3 Generalized Linear Models

In the previous chapter we examined association in two-way and three-way contingency tables. We primarily carried out tests of association and estimated certain measures of association of categorical variables. In this section we will consider the use of models to analyze categorical data.

- The inference procedures on two-way and three-way tables also result from analyzing covariate effects on the response in certain models.
- The earlier procedures primarily tested the null hypothesis of no association versus a general alternative.
- The CMH tests enabled us to test for specific alternatives including linear trends, mean location shifts, and general association.
- In three-way tables, we were able to adjust for the effect of a single covariate.

Statistics 659

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 estimates of the mean response.
- The generalized linear models that we will be using can be viewed as an extension of ordinary regression and ANOVA models for continuous response data.

3.1 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 distribution of 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.

3.1.1 Random Component

- For a sample of size n, the observations on the response variable Y are denoted $(Y_1, \ldots, Y_n)$.

- The GLMs treat $Y_1, \ldots, Y_n$ as being independent random variables with a probability distribution from an exponential family which can be written as

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

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

$$\begin{aligned} \mu_i = E(Y_i) &= b'(\theta_i) \\ \text{Var}(Y_i) &= a(\phi)b''(\theta_i) \end{aligned}$$

- The following distributions come from an exponential family:
  - Normal distribution
  - Binomial distribution
  - Poisson distribution
  - Negative binomial 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_i 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 $Y_1, \ldots, Y_n$.

- In ordinary regression, Y_i has a normal distribution with a constant variance σ^2.

3.1.2 Systematic Component

- The value of $\mu_i = E(Y_i)$ will depend on the value of the explanatory variables.

- The explanatory variables (or predictors) $x_1, \ldots, x_p$ enter linearly into the expression:

$$\eta = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (\alpha \text{ is often denoted as } \beta_0)$$

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

3.1.3 Link

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

$$\begin{aligned} g(\mu) &= \eta \\ &\text{or} \\ g(\mu_i) &= \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \end{aligned}$$

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

$$\mu_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

This is used in ordinary regression.

3.1.4 Examples of Generalized Linear Models

Random		Systematic		
Component	Link	Component	Model	Chapter
Normal	Identity	Continuous	Regression	
Normal	Identity	Categorical	ANOVA	
Normal	Identity	Mixed	ANCOVA	
Binomial	Logit	Mixed	Logistic Regression	4, 5, 8, 9
Poisson	Log	Mixed	Poisson Regression	3
			Log-Linear Models	7, 8
Negative binomial	Log	Mixed	Neg. Bin. Regression	3
Multinomial	Generalized Logit	Mixed	Multinomial Response	6
Multinomial	Cumulative Logit	Mixed	Proportional Odds Model	6

The usual multiple regression model is

$$Y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$, $x_1, \ldots, x_k$ are fixed constants, and $\alpha, \beta_1, \cdots, \beta_k, \sigma^2$ are unknown parameters.

- Random component: $Y \sim N(\mu, \sigma^2)$
- Systematic component: $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$
- Link: $\mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$

3.2 Generalized Linear Models for Binary Data

We will study in detail 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:

$$P(Y=1)=P(S)=\pi \quad \text{and} \quad P(Y=0)=P(F)=1-\pi.$$

For this model,

$$E(Y)=\pi \quad \text{and} \quad \text{Var}(Y)=\pi(1-\pi).$$

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

3.2.1 Linear Probability Model

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

$$\pi(x) = \alpha + \beta x$$

Problems with this model:

1. For certain x, $\pi(x) > 1$ or $\pi(x) < 0$.

2. Least squares is not optimal because $\mathrm{Var}(Y) = \pi(x)(1 - \pi(x))$.

3. Maximum likelihood estimators do not have closed form.

3.2.2 Logistic Regression Model

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

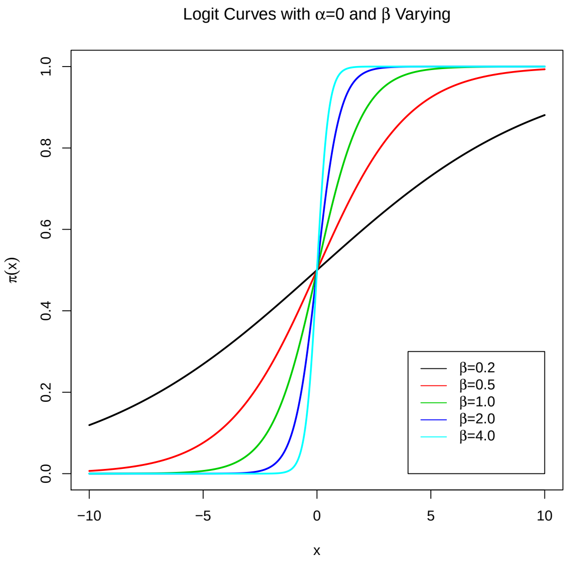
- are *monotonic* with $\pi(x)$ either increasing as x increases or decreasing as x increases.
- satisfy $0 \leq \pi(x) \leq 1$.
- often form an *S-shaped curve*.

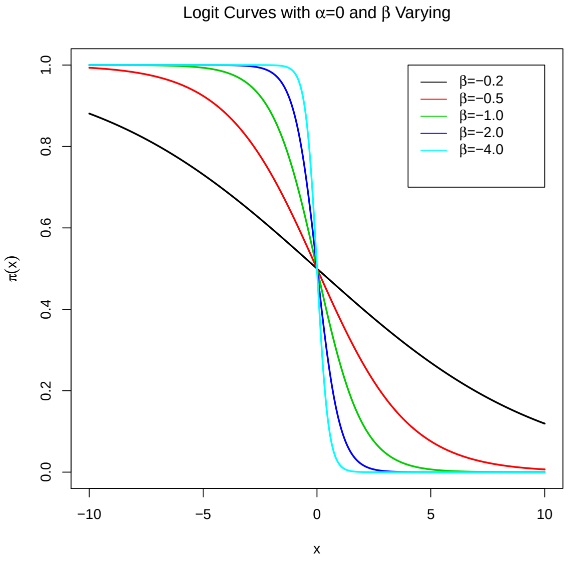
An model that satisfies the above is the *logistic regression function:*

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x.$$

We can solve for $\pi(x)$:

$$\pi(x) = \frac{\exp[\alpha + \beta x]}{1 + \exp[\alpha + \beta x]} = \frac{1}{1 + \exp[-(\alpha + \beta x)]}.$$


Logit Curves with α=0 and β Varying
π(x)
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
x
β=0.2
β=0.5
β=1.0
β=2.0
β=4.0


Logit Curves with α=0 and β Varying
π(x)
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
x
β=−0.2
β=−0.5
β=−1.0
β=−2.0
β=−4.0

