STA4030: Categorical Data Analysis
Generalized Linear Models

Instructor: Bojun Lu

School of Science and Engineering
CUHK(SZ)

June 30, 2020

Agenda

1. 7.1 Introduction to GLMs
2. 7.2 GLMs for Binary Data
3. 7.3 GLMs for Count Data
4. 7.4 Extension of Notation in GLMs

7.1 Introduction to GLMs

7.1.1 Definition of generalized linear models (GLMs)

Generalized linear models (GLMs) extend ordinary regression models to encompass non-normal response distributions and model functions of the mean. Three components specify a generalized linear model:

1. A *random component* identifies the response variable Y and its probability distribution.
2. A *systematic component* specifies explanatory variables used in a linear predictor function.
3. A *link function* specifies the function of $\mathrm{E}(Y)$ that the model equates the systematic component.

7.1 Introduction to GLMs

7.1.2 Components of generalized linear models

1. The *random component* of a GLM consists of a response variable Y with independent observations $(y_1, \ldots, y_N)$ from a distribution in the Natural Exponential Family. The family has probability density function or mass function form:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)].$$

Notes:

(i) Several important distributions are special cases, including the Poisson and Binomial.

(ii) The value of the parameter θ_i may vary for $i = 1, \ldots, N$, depending on values of explanatory variables.

(iii) The term $Q(\theta)$ is *natural parameter*.

7.1 Introduction to GLMs

Special cases in the Natural Exponential Family:

A: $y \sim \text{Binomial}(n, \pi)$

$$\begin{aligned} f(y;\pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \binom{n}{y} \left(\frac{\pi}{1-\pi}\right)^y (1-\pi)^n \\ &= (1-\pi)^n \binom{n}{y} \exp\left[y \log \frac{\pi}{1-\pi}\right], \end{aligned}$$

where

$$a(\theta) = (1-\pi)^n;\;\; b(y) = \binom{n}{y};\;\; Q(\theta) = \log \frac{\pi}{1-\pi}.$$

7.1 Introduction to GLMs

B: $y \sim \text{Poisson}(\mu)$

$$\begin{aligned} f(y;\mu) &= \frac{\mu^y}{y!}e^{-\mu} \\ &= e^{-\mu}\frac{1}{y!}\exp\left[y\log\mu\right], \end{aligned}$$

where

$$a(\theta) = e^{-\mu};\;\; b(y) = \frac{1}{y!};\;\; Q(\theta) = \log\mu.$$

7.1 Introduction to GLMs

C: $y \sim \text{Normal}(\mu, 1)$

$$\begin{aligned} f(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mu^2\right) \exp\left(-\frac{1}{2}y^2\right) \exp(y\mu), \end{aligned}$$

where

$$a(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mu^2\right); \;\; b(y) = \exp\left(-\frac{1}{2}y^2\right); \;\; Q(\theta) = \mu.$$

7.1 Introduction to GLMs

2. The *systematic component* of a GLM relates a vector $(\eta_1, \ldots, \eta_N)$ to the explanatory variables through a linear model.

Let x_{ij} denote the value of predictor j $(j = 1, \ldots, p)$ for observation i. Then

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, N.$$

This linear combination of explanatory variables is called the *linear predictor*.

Usually, one of the x_{ij} is equal to 1 for all i, e.g.

$$x_{i1} = 1, \text{ for } i = 1, \ldots, n.$$

The corresponding coefficient β_1 is an intercept in the model.

7.1 Introduction to GLMs

3. The *link function* of a GLM connects the random and systematic components.

Let $\mu_i = \mathbb{E}[Y_i], i = 1, \ldots, N$. The model links μ_i to $\eta_i = g(\mu_i)$, where the link function $g(\cdot)$ is a monotonic, differentiable function. Thus, $g(\cdot)$ links $\mathbb{E}[Y_i]$ to explanatory variables through the formula

$$g(\mu_i) = \eta_i = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, N.$$

The link allows the mean response to be non-linearly related to the linear predictor: $\eta_i = \log(\mu_i)$ makes sense for count data, for example, as $\mu_i > 0$.

7.1 Introduction to GLMs

The most important special case is $g(\mu) = \mu$, called the *identity link*, has $\eta_i = \mu_i$. It specifies a linear model for the mean itself. This is the link function for the ordinary regression with normally distributed Y.

