

Paper 349-2012

Enhanced Data Analysis using SAS® ODS Graphics and Statistical Graphics

Patricia A. Berglund, Institute For Social Research-University of Michigan, Ann Arbor, Michigan

ABSTRACT

This paper presents practical examples of enhanced data analysis through use of ODS Graphics and the Statistical Graphics (SG) procedures. SAS 9.3 ODS Graphics options and selected SG procedures are demonstrated with a variety of analytic techniques including examination of variable distributions and common regression methods. Procedures such as PROC MEANS, PROC SURVEYMEANS, PROC UNIVARIATE, PROC REG, PROC LOGISTIC, and PROC SURVEYLOGISTIC are used with accompanying ODS Graphics and SG tools for enhanced analysis. The analysis applications include unweighted and weighted analyses and, where appropriate, SAS SURVEY procedures for analysis of data derived from a complex sample design. The techniques presented here can be used in any operating system, and are intended for an intermediate level audience.

INTRODUCTION

The objective of this paper is to provide guidance on use of ODS Graphics tools and Statistical Graphics procedures for enhanced data analysis. The ODS Graphics options built into the SAS statistical procedures offer a rich array of tools to promote better understanding and interpretation of results while demanding minimal coding on the user's part. In addition to ODS Graphics, the SG procedures are excellent options for production of custom graphics using data directly within the procedure or data sets previously computed by a selected statistical procedure.

The analysis applications presented demonstrate use of both tools for analysis of variable distributions and regression modeling. In general, use of ODS Graphics is preferred but for some analyses, use of output analytic data sets as inputs for PROC SGPLOT is recommended. This two step approach can offer great flexibility in graph production for a wide range of needs.

The National Health and Nutrition Examination Survey (NHANES, 2005-2006) and the National Comorbidity Survey-Replication (NCS-R, 2001) data sets used in this paper are based on a complex sample design, therefore all variance estimates or standard errors should be calculated using an appropriate SAS SURVEY procedure. For more information on these two data sets, see the project websites-NHANES (<http://www.cdc.gov/nchs/>), NCS-R ([http://www.icpsr.umich.edu/cpes www.nchs](http://www.icpsr.umich.edu/cpes/www.nchs)).

Despite the need for corrected variance estimates, many of the ODS Graphics tools and SG procedures can be used to good advantage with complex sample design data sets. For example, the analysis applications illustrate how the analyst can selectively utilize ODS Graphics or SG procedures for diagnostic tasks while correctly estimating variance with the SURVEY procedures, when needed. For those working with SRS data, the concepts are readily applied to data analysis without the added complexity of the complex sample adjustments. The rationale for featuring complex sample design data is that many public release data sets are based on probability samples and the goal of this paper is to include a wide range of analytic tasks that analysts are likely to encounter in daily work.

OVERVIEW OF ODS GRAPHICS AND STATISTICAL GRAPHICS PROCEDURES

The ODS Graphics system allows the analyst to produce high quality graphics using pre-programmed tools built into numerous SAS analytic procedures. Related to ODS Graphics, the Statistical Graphics (SG) procedures produce a wide range of graphics via a streamlined coding process. Both approaches enable high quality and effective production of visual products appropriate for the statistical procedure of choice. Often, a precise visual display can reveal issues or patterns in the output where numbers are not as easily interpreted or understood. Therefore, the combination of tabular output paired with ODS Graphics or graphics produced by the SG procedures represents a powerful and effective approach to data analysis. ODS Graphics are included in procedures from SAS/STAT, Base SAS, SAS/QC and other modules. See the SAS documentation for a complete list of procedures.

SIMPLE RANDOM SAMPLE AND COMPLEX SAMPLE DESIGN DATA

Simple random sample (SRS) data is based upon the assumption that the sample is drawn from an infinite population by simple random sampling (SAS/STAT Documentation: Overview of Survey Procedures). In general, no special variance estimation methods are needed for analysis of SRS data and analyses are either un-weighted or weighted, depending on the data design and if weights are appropriate. Nearly all SAS analytic procedures assume a simple random sample.

On the other hand, data derived from a complex sample design requires that variances, hypothesis tests, and degrees of freedom incorporate the design features and probability weights for correctly adjusted estimation. This

can be accomplished in SAS via use of the SURVEY procedures. For more information on complex sample data analysis, see the SAS *"Introduction to Survey Sampling and Analysis Procedures"* of the SAS/STAT documentation or a text such as *Applied Survey Data Analysis* (Heeringa, West and Berglund, 2010).

The complex sample design should be carefully considered when using ODS Graphics or the SG procedures. For example, ODS graphs or diagnostic tools that make use of variance estimates/standard errors or hypothesis tests should take the complex sample into account via use of a SURVEY procedure and subsequently graphed using a pre-computed data set. This paper presents some examples of how to carry out this process. See Kish (1965), or Rust, (1985) for more information on analysis of complex sample design data.

ANALYSIS APPLICATIONS

DISTRIBUTIONAL ANALYSES

The analysis applications focus on obesity and related health issues such as high blood pressure and cholesterol levels among US adults (age 18 and over). Various demographic variables such as gender, region and age at interview are also included in the analyses. For general estimates of obesity status and BMI by selected demographic variables, the NCS-R data set is used while for linear and logistic regression models, the NHANES data is used as it provides additional detail on blood pressure readings and total cholesterol counts (from the Medical Examination Survey).

Examples 1-6 cover distributional analyses such as frequency tables for categorical or count variables and means or univariate analyses for continuous variables. The demonstrations feature use of ODS Graphics along with use of PROC SGPLOT with pre-computed data sets, as needed.

CATEGORICAL VARIABLES

EXAMPLE 1 - WEIGHTED FREQUENCY TABLES WITH ODS GRAPHICS

Example 1 presents weighted frequency plots for 2 variables, obese and obese6ca. Table 1.1 and Figures 1.1 and 1.2 show plots for the variable obese, a 2 category indicator of a respondent being obese or not (BMI \geq 30). Table 1.2 and Figures 1.3 and 1.4 pertain to the variable obese6ca, a 6 category obesity status variable.

Prior to beginning the analysis, the ODS Graphics tool must be enabled by use of the `ods graphics on ;` statement. This is assumed to be on for all analyses in this paper. Both variables are plotted using `type freqplot` with a `percent` scale and a `cumfreqplot` or cumulative frequency plot also with a `percent` scale. All plots are weighted with the Part 2 NCS-R weight. Use of `type=dot` option on the `cumfreqplot` statement allows the analyst to specify a dot for each cumulative frequency point.

```
ods graphics on ;
proc freq data=ncsr ;
    tables obese obese6ca /plots=(freqplot(scale=percent)
    cumfreqplot(scale=percent type=dot));
    weight ncsrwtlg ;
run ;
```

Table 1.1 and Figures 1.1 and 1.2 show that 25.29% of the total 5692 people are considered obese (BMI > 30). These percentages are weighted to correctly represent the US adult population in the early part of the century. Because the frequency tables analysis do not include standard errors or hypothesis tests, the complex sample variance estimation process is not required.

