Categorical Data Analysis
Lecture Note 6

Instructor: Seokho Lee

Hankuk University of Foreign Studies

6. Regression Models for Count Data

The use of statistical models has numerous advantages in the analysis of categorical data.

- Statistical models will enable us to describe the nature and strength of associations in terms of a small number of parameters.
- Explanatory variables can be continuous, categorical, or both.
- We can use models to analyze the effects of several covariates simultaneously.
- We can formulate questions about association in terms of model parameters. We can estimate these parameters to describe the strength of association and also the effects of covariates.
- We can determine which covariates significantly affect the response, while controlling for the effects of confounding variables.
- The model's predicted values smooth the data and improve the estimate of the mean response.

6.1. Components of a Generalized Linear Model (GLM)

The GLMs extend ordinary regression models by allowing the mean of a population to depend on a linear function of covariates through a nonlinear *link* function. The probability distribution of the response can be any member of an *exponential family of distributions*. Many well known families of distributions including the normal, binomial, and Poisson distributions are exponential families.

GLMs include the following models:

1. Ordinary regression and ANOVA models.
2. The logistic regression model for binary data with the binomial distribution.
3. Loglinear models and regression models for count data with the Poisson distribution.

A GLM has three components:

1. The *random component* identifies the response variable Y.
2. The *systematic component* specifies the explanatory variables using a *linear predictor*.
3. The *link function* relates the mean of the response to the linear predictor.

6.1.1. Random Comonent

- For a sample of size n, the observations on the response variable Y are denoted $(Y_1, \cdots, Y_n)$.
- The GLMs treat $Y_1, \cdots, Y_n$ as being independent random variables with a probability distribution from an exponential family which can be written as

$$f(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

  for specified functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$. These functions are related to the mean and variance of a random variable Y having this distribution:

$$\mu = E[Y] = b'(\theta), \qquad V(Y) = a(\phi)b''(\theta)$$

- The following distributions come from an exponential family:
  - Normal distribution
  - Binomial distribution
  - Poisson distribution

- The mean of Y_i is $\mu_i = E[Y_i]$.
- The variance of Y_i depends on the mean μ_i through the parameter θ and also on the *dispersion parameter* ϕ, which is either known or must be estimated.
- The random component of a GLM consists of identifying the response variable Y and selecting the probability distribution for $Y_1, \cdots, Y_n$.
- In ordinary regression, Y has a normal distribution with a constant variance σ^2.

6.1.2. Systematic Component

- The value of $\mu_i = E[Y_i]$ will depend on the value of the explanatory variables.
- The explanatory variables $x_1, \ldots, x_p$ enter linearly into the expression:

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- This linear combination of the covariates is called the **linear predictor**.

6.1.3. Link

- The **link function** $g(\cdot)$ relates the mean of the random component to the systematic component of the GLM.

$$g(\mu) = \eta$$

or

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

- The simplest link function is the *identity link*:

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

This is used in ordinary regression.

6.1.4. Examples of Generalized Linear Models

Random Component	Link	Systematic Component	Model
Normal	Identity	Continuous	Regression
Normal	Identity	Categorical	ANOVA
Normal	Identity	Mixed	ANCOVA
Binomial	Logit	Mixed	Logistic Regression
Poisson	Log	Mixed	Poisson Regression, Log-Linear
Multinomial	Generalized Logit	Mixed	Multinomial Response

The usual multiple regression model is

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$, $x_1, \ldots, x_p$ are fixed constants, and $\beta_0, \cdots, \beta_p, \sigma^2$ are unknown parameters.

- Random component: $Y \sim N(\mu, \sigma^2)$.
- Systematic component: $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
- Link: $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

6.2. Generalized Linear Models for Binary Data

We will study in details models for data where there are two possible outcomes which we call “Success” (S) and “Failure” (F). A random variable with two possible outcomes is known as a *Bernoulli variable*. Its distribution can be specified as follows:

$$\Pr(Y = 1) = \Pr(S) = p \qquad \text{and} \qquad \Pr(Y = 0) = \Pr(F) = 1 - p.$$

For this model,

$$E[Y] = p \qquad \text{and} \qquad Var(Y) = p(1 - p).$$

The systematic component will depend on an explanatory variable x. The probability of success is written as $\mathbf{p}(x)$ to indicate its dependence on x.

6.2.1. Linear Probability Model

A simple model relating p to x is a linear model:

$$p(x) = \beta_0 + \beta_1 x$$

Problems with this model:

- For certain x, $p(x) > 1$ or $p(x) < 0$.
- Least squares is not optimal because $Var(Y) = p(x)\{1 - p(x)\}$ (not constant variance).
- Maximum likelihood estimators do not have closed form.

6.2.2. Logistic Regression Model

In many cases, a nonlinear regression model is useful for binary data. Typically they

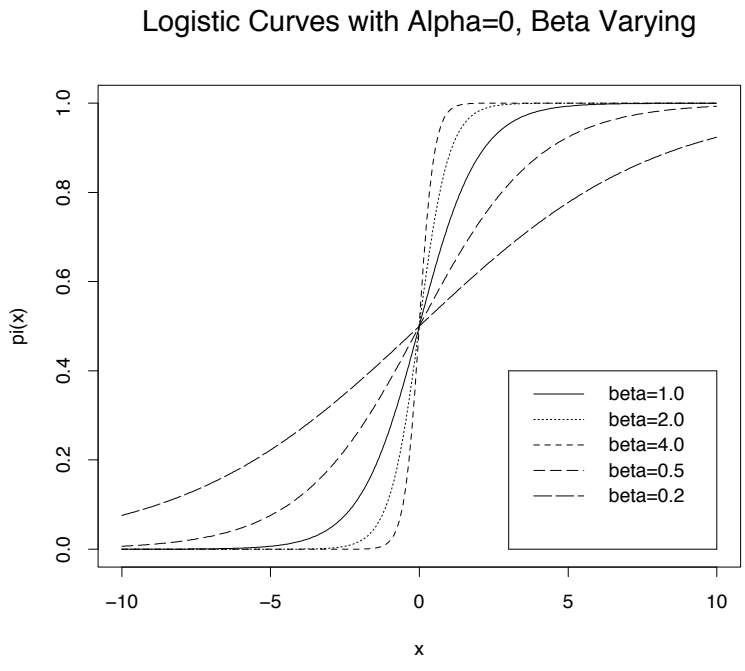
- are *monotonic* with $p(x)$ either increasing as x increases or decreasing as x increases.
- satisfy $0 \leq p(x) \leq 1$.
- often form an S-shaped curve.