The link function that transforms the mean to the natural parameter is called *canonical link*. For the canonical link,

$$g(\mu_i) = Q(\theta_i), \quad \text{and} \quad Q(\theta_i) = \sum_j \beta_j x_{ij}.$$

7.1 Introduction to GLMs

Random Component	Link Function	Systematic Component	Model
Normal	Identity	Continuous	Regression
Normal	Identity	Categorical	ANOVA
Normal	Identity	Mixed	ANCOVA
Binomial	Logit	Mixed	Logistic Regression
Binomial	Probit, etc.	Mixed	Binary Regression
Multinomial	Generalized Logit	Mixed	Multinomial Response
Poisson	Log	Mised	Loglinear

Table: Types of GLMs for Statistical Analysis

7.1 Introduction to GLMs

GLM generalizes ordinary regression. It lets the response Y be non-normal.

Previously we might have tried to transform the data to make it normal with constant variance, so we could apply ordinary least-squares regression. This could be unintuitive, fiddly and unsuccessful. With GLMs, this is no longer necessary.

The choice of link function is separate from the choice of random component. Of course, some combinations are more popular than others (see the table on the previous slide), but if a link is useful, i.e. a linear model for the predictors is useful for that link, then use it. It is not necessary that the link also stabilizes variance or produces normality.

7.1 Introduction to GLMs

7.1.3 Binomial logit models for binary data

Response variable Y is binary. Represent the success and failure outcomes by 1 and 0, and

$$P(Y=1)=\pi,\ \ P(Y=0)=1-\pi,\ \ \mathrm{E}(Y)=\pi.$$

This is a *Bernoulli distribution*, and it is the special case of the Binomial distribution with $n=1$. The probability mass function is:

$$\begin{aligned} f(y;\pi) &= \pi^{y}(1-\pi)^{1-y} = (1-\pi)[\pi/(1-\pi)]^{y} \\ &= (1-\pi)\exp\left[y\log\frac{\pi}{1-\pi}\right],\ \text{for } y=0,1 \end{aligned}$$

7.1 Introduction to GLMs

It belongs to the Natural Exponential Family with

$$\theta = \pi, a(\theta) = 1 - \pi, \; b(y) = 1 \text{ and } Q(\theta) = \log \frac{\pi}{1-\pi}.$$

The natural parameter $\log \frac{\pi}{1-\pi}$ is the log odds of response 1, the *logit* of π. This is the *canonical link*.

GLMs using the *logit link* are often called *logit models*.

7.1 Introduction to GLMs

7.1.4 Poisson log linear models for count data

The simplest distribution for count data is the Poisson distribution. Let Y denote a count and let $\mu = \mathrm{E}(Y)$. The Poisson probability mass function is

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp(y\log\mu),\ \ y = 0, 1, 2, \cdots .$$

hence it belongs to the Natural Exponential Family with

$$\theta = \mu,\ \mathrm{a}(\theta) = \exp(-\mu),\ \mathrm{b}(y) = \frac{1}{y!},\ \text{and } \mathrm{Q}(\theta) = \log\mu.$$

7.1 Introduction to GLMs

The natural parameter is $\log \mu$, so the canonical link function is the log link: $\eta = \log \mu$.

The model using the log link is:

$$\log \mu_i = \sum_j \beta_j x_{ij}, \; i = 1, \ldots, N.$$

This model is called the *Poisson loglinear model*.

7.2 GLMs for Binary Data

Let Y be a binary response variable. Each observation has one of two outcomes, denoted by 0 and 1, with $\mathbb{E}(Y) = \mathbb{P}(Y = 1)$. We denote $\mathbb{P}(Y = 1)$ by $\pi(\mathbf{x})$, reflecting its dependence on $\mathbf{x} = (x_1, \ldots, x_p)$ of predictors. In introducing GLMs for binary data, we use a single explanatory variable for simplicity.

7.2.1 Linear Probability model

For a binary response, the regression model

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

is called a *linear probability model*.

With independent observations, it is a GLM with binomial random component and identity link function.

7.2 GLMs for Binary Data

Advantage: a simple interpretation of β; β is the change in $\pi(x)$ for a one-unit increase in x.

Disadvantage: probabilities fall in $[0, 1]$, but linear functions take values over the entire real line. So, the estimate of $\pi(x)$ may fall outside the $[0, 1]$ for some x when fitting this model.

- Example 7.1

We illustrate the linear probability model with the data in Table 7.1. From a survey of 2,484 subjects to investigate snoring as a risk factor for heart disease. Those surveyed were classified according to their spouses' report of how much they snored. The model states that the probability of heart disease is linearly related to the level of snoring x.