Obesity Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Obese	4252.384	74.71	4252.384	74.71
Obese	1439.616	25.29	5692	100.00

Table 1.1

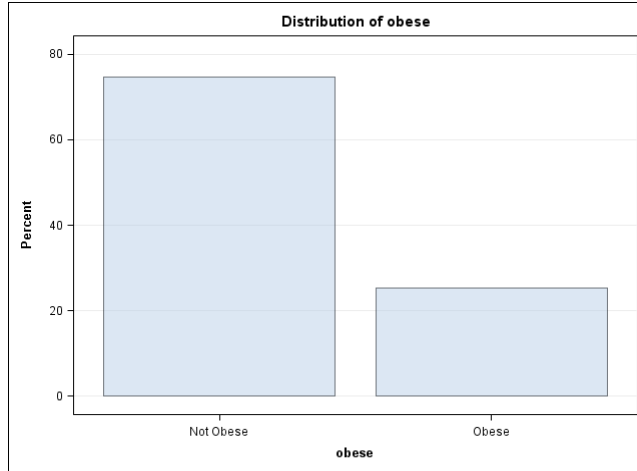


Figure 1.1

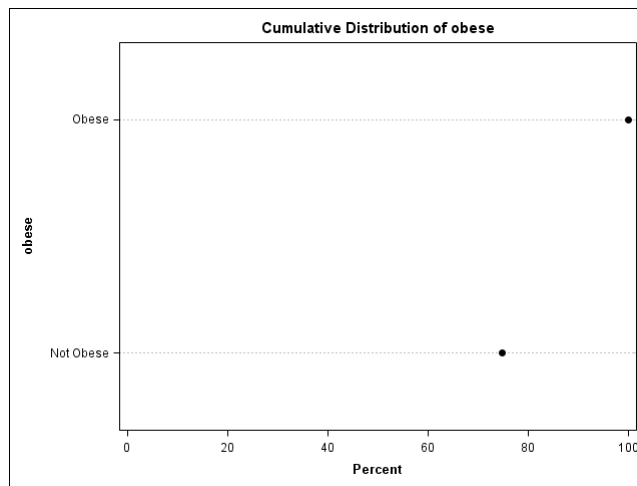


Figure 1.2

Table 1.2 and Figures 1.3-1.4 show the detail of 6 obesity status levels. Again, 25.29% of the total 5692 people are considered obese (BMI > 30) with 3.67% with BMI <=18.5, 38.29% BMI between 18.51 and 24.99, and 32.75% BMI in the 25-29.99 range.

	1=<18.5	2=18.5-24.9	3=25-29.9	4=30-34.9	5=35-39.9	6=40+		
OBESE6CA	Frequency	Percent	Cumulative Frequency	Cumulative Percent				
1	208.8624	3.67	208.8624	3.67				
2	2179.276	38.29	2388.138	41.96				
3	1864.246	32.75	4252.384	74.71				
4	895.0995	15.73	5147.484	90.43				
5	338.328	5.94	5485.812	96.38				
6	206.1886	3.62	5692	100.00				

Table 1.2

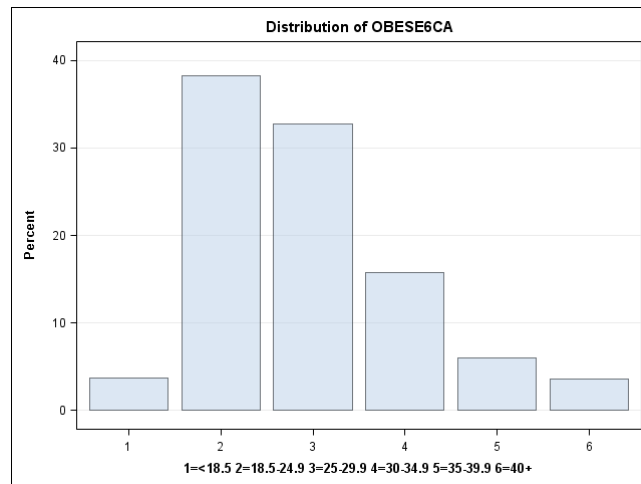


Figure 1.3

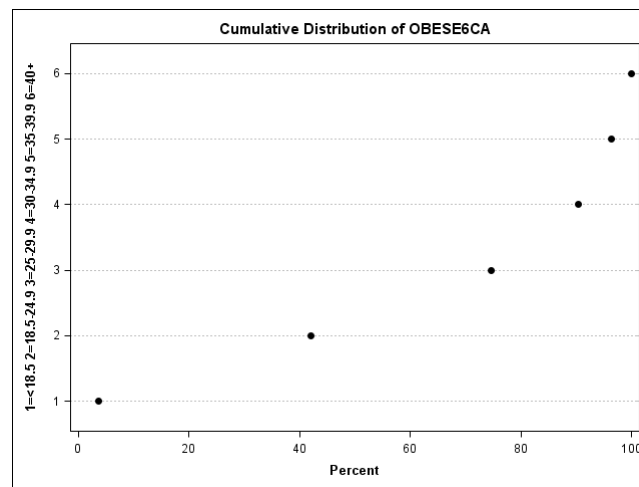


Figure 1.4

EXAMPLE 2 - FREQUENCY ANALYSIS OF MULTI-WAY TABLE WITH ODS GRAPHICS AND ODDS RATIO PLOT

Example 2 demonstrates a weighted frequency table analysis with odds ratios and complex sample design corrected confidence intervals produced by PROC SURVEYFREQ. This is required due to the complex sample design of the NCS-R data set. Partial output for the 4 age groups of 18-29, 30-44, 45-59, and 60+ is presented in tabular output (Table 2.1) along with the plot of Odds Ratios with 95% confidence limits in Figure 2.1.

```
proc surveyfreq data=ncsr order=formatted ;
  strata sestrat ; cluster seclustr ; weight ncsrwtlg ;
  tables ag4cat*sex*obese / plots(only)=oddsratioplot relrisk ;
  format ag4cat af. sex sexfr. obese obr. mde mdef. ;
run ;
```

The above code uses the SURVEYFREQ procedure with the design variables `sestrat` and `seclustr` along with the Part 2 weight, `ncsrwtlg`. Use of the `plots(only)=oddsratioplot relrisk;` statement in the `tables` statement requests an OR plot with corrected confidence limits.

