Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.3 Analysis of Deviance

ReCap. Parts I – IV. The General Linear Model
Part V. The Generalized Linear Model
16 Introduction: The Generalized Linear Model
16.1 Analysis of Count Data
    Binomial, Poisson, and Negative Binomial Counts
16.2 Goodness of Fit - Chisquare Statistic
    Goodness of Fit - G Statistic
    Testing Extrinsic Hypotheses
    Intrinsic Hypotheses: Two-way Contingency Test
16.3 Analysis of Deviance
    Intro
    Analysis of Deviance: Extrinsic Hypotheses
    Improvement in Fit Δ*G*
    Analysis of Deviance
    Example: Mutant frequency
    Example: Leaf type in two soil types.
16.4 GzLM for Normal Errors
16.5 Notation:
    Normal errors (GLM)

Data: Ch16.xls data.

ReCap (Ch 16.2) The G statistic is used in place of the classical chisquare statistic because it has better statistical properties.

Today: Model-based Analysis of Goodness of Fit - Analysis of Deviance

Wrap-up.

The analysis of deviance table is used to display improvement in fit due to adding a term to a generalized linear model.

The χ^2 distribution is used to declare a statistical decision about the improvement in fit.

Analysis of deviance and the GzLM can be applied to extrinsic hypotheses, such as a Mendelian ratio.

Analysis of deviance and the GzLM are applied to intrinsic hypotheses, such as comparing two proportions (two-way contingency test).

Intro.

Binomial, Poisson, and negative binomial counts will not meet the assumptions for GLM.

- The variance will depend on the mean and as a result, a plot of errors (residuals versus fits) will look like a cone.
- Counts are bounded at zero and as a result, the distribution of residuals will be asymmetrical for each fitted (model) value.

This problem will be serious if there are zero counts in the data, if fitted values are close to zero, or if the variance to mean ratio is large.

To analyze count data, we will use the general***ized*** linear model. The GzLM allows us to assume that the residuals arise from an appropriate distribution such a binomial, Poisson, or negative binomial (for overdispersed data).
The GzLM also allows us to compare proportions in a natural way, as ratios.

- How does the ratio of purple to white flower plants compare to the ratio expected from genetic theory?
- Does the proportion of animals with tumors increase with increasing dose of a suspected carcinogen?
- Is the risk of cancer higher in physicians who smoke than physicians who don't smoke ?
- Are accidental deaths of coal miners disproportionately common in some years?

When comparing proportions we take ratios (50% / 75% = 2/3) instead of differences (50% - 75% =?). With ratios, we avoid having to take a log transform, which is undefined for zero counts.

To compare proportions we use the odds ratio, defined as Odds = p / (1-p)
For the Mendel pea data the observed odds of a plant having purple flowers is:
Odds = (705/929)/(224/929) = 3.147 Odds = 3.147:1

The odds ratio has the nice property that it is reversible. The odds of plant having white flowers is 1 / 3.147 = 0.3177

Model-based Analysis of Goodness of Fit. Extrinsic Hypotheses.

Improvement in Fit ΔG

Examples of comparing observed proportion to proportion expected from theory.

Genetic analyses (*e.g.* 3:1 odds for Mendel pea data)

Sex ratio (Fisher's theory of 1:1).

From theory, the expected odds of purple flowered plants will be $Odds_{Mendelian} = 3{:}1$

To compare the observed odds to Mendelian odds we take the odds ratios

$$OR = Odds_{Observed} / Odds_{Mendelian}$$

$$OR = (3.147{:}1) / (3{:}1) \qquad OR = 3.147{:}1$$

Rearranging the odds ratios definition yields the statistical model that relates the observed (response) to the expected (explanatory) odds:

$$Odds_{Observed} \quad = Odds_{Mendelian} \quad * OR$$

Taking the logarithm gives us differences on an additive scale:

$$\begin{array}{llll}
 & \ln(Odds_{Obsserved}) & = \ln(Odds_{Mendelian}) & + \ln(OR) \\
 & \ln(Odds_{Obsserved}) & = \ln(Odds_{Mendelian}) & + \text{residual} \\
\text{The data equation is:} & \ln(3.147) & = \ln(3) & +\text{residual} \\
 & 1.146552 & = 1.0986 & + 0.04794 \\
\text{degrees of freedom} & n - 0 & = 0 & + 1
\end{array}$$

In this case n =1, and degrees of freedom from the model are zero because the value is from theory, the value is not an estimate from the data. Is the residual (difference between the observed and the theoretical expectation) too large to be due to chance? The observed odds are close to the Mendelian odds of 3:1, the odds ratio is close to 1 and ln(*OR*) is close to zero. Assuming the residual is distributed as chisquare, the p-value is 0.83 on one degree of freedom. We conclude that the residual is *not* too large to be due to chance.

