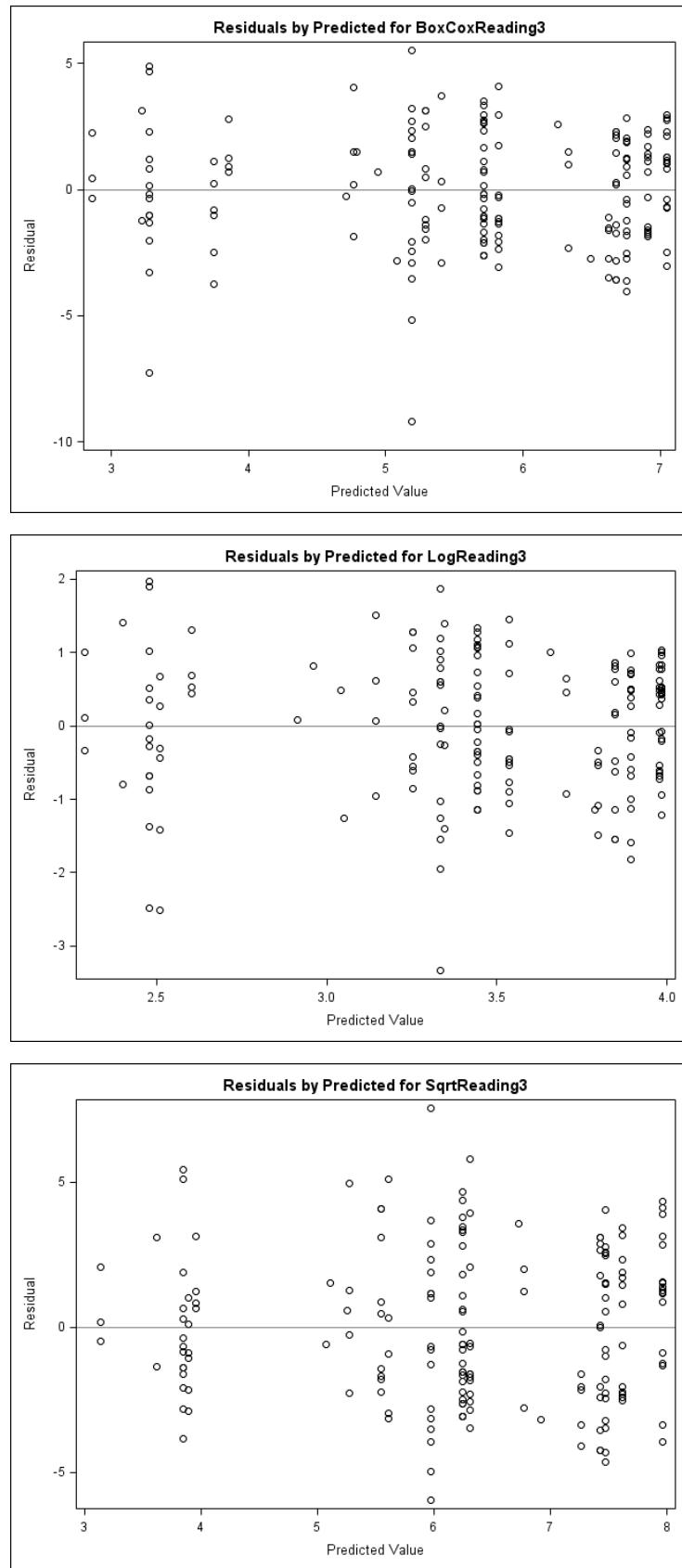
None of the normal quantile plots for the transformed data indicates any severe departures from normality. The same is true for the histograms that follow.





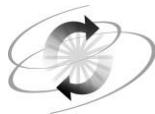
The residual plots for the transformed models do not show any evidence of nonconstant variances.







Remember that ANOVA is robust to departures from normality, especially with a large enough sample size. If you do not think the departures from the normality are gross for the model on the original scale of the data, and your sample size of 179 is large enough, then you can proceed with the analysis results from the original model. Otherwise, you might consider a model based on one of the demonstrated transformations.



## Exercises

---

### 1. Exploring Remedial Measures for Violations of the Assumptions in an ANOVA

The Marketing department of a catalog company wants to evaluate the effectiveness of different types of advertising. They have three advertisement designs in two sizes and collected data about the number of requests for catalogs that they receive as a result of these advertisements. These are the variables in the **STAT2.catalog** data set:

- Design** the advertisement design (*A*, *B*, or *C*)  
**Size** the size of the advertisement (*Small* or *Large*)  
**Requests** the number of catalog requests received

 Although it might be preferable to fit a Poisson regression model to the data, you are asked to fit a two-way ANOVA model to the data and explore remedial measures by transforming the dependent variable.

```
proc glm data=STAT2.catalog;
  class design size;
  model requests=design | size;
  output out=check r=residuals p=predicted;
run;
quit; *ST20Dd08_1.sas;
```

It can be shown that after you fit a two-way ANOVA model to the data using the code above, the following is true:

- The **Design\*Size** interaction term is significant.
- The residuals are not normally distributed.
- The variances do not seem to be the same across different groups. In fact, the variances seem to be a function of the means. The variance is larger for groups with larger means.
  - a. Use the MEANS procedure with the NWAY and NOPRINT options and the OUTPUT statement to output the mean and variance for each treatment combination to a data set. Use a DATA step to compute the ratio as the variance divided by the different powers of the mean. Examine the values to determine which relationship is the most stable (which ratio is the least variable across groups). (You can use PROC SQL to compare the minimum and maximum of these ratios to determine which relationship is the most stable.)
  - b. Because the ratio of the variance to the mean squared is the most stable, use a log transformation and evaluate the new model.
  - c. Does this transformation correct the violations of the assumptions of the model with the untransformed data? Why or why not?

**Advanced****2. Comparing Analyses: PROC REG with a Log Transformation versus PROC GENMOD with a Log Link**

The **STAT2.nonlin** data set contains the variables **X**, **Y**, and **LogY**, where **LogY** is equal to the natural log of **Y**.

- a. Use the SGPlot procedure to plot **Y** versus **X**.
- b. Look up the XAXIS, X2AXIS, YAXIS, and Y2AXIS Statements in the online documentation for PROC SGPlot. Request that the Y axis be shown on the log scale (base e). How does this affect the graph and what does it suggest?
- c. Use PROC REG to fit a regression line with **LogY** as the response variable and **X** as the predictor variable. Examine the residual plot and the fit plot to determine whether this is a good model. (Information about the fit plot can be found in the SAS 9 online documentation for PROC GENMOD.)
- d. Use PROC GENMOD to fit **Y** by **X**. Use the DIST=NORMAL and LINK=LOG options in the MODEL statement. Request the plots for the studentized residuals. Does this model appear to be a good fit? Compare this model to the previous model.

## D.5 Weighted Least Squares

---

### Objectives

- Introduce weighted least squares for modeling data with heteroscedasticity.
- Demonstrate how to find appropriate weights.
- Fit a weighted least squares model in PROC REG.

