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)

on chalk board

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 test whether the slopes of two or more regression lines differ.

The analysis demonstrates the use of categorical variables within the framework of the general linear model. The categorical variable is species (*D. persimilis* or *D. pseudoobscura*). It also demonstrates 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). Data by Theodosius Dobzhansky 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. One of the central questions in population biology is the origin and maintenance of genetic variability in natural populations.

Many students will not have had population genetics, so explain that interest here lines in factors that generate or erode genetic variability. Harsher environments at higher altitudes are expected to select for narrower range of phenotypes, hence reduce genetic variability with increasing altitude. Does this effect show up ? Do two species of fly differ in this effect ? In this case: Does genetic variability decrease at higher altitude, due to stronger selection in extreme environments ? Does the decrease in heterozygosity with altitude observed in the fruitfly *Drosophila pseudoobscura* occur also in *Drosophila persimilis*?

1. Construct model

Verbal model.

Inversion heterozygosity changes with altitude, depending on species.

		Heterozygosity (%)	
Elev (ft)	Elev (km)	D.persimilis	D. pseudoobscura
850	0.26	0.59	0.70
3000	0.91	0.37	0.69
4600	1.40	0.41	0.71
6200	1.89	0.40	0.70
8000	2.44	0.31	0.70
8600	2.62	0.18	0.62
10000	3.05	0.20	0.68

Graphical model.

Figure 14.1a

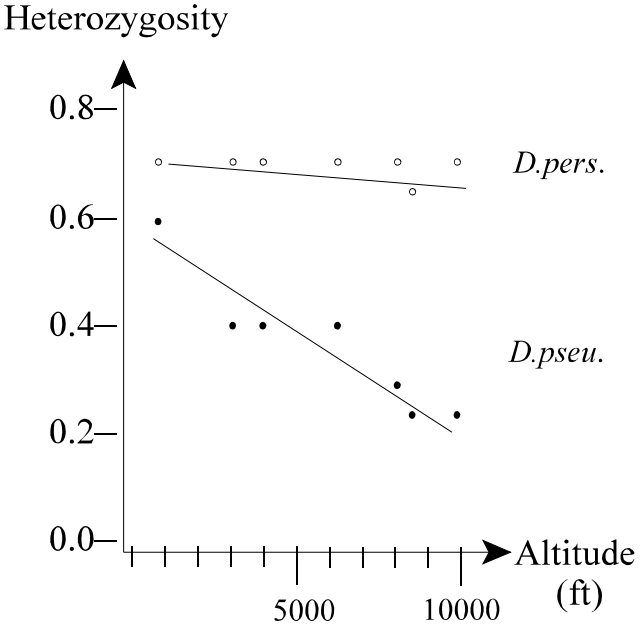
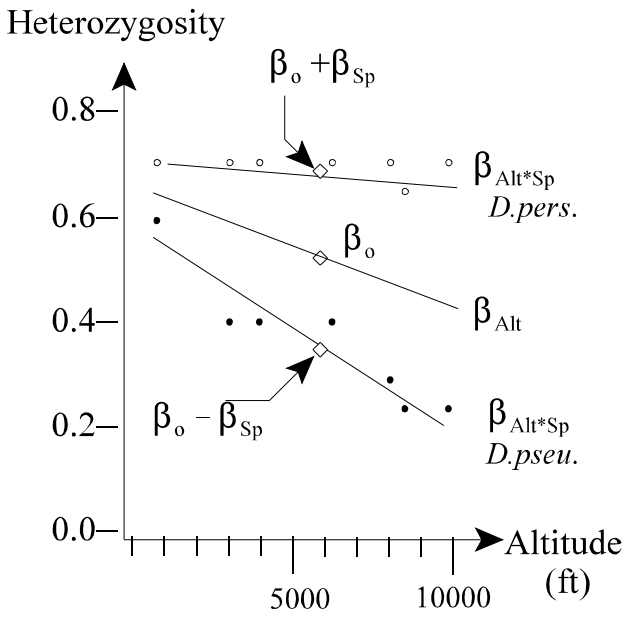


Figure 14.1b



1. Construct model

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

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

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

Begin modifying Fig L18a to L18b

The β_o parameter stands for the overall mean, just like an ANOVA.
The β_{Sp} parameter stands for a set of means, just like an ANOVA.

Add β_o
Add β_{Sp}

The β_{Alt} parameter stands for a single slope, just like a regression.
β_{Alt} is the heterozygosity gradient, regardless of species.

