Enhanced Data Analysis using SAS® ODS Graphics and Statistical Graphics

Patricia A. Berglund – University of Michigan, Institute for Social Research
Wisconsin and Illinois SAS User's Group
June 25, 2014

Presentation Topics

Overview of Presentation

- Use of ODS Graphics and the Statistical Graphics (SG) procedures for enhanced data analysis, based on paper 349-2012 SAS Global Forum, see support.sas.com to download paper
- How do these tools assist the analytic process?
  - Visual displays can highlight patterns and issues that might not be easily detected with tabular output
  - Tools are either built into procedures (ODS Graphics) or very streamlined (SG Procedures), resulting in little coding effort for analyst
- Applications include graphics and diagnostics for distributional analysis and linear/logistic regressions

Continued…..

Overview of Presentation

- Variety of applications demonstrated:
  - PROC SURVEYFREQ, PROC SURVEYMEANS and PROC UNIVARIATE for distributional analyses
  - PROC REG, PROC SURVEYREG, PROC LOGISTIC, and PROC SURVEYLOGISTIC for regression
- Effective use of selected procedures with ODS Graphics and/or SG procedures along with complex sample design adjustments for correct variance estimation
- All examples are also generalizable to data analysis based on a simple random samples

Introduction to ODS Graphics

- ODS Graphics are built into a number of procedures (SAS/STAT, Base SAS, SAS/OR, etc.)
- ODS Graphics are statistically appropriate for selected analytic technique, that is, they provide correct diagnostics for given analysis
- Added value is that they often reveal patterns and issues in the data that might be difficult to detect from tabular output
- No coding needed other than to request desired graphic from procedure, can be customized using options

Introduction to Statistical Graphics Procedures

- The SG set of procedures are newer alternatives to the SAS/GRAPH module
- SG procedures are currently part of Base SAS and work well with ODS Graphics as they are closely related and share syntax for some procedures
- Coding to produce high-resolution graphics has been simplified
- Applications in this presentation use PROC SGPLOT but other SG procedures such as PROC SGSCATTER and PROC SGPANEL are available

Data Sets Used in Analysis Applications

- The National Health and Nutrition Examination Survey (NHANES, 2005-2006) and the National Comorbidity Survey-Replication (NCS-R, 2001-2002) are used in applications
- Both are based on a complex sample design, therefore all variance estimates or standard errors should be calculated using an appropriate SAS SURVEY procedure
- Though the data sets require complex sample variance adjustments:
  - ODS Graphics/SG tools can still be used along with variance estimates that are design-based
  - Applications utilize ODS Graphics or SG procedures for diagnostic tasks while correctly estimating variance with the SURVEY procedures and graphing pre-computed results for plots/charts with variance estimates
  - Analysts using SRS data can skip the variance estimation in SURVEY procedures and use graphing tools within standard SAS procedures (PROC MEANS, PROC FREQ, PROC REG, PROC LOGISTIC, PROC PHREG, and so on)

Outline of Applications-Distributional Analyses

- Frequency Tables
  - » Frequency Plots, Odds Ratio Plots
  - » PROC SURVEYFREQ, PROC FREQ, both with ODS Graphics
- Means Analysis
  - » Mean Plots with Confidence Limits and Domain Analysis
  - » PROC SURVEYMEANS, PROC SGPLOT
- Univariate Analysis (demonstrated in paper, not shown here)
  - » Univariate Distributions with weighted and un-weighted histograms
  - » PROC UNIVARIATE w/ODS Graphics

Continued…

Outline of Applications-Regression Analyses

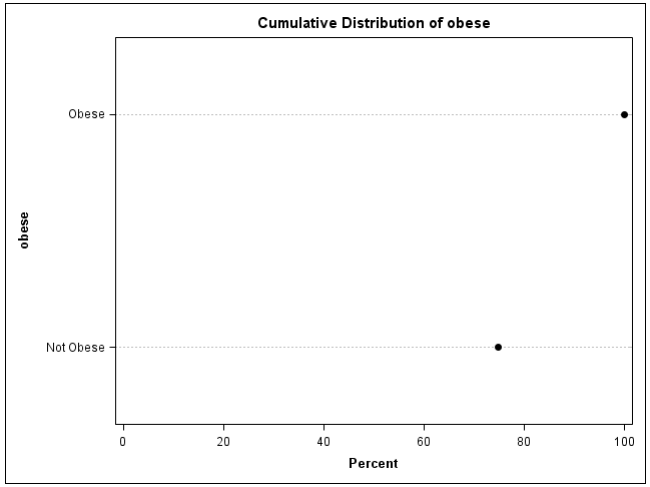
- Linear and Logistic Regression
  - » ODS Graphics are statistically appropriate for the selected analytic technique
  - » PROC REG/SURVEYREG
  - » PROC LOGISTIC/SURVEYLOGISTIC
- ODS Graphics and PROC SGPLOT used to enhance overall understanding and evaluation of model performance
- For any analysis/graphic using variance-related statistics, use of correct SURVEY procedure is demonstrated with subsequent graphing using pre-computed data set

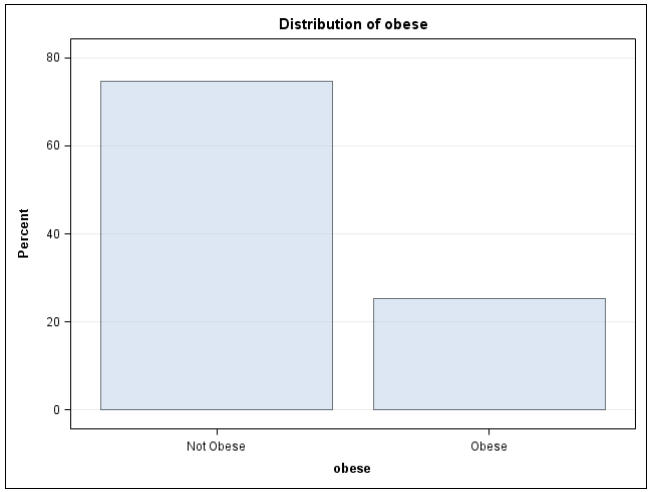
Distributions of Categorical and Continuous Variables

Frequency Table with ODS Graphics-Obese Indicator

- All distributional analyses use NCS-R data with a focus on demographics and body weight (obese indicator, and 6 category obese variable, continuous BMI)