112

The ordinary least squares method does not provide minimum variance parameter estimates when variances are not constant. However, if the errors are assumed to be independent, then weighted least squares can be used to reduce the influence of highly variable observations. The weighted least squares method is a direct application of generalized least squares. Parameter estimates are obtained by

minimizing the weighted residual sum of squares:  $\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2$ , where  $w_i$  is a set of nonnegative

weights assigned to the individual observations. Observations with small weights contribute less to the sum of squares and thus provide less influence to the estimation of parameters, and vice versa for observations with larger weights. Therefore, you want to assign larger weights to observations with smaller variances, and likewise assign smaller weights to observations whose large variances make them more unreliable. It can be shown that the best linear unbiased estimates are obtained if the weights are inversely proportional to the variances of the individual errors.

The weighted least squares regression is equivalent to the ordinary least squares regression with the square root of the weights applied to the  $\mathbf{X}$  matrix and the  $y$  vector in the regression model  $y = \mathbf{X}\beta + \varepsilon$ . In other words, if you multiply all the independent variables and the dependent variable by the square root of the weight, and perform the ordinary least squares regression on the transformed data, you obtain the same parameter estimates and standard errors as the weighted least squares regression with the WEIGHT statement.

## Obtaining the Weight

- Variances are usually unknown.
- Use of weights based on poor estimates of variances might be counterproductive.
- You might want to group your data to obtain estimated variances for different groups if replicates are not available.
- You might also use the information of how the variance changes as a function of the means to obtain the weights.

114

The variances of the individual errors are usually unknown. Multiple observations (or replicates), if present in your data, can be used to estimate this variance (Freund and Wilson 1998). If true replicates are not available, near neighbors (Montgomery and Peck 1982) can be used. However, these are all estimated variances, and the use of weights based on poor estimates of variances might be counterproductive.

Alternatively, knowledge about the distribution of residuals might provide a basis for determining weights. For example, the residual plot suggests that the variances of residuals might be a function of the means. The fact that the logarithmic transformation stabilized the variance suggests that the variances might be proportional to the squares of means. The estimated means can be obtained based on the predicted values from the unweighted regression model.



## Weighted Least Squares Using PROC REG

```

proc reg data=STAT2.cars2;
  model price=hwympg hwympg2 horsepower;
  output out=out p=pred;
  title 'Ordinary Least Squares Model';
run;
quit;

data out;
  set out;
  w=1/(pred*pred);
run;

proc reg data=out plots(unpack)=all;
  model price=hwympg hwympg2 horsepower;
  weight w;
  output out=wout p=wpred r=residual;
  title 'Weighted Least Squares Model';
run;
quit;                                *ST20Dd08.sas;

```

Selected REG procedure statement:

**WEIGHT** specifies a variable in the input data set with values that are relative weights for a weighted least squares fit. Values of the weight variable must be nonnegative. If an observation's weight is zero, the observation is deleted from the analysis. If a weight is negative or missing, then it is set to zero, and the observation is excluded from the analysis.

The DATA step was used to compute the weight as the reciprocal of the squares of the predicted values (or the estimated means). You assume that the variances are proportional to the squares of the means. If the weight value is proportional to the reciprocal of the variance for each observation, then the weighted estimates are the best linear unbiased estimates (BLUE).

## Partial Output from the First PROC REG Statement

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4448.69341	1482.89780	65.73	<.0001
Error	77	1737.20536	22.56111		
Corrected Total	80	6185.89877			

Root MSE	4.74985	R-Square	0.7192
Dependent Mean	18.64321	Adj R-Sq	0.7082
Coeff Var	25.47766		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	4.03949	2.17024	1.86	0.0665
Hwympg	1	-0.80407	0.21378	-3.76	0.0003
hwympg2	1	0.04350	0.01430	3.04	0.0032
Horsepower	1	0.09730	0.01614	6.03	<.0001

## Partial Output from the Second PROC REG Statement

Weight: w

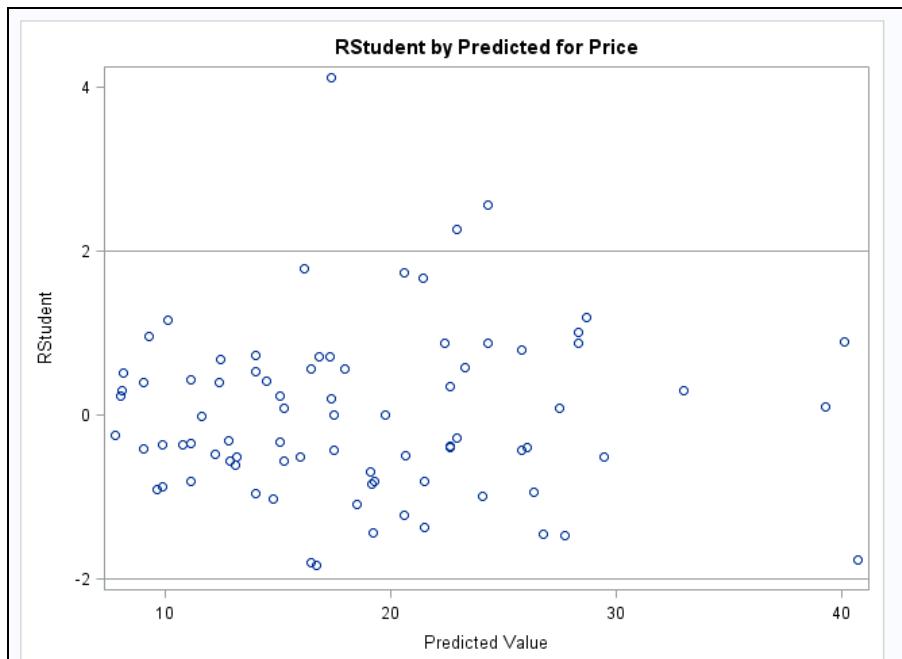
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11.62453	3.87484	79.01	<.0001
Error	77	3.77606	0.04904		
Corrected Total	80	15.40059			

Root MSE	0.22145	R-Square	0.7548
Dependent Mean	13.54827	Adj R-Sq	0.7453
Coeff Var	1.63452		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	1.86213	1.82175	1.02	0.3099
Hwympg	1	-0.53301	0.16186	-3.29	0.0015
hwympg2	1	0.03043	0.00948	3.21	0.0019
Horsepower	1	0.11508	0.01463	7.87	<.0001

A comparison of the results from the weighted least squares and unweighted least squares shows that the results are not identical. Because sums of squares are a function of the weights, they cannot be compared with each other from these two models. However, the R-square statistics are comparable because they are ratios and the effect of weights is canceled. The R-squared value is larger for the weighted analysis (0.7548) than the unweighted analysis (0.7192). This is because the effect of the largest residuals was reduced in the weighted analysis. Also, the estimated equation has smaller coefficients for **Hwympg2**. This indicates that the curve has a somewhat smaller curvature, which is presumably due to the lesser influence of the high or low **Price** values.

Finally, you might want to examine the plot of the studentized residuals.



The variances of the studentized residuals appear to be constant. However, three observations appear to have large studentized residuals.

