Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.3 Analysis of Continuous Data

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11), Part IV (Ch13, 14)
16 The Generalized Linear Model
16.1 Analysis of Count Data
Binomial, Poisson, and Negative Binomial Counts
Goodness of Fit - Chisquared Statistic
16.2 Analysis of Deviance
Goodness of Fit - G Statistic
Likelihood ratio tests
Data Equations
Improvement in fit ΔG
Analysis of Deviance Table
Analysis of Deviance - Mutant frequency
16.3 Analysis of Continuous Data

on chalk board

Ch16.xls

ReCap (Ch 16) We extend the model based approach we have learned to non-normal errors. This is called the generalized linear model. GLM (normal errors) is a special case of GzLM

Today: Analysis of Continuous Data

Wrap-up.
The General Linear Model is a special case of the Generalized Linear Model. Consequently, we can carry out any GLM as a GzLM.

The example today demonstrated the analysis of deviance for fly heterzygosity data already analyzed with ANOVA.

The notation for the GzLM differs from the GLM.

The computational routine for the GzLM differs from the GLM and so we obtain somewhat different estimates of parameters and of p-values for each term in the model.

The Generalized Linear Model (normal errors).

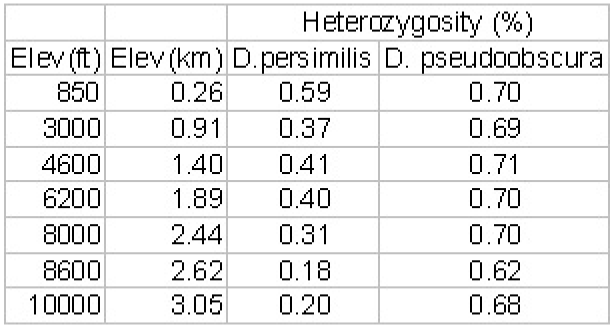
We can carry out hypothesis testing using goodness of fit (ΔG) and analysis of deviance within the framework of the Genera*lized* Linear Model. To do this we use the generic recipe for analysis of data with the General Linear Model, with only a few modifications (see Chapter 16 for modified recipe).

Fly Heterozygosity.

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

<u>Graphical model</u>.

Figure 14.1a

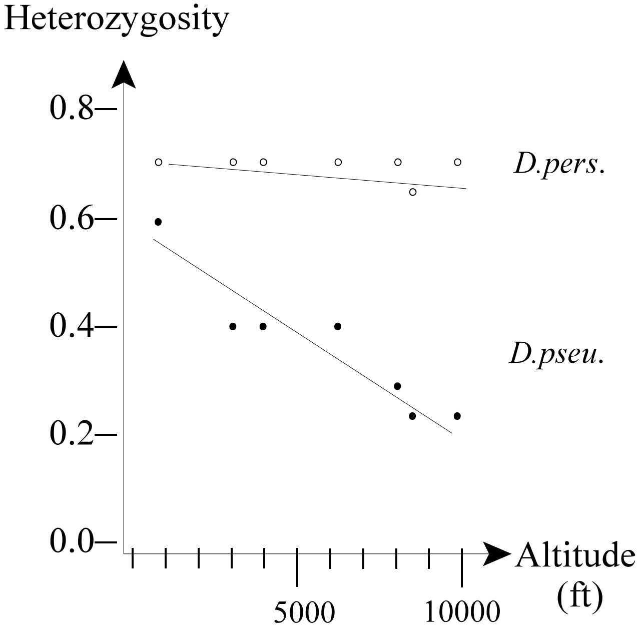
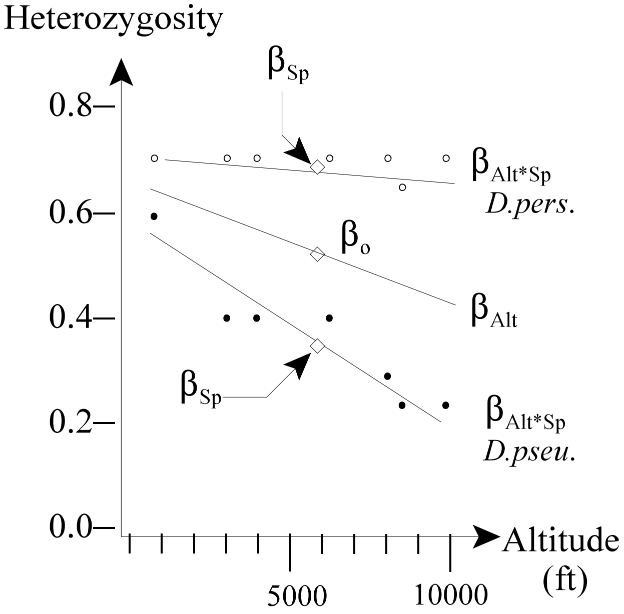


