GLMs Summary

← "marginal"

Tuesday, October 10, 2023 16:24

GLM: Generalized linear model, extends linear regression to other distributions
e.g. response var is generated from distribution in the exponential family (as opposed to normal, like in linear reg.)

Involves:
① exponential family of probability distributions
② $x_i'\beta$: A linear model
③ Link function & its inverse relates linear model to expectation:

$E[y_i \mid x_i] = \mu_i = g^{-1}(x_i'\beta)$
$Var[y_i \mid x_i] = v(\mu_i) = V(g^{-1}(x_i'\beta))$

does not involve noise variance (σ^2)

so find g such that $g(E[y_i \mid x_i]) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$

→ then, $\mu_i = g^{-1}\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}\right)$

→ should have a range $(-\infty, \infty)$ since β_0 and x_{ik} can be anything

→ domain is possible values of μ_i

Exponential family:

DENSITY function: $f_\theta(y) = \exp\left[\left[y\theta - b(\theta)\right]/a(\phi) + c(y,\phi)\right]$

b, a, c: known functions

ϕ = SCALE PARAMETER (doesn't generalize)

θ = the only unknown parameter
In GLM, will be a function of $x_i'\beta$

Two most common models: BINARY & POISSON outcomes

Estimating Equations

What do the above three score equations have in common?

Gaussian:

$$\sum_{i=1}^{n} x_{ik} \frac{1}{v_i}(y_i - x_i'\beta) = 0.$$

Logistic:

$$\sum_{i=1}^{n} x_{ik}\left(y_i - \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right) = 0.$$

Poisson:

$$\sum_{i=1}^{n} x_{ik}\left(y_i - e^{x_i'\beta}\right) = 0.$$

It can be shown that for any generalized linear model, the score equations we need to solve for the maximum likelihood estimate are given by

$$\sum_{i=1}^{n} \frac{\partial E(y_i;\beta)}{\partial \beta_k} \frac{1}{V(y_i;\beta)}\{y_i - E(y_i;\beta)\} = 0. \qquad (2)$$

46/57 Module 4: Generalized Estimating Equations

for BINARY outcomes

$E[y_i] = \mu_i = p_i \in (0,1)$
$Var[y_i] = V[\mu_i] = V[p_i] = p_i(1-p_i)$

$y_i \overset{ind}{\sim}$ Bernoulli (p_i) and $y_i \mid p_i$ are independent for $i = 1, \dots, n$

$y_i \in \{0,1\}$ $\quad P(y_i = 1)$

Probability mass function:

$f(y \mid p_i) = p_i^{y}(1-p_i)^{1-y}$

Link functions:

① LOGISTIC

$y_i \overset{ind}{\sim}$ Bernoulli (p_i)

$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$

always increasing

NO CHOICE HERE

$g(p_i)$ = log odds of success probability

When $p_i \to 0$, $\log\left(\frac{p_i}{1-p_i}\right) \to -\infty$
When $p_i \to 1$, $\log\left(\frac{p_i}{1-p_i}\right) \to \infty$

Likelihood function:

$L(\beta; y, x) = \log \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$ $\qquad$ where $p_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$ (logistic function)

$= \sum_{i=1}^{n} y_i \log p_i + (1-y_i)\log(1-p_i)$

$= \sum_{i=1}^{n} y_i \log\left(\frac{e^{\beta' x_i}}{1+e^{\beta' x_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\beta' x_i}}\right)$

$= \sum_{i=1}^{n} y_i \beta' x_i - \log(1 + \exp(\beta' x_i))$

Logistic function:

$p(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} = \frac{1}{1+\exp(-\eta)}$ $\qquad (-\infty, \infty) \to (0,1)$

Logit function

$g(p) = \log\left(\frac{p}{1-p}\right)$ $\qquad (0,1) \to (-\infty, \infty)$

Interpretation:

β_0 = baseline log odds = $\log\left(\frac{p_i}{1-p_i}\right)$ when $x = 0$

β_1 = the change in log odds per unit change in x_i (β_k = log odds ratio associated w/ covariate k while controlling for other covariates)
= the log odds ratio per unit change in x_i
$= \log\left[\frac{p_i'/(1-p_i')}{p_i/(1-p_i)}\right]$

e^{β_1} = odds ratio
$= \frac{p_i'(1-p_i')}{p_i/(1-p_i)}$ $\Rightarrow$ CI: $e^{(\beta_1 \pm 1.96 \cdot SE)}$

$p_i = \mu_i$ = predicted case estimated value (e.g. $P(Y = \text{pop} \mid \text{treatment} = \text{diff})$)
$= \frac{e^{\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}}}{1 + e^{\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}}}$ (e.g. baseline probability of _)

Inference:

Likelihood Ratio Tests:

$H_0: \beta_0 = 0$
to assess coefficients of interest

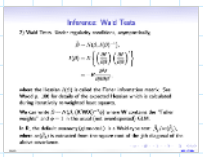
$\ell(\hat{\beta}_{full})$ = log-likelihood of full model
$\ell(\hat{\beta}_{reduced})$ = log-likelihood of reduced model

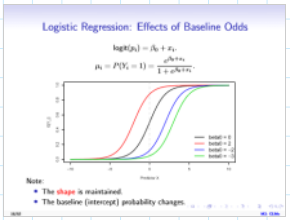
Reject H_0 if

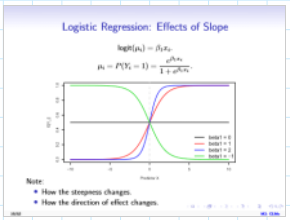
$-2\{\ell(\hat{\beta}_{reduced}) - \ell(\hat{\beta}_{full})\} > \chi^2_{p_1 - p_0}$

diff in # parameters between full & reduced

Wald Tests:
see slide 38 of GLM1 part 2







for POISSON outcomes

$E[y_i] = \mu_i = \lambda_i > 0$
$Var[y_i] = V[\mu_i] = V[\lambda_i] = \lambda_i$ $\Big\}$ MEAN = VARIANCE

$P(y_i = m) = \frac{\lambda_i^m e^{-\lambda_i}}{m!}$

$y_i \overset{ind}{\sim}$ Poisson (λ_i)

$\log(\lambda_i) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$

Link function is log()
- domain $(0, \infty)$
- Range $(-\infty, \infty)$
- Strictly increasing

Interpretation:

Consider a simple Poisson regression model with only one covariate:

$$y_i \overset{ind}{\sim} \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 x_i.$$

When $x_i = 0$, $\log(\lambda_i) = \beta_0$

- $e^{\beta_0} = \lambda_0$ is the baseline expected counts.

one covariate to start

λ_i: rate.

Now consider a unit change in x

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i \qquad \log(\lambda_i^*) = \beta_0 + \beta_1(x_i + 1).$$

Note that

$$\beta_1 = \log(\lambda_i^*) - \log(\lambda_i)$$

- $e^{\beta_1} = \lambda_i^*/\lambda_i$ is the relative change in count per unit change in x. Also called the relative rate. Also called the incident rate ratio.
- For continuous variables with $\beta_1 > 0$, the rate increases by $100 \times (e^{\beta_1} - 1)\%$ for every unit increase in x.
- For factors with $\beta_1 > 0$, the rate increases by $100 \times (e^{\beta_1} - 1)\%$ for level j relative to baseline.

Covariate impacts are multiplicative rather than additive (applies to log models in general) on the count scale.

$$E\,y_i = e^{\beta_0}e^{\beta_1 x_{i1}} \cdots e^{\beta_p x_{ip}}$$

Pct change

$$100\left(\frac{e^{\beta_0 + \beta_1(x_i + 1)} - e^{\beta_0 + \beta_1 x_i}}{e^{\beta_0 + \beta_1 x_i}}\right)$$

aka

$$100\left(e^{\beta_1} - 1\right)$$

INTERCEPT ONLY POISSON MODEL: "The log expected HIV case counts for a state without considering socioeconomic factors (the other covariates) is 6.65. The expected HIV case counts for a state without considering socioeconomic factors(the other covariates) is 773.0784."

