Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 18.7 Logistic ANCOVA.

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)	
18 Binomial Response Variables	
18.1 Logistic Regression (Dose-Response)	Ch18.xls
18.2 Single Factor. Prospective Analysis	
18.3 Single Factor. Retrospective Analysis	
18.4 Single Random Factor.	
18.5 Single Explanatory Variable. Ordinal Scale.	
18.6 Two Categorical Explanatory Variables	
18.7 Logistic ANCOVA	

on chalk board

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning
ReCap Part II (Chapters 5,6,7) Hypothesis testing and estimation
ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12,13,14,15) GLM with more than one explanatory variable
ReCap (Ch 16,17)
ReCap.
Response variables of interest in the natural and social sciences are often binomial: a series of trials (cases) that can be scored as yes/no, present/absent, etc.

We compared a binomial proportion in relation to one or more categorical explanatory variables.

We can extend this to ANCOVA.

Today: We will compare logistic regressions across categories (logistic ANCOVA).

Wrap-up.
The generalized linear model permits us to apply what we have learned about multiple explanatory variables to the analysis of binomial response variables in an ANCOVA design.

Context.
In a previous analysis we used an ANCOVA design to compare change in genotype with change in altitude in 2 species of fruitfly. One of the central ideas in quantitative genetics is that trait (phenotypic) variation depends on both genes and environment. In fact, when we assign variability in trait to genes versus environment we find that the heritability of a trait (the proportion of the phenotypic variance attributable to inheritance) is often small compare to interactive variability (Falconer, D.S. 1960 *Introduction to Quantitative Genetics*. Ronald Press, New York). In the example today we look at this interactive effect in more detail. Specifically we will analyze trait variability in relative to an environmental variable (altitude above sea level) in two genotypes. The data come from an unpublished study by E. Bottini, as reported in Sokal and Rohlf (1995 Box 17.15, 2012 Boxes 17.13 and 17.15).

Binomial Frequencies. Comparison of two logistic regressions

The data are frequencies of the A+BA phenotype in 13 Sardinian villages at elevations ranging from 10 m to 1000 m above sea level. The frequencies (*Fpres*) are reported by two genotypes ADA1 and ADA2. Each individual is scored as presence or absence of the A+BA phenotype. The underlying variability is binomial. The presence/absence data are conveniently gathered into ratios for each genotype in each village, as in the table shown here. The change in these proportions with altitude and genotype will be analyzed as odds ratios using a binomial error structure.

Fpres	N	Gtype	Elev(m)	Village
9	15	ADA2	1000	Fonni
34	100	ADA1	1000	Fonni
3	13	ADA2	797	Seulo
27	112	ADA1	797	Seulo
3	13	ADA2	796	Aritzu
13	107	ADA1	796	Aritzu
9	18	ADA2	648	Burcei
26	67	ADA1	648	Burcei
9	25	ADA2	590	Lanusei
43	112	ADA1	590	Lanusei
4	20	ADA2	550	Bitti
26	119	ADA1	550	Bitti
3	10	ADA2	442	Jerzu
16	45	ADA1	442	Jerzu
5	7	ADA2	345	Lode
34	119	ADA1	345	Lode
1	9	ADA2	228	Sedilo
29	104	ADA1	228	Sedilo
2	8	ADA2	185	Ottana
38	94	ADA1	185	Ottana
1	4	ADA2	45	Villasimius
25	56	ADA1	45	Villasimius
1	15	ADA2	15	Tortoli
38	107	ADA1	15	Tortoli
9	26	ADA2	10	Oristano
62	185	ADA1	10	Oristano

1. Construct Model

Verbal. Do the odds of the A+AB phenotype change with altitude, depending on genotype?

Response variable:

Fpres - presence of A+AB

N individuals

Odds of ADA1 or ADA2 genotype calculated from *Fpres* and *N*

Explanatory variables:

Altitude (ratio scale)

Gtype (2 categories)

Graphical model

Plot the proportion of A+AB versus altitude, for both genotypes.

Plot of odds of A+AB versus altitude, for both genotypes.

