Chapter 3: Generalized Linear Models

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:

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

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.

Linear Probability Model

A simple model relating the success probability π to the explanatory variable x is a linear model:

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

Problems with this model:

- For certain x, $\pi(x)$ could be more than one or less than zero.
- Least squares is not optimal because $\text{var}(Y) = \pi(x)(1 - \pi(x))$.
- Maximum likelihood estimators do not have closed form.

Logistic Regression Model

In many cases the success probability is a nonlinear function of the linear predictor $\eta = \alpha + \beta x$. Typically these functions

- are *monotonic*. Means either they increase as x increases or decrease as x increases.
- satisfy $0 \leq \pi(x) = \pi(\eta) \leq 1$.
- often form an *S-shaped curve*.

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

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

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 of a random variable X is defined as

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

where f is the probability density function of X.

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.

Several choices for modelling π in terms of x

- Logistic distribution: The CDF of the logistic distribution is

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

Thus,

$$\pi(x) = \pi(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

- Normal distribution: The CDF of the standard normal random variable is

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

Thus,

$$\pi(x) = \Phi(\eta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\eta} \exp(-r^2/2) dr.$$

- Gumble distribution: The CDF of the standard Gumbel distribution is

$$F(s) = 1 - \exp\{-\exp(s)\} \qquad -\infty < s < \infty.$$

Thus,

$$\pi(x) = F(\eta) = 1 - \exp\{-\exp(\eta)\}.$$

Link function

The link function on the mean of the response returns the linear predictor.

- For the logistic model the link function is the logit link

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

- When $\pi(x) = \Phi(\eta) = (1/\sqrt{2\pi})\int_{-\infty}^{\eta} \exp(-r^2/2)dr$, the link is called probit link which is the inverse-CDF of the normal distribution.
- When $\pi(x) = F(\eta) = 1 - \exp\{-\exp(\eta)\}$, the link function is called complementary log-log (cloglog):
$\text{cloglog}\{\pi(x)\} = \log[-\log\{1-\pi(x)\}] = \eta = \alpha + \beta x$

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

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

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

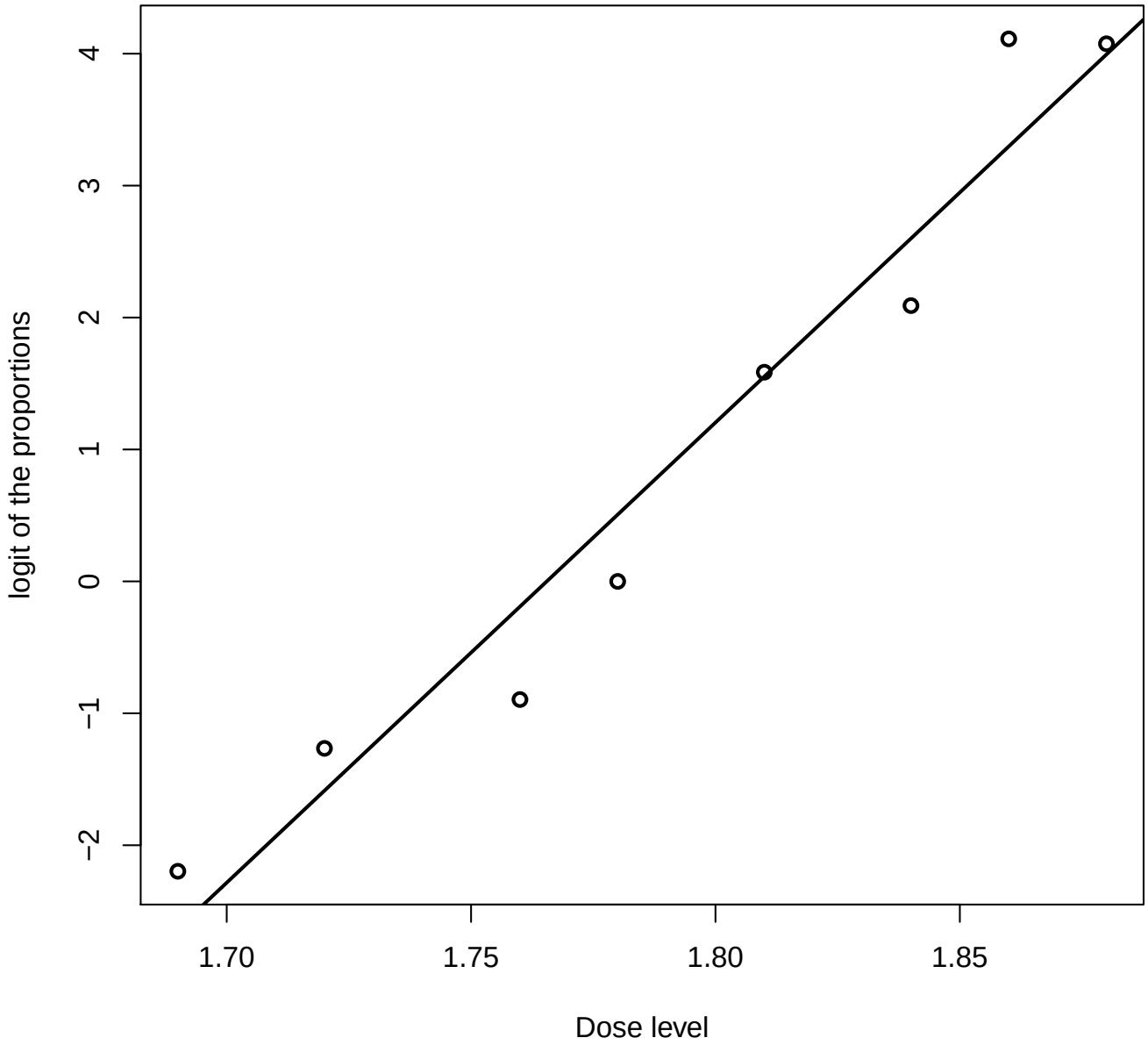
Code

```
x=c(1.69, 1.72, 1.76, 1.78, 1.81, 1.84, 1.86, 1.88)
y=c(0.1, 0.22, 0.29, 0.5, 0.83, 0.89, 0.9839, 0.9833)
y1=log(y/(1-y))
plot(x, y1, ylab="logit of the proportions", lwd=2)

lm(y1~x)
Call:
lm(formula = y1 ~ x)

Coefficients:
(Intercept)            x
     -61.58         34.88

 abline(a=-61.58, b=34.88, lwd=2)
```


