Outline for today

- What is a generalized linear model
- Linear predictors and link functions
- Example: fit a constant (the proportion)
- Analysis of deviance table
- Example: fit dose-response data using logistic regression
- Example: fit count data using a log-linear model
- Advantages and assumptions of `glm`
- Quasi-likelihood models when there is excessive variance

Review: what is a linear model

A model of the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

- *Y* is the response variable
- The *X* 's are the explanatory variables
- The β 's are the parameters of the linear equation
- The errors are normally distributed with equal variance at all values of the *X* variables.
- Uses least squares to fit model to data, estimate parameters
- `lm` in R

Review: fitting a linear model in R

Use the `lm` package in R

Simplest linear model: fit a constant (the mean)

```
z <- lm(y ~ 1)
```

Linear regression

```
z <- lm(y ~ x)
```

Single factor Anova

```
z <- lm(y ~ A)
```

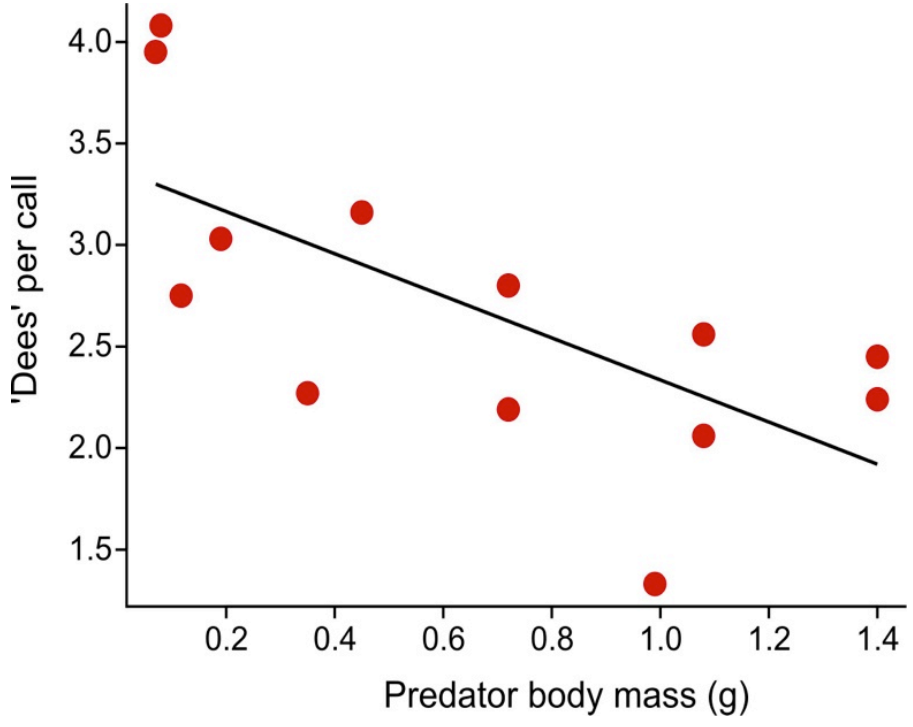
Review: what is a linear model

Eg: linear regression: $Y = \beta_0 + \beta_1 X$

The predicted *Y*, symbolized here by μ, is modeled as

$$\mu = \beta_0 + \beta_1 X$$

The part to the right of "=" is the linear predictor



What is a generalized linear model

A model whose predicted values are of the form

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

- The model still include a *linear predictor* (to right of "=")
- $g(\mu)$ is called the "link function"
- Wide diversity of link functions accommodated
- Non-normal distributions of errors OK (specified by link function)
- Unequal error variances OK (specified by link function)
- Uses maximum likelihood to estimate parameters
- Uses log-likelihood ratio tests to test parameters
- `glm` in R

The two most common link functions

1) Natural log (i.e., base *e*)

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

Usually used to model count data (e.g., number of mates, etc)

poisson distribution

$\eta = \log(\mu)$ is the link function.

The inverse function is $\mu = e^{\eta}$

The two most common link functions

2) Logistic or logit

$$\log \frac{\mu}{1-\mu} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

Used to model binary data (e.g., survived vs died)

The link function $\eta = \log \frac{\mu}{1-\mu}$ is also known as the log-odds

binomial distribution

The inverse function is $\mu = \frac{e^{\eta}}{1+e^{\eta}}$

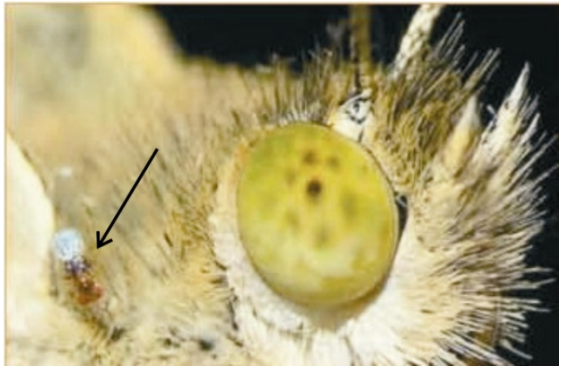
Example 1: Fit a constant to 0-1 data (estimate a proportion)

This example was used previously in Likelihood lecture:

The wasp, *Trichogramma brassicae*, rides on female cabbage white butterflies, *Pieris brassicae*. When a butterfly lays her eggs on a cabbage, the wasp climbs down and parasitizes the freshly laid eggs.

Fatouros et al. (2005) carried out trials to determine whether the wasps can distinguish mated female butterflies from unmated females. In each trial a single wasp was presented with two female cabbage white butterflies, one a virgin female, the other recently mated.

Y = 23 of 32 wasps tested chose the mated female. What is the proportion p of wasps in the population choosing the mated female?

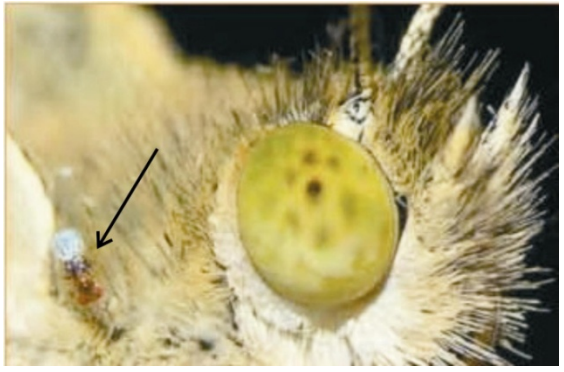


Number of wasps choosing the mated female fits a binomial distribution

Under random sampling, the number of "successes" in *n* trials has a binomial distribution, with *p* being the probability of "success" in any one trial.

To model these data, let "success" be "wasp chose mated butterfly"

Y = 23 successes
n = 32 trials



What is *p*?

Data are :
1 1 1 0 1 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 0 1 1 1 0 0 1 1

Use `glm` to fit a constant, and so obtain the ML estimate of *p*

The data are binary. Each wasp has a measurement of 1 or 0 ("success" and "failure"): 1 1 1 0 1 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 0 1 1 1 0 0 1 1

To begin, fit a model with only a constant. Use the link function appropriate for binary data:

$$\log \frac{\mu}{1-\mu} = \beta$$

μ here refers to the population proportion (p) but let's stick with μ here to use consistent notation for generalized linear models.

Fitting will yield the estimate, $\hat{\beta}$.
The estimate of proportion is then obtained using the inverse function:

$$\hat{\mu} = \frac{e^{\hat{\beta}}}{1+e^{\hat{\beta}}}$$

Use `glm` to fit a constant, and so obtain the ML estimate of *p*

```
z <- glm(choice ~ 1, family = binomial(link="logit"))
```

Formula structure is the same as when fitting a constant using `lm`.

`family` specifies the error distribution and link function.