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



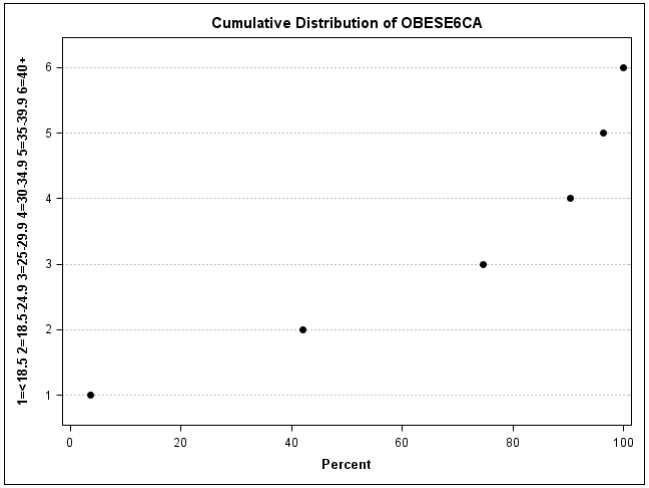


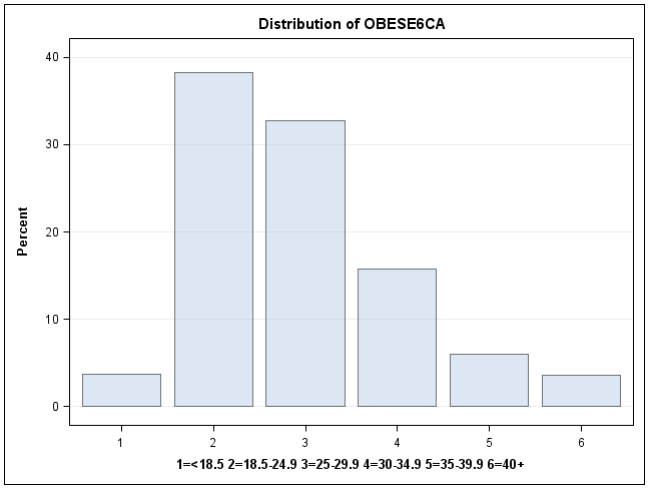
The FREQ Procedure

obese	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Obese	4252.384	74.71	4252.384	74.71
Obese	1439.616	25.29	5692	100.00

Frequency Table with ODS Graphics-Obesity Status

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





Frequency Table with Odds Ratio Plot

- Use of PROC SURVEYFREQ with an implied domain (sub-group) of age in groups by sex*obese indicator with OR Plot and CI's, output below is for age group 60+ only
- Confidence limits use a complex sample design adjusted SE

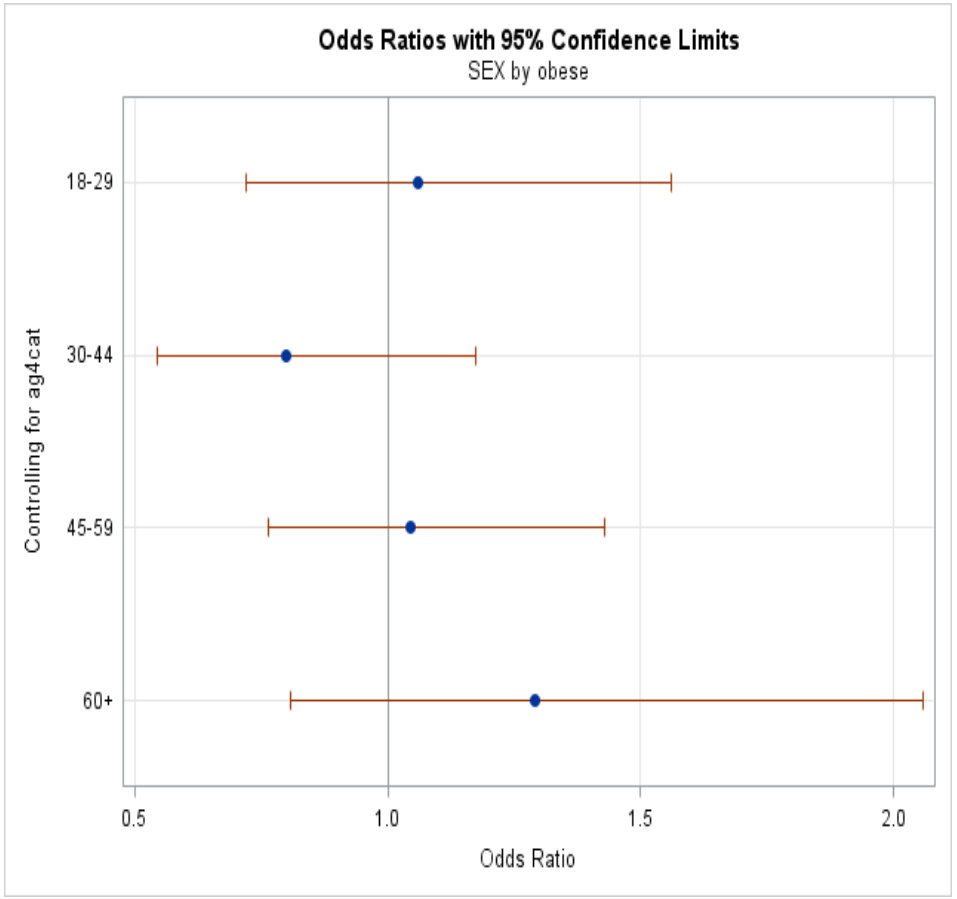
```
proc surveyfreq data=ncsr order=formatted ;
   strata sestrat ; cluster seclustr ; weight ncsrwtlg ;
   tables ag4cat*sex*obese
    / plots(only)=oddsratioplot relrisk ;
   format  ag4cat af. sex sexfr.
   obese obr. mde mdef.  ;
run ;
```

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	

Frequency Table with Odds Ratio Plot



- ODS Graphics available in PROC SURVEYFREQ, provides OR plot with design-based confidence limits
- Plot shows comparison of female to male considered obese (BMI > 30)
- Women are more likely to be obese, as compared to men, in each age group except 30-44 yrs of age, none are significant at .05 alpha level, CI's include 1.0

Means Analysis of Body Mass Index

- For a means analysis of BMI (continuous variable) by US region:
  - Use of PROC SURVEYMEANS to obtain design-based standard errors
  - ODS output statement creates an output data set with correct standard errors for use in plot of BMI by 4 regions

