**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 residual mean square 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 response variable (*e.g.* fish production from lakes) can be analyzed relative to a ratio scale explanatory variable (*e.g.* size of lake) controlled for a nominal scale variable (*e.g.* temperate versus tropical lakes). Or a nominal scale explanatory variable (*e.g.* experimental treatment versus control) can be tested controlling for the effects of a ratio scale explanatory variable (*e.g.* metabolic rate of the animal).

Of these two possibilities, the more commonly encountered is that of a classification (nominal scale) explanatory variable, controlled for a ratio scale variable. 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 use the interaction term to compare slope across a categorical variable. Today we will look at another application of ANCOVA, where we use the regression variable to control for its effects in comparing means across a categorical varible. This analysis assumes homogeneous slopes, which we can evaluate with the interaction term.

Data are from Table 9.1 in M.J. Crawley (1993) *GLIM for Ecologists*. The data are reported on p 288 in Crawley (2003) *Statistical Computing* as fruit biomass of the plant *Ipomopsis*. The Scarlet Gilia *Ipomopsis aggregata* subsp *weberi* is a rare and endangered plant endemic to the Park Mountain Range in Colorado and the Sierra Madre Range in Wyoming. The biomass of the fruit is a measure of seed production by each plant.

As reported in Crawley (2003) 40 plants were allocated to two treatments, grazed or not grazed by rabbits.

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

Seed production depends on plant size, which is measured as the diameter at the top of the root stock (in mm). Root diameter was measured before exposure to grazing.

At end of growing season, fruit biomass ($M_{fruit}$ = mg dry wt) was recorded 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

Verbal model.

Rabbit grazing reduces fruit biomass and hence seed production, once we control for the relation of production to root size.

## 1. Construct model

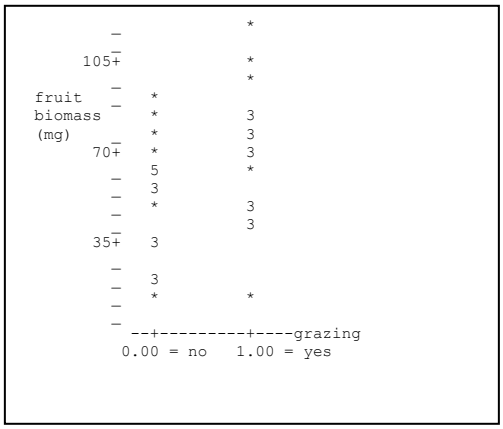
Table of variables

| Symbol | Name | Units | Scale type | Functional placement | Random/ Fixed |
|---|---|---|---|---|---|
| $M_{fruit}$ | Fruit | mg dry weight | ratio | Response | |
| *Root* | Root | mm (diameter) | ratio | Explanatory | Fixed |
| *Gr* | Grazed | Y/N | categorical | Explanatory | Fixed |

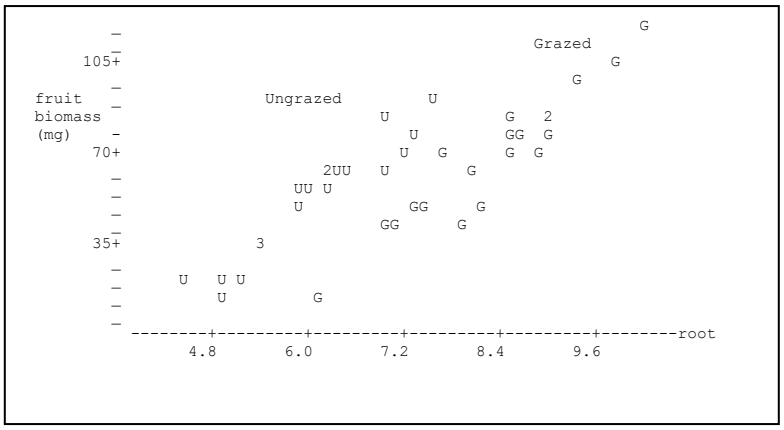
Root diameter is listed here as fixed because fruit production depends on plant size, hence root size. Random units, such a blocks, are not listed.

Graphical model.

Fruit biomass in relation to grazing pressure.




Fruit biomass in relation to root diameter




In the graph above, add a short horizontal line showing your visual estimate of the mean in the grazed and ungrazed plants. Comparing the means, we see that grazed plants (68 mg) have a higher mean than ungrazed (51 mg).
This unexpected result is due to the larger size of the grazed plants. It becomes clear that plants were not randomly assigned to one of the two treatment groups, a standard practice in experimental design.