7.2 GLMs for Binary Data

We treat the rows of Table 7.1 as independent binomial samples. Since no obvious choice of scores exists for categories of x, we used (0, 2, 4, 5), treating the last two levels as closer than the other adjacent pairs.

Table 7.1 Relationship between snoring and heart disease

	Heart Disease		Proportion	Linear	Logit
Snoring	Yes	No	Yes	Fit[a]	Fit[a]
Never	24	1355	0.017	0.017	0.021
Occasionally	35	603	0.055	0.057	0.044
Nearly every night	21	192	0.099	0.096	0.093
Every night	30	224	0.118	0.116	0.132

[a] Model fits refer to proportion of yes responses

7.2 GLMs for Binary Data

Software reports the ML fit,

$$\hat{\pi}(x) = 0.0172 + 0.0198x$$

with a standard error SE $= 0.0028$ for $\hat{\beta} = 0.0198$.

For non-snorers $(x = 0)$, $\hat{\pi}(x) = 0.0172$, the estimated proportion of subject having heart disease is 0.0172. We refer to the estimated values of $\mathbb{E}[Y]$ for a GLM as *fitted values*. Table 7.1 shows the sample proportions and the fitted values for this model, which suggest that the model fits well.

The interpretation is simple. The estimated probability of heart disease is about 0.02 for non-snorers. It increases $2(0.0198) = 0.04$ for occasional snorers, 0.04 for those who snore nearly every night, and 0.02 for those who always snore.

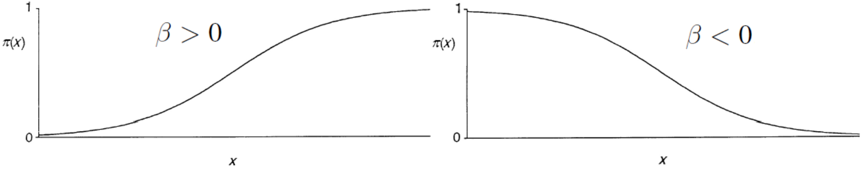
7.2 GLMs for Binary Data

Usually, binary data results from a *nonlinear* relationship between $\pi(x)$ and x.

The most important model that reflects nonlinear monotonic relationships between $\pi(x)$ and x is the *logistic regression* model:

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

As $x \to \infty, \pi(x) \uparrow 1 \quad \textit{if } \beta > 0; \quad \pi(x) \downarrow 0 \quad \textit{if } \beta < 0.$



7.2 GLMs for Binary Data

From the equation for $\pi(x)$ in the logistic regression model,

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

Thus, the appropriate link for the logistic regression model is the log odds transformation, the *logit*.

Note that the probability mass function of $Bin(n, \pi)$ is

$$f(\pi; y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} = \binom{n}{y} (1-\pi)^n \exp\left(y \log \frac{\pi}{1-\pi} \right).$$

Thus, logistic regression models are GLMs with binomial random component and logit link function. The logit is the natural parameter of the binomial distribution, so the logit link is the canonical link.

7.2 GLMs for Binary Data

For the snoring data in Example 7.1, software reports the logistic regression ML fit

$$\text{logit}[\hat{\pi}(x)] = -3.87 + 0.40x.$$

The positive $\hat{\beta} = 0.40$ reflects the increased incidence of heart disease at higher snoring levels.

Note: As $\pi(x)$ always fall in the $(0, 1)$ range, the logit can be any real number. So, the logit model does not have the structural problem.

7.2.3 Binomial GLM for 2 × 2 contingency tables

For a binary response Y, the simplest GLM is the one having a single explanatory variable X that is also binary. Label its values by 0 and 1.

7.2 GLMs for Binary Data

For a given link function, the GLM

$$\text{link}[\pi(x)] = \alpha + \beta x$$

has the effect of X described by

$$\beta = \text{link}[\pi(1)] - \text{link}[\pi(0)].$$

For the identity link,

$$\beta = \pi(1) - \pi(0) \quad \text{— the difference of proportions.}$$

For the log link,

$$\beta = \log[\pi(1)] - \log[\pi(0)] = \log \frac{\pi(1)}{\pi(0)} \quad \text{— log relative risk.}$$

7.2 GLMs for Binary Data

For the logit link,

$$\beta = \text{logit}[\pi(1)] - \text{logit}[\pi(0)] = \log\left(\frac{\pi(1)}{1-\pi(1)}\right) - \log\left(\frac{\pi(0)}{1-\pi(0)}\right)$$

$$= \log\left(\frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}\right)$$ — log odds ratio.

i.e. A measure of association for a 2×2 table is the effect parameter β in GLMs for binary data.