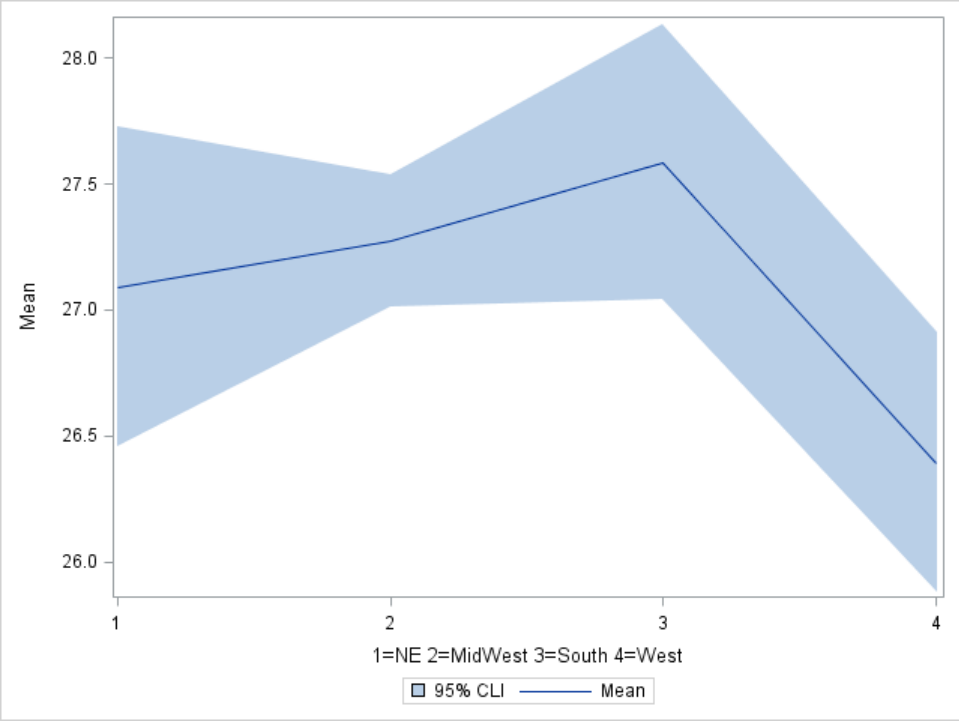
```
proc surveymeans data=ncsr;
  strata sestrat ;
  cluster seclustr ;
  weight ncsrwtlg ;
  var bmi ;
  domain region ;
  ods output domain=outstat ;
run ;
```

Means Analysis of Body Mass Index

- Read output data set from PROC SURVEYMEANS into PROC SGPLOT

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

Bands in plot show how CI's differ among the four regions.



Means Analysis of Body Mass Index By Age and Gender

```
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 ;
```

The SURVEYMEANS Procedure

Domain Statistics in ag4cat*SEX

1=17-29 2=30-44 3=45-59 4=60+	1=Male 2=Female	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
18-29	Male	bmi	Body Mass Index	508	25.889469	0.269224	25.3461527	26.4327859
	Female	bmi	Body Mass Index	725	25.187035	0.322050	24.5371110	25.8369591
30-44	Male	bmi	Body Mass Index	667	28.394008	0.356675	27.6742090	29.1138065
	Female	bmi	Body Mass Index	976	26.938839	0.376146	26.1797459	27.6979325
45-59	Male	bmi	Body Mass Index	542	28.298554	0.298454	27.6962483	28.9008589
	Female	bmi	Body Mass Index	812	28.065448	0.244743	27.5715336	28.5593579
60+	Male	bmi	Body Mass Index	319	26.902143	0.408545	26.0776663	27.7266202
	Female	bmi	Body Mass Index	550	27.267585	0.324843	26.6120265	27.9231445

Continued….

Listing of Outstat Data Set

- Code to produce list report of output data set to be used in next analysis

```
proc print data=outstat noobs ;
  var ag4cat sex varname mean stderr lowerclmean upperclmean ;
run ;
```

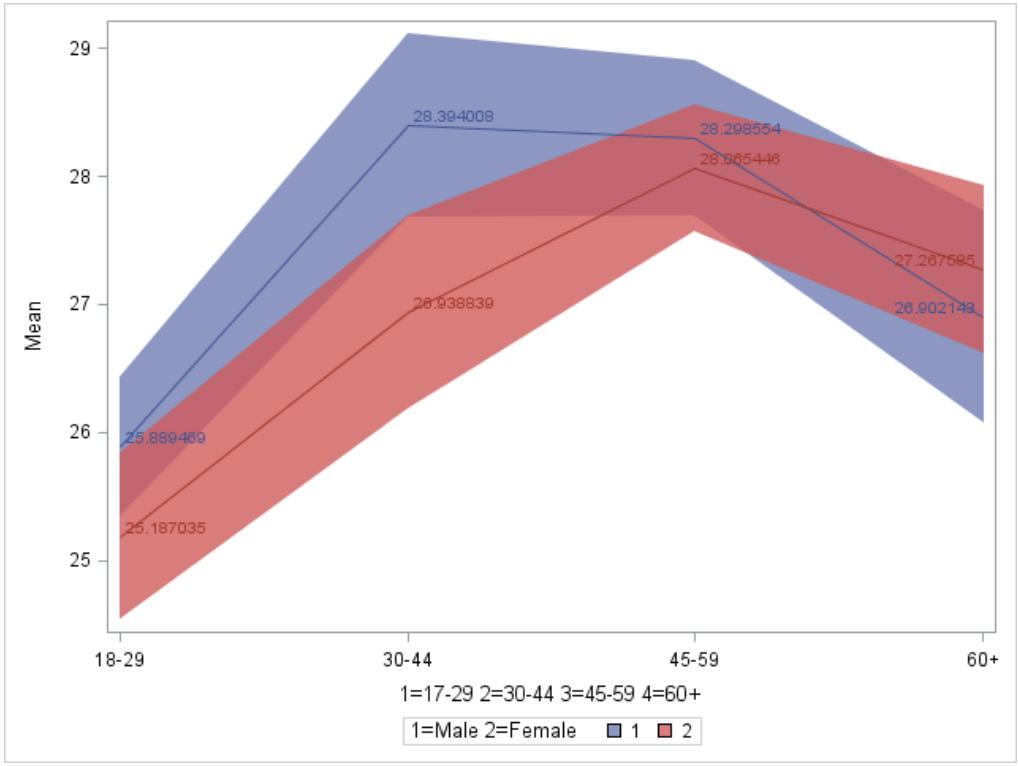
ag4cat	SEX	VarName	Mean	StdErr	LowerCLMean	UpperCLMean
18-29	Male	bmi	25.889469	0.269224	25.3461527	26.4327859
18-29	Female	bmi	25.187035	0.322050	24.5371110	25.8369591
30-44	Male	bmi	28.394008	0.356675	27.6742090	29.1138065
30-44	Female	bmi	26.938839	0.376146	26.1797459	27.6979325
45-59	Male	bmi	28.298554	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

Means Analysis of Body Mass Index with Domains

- Pre-computed data is used in PROC SGPLOT

```
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 ;
```

Plot shows how gender differs across the age groups with intersection at about age 50.