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

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

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

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

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

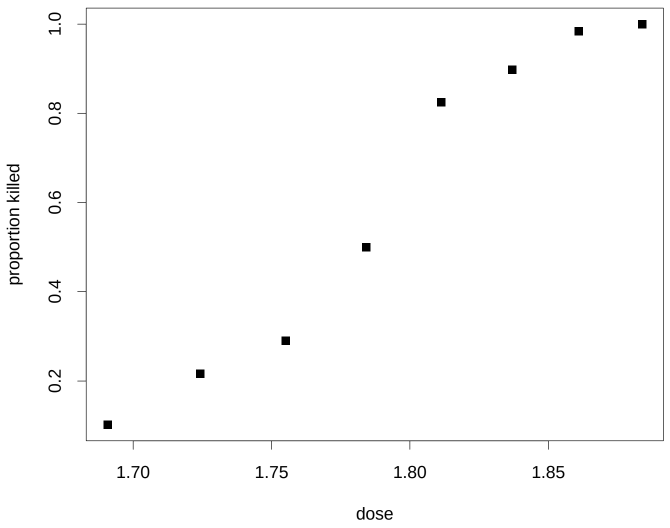
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

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	Number of	Number	Proportion
$(\log_{10}\mathrm{CS}_2\mathrm{mgl}^{-1})$	insects, n_i	killed, Y_i	killed, $\frac{y_i}{n_i}$
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	0.9833

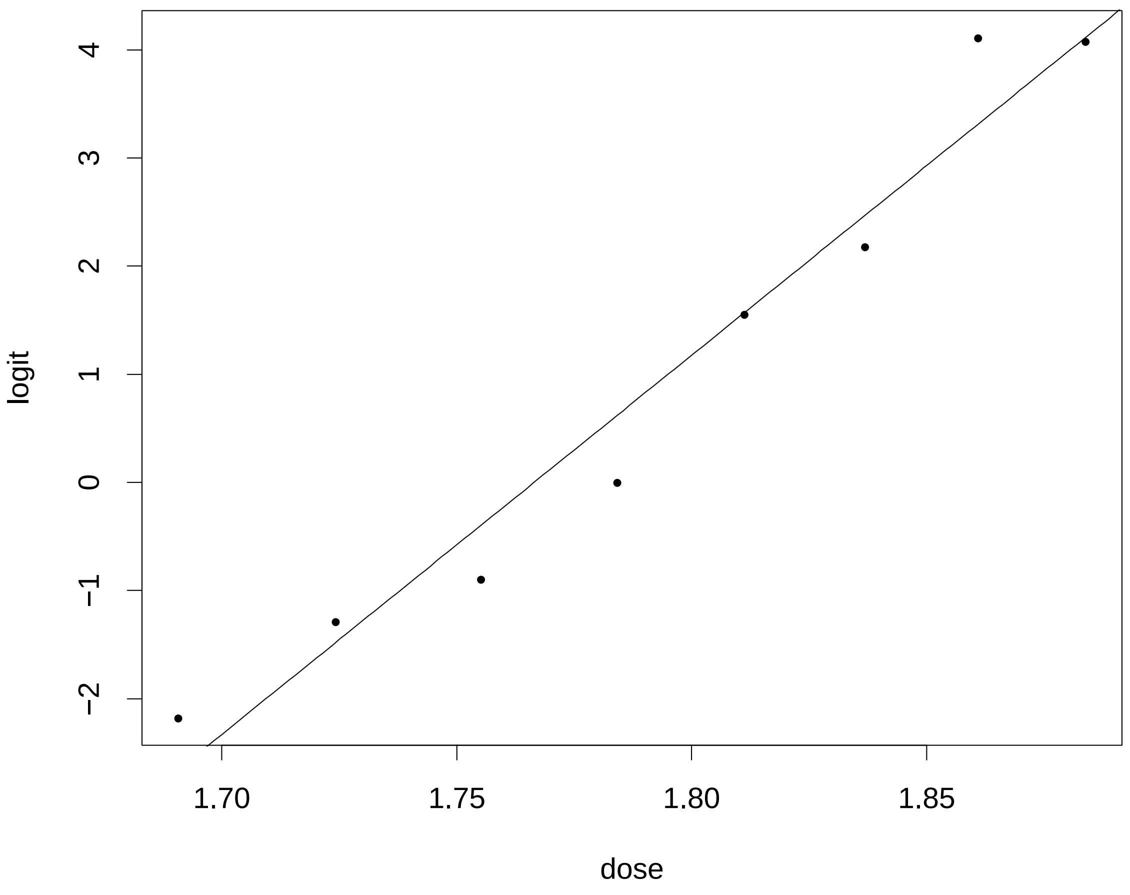
Beetle Mortality Data



Statistics 659

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

Diagnostic Plot for Beetle Mortality Data



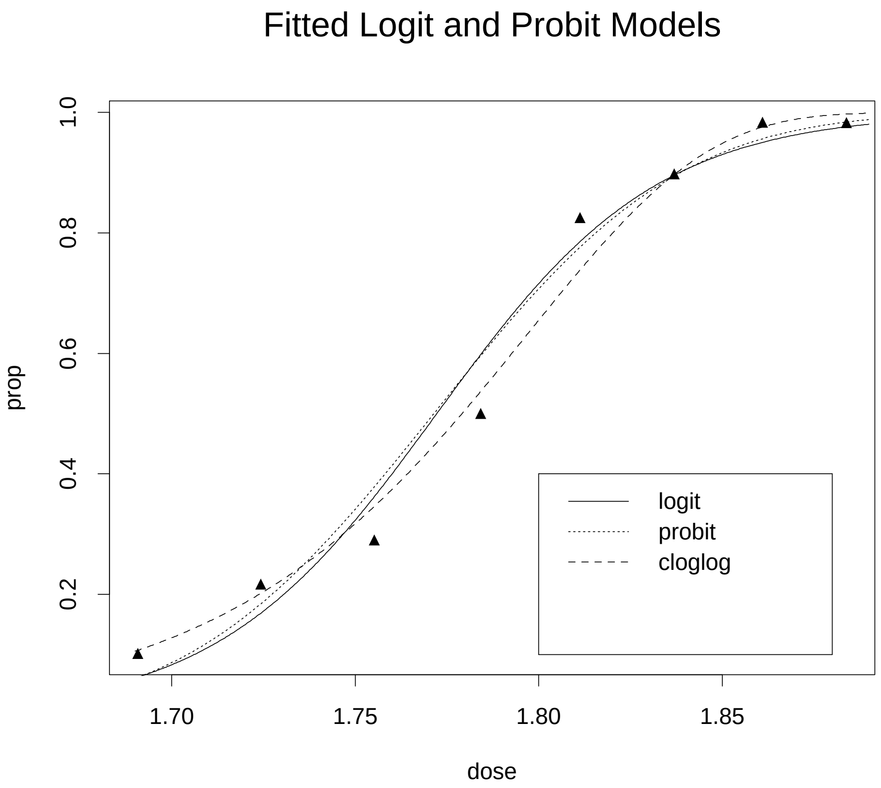
R was used to fit three binary regression models to these data.

```
> glm(prop~dose,family=binomial(link=logit),weight=insects)
Call:
glm(formula = prop ~ dose, family = binomial(link = logit), weights = insects)
Coefficients:
 (Intercept)    dose
   -59.18754 33.4007

Degrees of Freedom: 8 Total; 6 Residual
Residual Deviance: 8.639754
> glm(prop~dose,family=binomial(link=probit),weight=insects)
Call:
glm(formula = prop ~ dose, family = binomial(link = probit), weights = insects)
Coefficients:
 (Intercept)    dose
   -33.91668 19.1474

Degrees of Freedom: 8 Total; 6 Residual
Residual Deviance: 8.430316
> glm(prop~dose,family=binomial(link=cloglog),weight=insects)
Call:
glm(formula = prop ~ dose, family = binomial(link = cloglog), weights = insects)
Coefficients:
 (Intercept)    dose
   -36.89805 20.5354

Degrees of Freedom: 8 Total; 6 Residual
Residual Deviance: 6.185176
```

