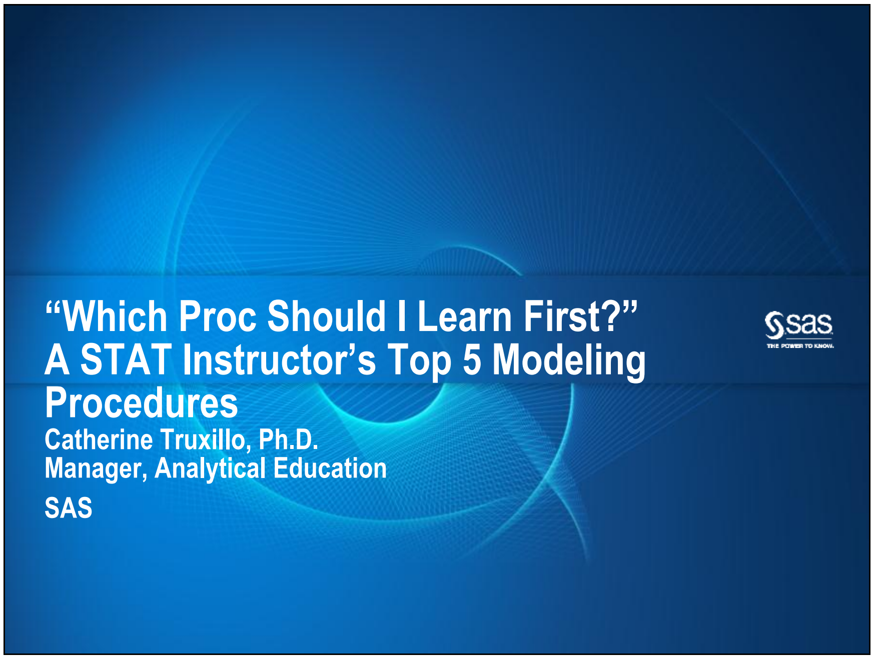


“Which Proc Should I Learn First?” A STAT Instructor’s Top 5 Modeling Procedures

Catherine Truxillo, Ph.D.
Manager, Analytical Education
SAS

The Target Audience for this Talk

I am new to SAS but have used other software for statistical analysis

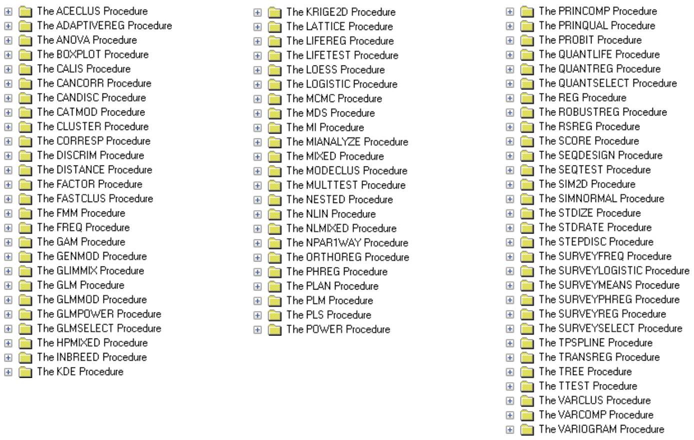
I am an experienced SAS programmer but I have not used many SAS/STAT procedures

I have used the same few SAS/STAT procedures for decades and want to modernize my use of SAS/STAT

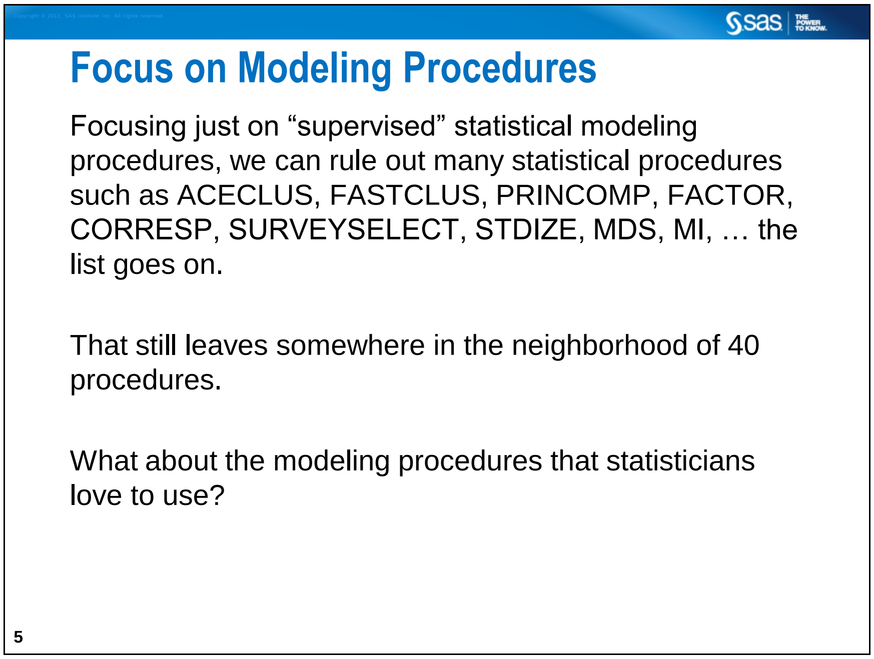
A Common Question

"Can you tell me when to use
each of the procedures
in SAS/STAT?"

SAS/STAT Has A Lot of Procedures



- The ACECLUS Procedure
- The ADAPTIVEREG Procedure
- The ANOVA Procedure
- The BOXPLOT Procedure
- The CALIS Procedure
- The CANCORR Procedure
- The CANDISC Procedure
- The CATMOD Procedure
- The CLUSTER Procedure
- The CORRESP Procedure
- The DISCRIM Procedure
- The DISTANCE Procedure
- The FACTOR Procedure
- The FASTCLUS Procedure
- The FMM Procedure
- The FREQ Procedure
- The GAM Procedure
- The GENMOD Procedure
- The GLIMMIX Procedure
- The GLM Procedure
- The GLMMOD Procedure
- The GLMPOWER Procedure
- The GLMSELECT Procedure
- The HPMIXED Procedure
- The INBREED Procedure
- The KDE Procedure
- The KRIGE2D Procedure
- The LATTICE Procedure
- The LIFEREG Procedure
- The LIFETEST Procedure
- The LOESS Procedure
- The LOGISTIC Procedure
- The MCMC Procedure
- The MDS Procedure
- The MI Procedure
- The MIANALYZE Procedure
- The MIXED Procedure
- The MODECLUS Procedure
- The MULTTEST Procedure
- The NESTED Procedure
- The NLIN Procedure
- The NLMIXED Procedure
- The NPAR1WAY Procedure
- The ORTHOREG Procedure
- The PHREG Procedure
- The PLAN Procedure
- The PLM Procedure
- The PLS Procedure
- The POWER Procedure
- The PRINCOMP Procedure
- The PRINQUAL Procedure
- The PROBIT Procedure
- The QUANTLIFE Procedure
- The QUANTREG Procedure
- The QUANTSELECT Procedure
- The REG Procedure
- The ROBUSTREG Procedure
- The RSREG Procedure
- The SCORE Procedure
- The SEQDESIGN Procedure
- The SEQTEST Procedure
- The SIM2D Procedure
- The SIMNORMAL Procedure
- The STDIZE Procedure
- The STDRATE Procedure
- The STEPDISC Procedure
- The SURVEYFREQ Procedure
- The SURVEYLOGISTIC Procedure
- The SURVEYMEANS Procedure
- The SURVEYPHREG Procedure
- The SURVEYREG Procedure
- The SURVEYSELECT Procedure
- The TPSPLINE Procedure
- The TRANSREG Procedure
- The TREE Procedure
- The TTEST Procedure
- The VARCLUS Procedure
- The VARCOMP Procedure
- The VARIOGRAM Procedure



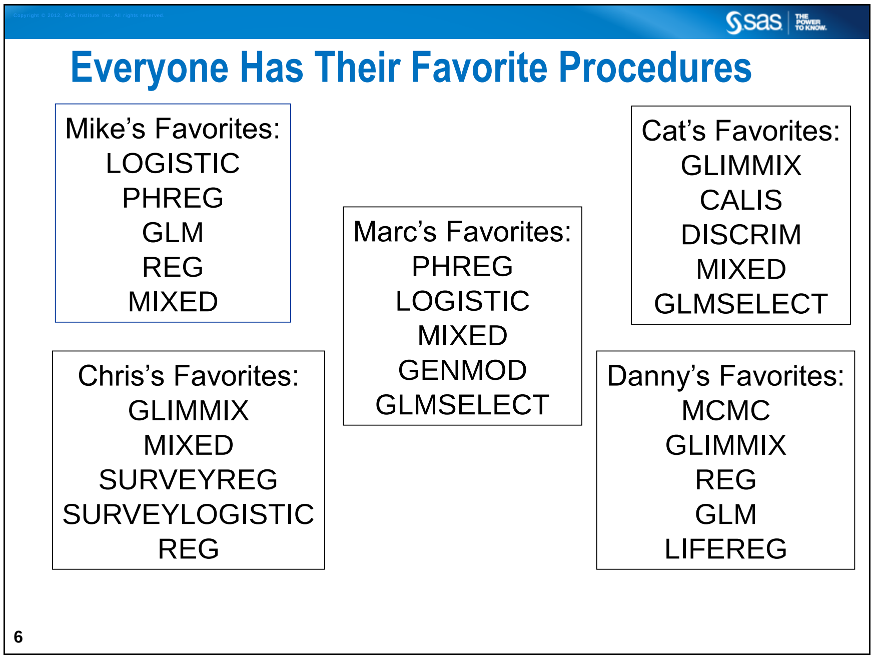
Focus on Modeling Procedures

Focusing just on “supervised” statistical modeling procedures, we can rule out many statistical procedures such as ACECLUS, FASTCLUS, PRINCOMP, FACTOR, CORRESP, SURVEYSELECT, STDIZE, MDS, MI, … the list goes on.

That still leaves somewhere in the neighborhood of 40 procedures.

What about the modeling procedures that statisticians love to use?