Regression Examples

Linear Regression with ODS Graphics Diagnostics Plots

- Regression examples use NHANES 2005-2006 data with selected MEC measurement variables:
  - total cholesterol and blood pressure (diastolic/systolic)
- Simple linear model regresses total cholesterol on BMI, among US adults (age 18+ is the domain indicator)
- Initial evaluation of model fit uses of PROC REG with the diagnostics panel of plots produced by ODS GRAPHICS
- Final model estimation done w/PROC SURVEYREG to adjust for complex sample design

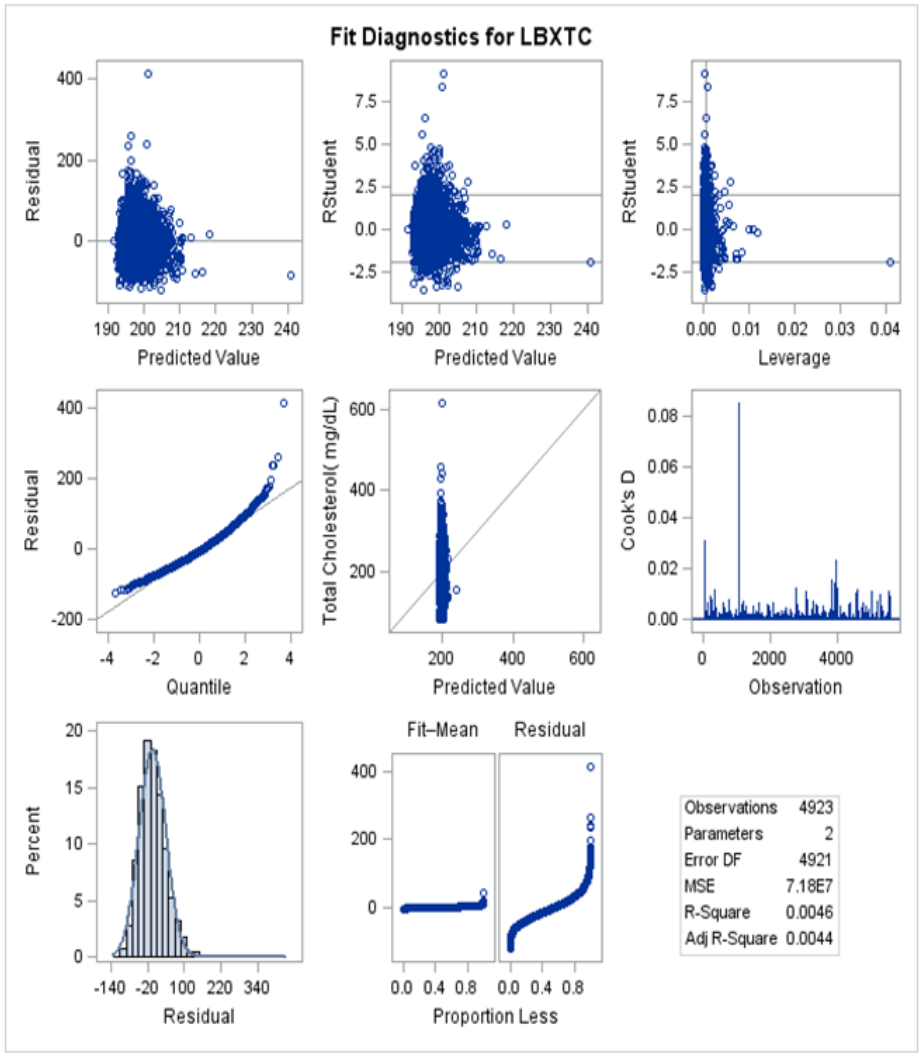
Linear Regression with ODS Graphics Diagnostic Plots

- Code below runs a linear regression and requests diagnostics available from PROC REG (ODS GRAPHICS is enabled)

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

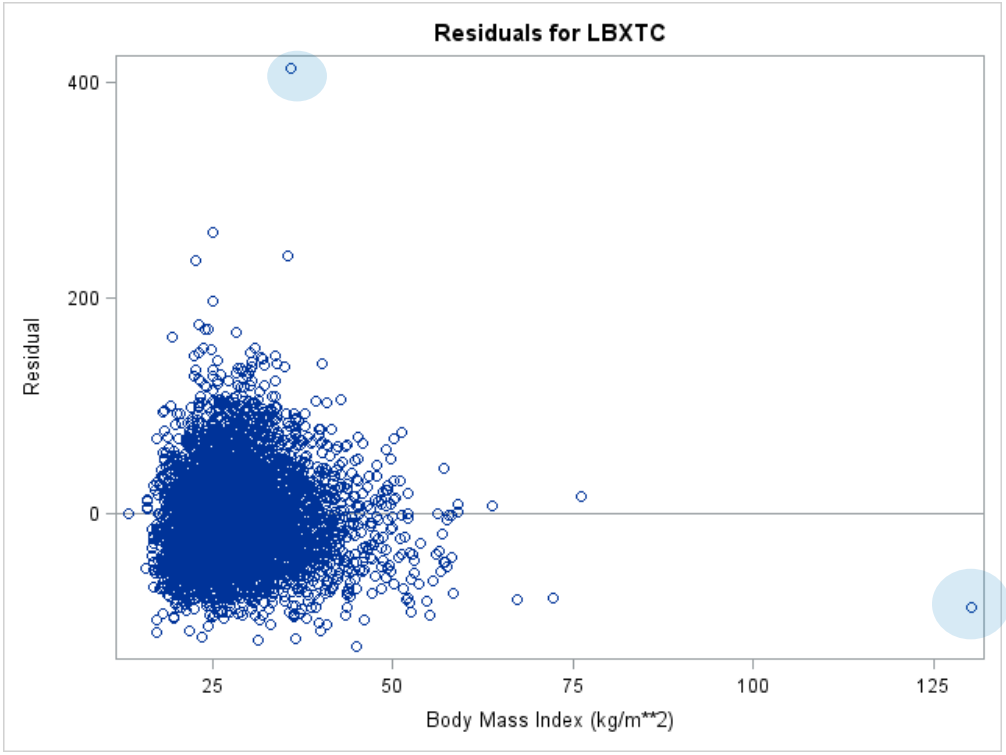
- Options:
  - » plots (maxpoints=10000)=(diagnostics) ;
    - » request default diagnostic plots, maxpoints=10000 increases maximum number of data points used in analysis, see panel on slide 23
    - » Plots: residuals for Total Cholesterol by BMI, outlier and leverage plot, see plots on slides 24 and 25

Linear Regression Diagnostics Panel



- Default Fit Diagnostics Panel includes 8 plots and model summary statistics
- Indication of data points that merit investigation in the plots on rows 1 and 2
- Potential influential points that may affect model fit
- Note the points at extremes on either or both Y and X axis
- Examine distribution of residuals for evidence of normal distribution or violations

