Block 14

Binomial and Poisson Regression

Binomial and Poisson Regression

Distributions for Counted Data

Bernoulli Distribution

Suppose the random variable Y has two possible values, perhaps called "success" or "failure", with probability of success equal to θ, where $0 \leq \theta \leq 1$. We label $Y = 1$ if success occurs, and $Y = 0$ otherwise. We will say that Y has a Bernoulli distribution with probability of success θ:

$$E(Y) = \theta, \quad \text{Var}(Y) = \theta(1 - \theta).$$

An important feature of the Bernoulli distribution is that the variance depends on the mean.

Binomial Distribution

The binomial distribution generalizes the Bernoulli. Suppose we have m random variables $B_1, B_2, \ldots, B_m$, such that

1. Each B_j has a Bernoulli distribution with the same probability θ of success, and
2. All the B_j's are independent.

Then if Y is the number of successes in the m trials, $Y = \sum B_j$, we say that Y has a binomial distribution with m trials and probability of success θ. The probability mass function is

$$P(Y = j) = \binom{m}{j} \theta^j (1 - \theta)^{m-j},$$

for $j \in \{0, 1, \ldots, m\}$. The mean and variance of the distribution are

$$E(Y) = m\theta, \quad \text{Var}(Y) = m\theta(1 - \theta).$$

Poisson Distribution

The Poisson distribution is the number of events of a specific type that occur in a fixed time or space. A Poisson variable Y can take the value of any non-negative integer $\{0, 1, 2, \dots\}$. The probability mass function of the Poisson distribution is given by:

$$P(Y = y) = \frac{\exp(-\lambda)\lambda^y}{y!}, \quad y = 0, 1, 2, \dots$$

The mean and variance are given by:

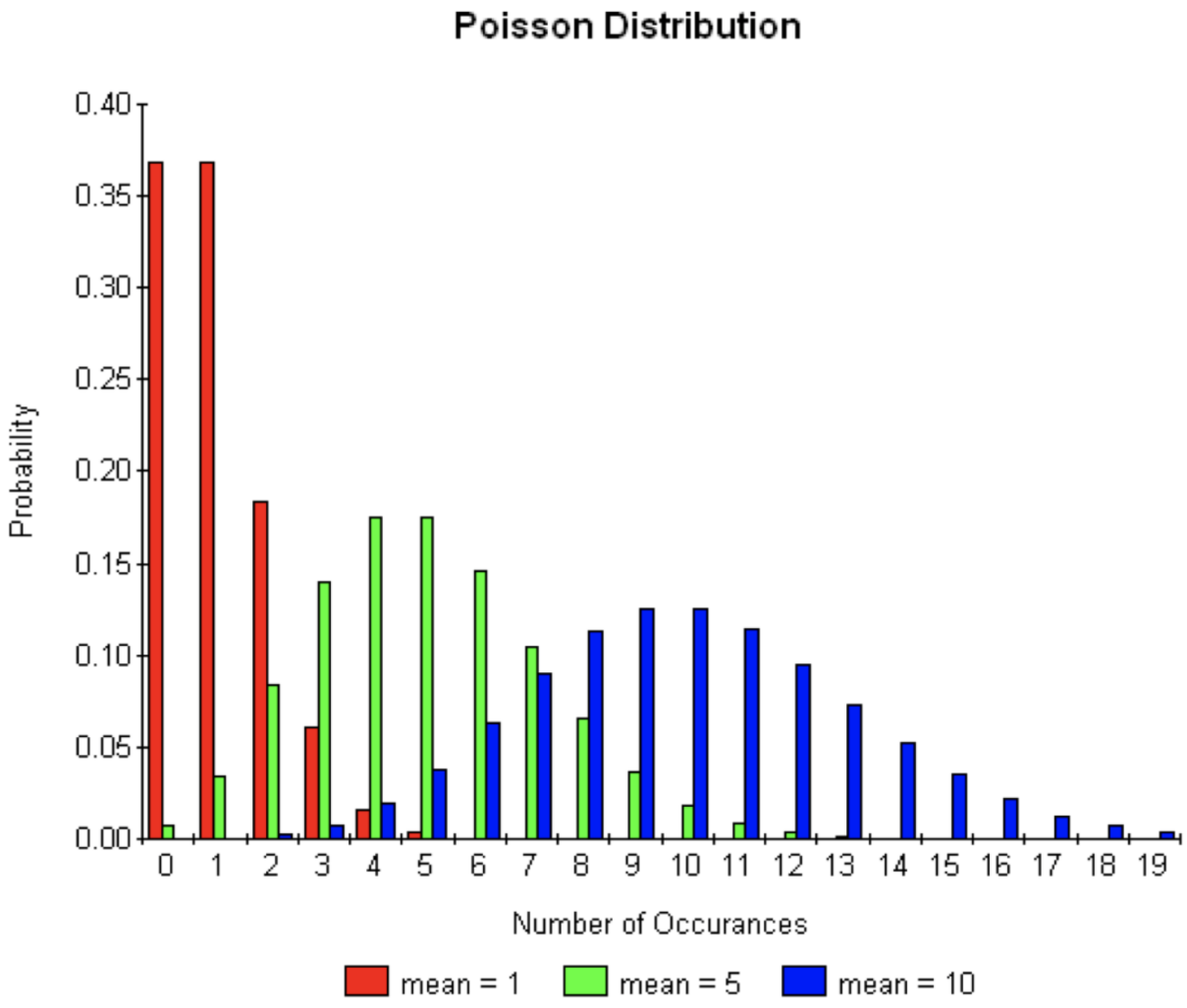
$$E(Y) = \lambda, \quad \text{Var}(Y) = \lambda.$$

