Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 18.1 Logistic Regression (Dose - Response)

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11), Part IV (Ch 13, 14)
18 Binomial Response Variables
18.1 Logistic Regression (Dose-Response)
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

Ch18.xls

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) GLMl with a single explanatory variable.
ReCap (Ch 12,13,14,15) GLM with more than one explanatory variable
ReCap (Ch 16,17)
ReCap (Ch 18)
Binomial data are analyzed within the framework of the generalized linear model. The response variable is the odds, calculated from the proportion of cases *p*.

Today: Binomial response variables.
Logistic regression (dose-response analysis).

Wrap-up.
We analyze dose-response data with logistic regression, in which the response variable is the odds, and relation of odds to dose is exponential.

Introduction.

Laboratory tests of known carcinogens are conducted at relatively high doses to produce measurable rates of response in a small sample of animals. These results must then be extrapolated to lower doses that correspond to anticipated human exposure levels. To do this we need a realistic model with good estimates of parameters.

Aflatoxin is a known carcinogen. What is the dose-response relation for aflatoxin B_1?

N = number of experimental animals fed aflatoxin B_1,

Ntumor = number developing liver tumors

Dose = amount fed to animals (ppb)

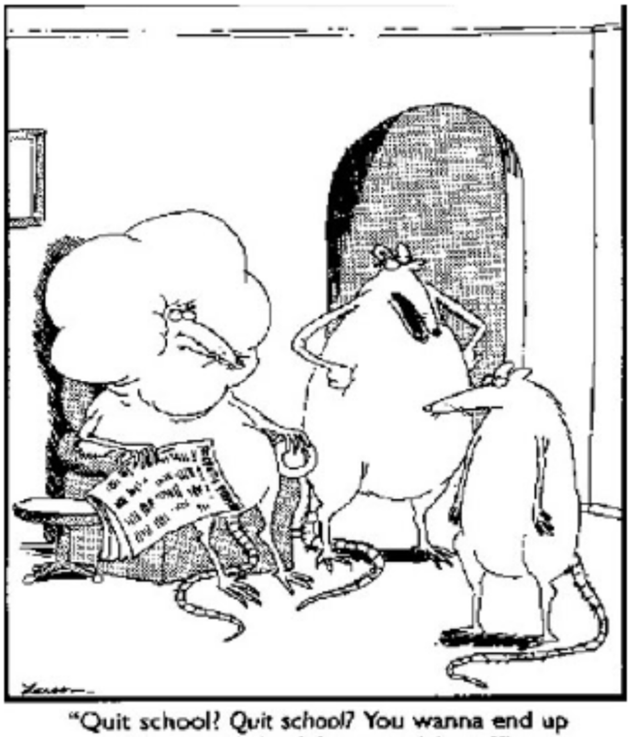
```
18   0   0
22   2   1
22   1   5
21   4  15
25  20  50
28  28 100

N Ntmr  Dose
```

Data from D.W. Gaylor (1987) . Linear nonparametric upper limits for low dose extrapolation. Pp. 63–66 in *American Statistical Association: Proceedings of the Biopharmaceutical Section*.

The data are binomial. Use of a normal error can be expected to produce biased parameter estimates. Regression of a binomial variable against an explanatory variable is called logistic regression, a special case of the generalized linear model.

Preliminary calculations-- proportions and odds.

Dose	Cases	Cases	Proportion	Odds	Variance
		w/ tumors	w/tumors		
ppb	N	Ntumor	p=Ntumor/N	p/(1-p)	Npq
0	18	0	0.00		
1	22	2	0.09	0.10	1.818
5	22	1	0.05	0.048	0.955
15	21	4	0.19	0.235	3.238
			0.50	1.00	5.0
50	25	20	0.80	4.0	4.0
100	28	28	1.00		



"Quit school? Quit school? You wanna end up like your father? A career lab rat?"

The odds of having a tumor increase with dosage. The increase is not a linear function of the dose.
The variance is maximum at p(1-p) = 0.5 · 0.5 = 0.25
In the example above the variance is 5 for 20 cases and p = 0.5.