Formal model

Distribution $Fpres \sim$ Binomial(N,π)

Link $Odds = e^{\eta}$

$\eta = \beta_{Ref} + \beta_{Gtype} \cdot Gtype + \beta_{Alt} \cdot Alt + \beta_{G \cdot A} Gtype \cdot Alt$

$e^{\beta_{Ref}}$	Odds, reference class. Village = Fonni at 1000 m. ADA2
$e^{\beta_{Gtype}}$	Odds ratio, ADA1 compared to ADA2
$e^{\beta_{Alt}}$	Change in odds with change in altitude. Ref group = ADA2
$e^{\beta_{Gtype \cdot Alt}}$	Change in odds with altitude, ADA1 compared to ADA2

With the logit link, we have a linear model with the same structural model as the GLM ANCOVA. While it is not a general linear model ancova, we could call it a binomial ancova for the sake of consistency

2. Execute analysis.

Data are already in model format

There are two columns for the response variable *Fpres* and *N*.

There are two columns for the explanatory variables, *Gtype* and *Alt*.

Village is included as accessory information.

```
Data A;
  Input Fpres N Gtype $ Alt Village $;
  Cards;
    9   15 ADA2 1000 Fonni
   34 100 ADA1 1000 Fonni
      .
      .
    9   26 ADA2   10 Oristano
   62 185 ADA1   10 Oristano
;
```

SAS data definition file

2. Execute analysis.

Execute analysis according to model

$$Odds = e^{\eta}$$

$$\eta = \beta_{Ref} + \beta_{Gtype} \cdot Gtype + \beta_{Alt} \cdot Alt + \beta_{G \cdot A} Gtype \cdot Alt$$

```
Proc Genmod;
  Classes Gtype;
  Model Fpres/N = Gtype Alt Gtype*Alt/
  Link=logit dist=binomial type1 type3;
  Output out=B p=fit resdev=res;
Proc Plot;
  Plot res*fit/vref=0;
  Plot res*Elev/vref=0;
```

SAS command file

```
Mod<-glm(Fpres/N~Gtype+Alt+Gtype*Alt, weight=N,
     family=binomial(link=logit),data=SRBX17_15)
plot(Mod)
anova(Mod)
summary(Mod)
```

R script file

2. Execute analysis.

Obtain fitted values.

```
                    Analysis Of Parameter Estimates
                                 Standard       Wald 95%
Parameter        DF  Estimate     Error   Confidence Limits

Intercept         1   -1.1958    0.2984   -1.7807   -0.6110
Gtype      ADA1   1    0.6091    0.3131   -0.0046    1.2228
Gtype      ADA2   0    0.0000    0.0000    0.0000    0.0000
Alt               1    0.0009    0.0005   -0.0000    0.0019
Alt*Gtype  ADA1   1   -0.0015    0.0005   -0.0025   -0.0004
Alt*Gtype  ADA2   0    0.0000    0.0000    0.0000    0.0000
Scale             0    1.0000    0.0000    1.0000    1.0000
```

SAS output file

$e^{\beta_{ref}} = e^{-1.1958} = 0.3025$ odds, reference group

$e^{\beta_{Gtype}} = e^{0.6091} = 1.838$ odds ratio, ADA1 relative to ADA2

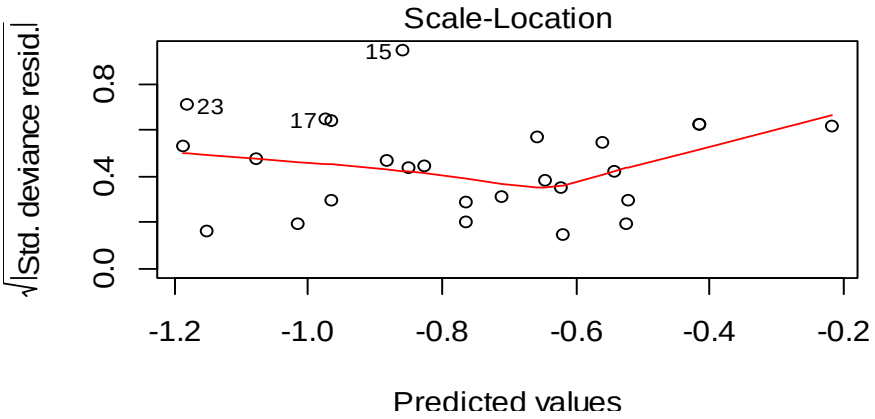
$e^{\beta_{Alt}} = e^{0.0009} = 1.0009$ change in odds with change in altitude ADA2