Of interest is the large p-value for this experiment, supporting strong agreement with theory. R.A. Fisher (1936) noted that Mendel's data from the later years of experiments on peas were biased toward agreement with theory "a possibility among others that Mendel was deceived by some assistant who knew too well what was expected."

Fisher, R. A. (1936). Has Mendel's work been rediscovered? Annals of Science 1:115–137.

Improvement in Fit Δ*G*.

An exact calculation, as above, is rarely possible. When a parameter is estimated, we use the deviance to calculate the improvement in fit. The fit of data to theory is $G = 0.393$; the fit to the observed is perfect: $G = 0$.

$Odds = OR * Odds_M$ $\quad G = 0.393 \quad$ (Odds in population are 3:1)

$Odds = 1 * Odds_M$ $\quad G = 0.0 \quad$ (Odds in sample are 3.147:1)

The improvement in fit is $\quad \Delta G = 0.393 - 0 = 0.393$

The improvement in fit is not statistically significant

($\Delta G = 0.393$ df = 1 p = 0.53)

Analysis of Deviance Table

The improvement in fit is tabulated in an Analysis of Deviance table.

The AnoDev table reports the change in fit due to adding a term to the model

Here is the analysis of deviance table for the Mendelian model of pea flower data.

Source	df	$G = 2*\ln L$	ΔG ----->	Pr>ChiSq
3:1 ratio		0.393		
Observed ratio	1	0.0	0.393	0.53

The model terms are listed as sources, just as in the ANOVA table. The AnoDev table has no residual term. In this example the sources are fit to the extrinsic hypothesis (2 colors in a 3:1 ratio). This is compared to the fit if the population odds are exactly the same as the observed odds ($G = 0$).

df. We have no df_{total} or $df_{residual}$. Degrees or freedom are listed according to the number of parameters estimated for each term in the model. In this case there is no *df* for the 3:1 ratio because this parameter is not estimated. The *df* of the observed ratio is 1 because one parameter is estimated from the data (the odds = $p / (1-p)$ = (705/929) / (224/929) = 705 / 224

G replaces Seq SS. The first *G* value is the fit of the model to the 3:1 ratio.

In this example the intercept is the log of the Odds ratio, ln(*Odds*) = 0.04794

The deviance if the 3:1 odds are true is $G = 0.393$.

The deviance if the observed odds are true is $G = 0$ (the fit is perfect)

ΔG There is no error term so we compute the change in fit: $\quad \Delta G = 0.393$

p-value We compute a p-value from a chisquare distribution.

Model-based Analysis of Goodness of Fit. Extrinsic Hypothesis.

Example - Mutant Frequency.

Data from Table 17.1 in Sokal and Rohlf 1995
The frequency of offspring of two phenotypes, wild and mutant. $f = [80\ 10]$

The proportion of wild type offspring: $f_W / N = p_W = 80/90 = 0.89$
The proportion of mutant offspring: $f_M / N = p_M = 10/80 = 0.11$

The odds of wild type offspring: $Odds_W = 80/10 = 8{:}1$
The odds of mutant offspring: $Odds_M = (Odds_W)^{-1} = 10/80 = 0.125 : 1$

Can the observed proportion of mutant offspring be explained by a simple recessive gene, which is expressed in 1 out of 4 offspring ?
The expected proportion of mutant offspring: $\mathrm{E}(p_M) = 1/4 = 0.25$

Equivalently, can the observed odds of mutant offspring be explained by a simple recessive gene expressed in 1 out of 4 offspring ?
The expected odds of mutant offspring: $\mathrm{E}(Odds_M) = 1{:}3 = 0.33{:}1$

1. Construct Model

Response variable is observed odds of mutant offspring.
Explanatory variable is Mendelian odds of mutant offspring, if a single recessive.