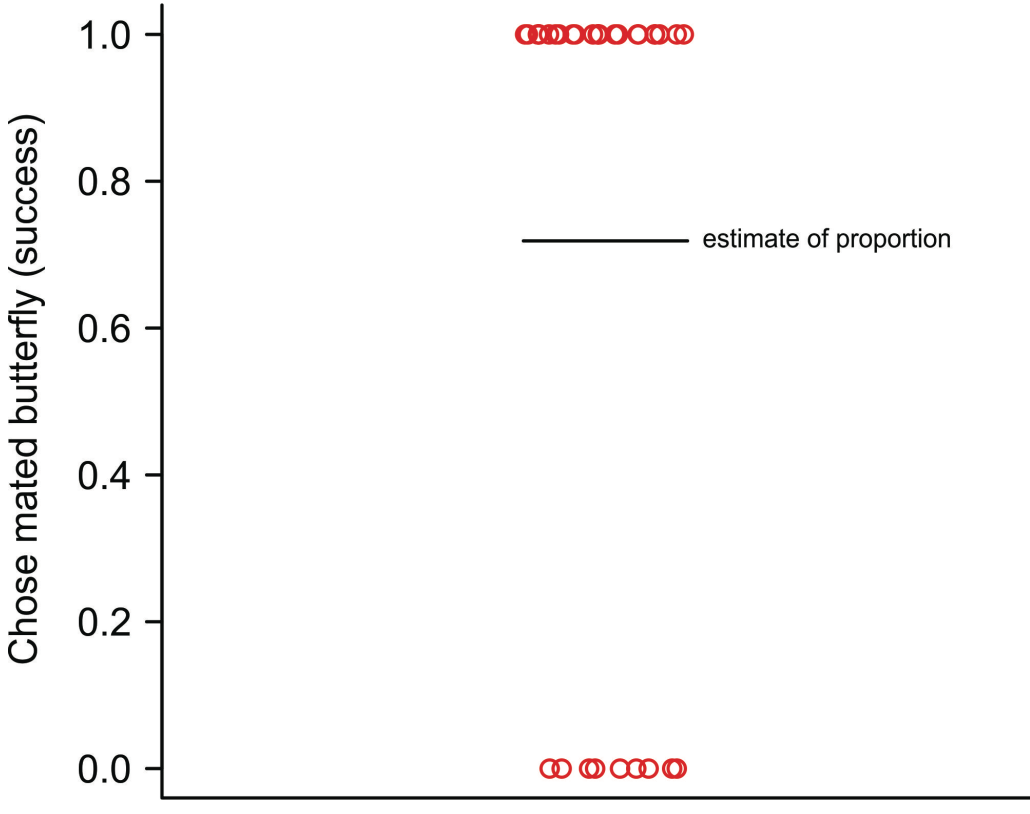
Use `glm` to fit a constant, and so obtain the ML estimate of *p*

```
summary(z)
 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.9383     0.3932   2.386    0.017 *
```

0.9383 is the estimate of β (the constant on the **logit** scale). Convert back to ordinary scale (plug into inverse equation) to get estimate of population proportion:

$$\hat{\mu} = \frac{e^{\hat{\beta}}}{1 + e^{\hat{\beta}}} = \frac{e^{0.9383}}{1 + e^{0.9383}} = 0.719$$

This is the ML estimate of the population proportion. Does it look familiar?



Use `summary()` for estimation, not hypothesis testing

```
summary(z)

 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
 (Intercept)    0.9383     0.3932   2.386    0.017 *
```

anova command carries out log liklihood ratio test

The *z*-value (Wald statistic) and *P*-value test the null hypothesis that $\beta = 0$. This is the same as a test of the null hypothesis that the true (population) proportion $\mu = 0.5$, because

$$\frac{e^0}{1 + e^0} = 0.5$$

Agresti (2002, *Categorical data analysis*, 2nd ed., Wiley) says that for small to moderate sample size, the Wald test is less reliable than the log-likelihood ratio test. So don't use it.

Use `summary()` for estimation, not hypothesis testing

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.9383     0.3932   2.386    0.017 *
```

```
# 95% confidence limits:
CI <- confint(z)       # on logit scale
exp(CI)/(1 + exp(CI)) # inverse logit

    2.5 %    97.5 %
0.5501812 0.8535933
```

$0.550 \leq p \leq 0.853$ is the same result we obtained last week (used more decimal places this week).

Use `anova()` to test a hypothesis about a proportion

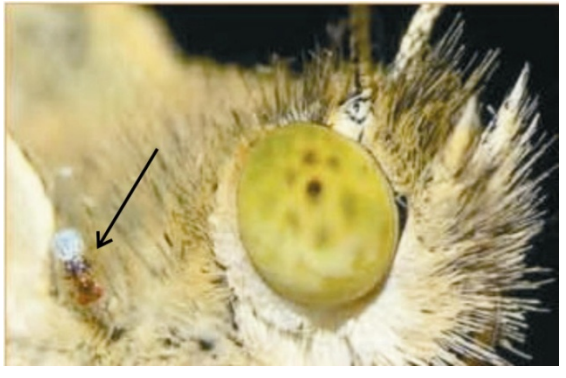
We calculated the log-likelihood ratio test for these data by hand in the likelihood lecture. Here we'll use `glm` to accomplish the same task.

"Full" model (β estimated from data):

```
z1 <- glm(y ~ 1, family = binomial(link="logit"))
```

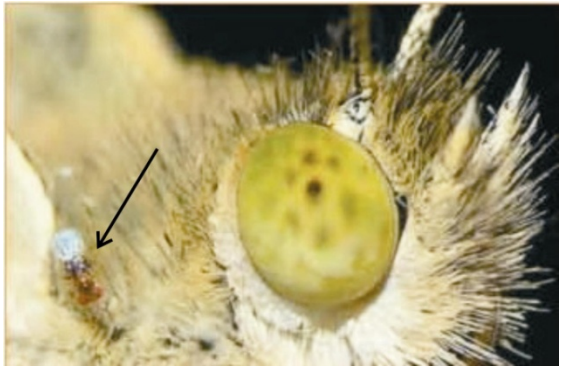
"Reduced" model (β set to 0 by removing intercept from model):

```
z0 <- glm(y ~ -1, family = binomial(link="logit"))
```



Use `anova()` to test a hypothesis about a proportion

```
anova(z0, z1, test  = "Chi") # Analysis of deviance
```



```
 Model 1: y ~ -1    # Reduced model
 Model 2: y ~ 1     # Full model
```

Analysis of deviance table:

```
   Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         32     44.361
2         31     38.024  1    6.337   0.01182 *
```

The deviance is the log-likelihood ratio statistic (*G*-statistic).
Deviance (*G*) has an approximate χ^2 distribution under the null hypothesis.

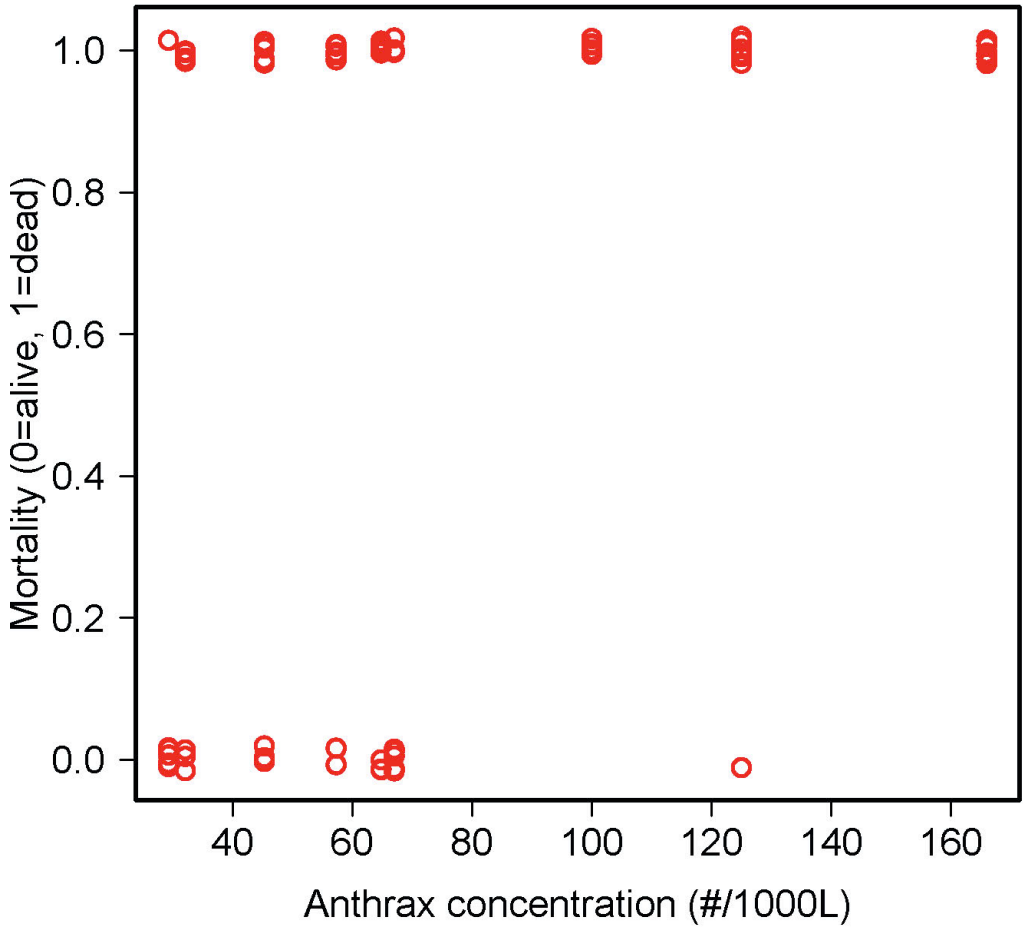
This is the same result we obtained doing the *G*-test by hand last week.

Example 2: Logistic regression

One of the most common uses of generalized linear models.

Goal is to model the relationship between a proportion and an explanatory variable

Data: 72 rhesus monkeys (*Macacus rhesus*) exposed for 1 minute to aerosolized preparations of anthrax (*Bacillus anthracis*). Want to estimate the relationship between dose and probability of death.

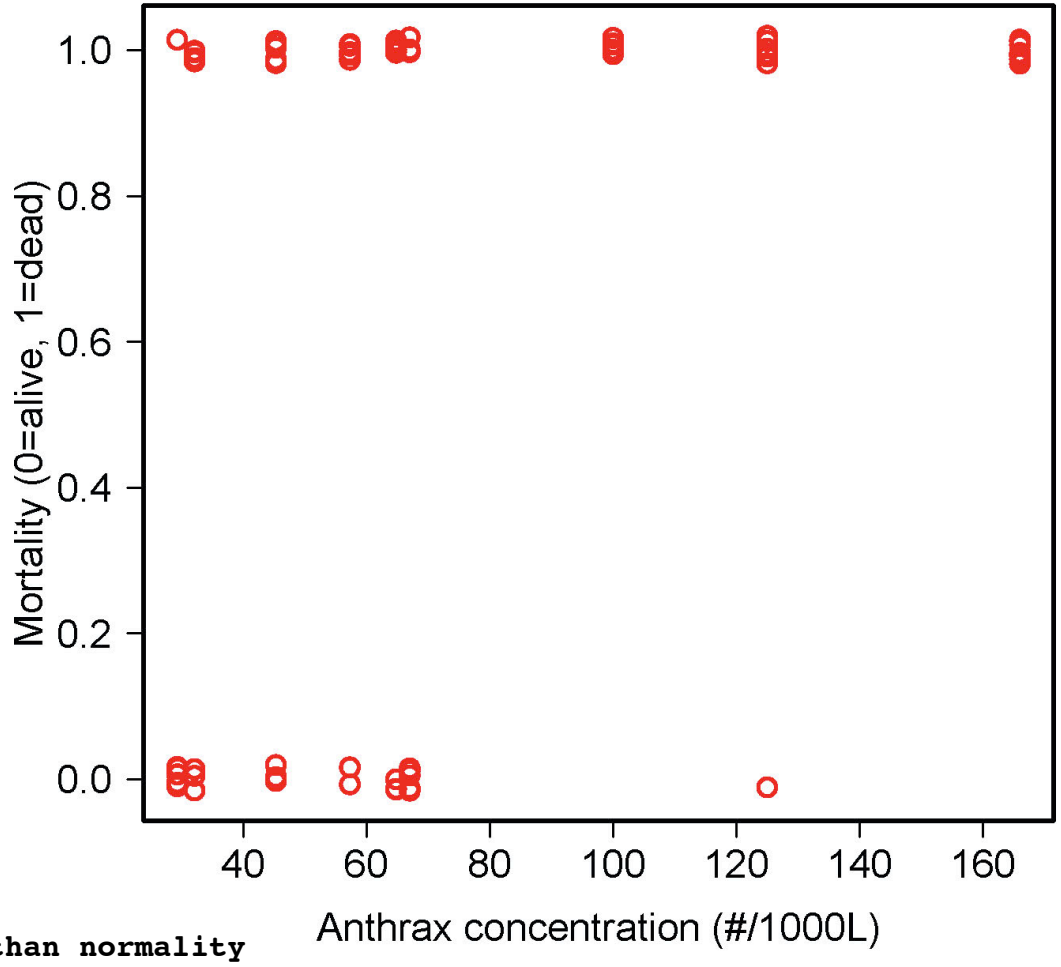


Logistic regression

Measurements of individuals are 1 (dead) or 0 (alive)

Ordinary linear regression model not appropriate because

- For each *X* the *Y* observations are binary, not normal
- For every *X* the variance of *Y* is not constant
- A linear relationship is not bounded between 0 and 1
- 0, 1 data can't simply be transformed



violation of homogeneity of variance is more serious than normality

The generalized linear model

$$g(\mu) = \beta_0 + \beta_1 X$$

μ is the probability of death, which depends on concentration *X*.

$g(\mu)$ is the link function.

Linear predictor (right side of equation) is like an ordinary linear regression, with intercept β_0 and slope β_1

Logistic regression uses the logit link function

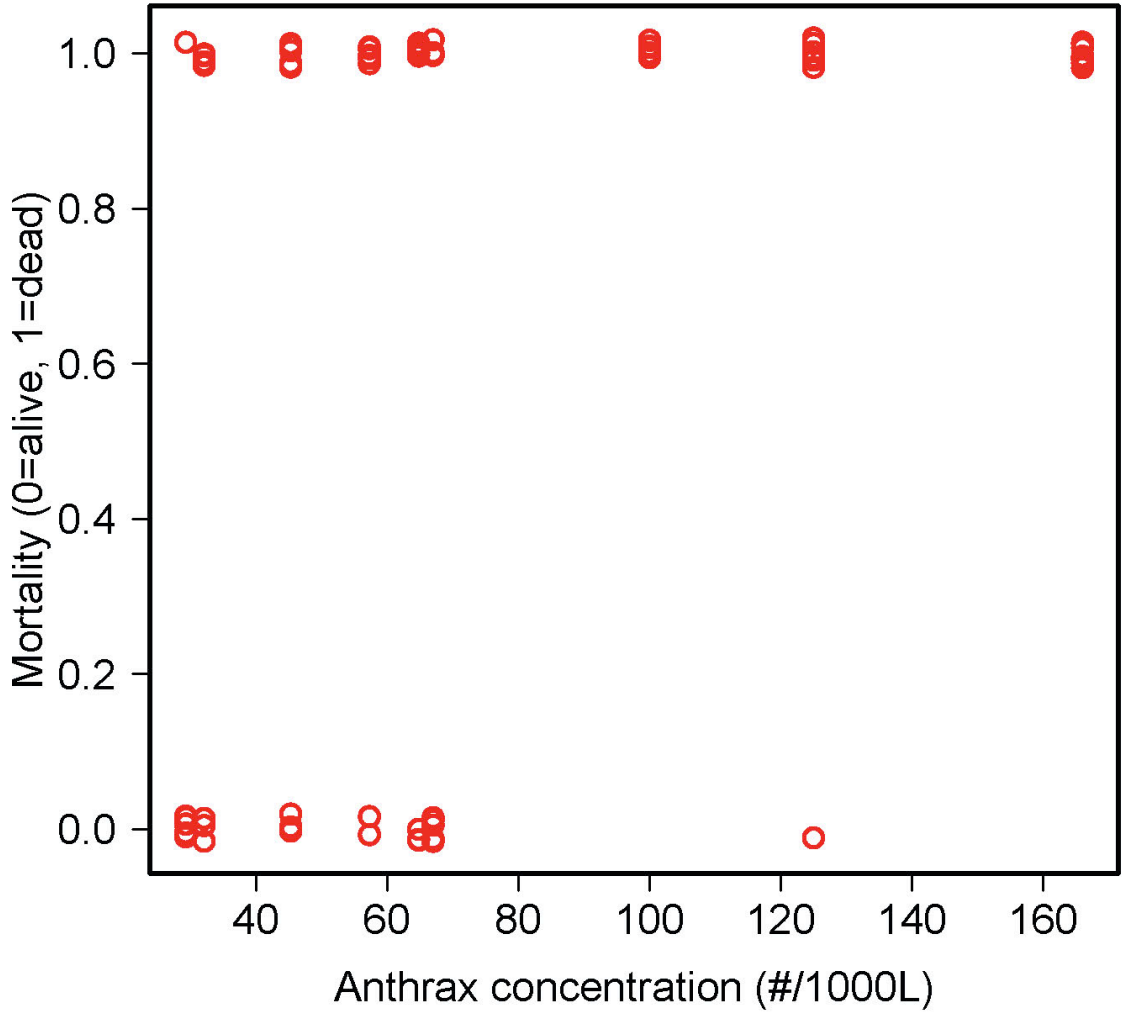
```
z <- glm(mortality ~ concentration,
              family = binomial(link = "logit"))
```

The generalized linear model

$$g(\mu) = \beta_0 + \beta_1 X$$

`glm` uses maximum likelihood: the method finds those values of β_0 and β_1 for which the data have maximum probability of occurring. These are the maximum likelihood estimates.

No formula for the solution. `glm` uses an iterative procedure to find the maximum likelihood estimates.



Use `summary()` for estimation