Table of SEX by obese						
Controlling for ag4cat=60+						
SEX	obese	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
1: Female	1: Obese	162	168.03348	21.53082	13.9296	1.6589
	2: Not Obese	445	528.27953	47.52591	43.7933	2.5740
	Total	607	696.31301	54.80012	57.7229	2.6669
2: Male	1: Obese	81	100.85459	14.36378	8.3606	1.0497
	2: Not Obese	286	409.13546	40.08864	33.9165	2.8170
	Total	367	509.99005	42.79496	42.2771	2.6669
Total	1: Obese	243	268.88807	27.42681	22.2903	1.9059
	2: Not Obese	731	937.41500	62.14784	77.7097	1.9059
	Total	974	1206	72.75483	100.000	

Table 2.1

Odds Ratio and Relative Risks (Row1/Row2)			
	Estimate	95% Confidence Limits	
Odds Ratio	1.2903	0.8090	2.0580
Column 1 Relative Risk	1.2203	0.8463	1.7596
Column 2 Relative Risk	0.9457	0.8545	1.0466
Sample Size = 5692			

Table 2.2

Figure 2.1 is the ODS Graphics plot of Odds Ratios and 95% CI's. It reveals that the 95% confidence limits all include 1.0 and are therefore non-significant while relative to men, women have elevated odds (> 1.0) in each age group except the 30-44 year old group. The plot enables an easy way to evaluate the relationships between age, gender and being obese while taking the complex sample design into account via PROC SURVEYFREQ.

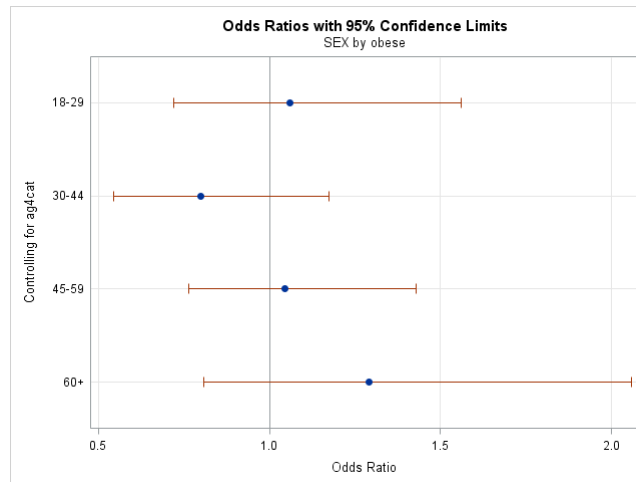


Figure 2.1

CONTINUOUS VARIABLES

EXAMPLE 3 - MEANS ANALYSIS OF A CONTINUOUS VARIABLE WITH PRE-COMPUTED DATA AND PROC SGPLOT

Example 3 demonstrates use of PROC SURVEYMEANS to obtain means and corrected standard errors graphed in PROC SGPLOT, using the pre-computed data. This two step approach is needed to obtain the complex sample corrected standard errors used in the confidence limits since PROC SURVEYMEANS does not offer ODS Graphics. The variable of interest is body mass index (bmi) and a domain or sub-group analysis using region of the country is requested. An ODS output data set called "outstat" is saved for use with PROC SGPLOT.

```
proc surveymeans data=ncsr;
  strata sestrat ; cluster seclustr ; weight ncsrwtlg ;
  var bmi ; domain region ;
  ods output domain=outstat ;
run ;
```

The SGPLOT syntax uses two statements: `band` for the confidence limits for the mean of BMI and `series` for the series plot of `y` and `x` (mean for BMI and region, respectively). Use of the `xaxis integer;` statement requests integers along the x axis. The code below creates Figure 3.1 which shows the highest mean BMI in the South region with the lowest mean BMI in the West region. The bands display the 95% confidence limits and show that the NorthEast and the South have the largest confidence intervals.

```
proc sgplot data=outstat;
  band x=region lower=lowerclmean upper=upperclmean / legendlabel="95% CLI" ;
  series y=mean x=region ;
  xaxis integer ;
run ;
```

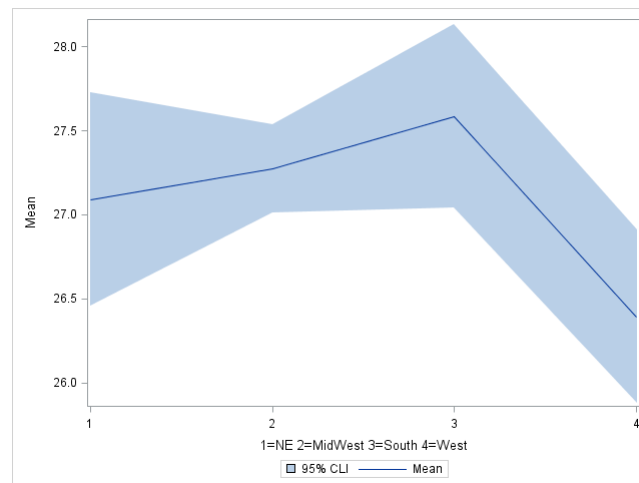


Figure 3.1

EXAMPLE 4 - MEANS ANALYSIS OF A CONTINUOUS VARIABLE WITH MULTIPLE DOMAINS USING PRE-COMPUTED DATA AND PROC SGPLOT

Example 4 builds on Example 3 by adding two crossed domain variables to the BMI analysis. This requires use of PROC SURVEYMEANS with the crossing operator in the domain statement: `domain ag4cat*sex;`, along with the `ods output domain=outstat;` statement to create a pre-computed data set. A print out of "outstat" is presented in Table 4.1.

```

proc surveymeans data=ncsr;
  strata sestrat ; cluster seclustr ; weight ncsrwtlg ;
  var bmi ; domain ag4cat*sex ; format ag4cat af. sex sexfo. ;
  ods output domain =outstat ;
run ;
proc print data=outstat noobs ;
  var ag4cat sex varname mean stderr lowerclmean upperclmean ;
run ;

```

Age	Sex	Variable Name	Mean	StdErr	LowerCLMean	UpperCLMean
18-29	Male	bmi	25.889469	0.269224	25.3461527	26.4327859
18-29	Female	bmi	25.187035	0.322050	24.5371110	25.8369591
30-44	Male	bmi	28.394008	0.356675	27.6742090	29.1138065
30-44	Female	bmi	26.938839	0.376146	26.1797459	27.6979325
45-59	Male	bmi	28.298554	0.298454	27.6962483	28.9008589
45-59	Female	bmi	28.065446	0.244743	27.5715336	28.5593579
60+	Male	bmi	26.902143	0.408545	26.0776663	27.7266202
60+	Female	bmi	27.267585	0.324843	26.6120265	27.9231445