The fitted models were:

- $\text{logit}(\hat{\pi}(x)) = -58.936 + 33.255x$
- $\text{probit}(\hat{\pi}(x)) = -33.791 + 19.076x$
- $\text{cloglog}(\hat{\pi}(x)) = -36.862 + 20.513x$

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



3.3 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 a Poisson random variable is positive, the Poisson loglinear model uses the log link:

$$\log \mu = \alpha + \beta x.$$

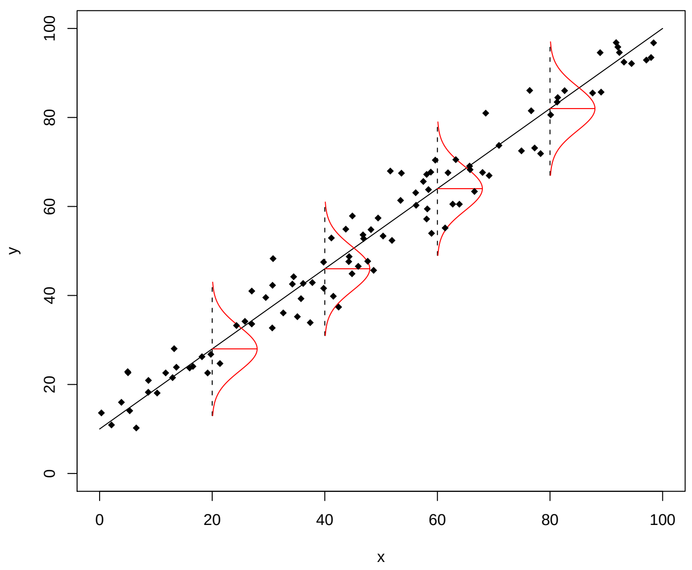
This implies that the mean satisfies the relationship

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}.$$

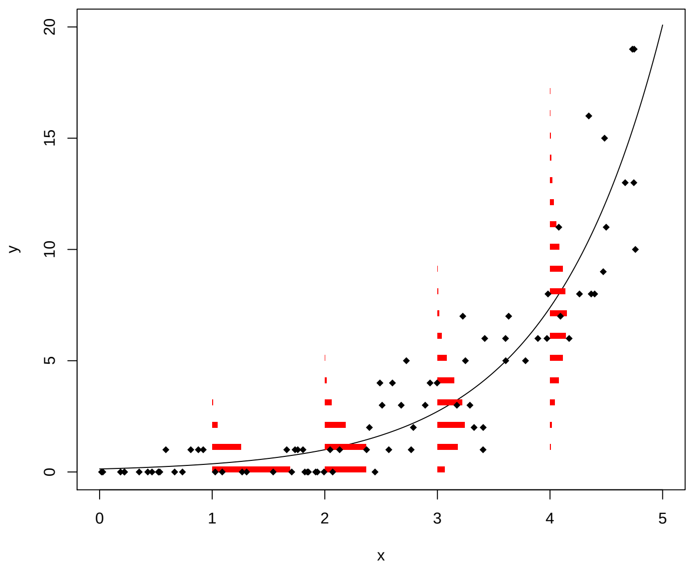
Thus,

$$Y \sim \text{Poisson}(\exp(\alpha + \beta x)).$$

If we increase x by 1 unit, the mean of Y is multiplied by a factor of e^{β}.



Simple Linear Regression



Poisson Regression

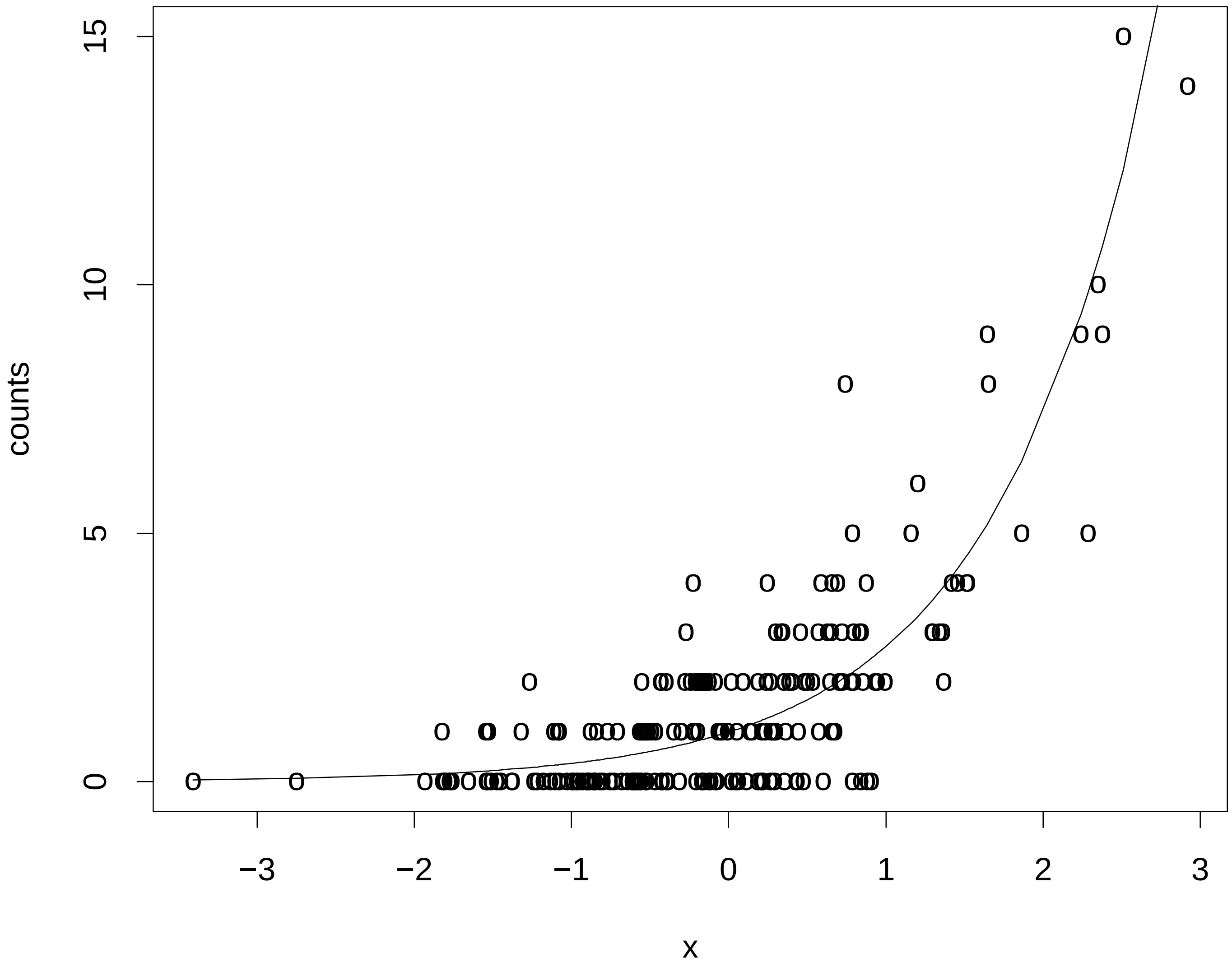
3.3.1 Simulation of Poisson Regression Data