In the graph showing root diameter, sketch a regression line, ignoring the groups. Then sketch a regression line for each of the two groups. We see that the lines look parallel for the two groups, but are not parallel to the line drawn through all the data. Comparing the slopes we find that the overall slope is $\hat{\beta}_{root} = 14.0$ mg/mm.
The ANCOVA estimate of the slope (14 mg/mm) from both groups differs markedly from the slope estimate for each group: 24 mg/mm for ungrazed, 23.3 mg/mm for grazed. This arises from the lateral offset: 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 averaged across both groups.

Graphical model.

In the context of statistical control the covariate (Root) and the *Root* x *Gr* term are not of interest. However, if grazing alters the relation of fruit biomass to plant size, we might expect the regression slopes to differ. The text example makes no mention of hypotheses concerning the interaction term.

Formal model

Write the formal model (GLM).

$$M_{Fruit} = \beta_o + \beta_{Root} * Root \; + \; \beta_{Gr} * Gr \; + \; \beta_{Root*Gr} * Root * Gr \; + \; \varepsilon$$

```
Above each term sketch a graph
showing data dots with means or lines
```

**2. Execute analysis.**

Place data in model format:

Column labelled $M_{fruit}$ the response variable fruit biomass (mg dry wt)
Column labelled Grazed with explanatory variable Gr: grazed=Y, ungrazed=N
Column labelled Root with explanatory variable Root = diameter

Code the model statement in a statistical package according to the GLM

$$M_{Fruit} \; = \beta_o \; + \; \beta_{Root} \cdot Root \; + \; \beta_{Gr} \cdot Gr \; + \; \beta_{Root \cdot Gr} \cdot \; Root \cdot \; Gr \; + \; \varepsilon$$

```
MTB > glm   'Mfruit' = 'root'   'Gr'     'root'*'Gr';
SUBC> covariate  'root'.
```

Fits and residuals are calculated in any of several ways:

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

The mean for grazed and ungrazed is expressed as a deviation from $\hat{\beta}_o = 59.4$ mg

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

The slope parameter for grazed and ungrazed together is $\hat{\beta}_{root} = 14.0$ mg/mm

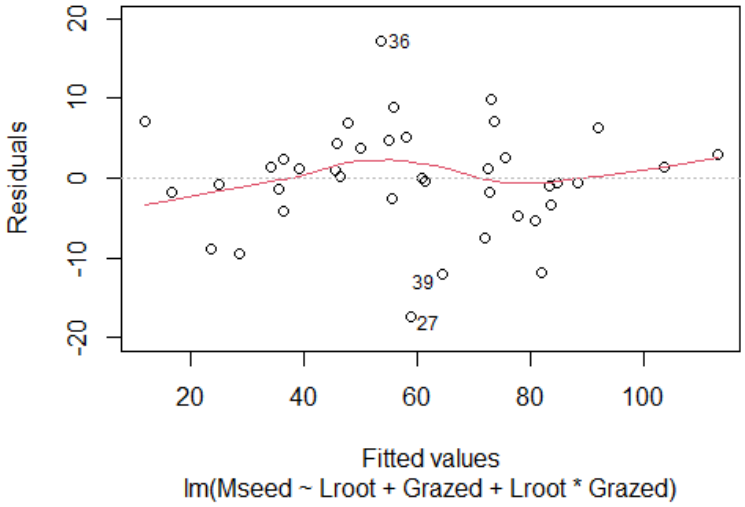
```
MTB > regress 'fruit' 1 'root'.

 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
```

The regression equations for each group differ substantially from the overall regression.

$M_{Gr=No} = -94.367 + 23.996\ Root$
$M_{Gr=Yes} = -125.28 + 23.254\ Root$

**3. Evaluate the model**

Plot residuals versus fits.




Straight line assumption questionable. Residual plot shows same pattern as response variable

Error model.

If n small, evaluate assumptions for p-values from chisquare (t, F) distributions.

