Model Based Statistics in Biology.
Part IV. The General Linear Model. Multiple Explanatory Variables.
Chapter 14.1 ANCOVA - Comparison of Slopes

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)

Brussard.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).

Today: Analysis of Covariance (ANCOVA)
ANCOVA is a special case of the GLM in which there are both nominal scale (categorical) and ratio scale explanatory variables.

Wrap-up.
ANCOVA is applied to data situations that require both ratio and nominal scale explanatory variables. One important use is to compare two or more regression lines.

The analysis demonstrates the use of categorical along with ratio scale variables within the framework of the general linear model. The categorical variable is species (*D. persimilis* or *D. pseudoobscura*). The ratio scale variable is change in heterozygosity with altitude. This example also shows the logic of interaction terms, which are examined before main effects.

Introduction.
The next application of the general linear model compares two functions expressed as straight lines. The data are from Dobzhansky (1948) as reported in Brussard (1984). Theodosius Dobzhansky collected data on inversion heterozygosity (assuming Hardy Weinberg equilibrium) of 3rd chromosome inversions from two species of fruit fly, from Yosemite Park, California. Inversion heterozygosity is a measure of genetic variability, which is the raw material that natural selection acts on to produce descent with modification (Darwinian evolution). Thus, the origin and maintenance of genetic variability is one of the central questions in population biology. What factors generate or erode genetic variability? Harsher environments at higher altitudes are expected to select for narrower range of phenotypes, hence reduce genetic variability. Does genetic variability decrease at higher altitude, due to stronger selection in extreme environments? Does the heterozygosity gradient differ in two co-occuring species of fruitfly *Drosophila persimila* and *Drosophila pseudoobscura*?

1. Construct model

Verbal model.

Inversion heterozygosity decreases with altitude, depending on species.

Graphical model. [species labels need to be switched]

Figure 14.1a

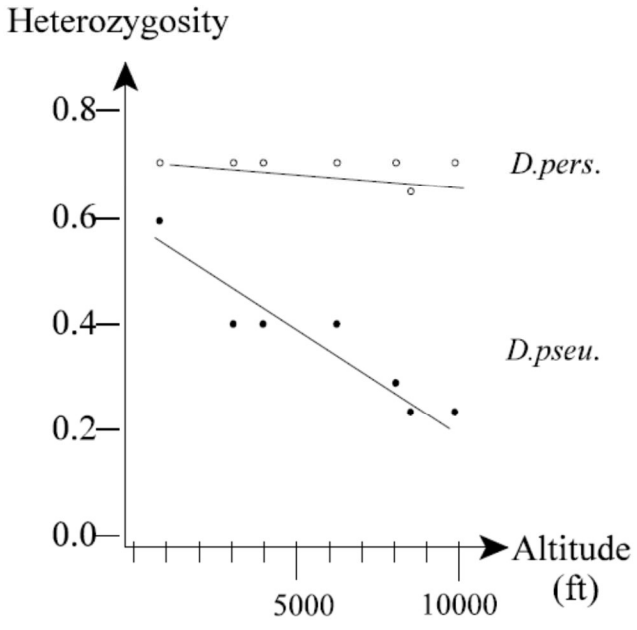
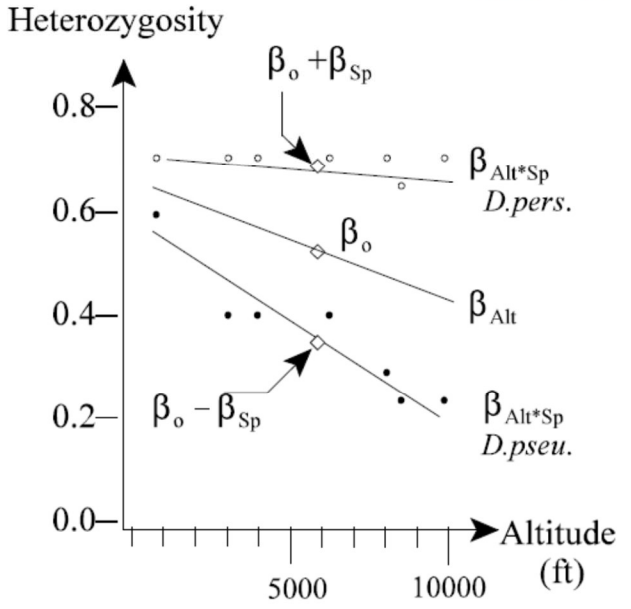