5



Everyone Has Their Favorite Procedures

Mike’s Favorites:
LOGISTIC
PHREG
GLM
REG
MIXED

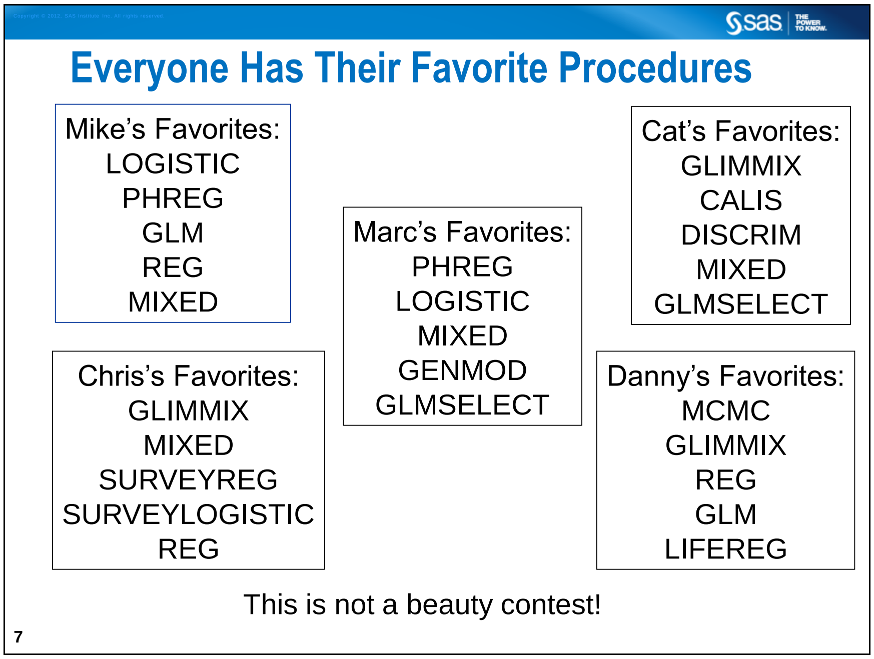
Chris’s Favorites:
GLIMMIX
MIXED
SURVEYREG
SURVEYLOGISTIC
REG

Marc’s Favorites:
PHREG
LOGISTIC
MIXED
GENMOD
GLMSELECT

Cat’s Favorites:
GLIMMIX
CALIS
DISCRIM
MIXED
GLMSELECT

Danny’s Favorites:
MCMC
GLIMMIX
REG
GLM
LIFEREG

6



Everyone Has Their Favorite Procedures

Mike's Favorites:
LOGISTIC
PHREG
GLM
REG
MIXED

Marc's Favorites:
PHREG
LOGISTIC
MIXED
GENMOD
GLMSELECT

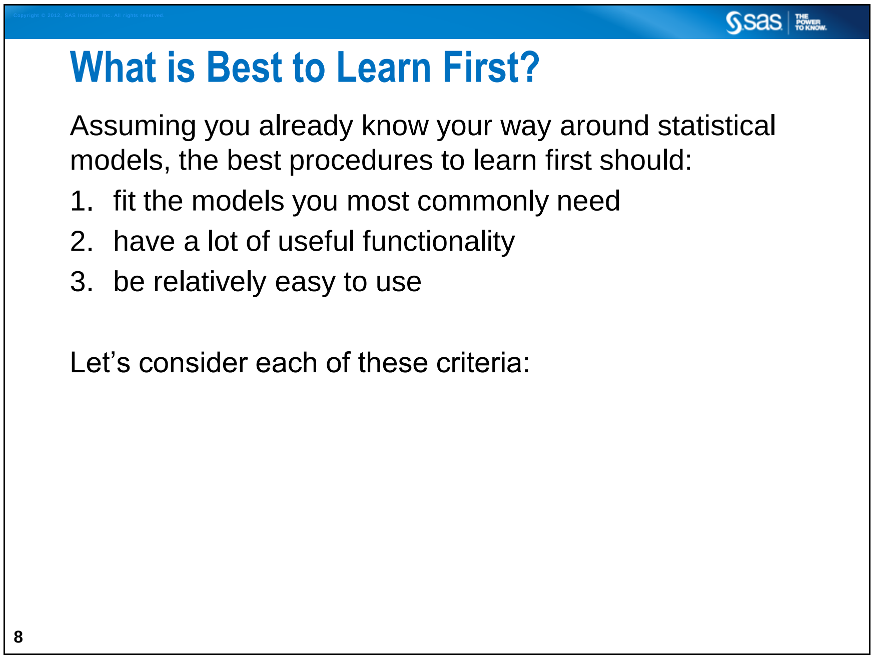
Cat's Favorites:
GLIMMIX
CALIS
DISCRIM
MIXED
GLMSELECT

Chris's Favorites:
GLIMMIX
MIXED
SURVEYREG
SURVEYLOGISTIC
REG

Danny's Favorites:
MCMC
GLIMMIX
REG
GLM
LIFEREG

This is not a beauty contest!

7



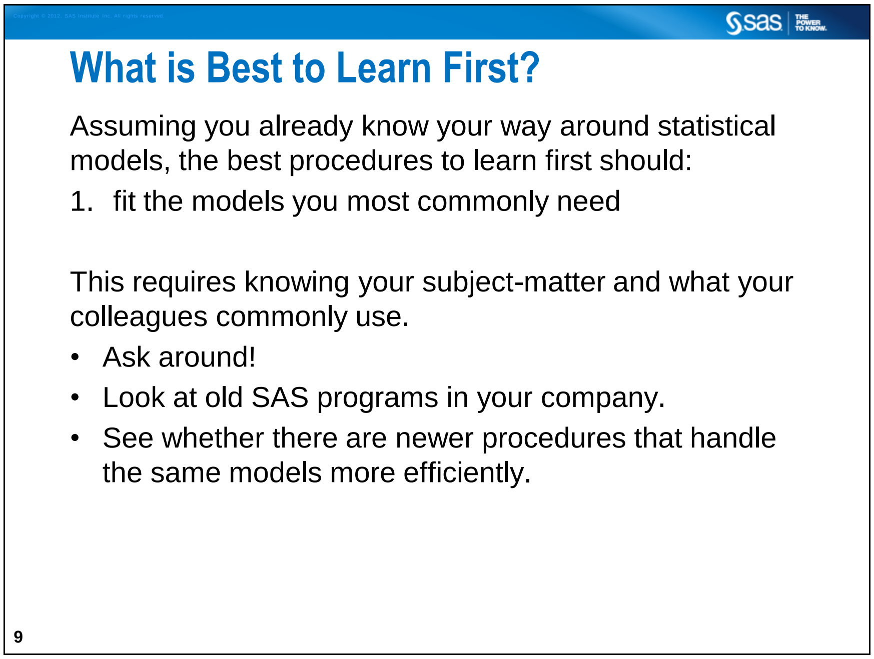
What is Best to Learn First?

Assuming you already know your way around statistical models, the best procedures to learn first should:

1. fit the models you most commonly need
2. have a lot of useful functionality
3. be relatively easy to use

Let's consider each of these criteria:

8



What is Best to Learn First?

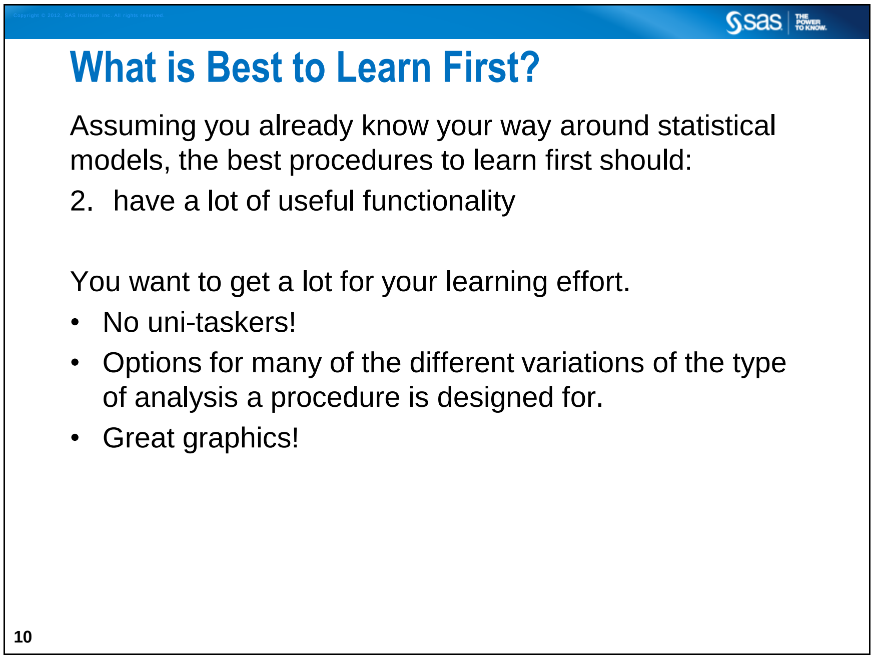
Assuming you already know your way around statistical models, the best procedures to learn first should:

1. fit the models you most commonly need

This requires knowing your subject-matter and what your colleagues commonly use.

- Ask around!
- Look at old SAS programs in your company.
- See whether there are newer procedures that handle the same models more efficiently.

9



What is Best to Learn First?

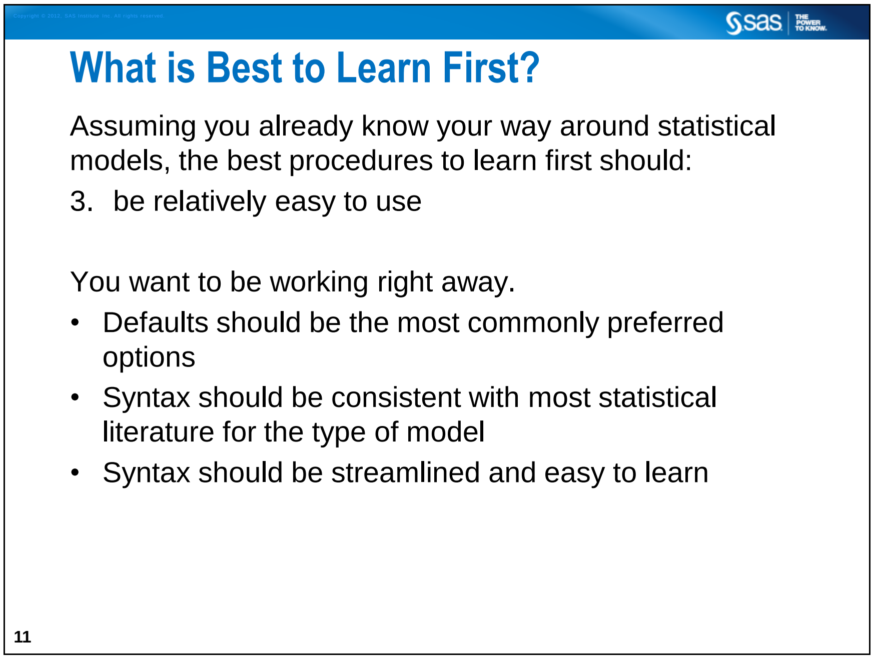
Assuming you already know your way around statistical models, the best procedures to learn first should:

2. have a lot of useful functionality

You want to get a lot for your learning effort.

- No uni-taskers!
- Options for many of the different variations of the type of analysis a procedure is designed for.
- Great graphics!

10



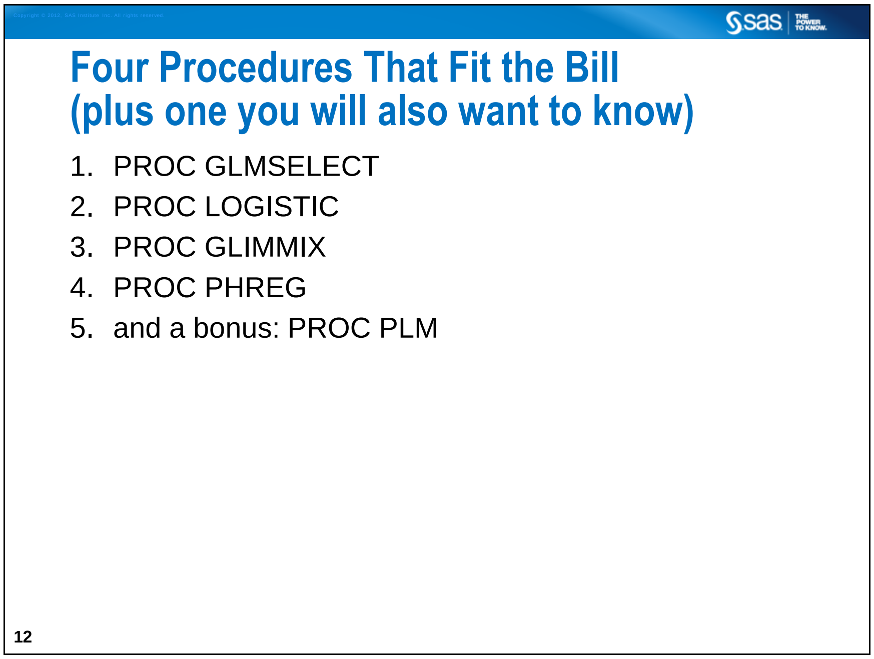
What is Best to Learn First?