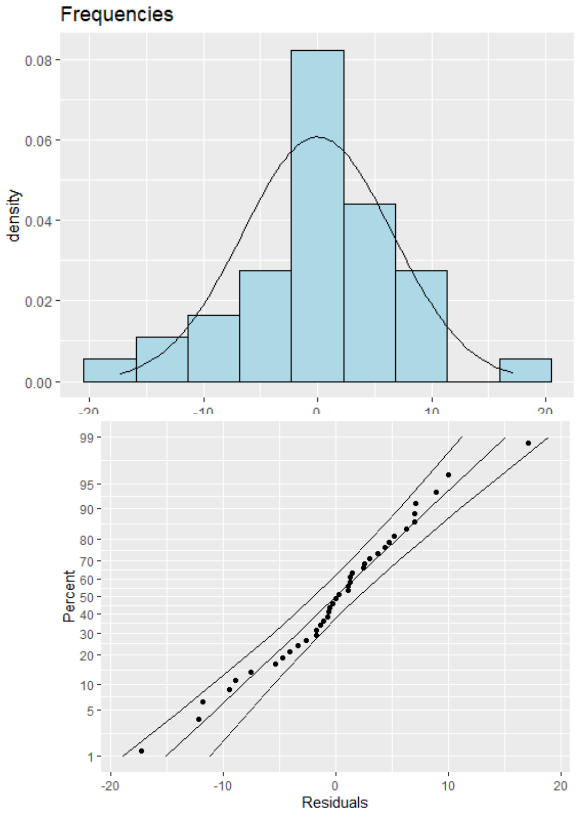
n = 40, so even substantial deviations will have little distorting effect on calculation of parameter estimates and p-values.

a. Homogeneous? No

Residuals show a spindle shape: high dispersion at the center, tapering to less dispersion at the extremes. Upon closer inspection we see that this is due to two diagonally oriented clusters of data points, one above the other.

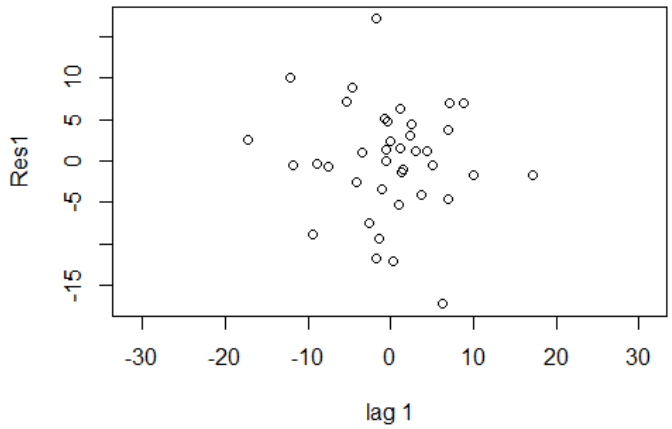
b. Normal?

The residuals look normal plotted as a histogram and in QQ plot




c. Independent?

Each residual plotted against its neighbor.




No evidence of non-independence.

d. Sum(res) = 0? Yes

The residuals are clearly not homogeneous, but this does not appear to produce values that can have undue influence on accurate estimates of the variable of interest fruit biomass. Estimates of Type I error (P-value) may be robust. Parameter estimates may not be accurate.

**4. Partition df and SS according to the model.**

| $M$ | = | $\beta_o$ | + | $\beta_{root}Root$ | + | $\beta_{Gr}Gr$ | + | $\beta_{Root\ x\ Gr}Root \cdot Gr$ | + | res |
|---|---|---|---|---|---|---|---|---|---|---|
| 40–1 | = | | | 1 | + | 1 | + | 1 | + | 39 |
| 23752.2 | = | | | 16800.4 | | 5266.7 | | 4.6 | + | 1680.5 |

**Calculate likelihood ratio for overall (omnibus) model.**

$1- R^2 = (1680/23752.2) = 0.07075$ $\quad LR = (1-R^2)^{-40/2} = 2.7 \times 10^{22}$

**5. State population and whether the data are a representative sample.**

The principles of sound experimental design are randomization, replication, and local control. In this case we have acceptable replication—20 grazed and 20 ungrazed plants. Randomization and local control were not achieved. Grazed plants were larger at the outset. This introduces an uncontrolled source of variance that is confounded with the presence/absence of grazing. The reason for larger plants in the group was not reported in the text example.

The sample of 40 plants will be taken as representative of a population based on multiple repeats of the same experimental protocol, keeping in mind the limitations of that protocol.

**5. Justify mode of inference.**

For ANCOVA with statistical control by the covariate, we normally assume that the slopes are homogeneous and that the interaction term is negligible. We will use likelihood ratios to evaluate this assumption. We will not use hypothesis testing to judge whether the interactive effect is important (cf. Chapter 11). Instead, we will focus on measures of evidence and on an accurate estimate of the difference in fruit biomass, accompanied by a measure of uncertainty.

**6. State hypothesis pairs and likelihood ratios.**

First, a check on whether slopes are parallel.

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

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

Next the factor of interest, Grazing or not.

