Generalized Linear Model

Hao Zheng
December 4, 2009
haozheng@stat.wisc.edu
What is GLM?

- GLM is a flexible generalization of ordinary least squares regression.
- “Generalized Linear Models”
  - John Nelder and Robert Wedderburn (1972)
- By using a link function, the response is sort of linearly related to the predictors.
  \[ g(EY) = X\beta \]
- Stat 851 – Generalized Linear Models
Why use GLM?

- Explore the relationship?
- Perform the prediction?

- If response y is binary, the result from logistic regression is probability, while the ordinary linear regression gives?

- Assumptions of residuals $y_i - g^{-1}(x_i\beta)$. 
Link Function

- canonical
  - \texttt{family=gaussian(link=“Identity”)}
- logit
  - \texttt{family=binomial(link=“logit”)}
- probit
  - \texttt{family=binomial(link=“probit”)}
- poission
  - \texttt{family=poisson(link=“log”)}
- gamma
  - \texttt{family=Gamma(link=“inverse”)}
- Inverse-gamma
  - \texttt{family=inverse.gaussian(link=“1/ mu^2”)}
- ...
Logistic Regression
Model

\[ \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + x_i \beta_1 \]

\[ \pi_i = \frac{\exp(\beta_0 + x_i \beta_1)}{1 + \exp(\beta_0 + x_i \beta_1)} \]
Maximization  
\[ \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \]

Y~Bernoulli(n, \( \pi_i \)), and  
\[ \pi_i = \frac{1}{\exp(x_i^T \beta)} \]

Log-likelihood  
\[ l(\beta) = \sum_{i=1}^{n} y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^{n} \log (1 - \pi_i) \]
\[ = \sum_{i=1}^{n} y_i \cdot x_i^T \beta - \sum_{i=1}^{n} \log (1 + \exp(x_i^T \beta)) \]

Newton-Raphson algorithm (or others) can be used to get the numerical solution
\[ \beta = \beta_0 - \left( \frac{\partial^2}{\partial \beta \partial \beta^T} l(\beta) \right)_{\beta_0}^{-1} \frac{\partial l(\beta)}{\partial \beta} \bigg|_{\beta_0} \]
Residuals

\[ \hat{e}_i = y_i - \hat{\pi}_i \]

Pearson Residuals

\[ \hat{r}_{pi} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)}} \]

Studentized Pearson Residuals

\[ \hat{r}_{spi} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)(1 - h_{ii})}} \]

where

\[ H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2} \]

\[ W = \text{diag}(\hat{\pi}_i (1 - \hat{\pi}_i)) \]
Deviance

\[ D^2 = -2 \sum_{i=1}^{n} \left( y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i) \right) \]

Deviance Residual

\[ D = \text{sign}(y_i - \hat{\pi}_i) \sqrt{-2 \sum_{i=1}^{n} \left( y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i) \right)} \]
The board of directors of a professional association conducted a random sample survey of 30 members to assess the effects of several possible amounts of dues increase. Two values are recorded: (1) The dollar increase in annual dues posited in the survey interview; (2) Whether or not the interviewee indicated that the membership will not be renewed at that amount of dues. Consider fitting a logistic regression model to these data treating (2) as the response and (1) as the covariate.

\[
Pr(\text{Renew} | \text{Amount in increase}) = \text{logit}(\beta_0 + \beta_1(\text{Amount in increase})).
\]

(a) Find the maximum likelihood estimates of \(\beta_0\) and \(\beta_1\) of the logistic regression model.

(b) Obtain a scatter plot of the data with both the fitted logistic response from part (a) and a lowess smooth superimposed. Does the fitted logistic response function appear to fit well?

\[
> \text{glm}(y \sim x, \text{binomial})
\]
> x
[1] 30 30 30 31 32 33 34 35 35 36 37 38 39 40 40 41 42 43 44 45 45 46 47 48 49 50 50
> y
[1] 0 1 0 0 0 0 1 0 0 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 1 1
> glm(y~x,binomial)

Call: glm(formula = y ~ x, family = binomial)

Coefficients:
(Intercept)       x
      -4.8075     0.1251

Degrees of Freedom: 29 Total (i.e. Null);  28 Residual
Null Deviance:    41.46
Residual Deviance: 37.46   AIC: 41.46

> pchisq(41.46-37.46,1,lower.tail=F)
[1] 0.04550026
```r
> fit = glm(y~x, binomial)
> par(mfrow=c(2,2))
> plot(fit$fitted.values, residuals(fit, type="working"))
> lines(lowess(fit$fitted.values, residuals(fit, type="working")))
> plot(fit$fitted.values, residuals(fit, type="pearson"))
> lines(lowess(fit$fitted.values, residuals(fit, type="pearson")))
> plot(fit$fitted.values, residuals(fit, type="deviance"))
> lines(lowess(fit$fitted.values, residuals(fit, type="deviance")))
```
> summary(fit)

Call:
glm(formula = y ~ x, family = binomial)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.7651  -1.0012   0.6502   0.9828   1.6455