Assuming you already know your way around statistical models, the best procedures to learn first should:

3. be relatively easy to use

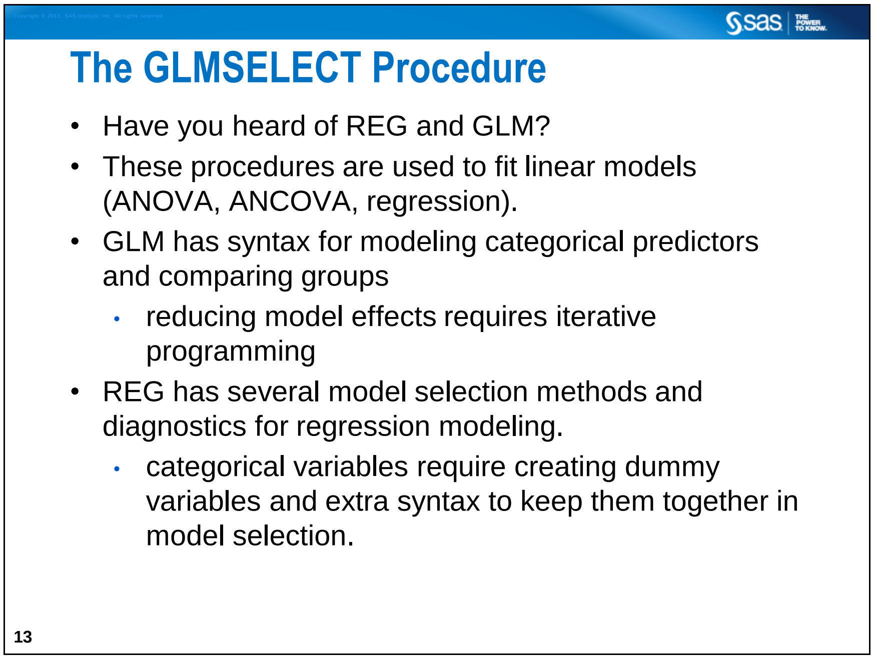
You want to be working right away.

- Defaults should be the most commonly preferred options
- Syntax should be consistent with most statistical literature for the type of model
- Syntax should be streamlined and easy to learn



Four Procedures That Fit the Bill (plus one you will also want to know)

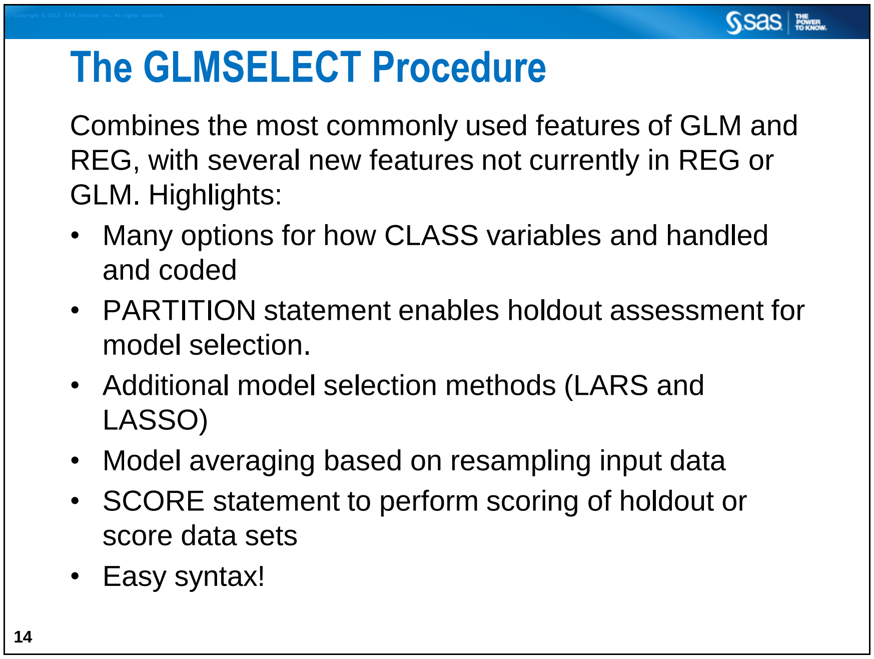
1. PROC GLMSELECT
2. PROC LOGISTIC
3. PROC GLIMMIX
4. PROC PHREG
5. and a bonus: PROC PLM



The GLMSELECT Procedure

- Have you heard of REG and GLM?
- These procedures are used to fit linear models (ANOVA, ANCOVA, regression).
- GLM has syntax for modeling categorical predictors and comparing groups
  - reducing model effects requires iterative programming
- REG has several model selection methods and diagnostics for regression modeling.
  - categorical variables require creating dummy variables and extra syntax to keep them together in model selection.

13

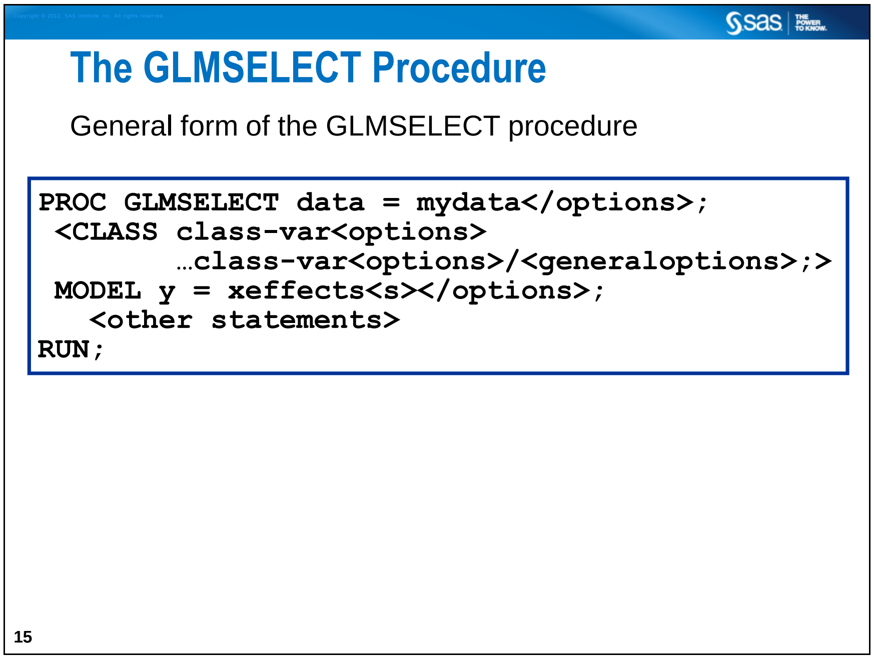


The GLMSELECT Procedure

Combines the most commonly used features of GLM and REG, with several new features not currently in REG or GLM. Highlights:

- Many options for how CLASS variables and handled and coded
- PARTITION statement enables holdout assessment for model selection.
- Additional model selection methods (LARS and LASSO)
- Model averaging based on resampling input data
- SCORE statement to perform scoring of holdout or score data sets
- Easy syntax!

14

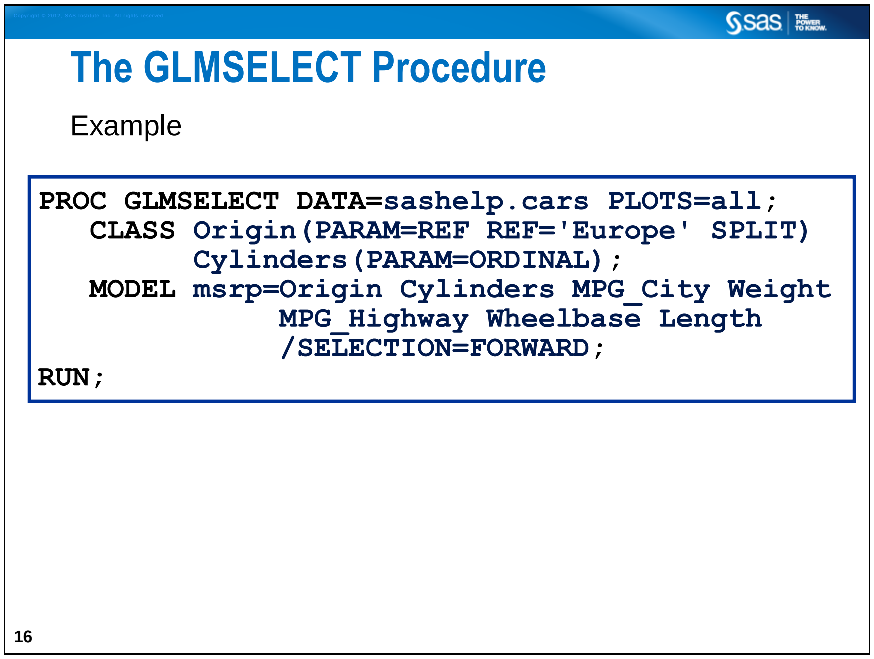


The GLMSELECT Procedure

General form of the GLMSELECT procedure

```
PROC GLMSELECT data = mydata</options>;
 <CLASS class-var<options>
        …class-var<options>/<generaloptions>;>
 MODEL y = xeffects<s></options>;
   <other statements>
RUN;
```

15

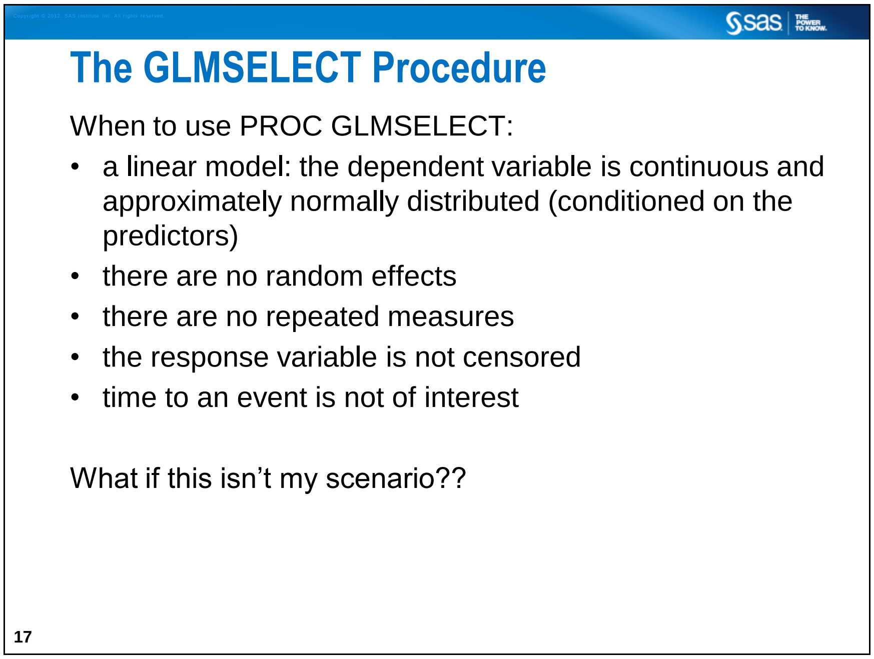


The GLMSELECT Procedure

Example

```
PROC GLMSELECT DATA=sashelp.cars PLOTS=all;
   CLASS Origin(PARAM=REF REF='Europe' SPLIT)
         Cylinders(PARAM=ORDINAL);
   MODEL msrp=Origin Cylinders MPG_City Weight
              MPG_Highway Wheelbase Length
              /SELECTION=FORWARD;
RUN;
```