7.2 GLMs for Binary Data

7.2.4 Probit Regression Model

$$\text{probit}[\pi(x)] = \alpha + \beta x.$$

The link function is called the probit link. It transfers probabilities to z-scores from $N(0, 1)$. That is, $\text{probit}[z]$ gives the inverse standard normal cumulative probability of z.

E.g.

probit(0.05) = -1.645

probit(0.50) = 0

probit(0.95) = 1.645

probit(0.975) = 1.96

7.2 GLMs for Binary Data

- Example 7.2: Snoring data revisited.

For the snoring and heart disease data with scores $\{0, 2, 4, 5\}$ as snoring level, the ML fit of the probit model is

$$\text{probit}[\hat{\pi}(x)] = -2.061 + 0.188x.$$

At

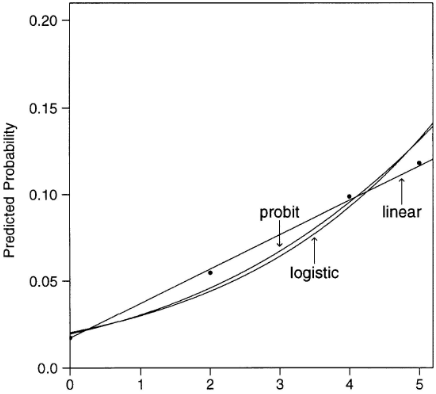
$$x = 0: \ \text{probit}(\hat{\pi}) = -2.061, \hat{\pi} = 0.020$$
$$x = 5: \ \text{probit}(\hat{\pi}) = -2.0671 + 0.188 \times 5 = -1.12, \hat{\pi} = 0.131$$

In practice, probit and logistic regression models provide similar fits.

7.2 GLMs for Binary Data

Table 7.2: Comparison among three methods.

	Heart Disease		Proportion	Linear	Logit	Probit
Snoring	Yes	No	Yes	Fit	Fit	Fit
Never	24	1355	0.017	0.017	0.021	0.020
Occasionally	35	603	0.055	0.057	0.044	0.046
Nearly every night	21	192	0.099	0.096	0.093	0.095
Every night	30	224	0.118	0.116	0.132	0.131



7.3 GLMs for Count Data

7.3.1 Poisson regression

The Poisson loglinear model with explanatory variable X is

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

For this model, the mean satisfies the exponential relationship

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

A 1-unit increase in x has a multiplicative impact of e^{β} on μ. The mean at $x+1$ equals the mean at x multiplied by e^{β}.

7.3 GLMs for Count Data

If $Y \sim \text{Poisson}(\mu)$, then the probability mass function is

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} = \frac{1}{y!}e^{-y}\exp(y\log\mu), \quad y = 0, 1, \ldots$$

Therefore, $Q(\mu) = \log\mu$, which is the natural parameter for Poisson distribution and the log link is the canonical link for a Poisson GLM.

A Poisson loglinear GLM assumes a Poisson distribution for Y and uses the log link.

7.3 GLMs for Count Data

- Example 7.3

We illustrate Poisson GLMs based on the following Table 7.3a from a study of nesting horseshoe crabs. Each female horseshoe crab had a male crab resident in her nest.

The study investigated factors affecting whether the female crab had any other males, called *satellites*, residing nearby.

Explanatory variables:

female crab's color, spine condition, weight, carapace width.

Response variable:

each crab's number of satellites.

7.3 GLMs for Count Data

Table 7.3: Number of crab satellites by female's characteristics

C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0
3	3	26.0	2.60	4	2	3	23.8	2.10	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0
4	2	21.0	1.85	0	2	1	26.0	2.30	14
2	3	29.0	3.00	1	4	3	24.7	2.20	0
1	2	25.0	2.30	3	2	1	22.5	1.60	1
4	3	26.2	1.30	0	2	3	28.7	3.15	3
2	3	24.9	2.10	0	1	1	29.3	3.20	4
2	1	25.7	2.00	8	2	1	26.7	2.70	5
2	3	27.5	3.15	6	4	3	23.4	1.90	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

7.3 GLMs for Count Data

In Table 7.3,

C – color (1, light medium; 2, medium; 3, dark medium; 4, dark)