The model for frequency is: $f = \mathrm{E}(p) \cdot N + \text{residual}$

The model for odds is $Odds_M = OR * \mathrm{E}(Odds)$

Our estimate of the odds ratios is: $OR = (10/80)/(1/3) = 0.375$

2. Execute Analysis.

$$\boxed{\begin{array}{llll} f & = \mathrm{E}(p) \cdot N & + & \text{residual} \\ 80 & = 0.75 \cdot 90 & + & 12.5 \\ 10 & = 0.25 \cdot 90 & + & -12.5 \end{array}}$$

3. Use Residuals to Evaluate Model.

We have too few residuals to undertake any diagnosis of homogeneity.

We can check independent trial assumption. The assumption of 90 independent trials could be checked by looking for runs of wild or mutant phenotypes in the data, in the order it was obtained. A quick check, if neighbors are known, is to plot scores (0/1, y/n, present/absent *etc*) against neighbors.

4. Population = ?

All possible outcomes, given random combination of wild and mutant alleles [W M] at this locus, for $N = 90$ offspring.

5. Mode of Inference Hypothesis testing.

6. State H_A / H_o with tolerance for Type I error.

H_A: $f \neq E(p) \cdot N$ $G > 0$ G will be too large to be due to chance
H_o: $f = E(p) \cdot N$ $G = 0$
$\alpha = 5\%$

7. AnoDev Calculate Improvement in Fit ΔG

The fit of the observed to expected is $G = 10.97$.
The improvement in fit is $\Delta G = 10.97 - 0 = 10.97$

f	$=$	$E(p) \cdot N$	$+$	residual	$\ln L = f \ln(f / e^{\beta} N)$
f	$=$	$e^{\beta} N$	$+$	residual	
80	$=$	$0.75 \cdot 90$	$+$	residual	13.592
10	$=$	$0.25 \cdot 90$	$+$	residual	-8.109
		$\Sigma f \ln(f / (E(p) \cdot N))$		$=$	5.483
		$G = 2 \Sigma f \ln(f / (E(p) \cdot N))$		$=$	10.97

In order to calculate the probability of the observed value of ΔG we need a distribution of outcomes. We use the chisquare distribution.

Here is the computation, using Minitab.
We have two data equations ($n = 2$)
$df = n - 1 = 1$

```
MTB> cdf 10.97;
SUBC>chisquare 1.
   10.97     0.999074
```

The degree of freedom is lost because once we compute the expected frequency of mutant phenotypes, $E(p) \cdot N = 22.5$, the expected frequency of wild types will not be free to vary. It must be $90 - 22.5 = 67.5$

8. Recompute p-value if assumptions not met.

We have no residuals with which to evaluate assumptions. Randomization based on 90 trials with 3:1 odds (*e.g.* HH in two coin tosses) will give same result as the chisquare p-value for the *G*-statistic. We would have to have some knowledge of the violation in order to recompute a p-value for this analysis.

9. Declare decision.

Using the Chi-square distribution with df = 1, we calculate that 99.91% of the *G*-statistics will be less than 10.97, if the data do indeed come from a population with an expected proportion 1 mutant in 4 offspring.
The p-value is 1 - 0.999074 = 0.00093, a very small probability.

$0.00093 = p < \alpha = 5\%$

Reject H_o Accept H_A that observed frequencies differ from 3:1 ratio
G = 10.97 df = 1 p = 0.00093

10. Report and interpret parameters of biological interest.

The observed proportion (8 /9)differs significantly from the theoretical proportion of (1/4)

Model-based Analysis of Goodness of Fit. Intrinsic Hypothesis.

Sokal and Rohlf 1995 Box 17.6

	Leaf Type		
Soil	pubescent	smooth	
Serpentine	12	22	34
Not Serpentine	16	50	66
Total	28	72	100

Example. Leaf Type.
Data from Sokal and Rohlf 1995.
Leaf type of 100 trees found in two soil types in an area of 400 square miles
Does the proportion of smooth leaves in serpentine soil (p_{serp} = 22/34 = 65%) differ significantly from the proportion ($p_{nonserp}$ = 50/66 = 76%) in non-serpentine soil?

