Applied Regression

Generalized Linear Model - GLM - Part-1

Module 9 Lecture - 9-1

GENERALIZED LINEAR MODEL

To understand generalized linear model, we first need to know, what is "exponential family".

Distributions that are members of the exponential family have the general form

$$f(y_i, \theta_i, \phi) = e^{\frac{(y_i\,\theta_i - b(\theta_i))}{a(\phi)} + h(y_i, \phi)}$$

where ϕ is a scale parameter and θ is called the natural location parameter.

For members of the exponential family,

$$E(Y) = \mu = \frac{db(\theta_i)}{d\theta_i} \quad \& \quad Var(Y) = \frac{d^2 b(\theta_i)}{d\theta_i^2}\, a(\phi) = \frac{d\mu}{d\theta_i} a(\phi)$$

Example:

$$\text{Normal:} \quad f(y_i, \theta_i, \phi) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\; exp\Big(-\frac{1}{2\,\sigma^2}\,(y-\mu)^2\Big)$$
$$= exp\Big[\frac{1}{\sigma^2}\Big(y\mu - \frac{\mu^2}{2}\Big) - \frac{y^2}{2\,\sigma^2} - \frac{1}{2}\,ln(2\,\pi\,\sigma^2)\Big]$$

$$\text{Hence} \quad \theta_i = \mu \;\; b(\theta_i) = \frac{\mu^2}{2},\;\; a(\phi) = \sigma^2,$$

$$\text{Implies} \quad E(Y) = \mu = \frac{db(\theta_i)}{d\theta_i} = \mu,\;\; Var(Y) = \frac{d^2 b(\theta_i)}{d\theta_i^2}\,a(\phi) = \frac{d\mu}{d\theta_i} a(\phi) = \sigma^2$$

$$\text{Binomial:} \quad f(y_i, \theta_i, \phi) = \binom{n}{k} \pi^y \, (1-\pi)^{n-y}$$
$$= exp\Big[y \, ln\Big(\frac{\pi}{1-\pi}\Big) \; + \; n \, ln(1-\pi) \; + \; ln\binom{n}{y} \Big]$$

$$\text{Hence} \quad \theta_i = ln\Big(\frac{\pi}{1-\pi}\Big), \;\; \pi = \frac{exp(\theta_i}{1 \; + \; exp(\theta_i)} \;\; b(\theta_i) = -n \, ln(1-\pi), \;\; a(\phi) = 1,$$

$$\text{Implies} \quad E(Y) = \mu = \frac{db(\theta_i)}{d\theta_i} = n \, \pi, \;\; Var(Y) = n \, \pi \, (1-\pi).$$

Similarly we can show that Poisson and Gamma are part of the exponential family.

The basic idea of a GLM is that an appropriate function of the expected value of the response variable is a linear function of the independent variables. . Let η_i be the linear predictor which is some function of expected value. i.e

$$\eta_i = g(E(Y_i)) = \mathbf{x}'\beta = \beta_\mathbf{0} + \beta_\mathbf{1}\,\mathbf{x_1} + \ldots + \beta_\mathbf{k}\,\mathbf{x_k}$$

The function g is called the link function. There are many possible choices of the link function. Note that for MLR model the link function is identity.

If we choose we say that $\eta_i = \theta_i$, then η_i is called the canonical link. Table below shows the canonical links for the most common choices of distributions employed with the GLM.

Distribution	Canonical Link	Name
Normal	$\eta_i = \mu_i$	Identity Link
Binomial	$\eta_i = ln\left(\frac{\pi}{1-\pi}\right)$	Logistic Link
Poisson	$\eta_i = ln(\lambda)$	Log Link
Exponential	$\eta_i = \frac{1}{\lambda_i}$	Reciprocal Link
Gamma	$\eta_i = \frac{1}{\lambda_i}$	Reciprocal Link

LOGISTIC REGRESSION

We start the GLM with logistic link function which is popularly known as logistic regression. This is a situation where the response variable has only two possible outcomes, generically called success and failure and denoted by 0 and 1.

Generally, when the response variable is binary, there is considerable empirical evidence indicating that the shape of the response function should be nonlinear. A monotonically increasing (or decreasing) S-shaped (or reverse S-shaped) function, such as shown in Figure (next slide), is usually employed.