$\beta_{Gr=0} > \beta_{Gr=1}$ (Grazing reduces fruit biomass)

$\beta_{Gr=0} \leq \beta_{root*Gr=1}$ (Grazing does not reduce fruit biomass)

**7. ANOVA**

```
MTB > glm 'seed' = 'root' 'grazing' 'root'*'grazing';
SUBC> covariate 'root'.

Analysis of Variance for seed

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
```

Root is listed first, which controls for this covariate in the sequential analysis.

The default ANOVA table in this package is the adjusted SS.

## 7. ANOVA

The default in the R package is the sequential SS.

```
SeedMod<-lm(Mass~RootDiameter+Grazed+Root*Grazed,data=CrwTb9_1)
anova(SeedMod)
                  SS          MS          Fratio
Root        1     16800.4     16800.4     359.912     < 2.2e-16
Grazed      1     5266.7      5266.7      112.8286    1.21E-12
Root:Grazed 1     4.6         4.6         0.0994      0.7544
Residuals   36    1680.5      46.7
```

First, the interaction term.

LR =( (4.6/1680.5)+1)$^{40/2}$ = 1.1 from sums of squares

LR = ((1/36)(0.10)+1)$^{40/2}$ = 1.1 from F and degrees of freedom

There is no evidence of an interactive effect. The slopes are parallel.

Next, the grazing effect.

LR =( (5266.7/1680.5)+1)$^{40/2}$ = 2.1 x $10^{-12}$ from sums of squares

There is strong evidence for a grazing effect in this analysis.

Looking at the adjusted SS table we see there is inadequate evidence.

LR =( (3.37/1680.5)+1)$^{40/2}$ = 6 times more likely than not, from sums of squares

We are in the uncomfortable position of results that depend on two different ways of controlling for size effects.

## 8. When assumptions are not met, decide whether to re-compute likelihood ratio.

Likelihood ratios and the F-statistics calculated from them depend on the error model. They also depend on the assumptions supporting either a sequential or adjusted analysis of variance, which in this case give different results.

We already have seen that the residuals plotted against the regression variable show a pattern. As we saw when we sketched the graphical model, the slope of the straight line is biased downward by the predominance of ungrazed plants at small root sizes, and the predominance of grazed plants at large root sizes. Rather than revising the statistical model, we can check our results against a completely different tactic, that of taking the data over the limited range of root sizes where both grazed and ungrazed plants occur. This occurs in the middle of the graph, at root sizes from 6.225 to 7.69 mm. Instead of revising the statistical model, we control for root size by executing the same ANCOVA model with balanced rather than unbalanced data.

## 9. Summary of results

```
             Df Sum Sq Mean Sq  F value    Pr(>F)
Root          1  850.7   850.7  29.7567 0.0001465 ***
Grazed        1 3603.6  3603.6 126.0551 1.011e-07 ***
Root:Grazed   1  159.9   159.9   5.5919 0.0357427 *
Residuals    12  343.1    28.6
```

LR = $((1/12)(5.59)+1)^{40/2}$ = 2097  from F and degrees of freedom

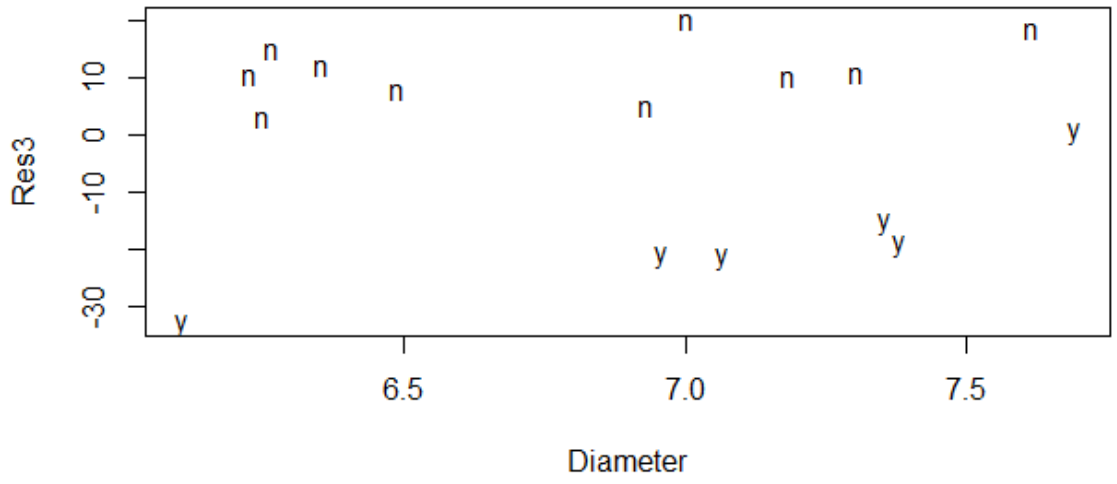
With balanced data we have evidence for a difference in regression lines for grazed and ungrazed plants.

