Chapter 7

Preparing Data for Multivariate Analysis

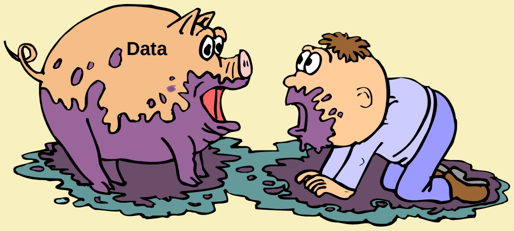
Section 7.1

Screening, Cleaning, and Preparing Data

Objectives

- Understand many of the most common data problems for multivariate analysis and the consequences of these problems.
- Screen for restricted range, small groups, and outliers.
- Clean and prepare data files for multivariate analysis using SAS.

Data Reality…



Things to Do Before You Begin

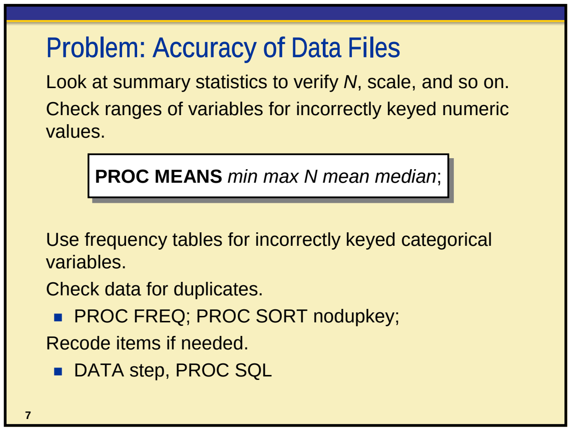
- ✓ Data files accurate?
- ✓ Outliers?
- ✓ Restricted ranges in continuous variables?
- ✓ Unequal cell sizes in categorical variables?
- ✓ Distributions?
- ✓ Collinearity/singularity in variables?
- ✓ Homogeneous covariance matrices?
- ✓ Extent and nature of missing data?

This is just a sampling!

Data Preparation is Key to Success

You should reasonably expect to spend more time cleaning, verifying, screening, and imputing your data than analyzing it. Data analysis is

- 90% perspiration
- 10% analysis
- 100% FUN with SAS!



Problem: Accuracy of Data Files

Look at summary statistics to verify *N*, scale, and so on.

Check ranges of variables for incorrectly keyed numeric values.

```
PROC MEANS min max N mean median;
```

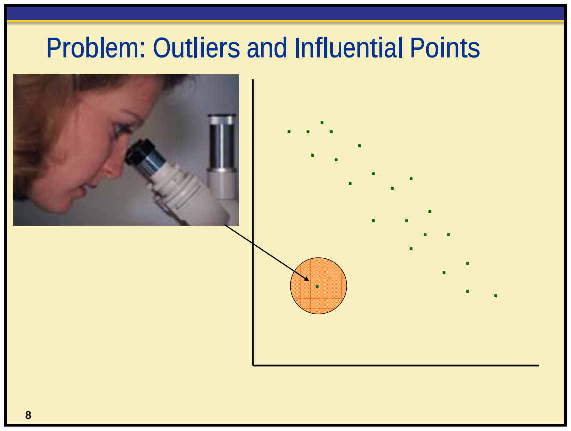
Use frequency tables for incorrectly keyed categorical variables.

Check data for duplicates.

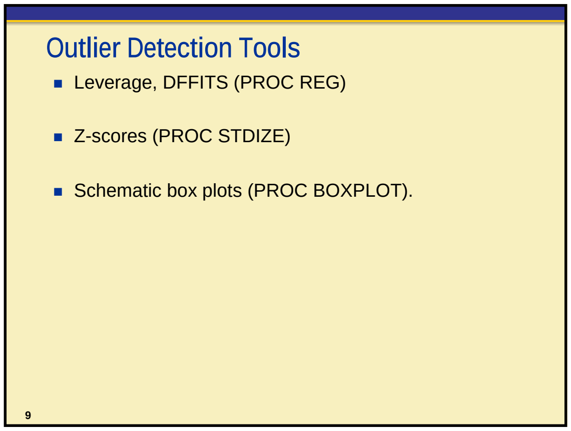
- PROC FREQ; PROC SORT nodupkey;

Recode items if needed.