Example: Occurrence of Anencephaly in Edinburgh 1956–66
Anencephaly is a serious disorder that causes the brain of a fetus not to develop properly. The number of children born with anencephaly in Edinburgh in the 132 months from 1955 to 1966 were:

#anencephaly	0	1	2	3	4	5	6	7	8	9+
#months	18	42	34	18	11	6	0	2	1	0

The Poisson distribution arises as:

- An approximation to the distribution of $Y \sim \text{Bin}(n, p)$ when p is small and n is large ($\lambda = np$).
- From a Poisson process.



Poisson Approximation to the Binomial Distribution

When n is large and p is small, we have $\lambda = np$,

$$\binom{n}{y} p^y (1-p)^{n-y} \approx \frac{\lambda^y e^{-\lambda}}{y!}.$$

Illustration

x	Poisson	Binomial	Binomial	Binomial
x	$\lambda = 0.5$	$n = 500, p = 0.001$	$n = 50, p = 0.01$	$n = 5, p = 0.1$
0	0.6065	0.6064	0.6050	0.5905
1	0.3033	0.3035	0.3056	0.3280
2	0.0758	0.0758	0.0756	0.0729
3	0.0126	0.0126	0.0122	0.0081
4	0.0016	0.0016	0.0015	0.0005

The Poisson distribution is often an appropriate model for "rare events."

Poisson Process

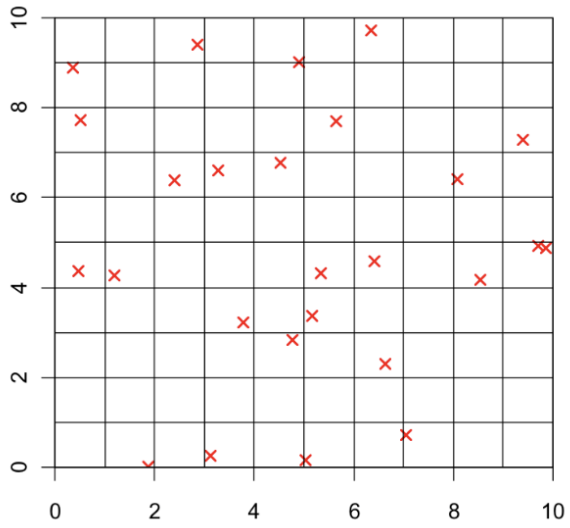
We are observing events (marked by x) happening over time:

Assume That:

- The rate of events λ is constant over time (rate = expected number of events per unit of time).
- The number of events in disjoint time intervals are independent.
- Events do not occur together.

Then we have a **Poisson process**.
The Poisson process is an appropriate model for events that are happening "randomly over time."
Let Y be the number of events in an interval of length t,

Then: $Y \sim \text{Poisson}(\lambda t)$.
In a similar manner, we may have a Poisson process in the plane:



Assume that:

- the rate of points λ is constant over the region (rate = expected number of points in an area of size one)
- the number of points in disjoint areas are independent
- points do not coincide

Then we have a **Poisson process** in the plane (spatial process). This is a model for "randomly occurring" points.
Let Y be the number of events in an area of size a,
Then: $Y \sim \text{Poisson}(\lambda a)$.