Add regression line though β_o

The $\beta_{Sp\ x\ Alt}$ parameter stands for the degree to which the gradients in each species β_{per} and β_{pseu} deviate from the average gradient β_{Alt}.

Draw angles that compare these two lines to the overall line β_{Alt}
Add question marks to query whether these are parallel to the overall line β_o

Which is to say,
we can draw a line describing change in heterozygosity with altitude for each species then examine whether or not these lines are parallel to the overall slope β_{Alt}. The parameter $\beta_{Alt\ x\ Sp}$ represents the degree to which the slopes in each class differ from the overall slope.

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.

In this analysis there are two explanatory variables. One is categorical (either of two species) while the other is on a ratio type of scale (altitude). Both terms appear in the model as the variable times a parameter or set of parameters. The parameter β_z is the altitudinal gradient in heterozygosity for both species together. The parameter β_z has two values, one for the mean heterozygosity of D. persimilis, the other for the mean of D. pseudoobscura. The β notation for the general linear model expresses the mean for each species as a positive or negative deviation from β_o the overall mean heterozygosity. The mean values are the sum of the parameters.

$$\beta_o + \beta_{SP} = \text{mean}(H_{pers})$$
$$\beta_o - \beta_{SP} = \text{mean}(H_{pseu})$$

1. Construct model

The model includes an interaction term as well as a term for each variable. This term expresses the interactive effects of its component variables on the response variable. In this analysis, where the term has a categorical and a ratio scale (regression) variable, interaction measures the heterogeneity of slopes (Figure 13.5a). The more divergent the slopes in each category, the larger the interaction term. The term is written as the product of its two component variables, with a parameter $\beta_{Alt \cdot SP}$ that represents a slope for each species. These slopes are expressed as deviations from the overall slope β_{Alt}.

The slope for each group is $\beta_{Alt} + \beta_{Alt \cdot SP}$ and $\beta_{Alt} - \beta_{Alt \cdot SP}$.

$$\beta_{Alt} + \beta_{Alt*SP} = \beta_{pers}$$
$$\beta_{Alt} - \beta_{Alt*SP} = \beta_{pseu}$$

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 Sp = Dpers or Dpseu

Code the model statement in statistical package according to the GLM

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

```
MTB > glm H = 'Alt' 'Sp'  'Alt'*'Sp';
SUBC> covariate  'Alt';
SUBC> fits c4;
SUBC> residuals c5.
MTB > name c8 'fits' c9 'res'
MTB > plot 'res' 'fits'
```

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

```
Proc GLM;
 Model H = z sp z*sp; Class sp;
 Output OUTDAT;
 res = res; fits = fit;
Proc Plot; Data OUTDAT;
 Plot res*fit;
```

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 heterozygosity for both species is $\hat{\beta}_o = 0.5186$ %

The mean for each species is expressed as a deviation from $\hat{\beta}_o$

$$\begin{cases} \hat{\beta}_o + \hat{\beta}_{SP} = \text{mean}(H_{pers}) = 0.5186 - 0.1671 = 0.351 \\ \hat{\beta}_o + \hat{\beta}_{SP} = \text{mean}(H_{pseu}) = 0.5186 + 0.1671 = 0.686 \end{cases}$$

2. Execute analysis.

The slope parameter for both species together is $\hat{\beta}_{Alt} = -0.07087\ \%\ \text{km}^{-1}$

The deviations from this slope are $\hat{\beta}_{Alt*Sp} = \begin{matrix} -0.05642\ \%\ \text{km}^{-1} \\ +0.05642\ \%\ \text{km}^{-1} \end{matrix}$

The slope for *D. persimilis* is more negative than the average slope by $-0.05642\ \%\ \text{km}^{-1}$

$$\hat{\beta}_{Alt} + \hat{\beta}_{Alt*SP} = \text{Slope}(H_{pers}) = -0.07087 - 0.05642 = -0.1273$$
$$\hat{\beta}_{Alt} + \hat{\beta}_{Alt*SP} = \text{Slope}(H_{pseu}) = -0.07087 + 0.05642 = -0.0145$$

These particular deviations happen to be symmetrical, but this will not always be the case. The deviations are symmetrical in this case because the overall slope $\hat{\beta}_{Alt}$ is half way between the slope for each species.