The weighted least squares regression is equivalent to the ordinary least squares regression with the square root of the weights applied to the **X** matrix and the **y** vector in the regression model  $y=X\beta+\epsilon$ . In other words, if you multiply all the independent variables and the dependent variable by the square root of the weight, and perform the ordinary least squares regression on the transformed data, you obtain the same parameter estimates and standard errors as the weighted least squares regression with the WEIGHT statement. Because the **X** matrix contains a column of ones for the intercept, you must create a new variable that has the transformed value (1\*square root of the weight), and use the NOINT option in the MODEL statement in PROC REG.

```
data wcars2;
  set out;
  Sqrtwgt=sqrt(w);
  WPrice=price*sqrtwgt;
  wHwmpg=hwmpg*sqrtwgt;
  WHwmpg2=hwmpg2*sqrtwgt;
  WHorsepower=horsepower*sqrtwgt;
  WIInt=1*sqrtwgt;
run;
```

```

proc reg data=wcars2;
  model wprice=wint whwympg whwympg2 whorsepower / noint;
  title 'Without WEIGHT Statement on Transformed Data';
run;
title; quit;
*ST20Dd08.sas;

```

Selected MODEL statement option:

**NOINT** suppresses the intercept term that is otherwise included in the model.

A DATA step is used to transform the dependent variable **Price** and all the independent variables (**Hwympg**, **Hwympg2**, and **Horsepower**) as well as the column of ones in the **X** matrix.

PROC REG Output

### Without WEIGHT Statement on Transformed Data

The REG Procedure  
Model: MODEL1  
Dependent Variable: WPrice

Number of Observations Read	81
Number of Observations Used	81

**Note:** No intercept in model. R-Square is redefined.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	81.40084	20.35021	414.97	<.0001
Error	77	3.77606	0.04904		
Uncorrected Total	81	85.17690			

Root MSE	0.22145	R-Square	0.9557
Dependent Mean	1.00160	Adj R-Sq	0.9534
Coeff Var	22.10949		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
WInt	1	1.86213	1.82175	1.02	0.3099
WHwympg	1	-0.53301	0.16186	-3.29	0.0015
WHwympg2	1	0.03043	0.00948	3.21	0.0019
WHorsepower	1	0.11508	0.01463	7.87	<.0001

The sums of squares and the R-square value are different from the weighted least squares model using the WEIGHT statement. This is because the sums of squares are computed differently using these two approaches. However, the parameter estimates and standard errors are identical. This illustrates how weighted least squares regression relates to the ordinary least squares model fit to the transformed data.

## D.6 Evaluating the Importance of Parameters

---

### Objectives

- Evaluate the importance of parameters by standardizing parameter estimates in PROC REG.



## Evaluating the Importance of Parameters

Presume that for the **STAT2.cars** data set, you chose to use the regression model with three variables (**Hwympg**, **Hwympg2**, and **Horsepower**). You might want to evaluate the relative importance among these three variables in explaining the variance in **Price**.

```
ods html close;
ods listing;
proc reg data=STAT2.cars2;
    model price=hwympg hwympg2 horsepower / stb;
run;
quit; *ST20Dd09.sas;
```

Selected MODEL statement option:

**STB** produces standardized regression coefficients. A standardized regression coefficient is computed by dividing a parameter estimate by the ratio of the sample standard deviation of the dependent variable to the sample standard deviation of the regressor.

Partial PROC REG Output

Variable	DF	Parameter Estimates				
		Parameter Estimate	Standard Error	t Value	Pr >  t	Standardized Estimate
Intercept	1	4.03949	2.17024	1.86	0.0665	0
Hwympg	1	-0.80407	0.21378	-3.76	0.0003	-0.45822
Hwympg2	1	0.04350	0.01430	3.04	0.0032	0.27668
Horsepower	1	0.09730	0.01614	6.03	<.0001	0.56031

The standardized estimate for each variable can be thought of as a relative measure of the importance of the predictor variables. The standardized estimate for **Horsepower** is the biggest (in absolute value), 0.56, which indicates that **Horsepower** affects the price more than other variables.

Now that you narrowed down the models under consideration, you can evaluate the models by checking to see whether the assumptions of the regression were met. In addition, influential observations should be identified and the model should be checked for collinearity.

## D.7 A Sample SAS Program for Comparing Model Fit

---

### Objectives

- Examine a program to compare fit statistics across several models.

120

The following SAS program was used to produce the comparisons of model fits presented in a previous section.

```

proc stdize data=STAT2.cars method=mean out=STAT2.cars2;
  var citympg hwympg fueltank weight;
run;

data STAT2.cars2;
  set STAT2.cars2;
  citympg2 = citympg*citympg;
  hwympg2 = hwympg*hwympg;
  fueltank2=fueltank*fueltank;
  fueltank3=fueltank2*fueltank;
  logprice=log(price);
run;

proc sql noprint;
  select avg(price)
    into :mean
    from STAT2.cars2;
quit;                                *ST20Dd10.sas;

proc reg data=STAT2.cars2;
  model logprice = hwympg hwympg2 horsepower;
  ods output ANOVA=ANOVATable;
  output out=out p=pred;
run;
```

```

quit;

data _null_;
  set ANOVATable;
  if source='Error' then call symput('var', MS);
run;

%let n=81;
%let p=4;
data lognormal(keep=MSElog R2log adjR2log);
  retain sumdiff 0 sumtotal 0;
  set out end=eof;
  estimate = exp(pred + &var/2);
  difference = price - estimate;
  sumdiff = sumdiff + difference**2;
  sumtotal = sumtotal + (price-&mean)**2;
  if eof then do;
    MSElog = sumdiff / (&n-&p);
    R2log = 1 - sumdiff/sumtotal;
    adjR2log = 1-(sumdiff / (&n-&p)) / (sumtotal/(&n-1));
    output;
  end;
run;                                         *ST20Dd10.sas;
ods output obstats=gamma_obstats;
proc genmod data=STAT2.cars2;
  model price = hwympg hwympg2 horsepower / dist=gamma link=log
                                             obstats;
  title 'Assumed Gamma Distribution';
run;

data gamma(keep=MSEgammalog R2gammalog adjR2gammalog );
  retain sumdiff 0 sumtotal 0;
  set gamma_obstats end=eof;
  sumdiff = sumdiff + resraw**2;
  sumtotal = sumtotal + (price-&mean)**2;
  if eof then do;
    MSEgammalog = sumdiff / (&n-&p);
    R2gammalog = 1 - sumdiff/sumtotal;
    adjR2gammalog = 1-(sumdiff / (&n-&p)) / (sumtotal/(&n-1));
    output;
  end;
run;

ods output obstats=gamma_obstats;
proc genmod data=STAT2.cars2;
  model price = hwympg hwympg2 horsepower / dist=gamma link=identity
                                             obstats;
  title 'Assumed Gamma Distribution';
run;

```

```

data gamma2(keep=MSEgammaiden R2gammaiden adjR2gammaiden) ;
retain sumdiff 0 sumtotal 0;
set gamma_obstats end=eof;
sumdiff = sumdiff + resraw**2;
sumtotal = sumtotal + (price-&mean)**2;
if eof then do;
  MSEgammaiden = sumdiff / (&n-&p);
  R2gammaiden = 1 - sumdiff/sumtotal;
  adjR2gammaiden = 1-(sumdiff / (&n-&p)) / (sumtotal/(&n-1));
  output;
end;
run;

data all;
  merge lognormal gamma gamma2;
run;

proc print;
run;
quit;
*ST20Dd10.sas;

```

## D.8 Incorrectly Treating Random Effects as Fixed

---

### Objectives

- Introduce the idea of Expected Mean Squares.
- Evaluate the effect of incorrectly treating random effects as fixed effects using expected mean squares in PROC MIXED.

