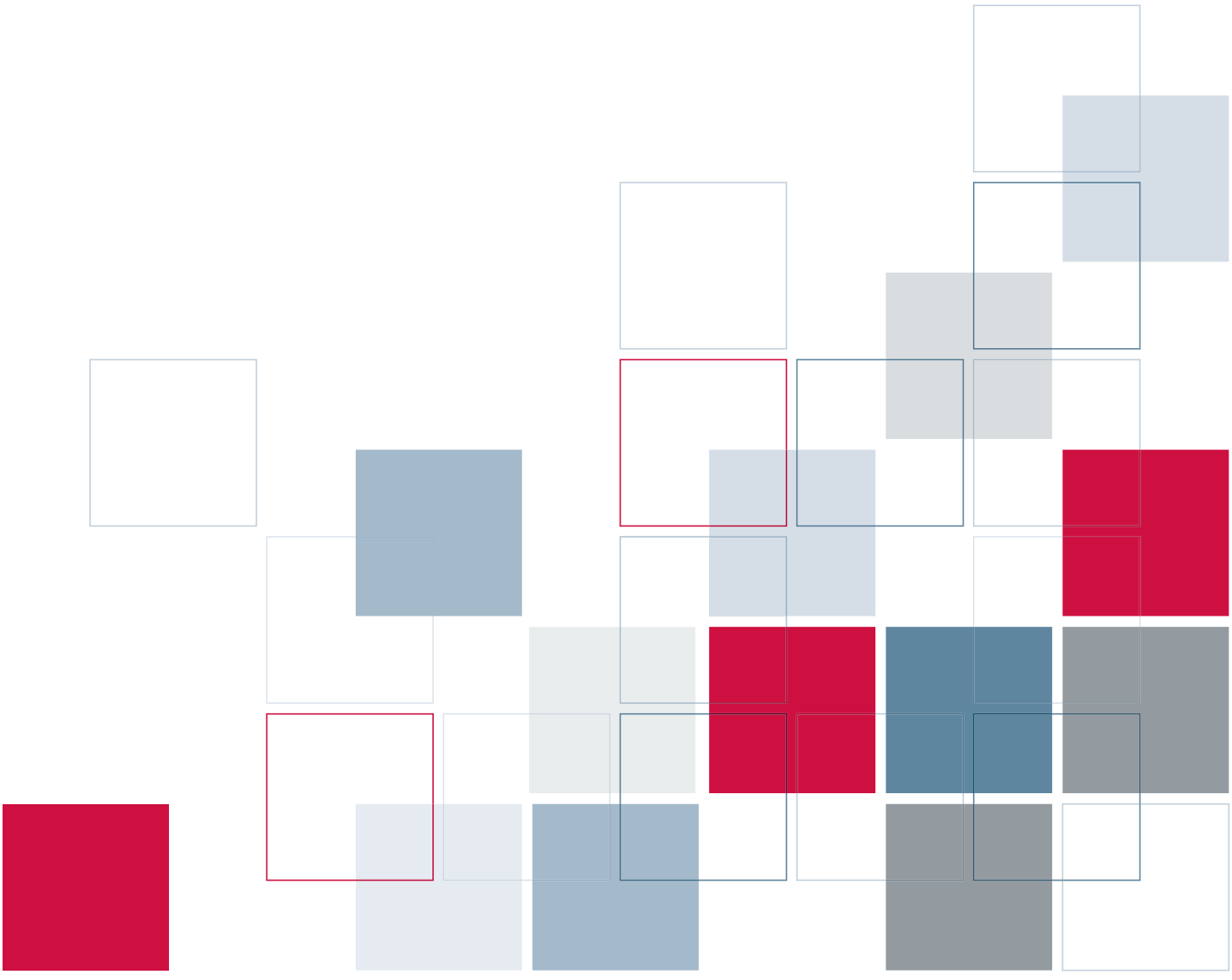
PASW® Bootstrapping 18



For more information about SPSS Inc. software products, please visit our Web site at *http://www.spss.com* or contact

SPSS Inc.
233 South Wacker Drive, 11th Floor Chicago, IL 60606-6412
Tel: (312) 651-3000 Fax: (312) 651-3668

Printed in the United States of America.

Preface

PASW Statistics 18 is a comprehensive system for analyzing data. The Bootstrapping optional add-on module provides the additional analytic techniques described in this manual. The Bootstrapping add-on module must be used with the PASW Statistics 18 Core system and is completely integrated into that system.

Installation

To install the Bootstrapping add-on module, run the License Authorization Wizard using the authorization code that you received from SPSS Inc. For more information, see the installation instructions supplied with the Bootstrapping add-on module.

Compatibility

PASW Statistics is designed to run on many computer systems. See the installation instructions that came with your system for specific information on minimum and recommended requirements.

Serial Numbers

Your serial number is your identification number with SPSS Inc. You will need this serial number when you contact SPSS Inc. for information regarding support, payment, or an upgraded system. The serial number was provided with your Core system.

Customer Service

If you have any questions concerning your shipment or account, contact your local office, listed on the Web site at *http://www.spss.com/worldwide*. Please have your serial number ready for identification.

Training Seminars

SPSS Inc. provides both public and onsite training seminars. All seminars feature hands-on workshops. Seminars will be offered in major cities on a regular basis. For more information on these seminars, contact your local office, listed on the Web site at *http://www.spss.com/worldwide*.

Technical Support

Technical Support services are available to maintenance customers. Customers may contact Technical Support for assistance in using PASW Statistics or for installation help for one of the supported hardware environments. To reach Technical Support, see the Web site at *http://www.spss.com*, or contact your local office, listed on the Web site at

http://www.spss.com/worldwide. Be prepared to identify yourself, your organization, and the serial number of your system.

Additional Publications

The *SPSS Statistics Statistical Procedures Companion*, by Marija Norušis, has been published by Prentice Hall. A new version of this book, updated for PASW Statistics 18, is planned. The *SPSS Statistics Advanced Statistical Procedures Companion*, also based on PASW Statistics 18, is forthcoming. The *SPSS Statistics Guide to Data Analysis* for PASW Statistics 18 is also in development. Announcements of publications available exclusively through Prentice Hall will be available on the Web site at *http://www.spss.com/estore* (select your home country, and then click Books).

Contents

Part I: User's Guide

Chapter
1

Introduction to Bootstrapping

When collecting data, you are often interested in the properties of the population from which you took the sample. You make inferences about these population parameters with estimates computed from the sample. For example, if the *Employee data.sav* dataset that is included with the product is a random sample from a larger population of employees, then the sample mean of $34,419.57 for *Current salary* is an estimate of the mean current salary for the population of employees. Moreover, this estimate has a standard error of $784.311 for a sample of size 474, and so a 95% confidence interval for the mean current salary in the population of employees is $32,878.40 to $35,960.73. But how reliable are these estimators? For certain "known" populations and well-behaved parameters, we know quite a bit about the properties of the sample estimates, and can be confident in these results. Bootstrapping seeks to uncover more information about the properties of estimators for "unknown" populations and ill-behaved parameters.

Figure 1-1
Making parametric inferences about the population mean

			Statistic	Std. Error
Current Salary	Mean		$34,419.57	$784.311
	95% Confidence Interval for Mean	Lower Bound	$32,878.40	
		Upper Bound	$35,960.73	
	Median		$28,875.00	

How Bootstrapping Works

At its simplest, for a dataset with a sample size of *N*, you take *B* "bootstrap" samples of size *N* with replacement from the original dataset and compute the estimator for each of these *B* bootstrap samples. These *B* bootstrap estimates are a sample of size *B* from which you can make inferences about the estimator. For example, if you take 1,000 bootstrap samples from the *Employee data.sav* dataset, then the bootstrap estimated standard error of $776.91 for the sample mean for *Current salary* is an alternative to the estimate of $784.311.

Additionally, bootstrapping provides a standard error and confidence interval for the median, for which parametric estimates are unavailable.

Figure 1-2
Making bootstrap inferences about the sample mean

			Statistic	Std. Error	Bootstrap[a]			
					Bias	Std. Error	95% Confidence Interval	
							Lower	Upper
Current Salary	Mean		$34,419.57	$784.311	$14.66	$776.91	$32,990.38	$36,026.06
	95% Confidence Interval for Mean	Lower Bound	$32,878.40					
		Upper Bound	$35,960.73					
	Median		$28,875.00		$-13.22	$536.63	$27,750.00	$29,850.00

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Support for Bootstrapping in the Product

Bootstrapping is incorporated as a subdialog in procedures that support bootstrapping. See Procedures That Support Bootstrapping for information on which procedures support bootstrapping.

When bootstrapping is requested in the dialogs, a new and separate `BOOTSTRAP` command is pasted in addition to the usual syntax generated by the dialog. The `BOOTSTRAP` command creates the bootstrap samples according to your specifications. Internally, the product treats these bootstrap samples like splits, even though they are not explicitly shown in the Data Editor. This means that, internally, there are effectively *B***N* cases, so the case counter in the status bar will count from 1 to *B***N* when processing the data during bootstrapping. The Output Management System (OMS) is used to collect the results of running the analysis on each "bootstrap split". These results are pooled, and the pooled bootstrap results displayed in the Viewer with the rest of the usual output generated by the procedure. In certain cases, you may see a reference to "bootstrap split 0"; this is the original dataset.

Chapter 2

Bootstrapping

Bootstrapping is a method for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient. It may also be used for constructing hypothesis tests. Bootstrapping is most useful as an alternative to parametric estimates when the assumptions of those methods are in doubt (as in the case of regression models with heteroscedastic residuals fit to small samples), or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors (as in the case of computing confidence intervals for the median, quartiles, and other percentiles).

Examples. A telecommunications firm loses about 27% of its customers to churn each month. In order to properly focus churn reduction efforts, management wants to know if this percentage varies across predefined customer groups. Using bootstrapping, you can determine whether a single rate of churn adequately describes the four major customer types. For more information, see the topic Using Bootstrapping to Obtain Confidence Intervals for Proportions in Chapter 3 on p. 10.

In a review of employee records, management is interested in the previous work experience of employees. Work experience is right skewed, which makes the mean a less desirable estimate of the "typical" previous work experience among employees than the median. However, parametric confidence intervals are not available for the median in the product. For more information, see the topic Using Bootstrapping to Obtain Confidence Intervals for Medians in Chapter 3 on p. 16.