logit of the proportions
4
3
2
1
0
−1
−2
1.70
1.75
1.80
1.85
Dose level

Code

```
> 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
```

Code

```
> 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
```

Three binary regression models were fit to these data. The fitted models were:

- $\text{logit}\{\hat{\pi}(x)\} = -59.19 + 33.40x$
- $\text{probit}\{\hat{\pi}(x)\} = -33.92 + 19.15x$
- $\text{cloglog}\{\hat{\pi}(x)\} = -36.90 + 20.53x$

Show the observed proportions and the fitted proportions for all three models.

The logistic model as the GLM

- Show that the probability mass function (pmf) of Y given X (or x) can be written

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

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

$$\mu = E(Y) = b'(\theta)$$
$$\text{var}(Y) = a(\phi)b''(\theta)$$

Connection with the linear model

The usual multiple regression model is

$$Y = \alpha + \beta X + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, and α, β, σ^2 are unknown parameters.

- Random component: $Y \sim N(\mu, \sigma^2)$
- Systematic component: $\alpha + \beta X$
- Link: $\alpha + \beta X = I(\mu) = \mu$, I stands for the identity link.

Examples of Generalized Linear Models

Random Component	Link	Systematic 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

Generalized Linear Models

The conditional probability distribution of the each of the above random variable Y given the explanatory variable X can be expressed as

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

where $\theta(X)$ is a function of X.

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 $\exp(\beta)$.

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 $Y_i \sim \text{Poisson}(\exp(x_i))$ 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.