Regression Model for Counts

The big idea is that the parameter for the counted distribution, θ for the binomial or λ for the Poisson, can depend on the values of predictors.

Binomial Regression

We assume that $\theta(x)$ depends on the values x of the regressors only through a linear combination $\beta' x$ for some unknown β:

$$\theta(x) = m(\beta' x),$$

where $\beta' x$ is called the linear predictor.

Many choices for m

First: why not use identity function?

Because $0 \leq \theta(x) \leq 1$

Nothing on $\theta(x) = \beta' x$ forces that to happen for all β and x.

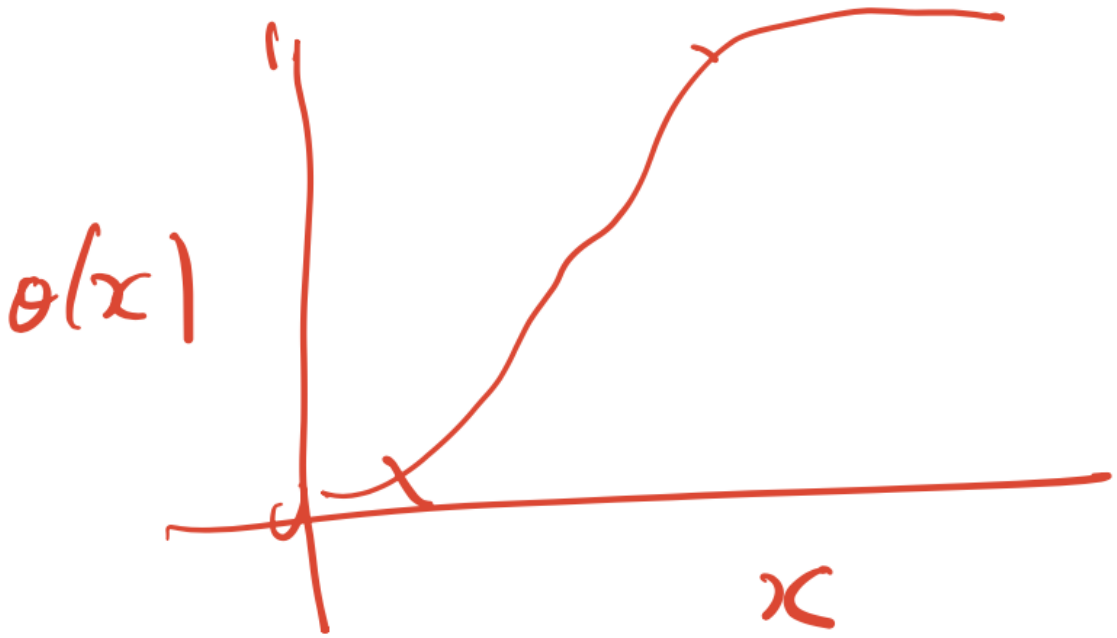
Fixes

Model $\theta(x)$ as a cdf:

$$\theta(x) = \int_{-\infty}^{x} f(s)\, ds$$

where $f(s) \geq 0$ and $\int_{-\infty}^{\infty} f(s)\, ds = 1$.

$f(s)$ is called the tolerance distribution.



Consider specific examples (one predictor)

a) $f(s)$ is $N(\mu, \sigma^2)$

$$\therefore \theta(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right) ds$$

$$= \Phi\left(\frac{x-\mu}{\sigma}\right)$$

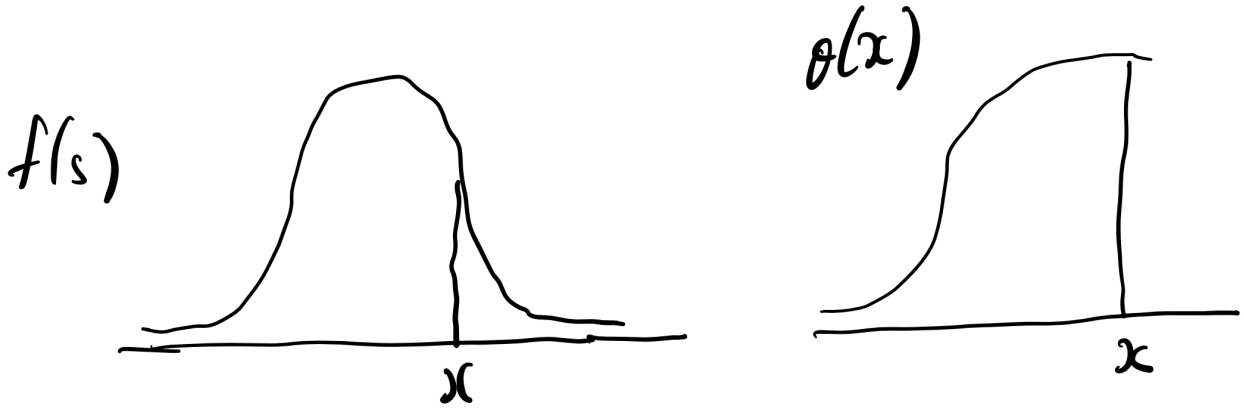
$$\therefore \Phi^{-1}(\theta(x)) = \beta_0 + \beta_1 x$$