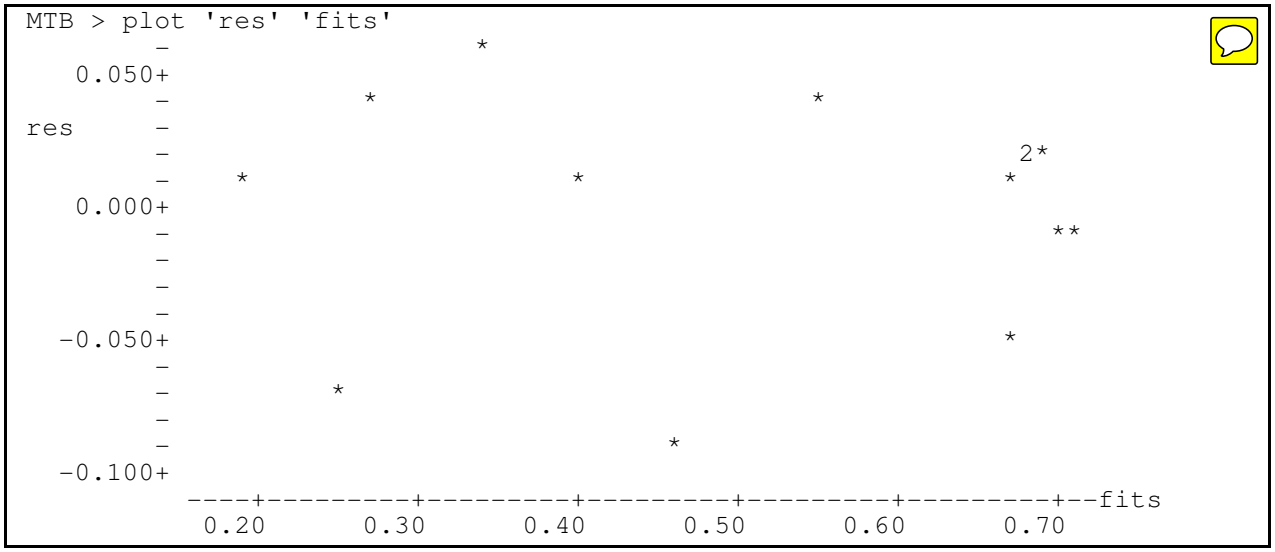
Compare to regression equation (one slope and one intercept) for each species:

$H_{pseu} = 0.580 - 0.127\ Alt$

$H_{pers} = 0.712 - 0.0145\ Alt$

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

3. Evaluate the model Plot residuals versus fitted values.



a. Straight line assumption acceptable. No bowls or arches in plot

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

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

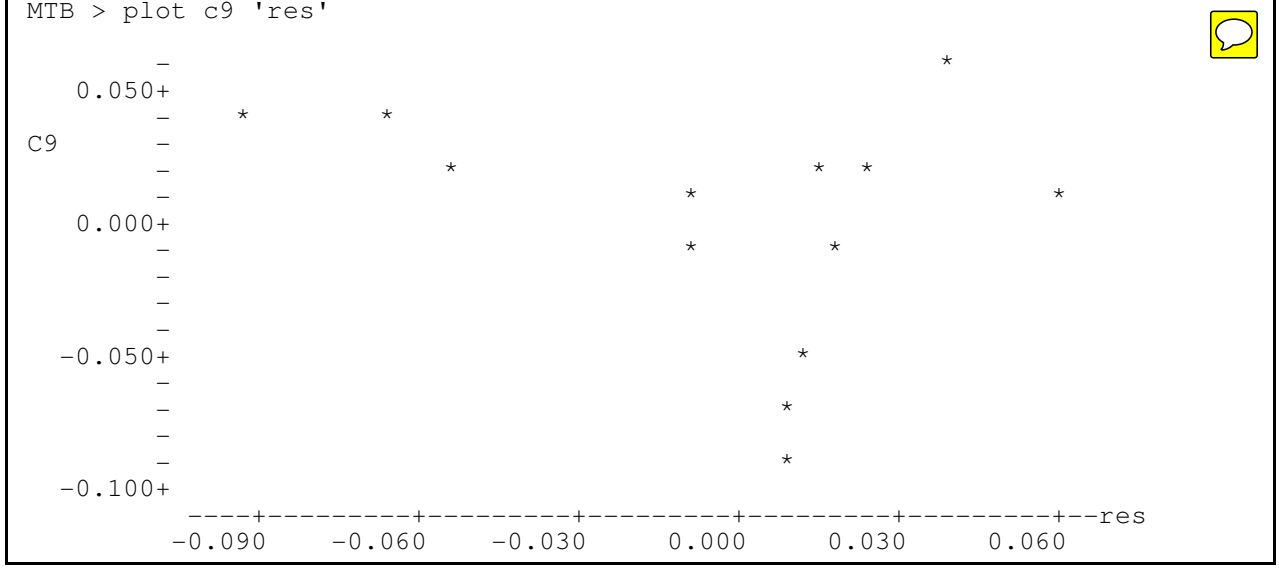
n = 14

Sum(res) = 0? Yes

3. Evaluate the model

Independent?

```
MTB > let c9 = lag('res')
MTB > plot c9 'res'
```



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

There is some indication that residuals are not completely independent (trends)

Normal?