Poisson data generation

Code

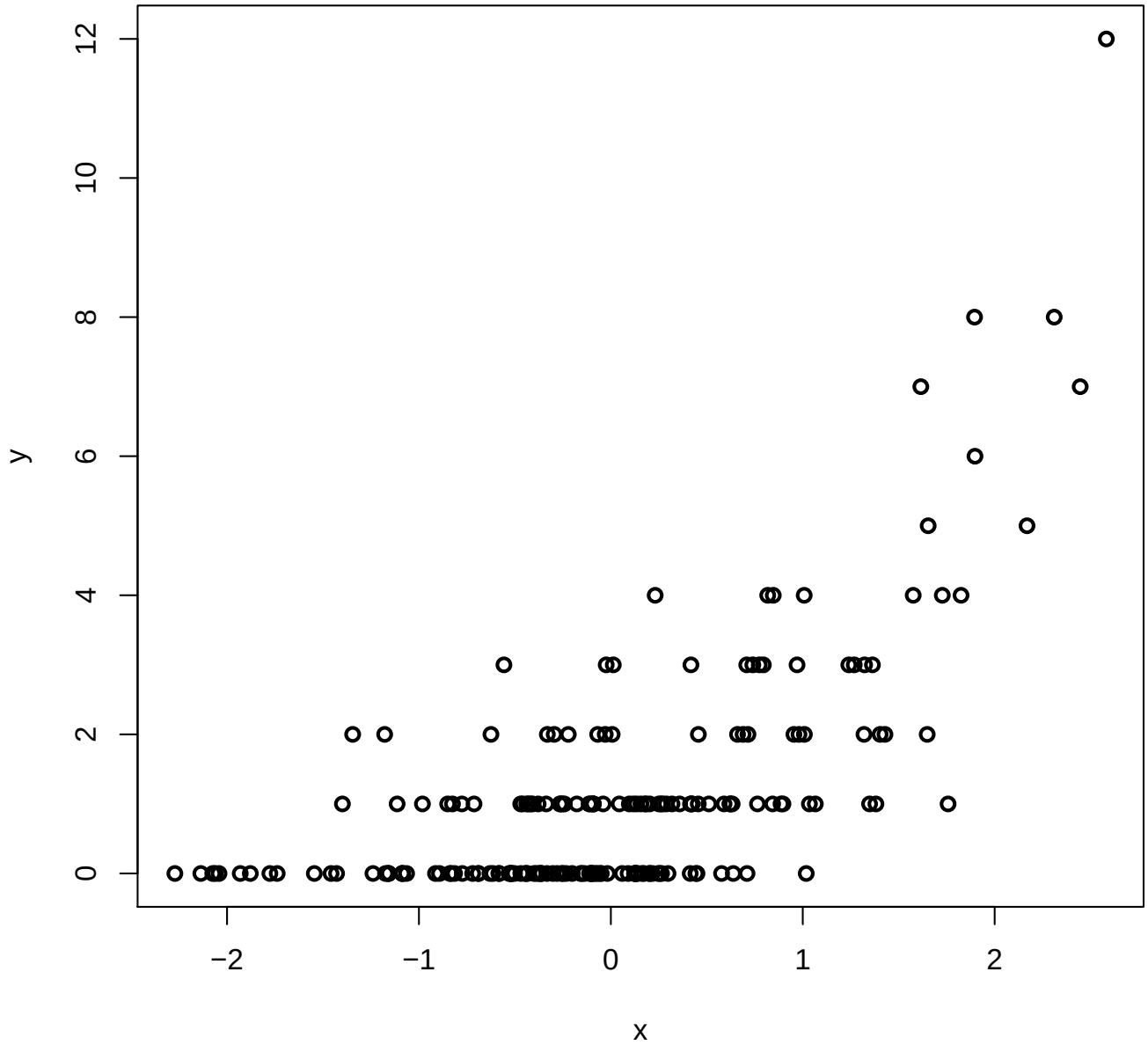
```
n=200
> set.seed(100)
> x=rnorm(n)
> mu=exp(-0.5+x)
> y=rpois(n, mu)
```