Table 4.1

The code below uses the pre-computed data set with PROC SGPLOT to produce Figure 4.1. This figure presents the mean of BMI by gender/age with design corrected confidence limits. Males have consistently higher mean BMI for all age groups except for age 60+ where mean BMI for women is about 27.3 and 26.9 for men. The highest mean BMI for both men and women occurs in middle age or between the ages of 30-44 (males) and 45-59 (women). The confidence bands overlap for each age*gender group as well. Use of the `group=sex` option requests bands for age groups by sex for the BMI confidence limits while use of the `series` statement with `datalabel` creates a series line with labels for the data points in the plot.

```

proc sgplot data=outstat ;
  band x=ag4cat lower=lowerclmean upper=upperclmean / group=sex
  transparency=.2 ; xaxis integer ;
  series y=mean x=ag4cat /group=sex datalabel ;
run ;

```

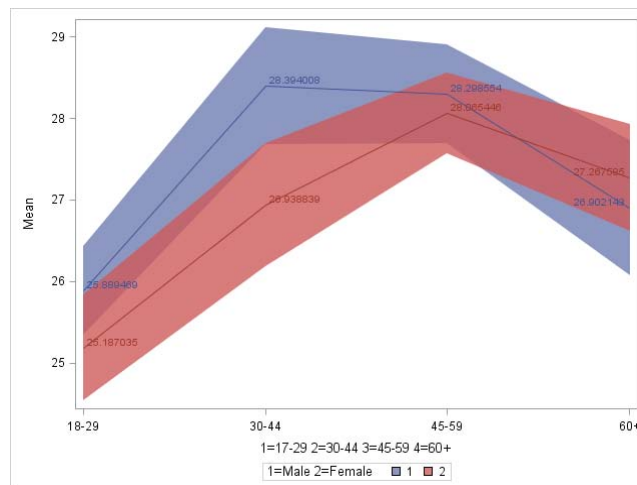


Figure 4.1

EXAMPLE 5 - WEIGHTED HISTOGRAM FOR A CONTINUOUS VARIABLE

Example 5 illustrates how to obtain a weighted histogram from PROC UNIVARIATE. Although UNIVARIATE does include an ODS Graphics histogram with an optionally superimposed normal curve, it does not provide a weighted histogram for use with non-integer weights (the FREQ statement requires integer weights). However, using the "old style" line printer plot option in the PROC UNIVARIATE statement will produce low resolution plots with the weight applied. This example uses the older approach for a correctly weighted histogram. By default, the use of the plot option will also create a normal probability plot (not shown here).

Figure 5.1 also includes a box-plot just to the right of the histogram. The code below uses the `plot` and `plotsize` statements in the procedure syntax to request a plot and with 40 rows used in the plot. Figure 5.1 shows that the mean BMI in US adults is about 27.1 (Boxplot: symbol + represents the mean, lower and upper lines the IQR and the dotted line in the middle represents the median). The rows of the weighted histogram have *'s (stars) which can represent a weighted count of up to 20 along with the weighted number of cases to the right.

```
proc univariate data=ncsr plot plotsize=40;
    var bmi ;
    weight ncsrwtlg ;
run ;
```

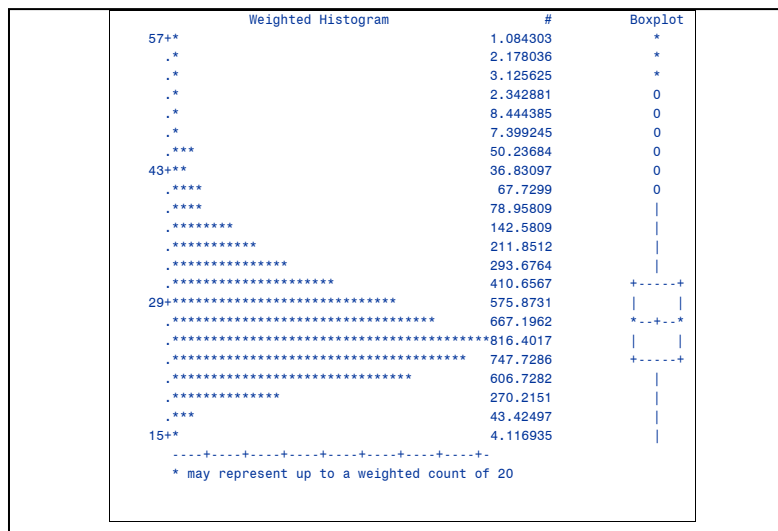


Figure 5.1

EXAMPLE 6 - UN-WEIGHTED HISTOGRAM FOR A CONTINUOUS VARIABLE USING ODS GRAPHICS

Example 6 provides an un-weighted comparison to the weighted histogram in Example 5. This is not technically correct as the histogram isn't weighted but serves as an example of how an un-weighted or FREQ statement weighted (using an integer weight only) histogram can be requested via ODS Graphics in PROC UNIVARIATE. The `histogram bmi/normal;` syntax requests a normal curve be superimposed on the histogram of the analysis variable, bmi. This analysis suggests that the distribution (un-weighted) of BMI in the NCS-R data set is not normally distributed, especially when not properly weighted.

```
proc univariate data=ncsr ;
    var bmi ; histogram bmi / normal ;
run ;
```

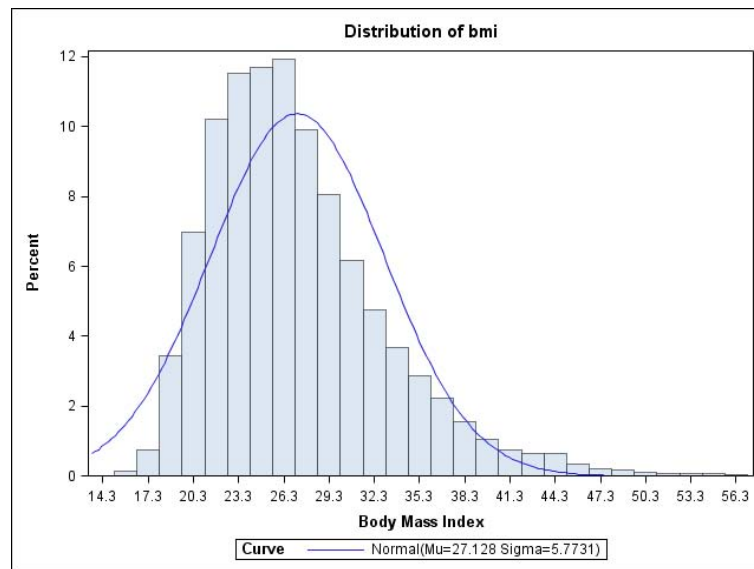


Figure 6.1

REGRESSION EXAMPLES

The regression examples use data from the 2005-2006 NHANES survey to take advantage of medical examination data from the MEC survey. This allows extension of previous examples using the NCS-R data set through use of the cholesterol and blood pressure measurements along with BMI and similar demographic variables already examined.

