Statistical Programs
College of Agriculture
and Life Sciences
http://www.uidaho.edu/ag/statprog/

Advanced Analyses

At times, it is advantageous to allow the coefficients of a regression analysis to be random. That is, the slope and intercept terms may possess random components to them. Such situation may arise when modeling data collected over several locations, years, or experiments.

Regression

Required Statements:

MODEL: Like Analysis of Variance, the MODEL statement for regression in PROC MIXED has dependent and independent variables and would have the form: MODEL Dependent var. = Independent var. The asterisk notation can also be used here which represents the interaction or multiplicative effects of more than one variable. An example might be MODEL y = X_1 X_1*X_2. The vertical bar notation is also applicable, but is not normally used. Nested terms are not typically used in regression models. Because the independent variables (regressor variables) are quantitative, the CLASS statement is not appropriate.

Additional Statements and Options:

MODEL Statement Options: In regression, it is usually of interest to examine predicted values and residuals. These values can be output into a SAS dataset with the option *outp*=<*name*> where name is the designation of the dataset. This dataset can then be used with plotting or summary procedures.

Some control over the model to be fitted is given by options such as INT and NOINT which tell SAS to fit a model with or without an intercept. SAS will fit an intercept by default. The SOLUTION option will print the estimated regression coefficients.

RANDOM: The syntax for the random statement is identical to that of ANOVA. Random effects for an intercept term may be specified with the keyword INTERCEPT.

CONTRAST: As with ANOVA, contrasts may be used concerning the regression coefficients by using the syntax outlined in handout #2.

Example 1 - Multiple Linear Regression

The data used in this example are taken from Hristov, A. N., W. J. Price, and B. Shafii. 2004. *A meta-analysis examining the relationship among dietary factors, dry matter intake, and milk and milk protein yield in dairy cows.* Journal of Dairy Science. 87: 2184-2196. The data was collected from a survey of past nutritional studies published in the Journal of Dairy Science (256 studies encompassing 846 different diets). One objective of the research was to regress the response, reported milk yields, on nutritional factors found in the study diets (DMI=Dry Matter Intake, MP = % metabolizable Protein, Fat=% fat). The slope coefficient for DMI and the intercept term were allowed to vary as random effects across studies.

The MIXED code to analyze the data was:

```
PROC MIXED DATA=NRC1 COVTEST;
        CLASS STUDY;
        WEIGHT WT_MILK;
        MODEL MILK=DMI FAT MP/SOLUTION OUTP = OVERALL;
        RANDOM INTERCEPT DMI/SUBJECT=STUDY TYPE=VC;
```

NOTES: In this example, the COVTEST option provides for significance tests (approximate Z-tests) on the random effects. A WEIGHT statement is also included to account for the different standard errors of each study. The SOLUTION option used with the model statement provides printed estimates of the regression coefficients and the OUTP option places the predicted values and residuals into a SAS dataset named OVERALL. In the RANDOM statement, the intercept term and coefficient for dry matter intake, DMI, are specified as random effects. The subject for the random effects is “study”. The TYPE=VC option requests that only variance components of the random effects be estimated. This option is the default and often is necessary in order to avoid computational limits caused by estimating too many parameters.

Modeling Correlation

In many field situations, samples are taken along transects. These subsamples may be spatially correlated. PROC MIXED allows the user to model such correlation.

Required Statements:

MODEL: The MODEL statement syntax is identical to the regular ANOVA situation..

RANDOM: The syntax for the random statement is identical to that of ANOVA.

REPEATED: The REPEATED statement is used to specify the spatial correlation variable and structure. Most common spatial models are available under the TYPE= option.

Additional Statements and Options:

CONTRAST: As with regular ANOVA, contrasts may be made on mean fixed effects using the syntax outlined in the previous handout.

Example 2 - Spatial Correlation ANOVA

In this example, the spatial correlation among subsamples taken within a location is modeled as a repeated measures process. The data used are derived from a study examining differences in biometrics of *Lepidium draba* (hoary cress) plants found in the continental United States and Europe (unpublished data, M. Schwarzlaender). In each continent, samples of *Lepidium* populations (24 in Europe and 26 in the US) were taken. Within each population, 30 quadrats were taken along a transect. Hence, the 30 transects in each population could potentially be spatially correlated. In addition, the populations taken represent a random sample of all populations present in each continent. These factors should be incorporated into any statistical model for the comparison of continents. The biometric used in this example will be *Lepidium* biomass.

```
PROC MIXED DATA=BIOMASS COVTEST;
        CLASS CONTINENT POPULATION;
        MODEL TOTAL_BIOMASS  =  CONTINENT;
        RANDOM POPULATION(CONTINENT)/
                SUBJECT =  POPULATION(CONTINENT);
        REPEATED /TYPE = SP(POW)(QUADRAT)
                SUBJECT =  POPULATION(CONTINENT);
        LSMEANS CONTINENT;
```

NOTES: As in the previous example, the COVTEST option allows for testing of the random effects. Because the populations sampled are unique to each continent, the random effect is specified as a nested term POPULATION(CONTINENT), read as populations within continent. The spatial correlation of quadrats is modeled using the REPEATED statement in conjunction with the TYPE=SP(POW) which specifies a spatial power model. The power model estimates the function $^{2\ d_{ij}}$ where is the spatial correlation and d_{ij} is the distance between two quadrats i and j. Thus, at $d_{ij} = 1$, measures the correlation between adjacent quadrats.