Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.1 GzLM for Data with Normal Errors

Part V. The Generalized Linear Model

ReCap The generalized linear model extends the model based approach we have learned to non-normal errors. GLM (normal error) is a special case of the GzLM

Today: GzLM with normal error and identity link (= GLM).

Wrap-up.

Because the General Linear Model is a special case of the Generalized Linear Model we can carry out any GLM as a GzLM with a normal error.

The example today demonstrated the analysis of deviance for normal errors and the identity link, for comparison with GLM output.

The results were similar for adjusted (Type III) analysis.

Computational routines for ANODEV and ANOVA differ in how parameters are estimated, and so can differ in estimates of these parameters and in likelihood ratios.

The Generalized Linear Model (normal errors).

The general linear model (GLM) is a special case of the generalized linear model (GzLM). A GLM is a GzLM with normally distributed errors and an additive link (the identity link) between the response variable and the structural model. To gain a preliminary understanding of the analysis of deviance we will compare the ANODEV table to the ANOVA table for the fly heterozygosity ANCOVA (Ch14.1).

Fly Heterozygosity.

1. Construct model

Verbal model:

Inversion heterozygosity changes with altitude, depending on species.

Formal model

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

H is inversion heterozygosity

Alt is elevation above sea level (feet)

Sp is two species of fruit fly, *Drosophila persimilis* and *D. pseudoobscura*

β_{Alt} is the vertical heterozygosity gradient (%/km) for both species.

$\beta_{Alt\cdot Sp}$ is the difference in the gradient between the two species.

With the GzLM we separate the structural (research) model from the error model in order to permit a flexible link between the two.
A GLM in this format is:

Distribution $H \sim Normal(\mu, \sigma)$

The errors (residuals) are distributed normally around mean μ with constant dispersion of σ

μ refers to the fitted model, not to the grand mean β_o

Link function $H = \eta$ This is called the identity link.

$$\eta = \beta_o + \beta_{Alt}Alt + \beta_{Sp}Sp + \beta_{Alt\cdot Sp}Alt \cdot Sp$$

A GzLM allows a choice of link functions. For example, we could use a log link ($H = e^{\eta}$) if we expected heterozygosity to change in an exponential fashion with change in altitude. For a GLM the link is always the identity link.

2. Execute model.

Place data in model format (a column of data for *H*, Altitude, and Species).

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

Specify the error distribution and the link function.

```
Proc Genmod;
   Class Sp;
   Model H = Sp Alt  Alt*Sp/
      link=identity
      dist=normal
      type1 type3;
```

2. Execute model.

For a GzLM we use scaled residuals to diagnose the assumptions. Deviance residuals are recommended because the make the variance constant if the error distribution is correct. A deviance residual is the contribution of a particular observation to the overall deviance. In the case of normal errors, the scaling is unity and hence the raw and deviance residuals are the same.

```
FlyMod < - glm(H ~ Sp + Alt + Alt*SP,
     family = Gaussian (link = identity))
anova(Flymod, test=”chisq”)  #Sequential deviances
Anova(Flymod, type=III)     #Adjusted deviances
```

3. Evaluate model

Residual vs fit and normal plots for GzLM are the same as from GLM routines

4. Partition df and Deviance according to model.

Source. The intercept is always a constant.
In this case the intercept is the Y-intercept for *D. persimilis.*

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

df. The degrees of freedom for each term are calculated in the same way as with the ANOVA table. In this case, 1 df for each term in the model. The ANODEV table uses Δdf , the change in df with the addition of each new term in the model.

Source	$G = -2^*\ln L$	Δdf	ΔG	$p > \chi^2$
Intercept	6.5402			
SP	26.5936	1	20.05	<.0001
Elev	35.9782	1	9.38	0.0022
Elev*SP	48.9910	1	13.01	0.0003
Residual		10		

G is the "goodness of fit." The fit improves with the addition of each term.
L is the likelihood of successive models
ln*L* is the log likelihood of successive models
The goodness of fit for the full (null) model (a constant of 0.7117) was
$G = 6.54/2$
The improvement in fit for the omnibus model (all terms) is
$\Delta G/2 = (48.99 - 6.54)/2 = 42.45$
The likelihood ratio is $\exp(42.45/2) = 1.6 \times 10^9$
There is very strong evidence for the reduced relative to the null model.