where

$$\beta_0 = -\frac{\mu}{\sigma}, \quad \beta_1 = \frac{1}{\sigma}.$$

Called probit model

- Very popular in some areas of biology.
- For example, $x = \mu$ is called LD50 because it corresponds to the dose that kills half of the animals.



b) Consider

$$f(s) = \beta_1 \frac{\exp(\beta_0 + \beta_1 s)}{[1 + \exp(\beta_0 + \beta_1 s)]^2}$$

logistic dist

$$\therefore \theta(x) = \int_{-\infty}^{x} f(s)\, ds$$

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

$$\therefore m^{-1}(\theta(x)) = \log\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0 + \beta_1 x$$

logit fct ratio $\equiv$ odds

c) Consider

$$f(s) = \beta_0 \exp\left[\beta_0 + \beta_1 s - \exp(\beta_0 + \beta_1 s)\right]$$

extreme value dist

$$\therefore \theta(x) = 1 - \exp\left[-\exp(\beta_0 + \beta_1 x)\right]$$

$$\therefore m^{-1}(\theta(x)) = \log\left(-\log\left(1 - \theta(x)\right)\right) = \beta_0 + \beta_1 x$$

clog log

$$glm(\text{family} = \text{"binomial"}, \text{link} = \text{"cloglog"})$$

Most common is logistic regression

Logistic regression models are not fit with OLS. Rather, maximum likelihood estimation is used, based on the binomial distribution.

Likelihood function

$$L(\beta_0, \beta_{iy}) = \prod_{i=1}^{n} \theta(x_i)^{y_i}(1 - \theta(x_i))^{1-y_i}$$

Taking the log-likelihood:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \left[y_i \log \theta(x_i) + (1 - y_i) \log(1 - \theta(x_i)) \right]$$

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \log(1 - \theta(x_i)) + \sum_{i=1}^{n} y_i \log \frac{\theta(x_i)}{1 - \theta(x_i)}$$

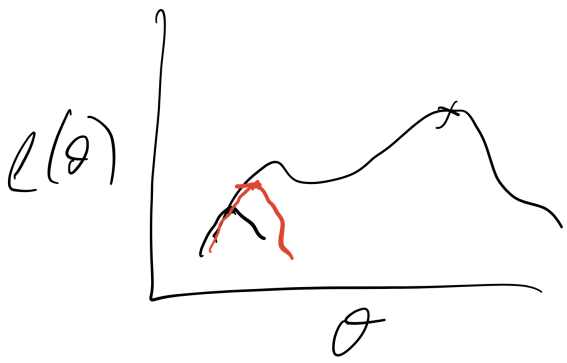
$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \log(1 - \theta(x_i)) + \sum_{i=1}^{n} y_i (\beta_0 + \beta x_i)$$

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} - \log(1 + e^{\beta_0 + \beta x_i}) + \sum_{i=1}^{n} y_i (\beta_0 + \beta x_i)$$

$$\frac{\partial \ell}{\partial \beta_j} = - \sum_{i=1}^{n} \frac{1}{1 + e^{\beta_0 + \beta x_i}} e^{\beta_0 + \beta x_i} x_j + \sum_{i=1}^{n} y_i x_{i,j}$$

$$= \sum_{i=1}^{n} \left[y_i - \theta(x_i, \beta_0, \beta) \right] x_{i,j}$$

Solve using numerical methods (Newton-Raphson ≡ IRLS).



If we consider two subjects with covariate values $x+\Delta$ and x, respectively, their odds ratio becomes:

$$\frac{\exp(\beta_0 + \beta_1(x + \Delta))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1 \Delta)$$

In particular, e^{β_1} is the odds ratio corresponding to one unit's increase in the value of the covariate.