16



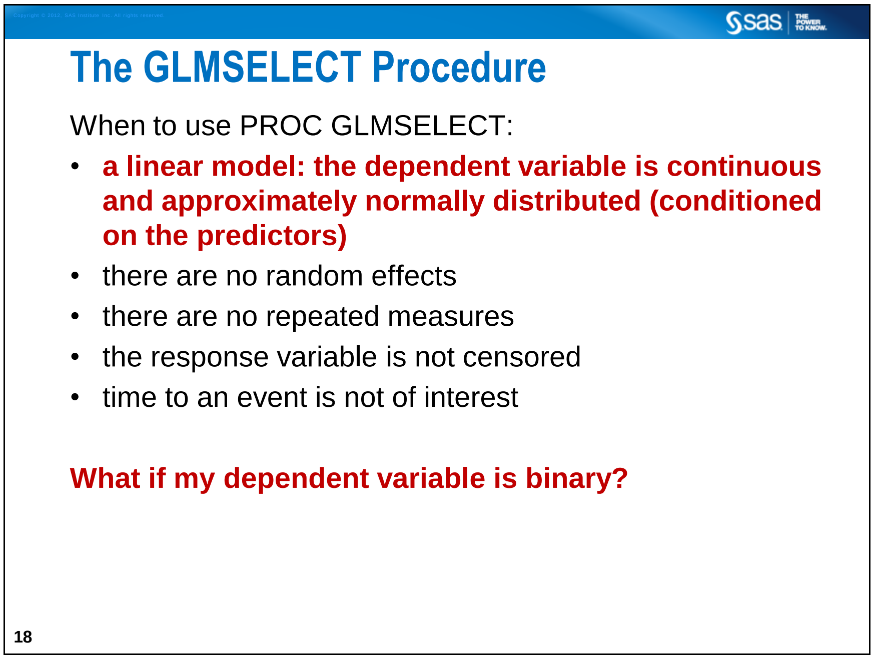
The GLMSELECT Procedure

When to use PROC GLMSELECT:

- a linear model: the dependent variable is continuous and approximately normally distributed (conditioned on the predictors)
- there are no random effects
- there are no repeated measures
- the response variable is not censored
- time to an event is not of interest

What if this isn't my scenario??

17



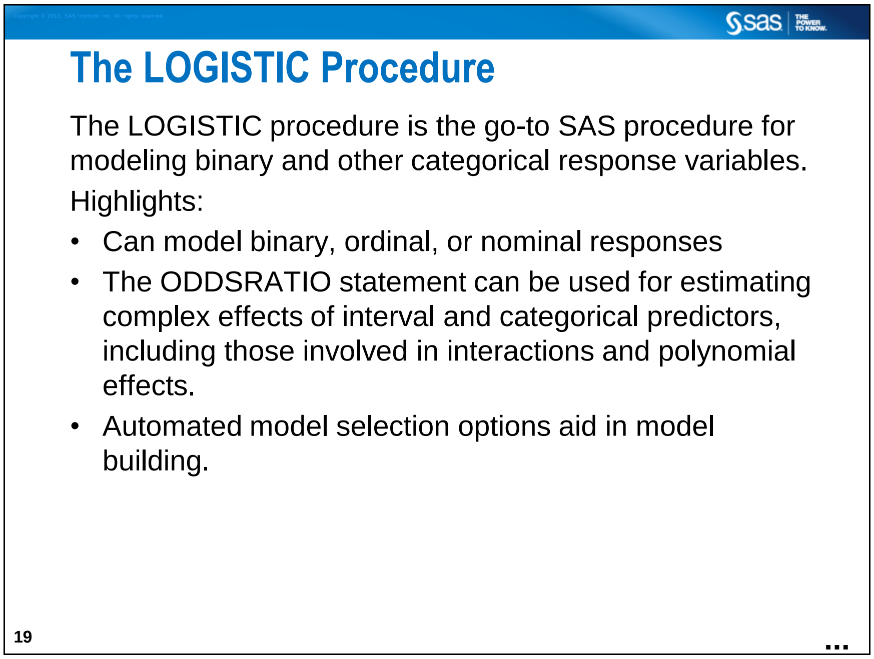
The GLMSELECT Procedure

When to use PROC GLMSELECT:

- **a linear model: the dependent variable is continuous and approximately normally distributed (conditioned on the predictors)**
- there are no random effects
- there are no repeated measures
- the response variable is not censored
- time to an event is not of interest

What if my dependent variable is binary?

18



The LOGISTIC Procedure

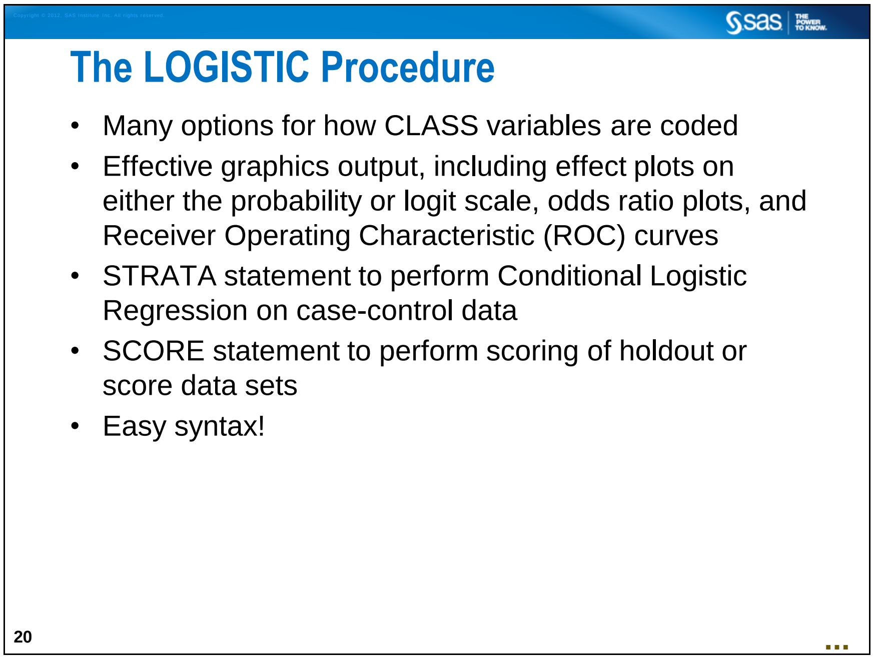
The LOGISTIC procedure is the go-to SAS procedure for modeling binary and other categorical response variables.

Highlights:

- Can model binary, ordinal, or nominal responses
- The ODDSRATIO statement can be used for estimating complex effects of interval and categorical predictors, including those involved in interactions and polynomial effects.
- Automated model selection options aid in model building.

19

...

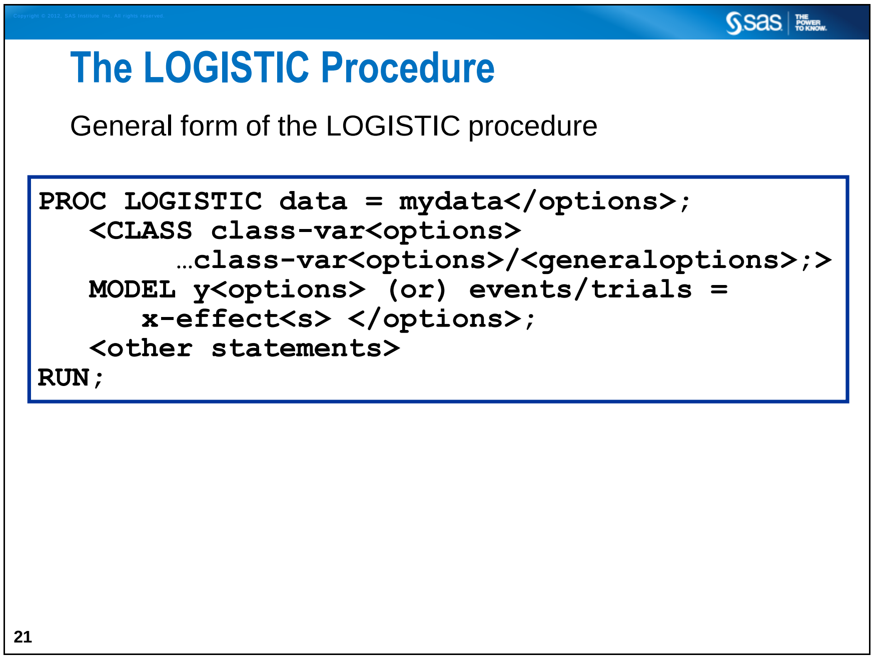


The LOGISTIC Procedure

- Many options for how CLASS variables are coded
- Effective graphics output, including effect plots on either the probability or logit scale, odds ratio plots, and Receiver Operating Characteristic (ROC) curves
- STRATA statement to perform Conditional Logistic Regression on case-control data
- SCORE statement to perform scoring of holdout or score data sets
- Easy syntax!

20

...