Figure 14.1b



1. Construct formal model

Variable Name	Symbol	Response or Explanatory	Scale	Fixed or Random
Heterozygosity	H	Response	Ratio	
Altitude	Alt	Explanatory	Ratio	
Species	Sp	Explanatory	Nominal	Fixed

The response variable is inversion heterozygosity in two species of fruit fly, *Drosophila persimilis* and *D. pseudoobscura* H_{per} = % H_{ps} = %

The ratio scale explanatory variable is altitude Alt = km

The nominal scale explanatory variable is species
Sp = *D. persimilis* or *D. pseudoobscura*

Formal model

GLM: $H = \beta_o + \beta_{Alt} \cdot Alt + \beta_{Sp} \cdot Sp + \beta_{Sp\ x\ Alt} \cdot Alt \cdot Sp + \varepsilon$

This looks just like a two way ANOVA.
This model has much in common with a two-way ANOVA.

Begin modifying Fig L18a to L18b

Add β_o
Add β_{Sp}

The β_o parameter stands for the overall mean, just like an ANOVA.
The β_{Sp} parameter stands for the contrast between means, just like an ANOVA.

The β_{Alt} parameter stands the heterozygosity gradient (%/km).

Add regression line though β_o

The $\beta_{Sp\ x\ Alt}$ parameter stands for the contrast in heterozygosity gradients, β_{per} versus β_{ps}

Draw angles that compare these two lines
Add question marks to query whether these are parallel.

The parameter $\beta_{Alt\ x\ Sp}$ represents the degree to which the slopes in each class differ

Sequential addition (L18b from model) went well.
In 1998 separate regression lines(L18a) erased before starting.
In 2000 L18b built up from L18a.
This went well and quickly.

2. Execute analysis.

Place data in model format:

Column labelled *H*, with response variable heterozygosity
Column labelled *Alt*, with explanatory variable Altitude
Column labelled *Sp*, with explanatory variable species labels Dper or Dps

Code the model statement in statistical package according to the GLM

$$H = \beta_o + \beta_{Sp} \cdot Sp + \beta_{Alt} \cdot Alt + \beta_{Alt \cdot Sp} \cdot Alt \cdot Sp + \varepsilon$$

```
MTB > glm H = 'Sp' 'Alt'  'Sp'*'Alt';
SUBC> covariate  'Alt'.
```

The ratio scale variable is labelled as a covariate in this package and SPSS.
Other packages (*e.g.* SAS) assume variables are on a ratio scale,
hence categorical variables must be defined in the model statement.

```
Proc GLM;
  Model H = Sp Alt Sp*Alt; Class Sp;
```

In R and SPlus the distinction between categorical and regression variable is determined by the definition of the explanatory variables in the data object (data=Brussard)

```
Hmodel < - lm(H ~ Sp + Alt + Sp*Alt,
     Data = Brussard)
```

In this example `Alt` exists as a numerical variable in the data object, while `Sp` exists as a factor (categorical variable) in the data object because it consists of letters, not numbers.
Fits and residuals are obtained depending on the package.

model statement output of fitted values and residuals (SAS)
parameters reported by GLM routine (SPSS, Minitab, R)
or direct calculation from model parameters

Statistical packages differ in how they report parameter estimates.
For our purposes, we can think of the parameter estimates as follows:

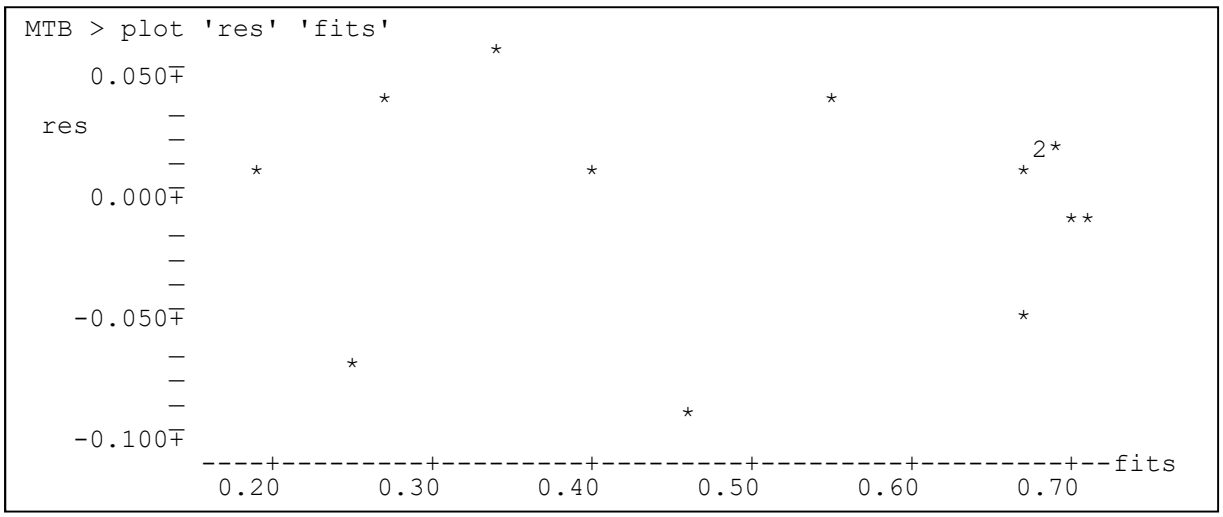
$\hat{\beta}_o$ is the mean heterozygosity for one of the species.
$\hat{\beta}_{Sp}$ as the contrast (difference) in mean heterozygosities
$\hat{\beta}_o + \widehat{\beta}_{Sp}$ is the mean heterozygosity of the other species
$\hat{\beta}_{Alt}$ is the heterozygosity gradient for the reference species.
$\hat{\beta}_{Akt\ x\ Sp}$ as the contrast in heterozygosity gradients
$\hat{\beta}_{Alt} + \hat{\beta}_{Akt\ x\ Sp}$ is the heterozygosity of the other species

The slope for *D. persimilis* is the more negative of the two.

$H_{per} = 0.580 - 0.127\ Alt$
$H_{ps} = 0.712 - 0.0145\ Alt$

The residuals for the GLM (both species) are computed from the fitted values.

3. Evaluate the model Plot residuals versus fitted values.



A. Straight line assumption is acceptable -- no bowls or arches in plot.
B. Sample size is small. (n = 14) so evaluate assumptions for p-values calculated from chisquare, t, or F distributions

