Lesson 11
General Linear Models

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \varepsilon$$

Continuous Response

Continuous or Dummy Variables

Random Error
$\varepsilon \sim NI(0,\sigma_\varepsilon)$

Multiple Regression
ANALYSIS - Experimental Design Data
ANACOVA - Experimental Design Data + continuous covariate

SAS Procedures

PROC ANOVA — Equal replication.

PROC GLM — Unequal replication, one random effect

PROC MIXED — Unequal replication, multiple random effect

PROC GENMOD — Unequal replication, one non-normal random effect

PROC GLM

```
PROC GLM options ;
      CLASS variable-list;
      MODEL dependents= independents / options; /* required */
      ABSORB variable-list;
      BY variable-list;
      FREQ variable;
      ID variable-list;
      WEIGHT variable;
      CONTRAST 'label' effect values... / options;
      ESTIMATE 'label' effect values... / options;
      LSMEANS effects / options;
      MANOVA H= effects E= effect M= equations...
             MNAMES= names PREFIX= name / options;
      MEANS effects / options;
      OUTPUT OUT= SAS-data-set keywords= names... ;
      RANDOM effects / options;
      REPEATED factorname levels (levelvalues)
               transformation<,...> / options;
      TEST H= effects E= effect / options;
```

```
PROC GLM options;
The following options can be used in the PROC GLM statement:

   DATA= SAS-data-set
   ORDER=FREQ|DATA|INTERNAL|FORMATTED
   OUTSTAT=SAS-data-set
   MANOVA
   MULTIPASS
   NOPRINT
```

```
CLASS variable-list;
  The CLASS statement names the classification variables to be
  used in the analysis.
  If the CLASS statement is used, it must appear before the
  MODEL statement.
  Classification variables can be either character or numeric.
  Only the first sixteen characters of a character variable are
  used.
  Class levels are determined from the formatted values of the
  CLASS variables.
```

```
MODEL  dependents= independents / options;
```

The MODEL statement names the dependent variables and independent effects.

If no independent effects are specified, only an intercept term is fit.

These options can be specified in the MODEL statement after a slash (/):

NOINT	INTERCEPT	NOUNI	SOLUTION	TOLERANCE
E	E1	E2	E3	E4
SS1	SS2	SS3	SS4	ALPHA= p
CLM	CLI	P	XPX	INVERSE
SINGULAR= value		ZETA= value		* | @ | |

When the bar (|) is used, the right- and left- hand sides becomes effects, and the cross of them becomes an effect. Multiple bars are permitted. A | B | C

You can also specify the maximum number of variables involved in any effect that results from bar evaluation by specifying that maximum number, preceded by an @ sign, at the end of the bar effect. A | B(A) | C@2

Crossed effects (interactions) are specified by joining class variables with asterisk: A*B B*C A*B*C

Assume the variable TEMP has values in the data set of {10,20,30} and that it has been read in as a number.

```
PROC GLM;
  MODEL resp = temp ;
```

Fit a <u>linear regression</u> between response and temp.

$$Y = \alpha_0 + \beta T + \varepsilon$$

```
PROC GLM;
  CLASS temp;
  MODEL resp = temp ;
```

Fit an ANOVA model where we compare mean responses among the three temperature classes.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$Z_{10} = \begin{cases} 1 & \text{if TEMP} = 10 \\ 0 & \text{otherwise} \end{cases}$$

$$Z_{20} = \begin{cases} 1 & \text{if TEMP} = 20 \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{ij} = \mu + \beta_{10} Z_{10} + \beta_{20} Z_{20} + \varepsilon_{ij}$$

Reference cell model

Assume, in addition to TEMP there is a second factor, say TIME set to values of {1,3,7,15}.

```
PROC GLM;
  MODEL resp = temp time ;
```

Fit a <u>multiple linear regression</u> between response and the two factors temp and time.

$$Y = \alpha_0 + \beta \mathrm{Temp} + \theta \mathrm{Time} + \varepsilon$$

```
PROC GLM;
  CLASS temp time;
  MODEL resp = temp|time ;
/* Alternate specification
MODEL resp=temp time
      temp*time ;          */
```

Fit an ANOVA model where we examine “main” effects of temperature and time as well as “interaction” effects.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$


```
PROC GLM;
  CLASS temp;
  MODEL resp = temp|time ;
/* Alternate specification
MODEL resp=temp time
      temp*time ;          */
```

Fit an ANACOVA model where we examine “main” effects of temperature and a regression on time as well as heterogeneity of slope effects.

