Non-normal regression models

Recall that in regression we started with a normal model:

\[ Y \sim \text{normal}(\mu, \sigma^2) \]

We allowed a parameter in the model to depend linearly on the value of covariates:

\[ \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \]

In fact, one can do this sort of thing with other statistical distributions.

**Logistic regression.** Suppose the response variable \( Y \) was binary: success/failure, survive/death, etc.:

\[ Y \sim \text{binomial}(1, p) \]

**ex.** (Griffith et al. 1989 *Science* 245:477-480) Pheasant release programs by state game agencies: many releases in different locales all over the country. Define

\[ Y_i = \begin{cases} 
0, & \text{release } i \text{ fails to establish population} \\
1, & \text{release } i \text{ succeeds} 
\end{cases} \]

Assume \( Y_i \sim \text{binomial}(1, p) \), except that one might expect \( p \) to vary from release to release, reflecting different conditions & circumstances. For instance, one might expect \( p \) to depend on the number of birds released, \( x \): more birds, higher chance of successful establishment.
One could try $p = \beta_0 + \beta_1 x$, but a problem is that $p$ has a natural range of $(0, 1)$, and a linear function would break down for high values of $x$ (with $\beta_1$ positive).

Try:

$$\log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 x$$

The log-odds ratio can be positive or negative, which makes it a convenient way to connect $p$ to a linear function of covariates. Solving for $p$ gives

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$
Data: \((y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\) (\(y_i\)'s are all 0's or 1's)
Unknown parameters: \(\beta_0, \beta_1\)

Likelihood: product of binomial probabilities, each with “# trials” of 1 and each with a different success probability:

\[
L = p_1^{y_1}(1 - p_1)^{1-y_1} p_2^{y_2}(1 - p_2)^{1-y_2} \cdots p_n^{y_n}(1 - p_n)^{1-y_n}
\]

where

\[
p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}
\]

ML estimates (numerical maximization) are \(\hat{\beta}_0\) and \(\hat{\beta}_1\)
Hypothesis test:

\[ H_0: \beta_1 = 0 \quad (x \text{ not in the model}) \]

maximized likelihood \( \hat{L}_0 \) uses \( \hat{p} = \frac{y_1+y_2+\cdots+y_n}{n} \)

\[ H_a: \beta_1 \neq 0 \quad (x \text{ in the model}) \]

maximized likelihood uses \( \hat{\beta}_0 \) and \( \hat{\beta}_1: \hat{L}_a \)

Test statistic:

\[ G^2 = -2 \log \left( \frac{\hat{L}_0}{\hat{L}_a} \right) \]

Rejection region: reject \( H_0 \) if \( G^2 \geq \chi^2_{\alpha} \), where the chi-square distribution has \( 2 - 1 = 1 \) df
Logistic regression in SAS:

PROC GENMOD (example posted at website)
  • for many distributions, binomial, gamma, Poisson, etc.
  • accommodates categorical predictor variables

PROC CATMOD
  • specifically for the multinomial distribution
  • accommodates categorical predictor variables

PROC LOGISTIC
  • quantitative predictor variables only (categorical variables must be coded as indicator variables)
  • various stepwise model selection routines
Poisson regression

$Y_i$: # shoots of a rare plant in a randomly placed sample plot (value of $Y_i$ could possibly be 0)

$x_{i1}, x_{i2}, \ldots x_{ik}$: values of $k$ predictor variables measured for that sample plot (soil pH, etc.)

Assume $Y_i \sim \text{Poisson}(\lambda_i)$ where

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_k x_{ik}}$$

and

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_k x_{ik}$$

Generalized linear models: the binomial, Poisson, normal, gamma, multinomial, & other distributions can be used as the basis of regression. The distributions are members of the generalized linear models (GLIM) family. (not to be confused with the normal-based general linear model in PROC GLM)

Models in this family can be fitted & analyzed with SAS: PROC GENMOD. More about generalized linear models later in course!