```
z <- glm(mortality ~ concentration,
             family = binomial(link = "logit"))

summary(z)
                Estimate Std. Error z value Pr(>|z|)
 (Intercept)    -1.74452    0.69206  -2.521  0.01171 *
 concentration   0.03643    0.01119   3.255  0.00113 **

 Number of Fisher Scoring iterations: 5
```

Numbers in red are the estimates of β_0 and β_1 (intercept and slope) which predict $\log(\mu / 1 - \mu)$.

Number of Fisher Scoring iterations refers to the number of iterations used before the algorithm used by glm converged on the maximum likelihood solution.

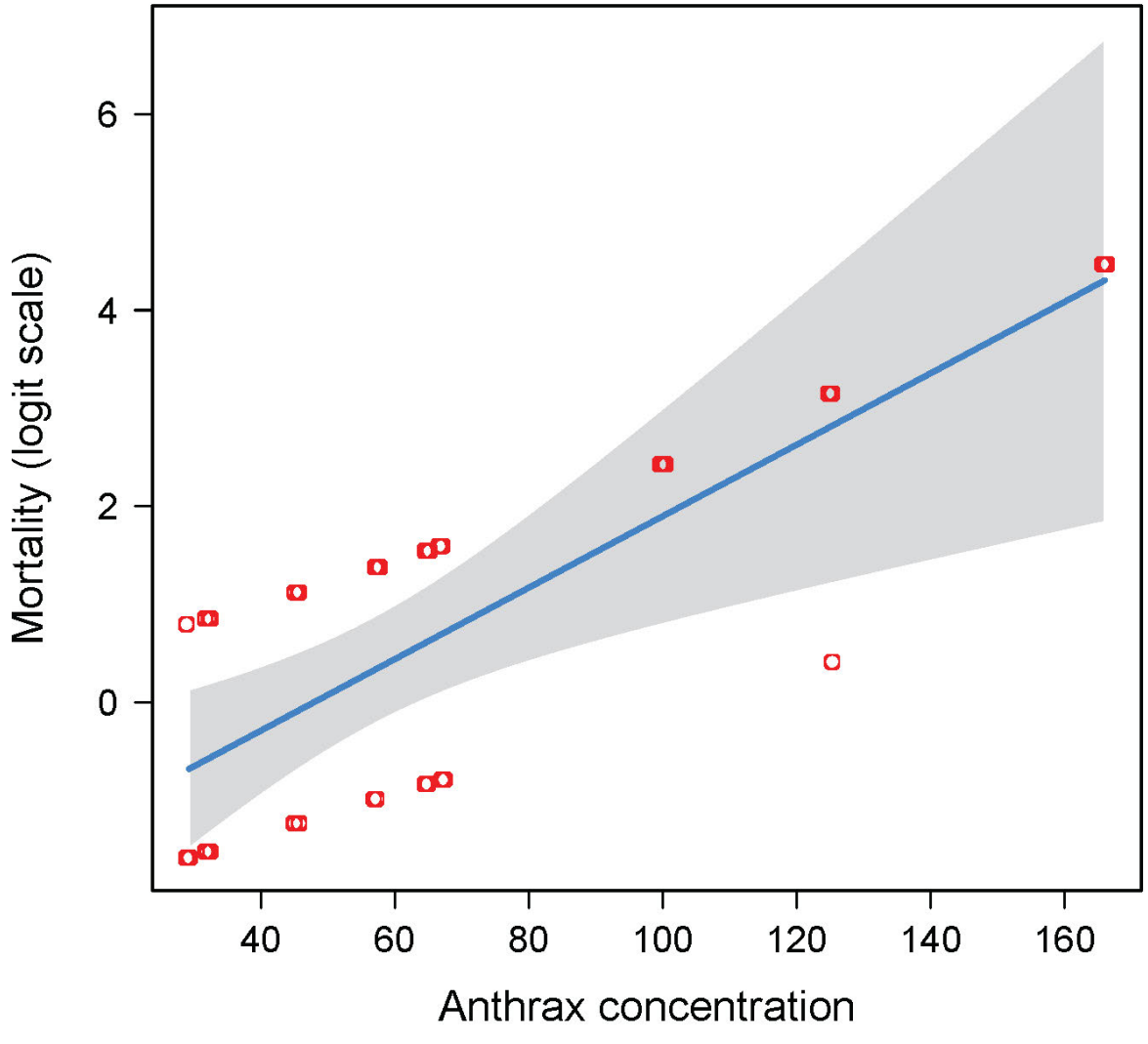
The generalized linear model

Use `predict(z)` to obtain predicted values on the logit scale

$\hat{\mu} = -1.7445 + 0.03643X$

`visreg(z)` uses `predict` to plot predicted values, with confidence limits on logit scale.

See that the function is a line. The points on this scale are not the logit-transformed data. R creates "working" values based on a transformation of the residuals calculated on the original scale.

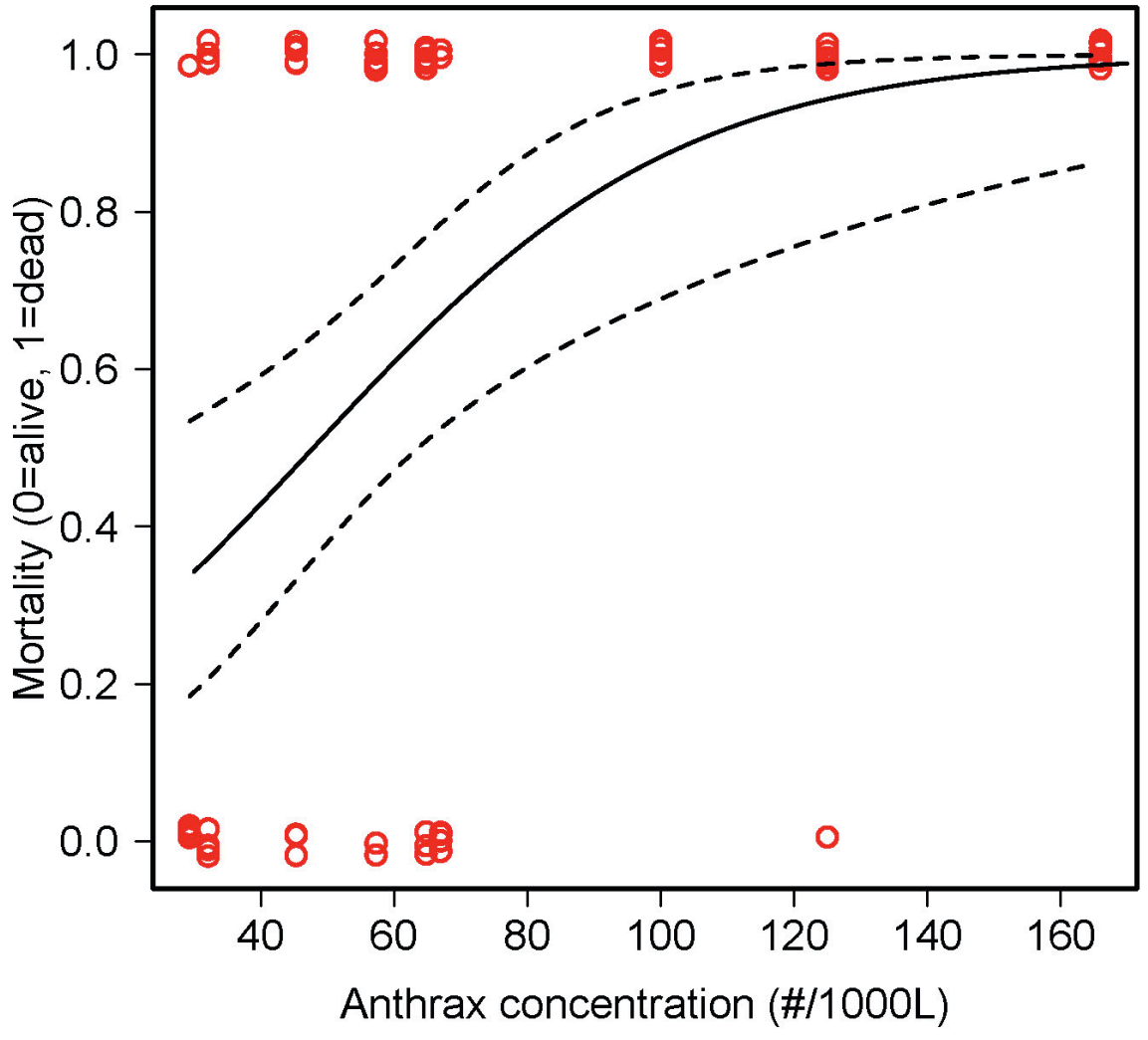


The generalized linear model

Use `fitted(z)` to obtain predicted values on the original scale

$$\hat{\mu} = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}}$$

Can calculate (approximate) confidence bands as in ordinary regression.



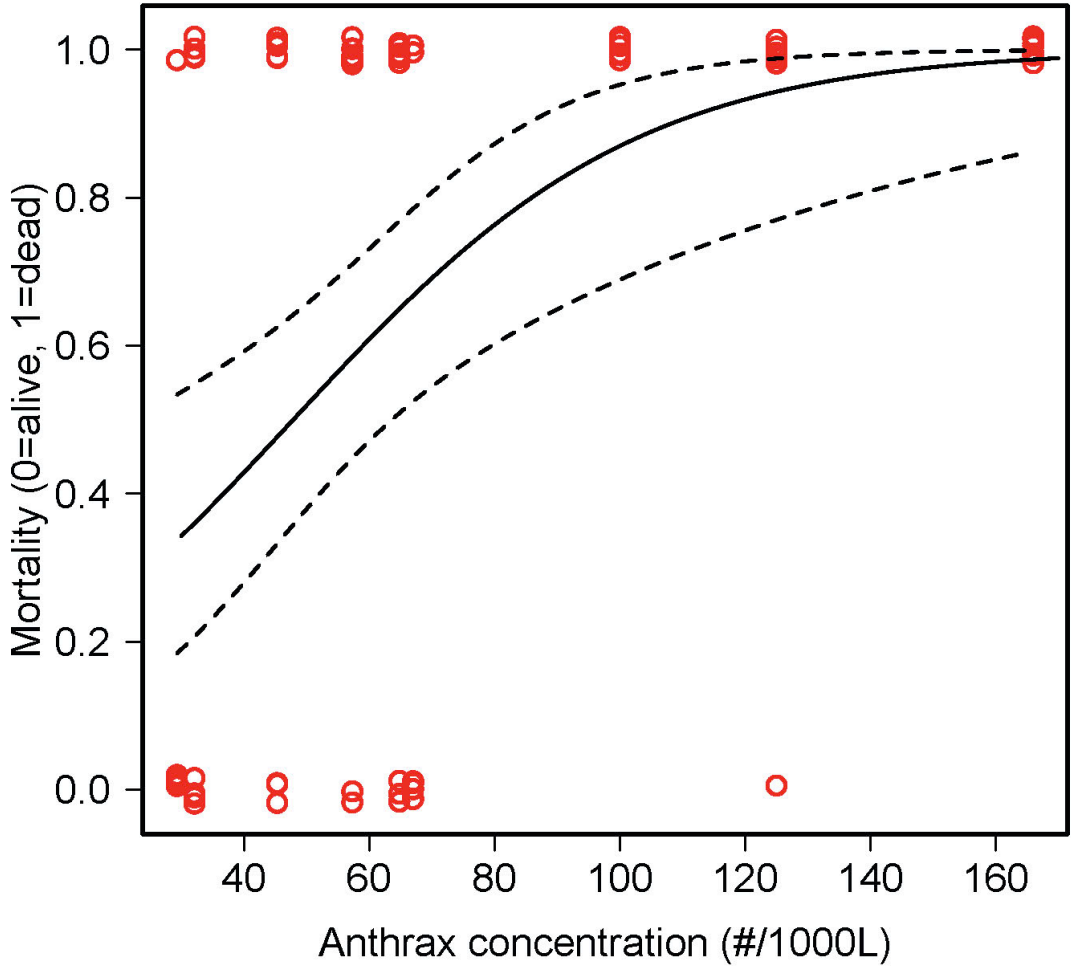
The generalized linear model

$$\mathrm{LD}_{50} = -\frac{\text{intercept}}{\text{slope}} = -\frac{0.03643}{-1.7445} = 47.88$$

The parameter estimates from the model fit can be used to estimate LD_{50}, the estimated concentration at which 50% of individuals are expected to die.

