There are four tutorials in this presentation:
1. Detecting Outliers in Regression
   (Influence Statistics)
2. Heterogeneity of Slopes
3. Contrasts
4. Proper Error Terms for Split Plots.

Except for a brief introduction, this presentation consists of transparencies used in the tutorial. Full size copies suitable for making transparencies are available from the author.

DETECTING OUTLIERS IN REGRESSION

Outliers or unusual observations can bias estimates and invalidate statistical tests. Therefore we must try to detect outliers. (What to do with suspected outliers is another question). A number of statistics which may be useful in detecting outliers are discussed in Belsley, Kuh and Welsch, Regression Diagnostics (Wiley 1980). Many of these are available in PROC REG.

The "Classical" Method for outlier detection is to compute residuals:
\[ (y - \hat{y}) \]
and plot against \( y \) (or the independent variables):
\[ \text{RESID} = \text{ACTUAL} - \text{PREDICTED} \]

Such plots may not be useful because:
(i) residuals do not all have same variances
(ii) outliers tend to pull regression line towards themselves.

An alternate set of statistics, the standardized residuals:
\[ \text{ST RESID} = \frac{\text{RESID}}{\sqrt{s^2(1-h_i)}} \]

where
\[ s^2 = \text{residual mean square (estimate of } \sigma^2) \]
\[ h_i = \text{diagonal element of } X(X'X)^{-1}X' \]

This helps, but usually not much.

An outlier may cause residual variance to be overestimated, hence the standardized outliers do not look bad. Solution: When standardizing a residual, estimate variance based on a regression estimate using all other observations. Sounds difficult, but it isn't. The variance estimate, omitting observation \( i \),
\[ s^2(1) = (n-m-2)^{-1} (\text{SSE} - (\text{RESID})^2/1-h_i) \]

This is used in place of \( s^2 \) in standardizing, the resulting "STUDENTIZED" residual is
\[ \text{STUD-R} = \frac{\text{RESID}}{\sqrt{s^2(1)(1-h_i)}} \]

Outliers may also pull the estimated regression plane towards themselves thus their residuals may not be large. Solution: re-estimate each residual using regression model estimates that do not use that observation and studentize that residual. This statistic has been dubbed DFFITS and is also not difficult to compute:
\[ \text{DFFITS} = \frac{\text{RESID}}{s(1) \sqrt{1-h_i}} \]

These four statistics do not exhaust all possibilities, but they indicate basic approaches to this problem. Other statistics include:

- COOK'S D: function of DFFITS, but does not use \( s^2(1) \)
- DFBETAS: shows how regression coefficient estimates are charged by deleting observation
- COVRATIO: effect on covariance of estimates by deleting observation, and finally
- \( h_i \): measures influence of observation. It is not in itself an outlier detection statistic.

All of these statistics (and others yet to be discovered) may be useful. But they are not universally applicable, particularly in the case of two or more outliers.

WHAT CAN YOU GET FROM SAS?
Use PROC REG:
Using R (RESIDUALS) option, in the model statement prints
- PREDICTED and RESID
- Standardized residuals (called STUDENT RESIDUAL)
- COOK'S D, and
- a simplified plot of the standardized residuals

Using INFLUENCE option in the model statement prints
- Studentized residual (STUD_R, called STUDENT)
- DFFITS
- DFBETAS (for each coefficient)
- COVRATIO
- \( h_i \) (HAT DIAG)

These cannot be plotted directly.

Using OUTPUT, you can make a new data set and plot and print
- RESID
- COOK'S D
- Standard error of residuals, which can be used to get Standardized residuals (called STUDENT RESIDUAL)

These statistics are illustrated on a simple generated example
\[ y_i = x_i + \varepsilon; \quad i=1,2,\ldots,11, \quad \varepsilon \sim N(0,1) \]
\[ y_1 = 21, \quad y_{12} = 17, \quad \text{an outlier which is } 40 \]
above the true regression.
EXAMPLE FOR ROW DIAGNOSTICS

SAS TUTORIAL
Testing for Equality of Slopes

We observe a dependent and one or more independent variables in several groups or treatments. We want to test if the regression coefficient(s) are the same for all the groups.

Example 1

It is desired to relate the circumference (CIRCUM) of trees to the number of rings (RINGS) in the trunk. The relationship between rings and circumference may be affected by where a tree is growing, data on rings and circumference were recorded for randomly selected trees on the north and south (SIDE, coded N and S) slopes of some foothills in Colorado.