Mean difference between Temp=10 and Temp=30 at average TIME.

Mean difference between Temp=20 and Temp=30 at average TIME.

Overall TIME slope.

$$Y_{ijk} = \mu + \beta_{10} Z_{10} + \beta_{20} Z_{20} + \beta \mathrm{TIME} + \theta_{10} Z_{10} \mathrm{TIME} + \theta_{20} Z_{20} \mathrm{TIME} + \varepsilon_{ijk}$$

Average response at Temp=30 and average TIME.

Change of slope between TEMP=10 and TEMP=30.

Change of slope between TEMP=20 and TEMP=30.

Hypothesis:

1. Strawberry yields are not effected by any fertilizer.
2. Strawberry yields are affected by fertilizer but type of fertilizer is not important.
3. Strawberry yields are affected by fertilizer but fertilizer rate is not important.

Response:
• Strawberry yeild - YIELD;

ExperimentalFactors:

• Strawberry variety - VARIETY;	levels=2
• Fertilizer type - FERTLIZ;	levels=2
• Application rate - RATE;	levels=3
• Replication - REPLICAT;	levels=2

Design Issues:
• Levels of VARIETY, FERTLIZ, and RATE are all crossed.
• RATE is a continuous factor.
• Replicates are nested within VARIETY by FERTLIZ by RATE combinations.


```
DATA berry;
  INPUT fertiliz $ variety $ rate replicat yield @@;
  DATALINES;
K Red   .3 1 9.1 K Red   .3 2 9.0
K Red   .6 1 8.7 K Red   .6 2 8.4
K Red   .9 1 8.0 K Red   .9 2 8.4
K Sweet .3 1 9.3 K Sweet .3 2 9.2
K Sweet .6 1 9.0 K Sweet .6 2 8.7
K Sweet .9 1 8.3 K Sweet .9 2 8.5
N Red   .3 1 8.4 N Red   .3 2 8.8
N Red   .6 1 8.8 N Red   .6 2 8.9
N Red   .9 1 9.0 N Red   .9 2 8.9
N Sweet .3 1 8.7 N Sweet .3 2 9.0
N Sweet .6 1 9.2 N Sweet .6 2 9.3
N Sweet .9 1 9.1 N Sweet .9 2 9.5
;
  run;
```


One-way ANOVA model with variety as a predictor

```
PROC GLM DATA=berry;
  CLASS variety;
  MODEL yield=variety;
  RUN;
```

One-way ANOVA model that considers the combination of variety and fertilizer to be a "treatment"

```
PROC GLM DATA=berry;
  CLASS variety fertiliz;
  MODEL yield=variety*fertiliz;
  RUN;
```


ANOVA model with rate, regarding .3, .6, and .9 as three different categories.

```
PROC GLM DATA=berry;
  CLASS rate;
  MODEL yield=rate;
  RUN;
```

Regression model based on rate.

```
PROC GLM DATA=berry;
  MODEL yield=rate;
  RUN;
```

Regression model based on rate and rate^2 (Note: This works in PROC GLM, but not in PROC REG)

```
PROC GLM DATA=berry;
  MODEL yield=rate rate*rate;
  RUN;
```

Analysis of covariance model with variety and rate, where rate is considered to be a continuous predictor, or covariate

```
PROC GLM DATA=berry;
  CLASS variety;
  MODEL yield=variety rate;
  RUN;
```

Model with variety and fertilizer as categorical predictors

```
PROC GLM DATA=berry;
  CLASS variety fertiliz;
  MODEL yield=variety fertiliz;
  RUN;
```