INTERCEPT ONLY POISSON MODEL WITH OFFSET: "Next, fit a Poisson model with offset log Pop100k. Interpret the intercept on the log-scale and back-transformed scale.
The log expected case rate per 100k is 2.68 without considering other covariates.

The expected case rate per 100k is 14.6 without considering other covariates."

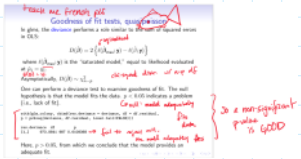
The usual plot of residual versus fitted is not useful in GLMs because of the relationship between the mean and variance.

The deviance residuals have an approximate normal distribution.

The deviance residual is

$$d_i^* = \text{sign}(y_i - \mu_i)\sqrt{d_i}.$$

where d_i is the ith term in the calculation of the deviance. For details.



In glm, the deviance performs a role similar to the sum of squared errors in OLS:

$$D(B) = 2\left(\ell(B_{sat}; y) - \ell(B; y)\right)$$

where $\ell(B_{sat}; y)$ is the "saturated model" equal to likelihood evaluated at $y_i = \frac{y_i}{\phi(1-\phi)}$.
Asymptotically, $D(B) \sim \chi^2_{n-p}$

One can perform a deviance test to examine goodness of fit. The null hypothesis is that the model fits the data. $p > 0.05$ indicates a problem (i.e., lack of fit).

estimate values, residual deviance = deviance, df = df residual, p = pchisq(deviance, df residual, lower.tail=FALSE)

```
> 1-pchisq(1.5, df=1)
[1] 0.2206572
```

Here, $p > 0.05$, from which we conclude that the model provides an adequate fit.

Overdispersion

In Poisson, the assumption that $\phi(\mu_i) = \phi(\mu_i)$ is often violated.

One can add an additional overdispersion parameter, also called a scale parameter.

One can adjust the parameter variances by the scale parameter:

$$\hat{\beta} \sim N(B, I(B)^{-1}\phi)$$

We will see this again in GEEs.

Section 3.1.5 in Wood describes three different estimators of ϕ.

We want overdispersion parameter to be close to 1.
>1 means overdispersed. And model is not a good fit
<1 means underdispersed

t values are only approximate.

Interpretation:

Estimates: for "log rate is"
The log of the expected count is [] in [baseline e.g. Urban master no mod].
The estimated # of _ and 95% CI for this baseline is $e^{()}$ and $e^{()}$ [—,1.96(SE)]

Interaction terms:
The multiplicative effect of the interaction between _ is e^{β} i.e. the rate increases by e^{β} relative to no interaction.

Ratios:
$e^{\text{Rate estimate}_1}/e^{\text{Rate estimate}_2}$ = # times as many _ compared to _

If y_i is the case count for state i, then y_i is independently distributed as a Poisson distribution with expected HIV case count λ_i.

The model is:

$$log(\lambda_i) = log(Pop100k_i) + \beta_0 + \beta_1 Poverty_i^* + \beta_2 NoDiploma_i^* + \beta_3 Uninsured_i^* + \beta_4 Vacant_i^*$$

$log(Pop100k)$ is the log of the population in hundred thousands. This variable is the offset to account for population size differences.

β_0 is the estimated log mean number of HIV cases for state i while adjusting for the log population size (Pop100k) for the average number of households living below federal poverty level, the average population 25 years and older without a HS diploma, the average number of uninsured, and the average number of vacant housing.

β_1 is the coefficient for the centered predictor variable Households living below the federal poverty level. e^{β_1} would be interpreted as the rate ratio of HIV cases for every one unit increase in Households living below the federal poverty level in a model controlling all other predictor variables.

GOF test:
H0: The model adequately fits the data.
H1: The model does not adequately fit the data.

Test statistic: Chi squared statistic of 6897.6641
Decision: Reject the null hypothesis
Conclusion: The model does not adequately fit the data.