122

Recall that data were collected by a school district to assess the reading skill progress of students in their first year of formal schooling. The data are stored in the **STAT2.school** data set. In a previous chapter, you used ANOVA models to evaluate the significance of some factors in explaining the difference in the average **Reading3** test scores. There is another variable, **Teacher**, in the data set. Teachers are nested within schools. Factors of interest are **School** and **Gender**.



## Comparing Expected Mean Squares for Fixed and Random Effects

In this data set, **Teacher(Material)** is a random effect. What would happen if you incorrectly specified **Teacher(Material)** as a fixed effect?

```
title 'Random Effect is Incorrectly Specified as Fixed Effect';
proc glimmix data=STAT2.scores;
  class material teacher;
  model score=material teacher(material);
  output out=checkvar variance=ResidualVariance;
run; *ST20Dd07.sas;
```

The OUTPUT statement requests that the residual variance estimate be output to the **Checkvar** data set.

PROC PRINT Output

### Random Effect is Incorrectly Specified as Fixed Effect

Obs	ResidualVariance
1	17.255

The estimate for the residual variance is the same as the one obtained from the previous model. This is because you have balanced data.

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
<b>Material</b>	3	100	6.99	0.0003
<b>Teacher(Material)</b>	16	100	13.05	<.0001

Now, the *F* value for the **Material** effect is 6.99 with the denominator degrees of freedom equal to 100. The *p*-value is 0.0003. You would incorrectly conclude that there is a significant difference in the average test scores among the four teaching materials.

Why is there a big difference in the **Material** effect between treating **Teacher(Material)** as random and treating it as fixed?

This question might be best answered by examining the expected mean squares table produced by the METHOD=TYPE3 option in the PROC MIXED statement. PROC MIXED is designed to fit linear mixed models and has many statements similar to PROC GLIMMIX as seen earlier.

```
title 'Expected Mean Squares for the Correct Model';
proc mixed data=STAT2.scores method=type3;
  class material teacher;
  model score=material;
  random teacher(material);
run; *ST20Dd07.sas;
```

Selected PROC MIXED option:

METHOD= specifies the estimation method for the covariance parameters. Possible values for METHOD are REML (the default), ML, MIVQUE0, TYPE1, TYPE2, and TYPE3.

The METHOD=TYPE*n* specifications apply only to variance component models with no SUBJECT= effects and no REPEATED statement. An analysis of variance table is included in the output, and the expected mean squares are used to estimate the variance components. The resulting method-of-moment variance component estimates are used in subsequent calculations, including standard errors computed from ESTIMATE and LSMEANS statements.

For balanced data, using REML (the default) or TYPE3 produces identical results for the covariance parameter estimates and tests for fixed effects in PROC MIXED. For unbalanced data, the results might differ between the two methods, but the differences between treating an effect fixed or random can still be illustrated using the METHOD=TYPE3 option.

-  The METHOD=TYPE3 option in the PROC MIXED statement produces the covariance parameter estimates that are based on expected mean squares table.

### Details

The expected mean squares are computed as follows. Consider the following model:

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$$

$\boldsymbol{\beta}_0$  represents the fixed effects and  $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\epsilon}$  represent the random effects. Random effects are assumed to be normally and independently distributed. For any  $\mathbf{L}$  in the row space of  $\mathbf{X} = (\mathbf{X}_0 | \mathbf{X}_1 | \mathbf{X}_2 | \cdots | \mathbf{X}_k)$ , the expected value of the sum of squares for  $\mathbf{L}\boldsymbol{\beta}$  is the following:

$$E(SS_L) = \mathbf{B}'_0 \mathbf{C}'_0 \mathbf{C}_0 \boldsymbol{\beta}_0 + SSQ(\mathbf{C}_1)\sigma_1^2 + \cdots + SSQ(\mathbf{C}_k)\sigma_k^2 + rank(\mathbf{L})\sigma_\epsilon^2$$

where  $\mathbf{C}$  is of the same dimensions as  $\mathbf{L}$  and is partitioned as the  $\mathbf{X}$  matrix. In other words,

$$\mathbf{C} = (\mathbf{C}_0 | \mathbf{C}_1 | \cdots | \mathbf{C}_k)$$

Furthermore,  $\mathbf{C} = \mathbf{ML}$ , where  $\mathbf{M}$  is the inverse of the lower triangular Cholesky decomposition matrix of  $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$ . SSQ(A) is defined as  $tr(\mathbf{A}'\mathbf{A})$ .

Partial PROC MIXED Output

### Expected Mean Squares for the Correct Model

Type 3 Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term
Material	3	361.691667	120.563889	Var(Residual) + 6 Var(Teacher(Material)) + Q(Material)	MS(Teacher(Material))
Teacher(Material)	16	3603.400000	225.212500	Var(Residual) + 6 Var(Teacher(Material))	MS(Residual)
Residual	100	1725.500000	17.255000	Var(Residual)	.

The Type 3 Analysis of Variance table includes an ANOVA table with the expected mean squares for each effect.

This table shows you the correct *F* statistics for each effect. For example, to test the **Material** effect, the *F* statistic should be computed as  $F = MS(Material) / MS(Teacher(Material))$ . A large *F* value provides evidence that the mean test scores among different teaching materials might be different.

Source	DF	Sum of Squares	Mean Square	Error Term	Error DF	F Value	Pr > F
Material	3	361.691667	120.563889	MS(Teacher(Material))	16	0.54	0.6647
Teacher(Material)	16	3603.400000	225.212500	MS(Residual)	100	13.05	<.0001
Residual	100	1725.500000	17.255000	.	.	.	.

The column labeled **Error Term** provides the correct denominator for the *F* statistics for each effect. The denominator for the **Material** effect is **MS(Teacher(Material))** rather than the default **MS(Residual)** in fixed effects models. The *p*-value of 0.6647 for this effect suggests that there are not significant differences in the average test scores among the four teaching materials. The test for **Teacher(Material)** tests the null hypothesis that the variance is not different from zero. The significant *p*-value of <0.0001 indicates that the variance among the teachers is significantly different from zero.

Covariance Parameter Estimates	
Cov Parm	Estimate
Teacher(Material)	34.6596
Residual	17.2550

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Material	3	16	0.54	0.6647

The results are identical to the same model using the PROC GLIMMIX with the default restricted maximum likelihood method because you have balanced data.