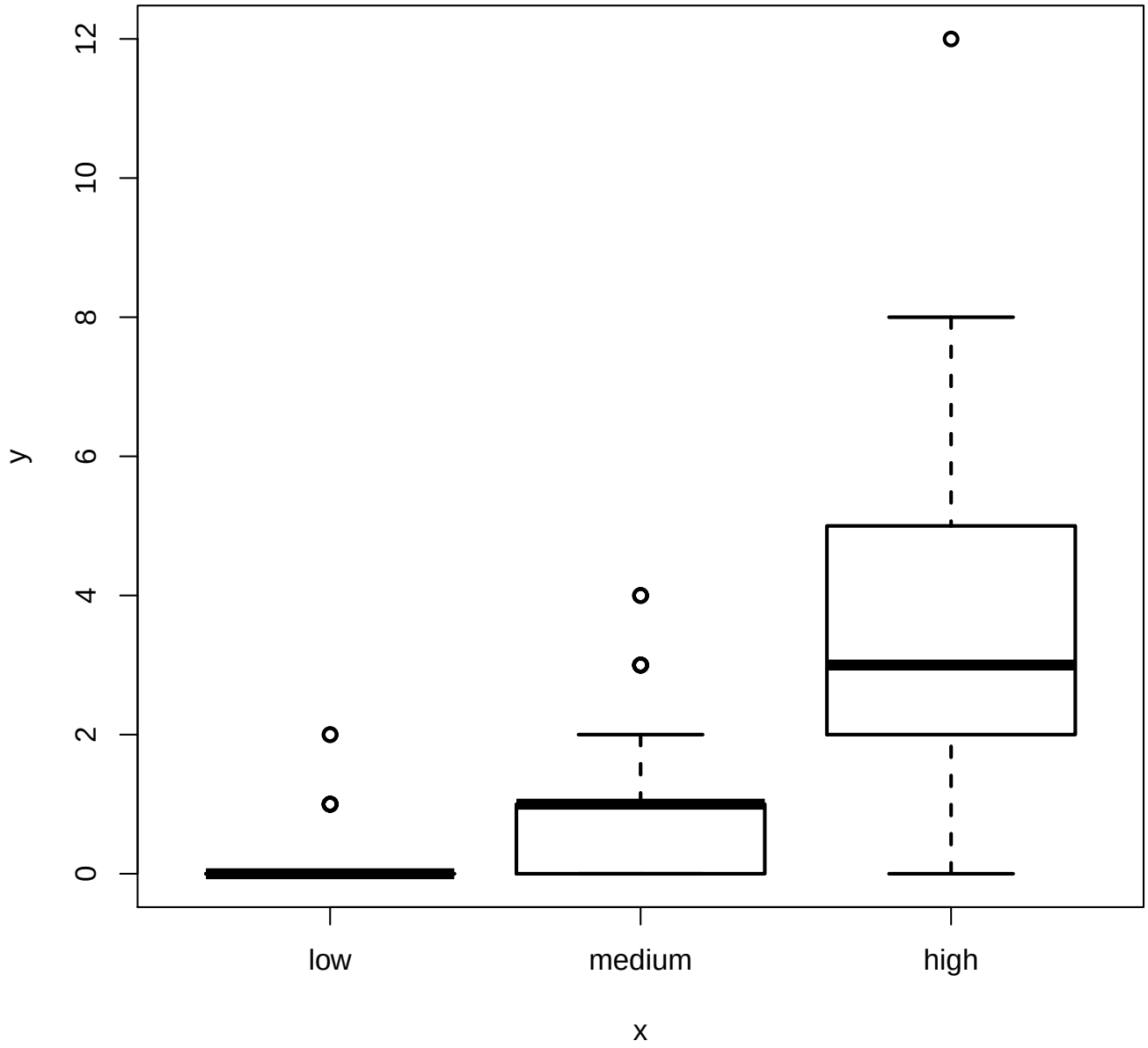
Look at the figures

Code

```
plot(x, y, lwd=2) # scatter plot

library(dplyr);
x.new <- cut(x, breaks=3, labels=c("low","medium","high"))
boxplot(y~x.new, lwd=2, xlab="x") # boxplot of y against a categorized x
```


y
0
2
4
6
8
10
12
−2
−1
0
1
2
x


y
0
2
4
6
8
10
12
low
medium
high
x

Poisson model fitting

Code

```
glm(y~x, family=``poisson''(link=log))

Call:  glm(formula = y ~ x, family = poisson(link = log))

Coefficients:
(Intercept)            x
    -0.3446       1.0102

Degrees of Freedom: 199 Total (i.e. Null);  198 Residual
Null Deviance:      404.6
Residual Deviance: 196.9  AIC: 461.4
```

Poisson Multiple Regression

We consider Poisson regression models with one or more explanatory variables.

- Poisson response: Y
- k predictors: $x = (x_1, \ldots, x_k)$
- Quantity to estimate: $\mu(x) = E(Y|x_1, \ldots, x_k)$

The Poisson regression model is

$$\log\{\mu(x)\} = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- The parameter β_j reflects the effect of a unit change in x_j on the log of the mean response, keeping the other x_is constant.
- Here, e^{β_j} is the multiplicative effect on the mean response of a one-unit increase in x_j, keeping the other x_is constant:

$$\begin{aligned}
&\log(\mu(x_1 + 1, x_2, \ldots, x_k)) - \log(\mu(x_1, x_2, \ldots, x_k)) \\
&= \alpha + \beta_1(x_1 + 1) + \cdots + \beta_k x_k - (\alpha + \beta_1 x_1 + \cdots + \beta_k x_k) \\
&= \beta_1
\end{aligned}$$

Poisson Regression Models with Qualitative Predictors

Similar to ordinary regression, we can have qualitative explanatory variables. Consider first a simple model

$$\log(\mu(x)) = \alpha + \beta x,$$

where the predictor x which takes on the values 0 or 1. We obtain the following table:

Explanatory Variable (X)	$\log(\mu(x))$	$\mu(x)$
$x = 1$	$\log(\mu(1)) = \alpha + \beta$	$\exp(\alpha + \beta)$
$x = 0$	$\log(\mu(0)) = \alpha$	$\exp(\alpha)$

We often call such a predictor a dummy variable.