```
library(MASS)
dose.p(z)

                Dose         SE
p = 0.5: 47.8805 8.168823
```

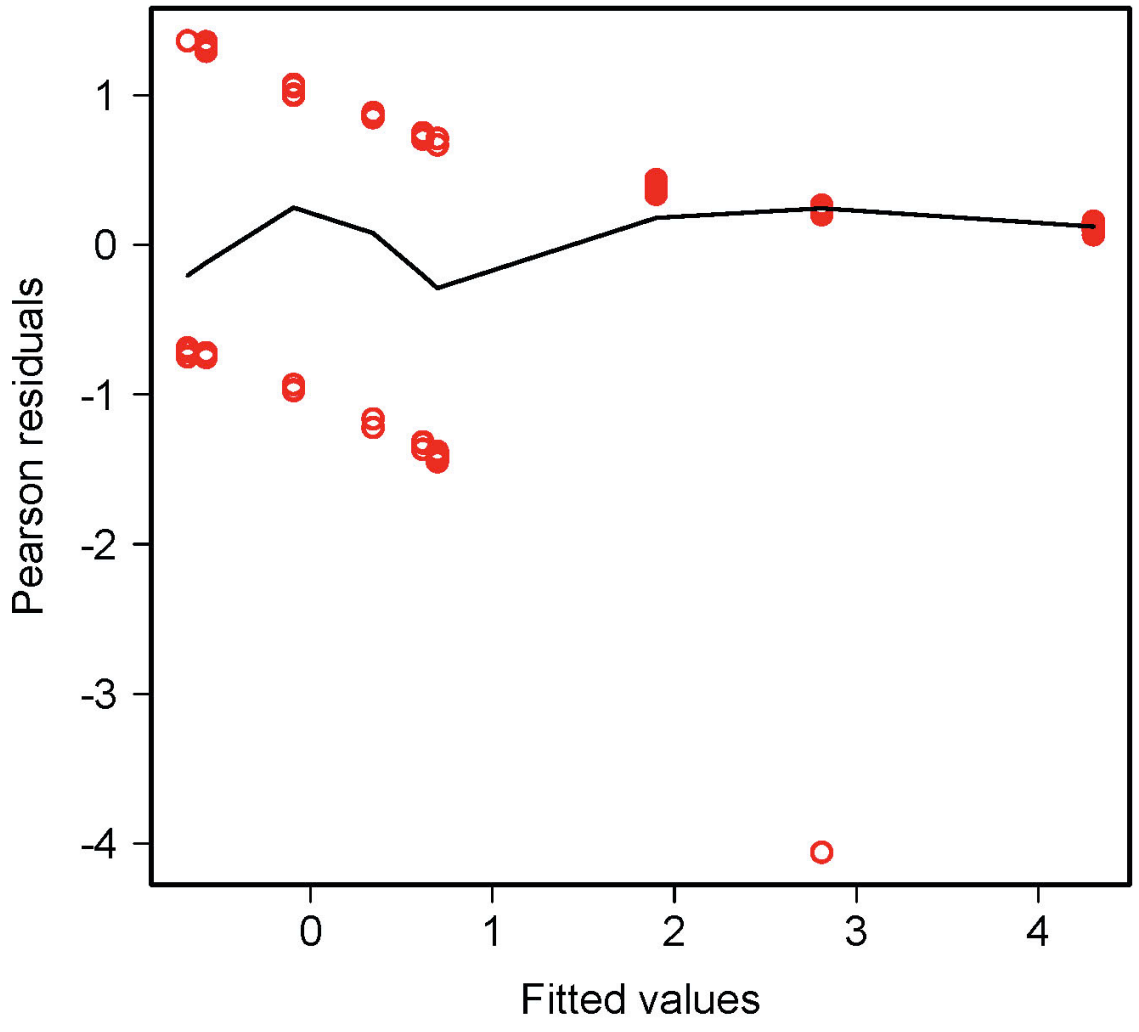


Residual plots in `glm`

Pearson residuals.
`plot(z)`

These are a rescaled version of the real "working" residuals.

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{\mu}_i)}}$$



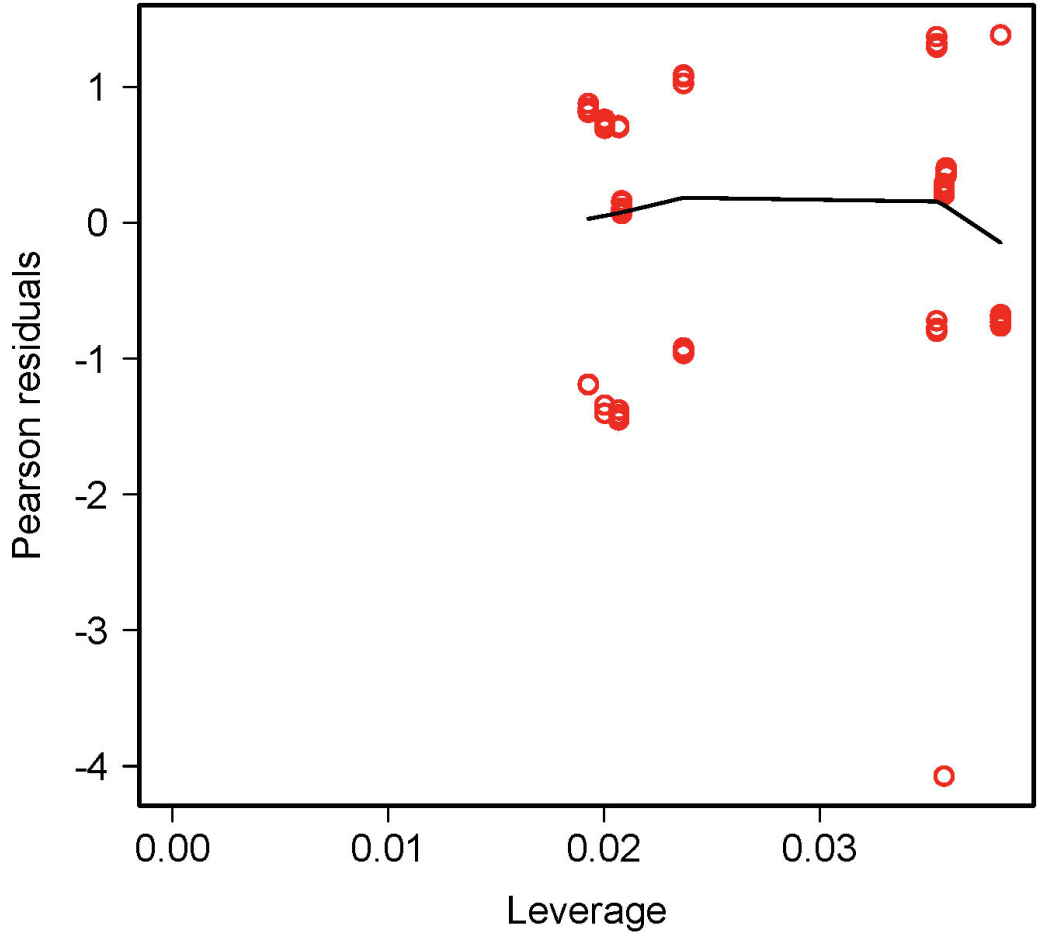
Leverage plot in `glm`

"Leverage" estimates the effect that each point has on the model parameter estimates. Obtained as