## Details

In statistical notation, the expected mean squares for the terms in this model are as follows:

Source	Expected Mean Squares
MS(material)	$\sigma^2 + 6\sigma_t^2 + 30 \left[ \sum_i (\alpha_i - \bar{\alpha}_i)^2 / 3 \right]$
MS(teacher(material))	$\sigma^2 + 6\sigma_t^2$
MS(Residual)	$\sigma^2$

Compare the expected mean squares from the correct model (shown above) to the expected mean squares from the incorrect model that assumes **Teacher(Material)** is a fixed effect.

```
title 'Expected Mean Squares for the Incorrect Model';
proc mixed data=STAT2.scores method=type3;
  class material teacher;
  model score=material teacher(material);
run;
title; *ST20Dd07.sas;
```

## Partial PROC MIXED Output

### Expected Mean Squares for the Incorrect Model

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
Material	3	361.691667	120.563889	Var(Residual) + Q(Material, Teacher (Material))	MS (Residual)	100	6.99	0.0003
Teacher (Material)	16	3603.400000	225.212500	Var(Residual) + Q(Teacher (Material))	MS (Residual)	100	13.05	<.0001
Residual	100	1725.500000	17.255000	Var(Residual)	.	.	.	.

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Material	3	100	6.99	0.0003
Teacher(Material)	16	100	13.05	<.0001

When you incorrectly specified **Teacher(Material)** as a fixed effect, the *F* statistic for the **Material** effect changed. The denominator of the *F* value for the **Material** effect is **MS(Residual)** rather than the correct **MS(Teacher(Material))**. The denominator degrees of freedom changed as well (from 16 to 100). This explains the difference in this test between the two models.

Covariance Parameter Estimates	
Cov Parm	Estimate
Residual	17.2550

The results for the residual variance are identical because you have balanced data.

## D.9 Solutions

---

### Solutions to Exercises

#### 1. Exploring Remedial Measures for Violations of the Assumptions in an ANOVA

- a. Use the MEANS procedure with the NWAY and NOPRINT options and the OUTPUT statement to output the mean and variance for each treatment combination to a data set. Use a DATA step to compute the ratio as the variance divided by the different powers of the mean. Examine the values to determine which relationship is most stable (which ratio is the least variable across groups). (You can use PROC SQL to compare the minimum and maximum of these ratios to determine which relationship is the most stable.)

```
*a;
proc means data=STAT2.catalog nway noprint;
  class design size;
  var requests;
  output out=transform mean=Mean var=Variance n=N;
run;

data transform;
  set transform;
  Sqrt_Transform=variance/mean;
  Log_Transform=variance/(mean**2);
  Inverse_Transform=variance/(mean**4);
run;

proc print data=transform;
  var design size n mean variance sqrt_transform
        log_transform inverse_transform;
run;                                              *ST20Dd08_1.sas;
```

#### PROC PRINT Output

Obs	Design	Size	N	Mean	Variance	Sqrt_Transform	Log_Transform	Inverse_Transform
1	A	Large	17	1.41176	0.6324	0.44792	0.31727	0.15919
2	A	Small	14	6.42857	22.7253	3.53504	0.54990	0.01331
3	B	Large	15	1.60000	1.5429	0.96429	0.60268	0.23542
4	B	Small	22	4.31818	11.5606	2.67719	0.61998	0.03325
5	C	Large	12	5.00000	18.0000	3.60000	0.72000	0.02880
6	C	Small	14	3.50000	10.1154	2.89011	0.82575	0.06741

The variable **Log\_Transform** seems to be the most stable. This is the variable that is the ratio of the variance to the mean squared. Therefore, a log transformation seems to be the most appropriate in this case.

Alternatively, you can use PROC SQL to compare the ratios.

```
proc sql;
  select max(sqrt_transform)/min(sqrt_transform) as SqrtRatio,
         max(log_transform)/min(log_transform) as LogRatio,
         max(inverse_transform)/min(inverse_transform) as
           InverseRatio from transform;
quit;                                         *ST20Dd08_1.sas;
```

SqrtRatio	LogRatio	InverseRatio
8.037209	2.602624	17.69272

- b. Because the ratio of the variance to the mean squared is the most stable, use a log transformation and evaluate the new model.

```
*b;
data catalog;
  set STAT2.catalog;
  LogRequests=log(requests);
run;

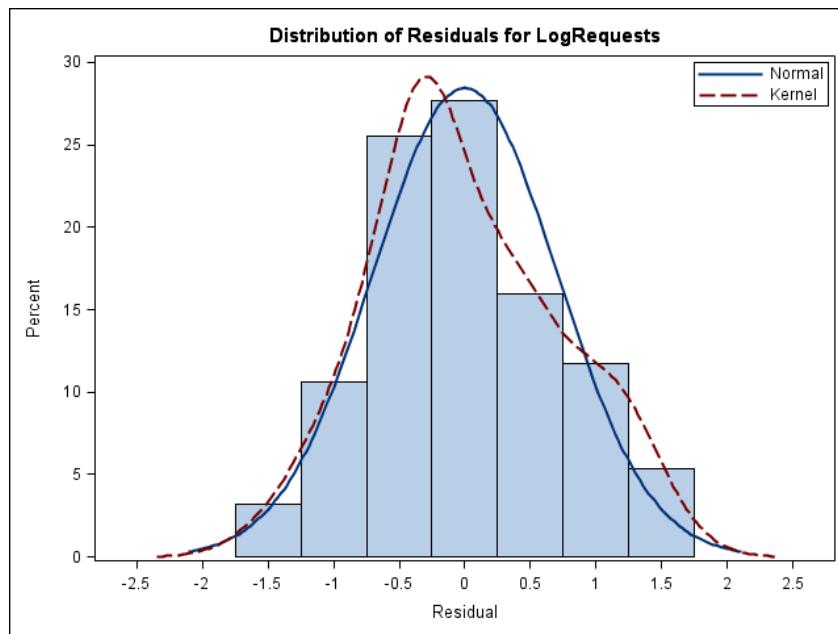
proc glm data=catalog plots(unpack)=diagnostics;
  class design size;
  model logrequests=design|size;
title 'Log Transformation';
run;
quit;                                         *ST20Dd08_1.sas;
```

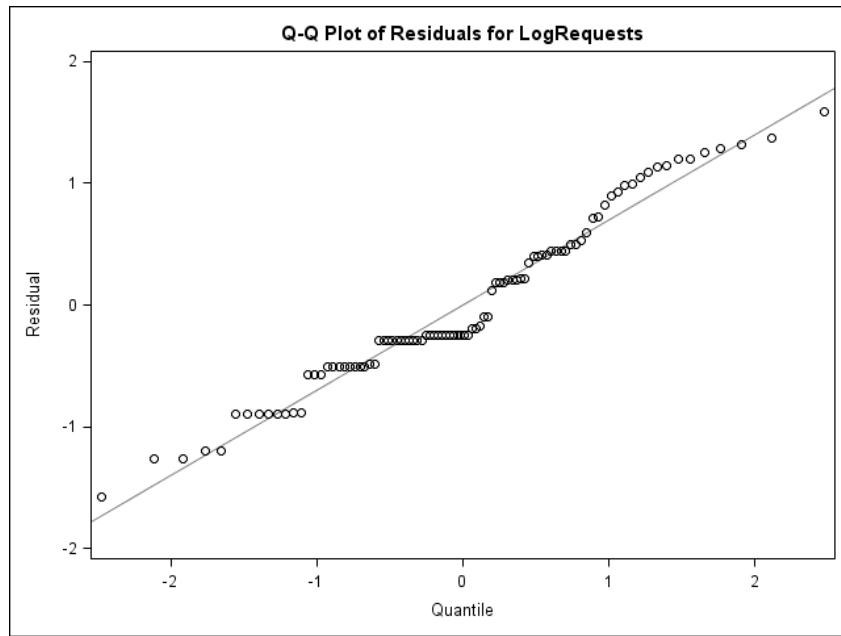
#### PROC GLM Output

Log Transformation		
The GLM Procedure		
Class Level Information		
Class	Levels	Values
Design	3	A B C
Size	2	Large Small
Number of Observations Read		94
Number of Observations Used		94

