Model Based Statistics in Biology.

Part IV. The General Linear Model. Multiple Explanatory Variables.

Chapter 14.2 ANCOVA - Statistical Control

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11)
ReCap Multiple Regression (Ch 12)
ReCap Multiple Categorical Variables (Ch 13)
14.1 Comparing Regression Lines
14.2 Statistical Control
14.3 Model Revision
14.4 More than two explanatory variables (to be written)

on chalk board

CrwTb9_1.xls
Ch14.xls

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning is based on models, including statistical analysis based on models.
ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to declare a decision.
Estimation is concerned with the specific value of an unknown population parameter.
ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12) GLM with more than one regression variable (multiple regression)
ReCap (Ch 13) GLM with more than one categorical variable (ANOVA).
ReCap (Ch 14) ANCOVA with GLM - Comparing regression lines.

Today: Statistical control, with ANCOVA.
Statistical control allows the effects of one variable to be removed,
in order to arrive at a better analysis of the effects of another variable.

Wrap-up.

Statistical control improves analysis be removing the effects of a secondary variable, to achieve lower error MS and better analysis of the variable of interest.

In ANCOVA either the ratio scale or the nominal scale explanatory variable can be the control variable. A ratio scale variable (e.g. fish production from lakes) can be analyzed relative to a ratio scale variable (e.g. size of lake) controlled for a nominal scale variable (e.g. temperate versus tropical lakes). Or a nominal scale variable (e.g. experimental treatment versus control) can be tested controlling for the effects of a ratio scale variable (e.g. metabolic rate of the animal).

Of these two possibilities, the more commonly encountered is that of a classification (nominal scale) variable, controlled for a ratio scale control. An example of this was worked through today.

Introduction.

ANCOVA is applied to data situations that have a mixture of both ratio and nominal scale explanatory variables. We have already looked at ANCOVA where we compare slopes of one or more regression lines, using the interaction term in the ANCOVA model. Today we will look at another application of ANCOVA, where we compare several groups (ANOVA explanatory variable) controlling for the effects of a second explanatory variable (regression variable on a ratio type of scale.). To do this analysis we will need to establish that the slopes are the same in the groups (no interaction term).

Data from Table 9.1 in M.J. Crawley (1993) *GLIM for Ecologists*.

The data consist of seed production in 40 plants allocated at random to two treatments, grazed and ungrazed.

The grazed plants were exposed to rabbits during the first two weeks of stem elongation, then protected from subsequent grazing.

The size of the plant was thought to influence seed production so the diameter at the top of the root stock (in mm) was measured <u>before</u> exposure to grazing.

At end of growing season, fruit production (M_{fruit} = mg dry wt) wasrecorded for each of the 40 plants.

fruit (mg)	root (mm)	grazed
59.77	6.225	n
60.98	6.487	n
14.73	4.919	n
19.28	5.13	n
34.25	5.417	n
35.53	5.359	n
87.73	7.614	n
63.21	6.352	n
24.25	4.975	n
64.34	6.93	n
52.92	6.248	n
32.35	5.451	n
53.61	6.013	n
54.86	5.928	n
64.81	6.264	n
73.24	7.181	n
80.64	7.001	n
18.89	4.426	n
75.49	7.302	n
46.73	5.836	n
80.31	8.988	y
82.35	8.975	y
105.1	9.844	y
73.79	8.508	y
50.08	7.354	y
78.28	8.643	y
41.48	7.916	y
98.47	9.351	y
40.15	7.066	y
116.1	10.25	y
38.94	6.958	y
60.77	8.001	y
84.37	9.039	y
70.11	8.91	y
14.95	6.106	y
70.7	7.691	y
71.01	8.515	y
83.03	8.53	y
52.26	8.158	y
46.64	7.382	y

1. Construct model

<u>Verbal model.</u>

Fruit production depends on grazing and root size.
Is the difference in fruit production between grazed and ungrazed plants significant <u>after</u> we control for the relation to root size?

1. Construct model

Graphical model.