The required SAS statements are:

```sas
PROC GLM DATA=SASDATA;
CLASS SIDE;
MODEL CIRCUM = RINGS SIDE RINGS*SIDE;
ESTIMATE 'INT NORTH' INTERCEPT 1 SIDE 1;
ESTIMATE 'INT SOUTH' INTERCEPT 1 RINGS 1;
ESTIMATE 'INT NORTH' INTERCEPT 1 SIDE 1;
ESTIMATE 'INT SOUTH' INTERCEPT 1 RINGS 1;
ESTIMATE 'INT NORTH' INTERCEPT 1 RINGS 1;
ESTIMATE 'INT SOUTH' INTERCEPT 1 RINGS 1;
```

The interaction (RINGS x SIDE) measures the difference in slope (RINGS) for the two SIDE’s.
DEPENCENT \fARLo..E: CIRCUtI
Sm OF' SQUM£S
1.16341222
7.96111111
22.84
0.0011
0.0001
6.31
0.0106
0.153049
3.4906
0.21441042
0.24569167
0.45
0.28
0.19
0.23
0.0001
0.0001
0.0001

NOTE: If slopes are not different, appropriate parameter estimates can only be obtained by redoing the analysis without the interaction.
The SAS statements are:

```sas
PROC GLM CLASS MAX1;
MODEL Y = T TWT MAX TRIAX T1;
ESTIMATE 'LIN MAX1' TIME 1 TIMESTEP 1;
ESTIMATE 'QUAD MAX1' TIME 1 TIMESTEP 1 TIMESTEP 4;
ESTIMATE 'QUAD MAX1' TIME 1;
ESTIMATE 'LIN MAX1' TIME 1 TIMESTEP 0 0;
ESTIMATE 'QUAD MAX1' TIME 1 TIMESTEP 0 0 0;
ESTIMATE 'QUAD MAX1' TIME 1;
```

The output:

The appropriate analysis requires a polynomial response surface with orthogonal polynomials, which are a special type of contrast. Using CONTRAST statements with PROC GLM:

```sas
PROC GLM CLASS CA PH;
MODEL Db' = CA PH;
CONTRAST 'PH Linear' PH -1 -1 1 1;
CONTRAST 'PH Quadratic' PH 1 -2 1;
CONTRAST 'CA Linear' CA -1 0 1;
CONTRAST 'CA Quadratic' CA 1 -2 1;
CONTRAST 'PHCA Linear by PHCA Quadratic' PH 3 1 -3 0 0 0 2 -1 1 1;
```

Contrasts can be computed to SAS by two methods:

(i) using the CONTRAST statement in PROC REG
(ii) using PROC REG

A factorial experiment was set up to compare the effect of different soil pH and calcium additives on the increase in trunk diameters for orange trees. pH levels are 4.5, 6, and 7, and levels of calcium supplement are 100, 200, and 300 pounds per acre resulting in a 4x3 factorial. At the end of a two-year period, 3 diameters were examined at each factor-level combination. The appropriate analysis requires a polynomial response surface with orthogonal polynomials, which are a special type of contrast.
Contrast coefficients \((c_i)\) are specified for each hypothesis desired. Note that the interaction contrast requires listing of coefficients for each cell.

Using PROC REG

Contrast coefficients are simply sets of coded independent variables. These variables are generated in the DATA step. The expression inside each set of parentheses is evaluated: if it is true then the value is 1, if false it is 0. The same result can be obtained by a set of IF statements. The linear by linear interaction contrast is simply the product of the two linear contrasts.

### CONTRAST COEFFICIENTS

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>SS</th>
<th>F Value</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>2</td>
<td>16.82</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>PH</td>
<td>3</td>
<td>21.94</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>CARPH</td>
<td>6</td>
<td>9.00</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>CONTRAST</td>
<td>DF</td>
<td>SS</td>
<td>F Value</td>
<td>PR &gt; F</td>
</tr>
<tr>
<td>PHIN</td>
<td>1</td>
<td>16.74</td>
<td>0.0004</td>
<td></td>
</tr>
<tr>
<td>PH QUADRATIC</td>
<td>1</td>
<td>48.79</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>CALIN</td>
<td>1</td>
<td>0.74</td>
<td>0.3770</td>
<td></td>
</tr>
<tr>
<td>CA QUADRATIC</td>
<td>1</td>
<td>20.90</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>PHIN CA LIN BY LIN</td>
<td>1</td>
<td>18.00</td>
<td>0.0002</td>
<td></td>
</tr>
</tbody>
</table>

---

### GENERAL LINEAR MODELS PROCEDURE

**DEPENDENT VARIABLE: DIAM**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>11</td>
<td>9.16309556</td>
<td>0.83482323</td>
<td>12.32</td>
<td>0.0001</td>
</tr>
<tr>
<td>ERROR</td>
<td>24</td>
<td>1.42644447</td>
<td>0.06777778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONNECTED TOTAL</td>
<td>25</td>
<td>10.90722222</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-SQUARE</td>
<td>C.V.</td>
<td>STD DEVI</td>
<td>DIAN MEAN</td>
<td>0.847519</td>
<td>3.4913</td>
</tr>
</tbody>
</table>