Log Transformation						
The GLM Procedure						
Dependent Variable: LogRequests						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	22.90171005	4.58034201	8.81	<.0001	
Error	88	45.73773170	0.51974695			
Corrected Total	93	68.63944175				
R-Square	Coeff Var	Root MSE	LogRequests Mean			
0.333652	79.95045	0.720935	0.901727			
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
Design	2	0.97322428	0.48661214	0.94	0.3960	
Size	1	11.01276351	11.01276351	21.19	<.0001	
Design*Size	2	10.91572226	5.45786113	10.50	<.0001	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
Design	2	1.68992843	0.84496421	1.63	0.2026	
Size	1	8.80365089	8.80365089	16.94	<.0001	
Design*Size	2	10.91572226	5.45786113	10.50	<.0001	

The model shows a significant *p*-value (less than 0.0001) for the **Design\*Size** interaction. The interaction term is significant.

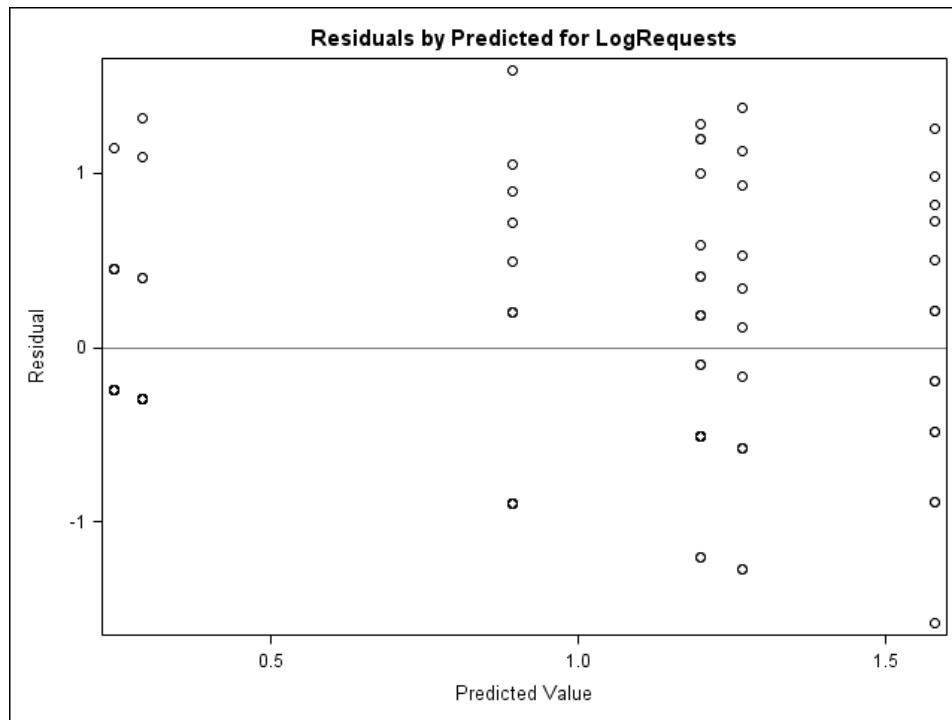




The residuals are slightly skewed to the right with slightly light tails, but neither the histogram nor the Q-Q plot of the residuals indicates severe departures from normality.

Now, evaluate the equal variance assumption.

Partial PROC GLM Graphics Output



The variance might or might not be equal for each group.

```

data catalog;
  set catalog;
  Group=compress(design||size);
run;

proc glm data=catalog;
  class group;
  model logrequests=group;
  means group/hovtest;
run;
title;
quit;                                *ST20Dd08_1.sas;

```

## Partial PROC GLM Output

Log Transformation					
The GLM Procedure					
Levene's Test for Homogeneity of logrequests Variance					
ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	5	4.0948	0.8190	2.53	0.0348
Error	88	28.5365	0.3243		
Log Transformation					
The GLM Procedure					
Level of group	N	-----logrequests-----			
		Mean	Std Dev		
ALarge	17	0.24464018	0.42028221		
ASmall	14	1.57983651	0.81706813		
BLarge	15	0.29213511	0.54725710		
BSmall	22	1.19953828	0.73227225		
CLarge	12	1.26664356	0.88730965		
CSmall	14	0.89386753	0.88224759		

The Levene's test shows a *p*-value of 0.035 for the homogeneity-of-variance test. Recall that ANOVA is robust against unequal variances when sample sizes are equal. Although the sample sizes are not equal in this example, they are fairly close. In addition, the summary statistics table shows that the standard deviations are pretty close to each other across groups. You can conclude that the variances are not different enough to cause concerns for the ANOVA model.

- c. Does this transformation correct the violations of the assumptions of the model with the untransformed data? Why or why not?

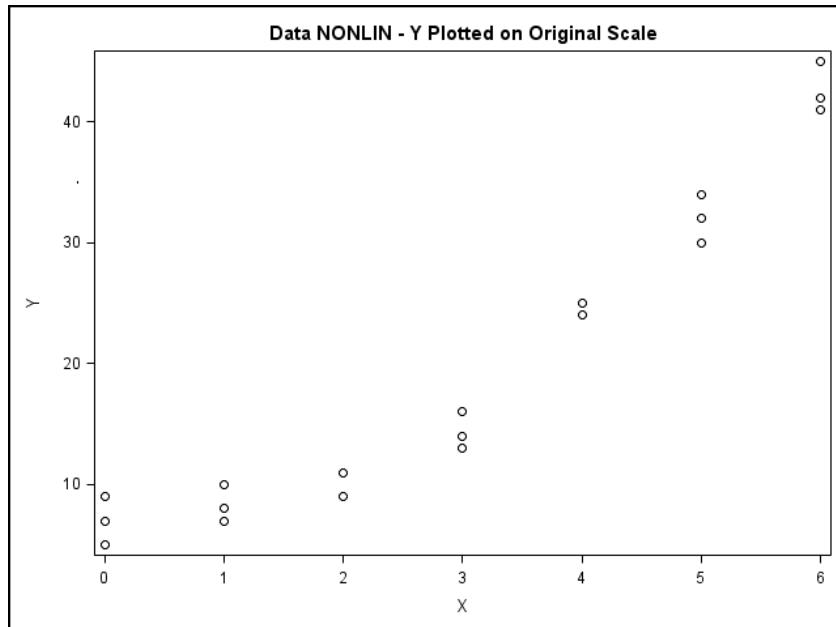
The log transformation seems to correct the violations of the assumptions of the model with the untransformed data. The residuals from the model with this transformation seem to be fairly normally distributed and the variances seem to be nearly homogeneous.