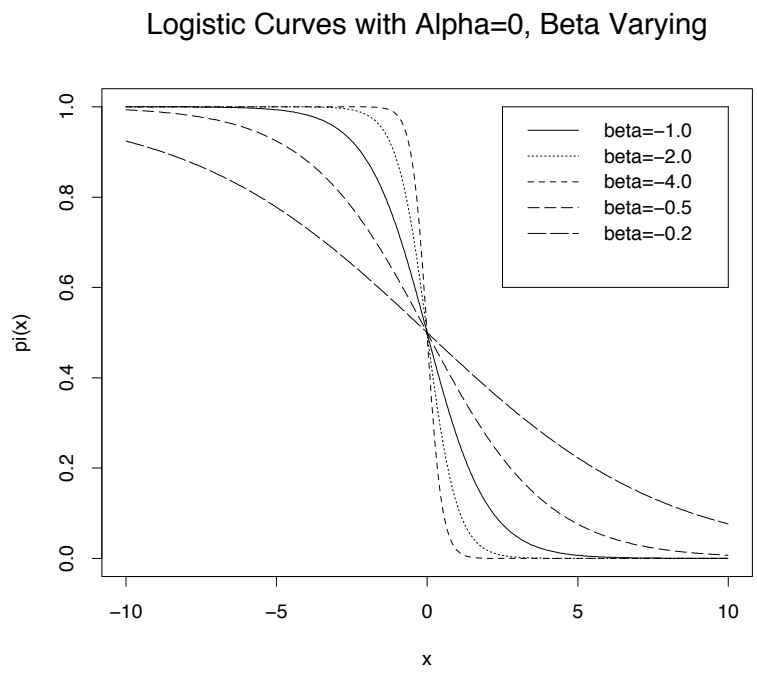
A model that satisfies the above is the *logistic regression function*:

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

We can solve $p(x)$:

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x)\}}.$$


Logistic Curves with Alpha=0, Beta Varying
pi(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
beta=1.0
beta=2.0
beta=4.0
beta=0.5
beta=0.2


Logistic Curves with Alpha=0, Beta Varying
pi(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
beta=−1.0
beta=−2.0
beta=−4.0
beta=−0.5
beta=−0.2

Remark: Because we often wish to use a monotone function $p(x)$ satisfying $0 \leq p(x) \leq 1$, it is convenient to use a cumulative distribution function (cdf) of a continuous random variable. Recall that a cdf is defined as

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

This form of a model is useful when a *tolerance distribution* applies to the subjects' responses. For instance, mosquitoes are sprayed with insecticide at various doses. The response is whether the mosquito dies. Each mosquito has a tolerance and the cdf $F(x)$ describes the distribution of tolerances.

- Logistic distribution: The cdf of a logistic random variable is

$$F(x) = \frac{1}{1 + e^{-1}}, \qquad -\infty < x < \infty$$

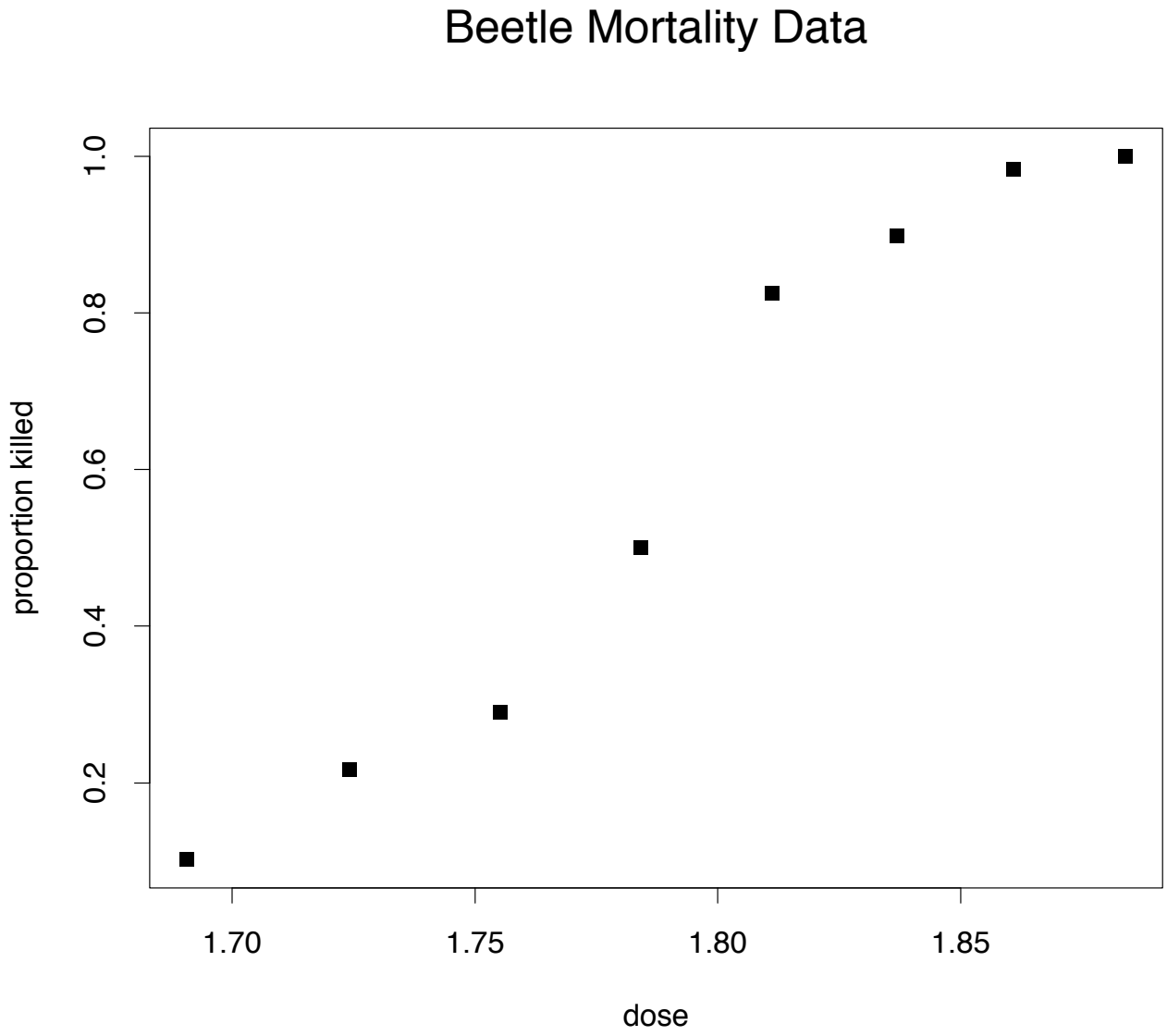
- Normal distribution: The cdf of a normal random variable is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