When a nominal predictor has k levels, we can represent the predictor in a regression model using $k-1$ dummy variables:

$$x_\ell = 1, \text{ if Category } \ell, \quad = 0, \text{ otherwise}, \quad \ell = 1, ..., k-1.$$

The resulting Poisson regression model is

$$\log(\mu(x)) = \alpha + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}.$$

This parameterization treats Category k as a reference category. Since

$$\log\left(\frac{\mu_\ell}{\mu_k}\right) = \log(\mu_\ell) - \log(\mu_k) = (\alpha + \beta_\ell) - \alpha = \beta_\ell,$$

the ratio of the mean response for category ℓ and that of the reference category k is

$$\frac{\mu_\ell}{\mu_k} = \exp(\beta_\ell).$$

Poisson Regression for Rate Data

It is often the case that the response of interest is the rate of occurrence of some event rather than the number of occurrences 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 log-linear 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

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

The following webpage has a nice data to illustrate the offset term. `https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot`

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:

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

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). We will use statistical software for categorical data to compute mles.

The MLE has excellent large sample properties under certain regularity conditions as $n \to \infty$.

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

The MLE is asymptotically normal:

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

The above statement says that for a large n, the distribution of $\frac{\hat{\theta}-\theta_0}{\sqrt{\mathcal{I}(\theta_0)^{-1}}}$ is close to $Normal(0,1)$.

The quantity $\mathcal{I}(\theta)$ is called Fisher's information and is defined as

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

This quantity can be estimated in several ways:

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

Maximum Likelihood Estimation

We replace $\mathcal{I}(\theta_0)^{-1}$ 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 $\hat{\theta}$.
- $-\dfrac{\partial^2 \ell(\hat{\theta})}{\partial \theta^2}$ measures the *curvature* of the log-likelihood function.
- The greater the curvature, the greater the information about θ and the smaller the SE.

Maximum Likelihood Estimation

The log-likelihood for the binomial distribution is

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

The first and second derivatives are

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta} \quad \text{and} \quad \frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}.$$

The mle is $\hat{\theta} = \frac{y}{n}$, and Fisher's information is

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

Then

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

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.

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

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

Wald Test

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

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

We define the *Wald statistic*:

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\mathcal{I}(\hat{\theta})^{-1}}} \sim N(0,1) \qquad \text{or} \qquad W = Z^2 = \frac{(\hat{\theta} - \theta_0)^2}{\mathcal{I}(\hat{\theta})^{-1}} \sim \chi^2_1.$$

Binomial(n, θ) example:

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

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*:

$$G^2 = Q_L = -2 \log \left\{ \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\hat{\theta})} \right\} = 2\{\ell(\hat{\theta}) - \ell(\theta_0)\} \xrightarrow{d} \chi^2_1 \text{ as } n \to \infty$$

- $\ell(\theta) = \log\{\mathcal{L}(\theta)\}$
- 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.

Let $Y \sim \text{Binomial}(n, \theta)$, y is the observed (realized) value of the random variable Y

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

Score Test

The score function is defined as

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

Recall that the mle is the solution to

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

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:

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

Bernoulli random sample

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta} \quad \text{and} \quad S = \frac{\left(\frac{y}{\theta_0} - \frac{n-y}{1-\theta_0}\right)^2}{\frac{n}{\theta_0(1-\theta_0)}} = \frac{n(\hat{\theta}-\theta_0)^2}{\theta_0(1-\theta_0)}$$

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 $\theta = (\theta_1, \ldots, \theta_p)^T$. We wish to test all components of θ together, $H_0 : \theta = \theta_0$. Then
  - The score function is now a vector of p partial derivatives of the log-likelihood function.
  - The MLE is determined by solving the resulting set of p equations in p unknowns.
  - Fisher's information is now a $p \times p$ matrix.
  - All three statistics are asymptotically equivalent and asymptotically have a chi-squared distribution with p d.f.

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

$$\text{Deviance} = D(M) = 2(\ell_S - \ell_M)$$

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 (M_0) holds is

$$\begin{aligned} Q_L &= 2(\ell_{M_1} - \ell_{M_0}) = 2(\ell_S - \ell_{M_0}) - 2(\ell_S - \ell_{M_1}) \\ &= D(M_0) - D(M_1) \end{aligned}$$

Deviance

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.