```
plot(z)
```

Leverage can range between 1/n (here, 0.013) and 1.

The graph here shows that even though one of the data points has a large residual, it does not have large leverage.



Advantages of generalized linear models

- More flexible than simply transforming variables. (A given transformation of the raw data may not accomplish both linearity and homogeneity of variance.)
- Yields more familiar measures of the response variable than data transformations. (E.g., how to interpret arcsine square root).
- Avoids the problems associated with transforming 0's and 1's. (For example, the logit transformation of 0 or 1 can't be computed.)
- Retains the same analysis framework as linear models.

Assumptions of generalized linear models

- Statistical independence of data points.
- Correct specification of the link function for the data.
- The variances of the residuals correspond to that expected from the link function.
- Later, I will show a method for dealing with excessive variance.

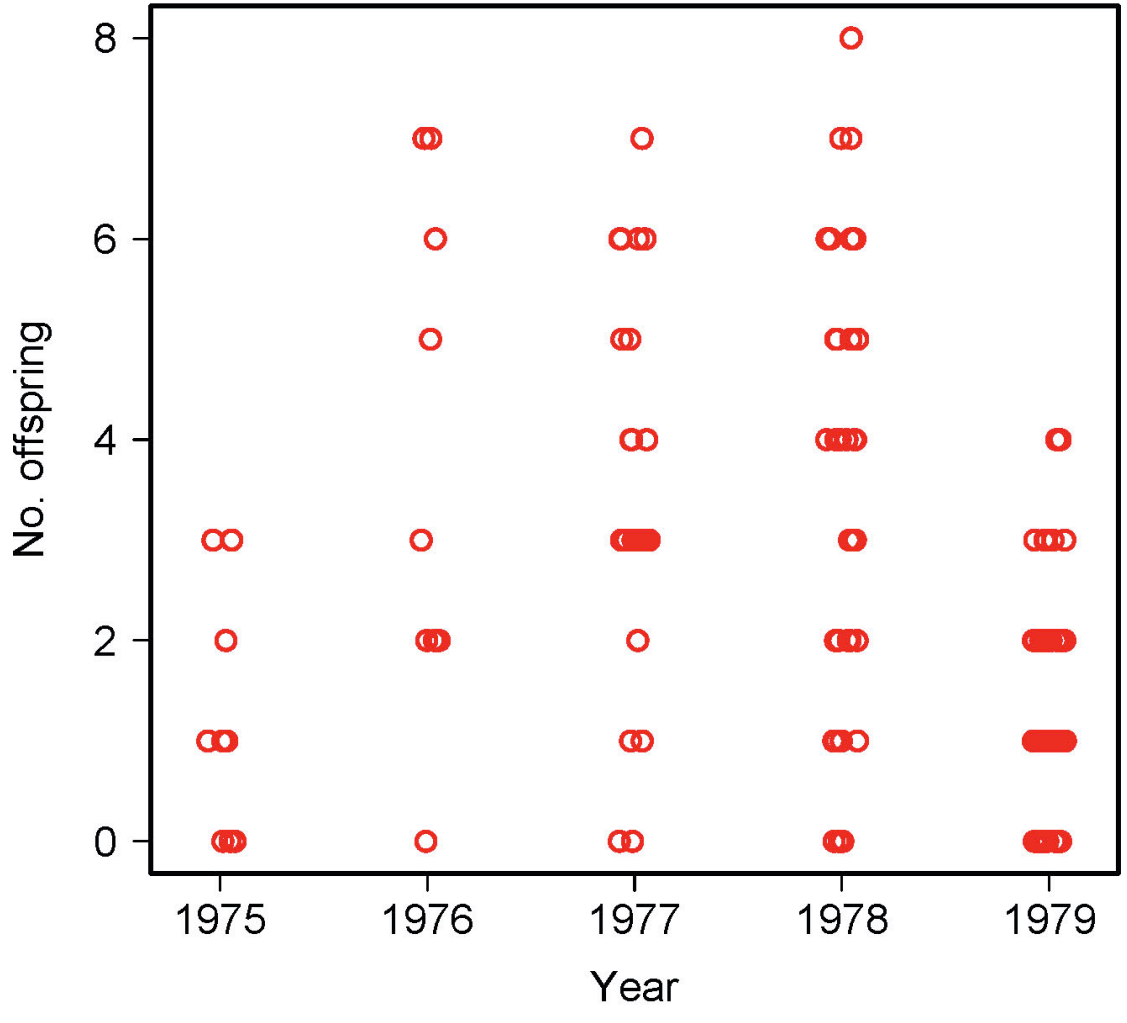
Example 3: Analyzing count data with log-linear regression

Estimate mean number of offspring fledged by female song sparrows on Mandarte Island, BC. Problem is similar to ANOVA, but ANOVA assumptions are not met.

http://commons.wikimedia.org/wiki/File:Song_Sparrow-27527-2.jpg

Data are discrete counts.

Variance increases with mean.



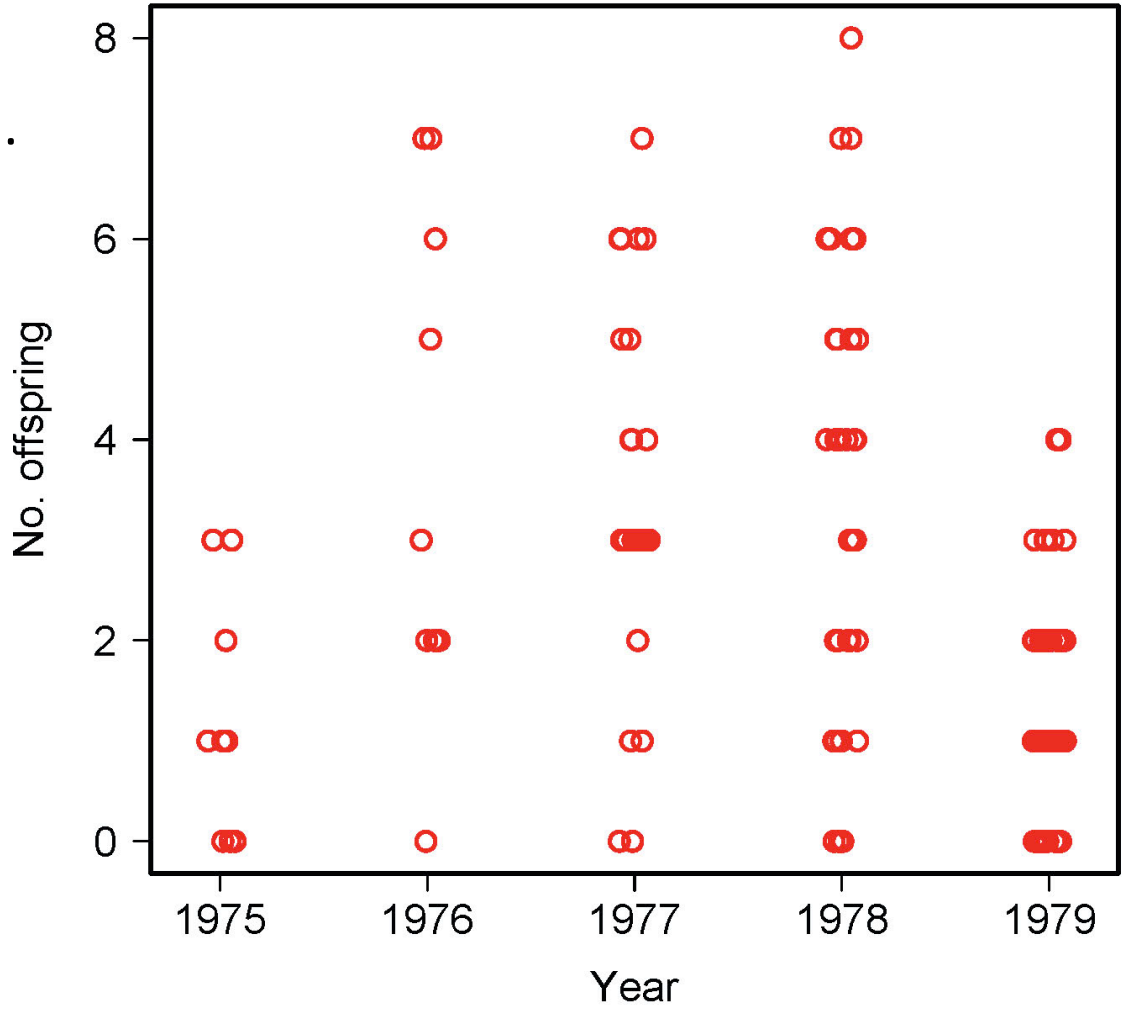
Example 3: Analyzing count data with log-linear regression

Estimate mean number of offspring fledged by female song sparrows on Mandarte Island, BC.

Data are discrete counts.
Variance tends to go up with mean.

Two solutions:
1. Transform data $X' = \log(X + 1)$

2. Generalized linear model.
Poisson distribution might be appropriate for error distribution.
So use log link function.



The generalized linear model

Log-linear regression (a.k.a. Poisson regression) uses the log link function

$$\eta = \log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

η is the response variable on the log scale (here, mean of each group on log scale)

Year is a categorical variable. So is analogous to single factor ANOVA

Categorical variables are modeled in R using “dummy” variables, same as with `lm`.

Use `summary()` for estimation