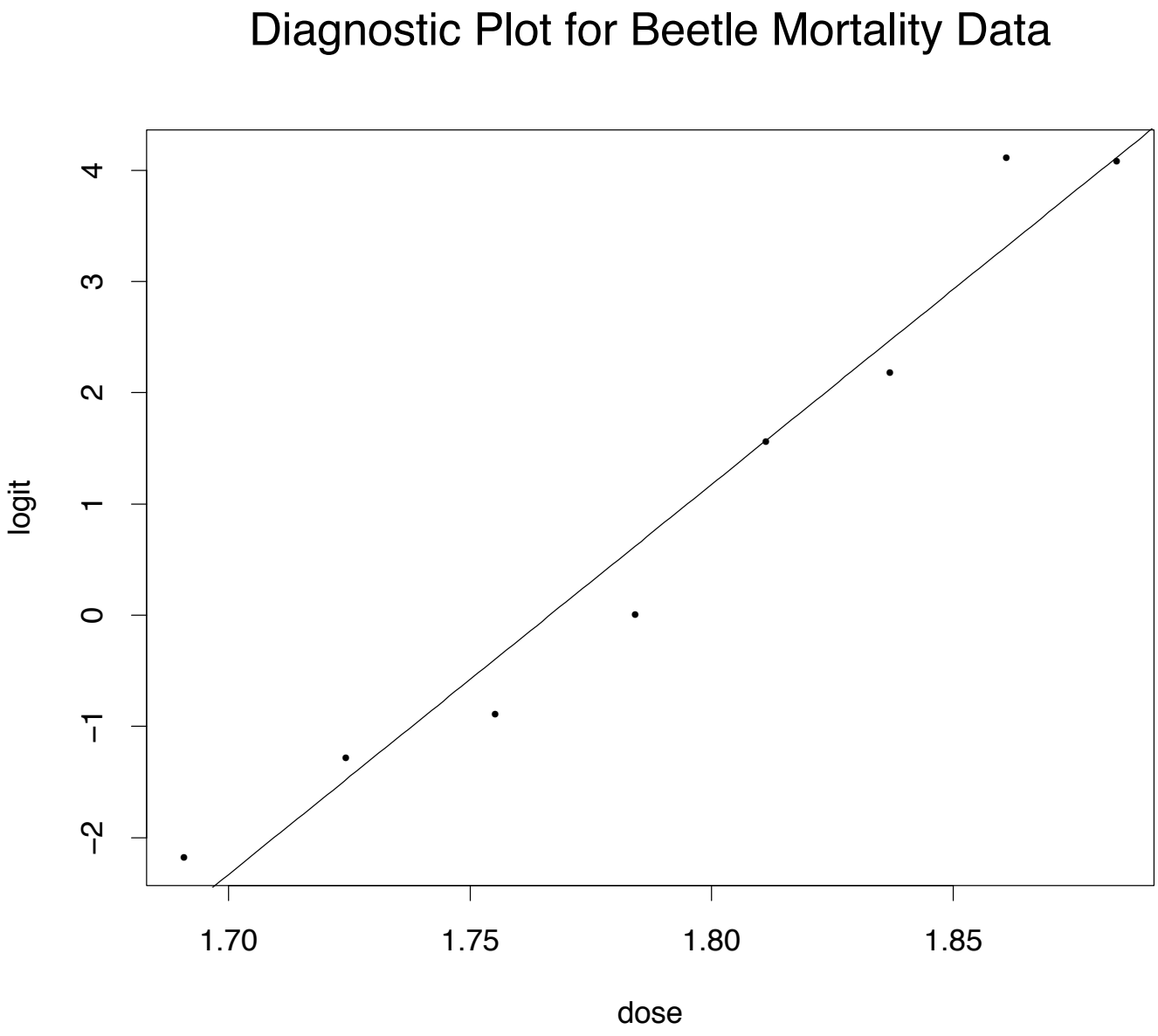
The logit transformation is obtained by finding the inverse of the logistic cdf. The same approach can be used to find the probit link by taking the inverse of the normal cdf.

Example: Beetles were treated with various concentrations of insecticide for 5 hrs. The data appear in the following table:

Dose x_i ($\log_{10} CS_2 mgl^{-1}$)	Number of insects, n_i	Number of killed, Y_i	Proportion of killed, Y_i/n_i
1.6907	59	6	.1017
1.7242	60	13	.2167
1.7552	62	18	.2903
1.7842	56	28	.5000
1.8113	63	52	.8254
1.8369	59	53	.8983
1.8610	62	61	.9839
1.8839	60	59	.9833


Beetle Mortality Data
proportion killed
0.2
0.4
0.6
0.8
1.0
1.70
1.75
1.80
1.85
dose

To see if the logistic model is plausible, we can plot logit($\hat{p}(x)$) versus dose. This plot should appear linear.