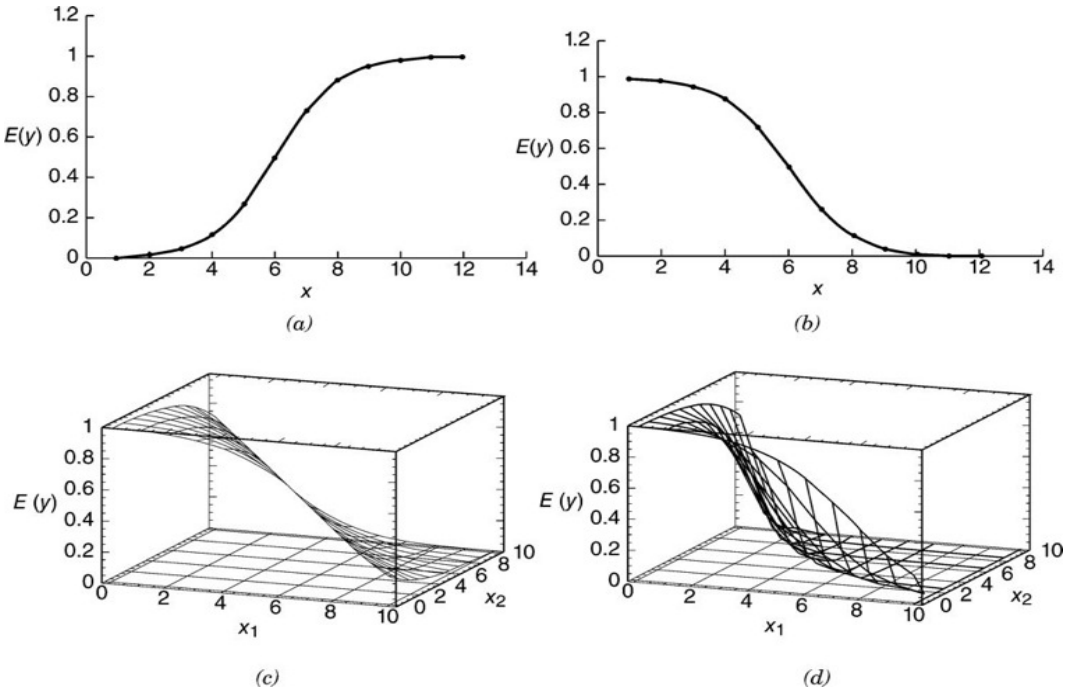
This function is called the logistic response function which corresponds to Logistic link function. The form is

$$E(y_i) = \pi = \frac{exp(x'\beta)}{1 + exp(x'\beta)} = \frac{1}{1 + exp(-x'\beta)}$$

It implies that $\eta = ln\left(\frac{\pi}{1-\pi}\right) = x'\beta$

This transformation is often called the logit transformation of the probability π and the ratio $\pi/(1-\pi)$ in the transformation is called the odds.

Logistic Curve



Sometimes the logit transformation is called the log-odds.

Estimation in Logistic Model

The regression parameters are estimated using maximum likelihood technique. Note that y_i follows Bernoulli distribution with

$$f(y_i) = \pi^{y_i} (1 - \pi)^{1-y_i} \quad \text{where} \quad i = 1, 2, \ldots, n.$$

Now the likelihood function is

$$L(y_1, y_2, \ldots, y_n, \beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi^{y_i} (1-\pi)^{1-y_i}$$

Now if you take log of $L(y, \beta)$ and then replace the π in terms of $x' \beta$ we get

$$ln(L(y, \beta)) = \sum_{i=1}^{n} y_i \, x_i' \, \beta - \sum_{i=1}^{n} ln[1 + exp(x_i' \, \beta)]$$

The above expression needs to be maximized for for some specific values of β which is called maximum likelihood estimates (or MLE of β). Here we need the help of the softwares like SAS or R to find those estimates.

Let $\hat{\beta}$ be the final estimate of the model parameters that the above algorithm produces. If the model assumptions are correct, then we can show that asymptotically (for large n)

$$E(\hat{\beta}) = \beta \quad \text{where} \quad Var(\hat{\beta}) = (X'VX)^{-1}$$

where V is an $n \times n$ diagonal matrix containing the estimated variance of each observation on the main diagonal; that is, the ith diagonal element of V is

$$V_{ii} = n_i \hat{\pi}(1 - \hat{\pi})$$

The estimated value of the linear predictor is $\hat{\eta} = x_i' \hat{\beta}$ and the fitted value of the logistic regression model is written as

$$\hat{y}_i = \hat{\pi} = \frac{exp(x_i' \hat{\beta})}{1 + exp(x_i' \hat{\beta})}$$

Example: The data concerning the proportion of coal miners who exhibit symptoms of severe pneumoconiosis and the number of years of exposure. The data are shown below in Table. The response variable of interest, y, is the proportion of miners who have severe symptoms. A reasonable probability model for the number of severe cases is the binomial, so we will fit a logistic regression model to the data.

