Linear Statistical Modeling Methods with SAS

Generalized Linear Models

Xuemao Zhang
East Stroudsburg University

March 27, 2024

Outline

- The Exponential Family of Distributions
- Generalized Linear Models
  - General Linear Models
  - Logistic Regression
  - Poisson Regression

The Exponential Family of Distributions

The **generalized linear model** (GLM) developed by Nelder and Wedderburn (1972) is a generalization of normal linear models. It requires that the response variables be from an **exponential family** and the expected responses be a function of the **linear predictors**.

For the scalar observation y, suppose the probability density function is given by

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

for some functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. This is called an **exponential family** with canonical parameter θ if ϕ is known.

It can be shown that

$$E(Y) = b^{'}(\theta).$$

Moreover, the variance of Y is related to its expected value by

$$Var(Y) = b^{''}(\theta)a(\phi),$$

where $b^{'}(\theta)$ is called the variance function and ϕ is called the dispersion parameter.

Natural exponential family of distributions

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

- θ is the natural parameter.
- ϕ is the dispersion parameter, often known.
- $\theta = g(\mu)$, where $\mu = E(Y)$, gives $\mu = g^{-1}(\theta)$; The link function g is called **Canonical/natural Link**.
- $E(Y) = b'(\theta)$
- $Var(Y) = \phi\, b''(\theta) = \phi V(\mu)$
- $V(\mu) = b''(\theta)$ is called the *variance function*.

Examples of Exponential Family - Normal Distribution

Normal pdf

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

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

- Natural parameter is $\theta = \mu$
- Dispersion parameter is $\phi = \sigma^2$
- $b(\theta) = \frac{\theta^2}{2}$

Examples of Exponential Family - Bernoulli or Binary

Binary pmf

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

$$
\begin{aligned}
p(y) = \pi^y (1-\pi)^{1-y} &= \exp\left\{y \log \pi + (1-y) \log(1-\pi)\right\} \\
&= \exp\left\{y \left(\log \pi - \log(1-\pi)\right) + \log(1-\pi)\right\} \\
&= \exp\left\{y \left(\log \frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\} \\
&= \exp\left\{\frac{y \left(\log \frac{\pi}{1-\pi}\right) - (-\log(1-\pi))}{1} + 0\right\},
\end{aligned}
$$

where $0 < \pi < 1$ is the population proportion.

- Natural parameter is $\theta = \log \frac{\pi}{1-\pi} = \log \frac{\mu}{1-\mu}$
- Dispersion parameter is $\phi = 1$
- $b(\theta) = \log(1 + e^{\theta})$

Examples of Exponential Family - Poisson

Poisson pmf

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

$$\begin{aligned} p(y) = \frac{\lambda^y e^{-\lambda}}{y!} &= \exp\left\{\log(\lambda^y) - \lambda - \log(y!)\right\} \\ &= \exp\left\{\frac{y\log(\lambda) - \lambda}{1} - \log(y!)\right\} \quad \lambda > 0, y = 0, 1, 2\ldots \end{aligned}$$

- Natural parameter is $\theta = \log(\lambda)$
- Dispersion parameter is $\phi = 1$
- $b(\theta) = \lambda = e^{\theta}$

The General Linear Model

Let $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)'$ be $n \times 1$ dimensional vectors. The classical general linear model can be rearranged to the following tripartite form:

1. The random component: $\mathbf{Y}$ has independent Normal distribution with constant variance σ^2 and $E(\mathbf{Y}) = \boldsymbol{\mu}$.
2. The systematic component: covariates in the form of an $n \times (k+1)$ design matrix $\mathbf{X} = (\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}) = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$ produce a linear predictor $\boldsymbol{\eta}$ given by
$$\boldsymbol{\eta} = \boldsymbol{X}\beta,$$
where β is a $(k+1) \times 1$ regression parameter vector.
3. The link between the random and systematic components is given by
$$\boldsymbol{\mu} = \boldsymbol{\eta}.$$

Components of a Generalized Linear Model

Generalized linear models generalize the classical linear models by allowing two extensions. First, the distribution in part 1 comes from an **exponential family** which includes the normal distribution as a special case. Secondly, the link between the random and systematic components is given by $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, where g is called the **link function** which is monotone and differentiable.

- **Random Component**: Probability distribution for **Y**
- **Systematic component**: Specifies explanatory variables in the form of a "linear predictor":
$$\boldsymbol{\eta} = \boldsymbol{X}\beta$$
- **Link function**: Connects $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, where $E(\mathbf{Y}) = \boldsymbol{\mu}$.

Random Component: Distribution of Y

To simplify the notations, we consider the relationships between the scalar random variables instead of using matrix notation.

- Ordinary regression: Normal
- Logistic regression: Bernoulli
- Poisson regression: Poisson

- Other possibilities: Binomial, Exponential, Gamma, Geometric ...

Systematic component: A regression-like equation called the *linear predictor*

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Link Function: The linear predictor is an increasing function of the expected value

The link function g is monotone and differentiable. If g is increasing:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- The function $g(\mu)$ is strictly increasing.
- The linear predictor is an increasing function of μ.
- So μ is an increasing function of the linear predictor.

General Linear Models (Normal Response)

- Link function

$$\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

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

- Natural parameter is $\theta = \mu$

- $E(Y) = \mu$
- The identity link: $\eta = g(\mu) = \mu$
- $\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

- It is our multiple linear regression model

Logistic Regression (Binary Response)

- Link function

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$$p(y) = \exp\left\{ \frac{y\left(\log\frac{\pi}{1-\pi}\right) - (-\log(1-\pi))}{1} + 0 \right\},$$

where $0 < \pi < 1$ is the population proportion.