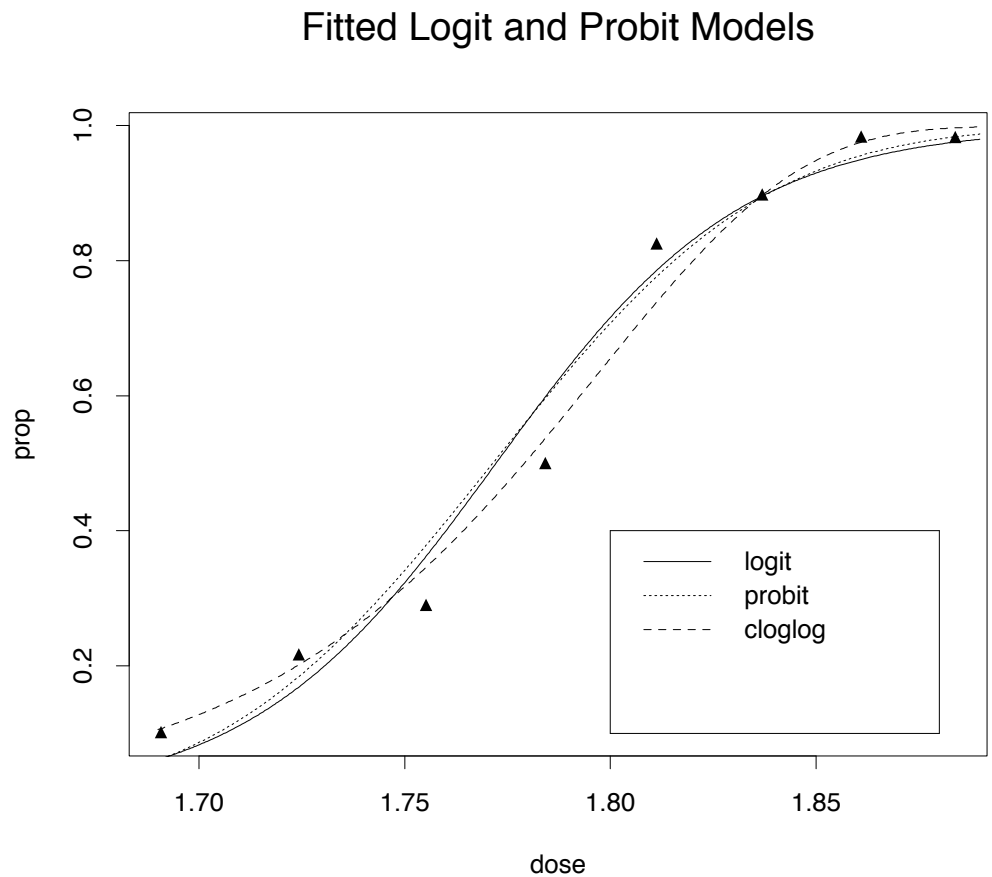
R was used to fit a logistic regression model to these data:

```
> glm(formula=prop~dose, family=binomial(link=logit),weight=insects)

Call:  glm(formula = prop ~ dose, family = binomial(link = logit), weights = insects)

Coefficients:
(Intercept)         dose
     -59.19        33.40

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     274.9
Residual Deviance: 8.645  AIC: 40.82

> glm(formula=prop~dose, family=binomial(link=probit),weight=insects)

Call:  glm(formula = prop ~ dose, family = binomial(link = probit), weights = insects)

Coefficients:
(Intercept)         dose
     -33.92        19.15

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     274.9
Residual Deviance: 8.437  AIC: 40.61

> glm(formula=prop~dose, family=binomial(link=cloglog),weight=insects)

Call:  glm(formula = prop ~ dose, family = binomial(link = cloglog), weights = insects)

Coefficients:
(Intercept)         dose
     -36.89        20.53

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     274.9
Residual Deviance: 6.195  AIC: 38.37
```

The fitted models were:

- $\text{logit}(\hat{p}(x)) = -59.19 + 33.40x$
- $\text{probit}(\hat{p}(x)) = -33.92 + 19.15x$
- $\text{cloglog}(\hat{p}(x)) = -36.89 + 20.53x$

The observed proportions and the fitted model appear in the following graph:



6.3. GLMs for Count Data: Poisson Regression

The Poisson distribution is commonly used for count data. Often we will need a model to relate counts to predictor variables. Since the mean of Poisson random variable is positive, the Poisson loglinear model uses the log link:

$$\log \mu = \beta_0 + \beta_1 x.$$

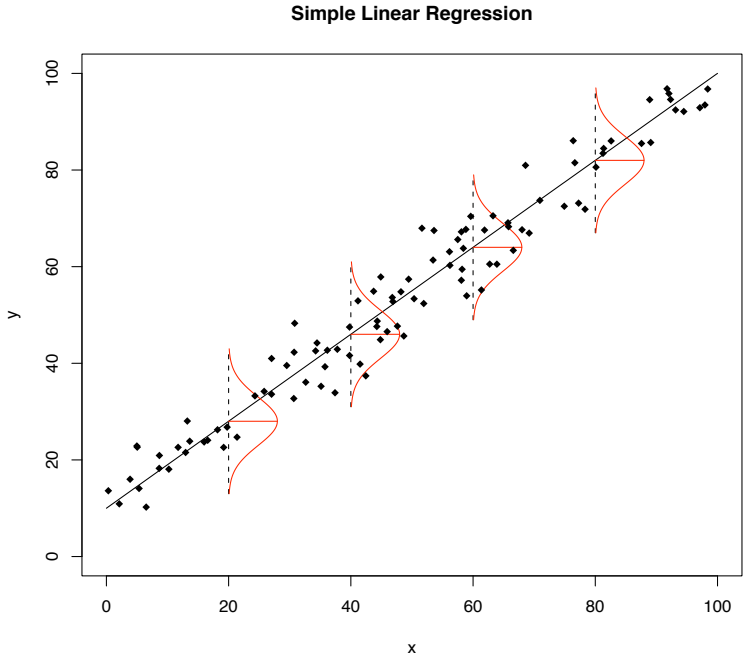
This implies that the mean satisfies the relationship

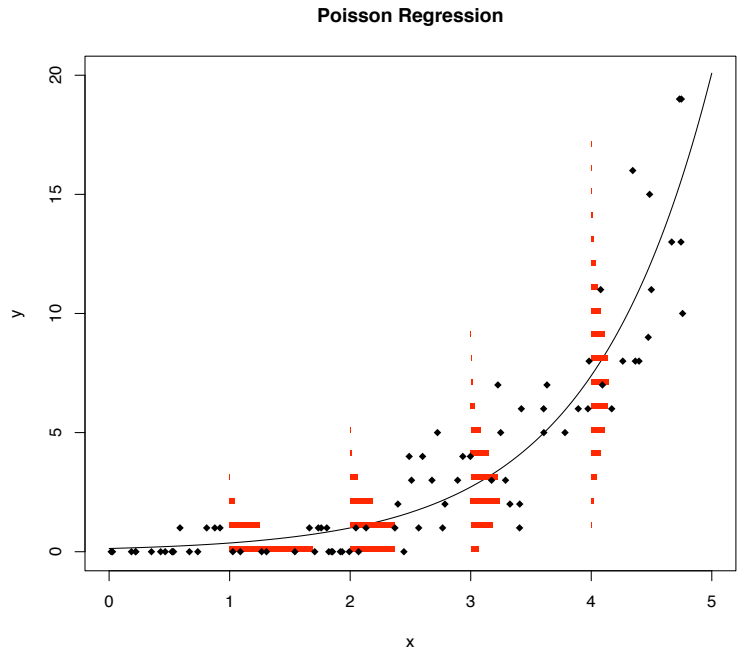
$$\mu = \exp(\beta_0 + \beta_1 x) = e^{\beta_0}(e^{\beta_1})^x.$$

Thus,

$$Y \sim \text{Poisson}(\exp(\beta_0 + \beta_1 x)).$$

If we increase x by 1unit, the mean of Y increases by a factor of e^{β_1}.