- Generate $x_1, \ldots, x_n$, a random sample from a normal distribution with mean 0 and variance 1.
- For each x_i, generate $Y_i \sim$ Poisson(e^{x_i}) random variable.
- Plot $(x_i, Y_i),\ i = 1, \ldots, n$.
- Plot $\mu = e^x$.

We also fit the data plotted on the next slide to a Poisson regression model using R.

```
> glm(formula=y˜x, family=poisson(link=log))
Call:
glm(formula = y ˜ x, family = poisson(link = log))

Coefficients:
 (Intercept)         x
  0.09811541 0.9275407

Degrees of Freedom: 200 Total; 198 Residual
Residual Deviance: 206.7074
```

Simulated Data from a Poisson Regression Model



3.3.2 Example — Demand for Medical Care

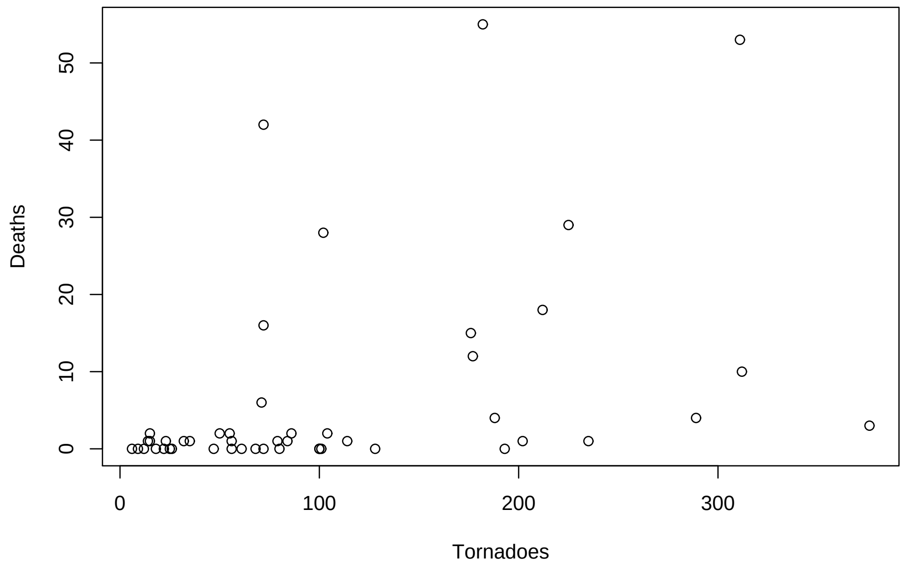
Researchers wish to model the number of physician and hospital outpatient visits (`ofp`) using the number of hospital stays (`hosp`), self-perceived health status (`healthpoor, healthexcellent`), the number of chronic conditions (`numchron`), age (`age`), marital status (`married`), and the number of years of education (`school`) as predictors. Since the effects differ across gender, the results are analyzed for only males.

3.3.3 Example — Deaths from Tornadoes

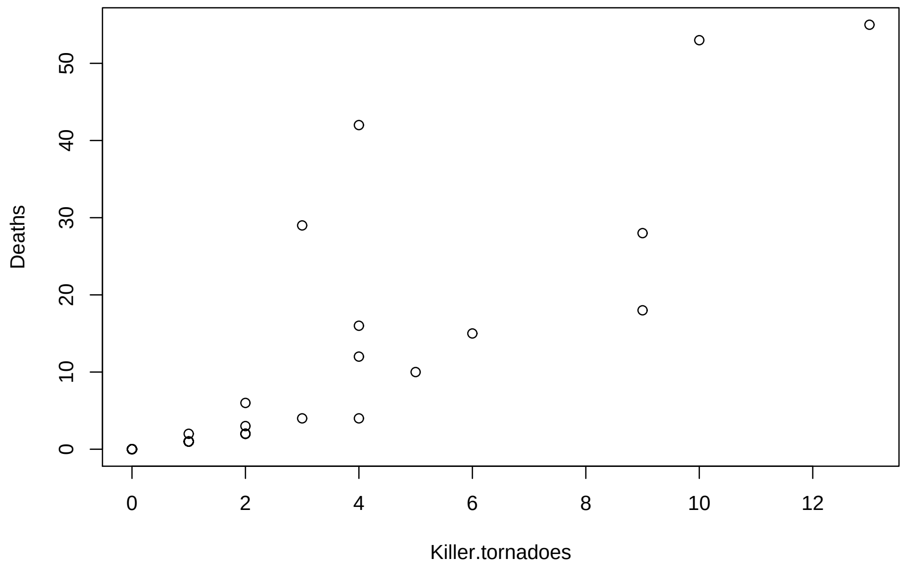
The Storm Prediction Center (an agency of NOAA) tracks the number and characteristics of tornadoes. We want to build a model relating the number of deaths to the number of tornadoes or to the number of killer tornadoes.

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.