The residuals look normal when plotted as a histogram.

Residuals are normal and homogeneous, although with some indication of non-independence.
The assumptions are considered to be met because there are no large violations.

```
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.

Does sample represent inversion heterozygosity in all fruit flies ? Not likely.
All fruit flies in Yosemite Park ? Again not likely.
Not enough is known to determine whether the flies were representative. As before, we will assume that the data come from a statistical population--all measurements that could have been obtained on this collection of flies, using the procedural statement.

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

Research question is whether heterozygosity gradient exists (yes/no question).
So hypothesis testing is appropriate. In particular, it is by no means evident from the data whether the apparent difference in gradient between the two species is more than chance.

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

Interaction term. Model I. Fixed effects (because both components are fixed)
Are the heterozygosity gradients the same ?
Is there variance due to the interaction term ?

$Var(\beta_{Sp\ x\ Alt}) > 0$ Same as $H_A: \beta_{per} \neq \beta_{pse}$
$Var(\beta_{Sp\ x\ Alt}) = 0$ Same as $H_o: \beta_{per} = \beta_{pse}$

The next two hypotheses, concerning β_{Alt} and β_{Sp} , cannot be examined if the slopes are heterogeneous. Thus, we need include this term in our model. Note that some statistical packages contain routines (ANCOVA commands) that leave out the interaction term.
For example the Minitab commands for executing ANCOVA are:

```
MTB > GLM '%H'  =  'Alt'  'Sp' 'Alt'*'Sp';
SUBC> covariate 'Alt'
SUBC> fits c4;
SUBC> res c5.
```

It is better to include the term and test it in a GLM command, rather than assuming it is not important.

Species term. Model I. Fixed effects.
Does the mean for D. *persimilis* differ from that for *D. pseudoobscura* ?
Is there variance due to the species term ?

H_A: $Var(\beta_{Sp}) > 0$ Same as $H_A: \beta_{per} \neq \beta_{pseu}$
H_o: $Var(\beta_{Sp}) = 0$ Same as $H_o: \beta_{per} = \beta_{pseu}$

Altitude term. Model I. Fixed effects. Is the slope less than zero ?
If so, there will be variance due to the altitude term.

H_A: $Var(\beta_{Alt}*Alt) > 0$ Same as $H_A: \beta_{Alt} \neq 0$
H_o: $Var(\beta_{Alt}*Alt) = 0$ Same as $H_o: \beta_{Alt} = 0$

Are there more specific hypotheses about parameters ?
Yes, the study was undertaken to find whether heterozygosity decreases in increasingly harsh environments at higher altitudes.

$H_A: \beta_{Alt} < 0$
$H_o: \beta_{Alt} \geq 0$

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

7. ANOVA - Calculate df, partition according to model.

GLM: $Y = \beta_o + \beta_{Alt} \cdot Alt + \beta_{Sp} \cdot Sp + \beta_{Sp\ x\ Alt} \cdot Alt \cdot Sp +$ res

Source: Total = Alt Sp Alt · Sp res

Compute total df: $n-1 = 14-1 = 13$

GLM:	Y	=	β_o	+	$\beta_{Alt} \cdot Alt$	+	$\beta_{Sp} \cdot Sp$	+	$\beta_{Sp\ x\ Alt} \cdot Alt \cdot Sp$	+	res
source	total	=			Alt	+	Sp	+	Alt ·Sp	+	res
df:	14−1	=			1	+	1	+	1	+	10

Add df to ANOVA table.

Source	df	SS	MS	F	---->	p
Alt	1					
Sp	1					
Alt*Sp	1					
Res	10					
Total	13					

Compute total SS

By hand: $SS_{total} = \Sigma Y^2 - n^{-1}(\Sigma Y)^2 = 0.395$

In Minitab: $SS_{total} = 13 \cdot Var(Hyz) = 13 \cdot 0.0395 = 0.5138$

```
MTB> let k1 = stdev('Hyz')*stdev('Hyz')
MTB> print k1
   k1    0.0395
```

Use Minitab to partition this SS_{total}.

$Hyz = \beta_o + \beta_{Sp} \cdot Sp + \beta_{Alt} \cdot Alt + \beta_{Alt * Sp} \cdot Alt \cdot Sp +$ res