Figure 14.1b



Formal 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*

Write formal model

$$H = \mu + Normalerror$$

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

The notation differs from that for the general linear model. However, if we substitute the second expression into the first, we obtain the same model as for the general linear model.
This new notation will be needed when we move from normal errors to other error distributions.

2. Execute model.

Place data in model format (column of data for H, Alt, and Sp).

```
Options linesize=80;
Title1 'Heterozygosity in relation to Elevation
    D. persimils, D. Pseudoobscura, Brussard 1984';
Data a;
  Input Elev H  SP $;
  Cards;
    850  0.59  Dper
   3000  0.37  Dper
   4600  0.41  Dper
   6200  0.40  Dper
   8000  0.31  Dper
   8600  0.18  Dper
  10000  0.20  Dper
    850  0.70  Dpse
   3000  0.69  Dpse
   4600  0.71  Dpse
   6200  0.70  Dpse
   8000  0.70  Dpse
   8600  0.62  Dpse
  10000  0.68  Dpse
;
```

Code the model in statistical package according to the structural model μ

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

```
Proc Genmod;
   Class Sp;
   Model H = Alt  Sp  Alt*Sp/
      link=identity
      dist=normal
      type1 type3;
OUTPUT out=RESPRED p=pred stdresdev=stdresdev;
PROC PLOT data=RESPRED; plot stdresdev*pred/vref=0;
```

The structural model consists of all of the explanatory terms, including interaction terms. The structural model is specified in a model statement, which has the same format as the model statement for the general linear model.

In addition to the structural model, we need to state the error structure and the function that links the response variable to the structural model. For the general linear model, the error structure is normal and the link is the identify link.

2. Execute model.

Here is the general linear model for the fly heterozygosity data, written as a general linear model.

$H = \mu + \epsilon$ [this is the identity link]

$\epsilon \sim N(0,\sigma)$ [the error is normal, with mean of zero and dispersion σ that will be estimated from the data]

$\mu = \beta_o + \beta_{Alt} \cdot Alt + \beta_{Sp} \cdot SP + \beta_{Alt \cdot Sp} \cdot Alt \cdot SP$ [this is the structural model]

Code the model in statistical package according to the structural model μ

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

```
Proc Genmod;
   Class Sp;
   Model H = Alt  Sp  Alt*Sp/
      link=identity
      dist=normal
      type1 type3;
OUTPUT out=RESPRED p=pred stdresdev=stdresdev;
PROC PLOT data=RESPRED; plot stdresdev*pred/vref=0;
```

We calculate the standardized (deviance) residuals because we expect these residuals to be homogeneous if our choice of error structure was correct. A deviance residual is the contribution of a particular observation to the overall deviance. The deviance residuals are computed by dividing the raw residuals by a factor that makes the variance constant, if the assumed error distribution is correct. In the case of normal errors, this factor is unity and hence the raw and deviance residuals are the same.

3. Evaluate model

a. Straight line assumption. OK - no bowls or arches
b. Homogeneity of variance. OK - no cones in deviance residuals

Note that we evaluate the homogeneity of the residuals, but we no longer evaluate whether the residuals are normal.

4. State population and whether sample is representative.

As before, we will assume that the data come from a statistical population--all measurements that could have been obtained, using the procedural statement.

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

Yes. We wish to know whether the apparent difference in gradient between the two species is more than chance.

6. State H_A / H_o pair, tolerance for Type I error

Interaction term. Are the heterozygosity gradients the same ?

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

Statistic - Non-Pearsonian chisquare (G-statistic)

7. Analysis of Deviance (instead of analysis of variance).
Here is the AnoDev table for the fly heterozygosity example.

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

Source	df	G = 2*lnL	ΔG	----->	Pr>ChiSq
Intercept	1	6.5402			
Alt	1	8.2761	1.74		0.1877
Sp	1	35.9782	27.70		<0.0001
Alt*Sp	1	48.9910	13.01		0.0003

For comparison, here is the ANOVA table.