Fruit production in relation
to grazing pressure.

```
         _             *
         _
      105+             *
         _             *
 fruit   _   *
production   *         3
 (mg)    _   *         3
       70+   *         3
         _   5         *
         _   3
         _   *         3
         _             3
       35+   3
         _
         _   3
         _   *         *
         _
           --+---------+----grazing
          0.00 = no   1.00 = yes
```

Fruit production also depends on root size

```
         _                                                      G
         _                                           Grazed
      105+                                                   G
         _                                               G
 fruit   _               Ungrazed         U
production                           U            G   2
 (mg)    -                              U         GG  G
       70+                             U   G      G  G
         _                     2UU   U        G
         _                  UU U
         _                  U           GG     G
         _                           GG      G
       35+              3
         _
         _      U   U U
         _          U         G
         _
           --------+---------+---------+---------+---------+--------root
                 4.8       6.0       7.2       8.4       9.6
```

Response variable = M_{fruit} = fruit production (mg dry wt)
Explanatory variable = Gr = ungrazed (0) or grazed (1)
Explanatory variable = root =diameter (mm)

Formal model

Write formal model (GLM)

```
Sketch a graph above each term
```

$$M_{fruit} = \$_o + \$_{root} * root + \$_{Gr} * Gr + \$_{Root*Gr} * root * Gr + ,$$

This is our preliminary model to test whether slope are parallel. If slopes are parallel (no interaction term) then we are going to revise the model by removing the interaction term, so we can test for grazing effects controlled for plant size (root diameter)

The goal is to remove the effects of root size, which is the regression variable. To do this, when need to show that root size has the same effect on seed production in both groups. In other words, we need to show that the slopes are the same. In statistical terms, we need to show that there is no interaction term.

Consequently, the analysis will proceed in 2 cycles through the generic recipe. First pass: slopes homogeneous ? Second pass: grazing effects ? (root effects removed if slopes homogeneous).

2. Execute analysis.

Place data in model format:

Column labelled M_{fruit} the response variable fruit production (mg dry wt)
Column labelled Graze with explanatory variable Gr: ungrazed=0, grazed=1
Column labelled Root with explanatory variable Root = diameter

2. Execute analysis.

Code the model statement in statistical package according to the GLM

$$M_{fruit} \;=\; \$_o \;+\; \$_{root}\cdot Root \;+\; \$_{Gr}\cdot Gr \;+\; \$_{root\cdot Gr}\cdot\, Root\cdot\, Gr \;+\; ,$$

```
MTB > glm   'Mfruit' = 'root'   'Gr'     'root'*'Gr';
SUBC> covariate  'root';
SUBC> fits c4;
SUBC> residuals c5.
```

Fits and residuals from:

model statement output of fitted values and residuals (as above)
or parameters reported by GLM routine
or direct calculation of parameters

Here are the parameter estimates.

The overall mean fruit production is $\hat{\beta}_o$ = 59.41 mg

The mean for grazed and ungrazed is expressed as a deviation from $\hat{\beta}_o$

$$\hat{\beta}_o + \hat{\beta}_{GR} = \begin{cases} \text{mean}\left(M_{\text{GR=no}}\right) & = \quad 59.41 \;-\; 8.53 \;=\; 50.88 \\ \text{mean}\left(H_{\text{GR=yes}}\right) & = \quad 59.41 \;+\; 8.53 \;=\; 67.94 \text{ mg} \end{cases}$$

The slope parameter for grazed and ungrazed together is $\hat{\beta}_{root}$ = 23.625 mg/mm

Note that the ANCOVA estimate of the slope differs from the slope estimate by simple regression, without the grazing term in the model.

```
MTB > regress 'fruit' 1 'root'.

 The regression equation is
 fruit = - 41.3 + 14.0 root

 Predictor        Coef       Stdev    t-ratio        p
 Constant       -41.31       10.73      -3.85    0.000
 root           14.026       1.464       9.58    0.000
```

This is because the ungrazed plants are smaller, hence to the left of the grazed plants in the graph. This lateral offset reduces the overall slope from around 23 mg/mm in each group to 14.0 mg/mm across all the data.

The deviation from the ANCOVA estimate of the overall slope are small.

$$\hat{\beta}_{Root*Gr} \quad = \quad \begin{array}{l} !\,0.371 \text{ mg/mm} \\ \\ +0.371 \text{ mg/ mm} \end{array}$$

$$\hat{\beta}_{root} + \hat{\beta}_{root*GR} = \begin{cases} Slope\left(H_{pers}\right) & = \quad 23.625 \;\; - \;\; 0.371 \; = \; 23.996 \\ Slope\left(H_{pseu}\right) & = \quad 23.625 \;\; + \;\; 0.371 \; = \; 23.254 \end{cases}$$

These particular deviations are symmetrical because there are only two groups.

Compare to regression equation (one slope and one intercept) for each species:
$H_{Gr=No}$ = –94.367 + 23.996 *Root*
$M_{Gr=Yes}$ = –125.28 + 23.254 *Root*

The GLM routine computes fitted and residual values.