- DATA step, PROC SQL

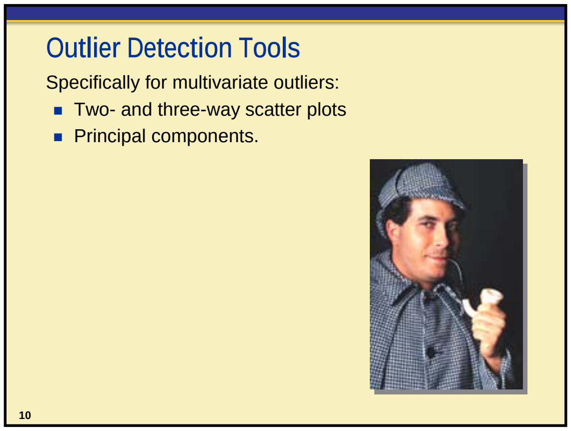


Problem: Outliers and Influential Points



Outlier Detection Tools

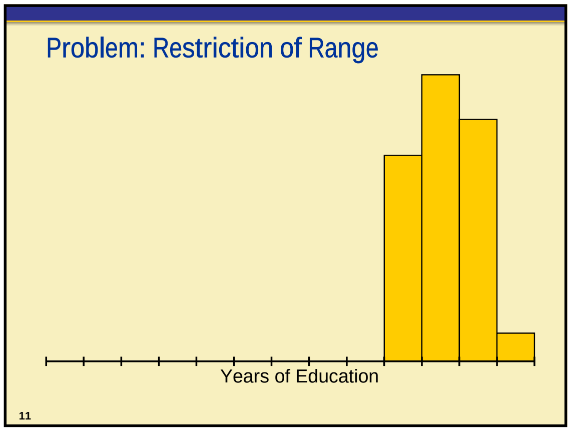
- Leverage, DFFITS (PROC REG)
- Z-scores (PROC STDIZE)
- Schematic box plots (PROC BOXPLOT).



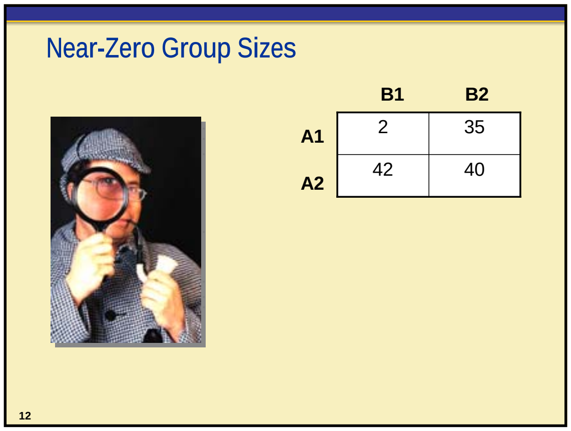
Outlier Detection Tools

Specifically for multivariate outliers:

- Two- and three-way scatter plots
- Principal components.