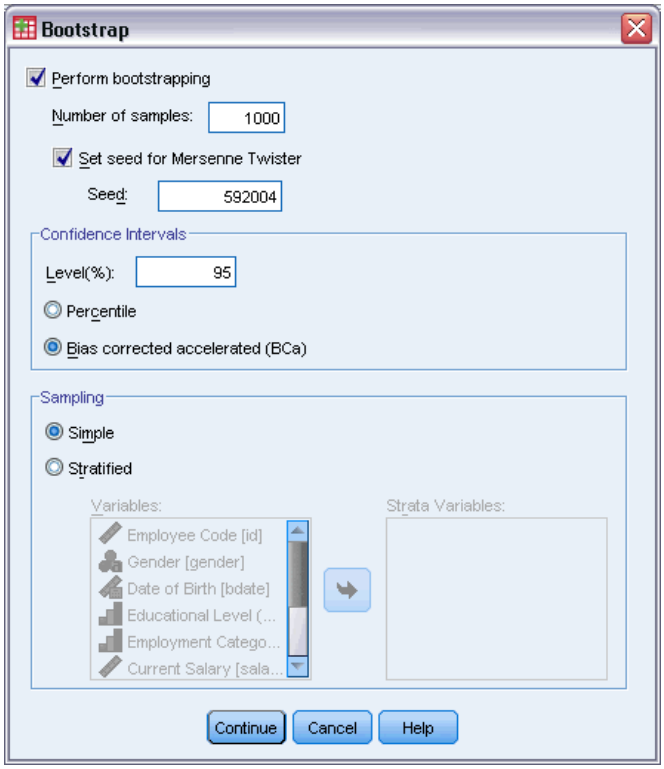
Management is also interested in determining what factors are associated with employee salary increases by fitting a linear model to the difference between current and starting salaries. When bootstrapping a linear model, you can use special resampling methods (residual and wild bootstrap) to obtain more accurate results. For more information, see the topic Using Bootstrapping to Choose Better Predictors in Chapter 3 on p. 19.

Many procedures support bootstrap sampling and pooling of results from analysis of bootstrap samples. Controls for specifying bootstrap analyses are integrated directly as a common subdialog in procedures that support bootstrapping. Settings on the bootstrap dialog persist across procedures so that if you run a Frequencies analysis with bootstrapping through the dialogs, bootstrapping will be turned on by default for other procedures that support it.

To Obtain a Bootstrap Analysis

- From the menus choose a procedure that supports bootstrapping and click Bootstrap.

Figure 2-1
Bootstrap dialog box



▶ Select Perform bootstrapping.

Optionally, you can control the following options:

Number of samples. For the percentile and BCa intervals produced, it is recommended to use at least 1000 bootstrap samples. Specify a positive integer.

Set seed for Mersenne Twister. Setting a seed allows you to replicate analyses. Using this control is similar to setting the Mersenne Twister as the active generator and specifying a fixed starting point on the Random Number Generators dialog, with the important difference that setting the seed in this dialog will preserve the current state of the random number generator and restore that state after the analysis is complete.

Confidence Intervals. Specify a confidence level greater than 50 and less than 100. Percentile intervals simply use the ordered bootstrap values corresponding to the desired confidence interval percentiles. For example, a 95% percentile confidence interval uses the 2.5th and 97.5th percentiles of the bootstrap values as the lower and upper bounds of the interval (interpolating the bootstrap values if necessary). Bias corrected and accelerated (BCa) intervals are adjusted intervals that are more accurate at the cost of requiring more time to compute.

Sampling. The Simple method is case resampling with replacement from the original dataset. The Stratified method is case resampling with replacement from the original dataset, *within* the strata defined by the cross-classification of strata variables. Stratified bootstrap sampling can be useful when units within strata are relatively homogeneous while units across strata are very different.

Procedures That Support Bootstrapping

The following procedures support bootstrapping.

Note:

- Bootstrapping does not work with multiply imputed datasets. If there is an *Imputation_* variable in the dataset, the Bootstrap dialog is disabled.
- Bootstrapping uses listwise deletion to determine the case basis; that is, cases with missing values on any of the analysis variables are deleted from the analysis, so when bootstrapping is in effect, listwise deletion is in effect even if the analysis procedure specifies another form of missing value handling.

Statistics Base Option

Frequencies

- The Statistics table supports bootstrap estimates for the mean, standard deviation, variance, median, skewness, kurtosis, and percentiles.
- The Frequencies table supports bootstrap estimates for percent.

Descriptives

- The Descriptive Statistics table supports bootstrap estimates for the mean, standard deviation, variance, skewness, and kurtosis.

Explore

- The Descriptives table supports bootstrap estimates for the mean, 5% Trimmed Mean, standard deviation, variance, median, skewness, kurtosis, and interquartile range.
- The M-Estimators table supports bootstrap estimates for Huber's M-Estimator, Tukey's Biweight, Hampel's M-Estimator, and Andrew's Wave.
- The Percentiles table supports bootstrap estimates for percentiles.

Crosstabs

- The Directional Measures table supports bootstrap estimates for Lambda, Goodman and Kruskal Tau, Uncertainty Coefficient, and Somers' d.
- The Symmetric Measures table supports bootstrap estimates for Phi, Cramer's V, Contingency Coefficient, Kendall's tau-b, Kendall's tau-c, Gamma, Spearman Correlation, and Pearson's R.
- The Risk Estimate table supports bootstrap estimates for the odds ratio.
- The Mantel-Haenszel Common Odds Ratio table supports bootstrap estimates and significance tests for ln(Estimate).

Means

- The Report table supports bootstrap estimates for the mean, median, grouped median, standard deviation, variance, kurtosis, skewness, harmonic mean, and geometric mean.

One-Sample T Test

- The Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Test table supports bootstrap estimates and significance tests for the mean difference.

Independent-Samples T Test

- The Group Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Test table supports bootstrap estimates and significance tests for the mean difference.

Paired-Samples T Test

- The Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates for correlations.
- The Test table supports bootstrap estimates for the mean.

One-Way ANOVA

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Multiple Comparisons table supports bootstrap estimates for the mean difference.
- The Contrast Tests table supports bootstrap estimates and significance tests for value of contrast.

GLM Univariate

- The Descriptive Statistics table supports bootstrap estimates for the Mean and standard deviation.
- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.
- The Contrast Results table supports bootstrap estimates and significance tests for the difference.
- The Estimated Marginal Means: Estimates table supports bootstrap estimates for the mean.
- The Estimated Marginal Means: Pairwise Comparisons table supports bootstrap estimates for the mean difference.
- The Post Hoc Tests: Multiple Comparisons table supports bootstrap estimates for the Mean Difference.

Bivariate Correlations

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates for correlations.

Note: If nonparametric correlations (Kendall's tau-b or Spearman) are requested in addition to Pearson correlations, the dialog pastes `CORRELATIONS` and `NONPAR CORR` commands with a separate `BOOTSTRAP` command for each. The same bootstrap samples will be used to compute all correlations.

Partial Correlations

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates for correlations.

Linear Regression

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates for correlations.
- The Model Summary table supports bootstrap estimates for Durbin-Watson.
- The Coefficients table supports bootstrap estimates and significance tests for the coefficient, B.
- The Correlation Coefficients table supports bootstrap estimates for correlations.
- The Residuals Statistics table supports bootstrap estimates for the mean and standard deviation.

Ordinal Regression

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

Discriminant Analysis

- The Standardized Canonical Discriminant Function Coefficients table supports bootstrap estimates for standardized coefficients.
- The Canonical Discriminant Function Coefficients table supports bootstrap estimates for unstandardized coefficients.
- The Classification Function Coefficients table supports bootstrap estimates for coefficients.

Advanced Statistics Option

GLM Multivariate

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

Linear Mixed Models

- The Estimates of Fixed Effects table supports bootstrap estimates and significance tests for the estimate.
- The Estimates of Covariance Parameters table supports bootstrap estimates and significance tests for the estimate.

Generalized Linear Models

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

Cox Regression

- The Variables in the Equation table supports bootstrap estimates and significance tests for the coefficient, B.

Regression Option

Binary Logistic Regression

- The Variables in the Equation table supports bootstrap estimates and significance tests for the coefficient, B.

Multinomial Logistic Regression

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

BOOTSTRAP Command Additional Features

The command syntax language also allows you to:

- Perform residual and wild bootstrap sampling (`SAMPLING` subcommand)

See the *Command Syntax Reference* for complete syntax information.

Part II: Examples

Chapter 3

Bootstrapping

Bootstrapping is a method for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient. It may also be used for constructing hypothesis tests. Bootstrapping is most useful as an alternative to parametric estimates when the assumptions of those methods are in doubt (as in the case of regression models with heteroscedastic residuals fit to small samples), or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors (as in the case of computing confidence intervals for the median, quartiles, and other percentiles).

Using Bootstrapping to Obtain Confidence Intervals for Proportions

A telecommunications firm loses about 27% of its customers to churn each month. In order to properly focus churn reduction efforts, management wants to know if this percentage varies across predefined customer groups.

This information is collected in *telco.sav*. For more information, see the topic Sample Files in Appendix A on p. 30. Use bootstrapping to determine whether a single rate of churn adequately describes the four major customer types.

Note: This example uses the Frequencies procedure and requires the Statistics Base option.

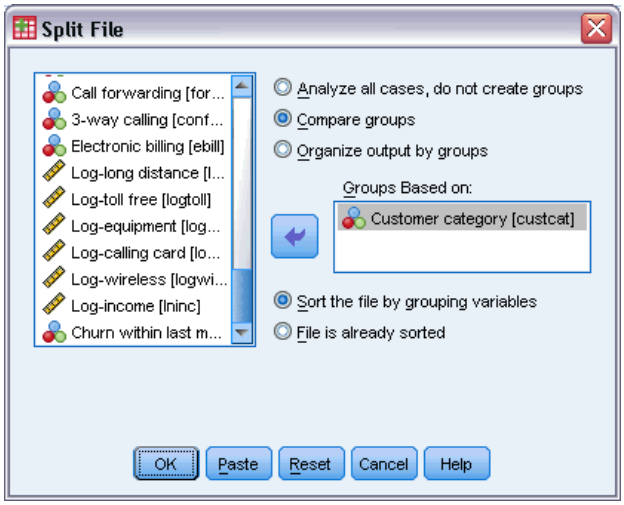
Preparing the Data

You must first split the file by *Customer category*.