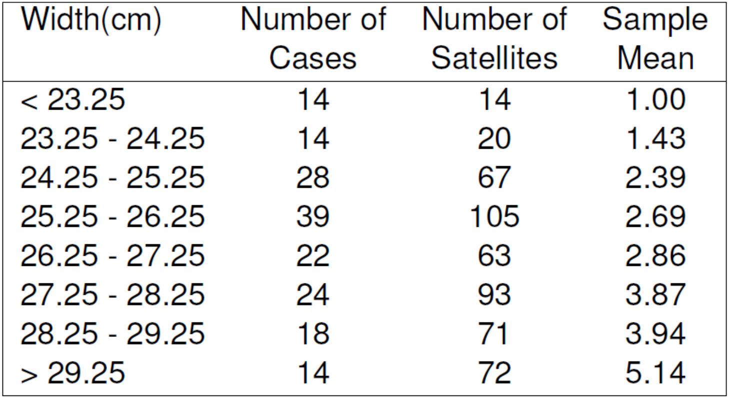
S – spine condition (1, both good; 2, one worn or broken; 3, both worn or broken)

W – carapace width (cm); Wt – weight (kg); Sa – number of satellites.

For now, we use width alone as a predictor. Figure 7.1 plots the response counts of satellites against width. The substantial variability makes it difficult to discern a clear trend. To get a clearer picture, we grouped the female crabs into width categories:

7.3 GLMs for Count Data

Table 7.4: Sample mean of number of satellites



Width(cm)	Number of Cases	Number of Satellites	Sample Mean
< 23.25	14	14	1.00
23.25 - 24.25	14	20	1.43
24.25 - 25.25	28	67	2.39
25.25 - 26.25	39	105	2.69
26.25 - 27.25	22	63	2.86
27.25 - 28.25	24	93	3.87
28.25 - 29.25	18	71	3.94
> 29.25	14	72	5.14

7.3 GLMs for Count Data

Figure 7.1

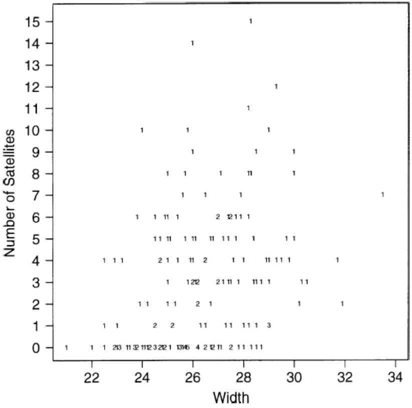
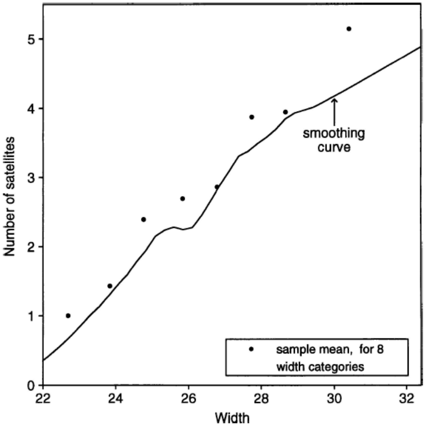


Figure 7.2



7.3 GLMs for Count Data

Figure 7.2 plots the sample mean number of satellites against the sample mean width for female crabs in each category. The plot shows a strong increasing trend.

For a female crab, let μ be the expected number of satellites and $x =$ width. From GLM software, the ML fit of the Poisson loglinear model is

$$\begin{aligned} \log \hat{\mu} &= \hat{\alpha} + \hat{\beta} x = -3.305 + 0.164x, \\ \text{or} \qquad \hat{\mu} &= \exp(-3.305 + 0.164x). \end{aligned}$$

The effect $\hat{\beta} = 0.164$ of width is positive, with SE $= 0.020$.

7.3 GLMs for Count Data

For instance, the fitted value at the mean width of x = 26.3 is

$$\hat{\mu}_x = \exp(\hat{\alpha} + \hat{\beta}x) = \exp[-3.305 + 0.164(26.3)] = 2.74.$$

In this model,

$$\exp(\hat{\beta}) = \exp(0.164) = 1.18$$

is the multiplicative effect on $\hat{\mu}$ for 1-cm increase in x.

The fitted value at $x + 1 = 27.3 = 26.3 + 1$ is

$$\hat{\mu}_{x+1} = \exp[-3.305 + 0.164(27.3)] = 3.23 = 1.18 \times 2.74 = 1.18\hat{\mu}_x.$$

A 1-cm increase in width yield an 18% increase in the estimated mean.