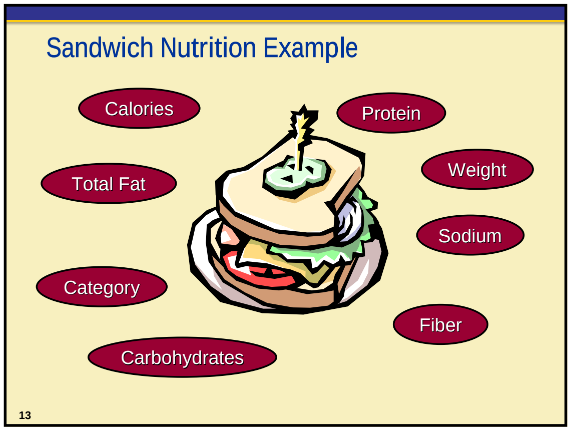


Problem: Restriction of Range



Near-Zero Group Sizes

	B1	B2
A1	2	35
A2	42	40



Sandwich Nutrition Example

Calories

Protein

Total Fat

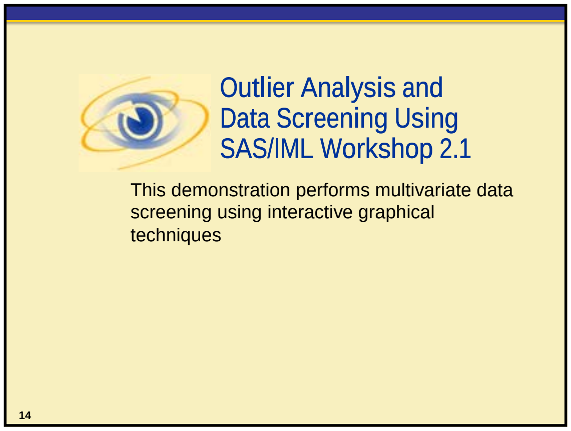
Weight

Sodium

Category

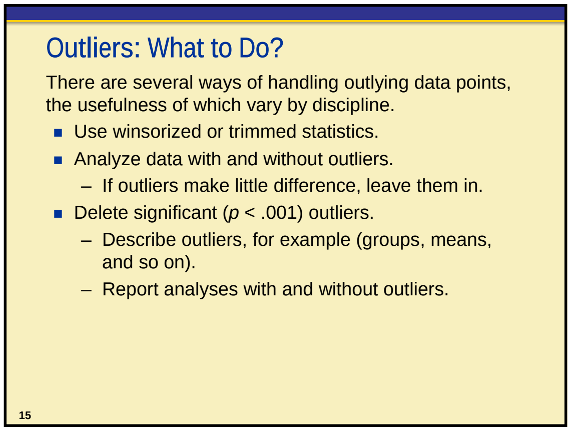
Fiber

Carbohydrates



Outlier Analysis and Data Screening Using SAS/IML Workshop 2.1

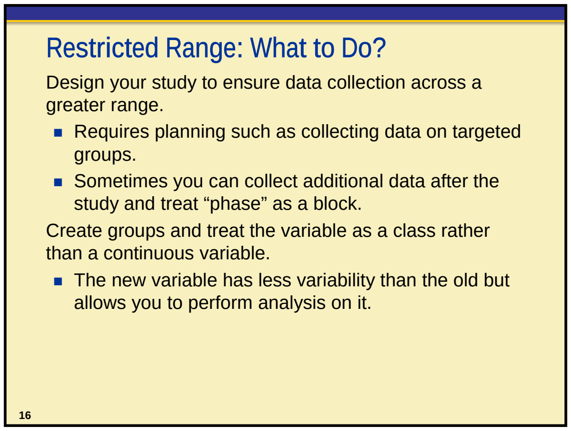
This demonstration performs multivariate data screening using interactive graphical techniques



Outliers: What to Do?

There are several ways of handling outlying data points, the usefulness of which vary by discipline.

- Use winsorized or trimmed statistics.
- Analyze data with and without outliers.
  - If outliers make little difference, leave them in.
- Delete significant (p < .001) outliers.
  - Describe outliers, for example (groups, means, and so on).
  - Report analyses with and without outliers.



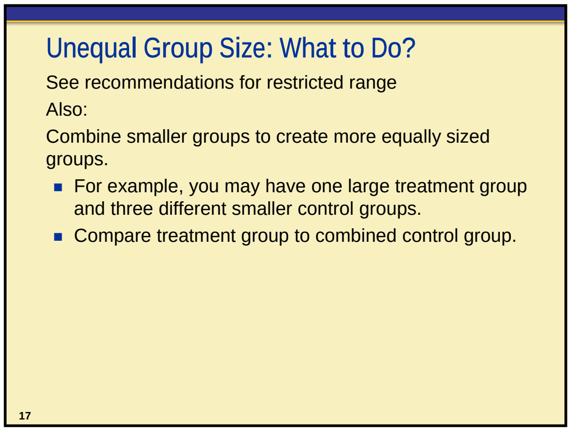
Restricted Range: What to Do?