▶ To split the file, from the Data Editor menus choose:

Data
  Split File...

Figure 3-1
Split File dialog box



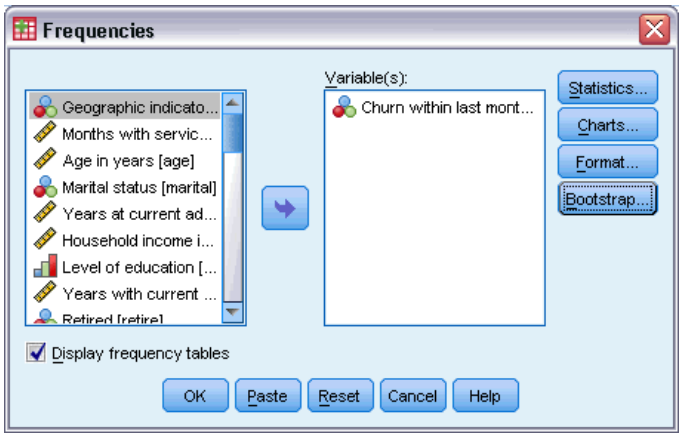
- Select Compare groups.
- Select *Customer category* as the variable on which to base groups.
- Click OK.

Running the Analysis

- To obtain bootstrap confidence intervals for proportions, from the menus choose:

  Analyze
  Descriptive Statistics
  Frequencies...

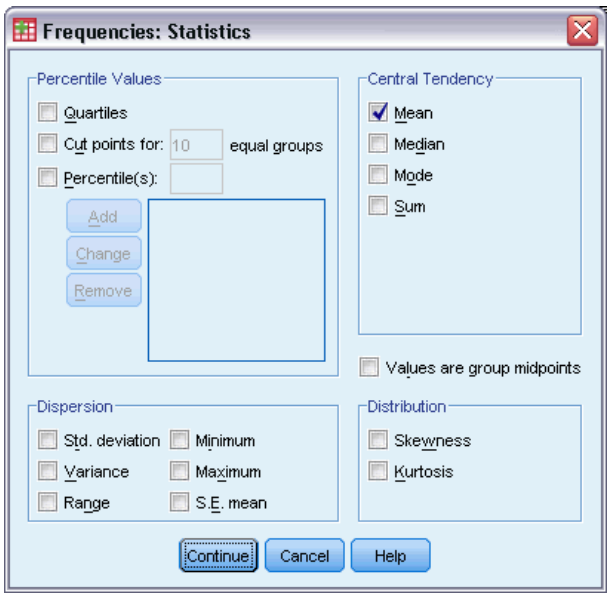
Figure 3-2
Frequencies main dialog



- Select *Churn within last month [churn]* as a variable in the analysis.

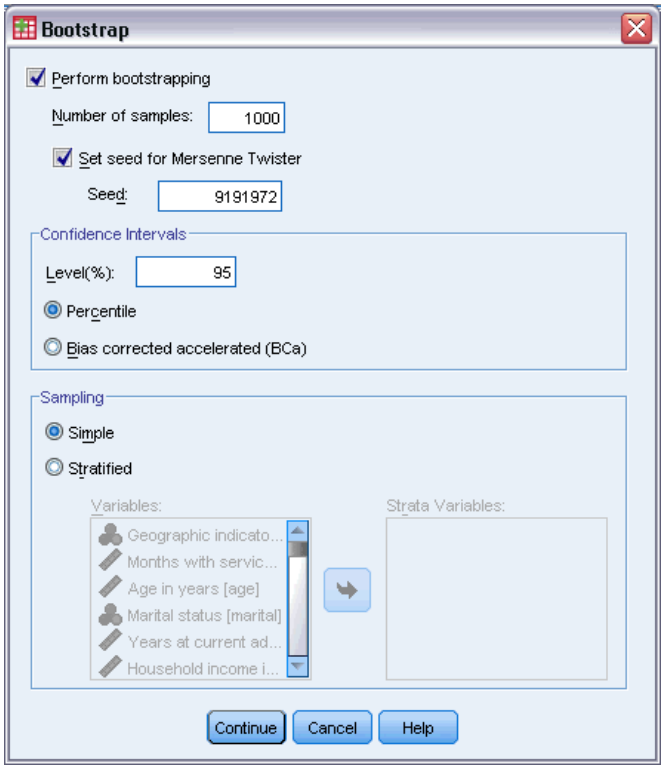
- Click Statistics.

Figure 3-3
Statistics dialog box



- Select Mean in the Central Tendency group.
- Click Continue.
- Click Bootstrap in the Frequencies dialog box.

Figure 3-4
Bootstrap dialog box



- Select Perform bootstrapping.
- To replicate the results in this example exactly, select Set seed for Mersenne Twister and type 9191972 as the seed.
- Click Continue.
- Click OK in the Frequencies dialog box.

These selections generate the following command syntax:

```
SORT CASES  BY custcat.
SPLIT FILE LAYERED BY custcat.
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES INPUT=churn
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
FREQUENCIES VARIABLES=churn
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
RESTORE.
```

- The `SORT CASES` and `SPLIT FILE` commands split the file on the variable *custcat*.

- The `PRESERVE` and `RESTORE` commands "remember" the current state of the random number generator and restore the system to that state after bootstrapping is over.
- The `SET` command sets the random number generator to the Mersenne Twister and the index to 9191972, so that the bootstrapping results can be replicated exactly. The `SHOW` command displays the index in the output for reference.
- The `BOOTSTRAP` command requests 1,000 bootstrap samples using simple resampling.
- The variable *churn* is used to determine the case basis for resampling. Records with missing values on this variable are deleted from the analysis.
- The `FREQUENCIES` procedure following `BOOTSTRAP` is run on each of the bootstrap samples.
- The `STATISTICS` subcommand produces the mean for variable *churn* on the original data. Additionally, pooled statistics are produced for the mean and the percentages in the frequency table.

Bootstrap Specifications

Figure 3-5
Bootstrap specifications

Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile

The bootstrap specifications table contains the settings used during resampling, and is a useful reference for checking whether the analysis you intended was performed.

Statistics

Figure 3-6
Statistics table with bootstrap confidence interval for proportion

Churn within last month

Customer category			Statistic	Bootstrap[a]			
				Bias	Std. Error	95% Confidence Interval	
						Lower	Upper
Basic service	N	Valid	266	0	0	266	266
		Missing	0	0	0	0	0
	Mean		.31	.00	.03	.26	.37
E-service	N	Valid	217	0	0	217	217
		Missing	0	0	0	0	0
	Mean		.27	.00	.03	.21	.34
Plus service	N	Valid	281	0	0	281	281
		Missing	0	0	0	0	0
	Mean		.16	.00	.02	.12	.20
Total service	N	Valid	236	0	0	236	236
		Missing	0	0	0	0	0
	Mean		.37	.00	.03	.31	.44

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The statistics table shows, for each level of *Customer category*, the mean value for *Churn within last month*. Since *Churn within last month* only takes values 0 and 1, with 1 signifying a customer who churned, the mean is equal to the proportion of churners. The Statistic column shows the values usually produced by Frequencies, using the original dataset. The Bootstrap columns are produced by the bootstrapping algorithms.

- Bias is the difference between the average value of this statistic across the bootstrap samples and the value in the Statistic column. In this case, the mean value of *Churn within last month* is computed for all 1000 bootstrap samples, and the average of these means is then computed.
- Std. Error is the standard error of the mean value of *Churn within last month* across the 1000 bootstrap samples.
- The lower bound of the 95% bootstrap confidence interval is an interpolation of the 25th and 26th mean values of *Churn within last month*, if the 1000 bootstrap samples are sorted in ascending order. The upper bound is an interpolation of the 975th and 976th mean values.

The results in the table suggest that the rate of churn differs across customer types. In particular, the confidence interval for *Plus service* customers does not overlap with any other, suggesting these customers are, on average, less likely to leave.

When working with categorical variables with only two values, these confidence intervals are alternatives to those produced by the One-Sample Nonparametric Tests procedure or the One-Sample T Test procedure.

Frequency Table

Figure 3-7
Frequency table with bootstrap confidence interval for proportion

Customer category			Frequency	Percent	Valid Percent	Cumulative Percent	Bootstrap for Percent[a]			
							Bias	Std. Error	95% Confidence Interval	
									Lower	Upper
Basic service	Valid	No	183	68.8	68.8	68.8	.0	2.8	63.2	74.4
		Yes	83	31.2	31.2	100.0	.0	2.8	25.6	36.8
		Total	266	100.0	100.0		.0	.0	100.0	100.0
E-service	Valid	No	158	72.8	72.8	72.8	.1	3.1	66.4	78.8
		Yes	59	27.2	27.2	100.0	-.1	3.1	21.2	33.6
		Total	217	100.0	100.0		.0	.0	100.0	100.0
Plus service	Valid	No	237	84.3	84.3	84.3	.0	2.1	80.1	88.3
		Yes	44	15.7	15.7	100.0	.0	2.1	11.7	19.9
		Total	281	100.0	100.0		.0	.0	100.0	100.0
Total service	Valid	No	148	62.7	62.7	62.7	.0	3.2	56.4	69.1
		Yes	88	37.3	37.3	100.0	.0	3.2	30.9	43.6
		Total	236	100.0	100.0		.0	.0	100.0	100.0

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The Frequency table shows confidence intervals for the percentages (proportion × 100%) for each category, and are thus available for all categorical variables. Comparable confidence intervals are not available elsewhere in the product.

Using Bootstrapping to Obtain Confidence Intervals for Medians

In a review of employee records, management is interested in the previous work experience of employees. Work experience is right skewed, which makes the mean a less desirable estimate of the "typical" previous work experience among employees than the median. However, without bootstrapping, confidence intervals for the median are not generally available in statistical procedures in the product.

This information is collected in *Employee data.sav*. For more information, see the topic Sample Files in Appendix A on p. 30. Use Bootstrapping to obtain confidence intervals for the median.