## 2. Comparing Analyses: PROC REG with a Log Transformation versus PROC GENMOD with a Log Link

The **STAT2.nonlin** data set contains the variables **X**, **Y**, and **LogY**, where **LogY** is equal to the natural log of **Y**.

- Use the SGPlot procedure to plot **Y** versus **X**.

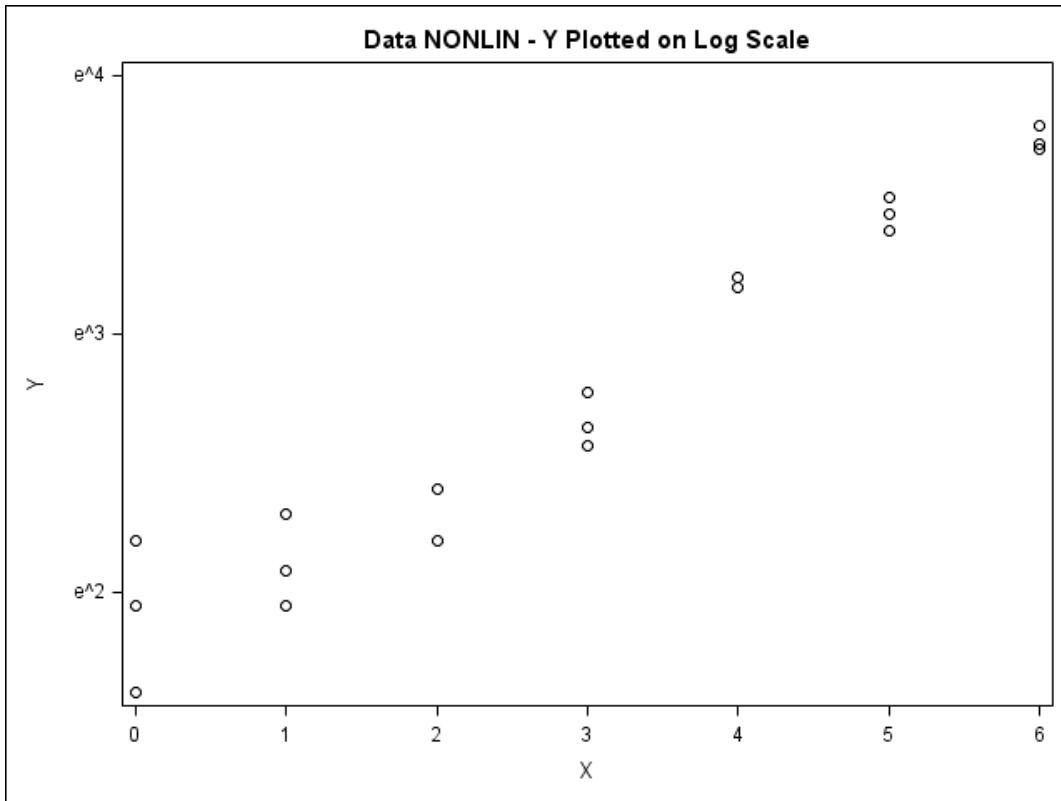
```
proc sgplot data=STAT2.nonlin;
  scatter y=y x=x;
title 'Data NONLIN - Y Plotted on Original Scale';
run; *ST20Ds02.sas;
```



There appears to be some curvature in the relationship between **X** and **Y**.

- Look up the XAXIS, X2AXIS, YAXIS, and Y2AXIS Statements in the online documentation for PROC SGPlot. Request that the Y axis be shown on the log scale (base e). How does this affect the graph and what does it suggest?

```
proc sgplot data=STAT2.nonlin;
  scatter y=y x=x;
  yaxis type=log logbase=e;
title 'Data NONLIN - Y Plotted on Log Scale';
run; *ST20Ds02.sas;
```



The plot looks nearly linear. This suggests modeling the log of Y as the outcome variable using OLS or modeling Y as the outcome variable in a gamma model with a log link.

- c. Use PROC REG to fit a regression line with **LogY** as the response variable and X as the predictor variable. Examine the residual plot and the fit plot to determine whether this is a good model.

```
proc reg data=STAT2.nonlin plots=all;
  model logy=x;
  title 'Regression Model on LogY';
run;                                *ST20Ds02.sas;
```

## Partial PROC REG Output

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.53078	8.53078	324.31	<.0001
Error	17	0.44718	0.02630		
Corrected Total	18	8.97796			

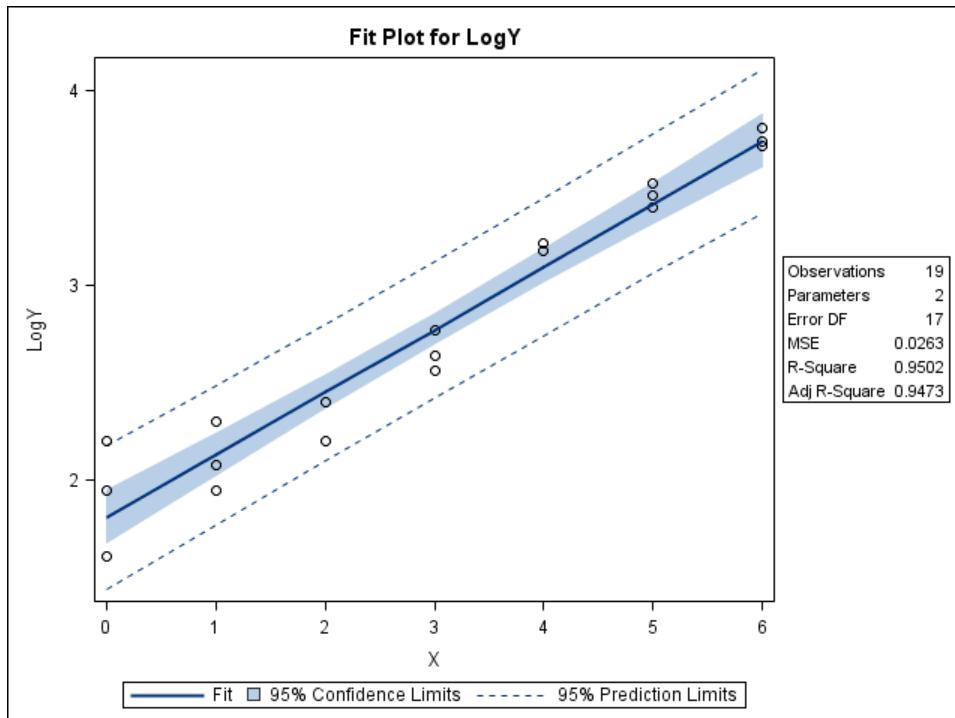
  

Root MSE	0.16219	R-Square	0.9502
Dependent Mean	2.77370	Adj R-Sq	0.9473
Coeff Var	5.84732		