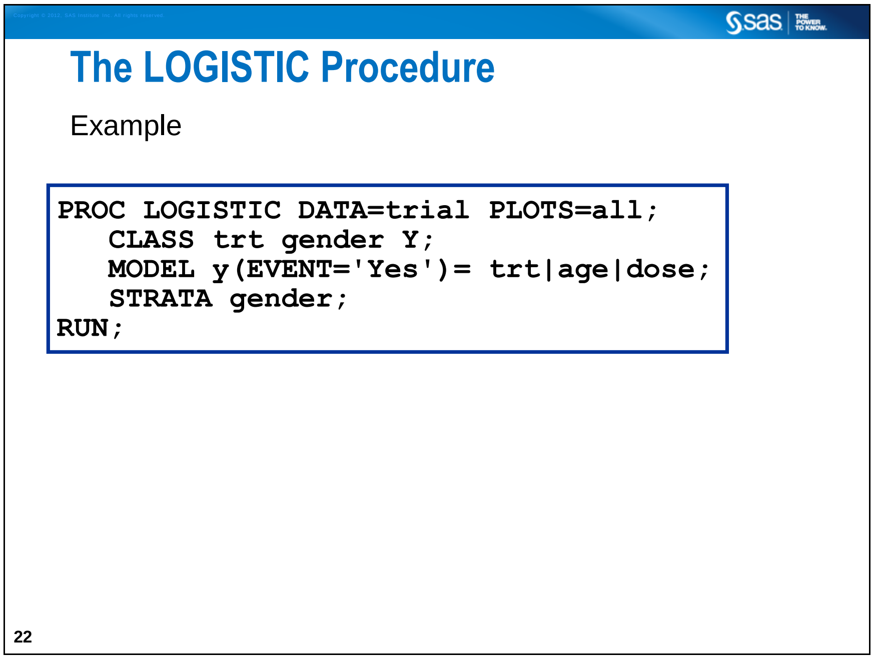


The LOGISTIC Procedure

General form of the LOGISTIC procedure

```
PROC LOGISTIC data = mydata</options>;
   <CLASS class-var<options>
        …class-var<options>/<generaloptions>;>
   MODEL y<options> (or) events/trials =
      x-effect<s> </options>;
   <other statements>
RUN;
```

21

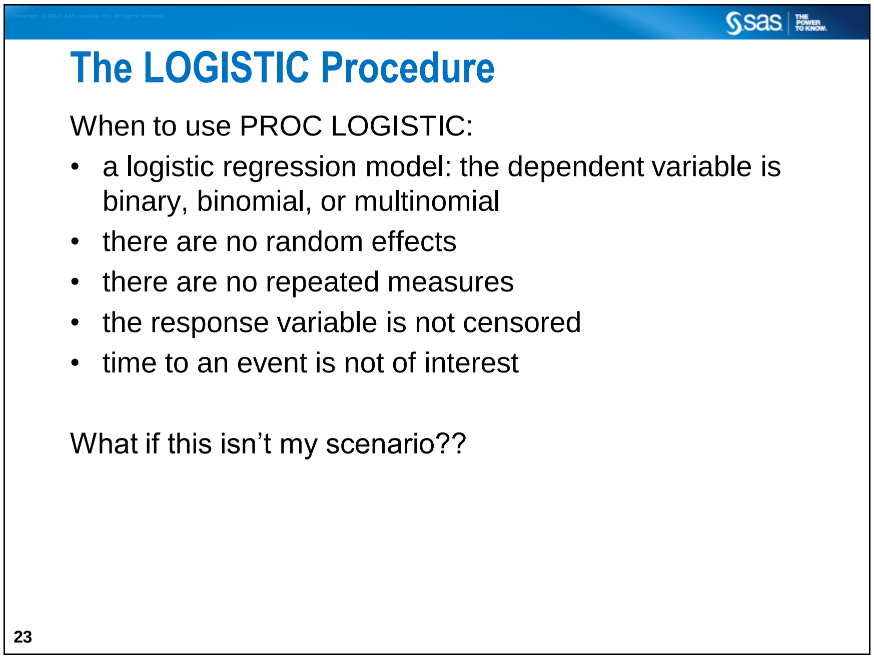


The LOGISTIC Procedure

Example

```
PROC LOGISTIC DATA=trial PLOTS=all;
   CLASS trt gender Y;
   MODEL y(EVENT='Yes')= trt|age|dose;
   STRATA gender;
RUN;
```

22



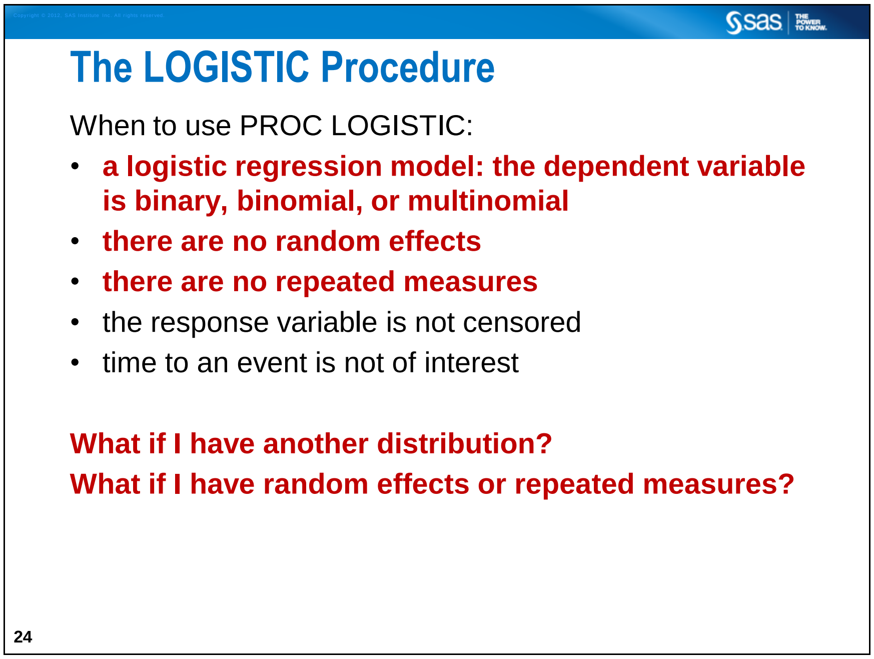
The LOGISTIC Procedure

When to use PROC LOGISTIC:

- a logistic regression model: the dependent variable is binary, binomial, or multinomial
- there are no random effects
- there are no repeated measures
- the response variable is not censored
- time to an event is not of interest

What if this isn't my scenario??

23



The LOGISTIC Procedure

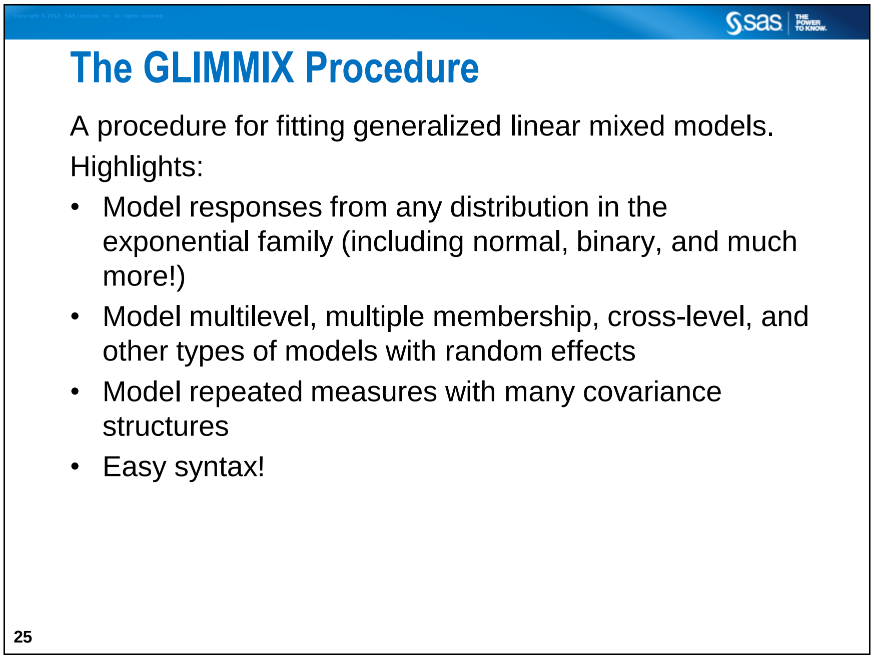
When to use PROC LOGISTIC:

- **a logistic regression model: the dependent variable is binary, binomial, or multinomial**
- **there are no random effects**
- **there are no repeated measures**
- the response variable is not censored
- time to an event is not of interest

What if I have another distribution?

What if I have random effects or repeated measures?

24



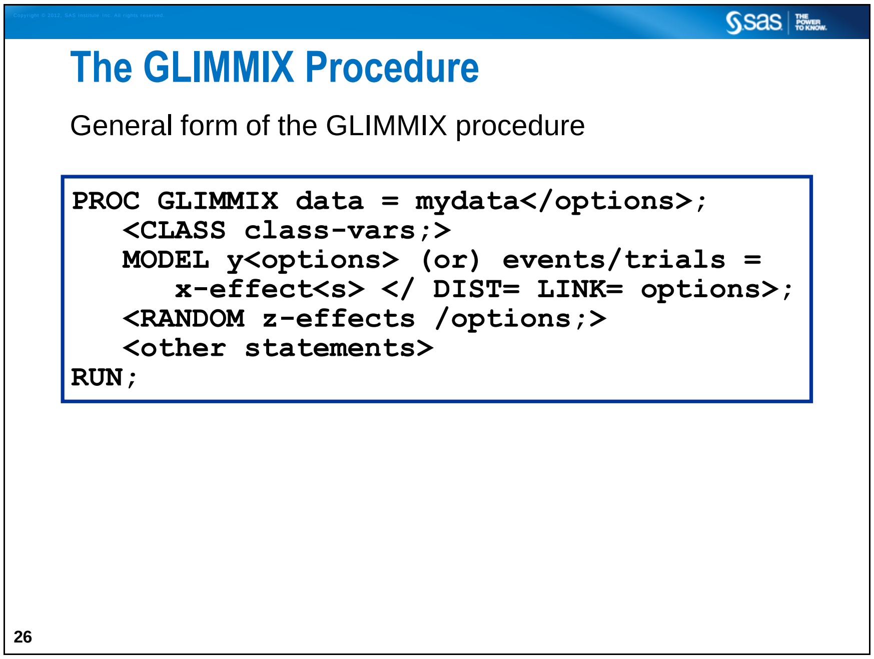
The GLIMMIX Procedure

A procedure for fitting generalized linear mixed models.

Highlights:

- Model responses from any distribution in the exponential family (including normal, binary, and much more!)
- Model multilevel, multiple membership, cross-level, and other types of models with random effects
- Model repeated measures with many covariance structures
- Easy syntax!

25

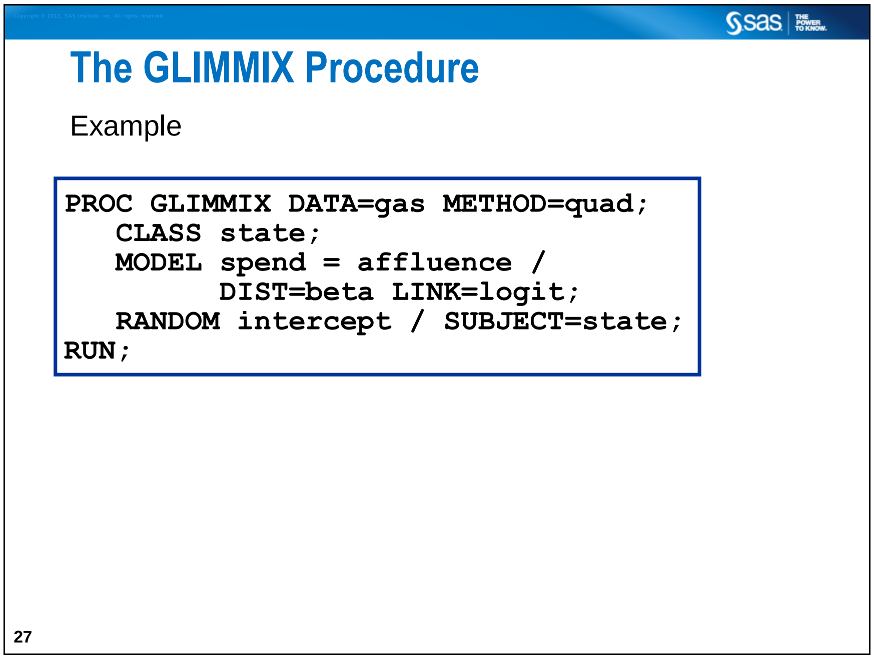


The GLIMMIX Procedure

General form of the GLIMMIX procedure

```
PROC GLIMMIX data = mydata</options>;
   <CLASS class-vars;>
   MODEL y<options> (or) events/trials =
      x-effect<s> </ DIST= LINK= options>;
   <RANDOM z-effects /options;>
   <other statements>
RUN;
```