1. Construct Model

N is number of experimental animals at each dosage.
Ntumor is number of animals that develop tumors.
p is the proportion of animals having tumors at each dosage. *Ntumor/N*
The odds of developing tumors are *p*/(1-*p*).

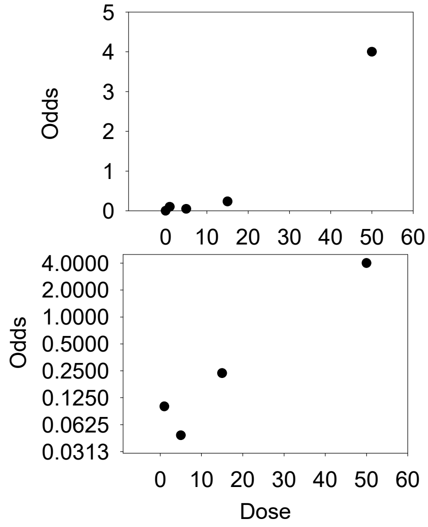
Verbal model.

The odds of developing tumors increase with dose.

Graphical model

The plot of odds versus dosage shows a curvilinear relation.

The plot of log odds versus dosage shows a linear relation. Dose has a multiplicative rather than additive effect on odds of developing tumors.



Response variable: odds of developing tumors
Explanatory variable: dose of aflatoxin B_1

$$Odds = \frac{p}{1-p}$$ Definition of odds

The formal model is written with multiplicative effect and binomial error.

Distribution	$N_{tumor} \sim Binomial(N, \pi)$	This is the error model,
	$\eta = \beta_o + \beta_{Dose} Dose$	This is the structural model,
Link	Odds = e^{η}	This is called a logit link.

We use a 3 line format to write the model.

- We need to state the error model.
- We no longer use the $Y = \sum \beta_x X + \varepsilon$ format because we no longer can use the unscaled residuals ε. We will be using scaled residuals when we use a non-normal error.
- We need to state the link between the response variable and structural model containing the explanatory variables.

1. Construct Model

For logistic regression we use a logit link.

- Odds are a natural way to interpret changes in risk.
- Odds ratios have the convenient property that we can invert them. If the odds of developing a tumor double with each increment in dose, then the odds are halved with each decrement in dose. This property does not apply to percentages.
- Graphical display (shown above) supports the idea of multiplicative effect on dose on odds of developing tumors.
- The logit link puts multiplicative effects on an additive scale, which is convenient for statistical analysis.
- The logit link avoids log transforming the response variable, which usually produce biased estimates of parameters.

2. Execute analysis.

Place data in model format:

The binomial response variable in two columns, cases and positives.

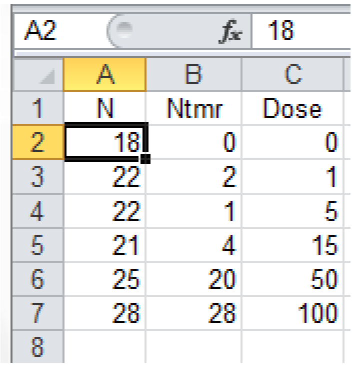
Column labelled *N*, with response variable, number of animals

Column labelled *Ntumor*, with response variable # of animals with tumors

Column labelled *Dose*, with explanatory variable Dose (in ppm)

```
Data Gaylor;
  Input N Ntumor Dose;
  Cards;
18   0   0
22   2   1
22   1   5
21   4  15
25  20  50
28  28 100
;
```

SAS data definition file



A2 fx 18

	A	B	C
1	N	Ntmr	Dose
2	18	0	0
3	22	2	1
4	22	1	5
5	21	4	15
6	25	20	50
7	28	28	100
8			

Spreadsheet format for input to graphics interface statistical packages –SPSS, Minitab, *etc.*

1. Construct Model

Code the GzLM model statement in the statistical package.

```
Proc Genmod;
  Model Ntumor/Ncase = Dose/
  Link=logit dist=binomial type1 type3;
```