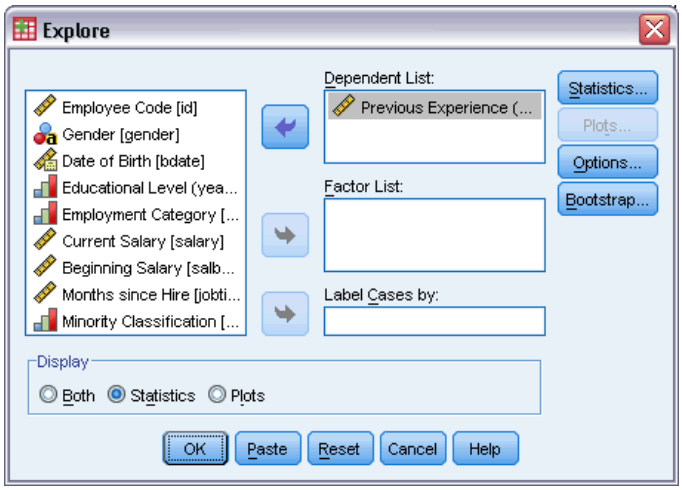
Note: this example uses the Explore procedure, and requires the Statistics Base option.

Running the Analysis

- To obtain bootstrap confidence intervals for the median, from the menus choose:

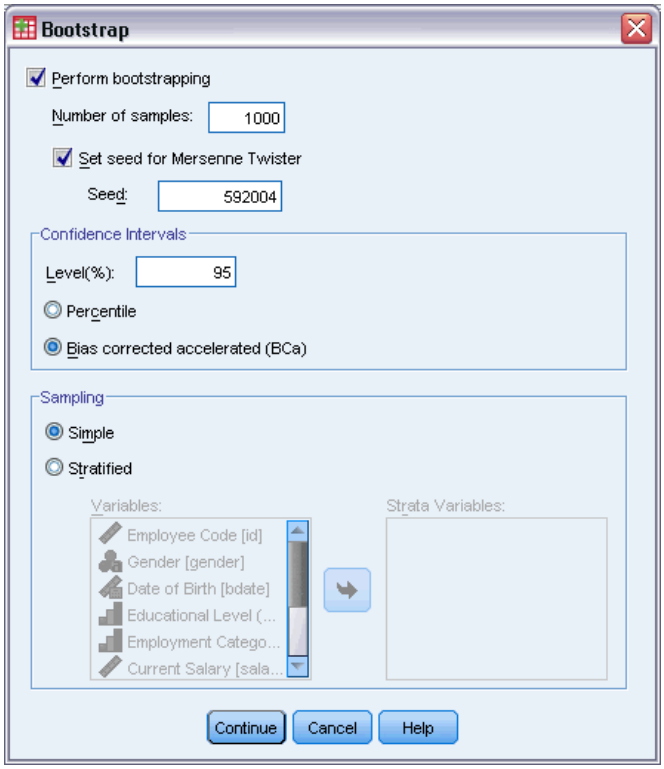
Analyze
  Descriptive Statistics
    Explore...

Figure 3-8
Explore main dialog



- Select *Previous Experience (months) [prevexp]* as a dependent variable.
- Select Statistics in the Display group.
- Click Bootstrap.

Figure 3-9
Bootstrap dialog box



- Select Perform bootstrapping.
- To replicate the results in this example exactly, select Set seed for Mersenne Twister and type 592004 as the seed.
- To obtain more accurate intervals (at the cost of more processing time), select Bias corrected accelerated (BCa).
- Click Continue.
- Click OK in the Explore dialog box.

These selections generate the following command syntax:

```
PRESERVE.
SET RNG=MT MTINDEX=592004.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=prevexp
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
EXAMINE VARIABLES=prevexp
  /PLOT NONE
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
RESTORE.
```

- The PRESERVE and RESTORE commands "remember" the current state of the random number generator and restore the system to that state after bootstrapping is over.
- The SET command sets the random number generator to the Mersenne Twister and the index to 592004, so that the bootstrapping results can be replicated exactly. The SHOW command displays the index in the output for reference.
- The BOOTSTRAP command requests 1000 bootstrap samples using simple resampling.
- The VARIABLES subcommand specifies that the variable *prevexp* is used to determine the case basis for resampling. Records with missing values on this variable are deleted from the analysis.
- The CRITERIA subcommand, in addition to requesting the number of bootstrap samples, requests bias-corrected and accelerated bootstrap confidence intervals instead of the default percentile intervals.
- The EXAMINE procedure following BOOTSTRAP is run on each of the bootstrap samples.
- The PLOT subcommand turns off plot output.
- All other options are set to their default values.

Descriptives

Figure 3-10
Descriptives table with bootstrap confidence intervals

					Bootstrap[a]			
							BCa 95% Confidence Interval	
			Statistic	Std. Error	Bias	Std. Error	Lower	Upper
Previous Experience (months)	Mean		95.86	4.804	-.01	4.86	86.39	105.20
	95% Confidence Interval for Mean	Lower Bound	86.42					
		Upper Bound	105.30					
	5% Trimmed Mean		84.64		.02	4.94	75.38	94.21
	Median		55.00		-.11	3.66	50.00	60.00
	Variance		10938.281		18.783	977.081	8954.509	13057.229
	Std. Deviation		104.586		-.015	4.689	94.644	114.245
	Minimum		0					
	Maximum		476					
	Range		476					
	Interquartile Range		121		-1	10	103	137
	Skewness		1.510	.112	.006	.110	1.284	1.768
	Kurtosis		1.696	.224	.040	.463	.823	2.876

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The descriptives table contains a number of statistics and bootstrap confidence intervals for those statistics. The bootstrap confidence interval for the mean (86.39, 105.20) is similar to the parametric confidence interval (86.42, 105.30) and suggests that the "typical" employee has roughly 7-9 years of previous experience. However, *Previous Experience (months)* has a skewed distribution, which makes the mean a less desirable indicator of "typical" current salary than the median. The bootstrap confidence interval for the median (50.00, 60.00) is both narrower and lower in value than the confidence interval for the mean, and suggests that the "typical" employee has roughly 4-5 years of previous experience. Using bootstrapping has made it possible to obtain a range of values that better represent typical previous experience.

Using Bootstrapping to Choose Better Predictors

In a review of employee records, management is interested in determining what factors are associated with employee salary increases by fitting a linear model to the difference between current and starting salaries. When bootstrapping a linear model, you can use special resampling methods (residual and wild bootstrap) to obtain more accurate results.

This information is collected in *Employee data.sav*. For more information, see the topic Sample Files in Appendix A on p. 30.

Note: this example uses the GLM Univariate procedure, and requires the Statistics Base option.

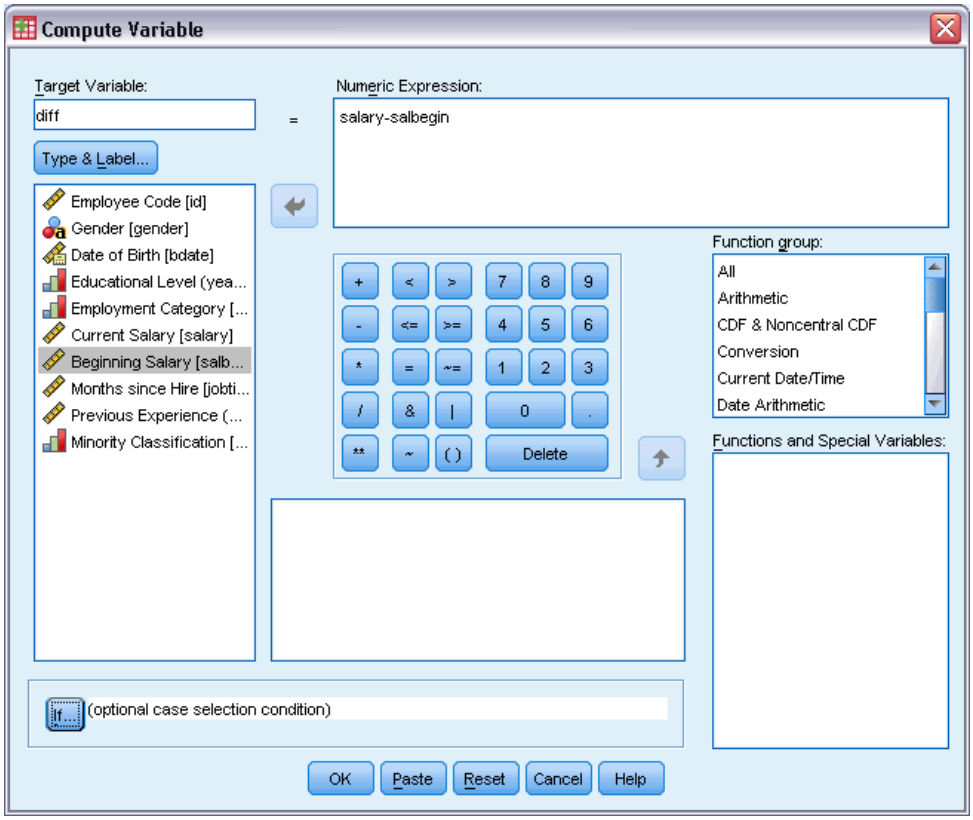
Preparing the Data

You must first compute the difference between Current salary and Beginning salary.

- From the menus choose:

Transform
  Compute Variable...

Figure 3-11
Compute Variable dialog box



- Type diff as the target variable..
- Type salary-salbegin as the numeric expression.
- Click OK.

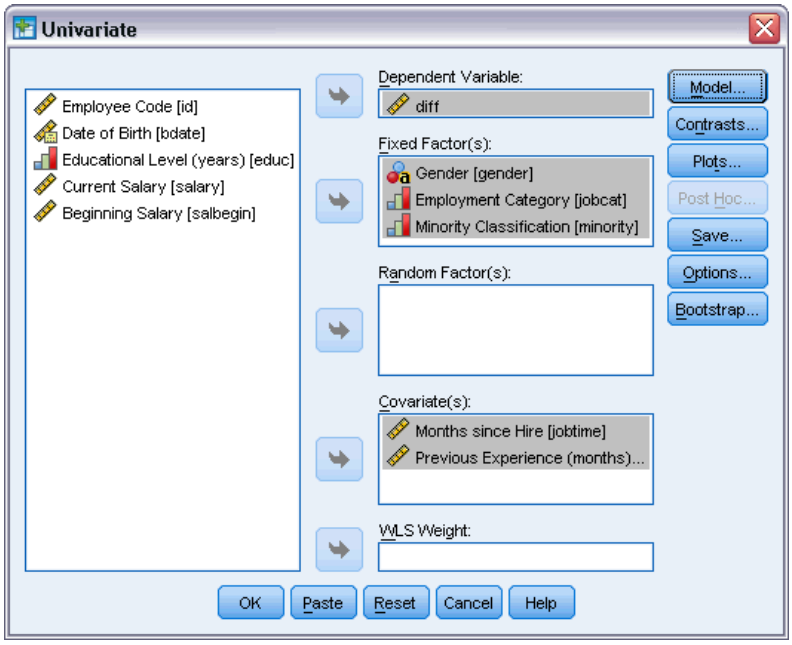
Running the Analysis

To run GLM Univariate with wild residual bootstrapping, you first need to create residuals.