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. Since the plot of the number of satellites against the carapace width does not reveal a clear trend, one can group the crabs into width categories and plot the mean number of satellites for the female crabs within each category.

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

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

The asymptotic standard error of $\hat{\beta}$ is $\widehat{se}(\hat{\beta}) = 0.0200$. 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:

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

The effect of a 1cm increase in width is a multiplicative effect of $\exp(\hat{\beta}) = \exp(0.164) = 1.18$ 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

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

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 in a width category
- k = number of female crabs having in the width category
- w = the average width of the female crabs in the width category

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

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

The Poisson regression model with an offset of $\log(k)$ results in the same fit as using the ungrouped data.
When we group the data by width categories, we obtain the fitted model,

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

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

Poisson model as a GLM

$$\begin{aligned} f(y;\mu) &= \exp(-\mu)\frac{\mu^y}{y!} \\ &= \exp\{-\mu + y\log(\mu) - \log(y!)\} \\ &= \exp\{y\theta - \exp(\theta) - \log(y!)\}, \end{aligned}$$

where $\theta = \log(\mu)$ so $\mu = \exp(\theta)$. So, in the exponential family form $b(\theta) = \exp(\theta)$, $\phi = 1$, $c(y, \phi) = -\log(y!)$.

Also, observe that $E(Y) = b'(\theta) = \exp(\theta) = \mu$, and $\text{var}(Y) = b''(\theta) = \exp(\theta)$.

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

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

Here $\hat{\theta}_i^S = \log(Y_i)$, $\hat{\theta}_i^M = \log(\hat{\mu}_i^M)$, if $\mu_i^M = \exp(\alpha + \beta X_i)$, then $\hat{\theta}_i^M = \hat{\alpha} + \hat{\beta} X_i$

Model Checking for GLMs Using Residuals

- 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

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

  where $\hat{\mu}_i$ we simply refer to $\hat{\mu}_i^M$

- 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.

Checking a Poisson Regression Model

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

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

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.

For the Poisson regression model, $G^2 =$ Residual deviance

- When the null hypothesis holds and the approximation is adequate, the expected values of G^2 and X^2 both equal approximately $df = n - p$. Thus, when the model fits, the ratio of deviance (or chi-squared) ratio to degrees of freedom will be close to one.
- In many cases, there are too few observations per value of the predictor(s) to ensure the adequacy of the chi-squared approximation. In such a case, we use large values of the ratio of deviance to degrees of freedom to indicate possible inadequacy of the model.
- The value of the deviance and its degrees of freedom will depend on how we define a single observation. This will be discussed in more detail when we test the fit of logistic regression models.

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

We could consider the fit in two ways:

- There are 173 crabs. The resulting goodness-of-fit statistics when treating these as separate observations are

$$X^2 = 544.16, \; X^2/df = 3.18 \quad \text{and}$$
$$G^2 = 567.88, \; G^2/df = 3.32 \quad \text{with } df = 64.$$

- 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

$$X^2 = 174.27, \; X^2/df = 2.72 \quad \text{and}$$
$$G^2 = 190.03, \; G^2/df = 2.97 \quad \text{with } df = 64.$$

Code for the two model fitting

Code

```
ac=read.csv("/Users/Samiran/Downloads/agresti_crab.csv")
out1=glm(satell~width, family=poisson, data=ac)
out1
Call:  glm(formula = satell ~ width, family = poisson, data = ac)

Coefficients:
(Intercept)        width
     -3.305        0.164

Degrees of Freedom: 172 Total (i.e. Null);  171 Residual
Null Deviance:      632.8
Residual Deviance: 567.9  AIC: 927.2
```

Code for the two model fitting

Code

```
newwidth= unique(ac$width)
newy=rep(0, length(newwidth))
for( i in 1: length(newy)) newy[i]= sum(ac$satell[ac$width==newwidth[i]])
newn=rep(0, length(newwidth))
for( i in 1: length(newy)) newn[i]= sum(as.numeric(ac$width==newwidth[i]))
out2=glm(newy~newwidth, family=poisson, offset=log(newn))
out2

Call:  glm(formula = newy ~ newwidth, family = poisson, offset = log(newn))

Coefficients:
(Intercept)     newwidth
     -3.305        0.164

Degrees of Freedom: 65 Total (i.e. Null);  64 Residual
Null Deviance:      254.9
Residual Deviance: 190  AIC: 402.5
```

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).

