Notes in Generalized Linear Models

Dennis (Shaoxiong) Zheng

Abstract

These notes of the course STAT 34700 Genearlized Linear Models in Winter 2020 Quarter at The University of Chicago are written based on the lectures taught by Professor Jingshu Wang as well as the textbook for this course *Foundations of Linear and Generalized Linear Models* by Alan Agresti.

The notes are wrtitten for review purpose. I follow the way of formal concept analysis (FCA) when naming the section names, so that one may look at the table of contents, lists of definitions, theorems and exmples, and then try to recall all the underlying details. If he can do so, he should be confident for the exam.

If you find any typo, please notify me at denniszheng@uchicago.edu. For more course notes, please visit my website.

Last update: February 24, 2020.

Contents

Part I
Genearlized Linear Models

1 General Setting of GLM

1.1 Structure

1.2 Estimation

1.3 Computation

1.4 Hypothesis Testing

1.5 Model Selection

2 Binary Data

3 Nominal Data

4 Ordinal Data

5 Count Data

5.1 Over-dispersion phenomenon

5.1.1 Definition: Actual $Var(y_i) > v^*(y_i)$

5.1.2 Detection: Plot $(y_i - \hat{\mu}_i)^2$ v.s. $\hat{v}^*(\hat{\mu}_i)$

Plot $(y_i - \hat{\mu}_i)^2$ v.s. $\hat{v}^*(\hat{\mu}_i)$. If the specification $Var(y_i) > v^*(y_i)$ is true, then the points should scatter around the line 45° line. If most points lie above the 45° line, then over-dispersion exists. Our assumption for the randomness of y_i is problematic.

5.1.3 Negative Binomial for Dispersed Counts

If $y_i \sim Poisson\,(\lambda_i)$ and $\lambda \sim \text{Gamma}\,(\mu_i, k_i)$, then $y_i \sim \text{NB}\,(\mu_i, k_i)$. We have

$$E(y_i) = \mu_i, Var(y_i) = \mu_i + \gamma_i \mu_i^2 > \mu_i$$

where $\gamma_i = 1/k_i$ is the dispersion parameter.

For NB GLM, We further assume $\gamma_i \equiv \gamma$ for all i, and the link is $\log(\mu_i) = X_i'\boldsymbol{\beta}$.

Note that when $\gamma = 0$, it is Poisson.

5.1.4 Beta-Binomial Model for dispersed Binary data

Asssume $ny \sim Bin(n, p)$ and $p \sim \text{Beta}\,(\alpha_1, \alpha_2)$. Let $\mu = \frac{\alpha_1}{\alpha_1+\alpha_2}$ and $\theta = \frac{1}{\alpha_1+\alpha_2}$, then the Beta-binomial distribution has the property

$$E(y) = \mu, \quad \text{Var}(y) = \left[1 + (n-1)\frac{\theta}{1+\theta}\right]\mu(1-\mu)/n > \mu(1-\mu) \text{ if } n > 1$$

The term $\rho = \frac{\theta}{1+\theta}$ is the measure of over-dispersion, It is also the correlation in completely dependent Bernoulli trials, where the variance function has the same form as here.

As $\theta \to 0$, the Beta distribution converges to a degenerate distribution at μ. Hence, the Beta-binomial distribuion converges to the $Bin(n, \mu)$.

We may use logit link,

$$\text{logit}\,(\mu_i) = X_i^T \beta$$

Both $\boldsymbol{\beta}$ and θ are unknown but we can estimate them using MLE.

Part II
Other Models and Methods

6 Quasi-Likelihood Methods

For a GLM $\eta_i = g(\mu_i) = X\boldsymbol{\beta}$, the likelihood equations are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{v\,(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = 0, \quad j = 1, \ldots, p$$

The choice of distribution for y_i determines the relation $v\,(\mu_i)$ between the variance and the mean.

An alternative approach, quasi-likelihood estimation does not assume the underlying distribution, but it assumes only a mean-variance relation.

6.0.1 Definition: replace $v\,(\mu_i)$ in score equations by other mean-variance relation.

Definition 1. Quasi-likelihood methods replace $v\,(\mu_i)$ in score equations by other mean-variance relation that typically involves another unknown dispersion parameter.