Model with variety, fertilizer, and interaction term.

```
PROC GLM DATA=berry;
  CLASS variety fertiliz;
  MODEL yield=variety fertiliz variety*fertiliz;
/* Equivalently:
  MODEL yield=variety fertiliz fertiliz*variety;
*/
  RUN;
```

Model with fertilizer, variety, and rate (considered to be categorical), as well as all of their interactions

```
PROC GLM DATA=berry;
 CLASS variety fertiliz rate;
 MODEL yield=variety fertiliz rate variety*fertiliz
  variety*rate fertiliz*rate variety*fertiliz*rate;
 RUN;
```


Model with the effects of rate considered to be nested within the type of fertilizer (for example, 0.3 units of K is not equivalent to 0.3 units of N, because N is distributed as a solid and K is distributed as a liquid)

```
PROC GLM DATA=berry;
  CLASS fertiliz rate;
  MODEL yield=fertiliz rate(fertiliz);
  RUN;
```

Model with fertilizer and a different slope with respect to rate foreachfertilizer

```
PROC GLM DATA=berry;
  CLASS fertiliz;
  MODEL yield=fertiliz rate(fertiliz);
  RUN;
```


Which Model to Use?

This depends on:

- **Experiment Design:**
  completely randomized design, randomized block design, split-plot design, etc.
- **Selection and Sampling Plan:**
  nested, repeated measures, destructive sampling, etc.
- **Type of Factor:**
  fixed, random, continuous
- **Model Parsimony:**
  simplest appropriate model that adequately describes results
- **Your "Expert":**
  statistical reference book, consultant, tradition, ...

The CRAFT of STATISTICS

TWO-FACTOR FACTORIAL in a COMPLETELY RANDOMIZED DESIGN

```
PROC GLM DATA=berry;
 CLASS fertiliz variety;
 MODEL yield=fertiliz variety fertiliz*variety/SOLUTION;
 RUN;
```

Levels Block

```
General Linear Models Procedure
Class Level Information
Class          Levels    Values
FERTILIZ            2    K N
VARIETY             2    Red Sweet
Number of observations in data set = 24
```

STA5106-11-17

Overall Model AOV Block

```
Dependent Variable: YIELD
                         Sum of         Mean
Source            DF    Squares       Square     F Value   Pr > F
Model              3  0.87166667   0.29055556       2.59   0.0816
Error             20  2.24666667   0.11233333
Corrected Total   23  3.11833333

   R-Square          C.V.          Root MSE        YIELD Mean
   0.279530      3.790707         0.3351617         8.8416667
```

STA5106-11-18

Types Sums of Squares test block

Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERTILIZ	1	0.37500000	0.37500000	3.34	0.0826
VARIETY	1	0.48166667	0.48166667	4.29	0.0515
FERTILIZ*VARIETY	1	0.01500000	0.01500000	0.13	0.7186

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERTILIZ	1	0.37500000	0.37500000	3.34	0.0826
VARIETY	1	0.48166667	0.48166667	4.29	0.0515
FERTILIZ*VARIETY	1	0.01500000	0.01500000	0.13	0.7186

The Type I sums of squares = sequential sums of squares.
- Are there fertilizer differences in expected yield?
- Are there significant variety differences after fertilizer differences have been taken into account.
- Is the effect of variety dependent on which fertilizer you look at? (or equivalently: Is the effect of fertilizer dependent on the variety considered?)

STA5106-11-19

Types Sums of Squares test block

Source	DF	Type I SS	Mean Square	F Value	Pr > F
FERTILIZ	1	0.37500000	0.37500000	3.34	0.0826
VARIETY	1	0.48166667	0.48166667	4.29	0.0515
FERTILIZ*VARIETY	1	0.01500000	0.01500000	0.13	0.7186

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERTILIZ	1	0.37500000	0.37500000	3.34	0.0826
VARIETY	1	0.48166667	0.48166667	4.29	0.0515
FERTILIZ*VARIETY	1	0.01500000	0.01500000	0.13	0.7186

The Type III sums of squares = partial sums of squares.
- Are there fertilizer differences in expected yield after accounting for variety effects?
- Are there significant variety differences after fertilizer effects have been taken into account.
- Is the effect of variety dependent on which fertilizer you look at? (or equivalently: Is the effect of fertilizer dependent on the variety considered?)