For the linear and logistic models in Examples 7 and 8, relationships between continuous total cholesterol and BMI (Example 7) and high blood pressure (binary indicator variable) and high cholesterol, gender and BMI (Example 8) are explored. Each regression example utilizes ODS Graphics tools for model diagnostics and results from both PROC SURVEYREG and PROC SURVEYLOGISTIC for final model correctly adjusted standard errors and confidence limits. Because many of the diagnostic tools built into ODS Graphics are useful for both SRS and complex sample design data, the combination usage of ODS Graphics along with adjusted variance estimates from the SURVEY procedures is demonstrated.

EXAMPLE 7 - LINEAR REGRESSION WITH ODS GRAPHICS DIAGNOSTICS

Example 7 presents a simple linear regression model with total cholesterol (lbxtc) regressed on BMI (bmx bmi), among those 18 and older (`age18p=1`). This example uses the 2005-2006 NHANES data set with the MEC weight (`wtmec2yr`).

EXAMPLE 7 STEP 1 - SIMPLE LINEAR REGRESSION WITH ODS GRAPHICS DIAGNOSTIC PANEL

Step 1 requests the full ODS Graphics diagnostic panel (Figure 7.1) for a general overview of model diagnostics. The syntax `plots (maxpoints=10000)=(diagnostics);` specifies a maximum number of data points of 10000 and the full diagnostics panel as output.

```
ods graphics on;
proc reg data=d1.chapter_exercises_nhanes0506 plots
(maxpoints=10000)=(diagnostics) ;
  weight wtmec2yr ;
  model lbxtc = bmxbmi ;
  where age18p=1 ;
run ;
```

Table 7.1 presents the dependent variable and sample information along with the Analysis of Variance table which indicates that the model is significantly better than one with only the intercept and no other predictors in explaining the outcome of total cholesterol, (F Value=22.88 and Pr>F=<.0001). Many other statistics are available from the full output but for this simple regression, the focus is on a few selected numbers and ODS Graphics tools for model diagnosis.

Dependent Variable: LBXTC Total Cholesterol(mg/dL)					
Number of Observations Read	5563				
Number of Observations Used	4923				
Number of Observations with Missing Values	640				
Weight: WTMEC2YR Full Sample 2 Year MEC Exam Weight					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1643385370	1643385370	22.88	<.0001

Table 7.1

The diagnostics panel (Figure 7.1) includes a number of diagnostic plots which indicate problems with influential observations (the top two rows include 6 plots which focus on model fit and potentially influential data points). The residual by predicted value plot shows at least one influential point to consider while the other predicted value and leverage plots and Cook's D all indicate a similar problem. The third row of plots indicates that the residuals are close to normally distributed but that spread of the model fit residuals is greater than the spread of the centered fit, again indicating poor model fit. Another indication of poor model fit is the Quantile plot (row 2 position 1) which indicates that the residuals are not normally distributed (deviation from the straight line).

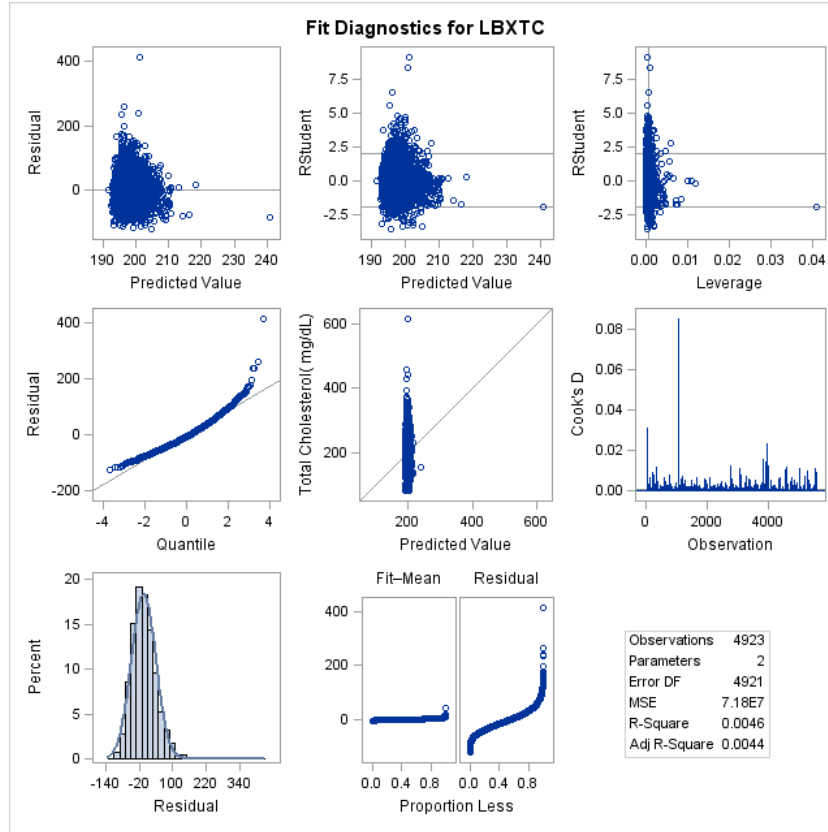


Figure 7.1

Figures 7.2 and 7.3 present plots with more detail regarding model fit. These plots are included when the full diagnostic plots are requested: `plots (maxpoints=10000)=(diagnostics)`; in the PROC REG statement. Once again, these two plots point to data points that may be influential in affecting model performance, (see points at the extremes of the Y and X axis).