- From the menus choose:

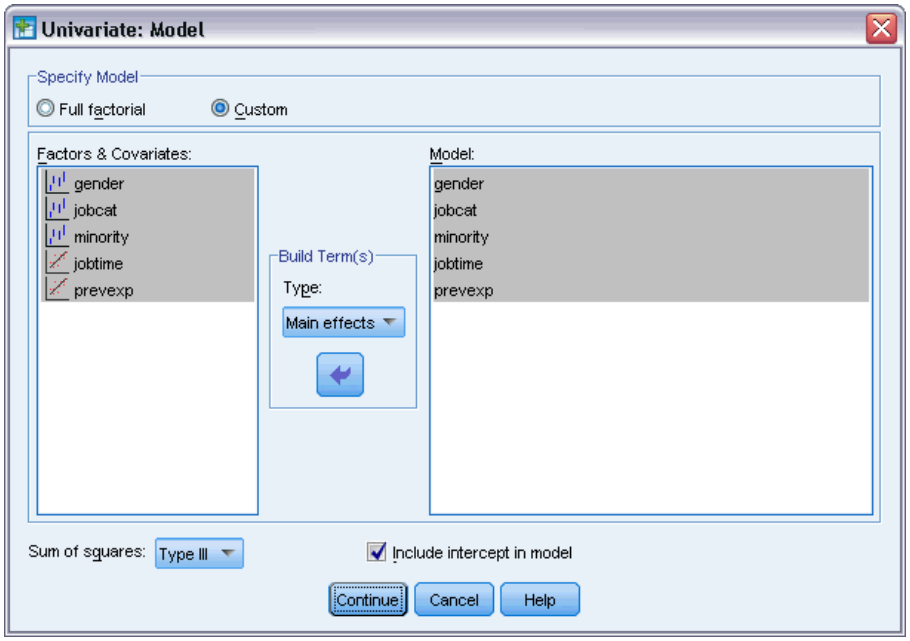
Analyze
  General Linear Model
    Univariate...

Figure 3-12
GLM Univariate main dialog



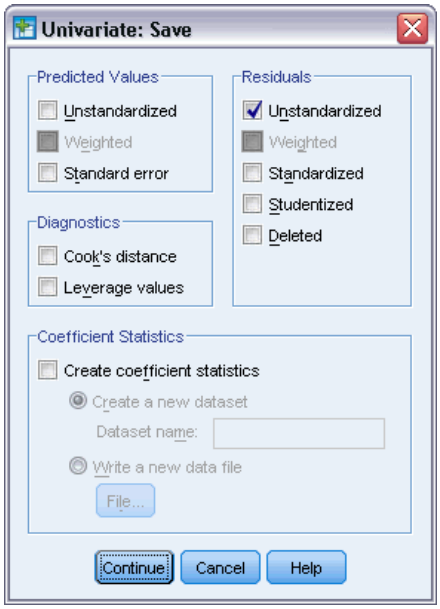
- Select *diff* as the dependent variable.
- Select *Gender [gender]*, *Employment Category [jobcat]*, and *Minority Classification [minority]* as fixed factors.
- Select *Months since Hire [jobtime]* and *Previous Experience (months) [prevexp]* as covariates.
- Click Model.

Figure 3-13
Model dialog box



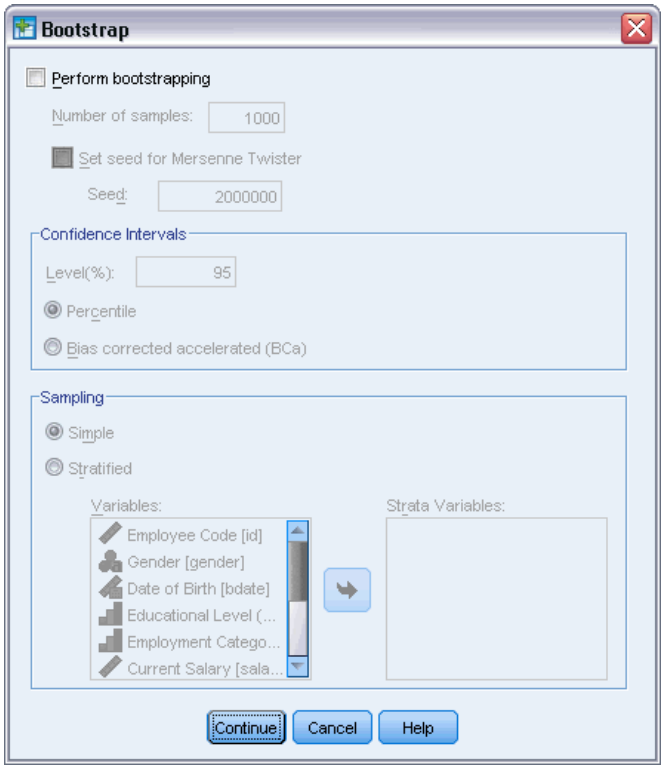
- Select Custom and select Main effects from the Build Terms dropdown.
- Select *gender* through *prevexp* as model terms.
- Click Continue.
- Click Save in the GLM Univariate dialog box.

Figure 3-14
Save dialog box



- Select Unstandardized in the Residuals group.
- Click Continue.
- Click Bootstrap in the GLM Univariate dialog box.

Figure 3-15
Bootstrap dialog box

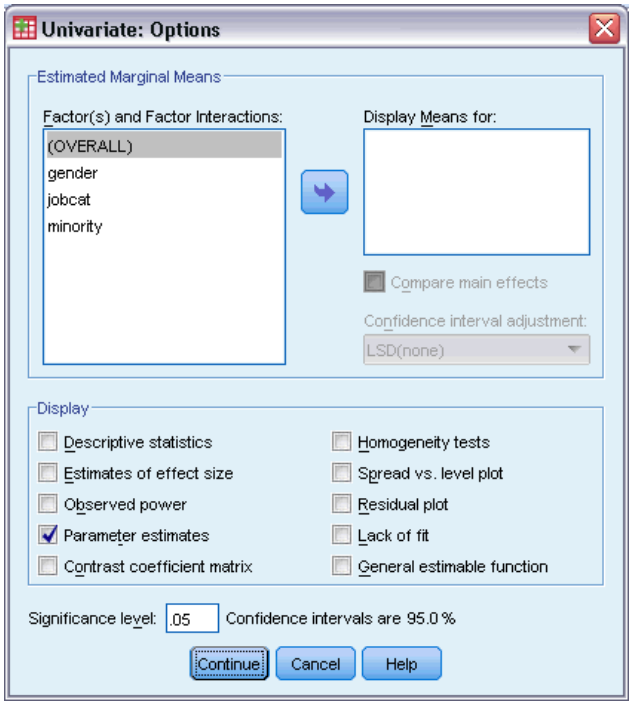


The bootstrap settings persist across dialogs that support bootstrapping. Saving new variables to the dataset is not supported while bootstrapping is in effect, so you need to make certain it is turned off.

- If needed deselect Perform bootstrapping.
- Click OK in the GLM Univariate dialog box. The dataset now contains a new variable, *RES_1*, which contains the unstandardized residuals from this model.
- Recall the GLM Univariate dialog and click Save.

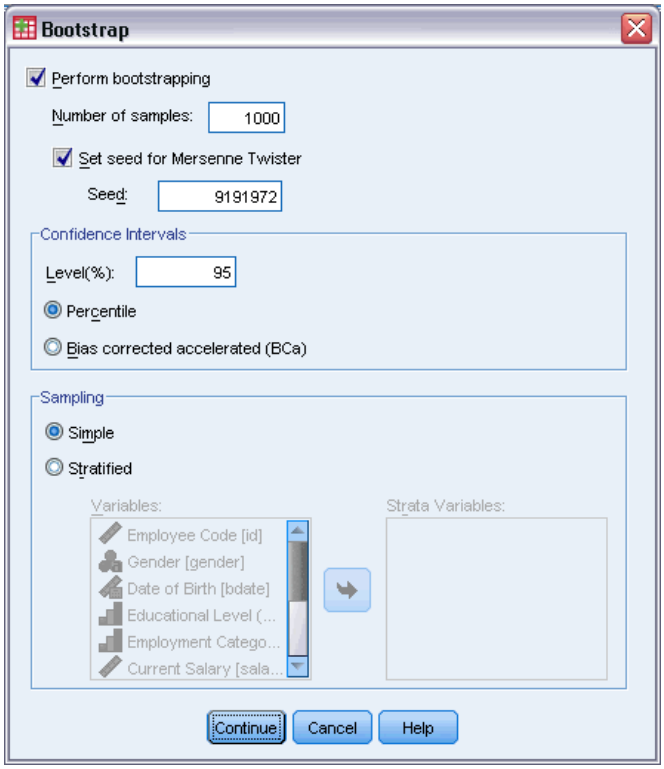
- Deselect Unstandardized, then click Continue and click Options in the GLM Univariate dialog box.

Figure 3-16
Options dialog box



- Select Parameter estimates in the Display group.
- Click Continue.
- Click Bootstrap in the GLM Univariate dialog box.

Figure 3-17
Bootstrap dialog box



- Select Perform bootstrapping.
- To replicate the results in this example exactly, select Set seed for Mersenne Twister and type 9191972 as the seed.
- There are no options for performing wild bootstrapping through the dialogs, so click Continue, then click Paste in the GLM Univariate dialog box.