This is the row by column contingency test, widely used the social sciences as well as in biology. It is typically presented as goodness of fit test of whether the row proportions differ between rows, or equivalently, whether the column proportions differ between columns. Cochran (1954) noted some of the problems with this approach, including lack of power, no measure of effect size, failure to meet the assumptions for using the $\chi 2$ distribution when the expected count in any one cell is small (*e.g.* less then 5) and difficulties when extending the analysis to several 2x2 tables. Cochran recommended analysis on a probit scale, *i.e.* taking the odds ratio across rows (or equivalently) across columns. A generation later McCullagh and Nelder (1972 *Generalized Linear Models*) similarly recommended comparing odds, using a binomial error.
Cochran, W.G. 1954. Some Methods for Strengthening the Common $\chi 2$ Tests. *Biometrics* 10: 417-451

1. Construct Model
Response variable is observed odds of pubescent leaf.
Explanatory variable is soil type.

The model is $Odds_{serpentine} = OR * Odds_{nonserpentine}$ $\qquad$ OR = (22/12)/(50/16) = 0.59

The odds ratio across leaf types is the same (OR = (22/50)/(12/16) = 0.59)
but this hardly justifies taking soil type as dependent on leaf type.

2. Execute Analysis.

$$\boxed{\begin{array}{lllll} Odds & = OR^{*}\,Odds_{ref} & & & \\ Odds_{serp} & = e^{\beta ref} & = & 1.83 = & 22/12 \\ Odds_{non} & = e^{\beta ref}\ e^{\beta non} & = & 3.13 = & 22/12 * 1.7 \end{array}}$$

We arbitrarily choose serpentine soil as the reference category.
The odds in the other category is the odds ratio times the reference odds.

3. Use Residuals to Evaluate Model.

There are no residuals to evaluate. This is a saturated model, which means there are as many parameters (2) as there are observations (2).

4. Population = ?

All possible outcomes, if the survey carried out the same way repeatedly in the same ecosystem.

5. Mode of Inference Hypothesis testing.

6. State H_A / H_o with tolerance for Type I error.

H_A: $OR \neq 1$ $\quad$ $G > 0$ $\quad$ G too large to be just chance
H_o: $OR = 1$ $\quad$ $G = 0$
$\alpha = 5\%$

7. AnoDev Calculate Improvement in Fit ΔG

The deviance is 1.332
This is the fit of each observation to a 72:28 ratio of smooth to pubescent

The deviance drops to zero when the observed data are fit to a 22:12 ratio for serpentine, and 50:16 for non serpentine.
The improvement in fit is $\Delta G = 1.332$.

Serpentine	f	p hat	fhat	f*ln(f/fhat)	G	
Pubescent	12	0.28	9.52	2.778		
Smooth	22	0.72	24.48	-2.350		
Total	34		34	0.428	0.856	= Gserp
NonSerpentine						
Pubescent	16	0.28	18.48	-2.306		
Smooth	50	0.72	47.52	2.544		
Total	66		66	0.238	0.476	= Gnonserp
					1.332	= SumG

8. Calculate the p-value.

The p-value from the chisquare distribution is
$p = 1 - 0.752 = 0.248$

9. Compare p to α to make decision.

$0.25 = p > \alpha = 5\%$

We accept chance as an explanation of the difference in proportion of smooth seeds in the two soil types.

$G = 1.332$ df = 1 $p = 0.25$
Accept H_o that observed proportions are due to chance.

10. Report and interpret parameters of biological interest.

The observed odds were 1.7 times higher in nonserpentine than serpentine soil, but this odds ratio is statistically indistinguishable from *OR* =1. We cannot conclude that the odds of encountering smooth leaves differs between soil types.

Extending what you have learned.

Set up a spreadsheet (see step 7) that calculates

- the marginal totals, given the four numbers.
- the number of leaves *N*, from the sum of the four numbers in the table.
- the odds ratio for leaf type across soils (see step 1)
- the G-statistic from the table ($G = 1.332$ in the example above)
- the p-value for the *G*-statistic, on a single degree of freedom.

What happens to *G* when you double all four numbers in the table ?

Minimum sample, given the odds ratio. Multiply the 4 numbers in the table by larger values than 2, until G becomes significant. What is the value of *N* at which *G* becomes significant?

Minimum odds ratio, given the sample size. For the table where $N = 100$, alter the ratio of leaf types in non-serpentine soil to more extreme values, keeping the row total constant (15 + 51, 14 + 52, etc).
What is the odds ratio at which *G* becomes significant?

Maximum odds ratio, given the sample size. For the table where $N = 100$, alter the ratio of leaf types in non-serpentine soil to more extreme values in the other direction, keeping the row total constant (17 + 49, 18 + 48, *etc*).
What is the odds ratio at which *G* becomes significant?