Residual Plot of Total Cholesterol by BMI

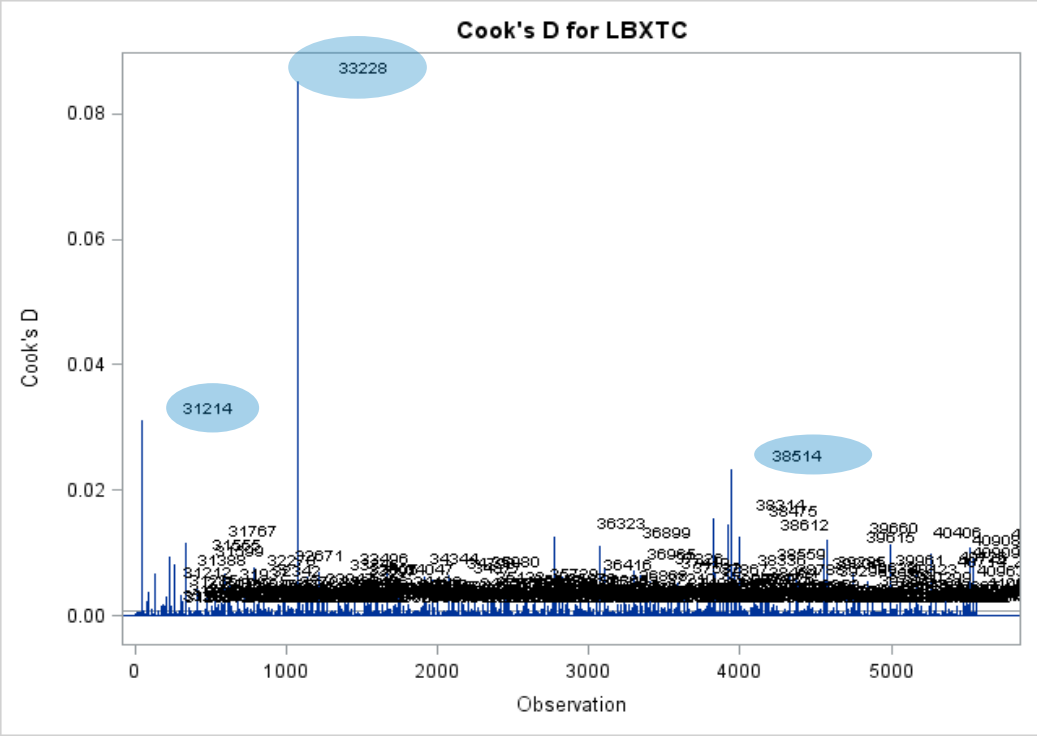


- The residual plot of total cholesterol by BMI indicates a few points with either high residual values or high BMI values, spur investigation of leverage/influence
- One strategy is to identify the points by observation or case ID and examine what happens when they are removed from the model

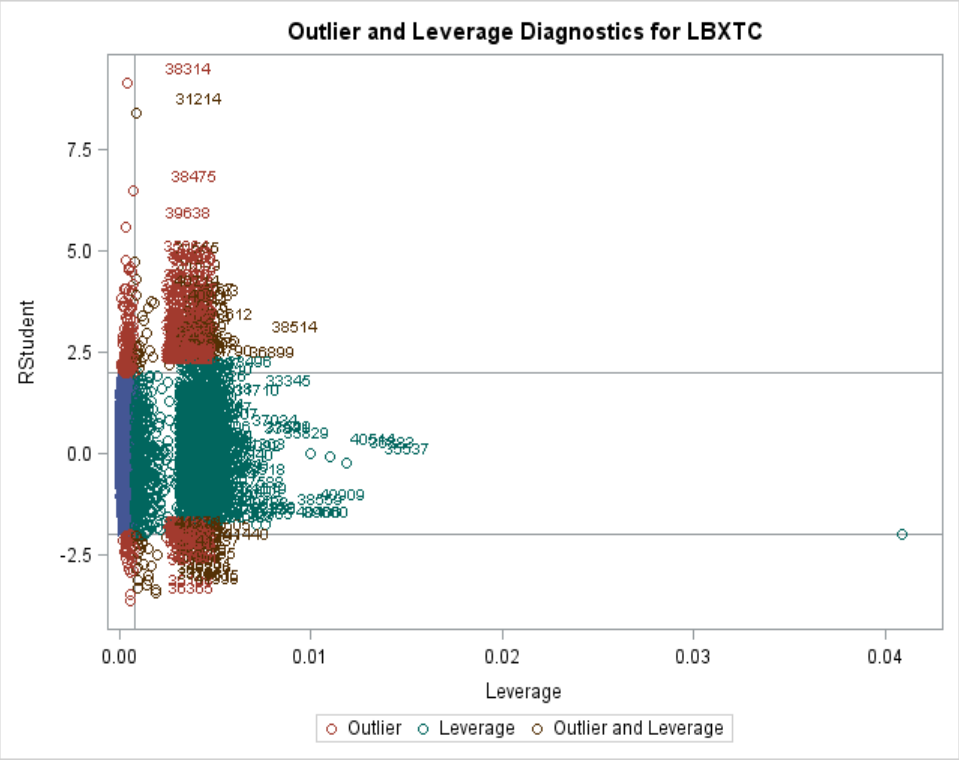
Identification of Influential Data Points

- The code below add labels using id=seqn (Case ID) to the plots, allows easy identification of cases
- Produces Cook's D (this slide) and leverage plot on next slide
- Note high value for sequence numbers in blue ellipses

```
proc reg data=d1.chapter_exercises_nhanes0506
  plots(only label   maxpoints=10000)=(cooksd RStudentByLeverage );
  weight wtmec2yr ;  id seqn ;  where age18p=1 ;
  model lbxtc = bmxbmi  ;
run ;
```