Design your study to ensure data collection across a greater range.

- Requires planning such as collecting data on targeted groups.
- Sometimes you can collect additional data after the study and treat “phase” as a block.

Create groups and treat the variable as a class rather than a continuous variable.

- The new variable has less variability than the old but allows you to perform analysis on it.



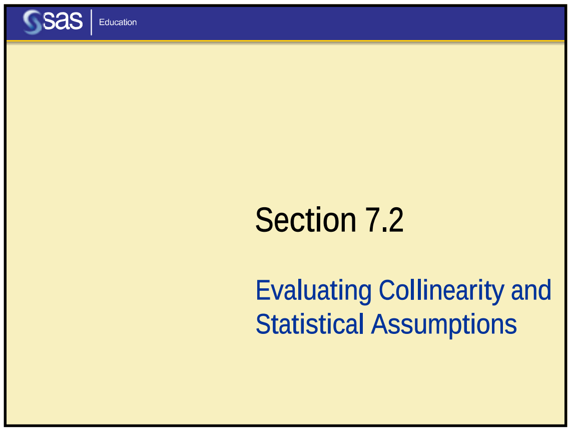
Unequal Group Size: What to Do?

See recommendations for restricted range

Also:

Combine smaller groups to create more equally sized groups.

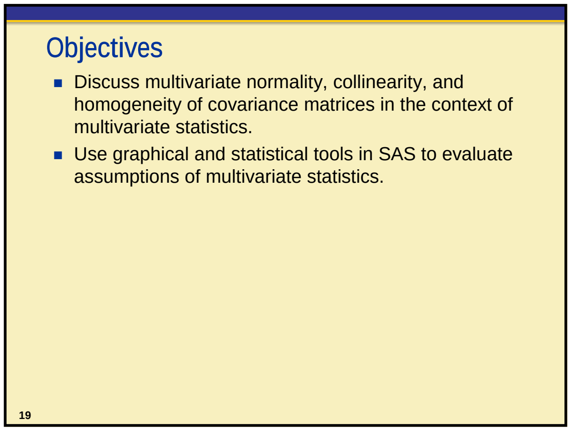
- For example, you may have one large treatment group and three different smaller control groups.
- Compare treatment group to combined control group.



sas Education

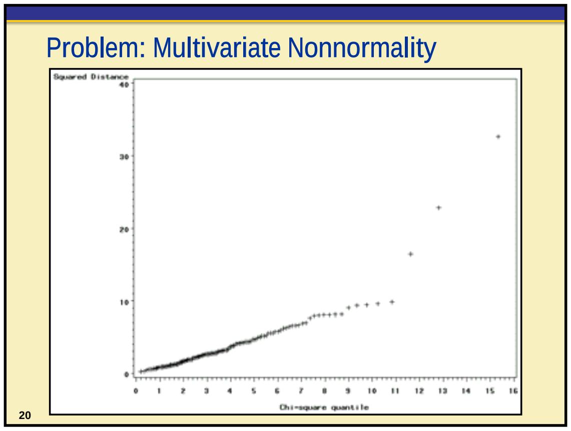
Section 7.2

Evaluating Collinearity and Statistical Assumptions

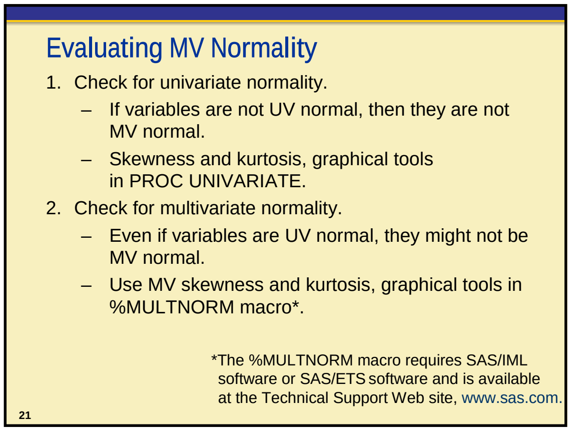


Objectives

- Discuss multivariate normality, collinearity, and homogeneity of covariance matrices in the context of multivariate statistics.
- Use graphical and statistical tools in SAS to evaluate assumptions of multivariate statistics.