Simple Linear Regression
y
x
0
20
40
60
80
100

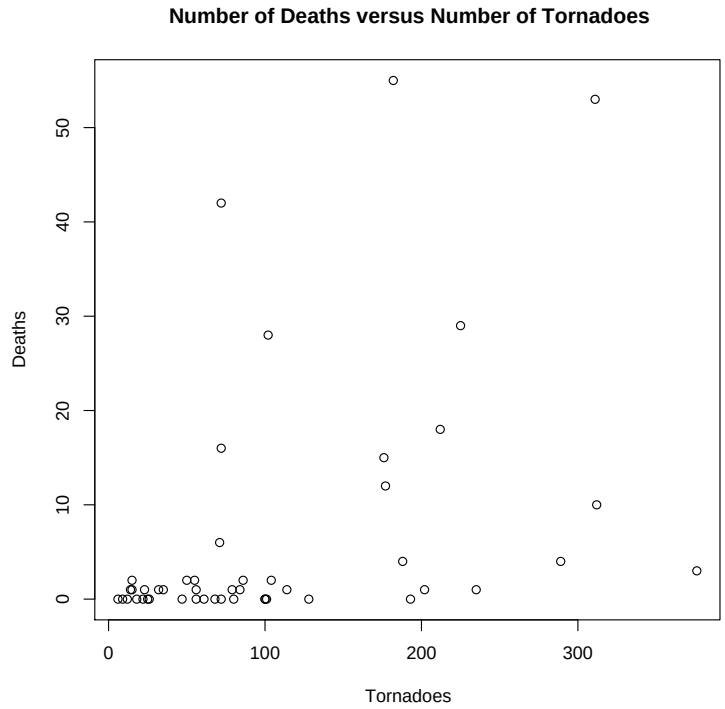

Poisson Regression
y
x
0
5
10
15
20
0
1
2
3
4
5

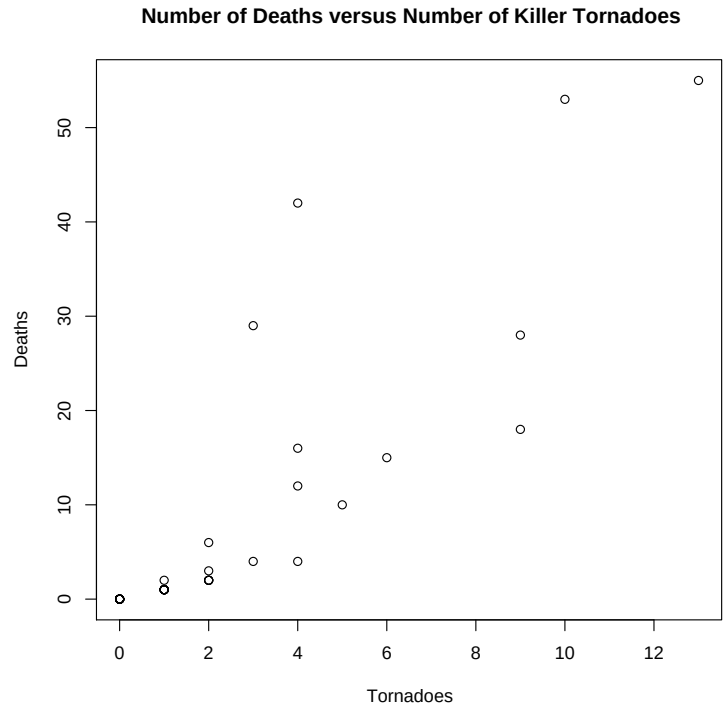
6.3.1. Example–Deaths from Tornadoes

The Storm Prediction Center (an agency of NOAA) tracks the number and characteristics of tornadoes. We want to build a model relating the number of deaths to the number of tornadoes or to the number of killer tornadoes.

```
> head(dat,10)
   Year Month.number     Month Tornado.season Killer.tornadoes Tornadoes Deaths
1  1996            1   January               1                1        35      1
2  1996            2  February               2                1        14      1
3  1996            3     March               3                2        71      6
4  1996            4     April               4                4       177     12
5  1996            5       May               5                1       235      1
6  1996            6      June               6                0       128      0
7  1996            7      July               6                1       202      1
8  1996            8    August               6                0        72      0
9  1996            9 September               6                0       101      0
10 1996           10   October               6                0        68      0
```

The scatter plots of the number of deaths versus the number of tornadoes and versus the number of killer tornadoes indicate that an ordinary linear regression model would not be appropriate.

