Additional Topics and Techniques

Weighted Regression: The regression analysis assumes a constant variance of responses across all levels of the regressors. If this is not the case **and** the data is replicated at each value of the regressor(s), it is possible to use a weighted regression to deal with the heteroskedasticity. This requires the data set to have a weight variable which is proportional to the inverse of the variance at each regressor value. Thus, before getting into the PROC REG procedure, the user must use a combination of DATA step and the PROC MEANS procedure to produce the weights.

Example:

```
PROC MEANS NOPRINT DATA=PHOTO;
     VAR PHOTO;
          OUTPUT OUT=VARIANCE VAR=VAR;
          BY IRRAD;

DATA PHOTO;
     MERGE PHOTO VARIANCE;
          BY IRRAD;
          WT = 1/VAR;

PROC REG DATA=PHOTO;
     WEIGHT WT;
     MODEL PHOTO = IRRAD;
```

The first step is to call PROC MEANS. The NOPRINT option is used because the printed output from MEANS is not of interest. The variances for each level of IRRAD are put in the variable VAR in the data set VARIANCE (NOTE: This requires more than one observation at each level of IRRAD which is not the actual case with our data). This new data set is then merged to the PHOTO data set by matching up the values of IRRAD. The weight variable, WT, is calculated by 1/variance. The last step calls PROC REG and invokes the WEIGHT statement with the WT variable. Typically if weighting is done in this manner and is appropriate, the regression estimates will change little in value, while their associated standard errors will be smaller.

Model Comparison: One of the most common uses of regression is the comparison of models based on different subsets of the data. This requires that the models be of the same mathematical form, i.e. parameter estimates from a linear model can not be statistcally compared to those of a quadratic model. Given this constraint, two or more models can be compared by combining the models into a single model via dummy variables. Once this model is fit, the user can compare appropriate combinations of parameters. In SAS, both PROC REG and PROC GLM can be used to accomplish this task.

PROC REG: In order to use PROC REG for dummy variable regression (DVR), the user must explicitly define the dummy variables and their associated interactions in a data step. This is done using the conditional IF statement with the subsetting variable in question.

Example:

```
data soy;
    input yield height trt;
    trt1=0;trt2=0;trt3=0;
    if trt=1 then trt1=1;
    else if trt=2 then trt2=1;
    else if trt=3 then trt3=1;

    t1_ht=trt1*height;
    t2_ht=trt2*height;
    t3_ht=trt3*height;

    cards;
          .
          .
          .
proc reg data=soy;
    model yield = trt1 trt2 trt3 t1_ht t2_ht
                  t3_ht/noint;

    int1_2: TEST trt1_trt2=0;
    int1_3: TEST trt1_trt3=0;

    slp1_2: TEST t1_ht_t2_ht=0;
    slp1_3: TEST t1_ht_t3_ht=0;

    line1_2: TEST trt1_trt2=0,
                  t1_ht_t2_ht=0;
    line1_3: TEST trt1_trt3=0,
                  t1_ht_t3_ht=0;
```

The actual comparisons in this example occur with the TEST statement. The form of TEST is: ***label: TEST <expression>;***. The label is an identifier for the contrast and will be printed on the output. The expression is any linear combination of parameters about which the hypothesis is to be tested. In the first case, the test is concerned with the difference between trt1 and trt2. If the difference is significantly different from zero, then the intercepts are different. Similar comparisons can be made for other parameters. These are one degree of freedom tests. The last two TEST statements compare both intercepts and slopes for the two treatments (whether the entire lines are different) with 2 degrees of freedom. The NOINT option is used because the intercepts are explicitly given in the model (trt1, trt2, trt3).

PROC GLM: The GLM procedure is somewhat less cumbersome than PROC REG for model comparison because the explicit definition of dummy variables is not required. The use of a CLASS statement in GLM allows SAS to define the dummy variables for us and the use of crossed terms in the model statement produces the appropriate interactions. Applying GLM this way was covered in the GLM workshop and more information can be found in handout #4 titled "Analysis of Covariance (ANCOVA) or Dummy Variable Regression (DVR)".

Example:

```
PROC GLM DATA=SOY;
     CLASS TRT;            <------defines dummy var's.
     MODEL YIELD = TRT TRT*HEIGHT/NOINT SOLUTION;
     CONTRAST 'int1_2' TRT 1 -1 0;
     CONTRAST 'int1_3' TRT 1 0 -1;

     CONTRAST 'slp1_2' TRT*HEIGHT 1 -1 0;
     CONTRAST 'slp1_3' TRT*HEIGHT 1 0 -1;

     CONTRAST 'line1_2' TRT 1 -1 0,
                        TRT*HEIGHT 1 -1 0;
     CONTRAST 'line1_3' TRT 1 0 -1,
                        TRT*HEIGHT 1 0 -1;
```

The model statement in GLM contains the terms TRT and TRT*HEIGHT. These codes correspond to the equivalent PROC REG variables (trt1, trt2, trt3) and (t1_ht, t2_ht, t3_ht), respectively. The NOINT option is again used because intercepts were given in the model. SOLUTION requests GLM to print out the parameter estimates. Testing of parameters is now accomplished using the CONTRAST statement and will produce identical results to the TEST's described earlier. The use of GLM makes model comparison easier, however, this step should not be followed until each model is appropriately selected and statistically examined on an individual basis -- this is best accomplished using PROC REG.