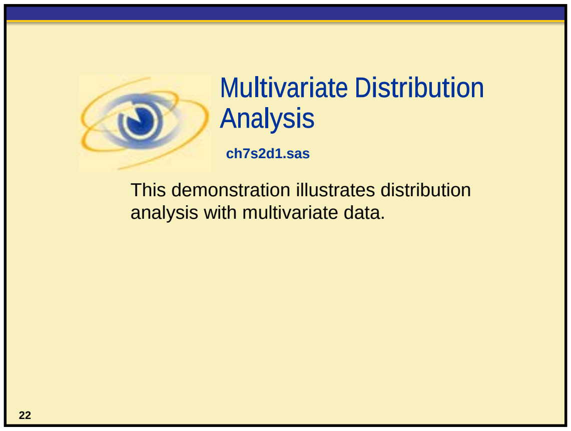
Problem: Multivariate Nonnormality



Evaluating MV Normality

1. Check for univariate normality.
   - If variables are not UV normal, then they are not MV normal.
   - Skewness and kurtosis, graphical tools in PROC UNIVARIATE.
2. Check for multivariate normality.
   - Even if variables are UV normal, they might not be MV normal.
   - Use MV skewness and kurtosis, graphical tools in %MULTNORM macro*.

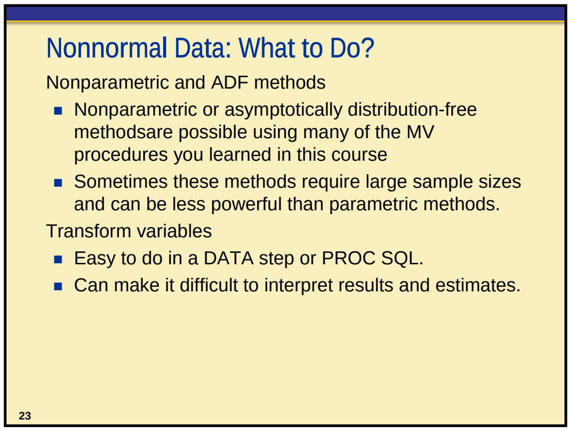
*The %MULTNORM macro requires SAS/IML software or SAS/ETS software and is available at the Technical Support Web site, www.sas.com.



Multivariate Distribution Analysis

ch7s2d1.sas

This demonstration illustrates distribution analysis with multivariate data.



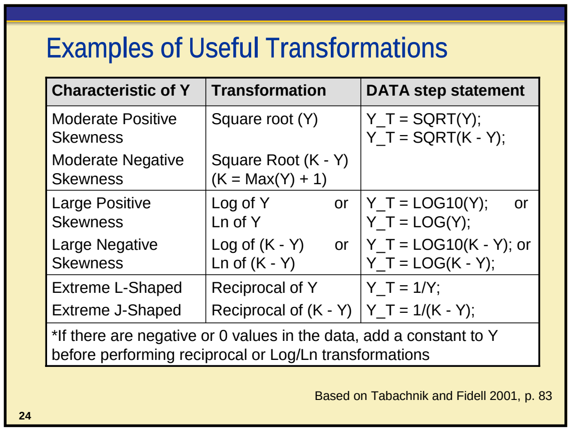
Nonnormal Data: What to Do?

Nonparametric and ADF methods

- Nonparametric or asymptotically distribution-free methodsare possible using many of the MV procedures you learned in this course
- Sometimes these methods require large sample sizes and can be less powerful than parametric methods.

Transform variables

- Easy to do in a DATA step or PROC SQL.
- Can make it difficult to interpret results and estimates.