Source	df	Seq SS	MS	F	---->	Pr>F
Alt	1	0.05991	0.0599	24.19		0.0006
Sp	1	0.39111	0.3911	157.91		<0.0001
Alt*Sp	1	0.03798	0.03798	15.33		0.0029
Res	10	0.02477	0.00248			
Total	13	0.51377				

7. Analysis of Deviance

Source. The intercept is the first parameter estimated in the model. In this case the intercept is the Y-intercept of the first species group,

D. persimilis. $H_{pers} = 0.712 - 0.0145\, Alt$

df. The degrees are freedom are calculated in the same way as with the ANOVA table, for the structural part of the model. df_{total} and $\text{df}_{\text{residual}}$ are not listed.

G replaces Seq SS. The first G value is the fit of the model to a single value, the intercept. In this example (normal error, identity link) the intercept is 0.7117.
The deviance associated with intercept term is G = 6.54.

ΔG This is the change in fit associated with each term in the model.
The fit, if we add Altitude, is G = 8.2761, the change in fit is ΔG = 1.74
The fit, after we add the species term, is G = 35.98, the change in fit due to species is then ΔG = 27.7
The fit, after we add the interaction term, is G = 48.99, the change in fit is ΔG = 13.01

7. Analysis of Deviance

p-value For each change in fit, we can compute a p-value from a Chisquare distribution.

It is evident that the AnoDev and ANOVA table give different p-values for each term. This stems in part from differences in the procedures used by GLM routines (ordinary least squares) and those used by GzLM routines (iteratively reweighted least squares). The differences also stem from the basis of comparison. The ANOVA table uses ratios of variances while the AnoDev table uses differences in deviations relative to a simpler model immediately above it in the table.

The differences stemming from sequential comparisons in the AnoDev table are removed by computing an adjusted SS and adjusted ΔG. The same strategy (called Type III analysis) is used for both ANOVA and Analysis of Deviance: what is the SS or ΔG value if the term is included last in the model.

7. Analysis of Deviance

Here are the ANOVA and AnoDev tables for Type III computation (each term entered last in the model).

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

Source	df	G = 2*lnL	ΔG ---->	Pr>ChiSq
Alt	1		17.21	<0.0001
Sp	1		5.78	0.0162
Alt*Sp	1		13.01	0.0003

Source	df	Adj SS	Adj MS	F ---->	Pr>F
Alt	1	0.05991	0.0599	24.19	0.0006
Sp	1	0.39111	0.0127	5.11	0.0473
Alt*Sp	1	0.03798	0.03798	15.33	0.0029
Res	10	0.02477	0.00248		
Total	13	0.51377			

Now the results from the ANOVA and AnoDev table are similar. The two tables produce the same decisions concerning the statistical significance of each term.

Te next step is to declare a decision. We no longer need to decide whether to use randomization to overcome problems with non-normal errors.

8. Assess table, based on evaluation of residuals.
Assumptions met, continue to step 9.

9. Declare decision about terms in model.
Reject H_o that slopes are equal. Accept H_A that slopes differ.

$$0.00029 = p < \alpha = 0.05.$$

10. Analysis of parameters of biological interest.

```
                         Analysis Of Parameter Estimates

                                Standard       Wald 95%          Chi-
Parameter          DF  Estimate    Error   Confidence Limits   Square  Pr > ChiSq

Intercept           1    0.7117   0.0348    0.6435    0.7798   418.68      <.0001
SP         Dper     1   -0.1316   0.0492   -0.2280   -0.0352     7.16      0.0075
SP         Dpse     0    0.0000   0.0000    0.0000    0.0000      .           .
Elev                1   -0.0000   0.0000   -0.0000    0.0000     0.70      0.4016
Elev*SP    Dper     1   -0.0000   0.0000   -0.0000   -0.0000    21.46      <.0001
Elev*SP    Dpse     0    0.0000   0.0000    0.0000    0.0000      .           .
Scale               1    0.0421   0.0079    0.0290    0.0609
```

[slopes not reported in Elev*Sp term because too few decimal places reported]
[redo with Elev = km in order to display parameter estimates]