6.1 Proportional Mean-variance Relation $a(\mu_i, \phi) = \phi v^*(\mu_i)$

6.1.1 Definition

Definition 2. Suppose the original standard GLM model specifies a mean-variance relation $v^*(\mu_i)$. To allow for the actual variance to differ from $v^*(\mu_i)$ to model over/under-dispersion, now we assume $a(\mu_i, \phi) = \phi v^*(\mu_i)$.

The proportional mean-variance relationship is the easiest for the computation of as cancels and does not affect solving the "score" equations. For instance,

- Counts data: $a(\mu_i, \phi) = \phi \mu_i$
- Grouped binary data: $a(\pi_i, \phi) = \phi \pi_i (1 - \pi_i)/n_i$

There are two common ways of overdispersion in binomial data,

- $P(Y_i = 1) = \pi_i$ depends on unobserved variable, i.e. $\pi_i = \mathbf{X}\boldsymbol{\beta} + \epsilon$. To deal with this, we can use a hierachical mixture model that lets π_i itself have distribution. See Section 5.1.4.
- Bernoulli trials at each i are positively correlated. See more examples in Section 6.2.1.

6.1.2 Consequences

- Generalized standard variance function

  $$w_i^* = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{Var(y_i)}$$

  becomes

  $$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{\phi v^*(\mu_i)} = w_i^*/\phi$$

  In matrix form

  $$\mathbf{W} = \mathbf{D^2}(\phi\mathbf{V})^{\mathbf{-1}} = \mathbf{W}^*/\phi$$

- The variance of $\hat{\boldsymbol{\beta}}$ becomes

  $$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{XWX'})^{-1} = \phi Var^*(\hat{\boldsymbol{\beta}})$$

  Thus, se($\hat{\boldsymbol{\beta}}$) is inflated by $\sqrt{\phi}$.

6.1.3 Estimate of ϕ

When $X^{*2} = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{v^*(\hat{\mu}_i)}$ is approximately chi-squared (why?)

$$E(X^{*2}/\phi) \approx n - p$$

Using the motivation of estimation by matching momoents,

$$\hat{\phi} = \frac{X^{*2}}{n-p}$$

Thus, se($\hat{\boldsymbol{\beta}}$) estimates is inflated by $\sqrt{\frac{X^{*2}}{n-p}}$.

Above reasoning works only in the case when the structural relation between $E(y_i)$ and X holds. If not (e.g. missing a explanatory variable), adjusting for overdispersion will not address the inadequacy.

6.2 Other Forms of Mean-Variance Relation

In addition to proportional assumption, we may set

- Counts data: $a(\mu_i, \phi) = \mu_i + \phi\mu_i^2$
- Grouped binary data: $a(\mu_i, \phi) = [1 + \rho(n_i - 1)]\, \pi_i(1 - \pi_i)/n_i$

Why are they in this form? Because they are consistent wity Poisson-Gamma and Beta-Binomial.

When to use it? Plot the standardized residuals or Peason residuals for the ordinary binomial model against the indices n_i, if there is an increasing trend in the spread (why?), then Beta-Binomial-type variance function may be more appropriate.

6.2.1 Overdispersion Caused by Correated Bernoulli Trials

Example 3 (Overdispersion in Binomial)**.** In an election, perhaps in each household the head of the household decides how to vote, and then everyone else in the household votes the same way. Then the sample proportion in household i voting for a particular candidate has

$$P(y_i = 1) = \pi_i, \quad P(y_i = 0) = 1 - \pi_i$$

That is, y_i can take only its extreme possible values, i.e. $y_i \sim Ber(p_i)$, and thus $Var(y_i) = \pi_i(1-\pi_i) > \pi_i(1-\pi_i)/n_i$.

Example 4 (Underdispersion in Binomial)**.** Suppose that the observations occur sequentially and $y_{ij}|y_{i1}, \ldots, y_{i,j-1}$ equals $1 - y_{i,j-1}$

When n_i is an even number, $y_i = \frac{\sum_j y_{ij}}{n_i} = \frac{1}{2}$, $P(y_i = 1/2) = 1$ so $Var(y_i) = 0$, so there is underdispersion.