SAS file

```
GaylorGzLM <- glm(Ntumor/N ~ Dose,
   family=binomial(link=logit), weights=N, data=Gaylor)
summary(GaylorGzLM)
plot(GaylorGzLM)
```

R command lines

2. Execute analysis.

```
MTB > BLogistic 'Ntumor' 'Ncase' = Dose;        as of 2002
SUBC>   ST;
SUBC>   Logit;
SUBC>   Brief 2.
```

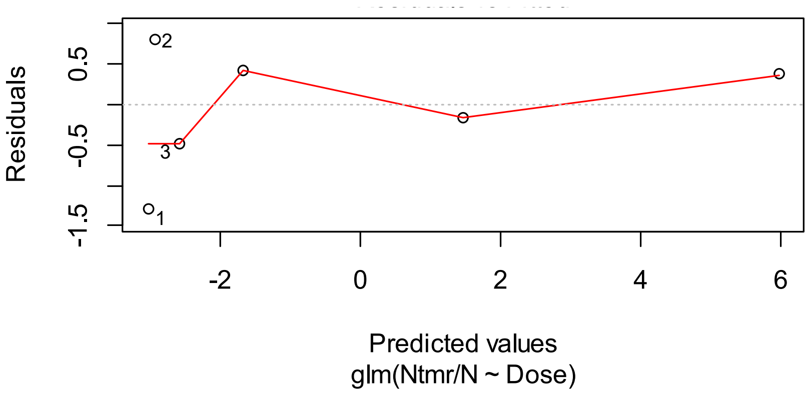
Minitab command lines

```
Click Stat
    Click Regression
       Click Binary Logistic Regression
           Click Success, place column of Ntumor,
           Click trials, place column of Ncase
           Click Model, place column with Dose
           Click Storage (optional)
           Click Pearson residuals, Event probability, ok
```

Minitab sequence to produce line commands

3. Use residuals to evaluate model.

There are several ways to scale the residuals. We will use the deviance residuals, which are the default in most cases.

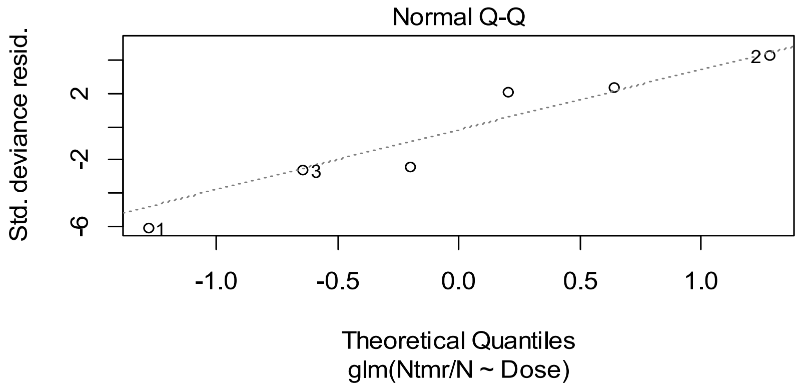


A. A straight line model is acceptable, no bowl or arch in residual plot.

B1. A binomial error structure was used to fit the model. This was expected to remove any change in vertical spread in the residual vs fit plot.

There are too few residuals to evaluate the homogeneity assumption.

B2. We expect the scaled residuals to be normally distributed. Confidence limits and p-values assume normal errors. The residuals are acceptably normal based on the quantile-quantile plot.



4. What is the evidence?

Estimate change in deviance.

Calculate *LR*

ANODEV Table Binomial model

Terms added sequentially, 1st to last

	Df	Resid.Dev	Δ df	Δ Deviance
NULL	5	116.5		
Dose	4	32.0	1	84.5

LR = $e^{84.5/2} = 2.2 \times 10^{18}$

Strong evidence (LR > 1000) that tumors increase with dose.

5. Mode of inference.

Frequentist? Yes. We take the population to be an infinite number of repetitions of the experimental protocol.