$e^{\beta_{Alt*Gtype}} = e^{-0.0015} = 0.9985$

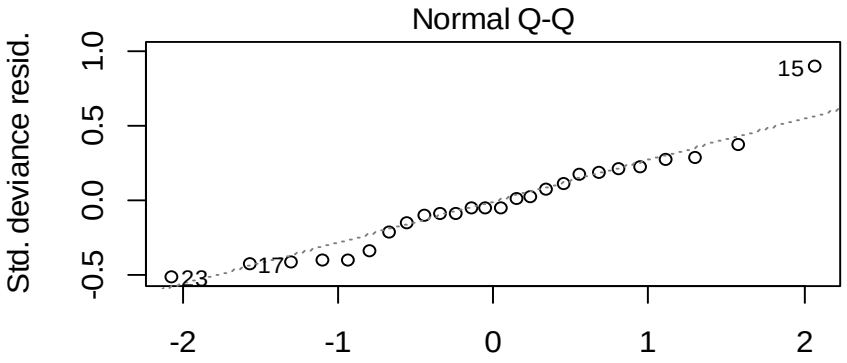
change in odds with change in altitudeADA1 relative to ADA2

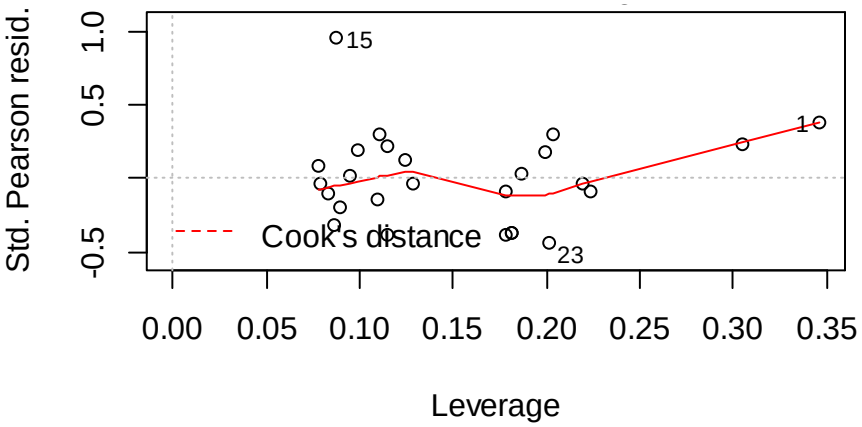
3. Evaluate model

A. Evaluate straight line assumption with residuals versus fit plot.
The straight line on a logit scale is acceptable--no bowls or arches.



B1. Distributional model (binomial) acceptable. No fans or spindles.
B2. The residuals are normally distributed except for one outlier.





The outlier exerts no leverage on the regression line.

4. What is the evidence?

The deviance for the full (null) model is: 67.7442
The deviance for the omnibus model (reduced model) is: 55.9706
The improvement in fit is: 11.7736
The likelihood ratio is: 360
The mechanistic model is 360 times more likely than the null model.
We continue with analysis of individual terms in the model.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	67.7442			
Gtype	67.6241	1	0.12	0.7290
Alt	63.5724	1	4.05	0.0441
Alt*Gtype	55.9706	1	7.60	0.0058

5. Analytic Mode.

We have many choices.

Exploratory? No. We have a model based on the biology.

Bayesian? No. We have insufficient prior information to set up a defensible prior probability.

Frequentist? Yes. We have survey and measurement protocols that are repeatable. We will infer to long run probabilities from these protocols.

Decision theoretic? No. We have no way of gauging Type I versus Type II error. Optimal power at fixed Type I error is not relevant. We do not need to control Type I error.

Evidentialist? Yes. We have little need of probabilities to temper judgement based on likelihood ratios because all of our comparisons will be single degree of freedom tests.