Example 5 (Completely dependent Bernoulli trials (exchangeability of trials))**.** Suppose one trail is completely dependent on another with a common correlation ρ between each pair of $\{y_{i1}, y_{i2}, \ldots, y_{in_i}\}$, as is often assumedi n cluster sampling. Then

$$\operatorname{var}(y_{it}) = \pi_i(1-\pi_i), \operatorname{cov}(y_{is}, y_{it}) = \rho\pi_i(1-\pi_i)$$

Hence

$$\begin{aligned}
\operatorname{var}(y_i) &= \operatorname{var}\left(\frac{\sum_{t=1}^{n_i} y_{it}}{n_i}\right) \\
&= \frac{1}{n_i^2}\left[\sum_{t=1}^{n_i} \operatorname{var}(y_{it}) + 2\sum_{s<t}\sum_{s<t} \operatorname{cov}(y_{is}, y_{it})\right] \\
&= \frac{1}{n_i^2}\left[n_i\pi_i(1-\pi_i) + n_i(n_i-1)\rho\pi_i(1-\pi_i)\right] \\
&= [1 + \rho(n_i - 1)]\frac{\pi_i(1-\pi_i)}{n_i}
\end{aligned}$$

Overdispersion occurs when $\rho > 0$.

6.2.2 QL for Correlated Bernoulli Trials

Motivated by the above example, we can set the mean-variance relation to be

$$v(\pi_i) = [1 + \rho(n_i - 1)]\, \pi_i(1-\pi_i)/n_i$$

with $|\rho| \leq 1$.

There is an approach to estimte ρ. By equating

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{[1 + \hat{\rho}(n_i - 1)]\, \hat{\pi}_i(1-\hat{\pi}_i)/n_i} = n - p$$

and then solve it for $\hat{\rho}$ and solve the score equation for $\hat{\boldsymbol{\beta}}$ iteratively.

6.3 Properties of Quasi-Likelihood Estimators

6.3.1 Estimating Equations

Definition 6 (Quasi-score Equations)**.** Quasi-score equations are defined as

$$\boldsymbol{u}(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}}\frac{(y_i-\mu_i)}{v\left(\mu_i\right)}=\mathbf{0}$$

Note that

$$\frac{\partial \mu_i}{\partial \beta_j}=\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}=\frac{\partial \mu_i}{\partial \eta_i}x_{ij}$$

So if $v(\mu_i)$ is the standard mean-variance relation, then this is the score equation.

Definition 7 (Estimating Equations)**.** If we use the quasi-score equations to determine $\hat{\boldsymbol{\beta}}$, then they are called estimating equaitons. It determines an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Definition 8 (Unbiased Estimating Function)**.** For any function $h(\mathbf{y};\boldsymbol{\beta})$, if $E[h(\mathbf{y};\boldsymbol{\beta})]=0$ for all $\boldsymbol{\beta}$, then the function is an unbiased estimating function.

6.3.2 QL Estimators: which solves the estimating equations.

Definition 9 (QL Estimators)**.** QL estimators maximize the a quasi-log-likelihood function. That is, it solves the estimating equations.

6.3.3 Property: Consistency

Theorem 10 (Consistency if $g\left(\mu_i\right)=\sum_j \beta_j x_{ij}$)**.** *Asuming $g\left(\mu_i\right)=\sum_j \beta_j x_{ij}$ is correct, QL estimator is consistent for $\boldsymbol{\beta}$ even if $v(\mu_i)$ is misspecified.*

6.3.4 Property: Under Correct Specification: Asymptotically Normality

Theorem 11 (Asymptotically Normality if $var(y_i)=v(\mu_i)$)**.** *When $var(y_i)=v(\mu_i)$, i.e. correct specification of the mean-variance relation, the QL estimators $\hat{\boldsymbol{\beta}}$ are aysmptotically normal with a model-based covariance matrix approximated by*

$$\boldsymbol{V}=\left[\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}}\left[v\left(\boldsymbol{\mu}_i\right)\right]^{-1}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\right]^{-1}$$

This is equivalent to the formula for the large-sample covariance matrix of the ML estimator in a GLM: $(X'WX)^{-1}$.