Number of Deaths versus Number of Tornadoes



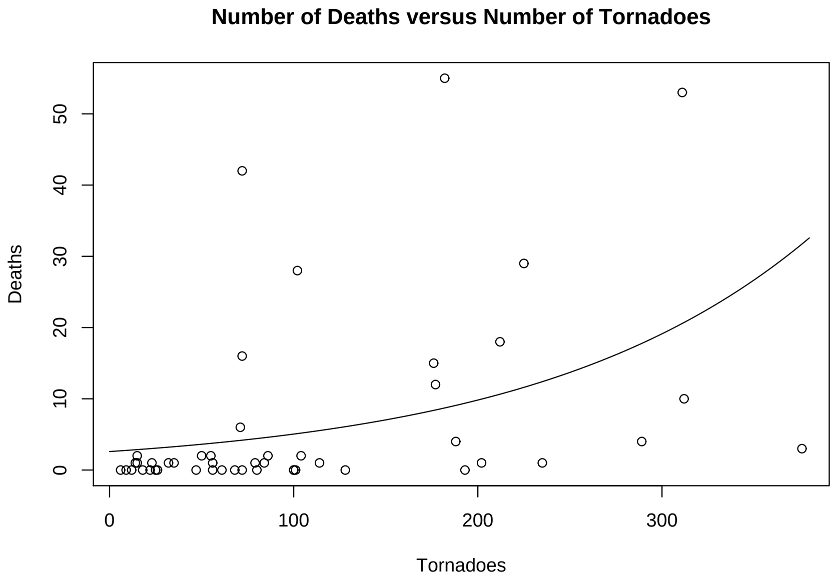
Number of Deaths versus Number of Killer Tornadoes

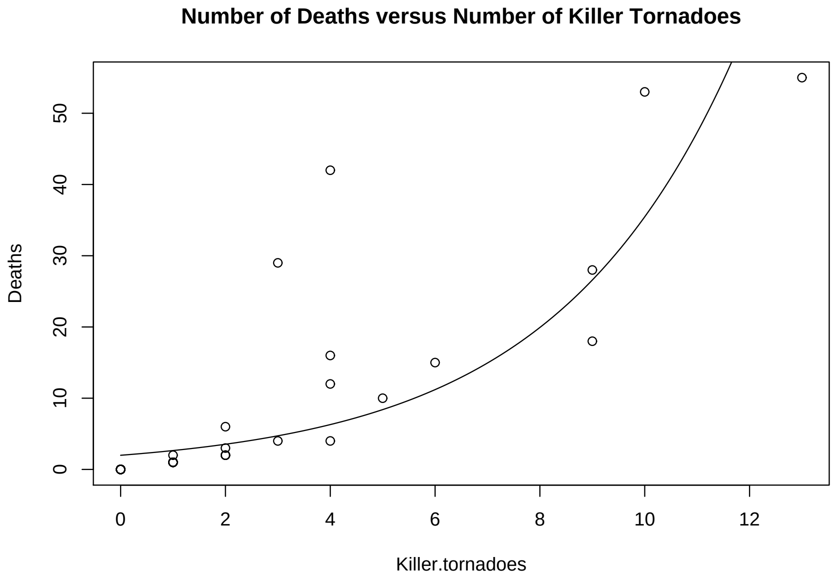


Three Poisson regression models were fit to the data using R. Later we will learn how to use this output to help us determine which of these models is best. We will also consider models with other explanatory variables including month.

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

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

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(Deaths~Killer.tornadoes,family=poisson(link=log))

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

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(Deaths~Tornadoes+Killer.tornadoes,family=poisson(link=log))