The Pneumoconiosis Data

Number of Years of Exposure	Number of Severe Cases	Total Number of Miners	Proportion of Severe Cases, y
5.8	0	98	0
15	1	54	0.0185
21.5	3	43	0.0698
27.5	8	48	0.1667
33.5	9	51	0.1765
39.5	8	38	0.2105
46	10	28	0.3571
51.5	5	11	0.4545

The table below gives the parameter estimates (in SAS).

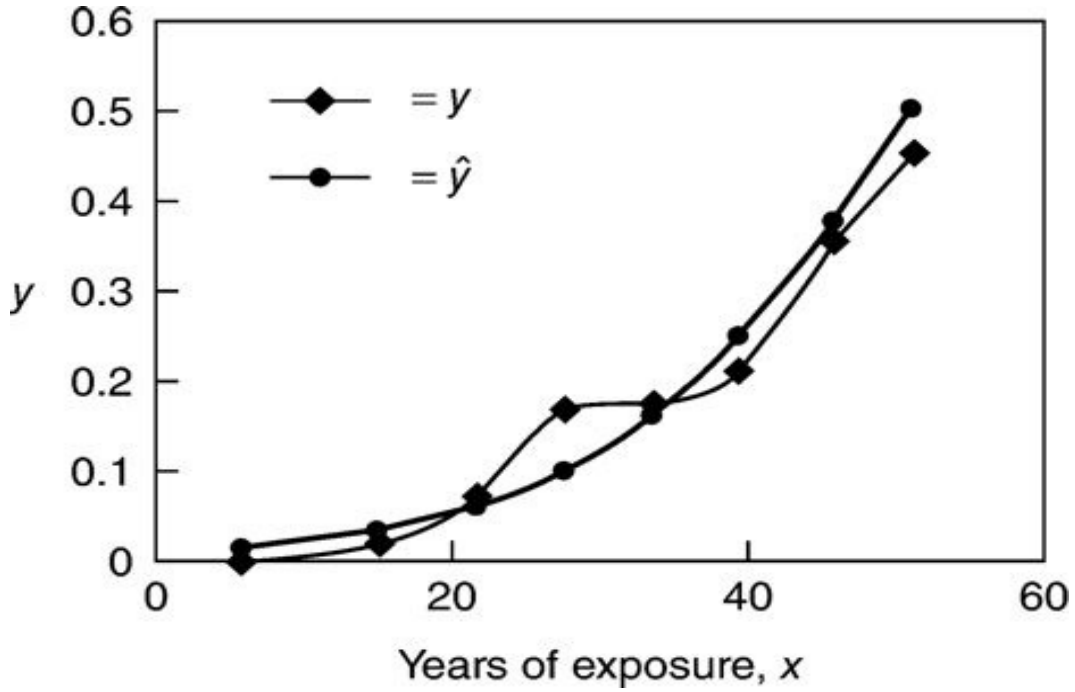
Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	SE	LCL	UCL	Chi-Square	p-value
Intercept	1	-4.7965	0.5686	-5.9109	-3.6821	71.16	$< .0001$
Years	1	0.0935	0.0154	0.0632	0.1237	36.71	$< .0001$

Hence the the fitted logistic regression model is

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{4.7965 - 0.0935x}}$$

A graph of the response variable versus the number of years of exposure is.