These selections generate the following command syntax:

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
 /SAMPLING METHOD=SIMPLE
 /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
 /METHOD=SSTYPE(3)
 /INTERCEPT=INCLUDE
 /PRINT=PARAMETER
 /CRITERIA=ALPHA(.05)
 /DESIGN=gender jobcat minority jobtime prevexp.
RESTORE.
```

In order to perform wild bootstrap sampling, edit the `METHOD` keyword of the `SAMPLING` subcommand to read `METHOD=WILD(RESIDUALS=RES_1)`.

The "final" set of command syntax looks like the following:

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
 /SAMPLING METHOD=WILD(RESIDUALS=RES_1)
 /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
 /METHOD=SSTYPE(3)
 /INTERCEPT=INCLUDE
 /PRINT=PARAMETER
 /CRITERIA=ALPHA(.05)
 /DESIGN=gender jobcat minority jobtime prevexp.
RESTORE.
```

- The `PRESERVE` and `RESTORE` commands "remember" the current state of the random number generator and restore the system to that state after bootstrapping is over.
- The `SET` command sets the random number generator to the Mersenne Twister and the index to 9191972, so that the bootstrapping results can be replicated exactly. The `SHOW` command displays the index in the output for reference.
- The `BOOTSTRAP` command requests 1000 bootstrap samples using wild sampling and *RES_1* as the variable containing the residuals.
- The `VARIABLES` subcommand specifies that the *diff* is the target variable in the linear model; it and the variables *gender*, *jobcat*, *minority*, *jobtime*, and *prevexp* are used to determine the case basis for resampling. Records with missing values on these variables are deleted from the analysis.
- The `CRITERIA` subcommand, in addition to requesting the number of bootstrap samples, requests bias-corrected and accelerated bootstrap confidence intervals instead of the default percentile intervals.
- The `UNIANOVA` procedure following `BOOTSTRAP` is run on each of the bootstrap samples and produces parameter estimates for the original data. Additionally, pooled statistics are produced for the model coefficients.

Parameter Estimates

Figure 3-18
Parameter estimates

Dependent Variable:diff

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	22789.014	2920.700	7.803	.000	17049.673	28528.355
[gender=f]	-4085.253	726.416	-5.624	.000	-5512.701	-2657.804
[gender=m]	0[a]	.	.	.	.	.
[jobcat=1]	-17717.706	939.798	-18.853	.000	-19564.463	-15870.949
[jobcat=2]	-13101.918	1780.683	-7.358	.000	-16601.061	-9602.776
[jobcat=3]	0[a]	.	.	.	.	.
[minority=0]	1332.363	819.349	1.626	.105	-277.705	2942.431
[minority=1]	0[a]	.	.	.	.	.
jobtime	145.539	32.586	4.466	.000	81.505	209.572
prevexp	-21.423	3.575	-5.993	.000	-28.447	-14.398

a.This parameter is set to zero because it is redundant.

The Parameter Estimates table shows the usual, non-bootstrapped, parameter estimates for the model terms. The significance value of 0.105 for *[minority=0]* is greater than 0.05, suggesting that *Minority Classification* has no effect on salary increases.

Figure 3-19
Bootstrap parameter estimates

Dependent Variable:diff

Parameter	B	Bootstrap[a]				
		Bias	Std. Error	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
Intercept	22789.014	-141.666	3158.690	.001	16109.048	28431.372
[gender=f]	-4085.253	-43.965	598.687	.001	-5258.788	-2979.228
[gender=m]	0	0	0		0	0
[jobcat=1]	-17717.706	29.668	1448.572	.001	-20454.119	-14869.484
[jobcat=2]	-13101.918	11.820	1724.019	.001	-16517.418	-9681.869
[jobcat=3]	0	0	0		0	0
[minority=0]	1332.363	18.269	635.451	.006	174.525	2621.938
[minority=1]	0	0	0		0	0
jobtime	145.539	1.275	33.908	.001	85.323	213.415
prevexp	-21.423	.085	2.681	.001	-26.550	-15.869

a.Unless otherwise noted, bootstrap results are based on 1000 wild bootstrap samples

Now look at the Bootstrap for Parameter Estimates table. In the Std. Error column, you see that the parametric standard errors for some coefficients, like the intercept, are too small compared to the bootstrap estimates, and thus the confidence intervals are wider. For some coefficients, like *[minority=0]*, the parametric standard errors were too large, while the significance value of 0.006 reported by the bootstrap results, which is less than 0.05, shows that the observed difference in salary increases between employees who are and are not minorities is not due to chance. Management now knows this difference is worth investigating further to determine possible causes.

Recommended Readings

See the following texts for more information on bootstrapping:

Davison, A. C., and D. V. Hinkley. 2006. *Bootstrap Methods and their Application.* : Cambridge University Press.

Shao, J., and D. Tu. 1995. *The Jackknife and Bootstrap.* New York: Springer.

Appendix

A

Sample Files

The sample files installed with the product can be found in the *Samples* subdirectory of the installation directory. There is a separate folder within the Samples subdirectory for each of the following languages: English, French, German, Italian, Japanese, Korean, Polish, Russian, Simplified Chinese, Spanish, and Traditional Chinese.

Not all sample files are available in all languages. If a sample file is not available in a language, that language folder contains an English version of the sample file.

Descriptions

Following are brief descriptions of the sample files used in various examples throughout the documentation.

- **accidents.sav.** This is a hypothetical data file that concerns an insurance company that is studying age and gender risk factors for automobile accidents in a given region. Each case corresponds to a cross-classification of age category and gender.
- **adl.sav.** This is a hypothetical data file that concerns efforts to determine the benefits of a proposed type of therapy for stroke patients. Physicians randomly assigned female stroke patients to one of two groups. The first received the standard physical therapy, and the second received an additional emotional therapy. Three months following the treatments, each patient's abilities to perform common activities of daily life were scored as ordinal variables.
- **advert.sav.** This is a hypothetical data file that concerns a retailer's efforts to examine the relationship between money spent on advertising and the resulting sales. To this end, they have collected past sales figures and the associated advertising costs..
- **aflatoxin.sav.** This is a hypothetical data file that concerns the testing of corn crops for aflatoxin, a poison whose concentration varies widely between and within crop yields. A grain processor has received 16 samples from each of 8 crop yields and measured the alfatoxin levels in parts per billion (PPB).
- **aflatoxin20.sav.** This data file contains the aflatoxin measurements from each of the 16 samples from yields 4 and 8 from the *aflatoxin.sav* data file.
- **anorectic.sav.** While working toward a standardized symptomatology of anorectic/bulimic behavior, researchers (Van der Ham, Meulman, Van Strien, and Van Engeland, 1997) made a study of 55 adolescents with known eating disorders. Each patient was seen four times over four years, for a total of 220 observations. At each observation, the patients were scored for each of 16 symptoms. Symptom scores are missing for patient 71 at time 2, patient 76 at time 2, and patient 47 at time 3, leaving 217 valid observations.

- **autoaccidents.sav.** This is a hypothetical data file that concerns the efforts of an insurance analyst to model the number of automobile accidents per driver while also accounting for driver age and gender. Each case represents a separate driver and records the driver's gender, age in years, and number of automobile accidents in the last five years.
- **band.sav.** This data file contains hypothetical weekly sales figures of music CDs for a band. Data for three possible predictor variables are also included.
- **bankloan.sav.** This is a hypothetical data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks.
- **bankloan_binning.sav.** This is a hypothetical data file containing financial and demographic information on 5,000 past customers.
- **behavior.sav.** In a classic example (Price and Bouffard, 1974), 52 students were asked to rate the combinations of 15 situations and 15 behaviors on a 10-point scale ranging from 0="extremely appropriate" to 9="extremely inappropriate." Averaged over individuals, the values are taken as dissimilarities.
- **behavior_ini.sav.** This data file contains an initial configuration for a two-dimensional solution for *behavior.sav*.
- **brakes.sav.** This is a hypothetical data file that concerns quality control at a factory that produces disc brakes for high-performance automobiles. The data file contains diameter measurements of 16 discs from each of 8 production machines. The target diameter for the brakes is 322 millimeters.
- **breakfast.sav.** In a classic study (Green and Rao, 1972), 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference with 1="most preferred" to 15="least preferred." Their preferences were recorded under six different scenarios, from "Overall preference" to "Snack, with beverage only."
- **breakfast-overall.sav.** This data file contains the breakfast item preferences for the first scenario, "Overall preference," only.
- **broadband_1.sav.** This is a hypothetical data file containing the number of subscribers, by region, to a national broadband service. The data file contains monthly subscriber numbers for 85 regions over a four-year period.
- **broadband_2.sav.** This data file is identical to *broadband_1.sav* but contains data for three additional months.
- **car_insurance_claims.sav.** A dataset presented and analyzed elsewhere (McCullagh and Nelder, 1989) concerns damage claims for cars. The average claim amount can be modeled as having a gamma distribution, using an inverse link function to relate the mean of the dependent variable to a linear combination of the policyholder age, vehicle type, and vehicle age. The number of claims filed can be used as a scaling weight.
- **car_sales.sav.** This data file contains hypothetical sales estimates, list prices, and physical specifications for various makes and models of vehicles. The list prices and physical specifications were obtained alternately from *edmunds.com* and manufacturer sites.
- **car_sales_uprepared.sav.** This is a modified version of *car_sales.sav* that does not include any transformed versions of the fields.