For the balanced data set, the graph of the residuals from the regression shows a stronger positive relation in grazed than in ungrazed plants. This can be attributed to large plants having greater capacity to withstand grazing than smaller plants.

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -57.177     23.778  -2.405   0.033234
Root             18.563      3.508   5.291   0.000191
Grazed yes     -124.799     39.082  -3.193   0.007729
Root:Grazed yes  13.237      5.598   2.365   0.035743
```

The slope estimate for ungrazed is 18.5 mg/mm
The slope estimate for grazed is 18.5 + 13.2 = 31.7 mg/mm.

The difference in fruit size is evident in the graph of the residuals from regression for the balanced data.




The residuals for balanced data show far less pattern than those for unbalanced data. The residuals show clear separation of grazed and ungrazed plants, with far low fruit biomass for grazed than ungrazed plants.

## 10. Conclusion

When slopes are homogeneous, a convenient way to compare effects of grazed and ungrazed plants is to use the points at which x = zero (the y-intercepts). When the slopes are heterogeneous, as with the balanced data, divergence in slopes amplifies the difference in y-intercepts.

When slopes are heterogeneious, an alternative to comparing the y-intercepts is to compare the residuals from regression for grazed and ungrazed plants. For the balanced data, the residuals from regression on root diameter were negative for grazed, positive for ungrazed.

Ungrazed average = 11.04 mg  Grazed average = -40.5 mg
Ungrazed - Grazed = -29.44 mg

The reduction due to grazing is substantial, compared to the average biomass of all ungrazed plots = 69.4 mg

-29/69  # = -42 %  Balanced data (Residuals from regression)
-36/69  # = -52 %  Unbalanced data (Comparison of y-intercepts)

The parameter estimates differ by a ratio of 36/29 = 1.24. The reduction in fruit biomass estimated from unbalanced data is high, compared to that estimated from balanced data.

**Estimate of grazing effect based on comparison of Y-intercepts, unbalanced data.**

The slopes are parallel. $\beta_{root*Gr=0} = \beta_{root*Gr=1}$

Because of the offset, we report the rate in each group, not the overall rate.

**9. 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 biomass appears to be less for ungrazed than for grazed.

| | |
|---|---|
| Ungrazed | 50.88 mg |
| Grazed | –67.94 mg |
| Difference | –17.06 mg |

This is because initially 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. A convenient point at which to do this is the point at which x = zero (the y-intercepts).

| | | | | | | |
|---|---|---|---|---|---|---|
| $\hat{\alpha}$ | = | $\beta_o$ | – | $\beta_{root}$ | * | mean(X) |
| $\hat{\alpha}_{Gr=no}$ | = | Mean($M_{Gr=no}$) | – | $\beta_{root}$ | * | Mean($root_{GR=no}$) |
| | = | 50.88 | – | 23.6 | * | 6.053) = –91.729 mg |
| $\hat{\alpha}_{Gr=yes}$ | = | Mean($M_{Gr=Yes}$) | – | $\beta_{root}$ | * | Mean($root_{GR=Yes}$) |
| | = | 67.94 | – | 23.6 | * | 8.309) = –127.82 mg |

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

| | |
|---|---|
| Grazed | –127.820) mg |
| Ungrazed | – ( –91.729 mg |
| Difference | – 36.091 mg |

By this accounting, the fruit biomass for grazed plants was less by 36 mg. This accounting is based on a regression line extrapolated beyond the data, from unbalanced data that distorts the regression. The regression slope for all data differs from the regression slope in each group, which are nearly the same.

**Estimate of grazing effect based on residuals from regression in each group.**
When grazing effects are controlled by sequential analysis, the covariate appears first, with grazing second. This is equivalent to taking the residuals from the covariate, then comparing the means of the residuals from the two groups. The analysis above used sequential sums of square.
The use of sequential SS suggests a different approach, which is to estimate the regression in each group, calculate the residuals from that regression in each group, and then compare the residuals, which are now controlled for initial size.
The tactic fails for all data (ANCOVA on unbalanced design)

```
          Df Sum Sq Mean Sq F value Pr(>F)
Grazed     1    0.0   0.000       0      1
Residuals 38 1680.5  44.222
```

This tactic eliminates all of the differences between the two groups. In retrospect the regression equation in each group estimates 4 parameters that together eliminate the intercept differences as well as the slope effects.