- Natural parameter is $\theta = \log\frac{\pi}{1-\pi} = \log\frac{\mu}{1-\mu}$

- $E(Y) = \mu = \pi$
- The logit link: $\eta = g(\mu) = \log\frac{\mu}{1-\mu}$
- $\eta = \log\frac{\mu}{1-\mu} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

Poisson Regression (Poisson Response)

- Link function

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$$p(y) = \exp\left\{\frac{y\log(\lambda) - \lambda}{1} - \log(y!)\right\}, \quad \lambda > 0, y = 0, 1, 2, \ldots$$

- Natural parameter is $\theta = \log(\lambda)$

- $E(Y) = \mu = \lambda$
- The log link: $\eta = g(\mu) = \log(\mu)$
- $\eta = \log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

Estimating and Testing Generalized Linear Models

GLMs are fit to data by the method of **maximum likelihood (ML)**, providing not only estimates of the regression coefficients but also estimated asymptotic (i.e., large-sample) standard errors of the coefficients. The ML estimates can be found using an **IRLS** (Iteratively Re-Weighted Least Squares) algorithm.

To test the null hypothesis

$$H_0 : \beta_i = 0, i = 0, 1, \dots, k$$

we can compute the **Wald statistic**

$$Z_0 = \frac{\widehat{\beta}_i - 0}{SE(\widehat{\beta}_i)},$$

where $SE(\widehat{\beta}_i)$ is the asymptotic standard error of the estimated coefficient $\widehat{\beta}_i$. Under the null hypothesis, Z_0 follows a standard normal distribution.

Estimating and Testing Generalized Linear Models

Some of the exponential families on which GLMs are based include an unknown dispersion parameter ϕ. This parameter generally is estimated by the "method of moments" although this parameter can, in principle, be estimated by maximum likelihood as well.

Furthermore, like in the general linear models, we need to deal with model selection and diagnostics for generalized linear models.

The **GENMOD** procedure fits generalized linear models. Please see https://support.sas.com/documentation/onlinedoc/stat/141/genmod.pdf for details.

Example 1: Logistic Regression

A systems analyst studied the effect of computer programming experience on ability to complete within a specified time a complex programming task, including debugging. Twenty-five persons were selected for the study. They had varying amounts of programming experience (measured in months of experience), as shown in Table 14. la, column 1. All persons were given the same programming task, and the results of their success in the task are shown in column 2. The results are coded in binary fashion: $Y = 1$ if the task was completed successfully in the allotted time, and $Y = 0$ if the task was not completed successfully.

The scatter plot of the data is not too informative because of the nature of the response variable, other than to indicate that ability to complete the task successfully appears to increase with amount of experience.

Let's fit logistic regression model.

Example 1: Logistic Regression

```
data ch14tab01;
input x y@@;
label x = 'Experience'
y = 'Success';
cards;
14 0 29 0
6 0 25 1
18 1 4 0
18 0 12 0
22 1 6 0
30 1 11 0
30 1 5 0
20 1 13 0
9 0 32 1
24 0 13 1
19 0 4 0
28 1 22 1
8 1
;
run;
```

Example 1: Logistic Regression

```
proc sgplot data=ch14tab01;
  scatter x=x y=y / group=y markerattrs=(symbol=circlefilled);
  xaxis label='Experience';
  yaxis label='Success';
run;
```

Example 1: Logistic Regression

- Profile likelihood confidence intervals for the regression parameters are computed using the **LRCI** option.

```
proc genmod data=ch14tab01 descending;
/*descending order of the response variable*/
model y = x  / dist = bin  link = logit  lrci;
output out = temp resdev=devresidual p = fittedp;
/* fitted probabilities (FITTEDP) and deviance residuals (DEVRESIDUA
run;
```

```
proc print data = temp;
var x y  fittedp devresidual;
run;
```

```
proc gplot data = temp;
plot y*x fittedp*x / overlay;
run;
```

Example 2: Binomial Outcomes

- A binomial outcome is the number of success in several iid Bernoulli trials.
- See Appendix page 8.

Example. In a study of the effectiveness of coupons offering a price reduction on a given product, 1,000 homes were selected at random. A packet containing advertising material and a coupon for the product were mailed to each home. The coupons offered different price reductions (5, 10, 15,20, and 30 dollars), and 200 homes were assigned at random to each of the price reduction categories. The predictor variable X in this study is the amount of price reduction, and the response variable Y is a binary variable indicating whether or not the coupon was redeemed within a six-month period.

Table 14.2 contains the data for this study. X_j denotes the price reduction offered by a coupon, n_j the number of households that received a coupon with price reduction X_j, $Y_{\cdot j}$ the number of these households that redeemed the coupon, and p_j the proportion of households receiving a coupon with price reduction X_j that redeemed the coupon.

Example 2: Binomial Outcomes

```
data ch14tab02;
input x n r p;
label x = 'Reduction'
n = 'no. households'
r = 'coupons redeemed'
p = 'proportion of coupons redeemed';
cards;
5 200 30 .150
10 200 55 .275
15 200 70 .350
20 200 100 .500
30 200 137 .685
;
run;
```

```
proc genmod data=ch14tab02;
model r/n = x / dist = bin link = logit lrci;
run;
```

License