26

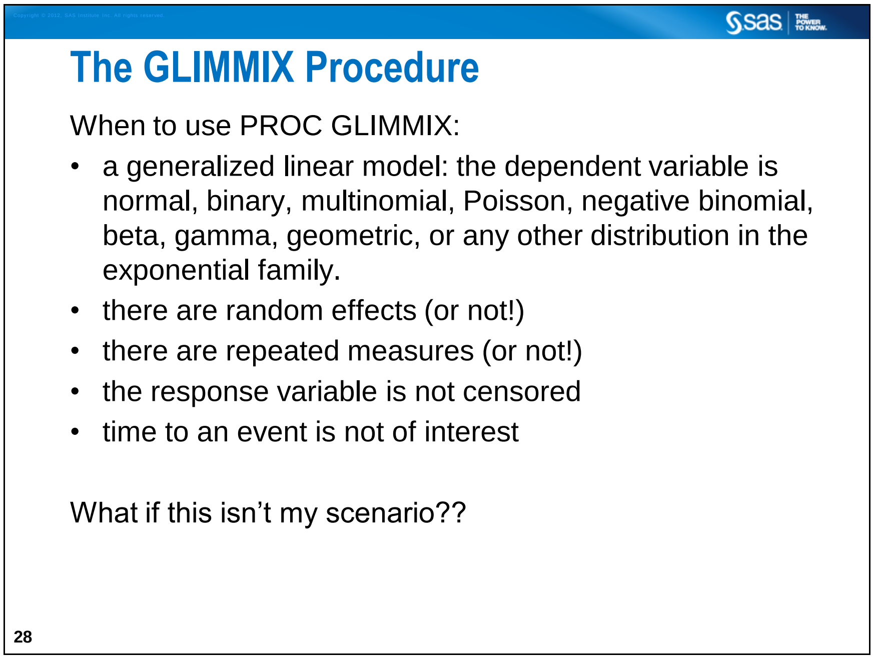


The GLIMMIX Procedure

Example

```
PROC GLIMMIX DATA=gas METHOD=quad;
   CLASS state;
   MODEL spend = affluence /
         DIST=beta LINK=logit;
   RANDOM intercept / SUBJECT=state;
RUN;
```

27



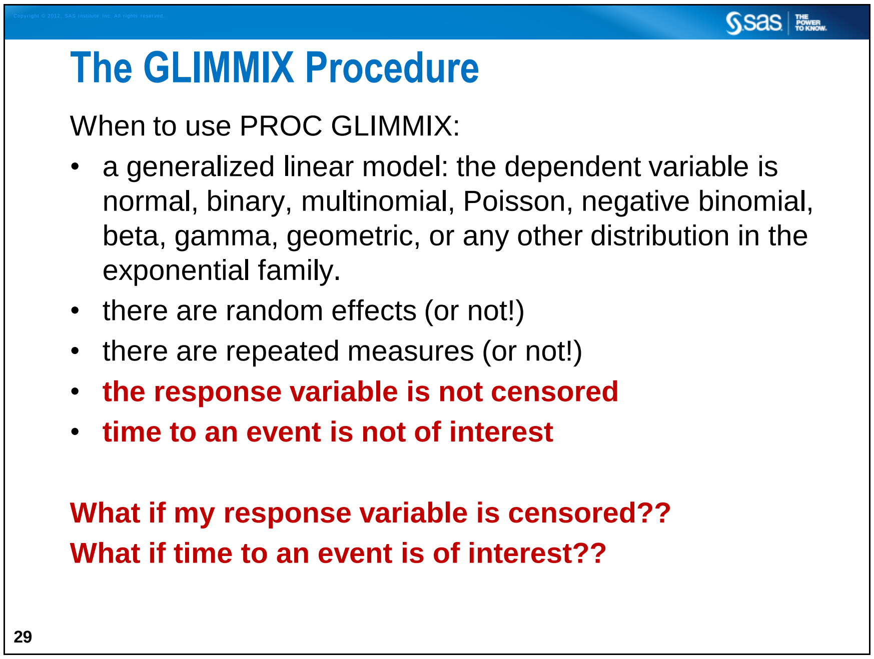
The GLIMMIX Procedure

When to use PROC GLIMMIX:

- a generalized linear model: the dependent variable is normal, binary, multinomial, Poisson, negative binomial, beta, gamma, geometric, or any other distribution in the exponential family.
- there are random effects (or not!)
- there are repeated measures (or not!)
- the response variable is not censored
- time to an event is not of interest

What if this isn’t my scenario??

28



The GLIMMIX Procedure

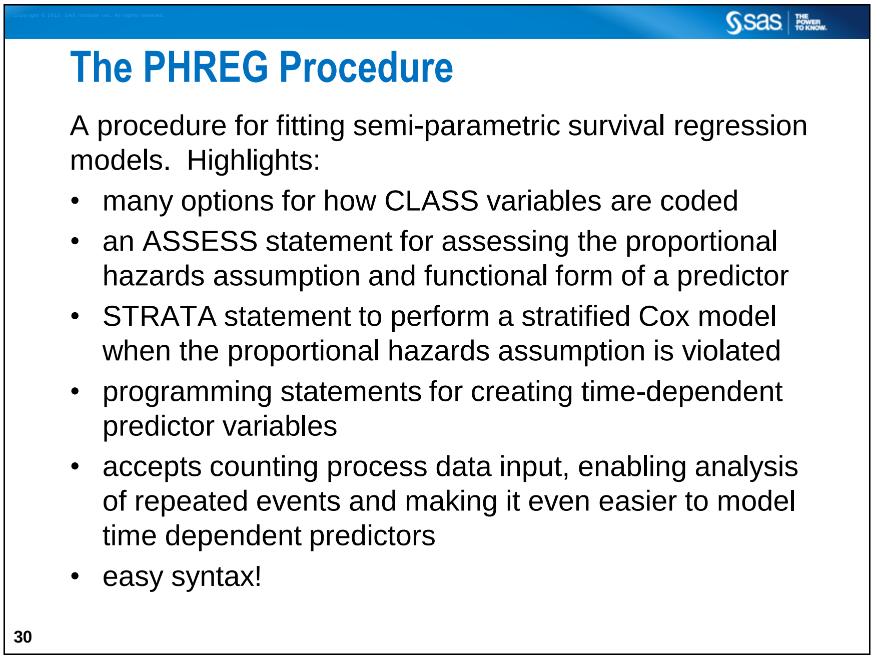
When to use PROC GLIMMIX:

- a generalized linear model: the dependent variable is normal, binary, multinomial, Poisson, negative binomial, beta, gamma, geometric, or any other distribution in the exponential family.
- there are random effects (or not!)
- there are repeated measures (or not!)
- **the response variable is not censored**
- **time to an event is not of interest**

What if my response variable is censored??

What if time to an event is of interest??

29

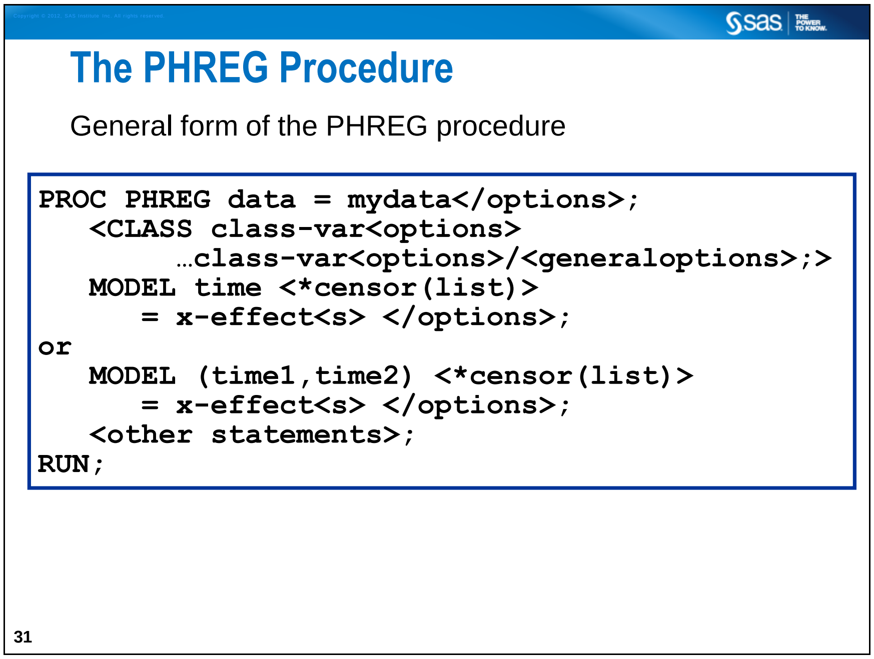


The PHREG Procedure

A procedure for fitting semi-parametric survival regression models. Highlights:

- many options for how CLASS variables are coded
- an ASSESS statement for assessing the proportional hazards assumption and functional form of a predictor
- STRATA statement to perform a stratified Cox model when the proportional hazards assumption is violated
- programming statements for creating time-dependent predictor variables
- accepts counting process data input, enabling analysis of repeated events and making it even easier to model time dependent predictors
- easy syntax!

30

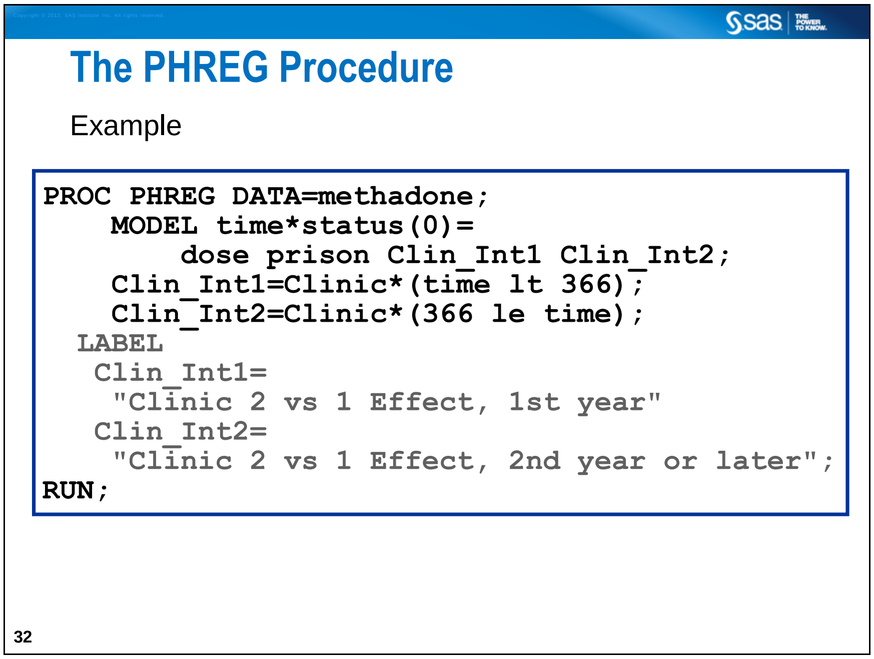


The PHREG Procedure

General form of the PHREG procedure

```
PROC PHREG data = mydata</options>;
   <CLASS class-var<options>
        …class-var<options>/<generaloptions>;>
   MODEL time <*censor(list)>
      = x-effect<s> </options>;
or
   MODEL (time1,time2) <*censor(list)>
      = x-effect<s> </options>;
   <other statements>;
RUN;
```