**SOURCE**

<table>
<thead>
<tr>
<th>DF</th>
<th>TYPE</th>
<th>SS</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>1</td>
<td>16.82</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>PH</td>
<td>3</td>
<td>21.94</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>CARPH</td>
<td>6</td>
<td>9.00</td>
<td>0.0001</td>
<td></td>
</tr>
</tbody>
</table>

**SOURCE**

<table>
<thead>
<tr>
<th>DF</th>
<th>TYPE</th>
<th>SS</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>1</td>
<td>16.82</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>PH</td>
<td>3</td>
<td>21.94</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>CARPH</td>
<td>6</td>
<td>9.00</td>
<td>0.0001</td>
<td></td>
</tr>
</tbody>
</table>

**CONTRAST**

<table>
<thead>
<tr>
<th>DF</th>
<th>SS</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHIN</td>
<td>1</td>
<td>16.74</td>
<td>0.0004</td>
</tr>
<tr>
<td>PH QUADRATIC</td>
<td>1</td>
<td>48.79</td>
<td>0.0001</td>
</tr>
<tr>
<td>CALIN</td>
<td>1</td>
<td>0.74</td>
<td>0.3770</td>
</tr>
<tr>
<td>CA QUADRATIC</td>
<td>1</td>
<td>20.90</td>
<td>0.0001</td>
</tr>
<tr>
<td>PHIN CA LIN BY LIN</td>
<td>1</td>
<td>18.00</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
The output from PROC REG

DEPENDENT VARIABLE: DIAM

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARE</th>
<th>F VALUE</th>
<th>PROB &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>11</td>
<td>9.18300556</td>
<td>0.83482323</td>
<td>12.322</td>
<td>0.0001</td>
</tr>
<tr>
<td>ERROR</td>
<td>24</td>
<td>1.46722222</td>
<td>0.06777778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORRECTED TOTAL</td>
<td>35</td>
<td>10.64777777</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-SQUARE | C.V. | STD DEV | DIAM MEAN | 0.60518 | 2.4013 | 0.260325 | 7.07177777 |

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>F VALUE</th>
<th>PROB &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>2</td>
<td>1.46722222</td>
<td>10.622</td>
<td>0.0004</td>
</tr>
<tr>
<td>PH</td>
<td>3</td>
<td>4.46683333</td>
<td>21.744</td>
<td>0.0001</td>
</tr>
<tr>
<td>CM PH</td>
<td>6</td>
<td>3.25500000</td>
<td>18.890</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Since these are orthogonal, the type II SS are independent, but the tests must be done sequentially starting with the highest order terms. Since the linear by linear interaction is significant the CA-linear term must be retained.

WHICH METHOD SHOULD YOU USE?

If data are balanced and contrasts are orthogonal both methods give identical results.

PROC GLM gets expensive for large experiments.

 Interaction contrasts in GLM require the listing of coefficients for each interaction cell.

Predicted values from PROC GLM are for the factorial, from PROC REG, for the chosen polynomial model.

Lack of fit tests can be done automatically with PROC GLM (not shown here).

If data are balanced but contrasts are not orthogonal, PROC REG adjusts each contrast for all other contrasts.

If data are not balanced PROC GLM must, of course, be used for the factorial. Contrasts are not (usually) orthogonal and are not fully adjusted. You may want to use PROC REG.

Lack of fit tests can be performed by combining results of the two analyses.
SAS TUTORIAL

GETTING PROPER ERROR TERMS FOR SPL1T PLOT DESIGNS

The split-plot design is a specialized design for a factorial experiment. While one factor is readily applied to large experimental units or main plots, another factor may be applied to smaller units or subplots within the larger unit.

Treatments applied to main plots are subject to larger experimental error than those applied to subplots; hence different variances must be used as denominators for the corresponding F ratios.

As a byproduct, this experimental design concentrates more information on treatments applied to the smallest areas than on the treatments applied to large areas.

EXAMPLE

Two cultivars (CULT) of grasses applied to main plots.

Three bacterial inoculations (INOC) applied to subplots.

Four replications.

A standard analysis of variance for this experiment is:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARE</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication</td>
<td>3</td>
<td>20,330,000</td>
<td>6,776,666</td>
<td>5.33</td>
<td>0.007</td>
</tr>
<tr>
<td>Cultivar</td>
<td>2</td>
<td>1,388,000</td>
<td>694,000</td>
<td>0.56</td>
<td>0.430</td>
</tr>
<tr>
<td>Replication X Cultivar (Error A)</td>
<td>6</td>
<td>5,388,000</td>
<td>898,000</td>
<td>0.69</td>
<td>0.624</td>
</tr>
<tr>
<td>INOC</td>
<td>2</td>
<td>1,212,000</td>
<td>606,000</td>
<td>0.48</td>
<td>0.647</td>
</tr>
<tr>
<td>Cultivar INOC (Error B)</td>
<td>12</td>
<td>7,128,000</td>
<td>594,000</td>
<td>0.43</td>
<td>0.767</td>
</tr>
</tbody>
</table>

Here Error B is equal to the default residual for INOC and CULT*INOC hence the F-values and associated p-values (PR > F) are obtained directly from the ANOVA table.