Code

```
data2=data.frame(newy, newn, newwidth)
head(data2)
  newy newn newwidth
1   23    3     28.3
2    5    3     22.5
3   38    6     26.0
4    0    1     24.8
5    6    3     23.8
6   16    6     26.5
```

A better chi-squared approximation can be obtained by grouping the data. The data were placed into the width categories:
$< 23.25, 23.25 - 24.25, \ldots, 28.25 - 29.25, > 29.25$. The resulting data is given in a table on the next slide. This results in categories with y_i and $\hat{\mu}_i$ much larger than they were in the original 66 width categories.

Code

```
newy2=c(14, 20, 67, 105, 63, 93, 71, 72)
newn2=c(14, 14, 28, 39, 22, 24, 18, 14)
newx2=0:7; out4=glm(newy2~newx2, family=poisson, offset=log(newn2))
out4
Call:  glm(formula = newy2 ~ newx2, family = poisson, offset = log(newn2))
Coefficients:
(Intercept)        newx2
     0.3640       0.1845
Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     72.38
Residual Deviance: 5.996  AIC: 56.44
```

The goodness-of-fit statistics were computed resulting in

$$X^2 = 5.78,\ X^2/df = 0.963 \quad \text{and} \quad G^2 = 5.996,\ G^2/df = 0.999, \quad \text{with } df = 6.$$

This indicates no lack of fit for the Poisson regression model. However, this analysis fails to indicate the presence of overdispersion.

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.

This table contains the 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

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	$\bar{y}$	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

Negative Binomial Model

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

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

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.

For handling extra zeros, one may consider the zero-inflated negative binomial distribution.

Further materials on Negative Binomial

The negative binomial distribution (NegBinom(r, p)):

$$f(Y = y; p, r) = \binom{y + r - 1}{y}(1 - p)^r p^y, \ y = 0, 1, 2, 3, \ldots,$$

where $p \in (0, 1)$ while r is a positive number. Sometime the probability mass function is written as

$$f(Y = y; p, r) = \frac{\Gamma(y + r)}{y!\Gamma(r)}(1 - p)^r p^y, \ y = 0, 1, 2, 3, \ldots,$$

The mean and variance are

$$\mu = E(Y) = \frac{pr}{1 - p}, \ \text{var}(Y) = \frac{pr}{(1 - p)^2}.$$

Clearly, $\text{var}(Y) = \mu + \mu^2/r$.

Data generation from a negative binomial distribution

Code

```
# Generates 10 random numbers from NegBinom(r=2, p=0.2)
set.seed(10)
rnbinom(10, 2, 0.2)
[1]  5  6  3  5  7 11 14  4 16 17

# r need not be a positive integer, here data are generated from
# NegBinom(r=4.21, p=0.1)
set.seed(10)
rnbinom(10, 4.21, 0.10)
[1] 32 11 32 28 36 58 29 57 29 22
```

Negative binomial and poisson connection

The negative binomial distribution can be obtained when the mean of the poisson distribution follows a Gamma distribution.
Suppose that conditional on λ, $Y \sim \text{Poisson}(\lambda)$ and λ follows a Gamma distribution with the shape parameter r and scale parameter $p/1(1-p)$ then the marginal distribution of Y is the NegBinom(r, p). That means when

$$f(y;\lambda) = \exp(-\lambda)\frac{\lambda^y}{y!},\ f(\lambda; r, p) = \exp\{-\lambda(1-p)/p\}\frac{\lambda^{(r-1)}\{(1-p)/p\}^r}{\Gamma(r)},$$

we obtain $Y \sim \text{NegBinom}(r, p)$.

negative binomial model fitting

Code

```
# Generates 30 random numbers from NegBinom(r=2.2, p=0.2)
set.seed(10)
y=rnbinom(30, 2.2, 0.4)

# Fit a NegBinom distribution to the data and obtain the MLE
library(MASS)
out=fitdistr(y, "Negative Binomial")

out
     size          mu
  2.7689726   2.9000031
 (1.4352723) (0.4448687)
# Here size=r
# Need some discussion about estimating p, in R, mu=r(1-p)/p
out2=fitdistr(y, "Poisson")
# Compare these two through AIC
c(AIC(out), AIC(out2))
[1] 132.1619 140.3683
```

Negative binomial continues

In the presence of explanatory variables, we intend to find how the mean of the response is governed by the explanatory variables. In the negative binomial case we write

$$\log(\mu) = \beta_0 + \beta_1^T X.$$

Thus, like the poisson model here the link function between the linear predictor and the mean is the log link. For the ease of understanding we write the Negative Binomial model in terms of μ and r.

$$f(y; \mu, r) = \frac{\Gamma(y + r)}{y!\Gamma(r)} \frac{\mu^y r^r}{(\mu + r)^{(\mu+r)}}.$$

In data generation, we can also specify the size r and mean μ in lieu of r and p.