In previous examples we have seen that a large likelihood ratio often results in a similarly small p-value. In this case the *p*-value on $G = 11.7736$ with 3 degrees of freedom is $p = 0.0082$. Given that *p*-values do not measure evidence and likelihood ratios do (Royall 1997) and recommendations against declaring significance at p = 5% (ASA 2019) why not use likelihood ratios instead of p-values? The answer is that likelihood ratios do not give the same result as *p*-values where the model of interest has explanatory variables with many parameters. Likeliood are not a replacement for p-values. They are a meaure of evidence for which we can calculate Type I error if we have need for it.

5. Analytic Mode.

Here is an example for an ANOVA with normal error and *p*-value from the *F*-distribution.

r^2	n	model parameters	F-ratio	p-value	Likelihood ratio	LR gradient
20%	30	1		0.013	28	28
20%	30	10		0.886	28	2.8

For the algebraically inclined, the table reflects the fact that the *LR* increases with increase in number of observations, while the *p*-value via the *F*-ratio is tempered by the number of parameters estimated. It is of interest to note that the LR gradient, defined as LR/Δdf, yields a conclusion similar to that from the *p*-value calculated from the *F*-ratio.

In the genotype analysis all of the terms in the model have a single degree of freedom. In the absence of a defined cost of Type I error, or even a ratio of Type I to Type II error, we will take a purely evidential approach, using only the likelihood ratios. We will not expect that this evidentialist approach will necessarily give us the same conclusion as a Neyman-Pearson decision theoretic approach aimed at rejecting a null hypothesis.

6. Population and sample. Hypotheses.

This is an observational study with many sources of uncontrolled variability. The results may not apply to other genotypes or other locations. The basis for inference is the probability model, which is logically applicable, and shown by residual diagnostics to be acceptable. The measurement protocol could be used to define a population of infinite number of repetitions of the experimental design with this genotype and location (Sardinia).
The sample will be haphazard, taken as representative. If necessary, we can infer to a population of random outcomes from data at hand, using a randomization test such as a permutation test or a jackknife.

7. ANODEV - Calculate improvement in fit due to explanatory variables.

Beneath each term in the model we list the df, the change in df, the deviance G^2, and the change in deviance. Here is a horizontal layout of the Anodev table.

Odds =	$\exp(\beta_o) \cdot \exp(\beta_{Alt}) \cdot \exp(\beta_{Gtype}) \cdot \exp(\beta_{Alt*Gtype})$			
Df	25	24	23	22
Δdf	1	1	1	1
Deviance	67.62	63.57	55.97	
ΔDev	0.12	4.05	7.6	
	Full model		Reduced model	

7. ANODEV table. The ANOVA table is replaced by the analysis of deviance table. The Anodev table shows the fit (deviance) and improvement in fit (change in deviance) for a sequence of models (Type I analysis). Alternatively, it displays the improvement in fit for terms when fitted last (Type III analysis).

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	67.7442			
Gtype	67.6241	1	0.12	0.7290
Alt	63.5724	1	4.05	0.0441
Alt*Gtype	55.9706	1	7.60	0.0058

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq	LR
Gtype	1	4.03	0.0447	7.5
Alt	1	0.62	0.4305	1.4
Alt*Gtype	1	7.60	0.0058	44.7

SAS output file with LR added

7. ANODEV – Interpretation

We begin with the interaction term. The interactive effect is 45 times more likely than no interactive effect. Given the good evidence we have for an interactive effect we do not interpret the main effects.

8. Re-compute LR if assumptions clearly violated and sample size is small.

Assumptions were met.

9. Statistical conclusion. The single degree of freedm interactive term is 45 times more likely than no interactive effect.

10. Biological conclusions.

Phenotypic expression depends strongly on environment. The change in odds with change in altitude differs for the two loci. The elevational gradient is small in both genotypes.

$e^{\beta_{Alt}} = e^{0.0009} = 1.0009$ Change in odds with change in altitude ADA2

$e^{\beta_{Alt*Gtype}} = e^{-0.0015} = 0.9985$ Change in odds with change in altitude, ADA1 relative to ADA2

Nature versus nurture has no basis in fact. Biologists recognize that individual variation depends as much on the interaction of nature (genes) and nurture (environment) as it does on either one.

Your turn

F-ratios can be back-calculated from R^2. Calculate the *F*-ratios in the table in step 5.