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

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
```

Statistics 659





3.3.4 Poisson Regression for Rate Data

It is often the case that the response of interest is the rate of occurence of some event rather than the number of occurences of that event.

- When analyzing the number of marriages by state, we would model the marriage rate (number of marriages per 100,000 residents) instead of the number of marriages.
- When analyzing the number of train accidents in the United States, we would model the number of train accidents per million miles travelled.

When the response Y has an index equal to k, the sample rate of outcomes is Y/k. The expected rate is μ/k. A loglinear model for the expected rate has the form

$$\log(\mu/k) = \alpha + \beta x.$$

This is equivalent to

$$\log(\mu) - \log(k) = \alpha + \beta x.$$

The adjustment term, $\log(k)$, to the log-link of the mean is called the *offset term*.

For this model the expected number of outcomes is

$$\mu = k \exp(\alpha + \beta x) = \exp(\log(k) + \alpha + \beta x).$$

3.4 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

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

- We observe $y_1, \ldots, y_n$ and view the likelihood as a function of θ.
- We can interpret $\ell(\theta)$ 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:

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

3.4.1 Maximum Likelihood Estimation

The value of θ in Θ that maximizes $\ell(\theta)$, 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.

3.4.2 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.
- The parameter space Θ satisfies certain conditions.

These properties hold as $n \to \infty$.

We denote the “true” value of θ by θ_0.

Asymptotic normality of the MLE

Fisher's information is used in obtaining the asymptotic variance of the MLE and 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:

$$\hat{V} = \mathcal{I}(\hat{\theta}) \quad - \quad \text{Plug in}$$

$$\hat{V} = -\frac{\partial^2 L(\hat{\theta})}{\partial\theta^2} \quad - \quad \text{Hessian or observed information.}$$

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 replace $I(\theta_0)^{-1}$ by its estimate to obtain

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

The maximum likelihood estimate has the following properties:

- In large samples the MLE has approximately the desired mean.
- The variance of the MLE is as small as possible.
- 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* (SE) 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 SE.

Example: The log-likelihood for the binomial distribution is

$$L(\theta) = \log(\ell(\theta)) = y\log(\theta) + (n-y)\log(1-\theta).$$

The first and second derivatives are

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

and

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

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

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

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

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

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

Confidence Interval for θ

When

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

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

$$\hat{\theta} \pm Z_{\alpha/2} SE(\hat{\theta}).$$

This interval is the **Wald** confidence interval for θ.

Example:

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

3.4.3 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 $\hat{\theta}$:

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

   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)

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

2. **Likelihood Ratio Test**

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

$$G^2 = Q_L = -2\,\log\left[\frac{\ell(\theta_0)}{\ell(\hat{\theta})}\right] = 2[L(\hat{\theta}) - L(\theta_0)] \xrightarrow{d} \chi_1^2 \text{ as } n \longrightarrow \infty.$$

- Now $\ell(\theta) \leq \ell(\hat{\theta})$ for all $\theta \in \Theta$, so $Q_L > 0$.
- When H_0 is true, we would expect $\hat{\theta}$ 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 $\ell(\theta_0) < \ell(\hat{\theta})$. We reject H_0 for large values of Q_L.

Example: $Y \sim$ Binomial(n, θ)

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

3. **Score Test**

The score function is defined as

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

Recall that the mle is the solution to

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

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

The score statistic is asymptotically normal:

$$Z = \frac{U(\theta_0)}{\sqrt{\mathcal{I}(\theta_0)}} \sim N(0,1) \qquad \text{or} \qquad S = Z^2 = \frac{U(\theta_0)^2}{\mathcal{I}(\theta_0)} \sim \chi^2_1$$

Example: Bernoulli random sample

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

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

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 SE 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 function 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 $(\theta, \theta_1, \ldots, \theta_p)$. 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.

3.4.4 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 $\hat{\mu}_i = y_i,\ i = 1, \ldots, n$. 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.

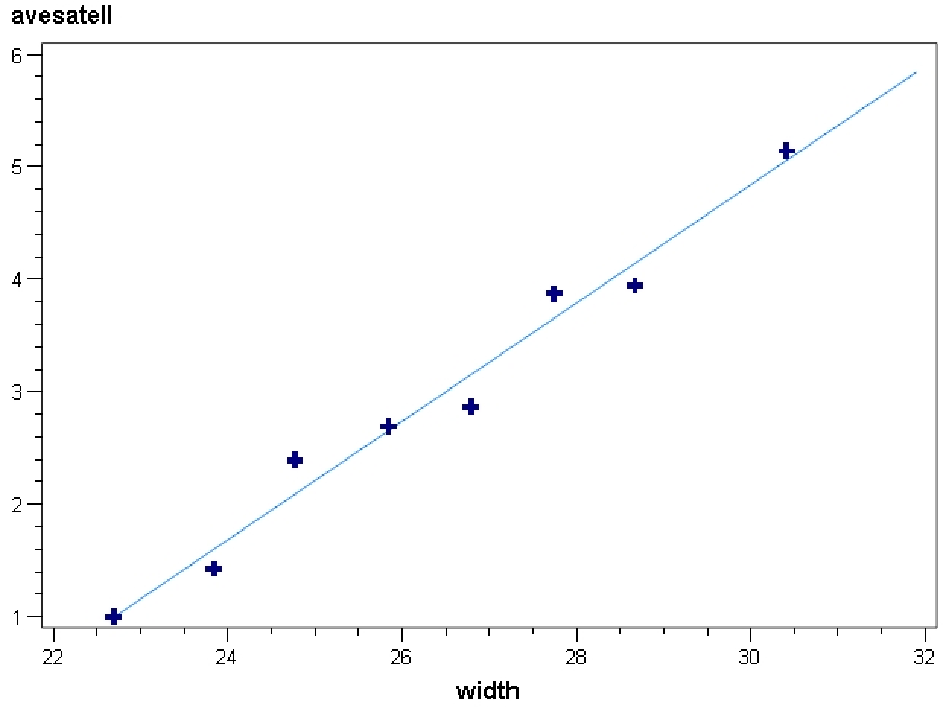