7.3 GLMs for Count Data

Figure 7.2 reveals that the number of satellites may grow approximately linearly with width. This suggests the Poisson regression model with identity link,

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

The ML fit: $\hat{\alpha} = -11.53$ and $\hat{\beta} = 0.55$ (SE $= 0.059$).

The effect of x on μ in this model is additive, rather than multiplicative. A 1-cm increase in width has predicted increase of $\hat{\beta} = 0.55$ in the expected number of satellites. For instance, the fitted value at the mean width of $x = 26.3$ and $x + 1 = 27.3$ are

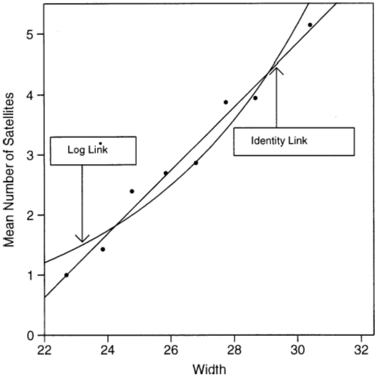
$$\hat{\mu}_x = -11.53 + 0.55(26.3) = 2.93,$$
$$\hat{\mu}_{x+1} = \hat{\mu}_x + 0.55 = 2.93 + 0.55 = 3.48.$$

On the average, an extra satellite results from roughly a 2-cm increase in width.

7.3 GLMs for Count Data

Figure 7.3 plots the fitted number of satellites against width, for the models with log link and with identity link. Though they diverge somewhat for relatively small and relatively large widths, they provide similar predictions over the portion of the width range in which most observations occur.

Figure 7.3:



7.3 GLMs for Count Data

7.3.2 Poisson regression for rate data

When events of a certain type occur over time, space, or some other index of size. It is often relevant to model the *rate* at which events occur.

e.g. In modeling numbers of auto thefts in 1995 for a sample of cities, we could form a rate for each city by dividing the number of thefts by the city's population size.

The model might describe how the rate depends on explanatory variables such as the city's unemployment rate, median income, and percentage of residents having completed high school.

7.3 GLMs for Count Data

When a response count Y has index (such as population size) equal to t, such that the expected value of Y is proportional to t, the sample rate of outcome is Y/t. The expected value of the rate is μ/t

A loglinear model for the expected rate has form

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

This model has equivalent representation

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

where the adjustment term, $-\log t$, to the log link for the mean is called an *offset*. Standard GLM software can fit models having offsets.

7.3 GLMs for Count Data

From the rate model, the expected number of outcome satisfies

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

The mean is proportional to the index t, with proportionality constant depending on the value of the explanatory variable. For a fixed value of x, doubling the population size t also doubles the expected number of auto thefts μ.

- Example 7.4

To illustrate Poisson regression models for rates, we use data dealing with motor vehicle accident rates for elderly drivers. The sample consisted of 16,262 Medicaid enrollers aged 65-84 years, with data on each subject for a period of somewhere between 0 and 4 years. The data are as follows.

7.3 GLMs for Count Data

Table 7.5: Total observation time and number of injurious accidents

	Male	Female
Total observation time (thousand years)	21.40	17.30
Total number of injurious accident	320	175
Sample rates of injurious accidents	320/21.4 =14.95	175/17.30 =10.12

Hence, the sample rates of injurious accidents are 14.95 and 10.12 crashes per thousand years of driving for males and females, respectively.

7.3 GLMs for Count Data

Let μ denote the expected number of injurious accidents, for an observation period of t thousand years. To model the effect of gender on the accident rate, we use the Poisson regression model with $x = 0$ for females and $x = 1$ for males.

The explanatory variable x $(= 0, 1)$ is a *dummy variable* for gender. The model is

$$\begin{aligned} \log(\mu/t) &= \alpha, && \text{if } x = 0 \ \ \text{(for females)}, \\ \log(\mu/t) &= \alpha + \beta, && \text{if } x = 1 \ \ \text{(for males)}. \end{aligned}$$

The rates are identical if $\beta = 0$.

7.3 GLMs for Count Data

The estimate of α is simply the sampling log(rate) for female, namely $\log(10.12) = 2.31$. The estimate of $\alpha + \beta$ is the sample log(rate) for males, namely $\log(14.95) = 2.70$. So, the estimated difference is

$$\hat{\beta} = 2.70 - 2.31 = 0.39, \ \ \exp(\hat{\beta}) = \exp(0.39) = 1.48.$$

The estimated accident rate for men was 1.48 times the rate for women. i.e. $14.95/10.12 = 1.48$, the sample rate being 48% higher for men.