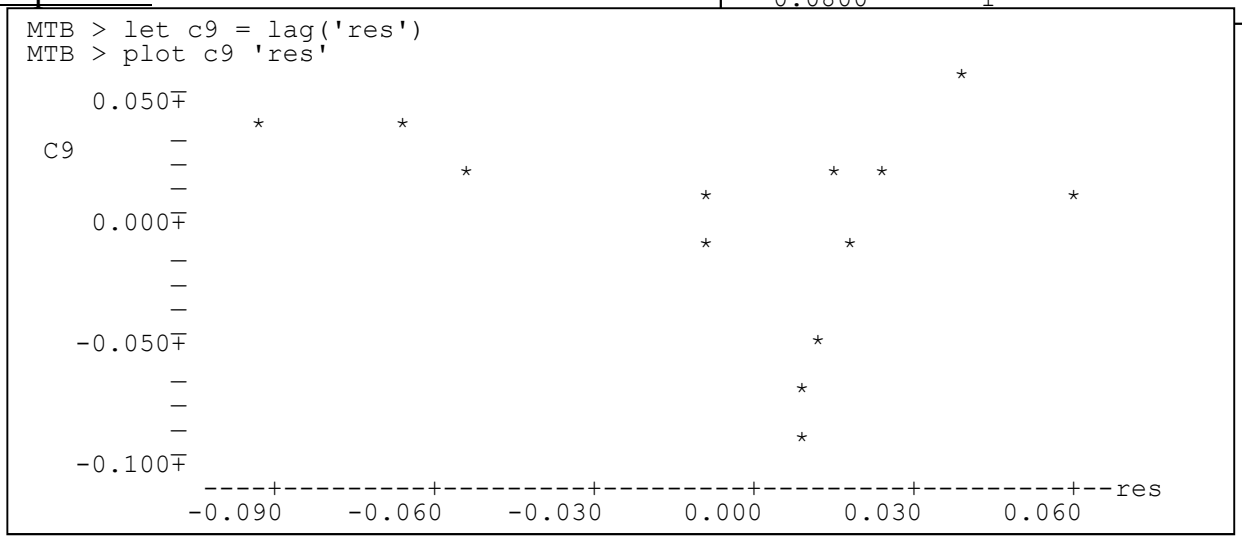
B1. Homogeneous error assumption (used in estimating parameters) is acceptable. Residuals do not change in any systematic way with fitted values (no cones).

```
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  *
```

B2 Normal?

The residuals look normal when plotted as a histogram.

B3. Independent?



Each residual is plotted against its neighbor, with data ordered by altitude.

There is some indication of trends -- residuals are not completely independent.

The assumptions are considered to be met because there are no large violations.

4. Partition df and SS according to model.

Y	=	β_o	+	$\beta_{Alt}Alt$	+	$\beta_{Sp}Sp$	+	$\beta_{Sp\ x\ Alt}Alt \cdot Sp$	+	res
14–1	=			1	+	1	+	1		+10
0.51377	=			0.05991		0.39111		0.03798		+ 0.02477

Calculate likelihood ratio for reduced model (all terms in the model)

$$LR = \frac{L(\beta_{Sp}, \beta_{Alt}, \beta_{SP\ x\ Alt}, \sigma\ |H)}{L(\sigma\ |H)}$$

$$LR = \left(\frac{0.02477}{0.51377}\right)^{(-14/2)} \quad LR = 165 > 100$$

The model has strong evidential support despite the small sample size.

5. State the populations and whether the sample is representative.

The chance set-up (Hacking 1965 p8, 114) consisted of repeated measurements (trials) with a procedure that generates a unique result (% heterozygosity) from a population of possible results (any value of heterozygosity from 0 to 1). On logical grounds this is the best estimate of the population of flies catchable by the investigators. Does the sample represent inversion heterozygosity relative to altitude in all fruit flies at the study sites? Or perhaps fruit flies at all sites in Yosemite Park ? Perhaps all fruit flies in the world? The results are taken as representative of genetic inheritance in fruit flies.

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

This is an observational study and so there are many potential sources of uncontrolled variability. An evidentialist approach is appropriate. Given the evidential support for the omnibus model it is of interest to calculate the evidential support for each term in the model.

10. Report and interpret parameters of biological interest.

In this model there are three terms, all fixed.
Interpretation begins with the interaction term.

Interaction term. This is a fixed effect because both of its components are fixed.
Is the model with the interaction term more likely than the model without the term?

$$LR = \frac{L(\ \beta_{Sp\ x\ Al}\ , \sigma\ |H)}{L(\sigma\ |H)}$$

In other words are the heterozygosity gradients the same ?

If there is little or no evidence of heterogeneous slopes we can interpret the component terms β_{Alt} and β_{Sp} independently of each other.

10. Report and interpret parameters of biological interest.

Species term. Model I. Fixed effects.

Does the mean for *D. persimilis* differ from that for *D. pseudoobscura* ?

$$LR = \frac{L(\beta_{Sp}, \sigma \mid H)}{L(\sigma \mid H)}$$

Altitude term. Model I. Fixed effects.