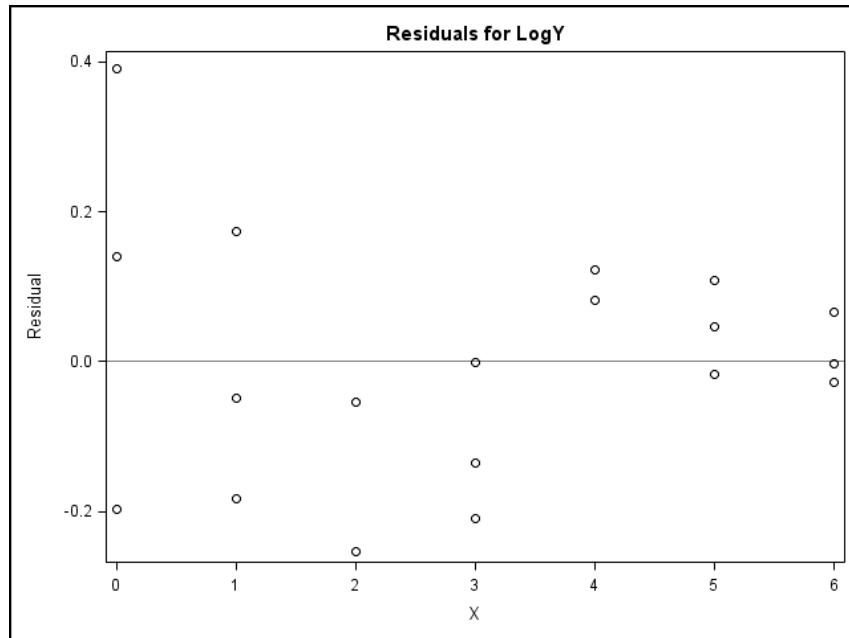
  

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	1.80607	0.06536	27.63	<.0001
X	1	0.32254	0.01791	18.01	<.0001

The model is  $\text{LogY} = 1.80607 + 0.3225 * X$ . To obtain the predicted value for Y, you must exponentiate the linear predictor with a low-bias adjustment factor.



The Fit Plot indicates that on the log scale, the predicted values fit the observed data well.



The regression line appears to fit the data fairly well. The residual plot shows some curvature, which might be problematic.

- d. Use PROC GENMOD to fit **Y** by **X**. Use the DIST=NORMAL and LINK=LOG options in the MODEL statement. Request the plots for the studentized residuals. Does this model appear to be a good fit? Compare this model to the previous model.

```
proc genmod data=STAT2.nonlin plots=stdreschi(xbeta);
  model y=x / dist=normal link=log obstats;
run; *ST20Ds02.sas;
```

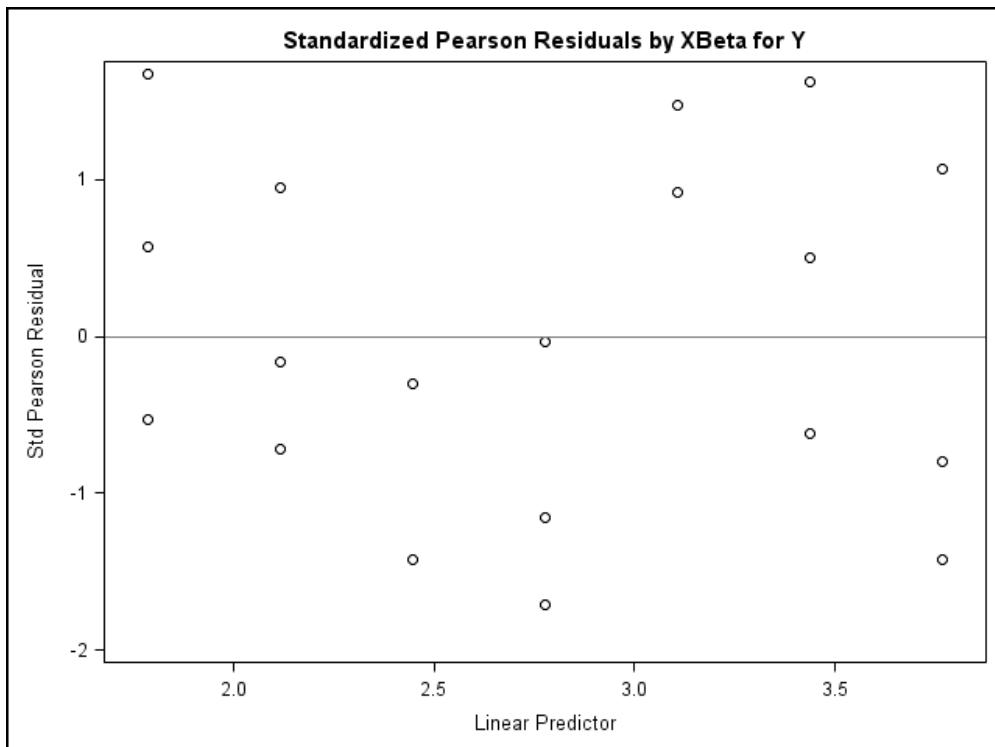
Partial PROC GENMOD Output

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	17	66.2512	3.8971
Scaled Deviance	17	19.0000	1.1176
Pearson Chi-Square	17	66.2512	3.8971
Scaled Pearson X2	17	19.0000	1.1176
Log Likelihood		-38.8255	
Full Log Likelihood		-38.8255	
AIC (smaller is better)		83.6510	
AICC (smaller is better)		85.2510	
BIC (smaller is better)		86.4843	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.7850	0.0727	1.6426	1.9275	602.99	<.0001
X	1	0.3305	0.0139	0.3033	0.3576	568.71	<.0001
Scale	1	1.8673	0.3029	1.3587	2.5663		

**Note:** The scale parameter was estimated by maximum likelihood.

The model is  $\log(E(Y)) = 1.785 + 0.3305 * X$ . The back-transformation results in  $E(Y) = e^{1.785 + 0.3305 * X}$ .



This model seems to fit your data slightly better than the previous model because the residual plot appears to be slightly better.



Using the transformed dependent variable to fit a linear regression model using PROC REG is not the same as using the link function (same type of transformation) in PROC GENMOD. In addition, you must apply a low-bias adjustment factor to the reverse transformation to obtain the low-bias estimates of the mean if you choose to transform the dependent variable. You need only to reverse-transform the link function to obtain the unbiased estimates of the mean if you use PROC GENMOD. This is automatically performed when you use the OBSTATS option in the MODEL statement in PROC GENMOD.

## Solutions to Polls and Quizzes

### D.01 Multiple Choice Poll – Correct Answer

You can use the following approach to determine the appropriate transformation of the dependent variable:

- a. theoretical knowledge or past studies
- b. how variances and means are related
- c. PROC TRANSREG
- d. trial and error
- e. c and d
- f. all of the above**

98

### D.02 Quiz – Correct Answer

Is the statement below true or false? Explain, your answer.

Although there are several methods to determine the appropriate transformation to stabilize variances and correct for nonnormality, the Box-Cox transformation in PROC TRANSREG provides an automated way of doing so.

**The answer is True. PROC TRANSREG does automate the selection of the appropriate transformation to stabilize variances and correct for nonnormality.**

103

### D.03 Poll – Correct Answer

When you transform the dependent variable to meet the model assumptions, you can always detransform the predicted value to obtain the unbiased estimates of the means on the original scale.

- True
- False