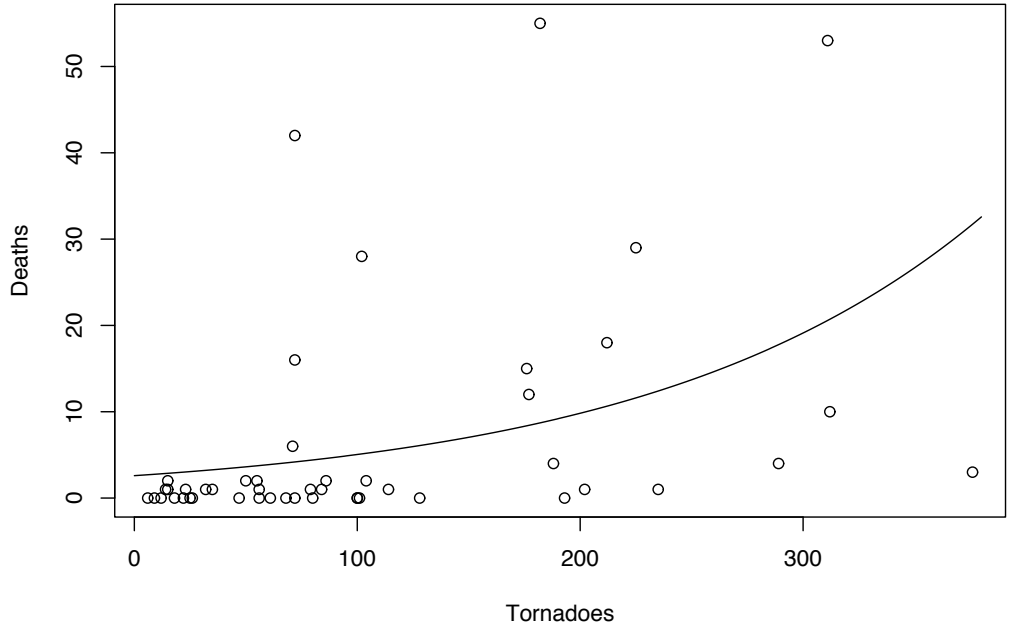




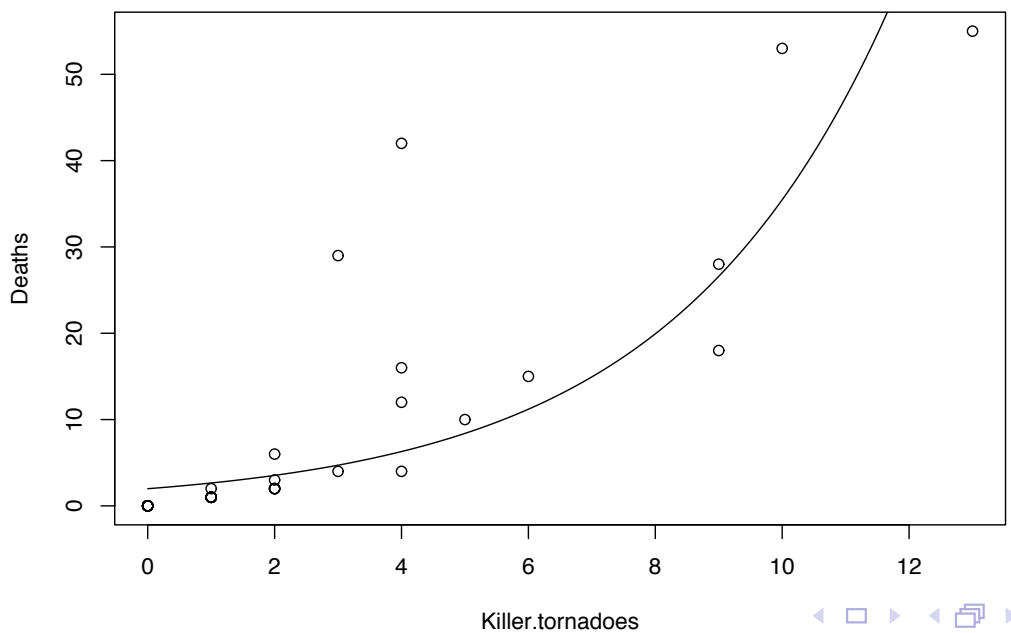
Three Poisson regression models were fit to the data using R. Later we will learn how to use this output to determine which of these is best. We will also consider models with other explanatory variables including month.

```
> glm(formula=Deaths~Tornadoes, family=poisson(link=log), data=dat)

Call:  glm(formula = Deaths ~ Tornadoes, family = poisson(link = log), data = dat)

Coefficients:
(Intercept)    Tornadoes
    0.95295      0.00666

Degrees of Freedom: 47 Total (i.e. Null);  46 Residual
Null Deviance:     810.3
Residual Deviance: 643.5  AIC: 745.8

> glm(formula=Deaths~Killer.tornadoes, family=poisson(link=log), data=dat)

Call:  glm(formula = Deaths ~ Killer.tornadoes, family = poisson(link = log), data = dat)

Coefficients:
     (Intercept)  Killer.tornadoes
          0.6876            0.2881

Degrees of Freedom: 47 Total (i.e. Null);  46 Residual
Null Deviance:     810.3
Residual Deviance: 275.4  AIC: 377.7

> glm(formula=Deaths~Tornadoes+Killer.tornadoes, family=poisson(link=log), data=dat)

Call:  glm(formula = Deaths ~ Tornadoes + Killer.tornadoes, family = poisson(link = log), data = dat)

Coefficients:
     (Intercept)         Tornadoes  Killer.tornadoes
        0.445621          0.002185          0.269227

Degrees of Freedom: 47 Total (i.e. Null);  45 Residual
Null Deviance:     810.3
Residual Deviance: 265.3  AIC: 369.5
```

Number of Deaths versus Number of Tornadoes



Number of Deaths versus Number of Killer Tornadoes



6.4. Inference for GLMs

Inference for GLMs is based on likelihood methods. Here we give a brief overview of estimation and testing from the likelihood point of view.

Model: We suppose that $Y_1, \ldots, Y_n$ are independent and (for now) identically distributed with probability mass function $f(y; \theta)$, where θ represents the unknown parameter. The *parameter space* Θ is the set of possible values of θ.

The **likelihood function** is defined as

$$\mathcal{L}(\theta) \;=\; \prod_{i=1}^{n} f(y_i; \theta) \;=\; f(y_1; \theta) \cdot f(y_2; \theta) \cdots f(y_n; \theta)$$

- We observe $y_1, \ldots, y_n$ and view the likelihood as a function of θ.
- We can interpret $\mathcal{L}(\theta)$ as the probability of observing $y_1, \ldots, y_n$ for a given value of θ.

Often we use the *log-likelihood function* for inference:

$$L(\theta) \;=\; \log \mathcal{L}(\theta) \;=\; \sum_{i=1}^{n} \log f(y_i; \theta)$$

6.4.1. Maximum Likelihood Estimation

The value of θ in Θ that maximizes $\mathcal{L}(\theta)$, or equivalently $L(\theta)$, is known as the **maximum likelihood estimate** (mle).

- Usually we maximize $L(\theta)$ for ease of computation.
- Calculus can often be used to find mles.
- Statistical software for categorical data computes mles.

6.4.2. Properties of Maximum Likelihood Estimators

The MLE has excellent large sample properties under certain regularity conditions:

- The density $f(y;\theta)$ is a smooth function of θ.
- These properties hold as $n \to \infty$.
- The parameter space Θ satisfies certain conditions.

We denote the "true" value of θ by θ_0.

Important properties of the MLE include the following:

1. $\hat{\theta}$ is asymptotically unbiased. That is

$$E[\hat{\theta}] \to \theta_0 \qquad \text{as} \qquad n \to \infty$$

2. $\hat{\theta}$ is asymptotically efficient. This means that in large sample, it has the smallest variance among all asymptotically unbiased estimators.