An analogy to the decomposition of sums of squares for linear models holds for the deviance in generalized linear models. Suppose that model M_0 is a special case of model M_1. Such a model is said to be *nested*. Given that M_1 holds and that both models have the same saturated model, the LR statistic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

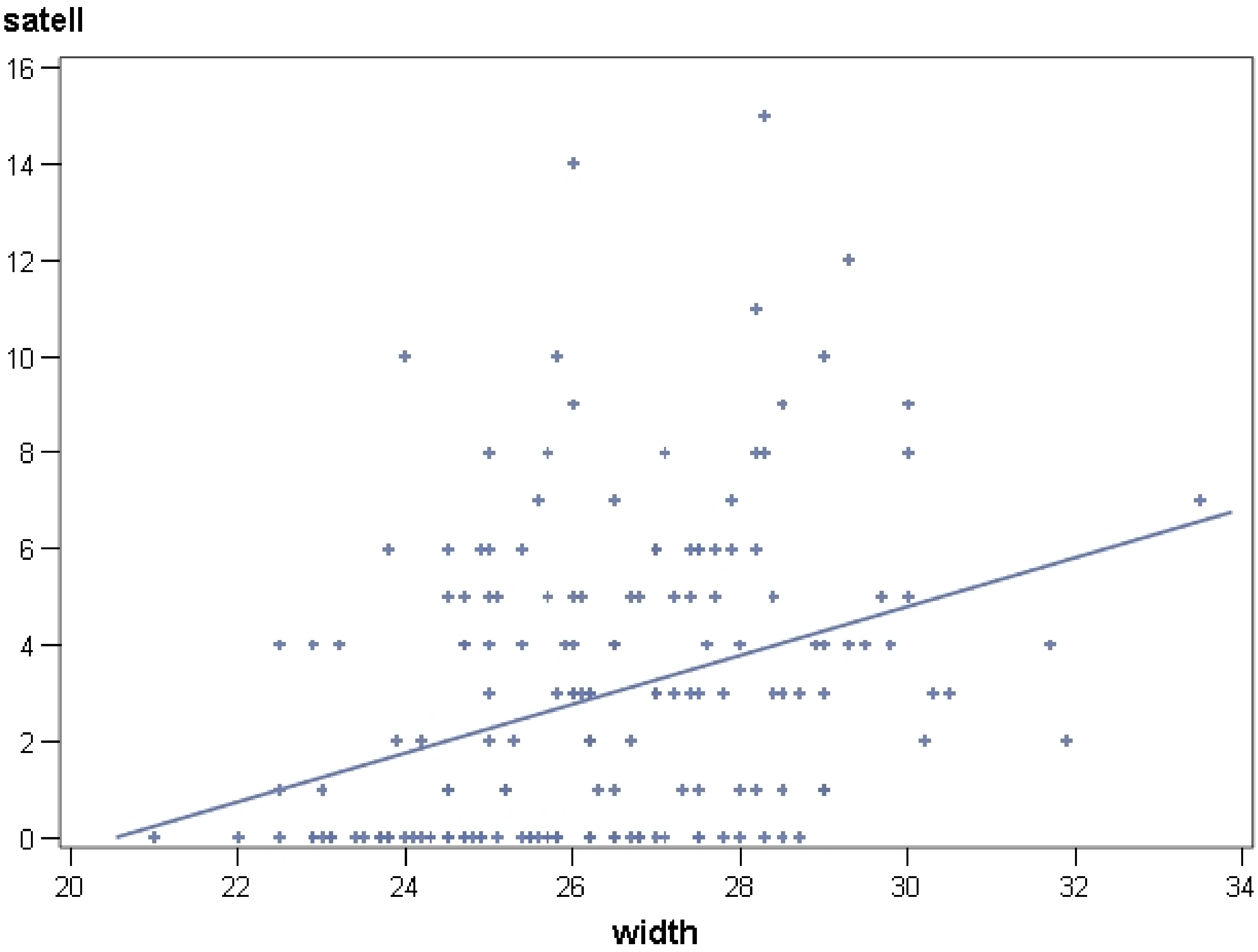
$$\begin{aligned} Q_L &= 2[L_{M_1} - L_{M_0}] = 2[L_S - L_{M_0}] - 2[L_S - L_{M_1}] \\ &= D(M_0) - D(M_1) \end{aligned}$$

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.

3.5 Analysis of Agresti's Crab Data

Agresti, Ch. 3, presents a data set from a study of nesting crabs. Each female in the study had a male crab accompanying her. Additional male crabs living near her are call *satellites*. The number satellites for each female crab is the response. Predictors include the color, spine condition, width, and weight of the female crab. The plot on the next page depicts the number of satellites as a function of carapace width. Since this plot does not reveal a clear trend, we group the crabs into width categories and plot the mean number of satellites for the female crabs within each category.




satell
0
2
4
6
8
10
12
14
16
20
22
24
26
28
30
32
34
width

The fitted model for the mean number of satellites for a female crab is

$$\log(\hat{\mu}) = \hat{\alpha} + \hat{\beta}_1 x = -3.3048 + 0.1640x.$$

The asymptotic standard error of $\hat{\beta}_1$ is $\widehat{se}(\hat{\beta}_1) = 0.0200$. The Wald chi-square statistic for testing $H_0 : \beta_1 = 0$ is $W = 67.51$ which gives strong evidence of an effect due to width on the mean number of satellites.

We can estimate the mean number of satellites for a female crab with a width of 30 cm:

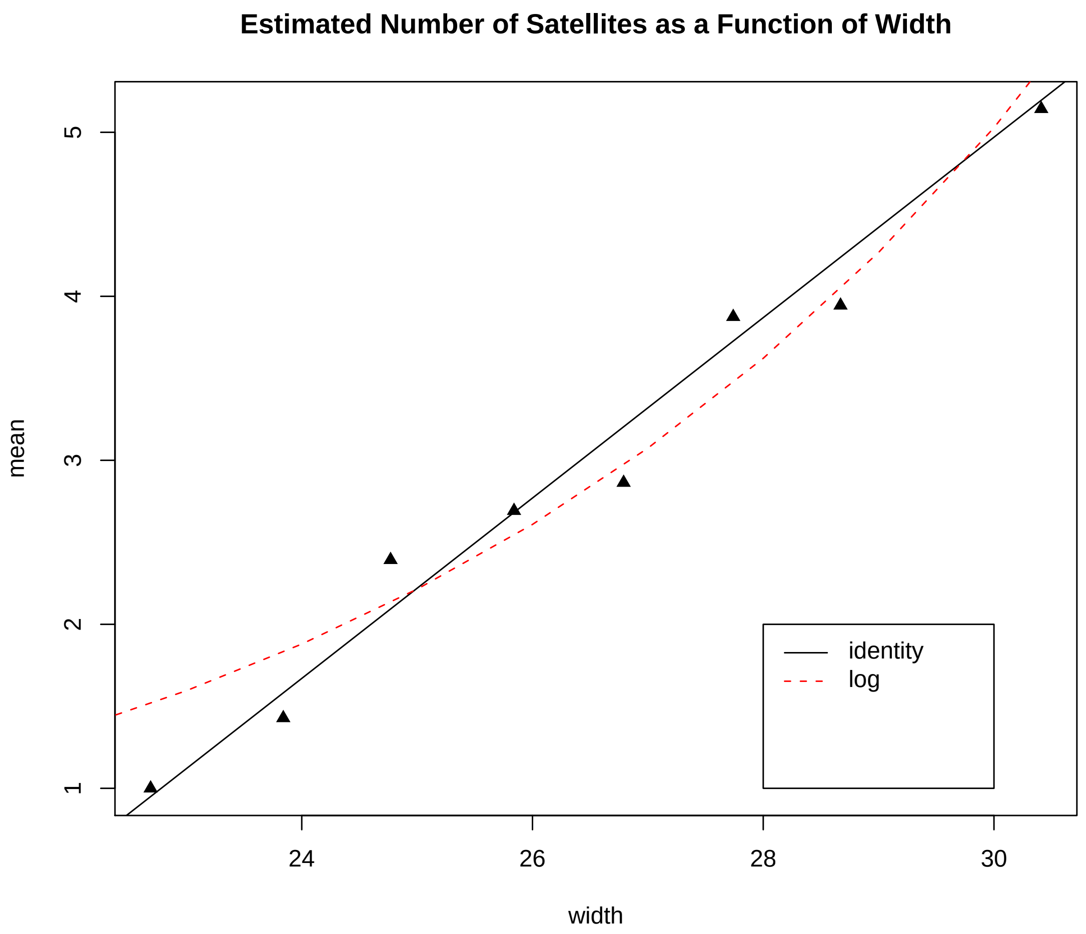
$$\hat{\mu} = \exp[\hat{\alpha} + \hat{\beta}_1 x] = \exp[-3.3048 + 0.1640(30)] = 5.03.$$

The effect of a 1cm increase in width is a multiplicative effect of $\exp(\hat{\beta}_1) = \exp(0.164) = 1.18$ on the mean number of satellites.

A Poisson regression model with identity link was also fit to the data resulting in a estimated mean response of

$$\hat{\mu} = \hat{\alpha} + \hat{\beta}_1 x = -11.5321 + 0.5495x.$$

The estimated mean number of satellites for a female crab with a width of 30 cm is $\hat{\mu} = -11.5321 + 0.5495(30) = 4.9529$ which is similar to the above fitted value.



We can set that both models produce similar estimates of the mean number of satellites over the middle part of the range of width, but the fit seems to be better for the identity link for small widths.

To illustrate the fitting of a Poisson regression model with an offset term we will fit the grouped horseshoe crab data. We consider the following variables:

- $Y =$ total number of satellites for females having a given width, w
- $k =$ number of female crabs having a given width w

If $\mu = E(Y)$, then μ/k is the expected number of satellites per female crab at that width. We fit the model

$$\log(\mu/k) = \alpha + \beta_1 w.$$

The Poisson regression model with an offset of $\log(k)$ results in

$$\log(\widehat{\mu/k}) = -3.5355 + 0.1727w$$

which is similar to the model we obtained from the complete data.

3.6 Model Checking for GLMs 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

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

- 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.

Statistics 659

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

- Deviance residual

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

- Pearson residual

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

See Simonoff, p. 133, for standardized versions of these residuals.

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

3.7 Checking a Poisson Regression Model

For the Poisson generalized linear model, the Pearson and the likelihood ratio statistics for goodness of fit are

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \qquad \text{and} \qquad G^2 = \sum y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right),$$

respectively. When the expected values $(\hat{\mu}_i)$ are large enough (≥ 5) and n is fixed, they have an approximate chi-squared distribution with $df = n - p$ where $p =$ the number of parameters in the model. We reject the Poisson regression model for large values of these statistics.

3.7.1 Examining the Fit of the Poisson Regression Model for the Crab Data

There are 66 distinct values of width for the 173 crabs. Each of these values has a total count of satellites y_i with fitted values $\hat{\mu}_i$. The above goodness-of-fit statistics were computed resulting in

$$X^2 = 174.27 \quad \text{and} \quad G^2 = 190.03 \quad \text{with } df = 64.$$

The validity of the large sample approximation using the chi-squared distribution is doubtful for a couple of reasons:

- Most of the expected frequencies are small.
- If we had more crabs in the sample, the number n of different settings would increase (not stay fixed).

A better chi-squared approximation can be obtained by grouping the data. The data were placed into the width categories: $< 23.25, 23.25 - 24.25, \ldots, 28.25 - 29.25, > 29.25$. This results in categories with $y_i, \hat{\mu}_i$ all larger than they were in the original 66 width categories. The goodness-of-fit statistics were computed resulting in

$$X^2 = 6.25 \quad \text{and} \quad G^2 = 6.52 \quad \text{with } df = 6.$$

This indicates no lack of fit for the Poisson regression model.

3.7.2 Residual Analysis