Interpretation of the Parameters:

$$\text{Odds} = \frac{\text{Probability of the event}}{\text{Probability of Non-Event}} \;\Rightarrow\; \text{Odds Ratio} = \frac{\text{Odds at } x = x_i + 1}{\text{Odds at } x = x_i}$$

$$\text{Now the log-Odds is } \eta = ln\Big(\frac{\pi}{1-\pi}\Big) = \beta_0 + \beta_1 x$$

.

$$\text{Hence} \quad \eta_{x_i+1} - \eta_{x_i} = ln(\text{Odds at } x_i + 1) - ln(\text{Odds at } x_i)$$

$$= ln\left(\frac{\text{Odds at } x_i + 1}{\text{Odds at } x_i}\right) = \beta_1$$

As a result $\hat{\beta}_1$ is the estimated log of Odds Ratio (log-Odds) (or estimated change in log-odds). So

$$\text{Est. Odds Ratio} = e^{\hat{\beta}_1} \;\Rightarrow\; \text{Est. \% Change in Odds} = 100(e^{\hat{\beta}_1} - 1)\%$$

In our example the linear predictor contains only one regressor variable (i.e Years Exposed). So the estimated Odds-ratio is

$$\widehat{\text{OR}} = e^{\hat{\beta}_1} = e^{0.0935} = 1.10$$

This implies that every additional year of exposure increases the odds of contracting a severe case of pneumoconiosis by 10%.

If the exposure time increases to 10 years, then the odds ratio becomes

$$\widehat{\text{OR}}(10) = e^{10\times\hat{\beta}_1} = e^{10\times 0.0935} = 2.55$$

This indicates that the odds more than double with a 10-year exposure.

Inference about the model parameters

To find a $100(1-\alpha)\%$ confidence interval we can use

$$\hat{\beta}_i \pm z_{\alpha/2}\sqrt{Var(\hat{\beta}_i)}$$

where $Var(\hat{\beta}_i)$ is the ith diagonal element of the variance-covariance matrix

$$Var(\hat{\beta}) = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) & . & Cov(\hat{\beta}_0,\hat{\beta}_n) \\ Cov(\hat{\beta}_0,\hat{\beta}_1) & Var(\hat{\beta}_1) & . & Cov(\hat{\beta}_1,\hat{\beta}_n) \\ . & . & Var(\hat{\beta}_i) & . \\ Cov(\hat{\beta}_0,\hat{\beta}_n) & Cov(\hat{\beta}_1,\hat{\beta}_n) & . & Var(\hat{\beta}_n) \end{pmatrix}$$

To test the significance regrading β_i we use the statistic (for large samples)

$z-stat = \frac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}}$ and we reject H_0 if $|z-stat| > z_{\alpha/2}$

Remark: Same statistic can be used to test one-sided hypothesis with appropriate rejection region on the same side.

In our example, Estimated Covariance Matrix is

$$Var(\hat{\beta}) = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} = \begin{pmatrix} 0.32330 & -0.008348 \\ -0.008348 & 0.0002380 \end{pmatrix}$$

To test whether risk of the severe pneumoconiosis increases over the years we need to test

$H_0 : \beta_1 \leq 0 \quad vs \quad H_1 : \beta_1 > 0$

$z - stat = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{Var(\hat{\beta}_1)}} = \frac{0.0935 - 0}{\sqrt{0.0002380}} = 6.06$

As $z - stat > z_{0.05} = 1.645$ we reject the H_0.

Remark: As $Z^2 = \chi^2_1$, for two sided hypothesis (i.e $H_0 : \beta_i = 0 \quad vs \quad \beta_i \neq 0$), it is possible to use $(z - stat)^2$ as the statistic and check if

$(z - stat)^2 > \chi^2_{0.05,1}$ to reject H_0. (as done by SAS in the previous table).

Also 95% confidence interval for β_1 is

$$\hat{\beta}_i \pm z_{\alpha/2} \sqrt{Var(\hat{\beta}_i)} = 0.0935 \pm 1.96 \times \sqrt{0.0002380} = (0.0632, 0.1237)$$

Testing Goodness of Fit

The goodness of fit of the logistic regression model can be assessed using two kinds of residuals (1) Deviance Residual (2) Pearson Residual.

Likelihood ratio test procedure, which uses the deviance residual, compares the current model to a saturated model, where each observation (or group of observations when $n_i > 1$) is allowed to have its own parameter (that is, a success probability). These parameters or success probabilities are y_i/n_i, where y_i is the number of successes and n_i is the number of observations.

Then ith Deviance Residual is defined as

$$d(y_i, \hat{\pi}_i) = \pm \left\{ 2 \left[y_i \, ln\left(\frac{y_i}{n_i \, \hat{\pi}_i}\right) + (n_i - y_i) \, ln\left(\frac{(n_i - y_i)}{n_i \, (1 - \hat{\pi}_i)}\right) \right] \right\}^{1/2}$$

where $\pm$ sign is the sign of $(y_i - n_i \, \hat{\pi}_i)$.