Another example

Code

```
set.seed(10)
> x=rnorm(30)
> lambda= exp(0.2+x*1)
> lambda=lambda*exp(-0.2+rnorm(30))
> y=rpois(30, lambda)
> out2=glm.nb(y~x)
> summary(out2)
Call:
glm.nb(formula = y ~ x, init.theta = 0.4451546934, link = log)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2376  -1.0554  -0.8485   0.2082   1.6750
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1758     0.3434   0.512    0.609
x             0.5812     0.4083   1.423    0.155
(Dispersion parameter for Negative Binomial(0.4452) family taken to be 1)
```

Another example

Code

```
    Null deviance: 28.668  on 29  degrees of freedom
Residual deviance: 25.409  on 28  degrees of freedom
AIC: 87.613

Number of Fisher Scoring iterations: 1


             Theta:  0.445
         Std. Err.:  0.229

 2 x log-likelihood:  -81.613
 # Theta is 1/r according to our notation
```

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,

$$\text{pr}(Y = y) = \begin{cases} \omega + (1-\omega)\exp(-\lambda) & \text{if } y = 0, \\ (1-\omega)\exp(-\lambda)\lambda^y/y! & \text{for } y > 0 \end{cases}$$

Thus the random variable stems from a mixture of two distributions, 1) that has probability mass fully concentrated on zero and 2) the other is the poisson distribution. The mixing weights are ω and $1-\omega$ respectively. If an observation is non-zero, that is sure to come from the Poisson distribution while a zero value of Y may come from either distributions. This model has two parameters (λ, ω), $\lambda > 0$ and $0 < \omega < 1$.

Zero-Inflated Poisson (ZIP) Model

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

where $\mu = (1-\omega)\lambda$. It is obvious that here $\text{var}(Y) > E(Y)$ for $\omega > 0$, and $\text{var}(Y) = E(Y)$ when $\omega = 0$ (i.e., Y is a simple poisson random variable).
If you are interested in knowing more about zero-inflation and its application, I would suggest you to read the following article [1]

[1] Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation by Xie at al. (2013).

Zero-Inflated Poisson Model

Like the GLM, we can model the mean of the second component of the ZIP distribution in terms of covariates X and also we can model the inflation parameter ω in terms of X. The standard models are

$$\begin{aligned} \lambda(x) &= \exp(\beta_0 + \beta_1^T X), \\ \omega(X) &= \frac{\exp(\gamma_0 + \gamma_1^T X)}{1 + \exp(\gamma_0 + \gamma_1^T X)}. \end{aligned}$$

These model parameters can be estimated via the maximum likelihood method.

ZIP model fitting

Code

```
 # simulated data example
 set.seed(20)
 n=200
 x=rnorm(n)
 lambda=exp(0.2+1.2*x)
 y=rpois(n, lambda)
 b=rbinom(n, 1, 0.8)
 y=y*b  # zero-inflated response
 mydata=data.frame(x, y)
#
barplot(table(mydata$y))
```

ZIP model fitting

Code

```
 #you need the following package
 library(pscl)
# Regular poisson model fitting
 out.p=glm(y~x, family=poisson, data=mydata)
# ZIP model fitting with a model for the inflation parameter
out.zip.m.infl=zeroinfl(y~x|x,  data=mydata)
# ZIP model fitting without any model for the inflation parameter
out.zip.nm.infl=zeroinfl(y~x|1,  data=mydata)
 #
```

R code for model comparisons

Code

```
# Compare the models via AIC (available under the base package)
c(AIC(out.p), AIC(out.zip.m.infl), AIC(out.zip.nm.infl))
[1] 635.7601 595.4074 594.7901
# Compare the models via BIC (available under the base package)
c(BIC(out.p), BIC(out.zip.m.infl), BIC(out.zip.nm.infl))
[1] 642.3567 608.6007 604.6851
# Compare the models via BIC (need the qpcR package)
library(qpcR)
c(AICc(out.p), AICc(out.zip.m.infl), AICc(out.zip.nm.infl))
[1] 635.8210 595.6125 594.9126
```

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 that need not be nested. Some of the more commonly used criteria are presented 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.

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 that has a increased protection against overfitting:

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

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.