6.3.5 Propertiy: Under Incorrect Specification: Asymptotic Covariance Matrix

If $var(y_i)\neq v(\mu_i)$, then the asymptotic covariance matrix of he QL estimator $\hat{\boldsymbol{\beta}}$ is not $\mathbf{V}$ as given in Theorem: Asymptotically Normality if $var(y_i)=v(\mu_i)$.

Theorem 12. *When $var(y_i)\neq v(\mu_i)$, i.e. incorrect specification of the mean-variance relation, the asymptotic covariance matrix of QL estimators $\hat{\boldsymbol{\beta}}$ is*

$$\operatorname{var}(\hat{\beta})\approx \mathbf{V}\left[\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \beta}\right)^{\mathrm{T}}\frac{\operatorname{var}\left(y_i\right)}{\left[v\left(\mu_i\right)\right]^2}\left(\frac{\partial \mu_i}{\partial \beta}\right)\right]\mathbf{V}$$

This method is called robust adjustment. The resulting standard errors are called robust standard errors.

Proof. To find the actual $Var(\hat{\boldsymbol{\beta}})$, try Taylor expansion for the quasi-score function at $\boldsymbol{\beta}$

$$u(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{u}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

Since $u(\hat{\boldsymbol{\beta}}) = 0$,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx -\left(\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1} \boldsymbol{u}(\boldsymbol{\beta})$$

so that

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) \approx \left(\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1} \operatorname{var}[\boldsymbol{u}(\boldsymbol{\beta})] \left(\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1}$$

Note $\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is the Hessian matrix for the quasi-log-likelihood. So $-\left[\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1}$ is the analog of an inverse observed information matrix for the specified model, and it approximates $\mathbf{V}$. Also,

$$\operatorname{var}[\boldsymbol{u}(\boldsymbol{\beta})] = \operatorname{var}\left[\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \beta}\right)^{\mathrm{T}} \frac{(y_i - \mu_i)}{v(\mu_i)}\right] = \sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \beta}\right)^{\mathrm{T}} \frac{\operatorname{var}(y_i)}{[v(\mu_i)]^2}\left(\frac{\partial \mu_i}{\partial \beta}\right)$$

In sumary, the actual asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) \approx \mathbf{V}\left[\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \beta}\right)^{\mathrm{T}} \frac{\operatorname{var}(y_i)}{[v(\mu_i)]^2}\left(\frac{\partial \mu_i}{\partial \beta}\right)\right]\mathbf{V}$$

If $\operatorname{var}(y_i) = v(\mu_i)$ then it simplies to $\mathbf{V}$. □

In practice, the $\operatorname{var}(y_i)$ is unknown. We can repalce μ_i by $\hat{\mu}_i$ and $\operatorname{var}(y_i)$ by $(y_i - \hat{\mu}_i)^2$, to get an estimator of the asymptotic covariance matrix. Note that $n\widehat{Var}(\hat{\boldsymbol{\beta}}) \xrightarrow{D} Cov(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))$. It is called called a sandwich estimator, because the empirical evidence is sandwiched between the model-based covariance matrices.

Proof. (Sketch by Professor Jingshu Wang)

Following

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx -\left(\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1} \boldsymbol{u}(\boldsymbol{\beta})$$

Since $\boldsymbol{u}(\boldsymbol{\beta}) = \sum u_i(\boldsymbol{\beta})$, we can write

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx -\left(\frac{1}{n}\sum \frac{\partial u_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1} \frac{1}{\sqrt{n}}\sum u_i(\boldsymbol{\beta})$$

Denote A, V by

$$\frac{1}{n}\sum \frac{\partial u_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \to E\left(\frac{\partial u_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = A$$

$$\frac{1}{\sqrt{n}}\sum u_i(\boldsymbol{\beta}) \approx N(0, V)$$

Thus

$$Var(\hat{\boldsymbol{\beta}}) \approx \frac{1}{n}A^{-1}VA^{-\top}$$

□

In practice, we estimate A and V by

$$\widehat{A} = \frac{1}{n}\sum_{i=1}^{n} \dot{u}_i(\hat{\boldsymbol{\beta}})$$

$$\widehat{V} = \frac{1}{n}\sum_{i} u_i(\hat{\boldsymbol{\beta}})u_i(\hat{\boldsymbol{\beta}})^T$$

7 Mixed Effect Models

8 Bayesian GLM

9 Survival Analysis