Wald tests and confidence intervals

- $\hat{\beta}_j$ = MLE for β_j
- $\text{se}(\hat{\beta}_j)$ = standard error for $\hat{\beta}_j$

To test the null hypothesis $H_{0j} : \beta_j = 0$, we use the Wald test statistic:

$$z = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

which is approximately $N(0, 1)$-distributed under H_{0j}.

95% confidence interval for β_j:

$$\hat{\beta}_j \pm 1.96 \cdot \text{se}(\hat{\beta}_j)$$

$OR_j = \exp(\beta_j)$ is the odds ratio for one unit's increase in the value of the j-th covariate, holding all other covariates constant.

We obtain a 95% confidence interval for OR_j by transforming the lower and upper limits of the confidence interval for β_j.

Consider the WCGS study with CHD as the outcome and age, cholesterol (mg/dL), systolic blood pressure (mmHg), body mass index (kg/m^2), and smoking (yes, no) as predictors.

R commands:

```
wcgs.mult = glm(chd69 ~ age + chol + sbp + bmi + smoke, data = wcgs,
                family = binomial, subset = (chol < 600))
summary(wcgs.mult)
```

R output (edited):

Variable	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-12.3110	0.9773	-12.598	$< 2e-16$
age	0.0644	0.0119	5.412	$6.22e-08$
chol	0.0107	0.0015	7.079	$1.45e-12$
sbp	0.0193	0.0041	4.716	$2.40e-06$
bmi	0.0574	0.0264	2.179	0.0293
smoke	0.6345	0.1401	4.526	$6.01e-06$

Odds Ratios with Confidence Intervals
R output (edited):

Variable	exp(coef)	Lower	Upper
(Intercept)	4.50×10^{-6}	6.63×10^{-7}	3.06×10^{-5}
age	1.067	1.042	1.092
chol	1.011	1.008	1.014
sbp	1.019	1.011	1.028
bmi	1.059	1.006	1.115
smoke	1.886	1.433	2.482

For a numerical covariate, it may be more meaningful to present an odds ratio corresponding to a larger increase than one unit.

This is easily achieved by refitting the model with a rescaled covariate.

If you (e.g.) want to study the effect of a ten-year increase in age, you fit the model with the covariate:

$$\mathsf{age\_10} = \frac{\mathsf{age}}{10}$$

R output (edited):

Variable	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-3.006	0.116	-12.598	$< 2e-16$
age_10	0.644	0.119	5.412	$6.22e-08$
chol_50	0.537	0.076	7.079	$1.45e-12$
sbp_50	0.965	0.205	4.716	$2.40e-06$
bmi_10	0.574	0.264	2.179	0.0293
smoke	0.634	0.140	4.526	$6.01e-06$

Odds Ratios with Confidence Intervals

R output (edited):

Variable	exp(coef)	Lower	Upper
(Intercept)	0.0494	0.0394	0.0621
age_10	1.9050	1.5085	2.4057
chol_50	1.7110	1.4746	1.9853
sbp_50	2.6240	1.7573	3.9180
bmi_10	1.7760	1.0595	2.9770
smoke	1.8860	1.4329	2.4824

An aim of the WCGS study was to study the effect on CHD of certain **behavioral patterns**, denoted A1, A2, B3, and B4.

Behavioral pattern is a categorical covariate with four levels and must be fitted as a factor in R.

R output (edited):

Variable	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-2.7527	0.2259	-12.19	$< 2e-16$
age_10	0.6064	0.1199	5.057	$4.25e-07$
chol_50	0.5330	0.0764	6.980	$2.96e-12$
sbp_50	0.9016	0.2065	4.367	$1.26e-05$
bmi_10	0.5536	0.2656	2.084	0.0372
smoke	0.6047	0.1411	4.285	$1.82e-05$
behcat2	0.0660	0.2212	0.298	0.7654
behcat3	-0.6652	0.2423	-2.746	0.0060
behcat4	-0.5585	0.3192	-1.750	0.0802

Here we may be interested in:

- Testing if behavioral patterns have an effect on CHD risk.
- Testing if it is sufficient to use two categories for behavioral patterns (A and B).

In general, we consider a logistic regression model:

$$\theta(x_1, x_2, \ldots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}.$$

Here we want to test the null hypothesis that q of the β_j's are equal to zero, or equivalently that there are q linear restrictions among the β_j's.