Outlier and Leverage Diagnostics for Total Cholesterol

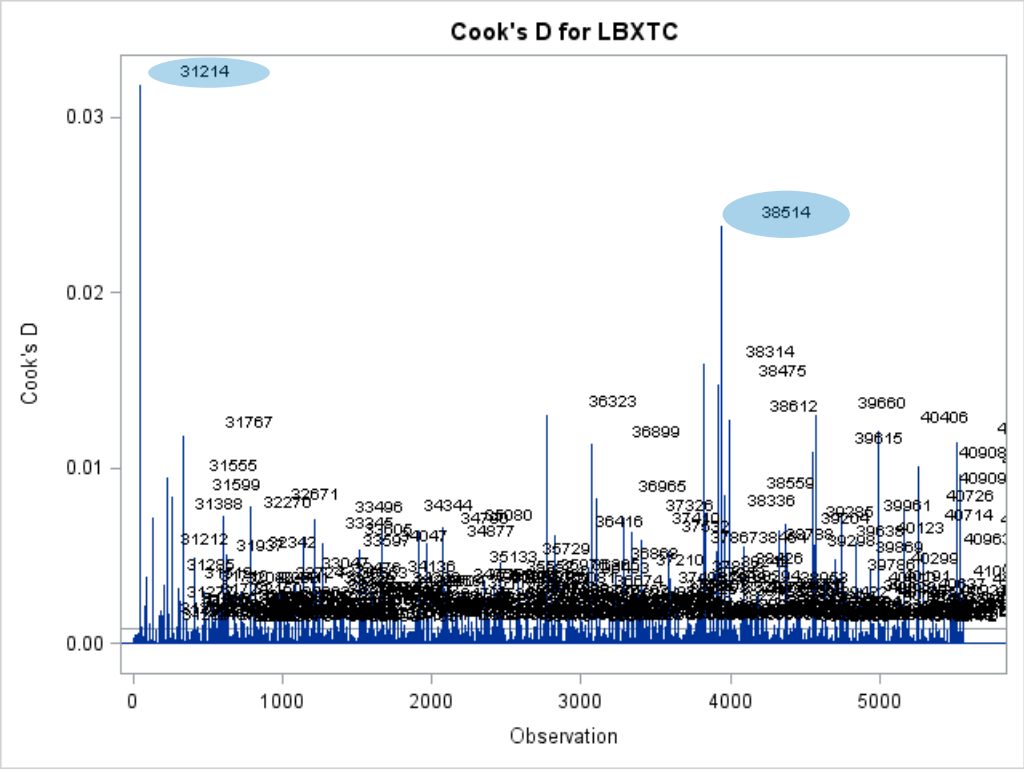


- This plot shows both outliers and leverage for total cholesterol
- Outliers are in red, points with high leverage are in green and points with both are in brick red
- Evidence of a number of data points with issues (outlier, leverage or both)

Regression with Influential Data Point Removed

```
proc reg data=d1.chapter_exercises_nhanes0506
      plots(only label maxpoints=10000)=(cooksd);
      weight wtmec2yr ;   id seqn ;
      where seqn ne 33228 and age18p=1 ;
      model lbxtc = bmxbmi  ;
run ;
```

The circled 2 IDs still show as having high values of Cook's D. They are also candidates for possible removal.



Final Model with PROC SURVEYREG

- Assuming all data cleaning and model fitting is complete (not shown here), final model is run in PROC SURVEYREG to correctly estimate variances (NHANES is based on a complex sample design)
- Recall that diagnostic tools presented in previous steps do not use variance estimates so these can be safely used for basic model fit evaluation with complex sample design data
- SURVEYREG code uses a DOMAIN statement for correct unconditional analysis to focus on the age 18+ domain but remove one ID with a WHERE statement (this is correctly done as it removes only one record)

```
proc surveyreg data=d1.chapter_exercises_nhanes0506 ;
   strata sdmvstra ; cluster sdmvpsu ; weight wtmec2yr ;
   model lbxtc = bmxbmi  ;
   where seqn ne 33228 ;
   domain age18p;
run ;
```

Final Model Estimation with PROC SURVEYREG

- The estimated regression coefficients from PROC SURVEYREG indicate that among those 18 and older:
  - A one unit (point) increase in BMI results in about a half point higher total cholesterol rating
  - This is a significant result at the .05 alpha level (p value <.0001)
  - Obviously, this model would benefit from additional covariates but use of model diagnostics from PROC REG along with correct variance estimates from PROC SURVEYREG are good options for enhanced analysis

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.

Logistic Regression with Diagnostics

- NHANES data is again used to model a binary outcome of High Blood Pressure (1=yes, defined as diastolic BP > 80 or systolic BP > 125, 0=no) predicted by BMI, gender, and an indicator of high cholesterol (total cholesterol > 200)
- Logistic regression diagnostic plots are obtained from PROC LOGISTIC with ODS GRAPHICS for model evaluation
- Final model is run using PROC SURVEYLOGISTIC and an Odds Ratio Plot is generated with confidence limits adjusted for the complex sample design using PROC SGPLOT
- Prior to modeling, a frequency table with ODS graphic is produced

Frequency Table of High Blood Pressure

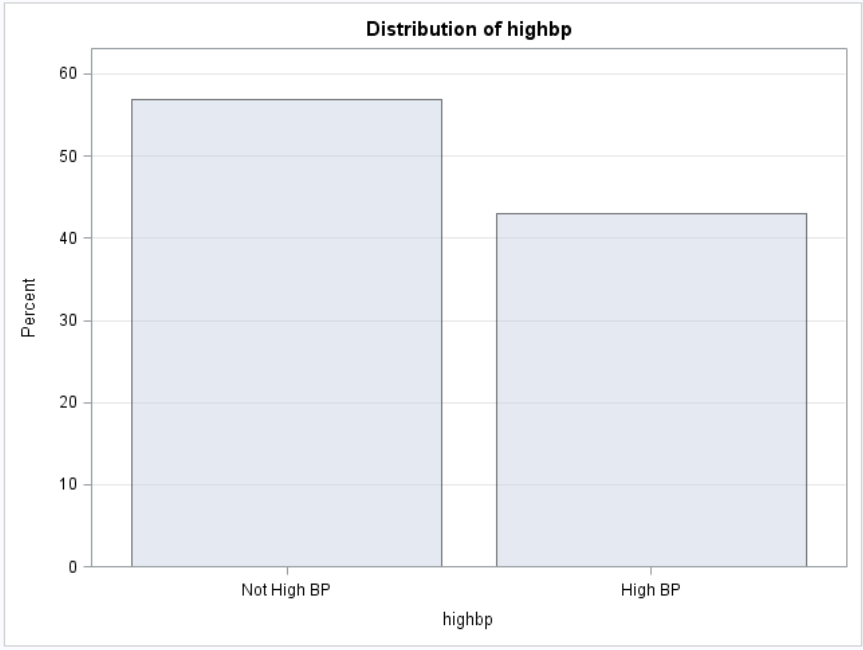
```
proc freq data=nhanes0506 ;
  tables highbp
      plots=(freqplot(scale=percent)) ;
  weight wtmec2yr ;
  format highbp bp. ;
  where age18p=1 ;
run ;
```

- About 43% of US adults (Age 18+) have high blood pressure (1=Yes, 0=No in graph to the right)
- What predicts a high blood pressure diagnosis?