Because the experiment is balanced, both Type I and Type III sums of squares are identical. Usually, the Type III sums of squares are used for inference, although the Type I sums of squares are used in specific situations. SAS can calculate Type II and Type IV sums of squares as well.


```
PROC GLM DATA=berry;
 CLASS fertiliz variety;
 MODEL yield=fertiliz variety fertiliz*variety/SOLUTION;
 OUTPUT OUT=results P=pred R=resid;
 RUN;
PROC PLOT DATA=results;
 PLOT resid*pred;
 RUN;
```

```
 Plot of RESID*PRED.  Legend: A = 1 obs, B = 2 obs, etc.

RESID  |
  0.5  + A                       A
       | A                       A                             A
       |                     A
       | A                   B   A                             B
  0.0  +                     B                                 A
       |                         A                             A
       | B
       |                     A   A                             A
 -0.5  +                         A
       | A
       |
       |
 -1.0  +

+-+--------+--------+--------+--------+--------+--------+-
       8.6       8.7       8.8       8.9       9.0       9.1
9.2
                                   PRED
```

Solutionblock

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error Of Estimate
INTERCEPT		9.13 B	66.75	0.001	0.137
FERTILIZ	K	-0.30 B	-1.55	0.137	0.194
	N	0.00 B	.	.	.
VARIETY	Red	-0.33 B	-1.72	0.100	0.194
	Sweet	0.00 B	.	.	.
FERTILIZ*VARIETY	K Red	0.10 B	0.37	0.719	0.274
	K Sweet	0.00 B	.	.	.
	N Red	0.00 B	.	.	.
	N Sweet	0.00 B	.	.	.

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

PROC GLM is a very diverse tool. It also provides the following capabilities:

Performing proper F tests with random and mixed ANOVA models
Performing multivariate analysis of variance
Comparing treatments, or combinations of treatments, with multiple comparison procedures
Calculating the averages of all observations with a particular combination of factor levels, or estimating what that average would have been had the experiment been balanced with least-squares means
Estimating the response under a given set of conditions, such as Treatment 1 with the covariate equal to 15
Contrasting one group of treatments versus another group of treatments, or testing for certain trends in the data with contrasts
Performing repeated-measures analysis of variance.

MEANS effects / options;

For any effect that appears on the right-hand side of the model and that does not contain any continuous variables, GLM can compute means of all continuous variables in the model.

You can use any number of MEANS statements, provided they appear after the MODEL statement.

These options can appear in the MEANS statement after a slash (/):

ALPHA=	E=	SCHEFFE
BON	ETYPE=	SIDAK
CLDIFF	GABRIEL	SMM | GT2
CLM	HOVTEST=	SNK
DEPONLY	HTYPE=	T | LSD
DUNCAN	KRATIO=	TUKEY
DUNNETT	LINES	WALLER
DUNNETTL	NOSORT	WELCH
DUNNETTU	REGWF	
	REGWQ	


```
PROC GLM DATA=berry;
 CLASS fertiliz variety;
 MODEL yield=fertiliz variety fertiliz*variety;
 MEANS fertiliz variety / waller ;
 RUN;
```

```
NOTE: Means from the MEANS statement are not adjusted for
      other terms in the model.  For adjusted means, use the
      LSMEANS statement.
ERROR: Must have at least 3 cells for WALLER.
ERROR: Must have at least 3 cells for WALLER.
```



```
PROC GLM DATA=berry;
 CLASS fertiliz variety;
 MODEL yield=fertiliz variety fertiliz*variety;
 MEANS fertiliz variety / lsd ;
 RUN;
```

```
General Linear Models Procedure

T tests (LSD) for variable: YIELD

NOTE: This test controls the type I comparisonwise error rate not the
      experimentwise error rate.

Alpha= 0.05  df= 20  MSE= 0.112333
Critical Value of T= 2.09
Least Significant Difference= 0.2854

Means with the same letter are not significantly different.

T Grouping           Mean      N  FERTILIZ

          A        8.9667     12  N
          A
          A        8.7167     12  K
```


END