7.3 GLMs for Count Data

To test whether the true rates are the same, we test

$$H_0 : \beta = 0 \quad \text{v.s.} \quad H_1 : \beta \neq 0$$

GLM software reports $\hat{\beta} = 0.39$, SE$(\hat{\beta}) = 0.09$, yielding the Wald test statistic $Z = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} = 4.33$, and P-value ≈ 0. So, there is strong evidence that the accident rate was higher for males (i.e. $\beta > 0$).

Note: the accident rates here do not take into account possibly different yearly levels of driving for the two groups.

7.3 GLMs for Count Data

7.3.3 Negative Binomial GLM

- Negative Binomial Distribution

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, \quad y=0,1,2,\ldots$$

where k and μ are parameters.

$$\mathrm{E}(y)=\mu, \quad \mathrm{Var}(y)=\mu+\mu^2/k=\mu+D\mu^2.$$

Usually, D is unknown. Estimating it helps summarize the extent of over-dispersion.

7.3 GLMs for Count Data

When k is fixed, the Negative Binomial distribution can be expressed in Natural Exponential Family form

$$f(y;k,\mu) = \left(\frac{k}{\mu+k}\right)^k \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \exp\left\{y \log\left(1-\frac{k}{\mu+k}\right)\right\},$$

where

$$a(\theta) = \left(\frac{k}{\mu+k}\right)^k, \; b(y) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}, \; Q(\theta) = \log\left(1-\frac{k}{\mu+k}\right).$$

Therefore, a model with Negative Binomial random component is a GLM. A variety of link functions can be applied. The most common one is the log link, but sometimes the identity link is adequate.

7.3 GLMs for Count Data

- Example 7.5: Horseshoe Crab Example revisited.

y = number of satellites, $x =$ shell width

1. The Poisson GLM with log link gives:

$$\log(\hat{\mu}) = -3.30 + 0.164x,$$

where $\hat{\beta} = 0.164, \quad \mathrm{SE}(\hat{\beta}) = 0.020, \quad 95\%\ \text{C.I. is} \quad (0.125, 0.203).$

2. The negative binomial GLM with log link has

$$\log(\hat{\mu}) = -4.05 + 0.192x, \quad \hat{D} = 1.1,$$

where $\hat{\beta} = 0.192, \quad \mathrm{SE}(\hat{\beta}) = 0.048, \quad 95\%\ \text{C.I. is} \quad (0.099, 0.285).$

7.3 GLMs for Count Data

e.g. at $x = 26.3$

For Poisson GLM:

$\hat{\mu} = 2.75 \quad \hat{\sigma}^2 = 2.75$

For Negative Binomial GLM:

$\hat{\mu} = 2.72 \quad \hat{\sigma}^2 = \hat{\mu} + D\hat{\mu}^2 = 10.86$

The greater SE($\hat{\beta}$) in the Negative Binomial GLM reflects the over-dispersion that is not captured by the Poisson GLM. The confidence interval for the Poisson GLM is unrealistically narrow, because it does not allow for the over-dispersion.

7.4 Extension of Notation in GLMs

7.4.1 The Exponential Dispersion Family

So far the random components we have considered have belonged to the Natural Exponential Family and only have one parameter. In fact, GLM can handle more general distributions which have a second, *dispersion* parameter.

This time, the random component of the GLM specifies that the N observations $(y_1, \ldots, y_N)$ on Y are independent, with probability mass or density function for y_i of form:

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

Distributions of this form consist the $\underline{\textit{Exponential Dispersion Family}}$.

ϕ is called $\underline{\textit{dispersion parameter}}$.

θ_i is the $\underline{\textit{natural parameter}}$.

7.4 Extension of Notation in GLMs

When ϕ is known, this form simplifies to the form of the Natural Exponential Family, which is

$$\begin{aligned} f(y_i;\theta_i) &= \exp\left\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\right\} \\ &= \exp[-b(\theta_i)/a(\phi)]\exp[c(y_i,\phi)]\exp\left[y_i\left\{\theta_i/a(\phi)\right\}\right] \\ &= a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)] \end{aligned}$$

with

$$a(\theta) = \exp[-b(\theta)/a(\phi)];\;\; b(y) = \exp[c(y,\phi)];\;\; \mathrm{Q}(\theta) = \theta/a(\phi).$$

Note:

1. This Exponential Dispersion Family contains several familiar distributions, such as Normal, Gamma, Inverse Gaussian, Binomial, Poisson, Negative Binomial, etc.
2. Usually, $a(\phi)$ has form $a(\phi) = \phi/\omega_i$ for a known weight ω_i.