For the Poisson generalized linear model, the Pearson residual is

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

The Pearson residual divided by its standard deviation is called the *adjusted residual*:

$$\tilde{r}_i^P = \frac{r_i^P}{\sqrt{(1-h_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-h_i)}}.$$

The term h_i is called the *leverage* of an observation i. See p.148 of Agresti or p. 132 of Simonoff a discussion of the “hat matrix” and leverage.

Following is a table of fitted values and residuals for the grouped crab data:

Width	n_i	y_i	$\hat{\mu}_i$	r_i^P	$\tilde{r}_i^P$
< 23.25	14	14	20.5	-1.44	-1.63
$23.25 - 24.25$	14	20	25.2	-1.01	-1.11
$24.25 - 25.25$	28	67	58.9	1.06	1.23
$25.25 - 26.25$	39	105	98.6	0.64	0.75
$26.25 - 27.25$	22	63	65.5	-0.31	-0.34
$27.25 - 28.25$	24	93	84.3	0.95	1.06
$28.25 - 29.25$	18	71	74.2	-0.37	-0.42
> 29.25	14	72	77.9	-0.67	-1.00

3.8 A Brief Look at Overdispersion

An assumption for Poisson regression is that the mean and variance of the responses are equal. Heterogeneity of the experimental units can cause the variance to be larger than the mean. This can occur in models where one or more important predictors is omitted. In our example, suppose that the responses are Poisson with the mean depending on four variables: width, weight, color, and spine condition. If we consider only one of these predictors, say width, the crabs with a given width will have differing values of weight, color, and spine condition resulting in different means. This will result in a larger variance than the Poisson model predicts.

We can carry out tests for overdispersion as outlined in Section 5.3 of Simonoff. Here we will check for overdispersion in the crab data by computing the sample mean and variance of the number of satellites for the crabs in the various weight categories:

Width	n_i	y_i	$\bar{x}$	s_i^2
< 23.25	14	14	1.00	2.77
$23.25 - 24.25$	14	20	1.43	8.88
$24.25 - 25.25$	28	67	2.39	6.54
$25.25 - 26.25$	39	105	2.69	11.38
$26.25 - 27.25$	22	63	2.86	6.88
$27.25 - 28.25$	24	93	3.87	8.81
$28.25 - 29.25$	18	71	3.94	16.88
> 29.25	14	72	5.14	8.29

3.8.1 Zero-Inflated Poisson Model

The zero-inflated Poisson model allows for an excessive number of zero observations relative to the Poisson distribution. For this distribution,

$$E(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = \mu + \frac{\omega}{1-\omega}\mu^2$$

where ω is the probability of observing a zero and $1-\omega$ is the probability of the observation coming from a Poisson distribution.

3.8.2 Negative Binomial Model

The negative binomial model is often used for regression models for overdispersed data. For this distribution,

$$E(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = \mu + k\mu^2,$$

where k is the negative binomial dispersion parameter that must be estimated from the data. As $k \to 0$, the distribution approaches a Poisson distribution.

3.9 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 from a collection of possible models. Some of the more commonly used criteria are present below.

1. -2 Log-likelihood or deviance

   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 the AIC.

2. Akaike Information Criterion

$$\text{AIC} = -2L + 2\nu$$

   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 modification of the AIC that increases the protection against overfitting is the corrected AIC:

3. Corrected Akaike Information Criterion

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

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

$$\text{AIC}^* = \text{Deviance} - 2L_S + 2\nu$$

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

$$\text{AIC} = \text{Deviance} + 2\nu.$$

Similarly we can consider the corrected AIC criterion:

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

4. Schwarz Criterion — A Bayesian argument yields the Bayesian information criterion:

$$\text{BIC} = \text{Deviance} + \nu \log(n).$$

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.
  - A difference of less than 2 means that the models are essentially equivalent.
  - A difference of more than 10 means that the model with larger AIC has a much poorer fit.

3.10 Poisson Regression—Analysis of Tornado Data

We wish to model the number of deaths resulting from tornadoes as a function of the number of tornadoes, the number of killer tornadoes, the year, and the month. Since we wish to compare the AIC$_C$ among the models, we can set the lowest value to zero and put the difference in the table. The AIC$_C$ difference values are given in the table on the next page for several potential models:

The smallest AIC$_C$ was found for the model with all the predictors. On the other, the model with Killer tornadoes and Month has only a small increase in AIC$_C$ and seems to be a reasonable model with a small number of predictors.

Statistics 659

Predictors	ν	Deviance	AIC	AIC$_C$ Diff.
None	1	810.3	910.6	650.9
Killer tornadoes	2	275.4	377.7	118.0
Tornadoes	2	643.3	745.8	486.1
Year	4	727.8	834.0	574.5
Month	12	374.4	496.4	237.7
Killer tornadoes, Tornadoes	3	265.3	369.5	109.9
Killer tornadoes, Year	5	271.0	379.3	119.7
Killer tornadoes, Month	13	140.1	264.4	5.83
Tornadoes, Year	5	594.5	702.7	443.21
Tornadoes, Month	13	226.9	351.2	92.70
Year, Month	15	291.6	419.9	161.73
Killer torn., Tornadoes, Year	6	257	367.3	107.9
Killer torn., Tornadoes , Month	14	133.3	259.6	1.29
Killer torn., Year, Month	16	133.2	263.4	5.53
Tornadoes, Year, Month	16	191.2	321.5	63.6
Killer torn., Tornadoes, Year, Month	17	125.4	257.7	0.0