Examples of Useful Transformations

Characteristic of Y	Transformation	DATA step statement

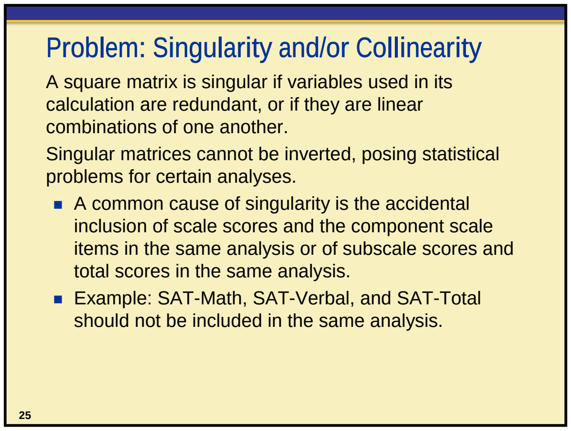
Moderate Positive Skewness
Moderate Negative Skewness
	Square root (Y)
	Square Root (K - Y)
	(K = Max(Y) + 1)
		Y_T = SQRT(Y);
		Y_T = SQRT(K - Y);

Large Positive Skewness
Large Negative Skewness
	Log of Y or
	Ln of Y
	Log of (K - Y) or
	Ln of (K - Y)
		Y_T = LOG10(Y); or
		Y_T = LOG(Y);
		Y_T = LOG10(K - Y); or
		Y_T = LOG(K - Y);

Extreme L-Shaped
Extreme J-Shaped
	Reciprocal of Y
	Reciprocal of (K - Y)
		Y_T = 1/Y;
		Y_T = 1/(K - Y);

*If there are negative or 0 values in the data, add a constant to Y before performing reciprocal or Log/Ln transformations		

Based on Tabachnik and Fidell 2001, p. 83

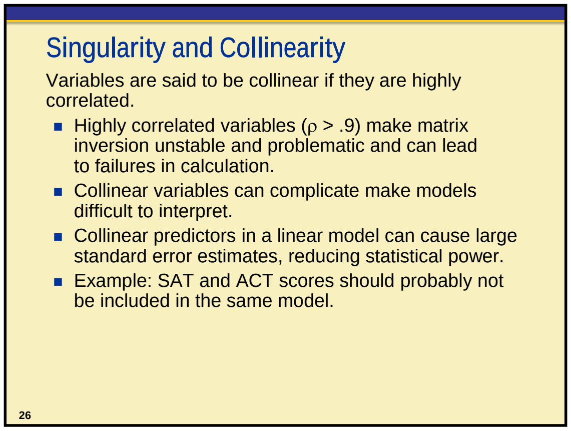


Problem: Singularity and/or Collinearity

A square matrix is singular if variables used in its calculation are redundant, or if they are linear combinations of one another.

Singular matrices cannot be inverted, posing statistical problems for certain analyses.

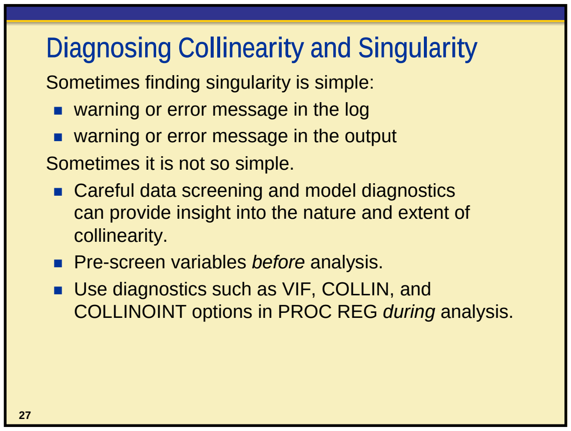
- A common cause of singularity is the accidental inclusion of scale scores and the component scale items in the same analysis or of subscale scores and total scores in the same analysis.
- Example: SAT-Math, SAT-Verbal, and SAT-Total should not be included in the same analysis.



Singularity and Collinearity

Variables are said to be collinear if they are highly correlated.