Error A is the REP* CULT interaction and the proper test is obtained from the TEST statement. The F and p-value for CULT in the ANOVA tables may not be used.

SAS statements required are:

```sas
PROC ANOVA;
CLASS REP CULT INOC;
MODEL DRYWT = REPLIC CULT REPLIC*INOC CULT*INOC;
TEST INOC=CULT;
TEST INOC=CULT*REP;
```

the resulting output:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARE</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>REP</td>
<td>3</td>
<td>20,330,000</td>
<td>6,776,666</td>
<td>5.33</td>
<td>0.007</td>
</tr>
<tr>
<td>Cultivar</td>
<td>2</td>
<td>1,388,000</td>
<td>694,000</td>
<td>0.56</td>
<td>0.430</td>
</tr>
<tr>
<td>Replication X Cultivar (Error A)</td>
<td>6</td>
<td>5,388,000</td>
<td>898,000</td>
<td>0.69</td>
<td>0.624</td>
</tr>
<tr>
<td>INOC</td>
<td>2</td>
<td>1,212,000</td>
<td>606,000</td>
<td>0.48</td>
<td>0.647</td>
</tr>
<tr>
<td>Cultivar INOC (Error B)</td>
<td>12</td>
<td>7,128,000</td>
<td>594,000</td>
<td>0.43</td>
<td>0.767</td>
</tr>
</tbody>
</table>

TESTS OF HYPOTHESES USING THE ANOVA HS FOR REPLIC AS AN ERROR TERM

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>ANOVA SS</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>REP</td>
<td>3</td>
<td>20,330,000</td>
<td>5.33</td>
<td>0.007</td>
</tr>
<tr>
<td>Cultivar</td>
<td>2</td>
<td>1,388,000</td>
<td>0.56</td>
<td>0.430</td>
</tr>
<tr>
<td>Replication X Cultivar (Error A)</td>
<td>6</td>
<td>5,388,000</td>
<td>0.69</td>
<td>0.624</td>
</tr>
<tr>
<td>INOC</td>
<td>2</td>
<td>1,212,000</td>
<td>0.48</td>
<td>0.647</td>
</tr>
<tr>
<td>Cultivar INOC (Error B)</td>
<td>12</td>
<td>7,128,000</td>
<td>0.43</td>
<td>0.767</td>
</tr>
</tbody>
</table>

Here Error B is equal to the default residual for INOC and CULT*INOC hence the F-values and associated p-values (PR > F) are obtained directly from the ANOVA table.

Error A is the REP* CULT interaction and the proper test is obtained from the TEST statement. The F and p-value for CULT in the ANOVA tables may not be used.

We consider a split-split-plot design with main-plot factor A, subplot factor B, and sub-sub-plot factor C to illustrate some principles of sums of squares computations with the ANOVA procedure. A standard analysis of variance is:

<table>
<thead>
<tr>
<th>Source</th>
<th>Replications</th>
<th>A</th>
<th>Replications xA (Error A)</th>
<th>B</th>
<th>AXB</th>
<th>Replications xB (Error B)</th>
<th>C</th>
<th>AxC</th>
<th>BxC</th>
<th>AXBxC</th>
<th>(Error C)</th>
</tr>
</thead>
</table>

Error (A) used to test A
Error (B) used to test B, AXB
Error (C) used to test C, AxC, BxC, AXBxC

SAS Statements
PROC ANOVA;
MODEL Y = REPS A/B/C;
give all 15 lines above but does not provide Error (B) and Error (C)

The analysis of variance and correct tests are given by these statements:

PROC ANOVA;
CLASS REP A B C;
MODEL RESPONSE=REP A REP*A B A*B REP*B(A) C A*C B*C A*B*C;
TEST HA=E=REP*A;;
TEST HB=A*B E=REP*B(A);

Another feature of PROC ANOVA (and PROC GLM) is that MODEL statements containing interaction terms without one or more of the corresponding main effect terms produces interaction sums of squares that contain the non-specified main effect sums of squares. For example, the statement

MODEL Y=A*B

produces
SS(AB)=SS(A) + SS(B) + SS(AB),
while the statement

MODEL Y=A*B

produces
SS(A*B)=SS(B) + SS(A*B).

Thus, in the above split-split-plot experiment, Error B may be obtained by specifying either

- ReplicationsxB(A)
- ReplicationsxB(B)

or

- ReplicationsxB(A)
- ReplicationsxBxA.