```
MTB> GLM      c4  =       c6       c5       C6*c5

MTB> GLM    'Hyz'   =    'sp'     'Alt'     sp'*'Alt'
```

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

Source	df	Seq SS	MS	F	---->	p
Alt	1	0.05991				
Sp	1	0.39111				
Alt*Sp	1	0.03798				
Res	10	0.02477				
Total	13	0.51377				

This partitioning is labelled Seq SS (also called Type I SS) in computer output.

7. ANOVA - Calculate variance, partition according to model.

The partitioning has been carried out in the order that the terms were written out in the model: first Altitude, then Species, then the interaction term. 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} would be only 0.01267, rather than 0.39111. Because the partitioning is sequential it is "first come first serve." That is to say, a term will generally be allocated a larger SS if it occurs early in the queue, rather than late.

In some situations we want the sequential SS. If we know the order in which we want to see the SS partitioned, then we simply write the model so as to obtain a partitioning according to this order. 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 significant species differences after altitude has been removed.

In most analyses of biological data we have no reason to order the terms. We are more interested in whether any one term can be considered significant. But the level of significance will depend on how large the SS, which in turn will depend on whether the term occurs early or late in the sequential partitioning. To adjust for this we use the SS allocated to each term when in occurs last in the model. This is called Adj SS in the GLM print-out. This tactic is the best way to examine all the terms, free of the effects of the other terms. It is also 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.

Source	df	Adj SS	MS	F	----> p
Alt	1	0.05991			
Sp	1	0.01267			
Alt*Sp	1	0.03798			
Res	10	0.02477			
Total	13	(do not add up)			

One consequence of this tactic of computing the Adj SS is that the sum of the SS allocated to each term will usually be less than the total SS, as in this example.

In the fly heterozygosity example we will use the Adj SS to compute the adjusted mean squares and then from this, the F-ratios. GLM command in Minitab has conveniently computed Adj MS for us. If we had wanted to use the Seq SS for computing F-ratios, we would need to specify this to Minitab.

The sequential partitioning of the Sum of Squares is also called Type I SS. The Adjusted SS is also 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.
Hence no effect on step 7 (straight line acceptable ?) and step 10 (errors meet assumptions?).

7. ANOVA - Table SS, MS, F-ratios

Calculate the correct MS = SS/df
if not already done by computer

Add all MS to table

Add one F-ratio to table

Calculate $F_{Sp * Alt} = 15.33$

This F-ratio measures how much the two slopes diverge.
As in a two-way ANOVA, we start with the interaction term. If the slopes diverge there is little point in trying to interpret the average slope β_{Alt}.

Calculate p-value.

Type I error is calculated from a theoretical F-distribution with df = 1,10. The result of this calculation is shown in the ANOVA table produced by the GLM command.

$$F_{Sp*Alt} = 15.33 \quad p = 0.003$$

The chance of obtaining an $F_{Sp * Alt}$ this large from our population of all possible measurements is p = 0.003.

8. Decide whether to recompute p-value.

Assumptions met, skip step.

9. Declare decision about terms.

As in a two-way ANOVA, we start with the interaction term. If the slopes are heterogeneous there is little point in trying to interpret the average slope β_{Alt}.

$$0.003 = p < \alpha = 0.05.$$

Reject H_o that slopes are equal. Accept H_A that slopes differ.

State decision with evidence.
The rate of decrease in heterozygosity with altitude differs in the two species.

$$(F_{Sp*Alt} = 15.33 \quad df = 1,10 \quad p = 0.003)$$

Because the slopes are unequal, there is no point in testing whether the average slope β_{Alt} differs from zero. It would be appropriate to test whether the slope for each species differs from zero (next step).

10. Report and interpret parameters of biological interest.
Interaction term significant hence the slopes differ in the two species.
Because the slopes differ, we examine each species separately.
Heterozygosity decreases with altitude in *D. persimilis*

	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

The equation for the gradient in heterozygosity in *D. persimilis* is:

$$H = 0.58 - 0.127\, Alt$$

Heterozygosity does not change with altitude in *D. pseudoobscura*:

	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

There is no change with altitude, so instead of an equation for the gradient, the heterozygosity in this species is adequately summarized by the mean

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

No conclusion can be drawn about the entire biological population of *D. persimilis* and *D. pseudoobscura*. Inference was made to a hypothetical population, all possible measurements with this measurement protocol. The conclusion about this hypothetical population can be used to form expectations about heterozygosity in other situations, despite the fact that only flies from one mountain were measured.