```
z <- glm(noffspring ~ year, family=poisson(link="log"))
summary(z)
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  0.24116    0.26726   0.902 0.366872
 year1976     1.03977    0.31497   3.301 0.000963 ***
 year1977     0.96665    0.28796   3.357 0.000788 ***
 year1978     0.97700    0.28013   3.488 0.000487 ***
 year1979    -0.03572    0.29277  -0.122 0.902898

(Dispersion parameter for poisson family taken to be 1)
```

Numbers in red are the parameter estimates on the log scale.
Intercept refers to mean of the first group (1975) and the rest of the coefficients are differences between each given group (year) and the first group.

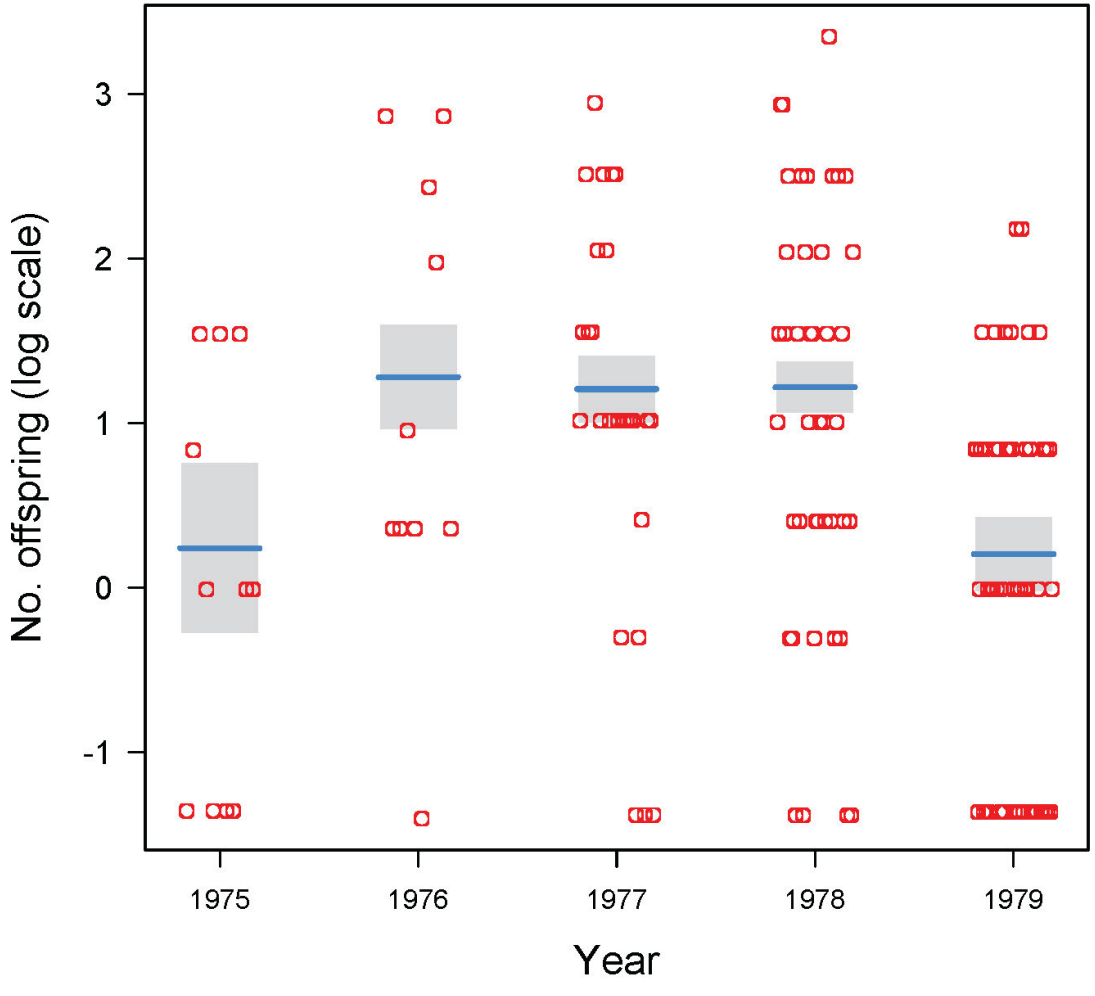
"Dispersion parameter" of 1 assumes that variance = 1(mean).

The generalized linear model

Predicted values on the transformed scale (here, log):
`predict(z)`

`visreg(z)` uses `predict` to plot the predicted values, with confidence limits on the transformed scale.

See that the "data points" on this scale are not just the transformed data (we can't take log of 0). R creates "working" values based on a transformation of the residuals calculated on the original scale.

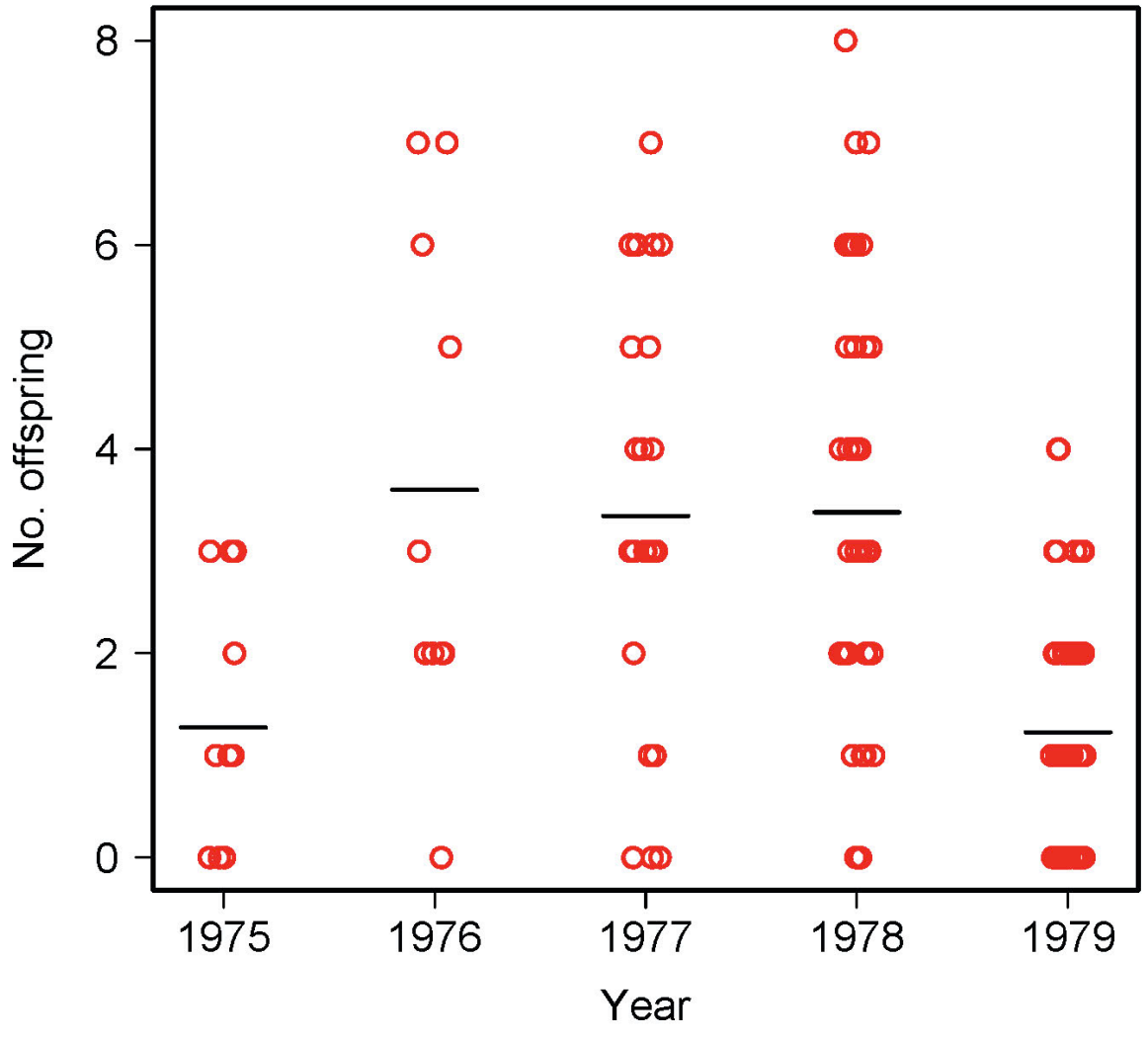


Predicted values on the original scale:
`fitted.values(z)`

$$\hat{\mu} = e^{\hat{\eta}}$$

I have plotted them with the original data.

Note that the fitted values aren't identical to the means of the original data. Fitted values are a transformation of means estimated on the log scale ("geometric means").



Use `anova()` to test hypotheses

Analysis of deviance table gives log-likelihood ratio test of the null hypothesis that there is no differences among years in mean number of offspring.

```
anova(z, test="Chisq")

 Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                   145    288.656
year   4   75.575       141    213.081 1.506e-15 ***
```

As with `lm`, default method is sequential fitting of terms ("Type 1 sums of squares")

Evaluating assumptions of the `glm` fit

Do the variances of the residuals correspond to that expected from the link function?

The $\log$ link function assumes that the *Y* values are Poisson distributed at each *X.*

A central property of the Poisson distribution is that the variance and mean are equal (`glm` dispersion parameter = 1).

(Similarly, when analyzing binary data, the logit link function also assumes a strict mean-variance relationship, specified by binomial distribution, enforced when dispersion parameter = 1.)

Evaluating assumptions of the `glm` fit

A central property of the Poisson distribution is that the variance and mean are equal (`glm` dispersion parameter = 1).

Let's check the sparrow data:

```
tapply(noffspring, year, mean)
tapply(noffspring, year, var)
```

```
     1975     1976     1977     1978     1979
 1.272727 3.600000 3.346154 3.380952 1.228070 # mean
 1.618182 6.044444 3.835385 4.680604 1.322055 # variance
```

Variances slightly, but not alarmingly, larger than means.