- **carpet.sav.** In a popular example (Green and Wind, 1973), a company interested in marketing a new carpet cleaner wants to examine the influence of five factors on consumer preference—package design, brand name, price, a *Good Housekeeping* seal, and a money-back guarantee. There are three factor levels for package design, each one differing in the location of the applicator brush; three brand names (*K2R*, *Glory*, and *Bissell*); three price levels; and two levels (either no or yes) for each of the last two factors. Ten consumers rank 22 profiles defined by these factors. The variable *Preference* contains the rank of the average rankings for each profile. Low rankings correspond to high preference. This variable reflects an overall measure of preference for each profile.
- **carpet_prefs.sav.** This data file is based on the same example as described for *carpet.sav*, but it contains the actual rankings collected from each of the 10 consumers. The consumers were asked to rank the 22 product profiles from the most to the least preferred. The variables *PREF1* through *PREF22* contain the identifiers of the associated profiles, as defined in *carpet_plan.sav*.
- **catalog.sav.** This data file contains hypothetical monthly sales figures for three products sold by a catalog company. Data for five possible predictor variables are also included.
- **catalog_seasfac.sav.** This data file is the same as *catalog.sav* except for the addition of a set of seasonal factors calculated from the Seasonal Decomposition procedure along with the accompanying date variables.
- **cellular.sav.** This is a hypothetical data file that concerns a cellular phone company's efforts to reduce churn. Churn propensity scores are applied to accounts, ranging from 0 to 100. Accounts scoring 50 or above may be looking to change providers.
- **ceramics.sav.** This is a hypothetical data file that concerns a manufacturer's efforts to determine whether a new premium alloy has a greater heat resistance than a standard alloy. Each case represents a separate test of one of the alloys; the heat at which the bearing failed is recorded.
- **cereal.sav.** This is a hypothetical data file that concerns a poll of 880 people about their breakfast preferences, also noting their age, gender, marital status, and whether or not they have an active lifestyle (based on whether they exercise at least twice a week). Each case represents a separate respondent.
- **clothing_defects.sav.** This is a hypothetical data file that concerns the quality control process at a clothing factory. From each lot produced at the factory, the inspectors take a sample of clothes and count the number of clothes that are unacceptable.
- **coffee.sav.** This data file pertains to perceived images of six iced-coffee brands (Kennedy, Riquier, and Sharp, 1996) . For each of 23 iced-coffee image attributes, people selected all brands that were described by the attribute. The six brands are denoted AA, BB, CC, DD, EE, and FF to preserve confidentiality.
- **contacts.sav.** This is a hypothetical data file that concerns the contact lists for a group of corporate computer sales representatives. Each contact is categorized by the department of the company in which they work and their company ranks. Also recorded are the amount of the last sale made, the time since the last sale, and the size of the contact's company.
- **creditpromo.sav.** This is a hypothetical data file that concerns a department store's efforts to evaluate the effectiveness of a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months. Half received a standard seasonal ad.

- **customer_dbase.sav.** This is a hypothetical data file that concerns a company's efforts to use the information in its data warehouse to make special offers to customers who are most likely to reply. A subset of the customer base was selected at random and given the special offers, and their responses were recorded.
- **customer_information.sav.** A hypothetical data file containing customer mailing information, such as name and address.
- **customers_model.sav.** This file contains hypothetical data on individuals targeted by a marketing campaign. These data include demographic information, a summary of purchasing history, and whether or not each individual responded to the campaign. Each case represents a separate individual.
- **customers_new.sav.** This file contains hypothetical data on individuals who are potential candidates for a marketing campaign. These data include demographic information and a summary of purchasing history for each individual. Each case represents a separate individual.
- **debate.sav.** This is a hypothetical data file that concerns paired responses to a survey from attendees of a political debate before and after the debate. Each case corresponds to a separate respondent.
- **debate_aggregate.sav.** This is a hypothetical data file that aggregates the responses in *debate.sav*. Each case corresponds to a cross-classification of preference before and after the debate.
- **demo.sav.** This is a hypothetical data file that concerns a purchased customer database, for the purpose of mailing monthly offers. Whether or not the customer responded to the offer is recorded, along with various demographic information.
- **demo_cs_1.sav.** This is a hypothetical data file that concerns the first step of a company's efforts to compile a database of survey information. Each case corresponds to a different city, and the region, province, district, and city identification are recorded.
- **demo_cs_2.sav.** This is a hypothetical data file that concerns the second step of a company's efforts to compile a database of survey information. Each case corresponds to a different household unit from cities selected in the first step, and the region, province, district, city, subdivision, and unit identification are recorded. The sampling information from the first two stages of the design is also included.
- **demo_cs.sav.** This is a hypothetical data file that contains survey information collected using a complex sampling design. Each case corresponds to a different household unit, and various demographic and sampling information is recorded.
- **dmdata.sav.** This is a hypothetical data file that contains demographic and purchasing information for a direct marketing company.
- **dietstudy.sav.** This hypothetical data file contains the results of a study of the "Stillman diet" (Rickman, Mitchell, Dingman, and Dalen, 1974). Each case corresponds to a separate subject and records his or her pre- and post-diet weights in pounds and triglyceride levels in mg/100 ml.
- **dischargedata.sav.** This is a data file concerning *Seasonal Patterns of Winnipeg Hospital Use*, (Menec , Roos, Nowicki, MacWilliam, Finlayson , and Black, 1999) from the Manitoba Centre for Health Policy.

- **dvdplayer.sav.** This is a hypothetical data file that concerns the development of a new DVD player. Using a prototype, the marketing team has collected focus group data. Each case corresponds to a separate surveyed user and records some demographic information about them and their responses to questions about the prototype.
- **flying.sav.** This data file contains the flying mileages between 10 American cities.
- **german_credit.sav.** This data file is taken from the "German credit" dataset in the Repository of Machine Learning Databases (Blake and Merz, 1998) at the University of California, Irvine.
- **grocery_1month.sav.** This hypothetical data file is the *grocery_coupons.sav* data file with the weekly purchases "rolled-up" so that each case corresponds to a separate customer. Some of the variables that changed weekly disappear as a result, and the amount spent recorded is now the sum of the amounts spent during the four weeks of the study.
- **grocery_coupons.sav.** This is a hypothetical data file that contains survey data collected by a grocery store chain interested in the purchasing habits of their customers. Each customer is followed for four weeks, and each case corresponds to a separate customer-week and records information about where and how the customer shops, including how much was spent on groceries during that week.
- **guttman.sav.** Bell (Bell, 1961) presented a table to illustrate possible social groups. Guttman (Guttman, 1968) used a portion of this table, in which five variables describing such things as social interaction, feelings of belonging to a group, physical proximity of members, and formality of the relationship were crossed with seven theoretical social groups, including crowds (for example, people at a football game), audiences (for example, people at a theater or classroom lecture), public (for example, newspaper or television audiences), mobs (like a crowd but with much more intense interaction), primary groups (intimate), secondary groups (voluntary), and the modern community (loose confederation resulting from close physical proximity and a need for specialized services).
- **healthplans.sav.** This is a hypothetical data file that concerns an insurance group's efforts to evaluate four different health care plans for small employers. Twelve employers are recruited to rank the plans by how much they would prefer to offer them to their employees. Each case corresponds to a separate employer and records the reactions to each plan.
- **health_funding.sav.** This is a hypothetical data file that contains data on health care funding (amount per 100 population), disease rates (rate per 10,000 population), and visits to health care providers (rate per 10,000 population). Each case represents a different city.
- **hivassay.sav.** This is a hypothetical data file that concerns the efforts of a pharmaceutical lab to develop a rapid assay for detecting HIV infection. The results of the assay are eight deepening shades of red, with deeper shades indicating greater likelihood of infection. A laboratory trial was conducted on 2,000 blood samples, half of which were infected with HIV and half of which were clean.
- **hourlywagedata.sav.** This is a hypothetical data file that concerns the hourly wages of nurses from office and hospital positions and with varying levels of experience.
- **insurance_claims.sav.** This is a hypothetical data file that concerns an insurance company that wants to build a model for flagging suspicious, potentially fraudulent claims. Each case represents a separate claim.

- **insure.sav.** This is a hypothetical data file that concerns an insurance company that is studying the risk factors that indicate whether a client will have to make a claim on a 10-year term life insurance contract. Each case in the data file represents a pair of contracts, one of which recorded a claim and the other didn't, matched on age and gender.
- **judges.sav.** This is a hypothetical data file that concerns the scores given by trained judges (plus one enthusiast) to 300 gymnastics performances. Each row represents a separate performance; the judges viewed the same performances.
- **kinship_dat.sav.** Rosenberg and Kim (Rosenberg and Kim, 1975) set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criterion from the first sort. Thus, a total of six "sources" were obtained. Each source corresponds to a 15×15 proximity matrix, whose cells are equal to the number of people in a source minus the number of times the objects were partitioned together in that source.
- **kinship_ini.sav.** This data file contains an initial configuration for a three-dimensional solution for *kinship_dat.sav*.
- **kinship_var.sav.** This data file contains independent variables *gender, gener*(ation), and *degree* (of separation) that can be used to interpret the dimensions of a solution for *kinship_dat.sav*. Specifically, they can be used to restrict the space of the solution to a linear combination of these variables.
- **mailresponse.sav.** This is a hypothetical data file that concerns the efforts of a clothing manufacturer to determine whether using first class postage for direct mailings results in faster responses than bulk mail. Order-takers record how many weeks after the mailing each order is taken.
- **marketvalues.sav.** This data file concerns home sales in a new housing development in Algonquin, Ill., during the years from 1999–2000. These sales are a matter of public record.
- **mutualfund.sav.** This data file concerns stock market information for various tech stocks listed on the S&P 500. Each case corresponds to a separate company.
- **nhis2000_subset.sav.** The National Health Interview Survey (NHIS) is a large, population-based survey of the U.S. civilian population. Interviews are carried out face-to-face in a nationally representative sample of households. Demographic information and observations about health behaviors and status are obtained for members of each household. This data file contains a subset of information from the 2000 survey. National Center for Health Statistics. National Health Interview Survey, 2000. Public-use data file and documentation. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accessed 2003.
- **ozone.sav.** The data include 330 observations on six meteorological variables for predicting ozone concentration from the remaining variables. Previous researchers (Breiman and Friedman, 1985), (Hastie and Tibshirani, 1990), among others found nonlinearities among these variables, which hinder standard regression approaches.
- **pain_medication.sav.** This hypothetical data file contains the results of a clinical trial for anti-inflammatory medication for treating chronic arthritic pain. Of particular interest is the time it takes for the drug to take effect and how it compares to an existing medication.