The **deviance** statistic is defined as twice the difference in log-likelihoods between this saturated model and the full model (which is the current model) that has been fit to the data.

The deviance is defined as

$$
\begin{aligned}
D &= 2.ln\left(\frac{\text{L(saturated model)}}{\text{L(Current model)}}\right) \\
&= 2\sum_{i=1}^{n}\left[y_i\, ln\left(\frac{y_i}{n_i\,\hat{\pi}_i}\right) + (n_i - y_i)\, ln\left(\frac{(n_i - y_i)}{n_i\,(1-\hat{\pi}_i)}\right)\right] \\
&= \sum_{i=1}^{n}(d(y_i, \hat{\pi}_i))^2
\end{aligned}
$$

where $\hat{\pi}_i = \frac{exp(x'\,\beta)}{1+exp(x'\,\beta)}$)

In calculating the deviance, note that if $y_i = 0$ then the first term is 0 and if $y_i = n_i$ then the second term is 0.

When the logistic regression model is an adequate fit to the data and the sample size is large, the deviance has a chi-square distribution with (n - p) degrees of freedom, where p is the number of parameters in the model. Small values of the deviance (or a large P value) imply that the model provides a satisfactory fit to the data, while large values of the deviance imply that the current model is not adequate. A good rule of thumb is to divide the deviance by its number of degrees of freedom. If the ratio D/(n - p) is much greater

than unity, the current model is not an adequate fit to the data.

Goodness of fit can also be assessed with a Pearson chi-square statistic that compares the observed and expected probabilities of success and failure at each group of observations. The expected number of successes is $n_i\,\hat{\pi}_i$ and the expected number of failures is $n_i\,(1-\hat{\pi}_i)$.

The Pearson Residual is defined as

$$r(y_i, \hat{\pi}_i) = \frac{(y_i - n_i\,\hat{\pi}_i)}{\sqrt{(n_i\,\hat{\pi}_i)\,(1-\hat{\pi}_i)}}$$

Then Pearson chi-square statistic is

$$\chi^2 = \sum_{i=1}^{n} (r(y_i, \hat{\pi}_i))^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\,\hat{\pi}_i)^2}{(n_i\,\hat{\pi}_i)\,(1-\hat{\pi}_i)}$$

The Pearson chi-square goodness-of-fit statistic can be compared to a chi-square distribution with n - p degrees of freedom. Small values of the statistic (or a large P value) imply that the model provides a satisfactory fit to the data. The Pearson chi-square statistic can also be divided by the number of

degrees of freedom n p and the ratio compared to unity. If the ratio greatly exceeds unity, the goodness of fit of the model is questionable.

When there are no replicates on the regressor variables, the observations can be grouped to perform a goodness-of-fit test called the Hosmer-Lemeshow test. In this procedure the observations are classified into g groups based on the estimated probabilities of success. Generally, about 10 groups are used (when g = 10 the groups are called the deciles of risk) and the observed number of successes O_j and failures ($N_j O_j$) are compared with the expected frequencies in each group,$N_j\,\bar{\pi}_j$ and $N_j\,(1-\bar{\pi}_j)$ where N_j is the number of observations in the jth group and the average estimated success probability in the jth group is $\bar{\pi}_j$.

The HosmerLemeshow statistic is really just a Pearson chi-square goodness-of-fit statistic comparing observed and expected frequencies:

$$HL = \sum_{i=1}^{g} \frac{(O_j - N_j\,\bar{\pi}_j)^2}{N_j\,\bar{\pi}_j(1-\bar{\pi}_j)}$$

If the fitted logistic regression model is correct, the HL statistic follows a chi-square distribution with g - 2 degrees of freedom when the sample size is

large. Large values of the HL statistic imply that the model is not an adequate fit to the data. It is also useful to compute the ratio of the HosmerLemeshow statistic to the number of degrees of freedom g - p with values close to unity implying an adequate fit.

For our example, here are the Goodness-of-fit statistics value. Observe that it is fairly close to 1 which is our usual cutoff. The p-values are much greater than significance level which indicates that there is no strong evidence that data is not fitting the model.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	p-value
Deviance	6.0508	6	1.0085	0.4175
Pearson	5.0283	6	0.8381	0.5402
HosmerLemeshow	5.0034	5		0.4155