Modeling excessive variance

Finding excessive variance (“overdispersion”) is common when analyzing count data. Excessive variance occurs because other variables not included in the model affect the response variable.

In the workshop we will analyze an example where the problem is more severe than in the case of the song sparrow data here.

Modeling excessive variance

Excessive variance can be accommodated in `glm` by using a different link function, one that incorporates a dispersion parameter (which must also be estimated). If the estimated dispersion parameter is >> 1 then there is likely excessive variance.

The `glm` procedure to accomplish over (or under) dispersion uses the relationship between mean and variance rather than an explicit probability distribution for the data. In the case of count data,

variance = (dispersion parameter) * (mean)

Method generates “quasi-likelihood” estimates that behave like maximum likelihood estimates.

Modeling excessive variance

Lets try it with the song sparrow data

```
z <- glm(noffspring ~ year, family = quasipoisson)

summary(z)
            Estimate Std. Error t value Pr(>|t|)
Intercept)   0.24116    0.29649   0.813  0.41736
year1976     1.03977    0.34942   2.976  0.00344 **
year1977     0.96665    0.31946   3.026  0.00295 **
year1978     0.97700    0.31076   3.144  0.00203 **
year1979    -0.03572    0.32479  -0.110  0.91259

Dispersion parameter for quasipoisson family taken to be
1.230689
```

The point estimates are identical with those obtained using `family=poisson` instead, but the standard errors (and resulting confidence intervals) are wider.

Modeling excessive variance

```
z <- glm(noffspring ~ year, family = quasipoisson)

summary(z)
            Estimate Std. Error t value Pr(>|t|)
Intercept)   0.24116    0.29649   0.813  0.41736
year1976     1.03977    0.34942   2.976  0.00344 **
year1977     0.96665    0.31946   3.026  0.00295 **
year1978     0.97700    0.31076   3.144  0.00203 **
year1979    -0.03572    0.32479  -0.110  0.91259

Dispersion parameter for quasipoisson family taken to be
1.230689
```

The dispersion parameter is reasonably close to 1 for these data. But typically it is much larger than 1 for count data, so use `family = quasipoisson`.

Other uses of generalized linear models

We have used `glm` to model binary frequency data, and count data.

The method is commonly used to model r x c (and higher order) contingency tables, in which cell counts depend on two (or more) categorical variables each of which may have more than two categories or groups (see R tips “Fit model” page).

`glm` can handle data having other probability distributions than the ones used in my examples, including exponential and gamma distributions.

Discussion paper for next week:

Whittingham et al (2006) Why do we still use stepwise modelling?

Download from "**assignments**" tab on course web site.

Presenters: Amelia and Sandra

Moderators: Tianyi and ... (need one additional person!)