Examples:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad (q = 4)$$

$$H_0 : \beta_1 = \beta_2 \quad \text{and} \quad \beta_3 = \beta_4 \quad (q = 2)$$

Deviance

In multiple linear regression, the residual sum of squares provides the basis for tests for comparing mean functions. In logistic and Poisson regression, the residual sum of squares is replaced by the deviance, which is often called G^2. The deviance is defined for logistic regression to be:

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)\right],$$

where $\hat{y}_i = m_i\hat{\theta}_i(x_i)$ are the fitted number of successes in m_i trials. The df associated with the deviance is equal to the number of cases n used in the calculation minus the number of elements of β.

Methodology for comparing models parallels the results in multiple linear regression. Write:

$$\beta' x = \beta_1' x_1 + \beta_2' x_2,$$

and consider testing:

$$H_0 : \theta(x) = m(\beta_1' x_1),$$

$$H_1 : \theta(x) = m(\beta_1' x_1 + \beta_2' x_2).$$

Obtain the deviance $G^2_{H_0}$ and degrees of freedom df_{H_0} under the null hypothesis, and then obtain $G^2_{H_1}$ and df_{H_1} under the alternative hypothesis. As with linear models, we will have evidence against the null hypothesis if:

$$G^2_{H_0} - G^2_{H_1} \quad \text{is large.}$$

To get p-value: Compare the difference with χ^2_{df} where:

$$df = df_{H_0} - df_{H_1}.$$

[Note: Not with F distribution as in MLR.]

Poisson Regression

When the data are to be modeled as if they are Poisson counts, the rate parameter is assumed to depend on the regression with linear predictors $\beta' x$ through the link function:

$$\log[\lambda(\beta' x)] = \beta' x.$$

Poisson regression models are often called log-linear models.

Interpretation

When all explanatory variables are zero, then:

$$E[Y_i | X_{i,1} = 0, \ldots, X_{i,p} = 0] = \lambda_i = e^{\beta_0}.$$

Thus, β_0 determines the expected response when all explanatory variables are zero.

More generally:

$$E[Y_i | X_{i,1} = x_1, \ldots, X_{i,p} = x_p] = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}.$$

If $X_{i,1}$ increases by one unit, we have:

$$E[Y_i | X_{i,1} = x_1 + 1, \ldots, X_{i,p} = x_p] = e^{\beta_0 + \beta_1 (x_1 + 1) + \cdots + \beta_p x_p} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} e^{\beta_1}.$$

Thus:

$$\frac{E[Y_i | X_{i,1} = x_1 + 1, \ldots, X_{i,p} = x_p]}{E[Y_i | X_{i,1} = x_1, \ldots, X_{i,p} = x_p]} = e^{\beta_1}.$$

Thus e^{β_p} is the multiplicative effect on the mean response for a one-unit increase in the associated explanatory variable when holding all other explanatory variables constant.

Offset

If not all counts are based on the same amount of time or space, we need to account for the amount of time or space used. To do this, we can include an **offset**.

Let T_i represent the amount of time or space. Then a Poisson regression model with an offset is:

$$Y_i \overset{ind}{\sim} Po(\lambda_i),$$

and

$$\log(\lambda_i) = \log(T_i) + \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p}.$$

The offset is $\log(T_i)$ and can be thought of as an explanatory variable with a known coefficient of 1. Note that:

$$\log E[Y_i/T_i] = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p}.$$

So we are effectively modeling the **rate**.

Maximum likelihood estimation is the usual method used to fit Poisson regression models. The deviance for Poisson regression is given by:

$$G^2 = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i) \right],$$

where $\hat{y}_i$ is the fitted value $\exp(\beta' x_i)$.

Goodness of Fit Tests

If a Poisson mean function is correctly specified, the residual deviance G^2 will be distributed as a $\chi^2(n-p')$ random variable, where n is the number of cells and p' is the number of regressors fit. If the mean function is not correctly specified, or if the Poisson assumption is wrong, then G^2 will generally be too large. A lack-of-fit test can be obtained by comparing the value of G^2 to the

relevant χ^2 distribution. The same idea can be used for binomial regression when the sample sizes are larger than 1.

An alternative to using G^2 for lack-of-fit testing is to use Pearson's χ^2 for testing, given by the familiar formula:

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

Like G^2, X^2 is compared with $\chi^2(n - p')$ to get significance levels. In large samples, the two tests will give the same inference, but in smaller samples, χ^2 is generally more powerful.