How good is the evidence for overall gradient?

$$LR = \frac{L(\beta_{Alt}, \sigma \mid H)}{L(\sigma \mid H)}$$

Are there more specific hypotheses about parameters ?

Yes, the study was undertaken to find whether heterozygosity decreases in increasingly harsh environments at higher altitudes. If there is little support for an interactive effect, we drop the species term and report the overall gradient β_{Alt}.

Sequential vs adjusted SS. When a covariate (regression variable) is included in the model, some of the explanatory variables may well be correlated. When they are, the partitioning of the SS will depend on the order in which terms are listed. The result is a sequential SS. In this example the focus is the altitudinal gradient and so a sequential SS could be used. The result is the SS for altitude, adjusted for S

Source	df	Seq SS	Adj SS
Sp	1	0.39111	0.01267
Alt	1	0.05991	0.05991
Alt*Sp	1	0.03798	0.03498
Res	10	0.02477	0.02477
Total	13	0.51377	does not add up

If the model had been written in a different order, then the partitioning would come out differently. If *Sp* had been the last term in the model, SS_{Sp} it would be only 0.01267, rather than 0.39111. With sequential partitioning it is "first come first serve." That is, a term will generally be allocated a larger SS if it occurs early in the queue, rather than later.

10. Report and interpret parameters of biological interest.

The sequential SS and is of interest in some cases. In the fly heterozygosity example, we may have been interested in whether the two species differ in heterozygosity after controlling for (removing the effects of) altitude. The altitude term is placed first in the list, so that we can examine whether there are species differences after altitude has been removed.

If we have no reason to adjust for order effects we use the SS allocated to each term when it occurs last in the model. This adjusted SS allows us to examine effects controlled for other terms. It is a conservative procedure. That is, it will allocate a relatively small SS to each term, generally smaller than if the term were listed early in the model.

One consequence of this tactic is that the sums of the squares no longer sum to the total SS.

The sequential partitioning of the sum of squares is called Type I SS. The Adjusted SS is called Type III SS. There are other ways to partition the SS, but Type I and Type III are the most commonly used. The residuals will be the same, regardless of how we partition the SS. So the choice of sequential or adjusted SS has no effect on diagnostics --straight line acceptable ? errors meet assumptions? If we use the Adj SS to compute the likelihood ratios, the SS ratio for each term is taken relative to the residual.

Source	Df	SS	1+SSratio	LR
Alt	1	0.05991	3.419	5457
Sp	1	0.1267	6.115	319745
Alt*Sp	1	0.03798	2.533	670
Res	10	0.02477		
Total	13			

$2.533 = 1+0.03798/0.02477$

$LR = 2.533^{14/2} = 670$

As with two-way ANOVA, we begin interpretation with the interaction term. The evidence for an interactive effect is substantial. An interactive effect is 670 times more likely than no interactive effect.

The heterozygosity gradient depends on species (LR = 670) so we report the gradient in each species.

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5801	0.0529	10.9711	0.0001
Alt	-0.1273	0.0262	-4.8619	0.0046

This output reports a t-statistic, which we can use to calculate the likelihood ratio.

10. Report and interpret parameters of biological interest.

Calculating the LR from a *t*-statistic.

The calculation begins by squaring the reported value of *t* for the altitude gradient.

t	t^2	$t^2/10$	$1+t^2/10$	LR
4.8619	23.64	2.36	3.36	4873

SS ratio = t^2/df, df = 10.

Heterozygosity decreases with altitude in *D. persimilis* ($LR = 4873$) so we report the regression equation for the heterozygosity gradient.

$H = 0.58 - 0.127\ Alt$

Heterozygosity does not change with altitude in *D. pseudoobscura* ($LR = 2$)

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7117	0.0243	29.2581	0.0000
Alt	-0.0144	0.0120	-1.1995	0.2841

t	t^2	$t^2/10$	$1+t^2/10$	LR
1.1195	1.25	0.13	1.13	2

There is no evidence of change with altitude ($LR = 2$), so instead of an equation for the gradient, the heterozygosity in this species is adequately summarized by the mean

$\text{mean}(H_{pseu}) = 0.686$