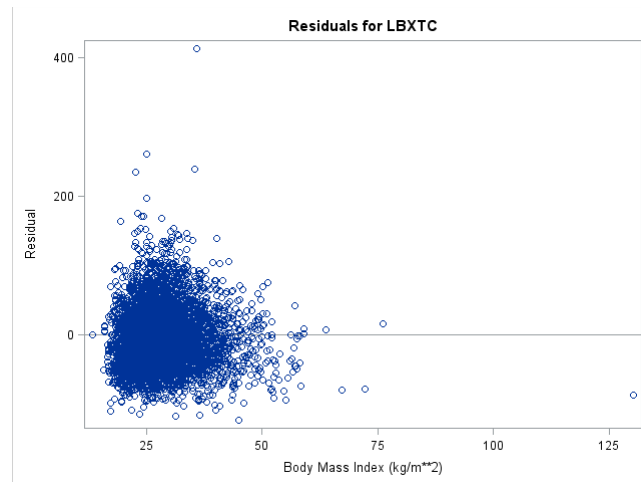


Figure 7.2

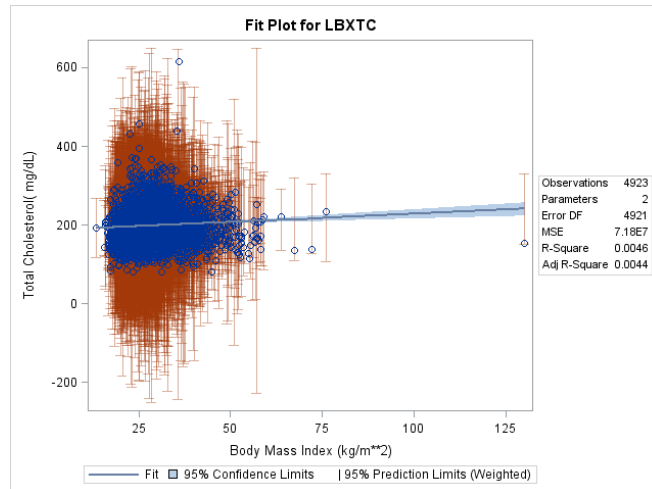


Figure 7.3

EXAMPLE 7 STEP 2 - PINPOINT INFLUENTIAL POINTS

Step 2 demonstrates how to identify an influential point by sequence number (seqn) through use of the `plots` (only `label maxpoints=10000`)=(`RStudentByLeverage CooksD`); and `id seqn`; statements. The `label` statement in the `plots` syntax requests a label for each observation and the `only` keyword requests only the plots listed in the parentheses along with the `id` used as the variable `seqn`. This code produces Figures 7.4 and 7.5. Both figures indicate at least one observation that may be unduly influencing the model fit (`seqn=33228`).

```
proc reg data=d1.chapter_exercises_nhanes0506 plots(only label
maxpoints=10000)=(cooksD RStudentByLeverage );
  weight wtmecl2yr ;
  id seqn ;
  where age18p=1 ;
  model lbxtc = bmx bmi ;
run ;
```

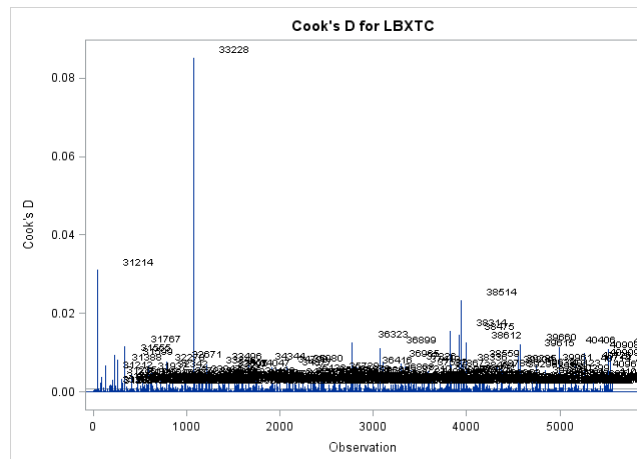


Figure 7.4

EXAMPLE 7 STEP 4 - PROC SURVEYREG FOR ESTIMATION OF FINAL MODEL

Step 4 assumes that all data cleaning is complete, model fit is acceptable, and final model estimation using PROC SURVEYREG is warranted. In this example, the data cleaning/investigation process is clearly not complete but for the sake of demonstration, assume that this is true. In Step 4, use of PROC SURVEYREG ensures that the complex sample is incorporated into the calculation of the SE's and the correct degrees of freedom are used in hypothesis testing.

Table 7.2 provides the regression parameters with correct standard errors in the Age 18+ domain, along with the note about the $df = 15$, using the "fixed rule" (# of clusters-# of strata) for number of degrees of freedom. This analysis indicates that a one point increase in BMI results in a .46 increase in the total cholesterol value, significant at the $\alpha = .05$ level. Another option might be to use a different number of BMI units such as 5 or 10 points instead of 1, resulting in a more broadly defined change in BMI to evaluate.

```
proc surveyreg data=d1.chapter_exercises_nhanes0506 ;
  strata sdmvstra ; cluster sdmvpsu ; weight wtmecl2yr ;
  model lbxtc = bmx bmi ;
  where seqn ne 33228 ; domain age18p;
run ;
```

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	185.002783	2.45934266	75.22	<.0001
BMXBMI	0.457437	0.08186690	5.59	<.0001

Note: The denominator degrees of freedom for the t tests is 15.

Table 7.2

In summary, use of PROC REG and PROC SURVEYREG for model fit evaluation and final model estimation provides the analyst with good tools for visual analysis of model fit and correct variance estimation for the final model.

EXAMPLE 8 LOGISTIC REGRESSION WITH ODS GRAPHICS AND PROC SGLOT

Example 8 demonstrates use of PROC LOGISTIC with ODS Graphics as well as PROC SURVEYLOGISTIC for corrected standard errors once model diagnostics are complete. Pre-computed data from SURVEYLOGISTIC is then used to produce an Odds Ratio plot in PROC SGLOT. A logistic regression approach is used with a binary outcome of high blood pressure (HP is defined as diastolic bp > 80 or systolic bp > 125) predicted by high cholesterol (total cholesterol > 200), and gender, among those 18 and older using NHANES data. This example starts with a frequency table with an ODS Graphics `freqplot` request to examine the characteristics of the outcome variable (Figure 8.1). About 43% of respondents have high blood pressure, among those 18+ and weighted with the `wtmecl2yr` weight variable. This is somewhat higher than previous prevalence estimates of high blood pressure and reflects some recent research in establishing lower blood pressure thresholds for defining high blood pressure.