In binomial regression with all or nearly all $m_i = 1$, neither G^2 nor X^2 provides a lack-of-fit test.

Transferring Knowledge about Linear Models

Most of the methodology developed in this book transfers to problems with binomial or Poisson responses. In this section, important connections are briefly summarized.

Scatterplots and Regression

Graphing data is just as important in binomial and Poisson regression as it is in linear regression. In problems with a binary response, plots of the response versus predictors or regressors are generally not very helpful because the response only has two values. Smoothers, however, can help look at these plots as well. Plots of predictors with color used to indicate the level of the response can also be helpful.

Simple and Multiple Regression

The general ideas in Chapters 2 and 3 apply to binomial and Poisson models, even if the details differ. With the counted data models, estimates $\hat{\beta}$ and $\text{Var}(\hat{\beta}|X)$ are computed using the appropriate maximum likelihood methods, not with the formulas in these chapters. Once these are found, they can be used in these formulas and methods given in the text. For example, a point estimate and standard error for a linear combination of the elements of β is given by

$$\hat{l} = a'\beta, \quad se(\hat{l}|X) = \hat{\sigma}\sqrt{a'(X'X)^{-1}a},$$

for linear regression. For the binomial and Poisson fit, we can replace $\hat{\sigma}$ by 1 and replace $(X'X)^{-1}$ by the covariance matrix of $\hat{\beta}$. Confidence intervals and tests use the standard normal rather than a t-distribution.

Testing and Analysis of Deviance

The t-tests discussed in Chapters 2, 3, and 6 are replaced by z-tests for binomial and Poisson models. The F-tests in Chapter 6 are replaced by χ^2 tests based on changes in deviance. The marginality principle, Section 6.2, is the guiding principle for testing with counted responses.

In linear models, the t-tests and F-tests for the same hypothesis have the same value, and so they are identical. With binomial and Poisson responses, the tests are identical only for very large samples, and in small samples, they can give conflicting summaries. The G^2 tests are generally preferred.

Variances

Failure of the assumptions needed for binomial or Poisson fitting may be reflected in overdispersion, meaning that the variation between observations given the predictors is larger than the value required by the model. One

general approach to overdispersion is to fit models that allow for it, such as:

- Estimate a dispersion parameter($\text{var}(y_i) = \theta(x)(1 - \alpha x)$).
- Negative binomial instead of binomial.
- Generalized linear mixed models.

Transformations

- Transformation of responses not relevant.
- Transformation of covariates still relevant.

Regression Diagnostics

- Generalizations of Pearson residuals can be defined.
- These can be used for diagnostic purposes.

$$Y = X\beta + e$$

$$g(E(Y)) = X\beta$$

$$\text{var}(Y \mid X) = \theta(x)(1 - \theta(X))$$

Variable Selection

Many previous ideas can be generalized for binomial and Poisson regression models, such as:

- Information criteria
- Lasso
- FS / BE / Stepwise

Generalized Linear Models

The models for

- Multiple linear regression
- Logistic regression
- Poisson regression

are the most common **generalized linear models (GLMs)**.

A GLM consists of three parts:

- A family of distributions *(exponential family)*
- A linear predictor
- A link function

Example GLM: (standard) Poisson Regression

The three parts are for Poisson regression:

- Family: The observations Y_i are independent and Poisson distributed with means $\mu_i = E(Y_i)$.
- The linear predictor: A linear expression in regression parameters and covariates

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

- The link function: Linking μ_i and η_i

$$\eta_i = g(\mu_i) = \log(\mu_i)$$

For the multiple linear regression model, the family is normal, and the link function is an identity function:

$$\eta_i = g(\mu_i) = \mu_i$$

For logistic regression: binary/binomial family and link function is the logit function:

$$\eta_i = g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$$

Other Link Functions

Other link functions may also be specified:

- **For binary responses:**
  - Complementary log-log link:

$$\eta_i = g(\mu_i) = \log(-\log(1-\mu_i))$$

  - Probit link:

$$\eta_i = g(\mu_i) = \Phi^{-1}(\mu_i) \quad \text{where } \Phi(z) \text{ is the cumulative } N(0,1)\text{-distribution}$$

- **For Poisson responses:**
  - Identity link:

$$\eta_i = g(\mu_i) = \mu_i$$

  - Square root link:

$$\eta_i = g(\mu_i) = \sqrt{\mu_i}$$

Statistical Inference in GLMs

Statistical inference in GLMs is performed as illustrated for logistic regression and Poisson regression:

- **Estimation:**
  - Maximum likelihood (MLE)
- **Testing:**
  - Wald tests
  - Deviance/likelihood ratio tests

A particular feature of the GLMs is the **variance function** $V(\mu)$, which is specific for each family of distributions. The variance functions describe how the variance depends on the mean μ.