31

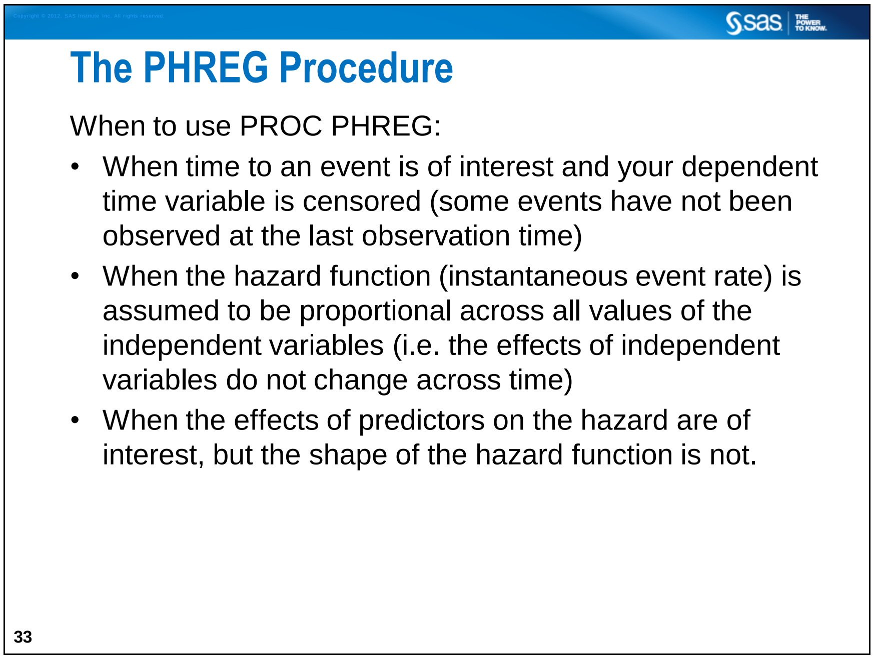


The PHREG Procedure

Example

```
PROC PHREG DATA=methadone;
    MODEL time*status(0)=
        dose prison Clin_Int1 Clin_Int2;
    Clin_Int1=Clinic*(time lt 366);
    Clin_Int2=Clinic*(366 le time);
  LABEL
   Clin_Int1=
    "Clinic 2 vs 1 Effect, 1st year"
   Clin_Int2=
    "Clinic 2 vs 1 Effect, 2nd year or later";
RUN;
```

32

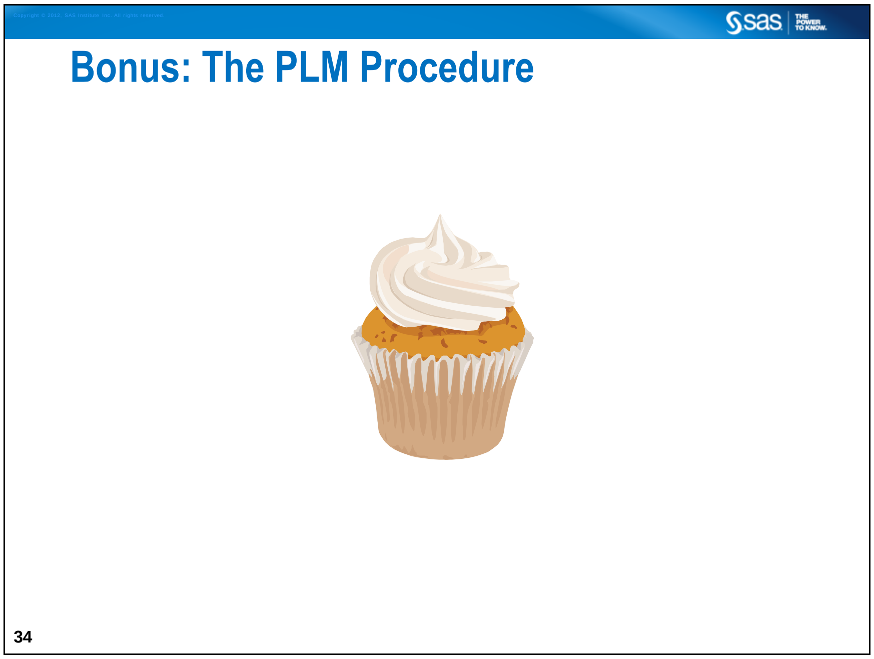


The PHREG Procedure

When to use PROC PHREG:

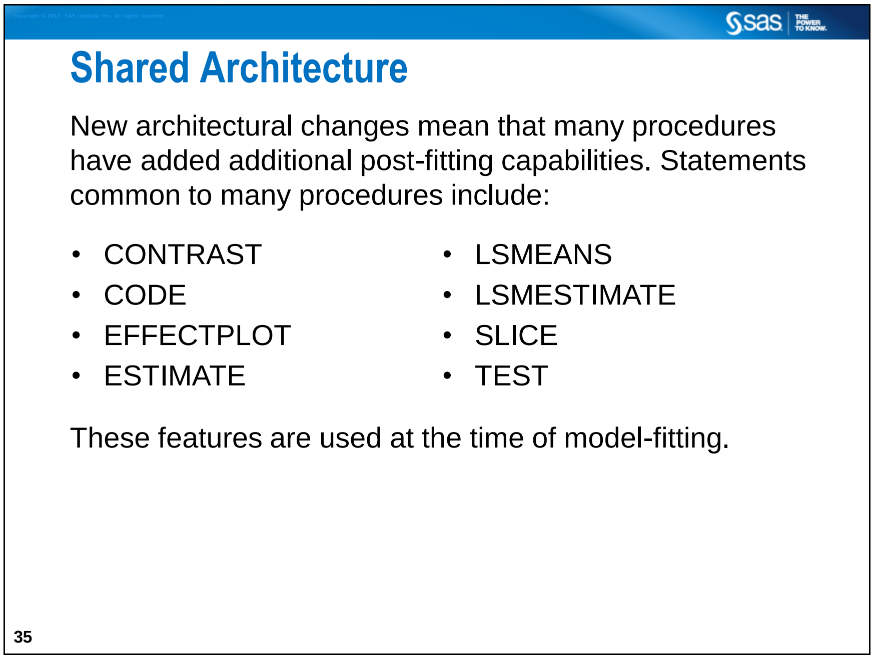
- When time to an event is of interest and your dependent time variable is censored (some events have not been observed at the last observation time)
- When the hazard function (instantaneous event rate) is assumed to be proportional across all values of the independent variables (i.e. the effects of independent variables do not change across time)
- When the effects of predictors on the hazard are of interest, but the shape of the hazard function is not.

33



Bonus: The PLM Procedure

34



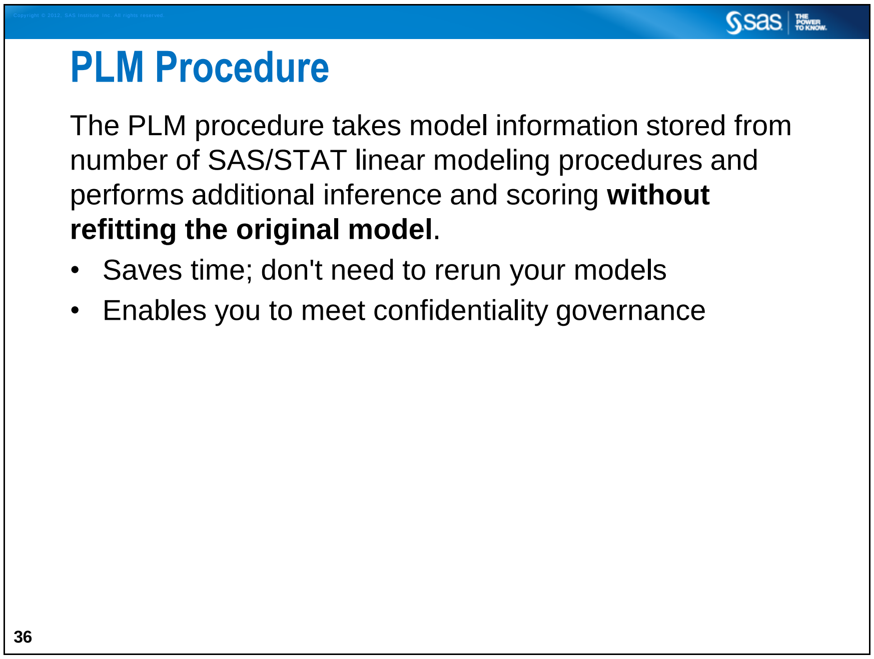
Shared Architecture

New architectural changes mean that many procedures have added additional post-fitting capabilities. Statements common to many procedures include:

- CONTRAST
- CODE
- EFFECTPLOT
- ESTIMATE
- LSMEANS
- LSMESTIMATE
- SLICE
- TEST

These features are used at the time of model-fitting.

35

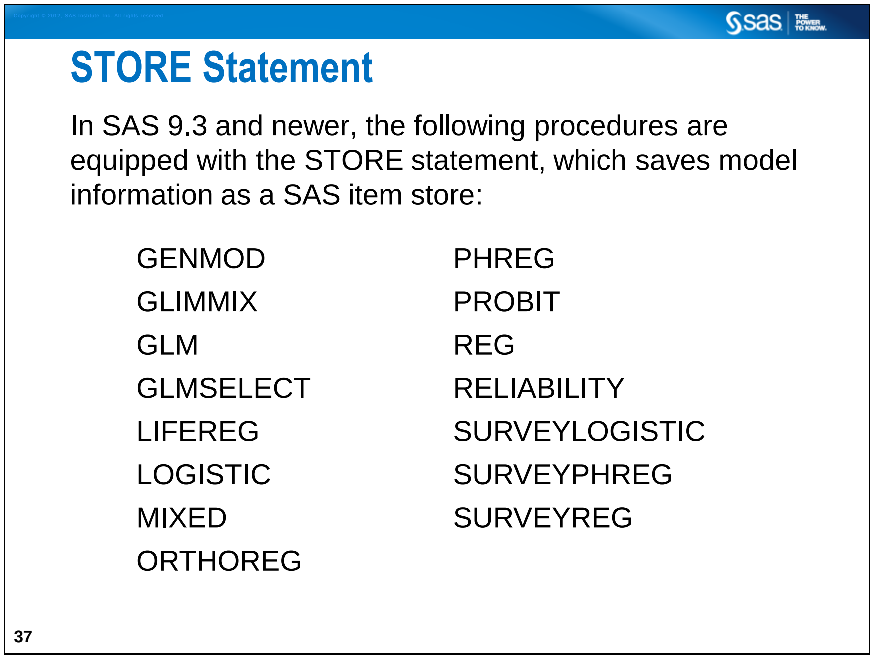


PLM Procedure

The PLM procedure takes model information stored from number of SAS/STAT linear modeling procedures and performs additional inference and scoring **without refitting the original model**.