EXAMPLE 8 STEP 1 - FREQUENCY TABLE WITH PLOT

```
proc freq ;
  tables highbp / plots=(freqplot(scale=percent)) ;
  weight wtmecl2yr ; where age18p=1 ;
run ;
```

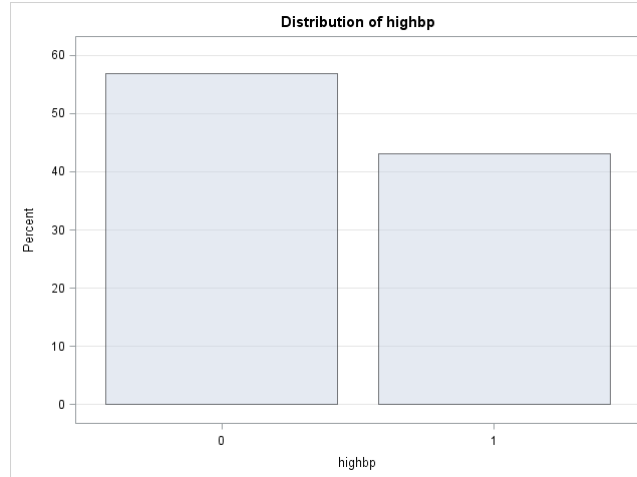


Figure 8.1

EXAMPLE 8 STEP 2 - PROC LOGISTIC WITH ODS GRAPHICS DIAGNOSTIC PLOTS

Step 2 continues the analysis by use of PROC LOGISTIC with ODS Graphics plots- (phat, roc and effect). The code below requests only the three plots listed in the parentheses and also requests the lackfit option on the model statement.

Table 8.1 contains selected output from the logistic regression run and indicates that the model fit is significantly better than a model with only the intercept and no other predictors (LLR ChiSquare is significant at the alpha=.05 level), the c statistic is .643 which is low but acceptable, and the Hosmer-Lemeshow lackfit table indicates a non-significant result meaning that the model fit is adequate and a good correspondence between observed and expected counts in the ten groups. (Table 8.1).

```
proc logistic data=nhanes0506 plots(only maxpoints=10000 label)=(phat roc
effect) ;
  weight wtmecl2yr ; where age18p=1 ;
  class riagendr high_chol (ref='Not High Cholesterol') ;
  model highbp (event='High BP') = bmx bmi riagendr high_chol / lackfit ;
  format highbp bp. riagendr sex. high_chol hc. ;
run ;
```

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11720944.6	3	<.0001
Score	11277982.2	3	<.0001
Wald	10638341.3	3	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	64.3	Somers' D	0.286
Percent Discordant	35.7	Gamma	0.286
Percent Tied	0.0	Tau-a	0.140
Pairs	4555870	c	0.643

Partition for the Hosmer and Lemeshow Test					
Group	Total	highbp = High BP		highbp = Not High BP	
		Observed	Expected	Observed	Expected
1	432	91	103.36	341	328.64
2	433	117	130.03	316	302.97
3	433	150	148.47	283	284.53
4	432	153	162.26	279	269.74
5	433	186	175.41	247	257.59
6	431	194	188.34	237	242.66
7	432	209	204.39	223	227.61
8	432	220	220.77	212	211.23
9	432	234	241.68	198	190.32
10	431	272	281.98	159	149.02

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.8352	8	0.4497

Table 8.1

Figures 8.2-8.4 present diagnostic plots requested by this code: `plots(only maxpoints=10000 label)=(phat roc effect);`

The ROC curve (Figure 8.2) indicates that the area under the curve or explained by the model is .64 which is low but acceptable for this demonstration.

Figure 8.3 includes four plots which focus on potential influential observations and they indicate that a few observations may merit further investigation. Again for the sake of demonstration though, the analysis continues with the SURVEYLOGISTIC step.

Figure 8.4 illustrates the predicted probabilities by cholesterol status (high or not) and gender. This plot shows the highest predicted probability of high blood pressure for men with high cholesterol while women without high cholesterol have the lowest predicted probability of high blood pressure.