Decision theoretic? Yes. Type I error and Type II error have identifiable costs because aflatoxins regularly occur in improperly stored staple commodities including corn, cotton seed, millet, peanuts, rice, sesame seeds, sorghum, tree nuts, wheat, and a variety of spices. The United States Food and Drug Administration (FDA) action levels for aflatoxin present in food or feed is 20 to 300 ppb. The upper level is about 1/3 of the minimum dose in this study.

6. Population, sample, hypotheses.

Inference is to a population of many repeats of the same experiment with similar animals and the same suspected carcinogen, dosages, experimental protocol, and sample sizes. The sample was large enough to establish carcinogenicity at doses of 1 ppm or more.

7. Analysis of Deviance

ANODEV Table Binomial model

Terms added sequentially, 1st to last

	Df	Resid.Dev	Δ df	Δ Deviance
NULL	5	116.5		
Dose	4	32.0	1	84.5

Calculate Type I error from ΔDeviance using $\chi 2$ distribution with 1 df.

p $= 4 \times 10^{-11}$

8. **Recompute Type I error if sample small and assumptions not met.**
   Assumptions met.

9. **Statistical conclusion.**
   We reject the null hypothesis at a fixed Type I error = 5%.
   ΔDeviance = 4.7, p = 0.03 on 1 degree of freedom, from $\chi 2$ distribution

10. **Science conclusion.**
    Interpret the parameters

```
Logistic Regression Table
                                                     Odds       95% CI
Predictor        Coef    SE Coef        Z      P    Ratio    Lower    Upper
Constant      -3.0360     0.4823    -6.30 0.000
Dose          0.09009    0.01456     6.19 0.000     1.094     1.06     1.13
                                                                Minitab output
```

$e^{\beta_o} = e^{-3.0360} = 0.048$ Model odds for zero dose

$e^{\beta_{Dose}} = e^{0.09009} = 1.094$ Change in odds for each dose increment

The confidence limits for the odds ratio of 1.094 are narrow: 1.06 to 1.13
The confidence limit excludes the null hypothesis (OR = 1.00) that the odds do not change with dose.

With these estimates we can compute the expected odds at any dose within the range of the study.

$$e^{\beta_o + \beta_{Dose} \cdot 50} = e^{-3.036+0.09009 \cdot 50} = e^{1.469} = 4.34$$

The expected odds of developing a tumor at a dose of 50 ppm are 4.34 to 1.
The expected odds of developing a tumor at a dose of 33 ppm are 0.94 to 1.

We can compare the dose response relation (OR = 1.094) with other dose-response relations because we have a standard error on the log odds ratio
ln OR = 0.09009 ± 0.01456

Extra (beyond the curriculum)

In this analysis we used a binomial error, for which the response variable is the odds ratio. What if we had used the proportion p instead? Proportions are multiplied, not added. So we use the log link.

Distribution $N_{tumor} \sim Poisson(\lambda)$
$\eta = \beta_o + \beta_{Dose} Dose$
Link $Y = e^{\eta}$ This is called a log link.

We obtain different parameter estimates and a different ANODEV table.

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

Intercept   1   -1.4209     0.1893   -1.7918   -1.0500    56.37      <.0001
Dose        1    0.0142     0.0079   -0.0013    0.0297     3.23      0.0722
Scale       0    1.0000     0.0000    1.0000    1.0000
```

SAS output

We obtain 95% confidence limits that include zero (no dose-response relation).

We obtain less convincing evidence:

Poisson $LR = e^{(3.23/2)} = 5.03$

Binomial $LR = e^{84.5/2} = 2.2 \times 10^{18}$

Extra: Your turn.

Before the advent of exponentially increasing computing power in the late 20th century, and the rapid advances in software in the present century, analysts were forced to rely on a crude approximation for binomial data – a normal error structure with proportions as the response variable. Run the analysis of the tumor data with a normal error, identity link, and proportion of animals with tumors as the response variable.

Compare the analysis to the binomial analysis with respect to weight of evidence (the likelihood ratio), width of 95% confidence limits, and whether the 95% limits exclude zero (no dose-response effect).