7.4 Extension of Notation in GLMs

7.4.2 Characteristics of three common distributions in the exponential dispersion family

1. Normal $N(\mu, \sigma^2)$.

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$
$$= \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$
$$= \exp\left\{\left[y\mu - \frac{\mu^2}{2}\right]\bigg/\sigma^2 - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\}.$$

Thus, $\theta = \mu,\ b(\theta) = \mu^2/2$, the dispersion parameter $\phi = \sigma^2$. Canonical link $\theta(\mu) = \mu$ is the identity link, and

$$c(y;\phi) = -\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right].$$

7.4 Extension of Notation in GLMs

2. Poisson(μ)

$$f(y;\mu) = \frac{\mu^y}{y!}e^{-\mu} = \exp\left\{y\log\mu - \log y! - \mu\right\}$$
$$= \exp\left\{[y\log\mu - \mu] - \log y!\right\}.$$

Thus, $\theta = \log\mu,\ b(\theta) = \mu = \exp(\theta),\ c(y;\theta) = -\log y!$

Dispersion parameter: $\phi = \sigma^2 = 1,\ a(\phi) = 1.$

Canonical link $\theta(\mu) = \log\mu$ is the log link.

7.4 Extension of Notation in GLMs

3. Binomial $B(n, \pi)$

$$\begin{aligned} f(y;\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= \binom{n}{y}\left(\frac{\pi}{1-\pi}\right)^y(1-\pi)^n \\ &= \exp\left\{\left[y\log\frac{\pi}{1-\pi} + n\log(1-\pi)\right] + \log\binom{n}{y}\right\}. \end{aligned}$$

Thus, $\theta = \log[\pi/(1-\pi)],\ b(\theta) = n\log(1-\pi) = -n\log(1+e^\theta)$.

Dispersion parameter: $\phi = \sigma^2 = 1,\ a(\phi) = 1$.

Canonical link $\theta(\pi) = \text{logit}(\pi)$ is the logit link, and $c(y;\phi) = \log\binom{n}{y}$.

7.4 Extension of Notation in GLMs

7.4.3 Maximum Likelihood Methods

Why do we require that the distribution of Y belong to the Exponential Dispersion Family?

Because the family is sufficiently large to contain most of our favourite distributions and regular enough to find estimate parameters and fit models easily.

Consider the loglikelihood function and its derivatives

$$l_i(y_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi); \frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}; \frac{\partial^2 l_i}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)}$$

7.4 Extension of Notation in GLMs

Since distribution f is in the Exponential Dispersion Family, it satisfies certain regularity conditions which allow us to write

$$\mathbb{E}\left[\frac{\partial l_i}{\partial \theta_i}\right] = 0 \text{ and } \mathbb{E}\left[\frac{\partial^2 l_i}{\partial \theta_i^2}\right] = -\mathbb{E}\left[\left(\frac{\partial l_i}{\partial \theta_i}\right)^2\right]$$

hence $\mathbb{E}[Y_i] = b'(\theta_i)$ and $Var(Y_i) = b''(\theta_i)a(\phi)$.

Now, the N linear predictors $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$, for $i = 1, \ldots, N$, are related to $\mathbb{E}[Y_i] = \mu_i$ through the link function g, i.e. $g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}$. When we fit an GLM for a given link function, random component and explanatory variables, it is the β_js we fit. We do so via the usual Maximum Likelihood methods.

7.4 Extension of Notation in GLMs

The equations we need to solve are

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j} = 0 \text{ for } j = 1, \ldots, p.$$

Apply the chain rule a few times to rewrite

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

and work out each part

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}; \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{Var(Y_i)}; \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

7.4 Extension of Notation in GLMs

Leading to the equations which must be solved to find the MLEs $\hat{\beta}_1, \ldots, \hat{\beta}_p$ of $\beta_1, \ldots, \beta_p$:

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \text{ for } j = 0, 1, \ldots, p$$

(It looks like there are no β_js in this equation, so how can we solve for them? However, remember $\mu_i = g^{-1}(\sum_{j=1}^{N} \beta_j x_{ij})$, so the β_js are involved implicitly.)

Usually these equations are nonlinear can are solved by numerical approaches like the Newton-Raphson method or the Fisher Scoring method.