The FREQ Procedure

highbp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not High BP	1.0878E8	56.94	1.0878E8	56.94
High BP	82268379	43.06	1.9105E8	100.00

Frequency Missing = 26654447.872



Logistic Regression with ODS Graphics

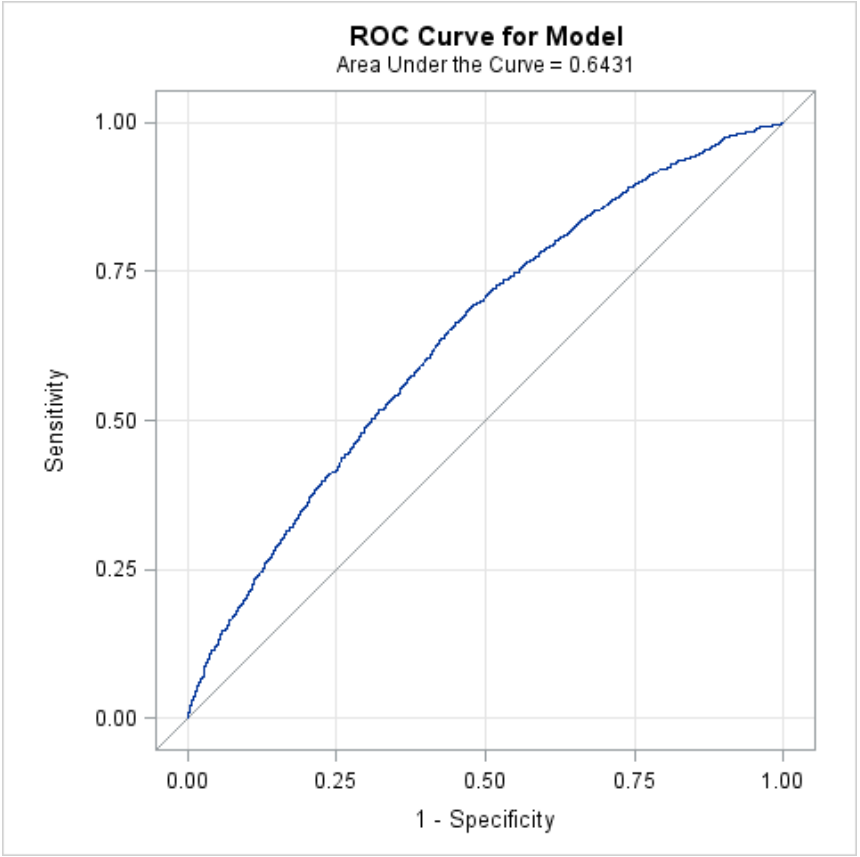
```
proc logistic data=nhanes0506  plots(only maxpoints=10000 label)=(phat roc effect)  ;
  weight wtmec2yr ;  where age18p=1 ;
  class riagendr high_chol  (ref='Not High Cholesterol')  ;
  model highbp (event='High BP') = bmxbmi riagendr high_chol / lackfit ;
  format highbp bp. riagendr sex. high_chol hc. ;
run ;
```

- Code highlighted in red shows plot requests of the following ODS Graphics Plots:
  - Roc (Receiver Operating Curve, slide 33)
  - Phat (Predicted Probability Diagnostics, slide 34)
  - Effect (Predicted Probabilities by Predictor Variables, slide 35)

- /lackfit on model statement requests Hosmer-Lemeshow lackfit option (not shown here, see paper for output and details)

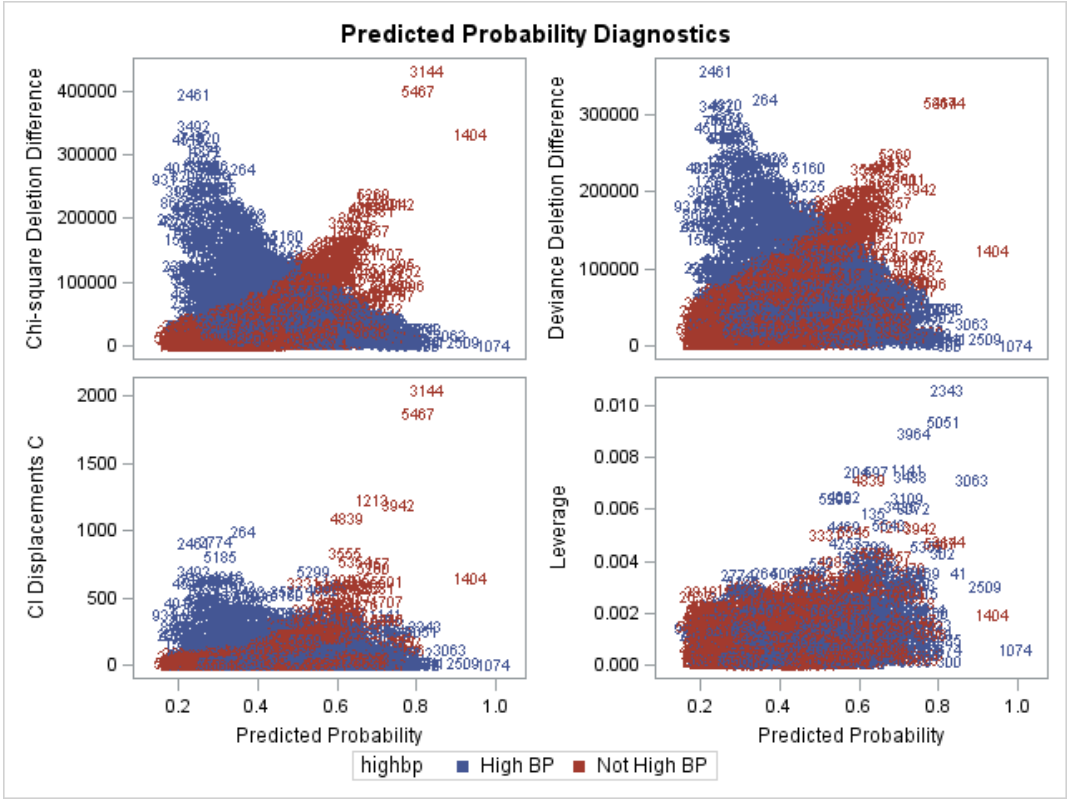
ROC Curve

- The ROC curve plots sensitivity*1-specificity or true positives*false positives
- Area Under the Curve is measured by the “c” or concordance statistic
- The AUC is .64 which is lower than .7 but acceptable for this model/demonstration
- The ROC plot provides a visual tool for model evaluation