5. Choose mode of inference. Is hypothesis testing appropriate?

We will calculate the evidential support for each term in the model as a likelihood ratio. Instead of controlling Type I error, we will use Fisher's 4 levels of definite support to interpret the results relative to Type I error.

5. State population and whether the sample is representative.

Inference is to a prospective population generated by Hacking's (1985) definition: many repeats of the collection protocol and the protocol for measuring heterozygosity.

6. State test statistic and treatment of Type I error.

Test statistic – the non-Pearsonian chisquare (G-statistic)
Type I error sorted at 4 levels: high ($p > 0.1$, moderate ($0.1 < p \leq 0.05$), Low ($0.05 < p \leq 0.01$), very low ($p < 0.01$).

6. State H_A / H_o pairs.

Interaction term. Are the heterozygosity gradients the same ?

Deviance($\beta_{Alt}Alt$) $>$ 0 Same as H_A:$\beta_{pers} \neq \beta_{pse}$

Deviance($\beta_{Alt}Alt$) = 0 Same as H_o: $\beta_{pers} = \beta_{pse}$

7. ANODEV table.

Here is the analysis of deviance table with F-ratio tests.

```
            ΔDf ΔDeviance   Df    Deviance  Pr(>F)
NULL                        13    0.51377
SP            1  0.39111    12    0.12266 157.907 1.9e-07 ***
Elev_km       1  0.05991    11    0.06274  24.189 0.00060 ***
SP:Elev_km    1  0.03798    10    0.02477  15.332 0.00288 **
```

ΔDeviance is the improvement in fit. For example 0.51377-0.12266 = 0.3911

Here is a comparison to the ANOVA table.

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

7. ANODEV table.

Sequential analysis in an ANOVA table with a covariate produces different results, depending on the order in which terms were listed. This dependency is removed by obtaining an adjusted SS. The same tactic (called Type III analysis) is used for analysis of deviance: what is the ΔG value if the term is estimated as if it were last in the model?

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

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			

The Type I error estimates from the ANOVA and ANODEV table are similar but not exactly the same because the test statistic differs.

Source	df	Δdf	G = -2*lnL	ΔG	LR=exp(G/2)
Intercept		1	6.5402		
Alt	13	1	8.2761	1.74	2.4
Sp	12	1	35.9782	27.70	10^6
Alt*Sp	11	1	48.9910	13.01	668
Residual	10				

The analysis of deviance, unlike the ANOVA table, yields a measure of the evidence, the likelihood ratio.

8. Evaluate sensitivity to deviations from assumptions.

Assumptions were met (see Ch 14.1).

9. Report statistical conclusion.

We begin with the interaction term $LR = \exp(13.01/2) = 668$
The model with interactive effect is 670 times more likely than the model without that term. There is strong evidence for differing heterozygosity gradients in the two species.

$G = 13.0$, $p = 0.0003$ from chisquare distribution with 1 degrees of freedom. Type I error from the likelihood ratio is well below Fisher's most conservative level. $0.0003 = p < \alpha = 0.01$.

Given the interactive effect, no conclusions are made about the main effect, the gradient in heterozygosity regardless of species.

10. Report science conclusion. Analysis of parameters of biological interest.

Given the evidence, we would report the gradient for each species, not the gradient for both combined.
The heterozygosity gradient in *D. pseudoobscura* is – 0.127 % / km
$H_{pers} = 0.580 - 0.127$ Alt
The gradient in *D. persimilis* is statistically indistinguishable from zero.
$H_{pseu} = 0.712 - 0.0145$ Alt The regression is no better than the mean $H = 0.686$.