   Fisher's information is defined as

$$\mathcal{I}(\theta) = E\left[-\frac{\partial^2 L(\theta)}{\partial \theta^2}\right]$$

   This quantity can be estimated in several ways:

$$\mathcal{I}(\hat{\theta}) \qquad - \qquad \text{Plug in}$$

   or

$$-\frac{\partial^2 L(\hat{\theta})}{\partial \theta^2} \qquad - \qquad \text{Hessian or observed information}$$

3. The MLE is asymptotically normal:

$$\frac{\hat{\theta} - \theta_0}{\sqrt{\mathcal{I}(\theta_0)^{-1}}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$

We usually interpret this to mean

$$\hat{\theta} \overset{\text{approx.}}{\longrightarrow} N(\theta_0, \hat{V})$$

where

$$V = \mathcal{I}(\theta_0)^{-1}$$

- The first result says that in large samples the MLE has approximately the desired mean.
- The second result means that the variance of the MLE is as small as possible.
- The third result says that we can use a relatively simple distribution to provide confidence intervals for θ. In general, the actual sampling distribution of $\hat{\theta}$ is very messy.
- $\sqrt{\hat{V}} = \left[\sqrt{\mathcal{I}(\hat{\theta})}\right]^{-1}$ provides the *asymptotic standard error* (ASE) for $\hat{\theta}$.
- $-\dfrac{\partial^2 L(\hat{\theta})}{\partial \theta^2}$ measures the *curvature* of the log-likelihood function.
- The greater the curvature, the greater the information about θ and the smaller the ASE.

Example: The log-likelihood for the binomial distribution is

$$L(\theta) \;=\; \log \mathcal{L}(\theta) \;=\; \log \binom{n}{y} + y \log \theta + (n-y) \log(1-\theta)$$

The first and second derivatives with respect to θ are

$$\frac{\partial L(\theta)}{\partial \theta} \;=\; \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

$$\frac{\partial^2 L(\theta)}{\partial \theta^2} \;=\; -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$$

Fisher's information is

$$\mathcal{I}(\theta) \;=\; -E\left[\frac{\partial^2 L(\theta)}{\partial \theta^2}\right] \;=\; \frac{n\theta}{\theta^2} - \frac{n(1-\theta)}{(1-\theta)^2} \;=\; \frac{n}{\theta(1-\theta)}.$$

Then

$$V \;=\; \mathcal{I}(\theta_0)^{-1} \;=\; \frac{\theta_0(1-\theta_0)}{n}$$

and

$$ASE \;=\; \sqrt{\mathcal{I}(\hat{\theta})^{-1}} \;=\; \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

The asymptotic properties of the MLE imply that

$$\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$

and that

$$\hat{\theta} \overset{\text{approx.}}{\longrightarrow} N\left(\theta_0, \frac{\theta_0(1-\theta_0)}{n}\right).$$

This is equivalent to our earlier normal approximation to the binomial distribution.

Confidence Interval for θ

When

$$\hat{\theta} \overset{\text{approx.}}{\longrightarrow} N(\theta_0, \hat{V}),$$

we can form an approximate $100(1-\alpha)\%$ confidence interval for θ:

$$\hat{\theta} \pm z_{\alpha/2} ASE(\hat{\theta})$$

Example:

$$\frac{y}{n} \pm z_{\alpha/2} \sqrt{\frac{\frac{y}{n}\left(1-\frac{y}{n}\right)}{n}}$$

6.4.3. The Likelihood Approach to Hypothesis Testing

We consider testing $H_0 : \theta = \theta_0$. More generally we could test $H_0 : \theta \in \Theta_0$.

There are three likelihood-based approaches to hypothesis testing:

- Likelihood ratio test
- Wald test
- Score test

1 Wald Test

The Wald test is based on the asymptotic normality of $\hat{\theta}$:

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathcal{I}(\theta_0)^{-1}}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty.$$

We define the *Wald statistic*:

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\mathcal{I}(\hat{\theta})^{-1}}} \sim N(0,1) \qquad \text{or} \qquad W = Z^2 = \frac{(\hat{\theta} - \theta_0)^2}{\mathcal{I}(\hat{\theta})^{-1}} \sim \chi^2(1)$$

Example: Binomial(N, θ)

$$Z = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{\theta}(1-\hat{\theta})}} \qquad \text{or} \qquad W = \frac{n(\hat{\theta} - \theta_0)^2}{\hat{\theta}(1-\hat{\theta})}$$

❷ Likelihood Ratio Test

We wish to compare the likelihood under H_0, $\mathcal{L}(\theta_0)$, to the largest likelihood, $\mathcal{L}(\hat{\theta})$, using the *likelihood ratio statistic*:

$$G^2 = Q_L = -2\log\left[\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\hat{\theta})}\right] = 2[L(\hat{\theta}) - L(\theta_0)] \xrightarrow{d} \chi^2(1) \quad \text{as} \quad n \to \infty.$$

- Now $\mathcal{L}(\theta) \leq \mathcal{L}(\hat{\theta})$ for all $\theta \in \Theta$, so $Q_L > 0$.
- When H_0 is true, the value of $\hat{\theta}$ to be close to θ_0 and the ratio inside Q_L to be close to 1.
- When H_0 is false, the value of $\hat{\theta}$ would differ from θ_0 and $\mathcal{L}(\theta_0) < \mathcal{L}(\hat{\theta})$. We reject H_0 for large values of Q_L.
- The *nested* models can be compared with likelihood ratio test, which will be discussed later.

Example: $Y \sim \text{Binomial}(n, \theta)$

$$\begin{aligned} Q_L &= -2[\log \mathcal{L}(\theta_0) - \log \mathcal{L}(\hat{\theta})] \\ &= -2[y\log\theta_0 + (n-y)\log(1-\theta_0) - y\log\hat{\theta} - (n-y)\log(1-\hat{\theta})] \\ &= 2\left[y\log\left(\frac{\hat{\theta}}{\theta_0}\right) + (n-y)\log\left(\frac{1-\hat{\theta}}{1-\theta_0}\right)\right] \end{aligned}$$

3 Score Test

The score function is defined as

$$U(\theta) = \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \frac{\partial L(\theta)}{\partial \theta}.$$

Recall that the mle is the solution to

$$U(\theta) = \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = 0.$$

We evaluate the score function at the hypothesized value θ_0 and see how close it is to zero.