3. Evaluate the model Plot residuals versus fitted values.

```
MTB > plot 'res' 'fits'
  res     _
          _                        *
          _
        12+
          _                         *        *
          _   *                 *            *        *
          _                    * *  **
          _              **                 * *             *    *
         0+          *      *  2      **          2 *
          _     *         *         *       *     *
          _               *                    **
          _                                 *
          _         * *
       -12+                             *        *
          _
          _                          *
          _
            ------+---------+---------+---------+---------+---------+fits
                 20        40        60        80       100       120
```

1. Straight line assumption acceptable. No bowls or arches in plot
2. If n small, evaluate assumptions for p-values from chisquare (t, F) distributions.
   n = 40, so even substantial deviations will have little distorting effect on calculation of p-values.

2a. Homogeneous? Yes
Residuals do not change in any systematic way with fitted values (no cones).
b. Sum(res) = 0? Yes

3. Evaluate the model

c. Independent?

Each residual plotted against its neighbor, data presumably in order it was taken.

```
MTB > let c11 = lag('res')
MTB > plot c11 'res'
C11     _
        _                           *
        _
      12+
        _            *         *
        _                     *         *       *  *
        _                            **   *     *
        _    *                       ** **  *
       0+            *   * *     *   *  * *  *
        _                       *   *   *           *         *
        _                               *  *    *
        _                         *
        _              *           *
     -12+                          *  *
        _
        _                                      *
        _
          --------+---------+---------+---------+---------+--------res
              -14.0      -7.0       0.0       7.0      14.0
```

No evidence of non-independence.

d. Normal?

The residuals look normal when plotted as a histogram.

Residuals are normal, homogeneous, and independent.

```
MTB > hist 'res';
SUBC> increment .04.

Histogram of res   N = 14

Midpoint   Count
 -0.0800       2  **
 -0.0400       1  *
  0.0000       7  *******
  0.0400       3  ***
  0.0800       1  *
```

4. State population and whether sample is representative.

Population might be that from which the plants were selected.
In this example, the population will be taken as all possible measurements, given the protocol.

5. Decide on mode of inference. Is hypothesis testing appropriate?

It is clear that fruit production depends on root size. It is not clear whether fruit production depends on grazing, after controlling for effects of root size. Hypothesis testing appropriate.

6. State H_A H_o pairs, test statistic, distribution, tolerance for Type I error.

Terms in model.

We begin with the interaction term. Are slopes parallel ?

H_A: var($\$_{Root*Gr}$) > 0

H_o: var($\$_{Root*Gr}$) = 0

This is equivalent to following hypotheses concerning parameters

$\$_{root*Gr=0} \neq \$_{root*Gr=1}$ (slope not parallel)

$\$_{root*Gr=0} = \$_{root*Gr=1}$ (slopes parallel)

If slopes are homogeneous (H_o rejected) then test for effects of grazing pressure.

$\$_{Gr=0} \neq \$_{Gr=1}$ (group means differ)

$\$_{Gr=0} = \$_{Gr=1}$ (group means do not differ)

State test statistic	F-ratio
Distribution of test statistic	F-distribution
Tolerance for Type I error	5% (conventional level)

7. ANOVA

```
MTB > glm 'fruit' = 'root' 'grazing' 'root'*'grazing';
SUBC> covariate 'root';
SUBC> fits c8;
SUBC> residuals c9.

Factor   Levels Values
grazing       2     0     1

Analysis of Variance for fruit

Source         DF     Seq SS     Adj SS     Adj MS        F       P
root            1    16800.4    18791.6    18791.6   402.57   0.000
grazing         1     5266.7      157.1      157.1     3.37   0.075
grazing*root    1        4.6        4.6        4.6     0.10   0.754
Error          36     1680.5     1680.5       46.7
Total          39    23752.2
```

8. When assumptions not met, decide whether to re-compute p-value.

Assumptions met, continue to next step.

9. Declare decision $\$_{root*Gr=0} = \$_{root*Gr=1}$ (slopes are parallel)

$F_{1,36} = 0.10$, for which $p = 0.754$

Interactive effect not significant, so we examine the grazing term.

It is close to the 5% criterion. Note, however the substantial difference between the Seq SS and the Adj SS of the grazing term.

The F-ratio for the grazing term, controlled for root size in sequential SS is:

F = (5266.7/1) / (1680.5/36) = 112.86 $p < 0.001$

Back to step 1.

1. Construct Model

$$Mfruit = \$_o + \$_{root} * Root + \$_{Gr} * Gr + ,$$

This is our model to test for grazing effects controlled for plant size (root diameter). The interaction term has been removed.
This simplified model is often used in ‘ANCOVA’ routines aimed at statistical control.