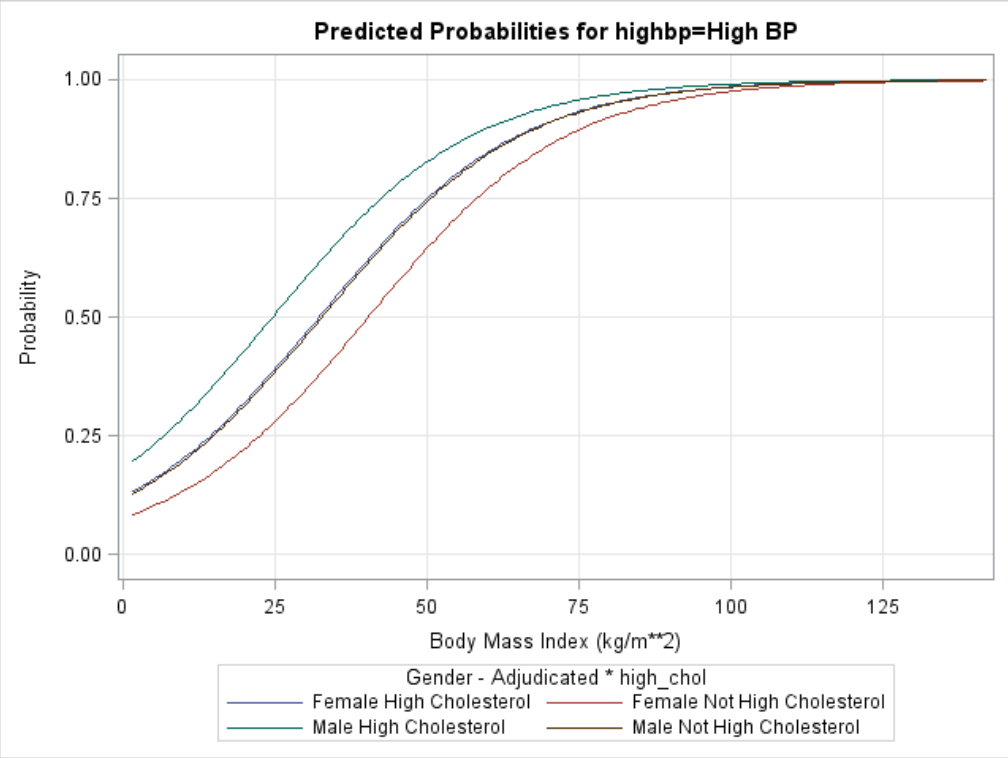
Predicted Probability Diagnostics

- In each of the 4 plots, look for outlying observations, indicating influence on model fit, when a case is removed
- The plots show a few points to consider (obs=3144, 5467) but for the purpose of demonstration the model is used without further data investigation



Predicted Probabilities for High BP

- The predicted probabilities curves are displayed for gender/high cholesterol groups by BMI (continuous on X)
- The plot illustrates the impact of all covariates: BMI, gender and high cholesterol
- The plot shows that males with high cholesterol have the highest predicted probabilities while women w/normal cholestrol are estimated to have the lowest predicted probabilities of high BP



Logistic Regression using PROC SURVEYLOGISTIC

- With model diagnostics complete, use of PROC SURVEYLOGISTIC with an ODS OUTPUT data set is illustrated
- The output data set called "domainors" is then used in PROC SGPLOT to produce a complex sample corrected OR plot with confidence limits

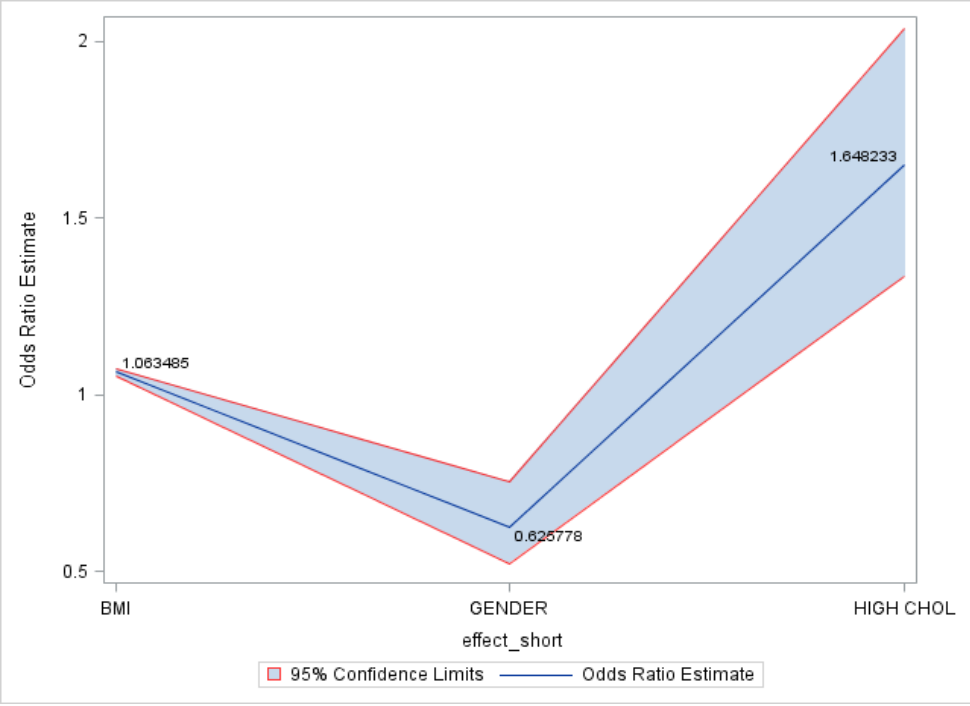
```
proc surveylogistic data=nhanes0506 ;
  ods output oddsratios=domainors ;
  strata sdmvstra ; cluster sdmvpsu ; weight wtmec2yr ;
  domain age18p ;
  class riagendr high_chol (ref='Not High Cholesterol') ;
  model highbp (event='High BP') = bmxbmi riagendr high_chol ;
  format highbp bp. riagendr sex. high_chol hc. ;
run ;
```

Odds Ratio Plot using Pre-Computed Data

effect_short	Effect	OddsRatioEst	LowerCL	UpperCL	age18p	Domain
BMI	BMXBMI	1.086	1.073	1.098	.	
GENDER	RIAGENDR Female vs Male	0.639	0.535	0.762	.	
HIGH CHOL	high_chol High Cholesterol vs Not High Cholesterol	1.930	1.602	2.326	.	
BMI	BMXBMI	1.097	1.054	1.142	0	1=Age >= 18 0=Age < 18=0
GENDER	RIAGENDR Female vs Male	0.731	0.446	1.197	0	1=Age >= 18 0=Age < 18=0
HIGH CHOL	high_chol High Cholesterol vs Not High Cholesterol	1.258	0.558	2.836	0	1=Age >= 18 0=Age < 18=0
BMI	BMXBMI	1.063	1.053	1.074	1	1=Age >= 18 0=Age < 18=1
GENDER	RIAGENDR Female vs Male	0.626	0.520	0.754	1	1=Age >= 18 0=Age < 18=1
HIGH CHOL	high_chol High Cholesterol vs Not High Cholesterol	1.648	1.334	2.036	1	1=Age >= 18 0=Age < 18=1

```
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 ;
```

The plot includes only those 18+ and presents the BMI, gender and high cholesterol odds ratios and confidence limits. This is a visually appealing approach which is easy to understand.



Summary

- Use of ODS Graphics and SG procedures allow the analyst to enhance data analysis through use of a variety of high resolution graphics
- Minimal coding is required for both tools and effective and informative graphics add to the analysis and interpretation of results
- Graphics are often easier to grasp than tabular output and promote diversity in publications or reports
- Tools are appropriate for all types of analysis including SRS and Complex Sample Design data (with modifications for variance estimation)

Thank you for attending!
Your comments are welcome.

Patricia A. Berglund
pberg@umich.edu