- **patient_los.sav.** This hypothetical data file contains the treatment records of patients who were admitted to the hospital for suspected myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **patlos_sample.sav.** This hypothetical data file contains the treatment records of a sample of patients who received thrombolytics during treatment for myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **polishing.sav.** This is the "Nambeware Polishing Times" data file from the Data and Story Library. It concerns the efforts of a metal tableware manufacturer (Nambe Mills, Santa Fe, N. M.) to plan its production schedule. Each case represents a different item in the product line. The diameter, polishing time, price, and product type are recorded for each item.
- **poll_cs.sav.** This is a hypothetical data file that concerns pollsters' efforts to determine the level of public support for a bill before the legislature. The cases correspond to registered voters. Each case records the county, township, and neighborhood in which the voter lives.
- **poll_cs_sample.sav.** This hypothetical data file contains a sample of the voters listed in *poll_cs.sav*. The sample was taken according to the design specified in the *poll.csplan* plan file, and this data file records the inclusion probabilities and sample weights. Note, however, that because the sampling plan makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*poll_jointprob.sav*). The additional variables corresponding to voter demographics and their opinion on the proposed bill were collected and added the data file after the sample as taken.
- **property_assess.sav.** This is a hypothetical data file that concerns a county assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties sold in the county in the past year. Each case in the data file records the township in which the property lies, the assessor who last visited the property, the time since that assessment, the valuation made at that time, and the sale value of the property.
- **property_assess_cs.sav.** This is a hypothetical data file that concerns a state assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties in the state. Each case in the data file records the county, township, and neighborhood in which the property lies, the time since the last assessment, and the valuation made at that time.
- **property_assess_cs_sample.sav.** This hypothetical data file contains a sample of the properties listed in *property_assess_cs.sav*. The sample was taken according to the design specified in the *property_assess.csplan* plan file, and this data file records the inclusion probabilities and sample weights. The additional variable *Current value* was collected and added to the data file after the sample was taken.
- **recidivism.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender and records their demographic information, some details of their first crime, and then the time until their second arrest, if it occurred within two years of the first arrest.
- **recidivism_cs_sample.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender, released from their first arrest during the month of June, 2003, and records their demographic information, some details of their first crime, and

the data of their second arrest, if it occurred by the end of June, 2006. Offenders were selected from sampled departments according to the sampling plan specified in *recidivism_cs.csplan*; because it makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*recidivism_cs_jointprob.sav*).

- **rfm_transactions.sav.** A hypothetical data file containing purchase transaction data, including date of purchase, item(s) purchased, and monetary amount of each transaction.
- **salesperformance.sav.** This is a hypothetical data file that concerns the evaluation of two new sales training courses. Sixty employees, divided into three groups, all receive standard training. In addition, group 2 gets technical training; group 3, a hands-on tutorial. Each employee was tested at the end of the training course and their score recorded. Each case in the data file represents a separate trainee and records the group to which they were assigned and the score they received on the exam.
- **satisf.sav.** This is a hypothetical data file that concerns a satisfaction survey conducted by a retail company at 4 store locations. 582 customers were surveyed in all, and each case represents the responses from a single customer.
- **screws.sav.** This data file contains information on the characteristics of screws, bolts, nuts, and tacks (Hartigan, 1975).
- **shampoo_ph.sav.** This is a hypothetical data file that concerns the quality control at a factory for hair products. At regular time intervals, six separate output batches are measured and their pH recorded. The target range is 4.5–5.5.
- **ships.sav.** A dataset presented and analyzed elsewhere (McCullagh et al., 1989) that concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the ship type, construction period, and service period. The aggregate months of service for each cell of the table formed by the cross-classification of factors provides values for the exposure to risk.
- **site.sav.** This is a hypothetical data file that concerns a company's efforts to choose new sites for their expanding business. They have hired two consultants to separately evaluate the sites, who, in addition to an extended report, summarized each site as a "good," "fair," or "poor" prospect.
- **siteratings.sav.** This is a hypothetical data file that concerns the beta testing of an e-commerce firm's new Web site. Each case represents a separate beta tester, who scored the usability of the site on a scale from 0–20.
- **smokers.sav.** This data file is abstracted from the 1998 National Household Survey of Drug Abuse and is a probability sample of American households. Thus, the first step in an analysis of this data file should be to weight the data to reflect population trends.
- **smoking.sav.** This is a hypothetical table introduced by Greenacre (Greenacre, 1984). The table of interest is formed by the crosstabulation of smoking behavior by job category. The variable *Staff Group* contains the job categories *Sr Managers*, *Jr Managers*, *Sr Employees*, *Jr Employees*, and *Secretaries*, plus the category *National Average*, which can be used as supplementary to an analysis. The variable *Smoking* contains the behaviors *None*, *Light*, *Medium*, and *Heavy*, plus the categories *No Alcohol* and *Alcohol*, which can be used as supplementary to an analysis.

- **storebrand.sav.** This is a hypothetical data file that concerns a grocery store manager's efforts to increase sales of the store brand detergent relative to other brands. She puts together an in-store promotion and talks with customers at check-out. Each case represents a separate customer.
- **stores.sav.** This data file contains hypothetical monthly market share data for two competing grocery stores. Each case represents the market share data for a given month.
- **stroke_clean.sav.** This hypothetical data file contains the state of a medical database after it has been cleaned using procedures in the Data Preparation option.
- **stroke_invalid.sav.** This hypothetical data file contains the initial state of a medical database and contains several data entry errors.
- **stroke_survival.** This hypothetical data file concerns survival times for patients exiting a rehabilitation program post-ischemic stroke face a number of challenges. Post-stroke, the occurrence of myocardial infarction, ischemic stroke, or hemorrhagic stroke is noted and the time of the event recorded. The sample is left-truncated because it only includes patients who survived through the end of the rehabilitation program administered post-stroke.
- **stroke_valid.sav.** This hypothetical data file contains the state of a medical database after the values have been checked using the Validate Data procedure. It still contains potentially anomalous cases.
- **survey_sample.sav.** This hypothetical data file contains survey data, including demographic data and various attitude measures.
- **tastetest.sav.** This is a hypothetical data file that concerns the effect of mulch color on the taste of crops. Strawberries grown in red, blue, and black mulch were rated by taste-testers on an ordinal scale of 1 to 5 (far below to far above average). Each case represents a separate taste-tester.
- **telco.sav.** This is a hypothetical data file that concerns a telecommunications company's efforts to reduce churn in their customer base. Each case corresponds to a separate customer and records various demographic and service usage information.
- **telco_extra.sav.** This data file is similar to the *telco.sav* data file, but the "tenure" and log-transformed customer spending variables have been removed and replaced by standardized log-transformed customer spending variables.
- **telco_missing.sav.** This data file is a subset of the *telco.sav* data file, but some of the demographic data values have been replaced with missing values.
- **testmarket.sav.** This hypothetical data file concerns a fast food chain's plans to add a new item to its menu. There are three possible campaigns for promoting the new product, so the new item is introduced at locations in several randomly selected markets. A different promotion is used at each location, and the weekly sales of the new item are recorded for the first four weeks. Each case corresponds to a separate location-week.
- **testmarket_1month.sav.** This hypothetical data file is the *testmarket.sav* data file with the weekly sales "rolled-up" so that each case corresponds to a separate location. Some of the variables that changed weekly disappear as a result, and the sales recorded is now the sum of the sales during the four weeks of the study.
- **tree_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.

- **tree_credit.sav.** This is a hypothetical data file containing demographic and bank loan history data.
- **tree_missing_data.sav** This is a hypothetical data file containing demographic and bank loan history data with a large number of missing values.
- **tree_score_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree_textdata.sav.** A simple data file with only two variables intended primarily to show the default state of variables prior to assignment of measurement level and value labels.
- **tv-survey.sav.** This is a hypothetical data file that concerns a survey conducted by a TV studio that is considering whether to extend the run of a successful program. 906 respondents were asked whether they would watch the program under various conditions. Each row represents a separate respondent; each column is a separate condition.
- **ulcer_recurrence.sav.** This file contains partial information from a study designed to compare the efficacy of two therapies for preventing the recurrence of ulcers. It provides a good example of interval-censored data and has been presented and analyzed elsewhere (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** This file reorganizes the information in *ulcer_recurrence.sav* to allow you model the event probability for each interval of the study rather than simply the end-of-study event probability. It has been presented and analyzed elsewhere (Collett et al., 2003).
- **verd1985.sav.** This data file concerns a survey (Verdegaal, 1985). The responses of 15 subjects to 8 variables were recorded. The variables of interest are divided into three sets. Set 1 includes *age* and *marital*, set 2 includes *pet* and *news*, and set 3 includes *music* and *live*. *Pet* is scaled as multiple nominal and *age* is scaled as ordinal; all of the other variables are scaled as single nominal.
- **virus.sav.** This is a hypothetical data file that concerns the efforts of an Internet service provider (ISP) to determine the effects of a virus on its networks. They have tracked the (approximate) percentage of infected e-mail traffic on its networks over time, from the moment of discovery until the threat was contained.
- **waittimes.sav.** This is a hypothetical data file that concerns customer waiting times for service at three different branches of a local bank. Each case corresponds to a separate customer and records the time spent waiting and the branch at which they were conducting their business.
- **webusability.sav.** This is a hypothetical data file that concerns usability testing of a new e-store. Each case corresponds to one of five usability testers and records whether or not the tester succeeded at each of six separate tasks.
- **wheeze_steubenville.sav.** This is a subset from a longitudinal study of the health effects of air pollution on children (Ware, Dockery, Spiro III, Speizer, and Ferris Jr., 1984). The data contain repeated binary measures of the wheezing status for children from Steubenville, Ohio, at ages 7, 8, 9 and 10 years, along with a fixed recording of whether or not the mother was a smoker during the first year of the study.
- **workprog.sav.** This is a hypothetical data file that concerns a government works program that tries to place disadvantaged people into better jobs. A sample of potential program participants were followed, some of whom were randomly selected for enrollment in the program, while others were not. Each case represents a separate program participant.

Bibliography

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Blake, C. L., and C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., and J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580–598.

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

Davison, A. C., and D. V. Hinkley. 2006. *Bootstrap Methods and their Application.* : Cambridge University Press.

Green, P. E., and V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., and Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Greenacre, M. J. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, 469–506.

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., and R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Kennedy, R., C. Riquier, and B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56–70.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Menec , V., N. Roos, D. Nowicki, L. MacWilliam, G. Finlayson , and C. Black. 1999. *Seasonal Patterns of Winnipeg Hospital Use*. : Manitoba Centre for Health Policy.

Price, R. H., and D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, 579–586.

Rickman, R., N. Mitchell, J. Dingman, and J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, 54–58.

Rosenberg, S., and M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, 489–502.

Shao, J., and D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, and H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, 363–368.

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, and B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, 366–374.

Index