- For the Poisson distribution: $V(\mu) = \mu$
- For binary data: $V(\mu) = \mu(1 - \mu)$
- For normal data, we define $V(\mu) = 1$ since the variance does not depend on the mean

$$\text{thus} \quad \text{Var}(Y_i) = \sigma^2 = \sigma^2 V(\mu_i)$$

As discussed previously, there may be overdispersion relative to a Poisson model. This could be allowed for by specifying a model:

$$\text{Var}(Y_i) = \phi V(\mu_i)$$

Example: Number of Sexual Partners

Study of sexual habits, Norwegian Institute of Public Health, 1988

$$Y_i = \text{no. sex-partners}, \quad i = 1, \ldots, n = 8553$$

A Poisson-regression found that the expected value increased with:

- Age, being single, having had HIV-test and was higher for men.

However, the data was overdispersed. A "Pearson X^2" statistic is:

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = 51927$$

which is large compared with residual degrees of freedom 8544. An overdispersion term is estimated as:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = \frac{51927}{8544} = 6.08$$

and should have been close to 1 if the Poisson model was correct. Standard errors and inference need correction for overdispersion!

Correction for Overdispersion

An overdispersed Poisson model is given by:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

$$\text{Var}(Y_i) = \phi \mu_i$$

This model can be fitted as a standard Poisson-regression, but the standard errors must be corrected to:

$$se^* = se\sqrt{\hat{\phi}}$$

where se is the standard error from the Poisson-regression and the overdispersion $\hat{\phi}$ is estimated as on the previous slide. Similarly, the z-values become:

$$z^* = \frac{z}{\sqrt{\hat{\phi}}}$$

and p-values must be corrected correspondingly.

Binomial Regression

Additional points

Recall, if $Y \sim \text{Bin}(n, \pi)$, then:

$$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

Now, if we have $Y_1, Y_2, \ldots, Y_N$ independent random variables, where each corresponds to the number of successes in N different subgroups/strata, where:

$$N = \sum n_i$$

Thus, $Y_i \sim \text{Bin}(n_i, \pi_i)$.

The likelihood function is:

$$\ell(\pi_1, \ldots, \pi_N; Y_1, \ldots, Y_N) = \sum_{i=1}^{N}\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i}\right].$$

Model Formulation

- Goal: To describe $P_i = \frac{y_i}{n_i}$ in each subgroup as a function of predictors.
- Note: $\mathbb{E}(Y_i) = n_i \pi_i$

$$\mathbb{E}(P_i) = \pi_i$$

- Then define:

$$g(\pi_i) = X_i' \beta$$

  where g is the logit link function, for example.

Estimation

- When $p' \leq 2$ (intercept and slope), just use Maximum Likelihood Estimation (MLE).
- Example data:

x_i (dose)	n_i	y_i
$\vdots$	$\vdots$	$\vdots$

  where N strata are observed.

When $p' > 2$ (more predictors):

- Group observations into covariate patterns.
- Then model as response:

$$Y_i = \#\text{ successes for covariate pattern } i \sim \text{Bin}(n_i, \pi_i)$$

- **Note:** If each observation has a different covariate pattern:

$$n_i = 1 \implies y_i \text{ is binary.}$$

- Proceed as usual using MLE under independent Bernoulli random variables.

Overdispersion

- Assume:

$$Y_i \text{ (independent)} \sim \text{Bin}(n_i, \pi_i)$$

- Variance:

$$\text{var}(Y_i) = n_i\pi_i(1-\pi_i)$$

- What if:

$$\text{var}(Y_i) > n_i\pi_i(1-\pi_i)$$

- **Causes:**
  - Lack of independence of Y_i's.
  - Wrong distributional assumption.

Can adjust for this:

- Assume:

$$\text{var}(Y_i) = n_i\pi_i(1-\pi_i)\phi$$

  where ϕ is the overdispersion parameter (estimated using MLE).

Other Approaches:

- Use negative binomial distribution.
- Use clustered data techniques.