Coefficients:
             Estimate Std. Error   z value Pr(>|z|)
(Intercept) -4.80751   2.65576  -1.8100  0.0703 .
x           0.12508   0.06676   1.8741  0.0610 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 37.465  on 28  degrees of freedom
AIC: 41.465

Number of Fisher Scoring iterations: 4
Poisson Regression
A health insurance company collected information on 788 of its subscribers who had made claims resulting from ischemic (coronary) heart disease. Data were obtained on totals costs of services provided for these 788 subscribers and the nature of the various services for the period of January 1, 1998 through December 31, 1999. Each line in the dataset has identification number and provides information on 9 other variables for each subscriber. The 10 variables are (in the order of columns that appear in the dataset):

- Identification number: 1-788.
- Total cost: Total cost of claims by subscriber (dollars).
- Age: Age of subscriber (years).
- Gender: Gender of subscriber: 1 = male, 0 = female.
- Interventions: Total number of interventions or procedures carried out.
- Drugs: Number of tracked drugs prescribed.
- Emergency room visits: Number of emergency room visits.
- Complications: Number of other complications that arose during heart disease treatment.
- Comorbidities: Number of other diseases that the subscriber had during period.
- Duration: Number of days of duration of treatment condition.

Consider modeling the number of emergency room visits as a function of other variables in the dataset.

(a) Obtain the fitted Poisson model using all the variables. State the estimated coefficients, their estimated standard errors, and the estimated response function.

(b) Comment on the adequacy of the fit of the Poisson regression model based on the deviance residuals from the fit.
Model

\[ P(y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \]

Link function

\[ \log(\lambda_i) = x_i^T \beta \]

Likelihood

\[ \prod_{i=1}^{n} \frac{\exp\left(y_i \cdot x_i^T \beta - \exp(x_i^T \beta)\right)}{y_i!} \]

Estimation based on

\[ \sum_{i=1}^{n} (e^{x_i^T \beta} - y_i) \cdot x_i = 0 \]
```r
> summary(glm(visits~.,data=dat,family=poisson))

Call:
glm(formula = visits ~ ., family = poisson, data = dat)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.6851 -1.0341 -0.2372  0.5845  5.7602

Coefficients:         Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.994e-01  1.761e-01  2.837  0.00456 **
cost          1.495e-05  2.855e-06  5.237  1.63e-07 ***
age           6.724e-03  2.967e-03  2.266  0.02346 *
gender        1.819e-01  4.400e-02  4.135  3.55e-05 ***
intervention  1.007e-02  3.808e-03  2.646  0.00816 **
drug          1.932e-01  1.268e-02 15.234  < 2e-16 ***
complication  6.125e-02  5.995e-02  1.022  0.30689
comorbidities -8.999e-04  3.685e-03 -0.244  0.80708
duration      3.529e-04  1.899e-04  1.859  0.06308 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1485.0  on 787  degrees of freedom
Residual deviance: 1043.6  on 779  degrees of freedom
AIC: 3271.0

Number of Fisher Scoring iterations: 5
```r
> mle.stepwise(fit2,type="Stepwise")

Call:
  mle.stepwise(formula = fit2, type = "Stepwise")

Stepwise selection procedure

F.in:  4
F.out:  4

Last 8 iterations:
  (Intercept) cost age gender intervention drug complication comorbidities duration
[1,]  1  0  0  0   0  0  0  0  0  1  16.810
[2,]  1  0  0  0   0  1  0  0  0  1  298.600
[3,]  1  1  0  0   0  1  0  0  0  1  90.340
[4,]  1  1  0  1   0  1  0  0  0  1  12.550
[5,]  1  1  1  1   0  1  0  0  0  1  4.579
[6,]  0  1  1  1   0  1  0  0  0  1  1.676
[7,]  0  1  1  1   1  1  0  0  0  1  4.151
[8,]  0  1  1  1   1  1  0  0  0  0  2.342
```
> step(fit2)
Start: AIC=3271.03
visits ~ cost + age + gender + intervention + drug + complication + comorbidities + duration

Df Deviance AIC
- comorbidities 1 1043.7 3269.1
- complication 1 1044.6 3270.1
<none> 1043.6 3271.0
- duration 1 1047.1 3272.5
- age 1 1048.8 3274.2
- intervention 1 1050.3 3275.8
- gender 1 1060.2 3285.6
- cost 1 1070.2 3295.6
- drug 1 1238.5 3463.9

Step: AIC=3269.09
visits ~ cost + age + gender + intervention + drug + complication + duration

Df Deviance AIC
- complication 1 1044.7 3268.1
<none> 1043.7 3269.1
- duration 1 1047.5 3270.9
- age 1 1048.8 3272.2
- intervention 1 1050.4 3273.8
- gender 1 1060.5 3283.9
- cost 1 1070.2 3293.6
- drug 1 1245.3 3468.7

Step: AIC=3268.1
visits ~ cost + age + gender + intervention + drug + duration
Thank you!

Good luck to your exams!