The score function is asymptotically normal:

$$Z = \frac{U(\theta_0)}{\sqrt{\mathcal{I}(\theta_0)}} \sim N(0,1) \quad \text{or} \quad S = Z^2 = \frac{U(\theta_0)^2}{\mathcal{I}(\theta_0)} \sim \chi^2(1)$$

Example: Binomial random sample

$$U(\theta) = \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

$$S = \frac{\left(\frac{y}{\theta_0} - \frac{n-y}{1-\theta_0}\right)^2}{\frac{n}{\theta_0(1-\theta_0)}} = \frac{n(\hat{\theta} - \theta_0)^2}{\theta_0(1-\theta_0)}$$

Comments

- The above tests all reject for large values based on chi-squared critical values.
- The three tests are asymptotically equivalent. That is, in large samples they will tend to have similar values and lead to the same decision.
- For moderate sample sizes, the LR test is usually more reliable than the Wald test.
- A large difference in the values of the three statistics may indicate that the distribution of $\hat{\theta}$ may not be normal.
- The Wald test is based on the behavior of the log-likelihood at the mle $\hat{\theta}$. The ASE of $\hat{\theta}$ depends on the curvature of the log-likelihood function at $\hat{\theta}$.
- The score test is based on the behavior of the log-likelihood at θ_0. It uses the derivative (or slope) of the log-likelihood at the null value, θ_0. Recall that the slope at $\hat{\theta}$ equals zero.
- Many commonly used test statistics are score statistics:
  - Pearson χ^2 statistic for independence in a 2-way table.
  - Cochran-Armitage M^2 statistic for testing a linear trend alternative to independence
  - Cochran-Mantel-Haenszel statistic for testing conditional independence in a 3-way table

- The LR statistic combines information about the log-likelihood function both at $\hat{\theta}$ and at θ_0. Thus, the LR statistic uses more information than the other two statistics and is usually the most reliable among the three.
- These statistics can be used for multiparameter models. Often we have a parameter vector $(\theta, \beta_1, \cdots, \beta_p)$. We wish to test $H_0 : \theta = \theta_0$. The following are the differences that hold for this model:
  - The score function is now a vector of $p + 1$ partial derivatives of the log-likelihood function.
  - The MLE is determined by solving the resulting set of $p+1$ equations in $p+1$ unknowns.
  - Fisher's information is now a $(p + 1) \times (p + 1)$ matrix.
  - All three statistics are asymptotically equivalent and asymptotically have a chi-squared distribution with 1 d.f..

6.4.4. Deviance

The analysis of generalized linear models is facilitated by the use of the deviance. Let L_M denote the maximized log-likelihood of the model of interest. The *saturated* model is defined to be the most complex model which has a separate parameter for each observation and $\hat{\mu}_i = y_i$, $i = 1, \cdots, n$. Let L_S denote the maximized log-likelihood of the saturated model.

The **deviance** $D(M)$ is defined to be

$$\text{Deviance} \;=\; D(M) \;=\; 2[L_S - L_M]$$

The deviance is the LR statistic for comparing model M to the saturated model. Often the deviance has an approximately chi-squared distribution.

An analogy to the decomposition of sums of squares for linear models holds for the deviance in generalized linear models. Suppose that model M_0 is a special case of model M_1. Such a model is said to be *nested*. Given that model M_1 holds, the LR statistic for testing that the simpler model holds is

$$\begin{aligned} Q_L &= 2[L_{M_1} - L_{M_0}] \;=\; 2[L_S - L_{M_0}] - 2[L_S - L_{M_1}] \\ &= D(M_0) - D(M_1) \end{aligned}$$

Thus, one can compare models by comparing deviances. For large samples, this statistic is approximately chi-squared with *df* equal to the difference in residual *df* for the two models.

6.5. Choosing the "Best" Model

- When the models are nested (i.e., all the explanatory variables in the smaller model are also contained in the larger model), one can use a LR test to choose between the two models.
- There are various criteria one can use to select a model form a collection of possible models. Some of the more commonly used criteria are present below.

Since the log-likelihood tends to be larger and the deviance tends to be smaller for models with more variables, we should consider measures that penalize the log-likelihood for the number of parameters in the model. The goal is to balance the goodness of fit of the model with simplicity of the model. One such measure is AIC:

1. Akaike Information Criterion

$$AIC = -2L + 2\nu$$

where ν is the number of parameters in the model. When comparing models, we choose the model with a smaller value of AIC.

AIC has a tendency to overfit models; that is, AIC can lead to models with too many variables. A version that increases the protection against overfitting is the corrected AIC:

2. Corrected Akaike Information Criterion

$$AIC_C = -2L + 2\nu\left(\frac{n}{n-\nu-1}\right)$$

We note that the AIC criterion can be written in terms of the deviance:

$$AIC^* = Deviance - 2L_s + 2\nu$$

Since the likelihood of the saturated model will be the same for all models being compared, we can order the models based on the sum of the deviance and twice the number of parameters:

$$AIC = Deviance + 2\nu$$

Similarly we can consider the corrected AIC criterion:

$$AIC_C = Deviance + 2\nu\left(\frac{n}{n-\nu-1}\right)$$

3. Schwarz Criterion - A Bayesian argument yields the Bayesian information criterion:

$$BIC = Deviance + \nu \log n.$$

Comments:

- *AIC*, AIC_C, and *BIC* penalize the log likelihood for the number of parameters in the model.
- Smaller values of *AIC*, AIC_C, or *BIC* indicate a more preferable model.
- For large sample sizes, the models chosen by *AIC* and AIC_C will be virtually the same.
- For large sample sizes, *BIC* will produce a larger penalty for additional variables and will tend to choose models with fewer predictors.
- One can produce a list of models to obtain a single “best” model using these criteria. It is more useful to use the criteria for comparing models.

6.6. Model Checking Using Residuals

Residuals are based on chi-squared statistics for testing lack of fit in a generalized linear model. Consider the two statistics for lack-of-fit.

- Likelihood ratio (deviance) statistic

$$\begin{aligned} \textit{Deviance} &= 2\sum_{i=1}^{n}\left[y_i(\hat{\theta}_i^S - \hat{\theta}_i^M) - b(\hat{\theta}_i^S) + b(\hat{\theta}_i^M)\right] \Big/ a_i(\phi) \\ &= \sum_{i=1}^{n} d_i \end{aligned}$$

- Generalized Pearson statistic

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{V}(y_i)}$$

Under H_0 that the model is correct, both statistics should have an approximate chi-squared distribution with $n - p - 1$ degrees of freedom.

We can use the terms in either sum to define a residual to assess lack of fit

- Deviance residual

$$r_i^D = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$$

- Pearson residual

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}}$$

Model checking can be carried out using plots of these residuals.