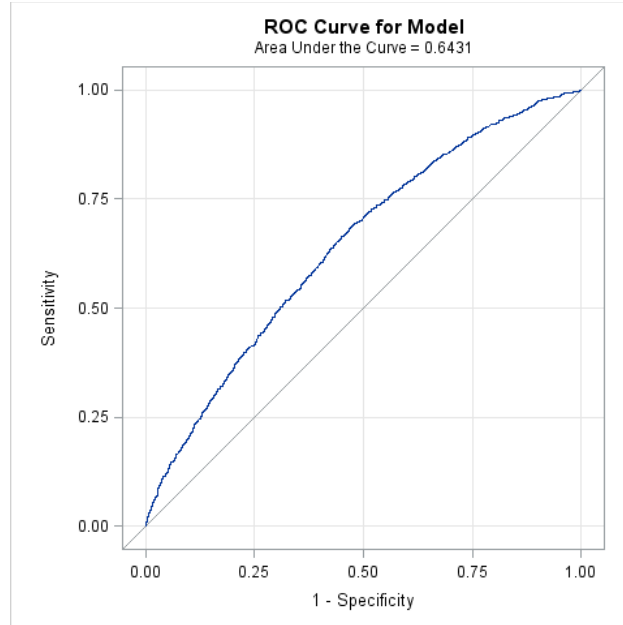


Figure 8.2

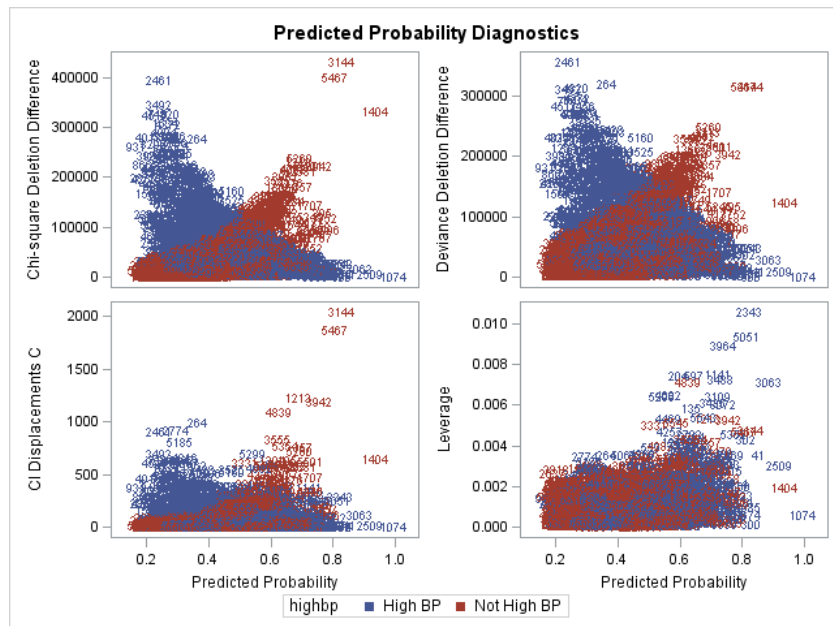


Figure 8.3

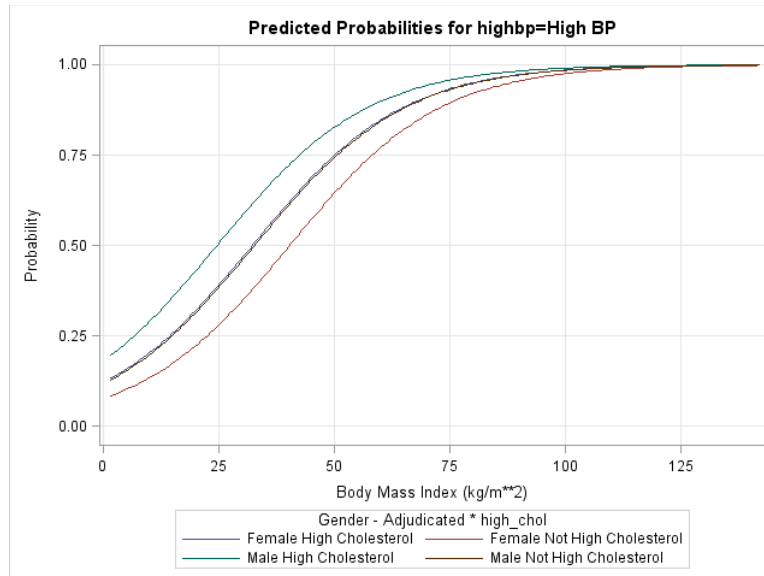


Figure 8.4

EXAMPLE 8 STEP 3 - LOGISTIC REGRESSION USING PROC SURVEYLOGISTIC

Once model fit evaluation is complete, use of PROC SURVEYLOGISTIC for complex sample corrected standard errors and confidence limits is required.

Table 8.2 presents odds ratios and corrected confidence limits (based on the default Taylor Series Linearization method). These statistics are now correctly estimated and can be interpreted and used in an odds ratio plot generated by PROC SGPLOT.

Based on Table 8.2, we see that the odds ratios are all statistically significant and indicate that compared to men, women are about .63 as likely to have high blood pressure, a one unit increase in BMI results in being 1.063 times more likely to have high blood pressure, and those with high cholesterol are about 1.65 times more likely, relative to those without high cholesterol, to suffer from high blood pressure, each while holding all other predictors constant.

```
proc surveylogistic data=nhanes0506 ;
  ods output oddsratios=domainors ;
  strata sdmvstra ; cluster sdmvpsu ; weight wtmec2yr ; domain agel8p ;
  class riagendr high_chol (ref='Not High Cholesterol') ;
  model highbp (event='High BP') = bmx bmi riagendr high_chol ;
  format highbp bp. riagendr sex. high_chol hc. ;
run ;
```

Variance Estimation	
Method	Taylor Series
Variance Adjustment	Degrees of Freedom (DF)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
BMXBMI	1.063	1.053	1.074
RIAGENDR Female vs Male	0.626	0.520	0.754
high_chol High Cholesterol vs Not High Cholesterol	1.648	1.334	2.036

Table 8.2

Odds Ratios and confidence limits are presented in Figure 8.5. This figure uses pre-computed data from PROC SURVEYLOGISTIC with PROC SGPLOT. The `band` statement sets up the confidence bands around the odds ratios and the `series` statement defines the y and x axis and requests data labels for the series points. The output data set "domainors1" contains the domain analysis results from the SURVEYLOGISTIC procedure but has been reduced to just adults for this demonstration (code not shown here). The figure re-iterates the results from Table 8.2 but is perhaps more clearly understood when presented visually.

```
proc sgplot data=domainors1 ;
  where age18p=1 ;
  band x=effect_short lower=lowercl upper=uppercl / outline
  lineattrs=(color=red) fill legendlabel="95% Confidence Limits" ;
  series y=oddsratioest x=effect_short / datalabel ;
run ;
```

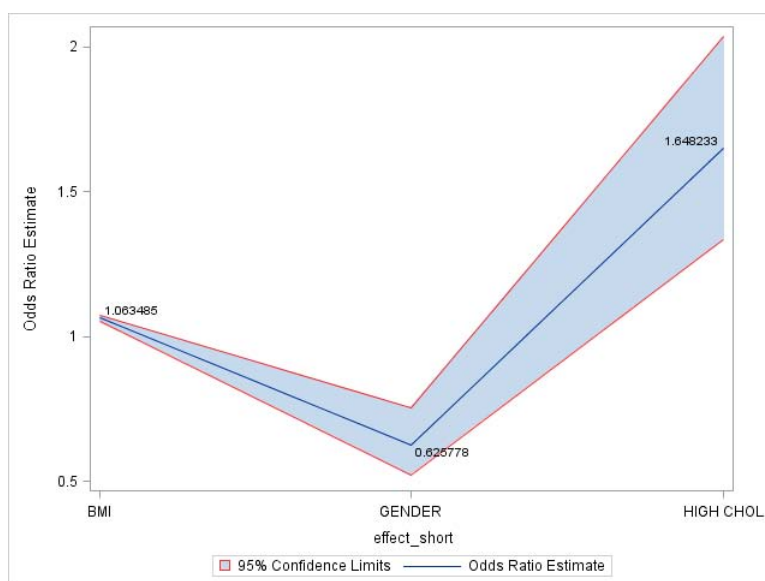


Figure 8.5

CONCLUSION

ODS Graphics and SG procedures offer very effective tools for enhanced data analysis and when used appropriately, can assist the analyst in interpretation of results and eliminate the need for time-consuming coding. This paper has presented a number of applications using complex sample design data but the concepts and approaches are applicable to analysis of simple random sample data as well.

REFERENCES

- Berglund, P. (2008) "Getting the Most out of the SAS® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS® v9.2", SAS Global Forum 2008.
- Heeringa, S. (1996) "National Comorbidity Survey (NCS): Procedures for Sampling Error Estimation".
- Heeringa, S., West, B.T., Berglund, P.A, "Applied Survey Data Analysis", Chapman Hall CRC Press, 2010.
- Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B-E., Walters, E.E., Zaslavsky, A., Zheng, H. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *The International Journal of Methods in Psychiatric Research*, 13(2), 69-92.
- Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Rust, K. (1985). Variance Estimation for Complex Estimation in Sample Surveys. *Journal of Official Statistics*, Vol 1, 381-397. (CP)

CONTACT INFORMATION

Your comments are welcome and can be emailed to Patricia Berglund at pberg@umich.edu.

Patricia Berglund
Institute for Social Research
University of Michigan
426 Thompson St.
Ann Arbor, MI 48106
pberg@umich.edu

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 registered trademarks or trademarks of their respective companies.