In this example, we tested the assumption of no interaction, rather than blindly assuming it to be true.

2. Execute analysis.
3. Evaluate model

Plot residuals vs fits

```
MTB > plot 'res' 'fits'
  res     _
          _                        *
          _
        12+
          _                         *        *
          _   *                 *            *
          _                    * * * *                *
          _              **                  *                  *
         0+     *        *     2      **    *      **       *
          _          *    *         *       *     *
          _              *                    ** *
          _                                 *
          _         *  *
       -12+                             *        *
          _
          _                          *
          _
            ------+---------+---------+---------+---------+---------+fits
                 20        40        60        80       100       120
```

1. Straight line acceptable, no bowls or arches.
2. a. Var(error) = constant ? Yes. No cones.
   b. E(error) = 0
   c. Independent errors ? Yes (not shown)
   d. Normal errors ? Yes. Histogram OK, so no further diagnosis

4. Population, sample
No change

5. Hypothesis testing? Yes.

```
MTB > hist 'res'
    Histogram of res   N = 40
Midpoint   Count
     -15       1  *
     -10       4  ****
      -5       6  ******
       0      17  *****************
       5       9  *********
      10       2  **
      15       1  *
```

6. State hypothesis H_A / H_o

Terms in model. Only one term will be examined, the grazing effect.

H_A: Var($\$_{Gr}$) > 0

H_o: Var($\$_{Gr}$) = 0

Equivalent to following hypotheses for parameters.

H_A: $\$_{Gr=0} \neq \$_{Gr=1}$ (grazing affect growth, controlled for size)

H_o: $\$_{Gr=0} = \$_{Gr=1}$

We can state a more specific hypothesis about the parameter, based on the biology.

H_A: $\$_{Gr=0} > \$_{Gr=1}$ (grazing reduces growth, controlled for size)

H_o: $\$_{Gr=0} \leq \$_{Gr=1}$

We are not interested in testing whether seed production depends on root size, it is obvious from the plot that it does.

7. ANOVA Table.

```
MTB > glm 'fruit' = 'root' 'grazing';
SUBC> covariate 'root';
SUBC> fits c8;
SUBC> residuals c9.
   Factor   Levels Values
   grazing       2     0     1
Analysis of Variance for fruit

Source     DF     Seq SS     Adj SS     Adj MS       F      P
root        1      16800      19155      19155  420.60  0.000
grazing     1       5267       5267       5267  115.64  0.000
Error      37       1685       1685         46
Total      39      23752
```

8. Recompute Type I error?

No need to do this. Type I error is far from the 5% criterion.

9. Declare decision.

Reject H_o. The observed difference in growth, controlled for root size, is not due to chance. $F_{1,37} = 115.64$ $p < 0.00001$

10. Analysis of parameters of biological interest.

```
           grazing       N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
 fruit           0      20     50.88     54.24     50.84     21.76      4.87
                 1      20     67.94     70.85     68.21     24.97      5.58
```

When root size is not taken into account, the fruit production appears to be less for ungrazed than for grazed.

Ungrazed	50.88 mg
Grazed	$-\underline{67.94\text{ mg}}$
Difference	–17.06 mg

This is because the grazed plants were larger than the ungrazed plants.

To compare grazed vs ungrazed, controlled for size, we calculate the vertical separation between the two regression lines. The most convenient point at which to do this is the point at which x = zero (the y-intercepts).

$$
\begin{array}{lllllll}
\hat{\alpha} & = & \$_{o} & - & \$_{root} & * & \text{mean(X)} \\
 & & & & & & \\
\hat{\alpha}_{Gr=no} & = & \text{Mean}(\text{M}_{\text{Gr=no}}) & - & \$_{root} & * & \text{Mean}(\text{root}_{\text{GR=no}}) \\
 & = & 50.88 & - & 23.6 & * & 6.053) \\
 & = & -91.729\text{ mg} & & & & \\
 & & & & & & \\
\hat{\alpha}_{Gr=yes} & = & \text{Mean}(\text{M}_{\text{Gr=Yes}}) & - & \$_{root} & * & \text{Mean}(\text{root}_{\text{GR=Yes}}) \\
 & = & 67.94 & - & 23.6 & * & 8.309) \\
 & = & -127.82\text{ mg} & & & &
\end{array}
$$

The intercept for grazed is below that for ungrazed.
The vertical separation between the two regression lines is:

Ungrazed	–91.729 mg
Grazed	$\underline{-(-127.820)}$ mg
Difference	36.091 mg

When root size is taken into account, the fruit production for grazed plants is less than for ungrazed. The fruit production for grazed plants was less by 36 mg.