- Highly correlated variables (ρ > .9) make matrix inversion unstable and problematic and can lead to failures in calculation.
- Collinear variables can complicate make models difficult to interpret.
- Collinear predictors in a linear model can cause large standard error estimates, reducing statistical power.
- Example: SAT and ACT scores should probably not be included in the same model.



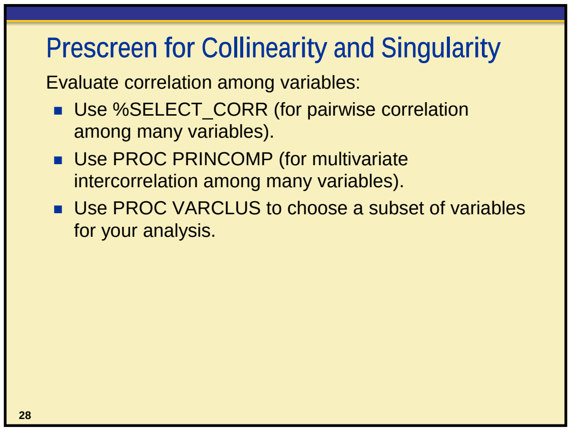
Diagnosing Collinearity and Singularity

Sometimes finding singularity is simple:

- warning or error message in the log
- warning or error message in the output

Sometimes it is not so simple.

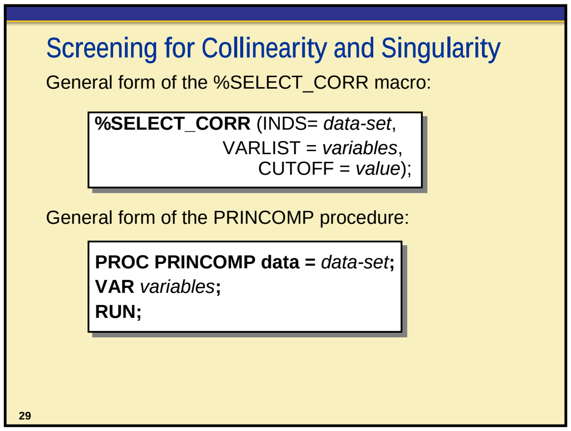
- Careful data screening and model diagnostics can provide insight into the nature and extent of collinearity.
- Pre-screen variables *before* analysis.
- Use diagnostics such as VIF, COLLIN, and COLLINOINT options in PROC REG *during* analysis.



Prescreen for Collinearity and Singularity

Evaluate correlation among variables:

- Use %SELECT_CORR (for pairwise correlation among many variables).
- Use PROC PRINCOMP (for multivariate intercorrelation among many variables).
- Use PROC VARCLUS to choose a subset of variables for your analysis.



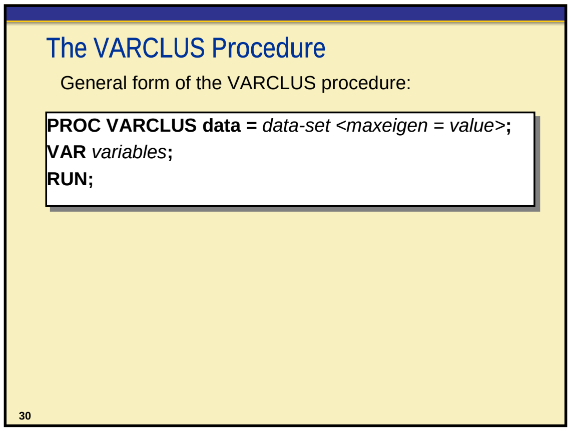
Screening for Collinearity and Singularity

General form of the %SELECT_CORR macro:

```
%SELECT_CORR (INDS= data-set,
                  VARLIST = variables,
                     CUTOFF = value);
```

General form of the PRINCOMP procedure:

```
PROC PRINCOMP data = data-set;
VAR variables;
RUN;
```

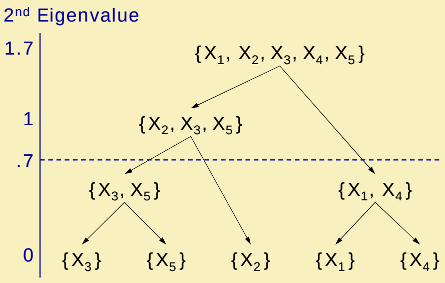


The VARCLUS Procedure

General form of the VARCLUS procedure:

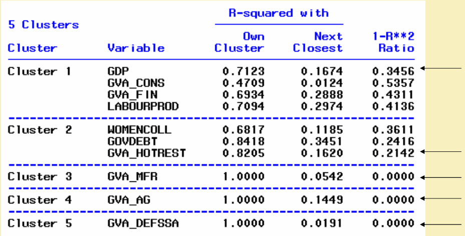
```
PROC VARCLUS data = data-set <maxeigen = value>;
VAR variables;
RUN;
```

Variable Clustering



Cluster Representatives

$$1 - R^2 \text{ ratio } = \frac{1 - R^2_{\text{own cluster}}}{1 - R^2_{\text{next closest}}}$$



```
                                   R-squared with
5 Clusters                        ----------------
                                    Own      Next     1-R**2
Cluster      Variable             Cluster   Closest    Ratio
------------------------------------------------------------
Cluster 1    GDP                   0.7123    0.1674    0.3456
             GVA_CONS              0.4709    0.0124    0.5357
             GVA_FIN               0.6934    0.2888    0.4311
             LABOURPROD            0.7094    0.2974    0.4136
------------------------------------------------------------
Cluster 2    WOMENCOLL             0.6817    0.1185    0.3611
             GOVDEBT               0.8418    0.3451    0.2416
             GVA_HOTREST           0.8205    0.1620    0.2142
------------------------------------------------------------
Cluster 3    GVA_MFR               1.0000    0.0542    0.0000
------------------------------------------------------------
Cluster 4    GVA_AG                1.0000    0.1449    0.0000
------------------------------------------------------------
Cluster 5    GVA_DEFSSA            1.0000    0.0191    0.0000
```

Diagnosing Collinearity or Singularity in Your Data

ch7s2d2.sas

This demonstration illustrates screening data for collinearity and singularity.

Collinearity and Singularity: What to Do?

- If some variables are linear combinations of other variables, do not include all of them in the same analysis.
- If some variables are moderately or highly collinear, consider using only a few of the variables or combining the variables into a single score (VARCLUS or PRINCOMP).

Problem: Heterogeneous Variance-Covariance Matrices

Just as unequal variances can cause problems with both type-I and type-II error in ANOVA, unequal variances and covariances can cause problems in MANOVA.

- If groups are equally sized and covariance matrices are moderately unequal, MANOVA is robust.
- For unequal group sizes and/or large differences in covariance matrices, inference is likely to be incorrect.

Diagnosis of Heterogeneous Variance-Covariance Matrices

General form of the DISCRIM procedure:

```
PROC DISCRIM DATA = data-set-name POOL = test;
        CLASS class-var;
        VAR continuous-vars;
RUN;
```

Note: This test is sensitive to multivariate nonnormality

Evaluating Homogeneity of Covariance Matrices

ch7s2d3.sas

This demonstration illustrates tests for homogeneity of covariance matrices.

Ready to Analyze Some Data!

There are many exciting and fun ways you can use the SAS System for multivariate statistical analysis.

Always remember, your results can only be as good as the data that went into the analysis!

Beware of Famous Last Words…

- “Just delete everything more than three standard deviations from the mean.”
- “We don’t have data quality problems because the data entry people check for errors.”
- “Let’s include scales and items in the same analysis to learn about global and specific characteristics.”
- “You can make any inference you want from a sample of size 30.”
- “Just stick all the variables into a factor analysis and see what you get.”

Check, Explore, Evaluate, Transform, Replace

Remember that data preparation is key to the success of multivariate statistical analysis!

- Check your data for accuracy
- Explore your data for unusual points and patterns
- Evaluate statistical assumptions in your data
- Transform variables as necessary
- Determine the extent and nature of collinearity in your data
- Handle missing values in a reasonable way
- Use graphical and statistical tools for data preparation.

Multivariate Success!