- Saves time; don't need to rerun your models
- Enables you to meet confidentiality governance

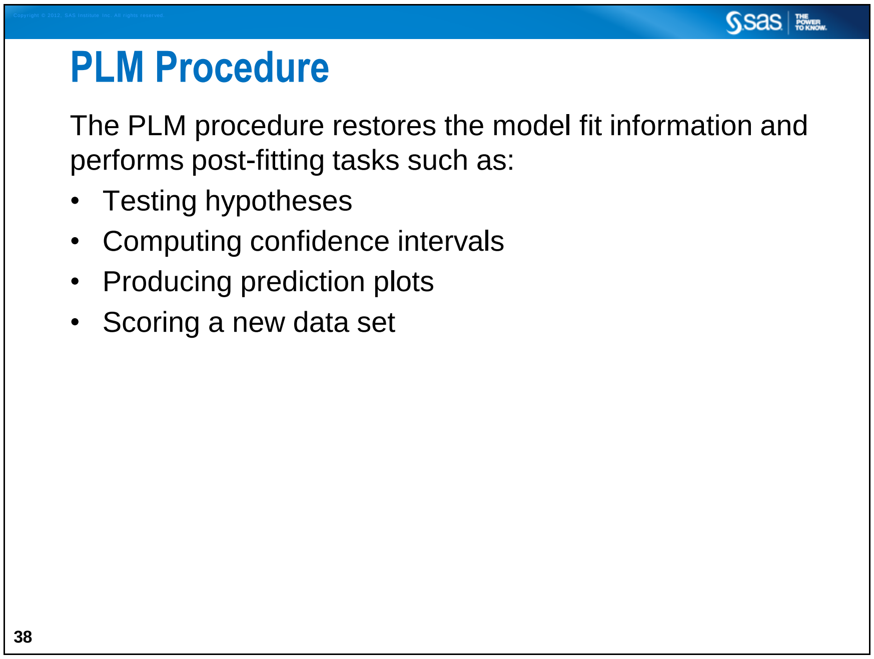
36



STORE Statement

In SAS 9.3 and newer, the following procedures are equipped with the STORE statement, which saves model information as a SAS item store:

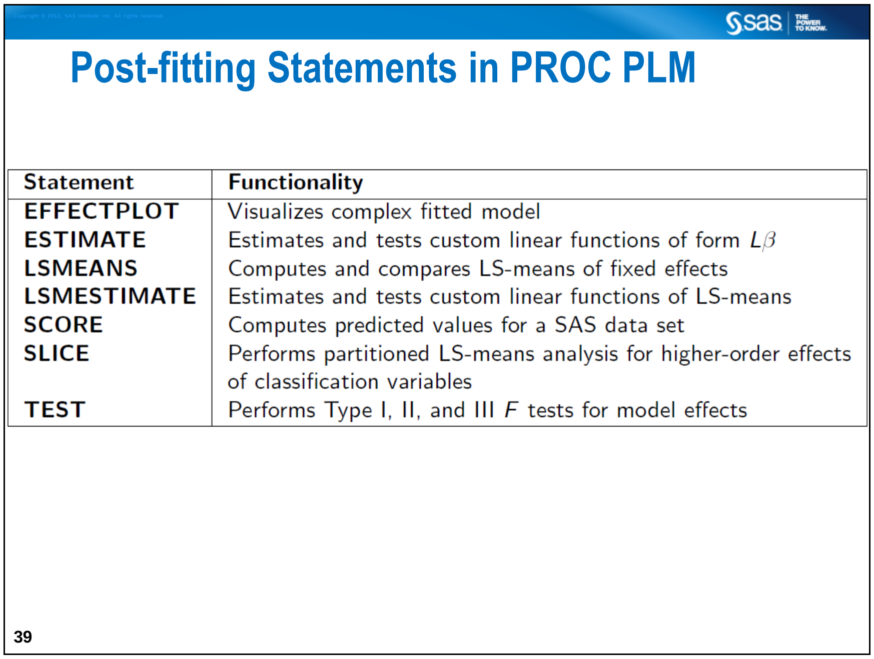
GENMOD
GLIMMIX
GLM
GLMSELECT
LIFEREG
LOGISTIC
MIXED
ORTHOREG
PHREG
PROBIT
REG
RELIABILITY
SURVEYLOGISTIC
SURVEYPHREG
SURVEYREG



PLM Procedure

The PLM procedure restores the model fit information and performs post-fitting tasks such as:

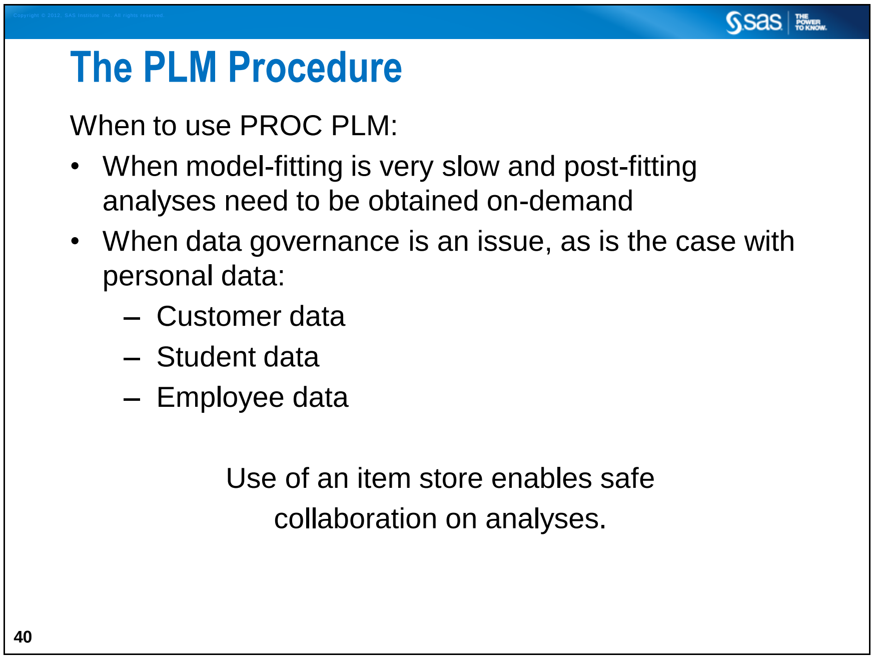
- Testing hypotheses
- Computing confidence intervals
- Producing prediction plots
- Scoring a new data set



Post-fitting Statements in PROC PLM

Statement	Functionality
EFFECTPLOT	Visualizes complex fitted model
ESTIMATE	Estimates and tests custom linear functions of form $L\beta$
LSMEANS	Computes and compares LS-means of fixed effects
LSMESTIMATE	Estimates and tests custom linear functions of LS-means
SCORE	Computes predicted values for a SAS data set
SLICE	Performs partitioned LS-means analysis for higher-order effects of classification variables
TEST	Performs Type I, II, and III F tests for model effects

39



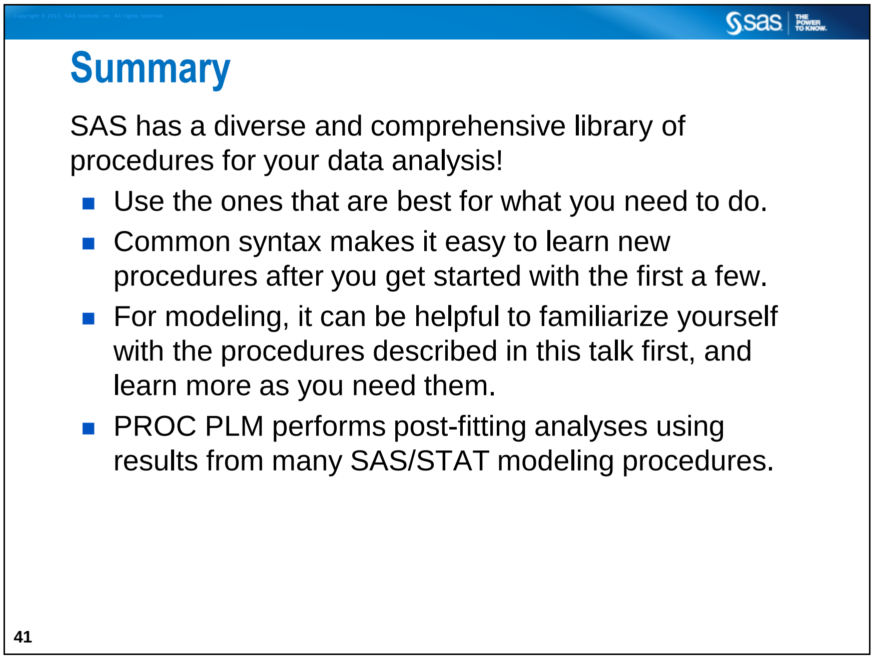
The PLM Procedure

When to use PROC PLM:

- When model-fitting is very slow and post-fitting analyses need to be obtained on-demand
- When data governance is an issue, as is the case with personal data:
  - Customer data
  - Student data
  - Employee data

Use of an item store enables safe collaboration on analyses.

40

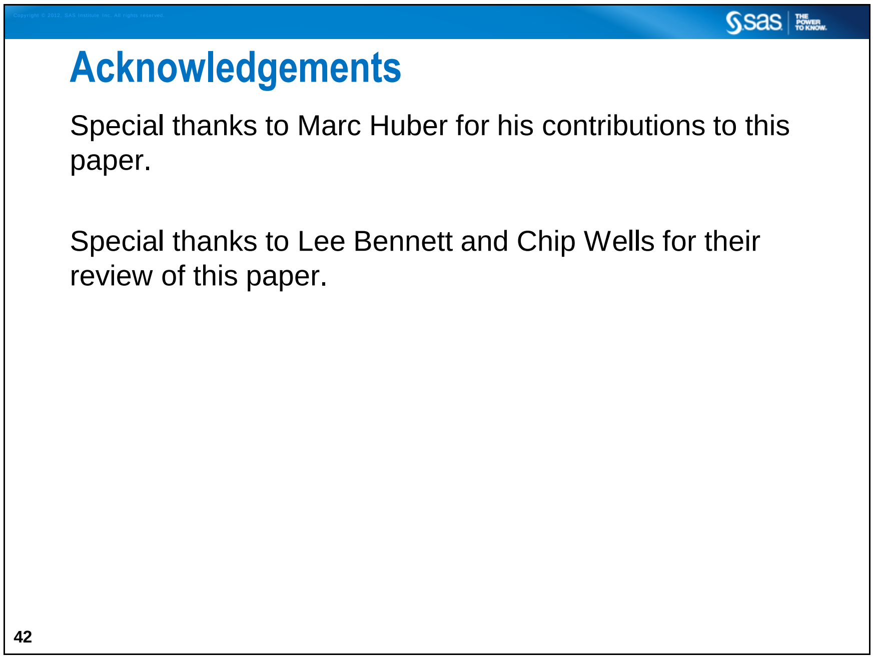


Summary

SAS has a diverse and comprehensive library of procedures for your data analysis!

- Use the ones that are best for what you need to do.
- Common syntax makes it easy to learn new procedures after you get started with the first a few.
- For modeling, it can be helpful to familiarize yourself with the procedures described in this talk first, and learn more as you need them.
- PROC PLM performs post-fitting analyses using results from many SAS/STAT modeling procedures.

41



Acknowledgements

Special thanks to Marc Huber for his contributions to this paper.

Special thanks to Lee Bennett and Chip Wells for their review of this paper.

42