Chapter 4: Generalized Linear Models-II

Dipankar Bandyopadhyay

Department of Biostatistics,
Virginia Commonwealth University

BIOS 625: Categorical Data & GLM

[Acknowledgements to Tim Hanson and Haitao Chu]

4.3.3 Overdispersion for Poisson GLMs

- If data are truly Poisson, then we should have roughly $E(Y_i) = \text{var}(Y_i) = \mu_i$. Data can be grouped into like categories and this can be informally checked.
- For the horseshoe crab data we have the following (Table 4.4):

Width (cm)	Sample mean	Sample variance
< 23.25	1.0	2.8
23.25 – 24.25	1.4	8.9
24.25 – 25.25	2.4	6.5
25.25 – 26.25	2.7	11.4
26.25 – 27.25	2.9	6.7
27.25 – 28.25	3.9	8.9
28.25 – 29.25	3.9	16.9
> 29.25	5.1	8.3

- The sample variance tends to be 2-3 times as much as the mean. This is an example of overdispersion. There is greater variability in the data than we expect under our sampling model.
- Fixes:
  - Find another sampling model!
  - Include other important, explanatory covariates.
  - Random effects as a proxy to unknown, latent covariates.
  - Quasi-likelihood approach.
- We'll explore a common approach to the first fix above...

Review: PhD 2011 Exam Question on Linear Models, Q1

We say $Y \sim NB(\alpha, \beta)$, if $Y|\lambda \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(\alpha, \beta)$ with density $f(\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\lambda/\beta}, \alpha, \beta > 0$.
Proof: The probability mass function for the Negative Binomial is

$$\begin{aligned}
h(y) &= \int_0^\infty g(y|\lambda)f(\lambda)d\lambda \\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\frac{\lambda^{\alpha-1}e^{-\lambda/\beta}}{\Gamma(\alpha)\beta^\alpha}d\lambda \\
&= \frac{1}{\Gamma(y+1)\Gamma(\alpha)\beta^\alpha}\int_0^\infty \lambda^{y+\alpha-1}\exp\left[-\lambda\left(\frac{\beta}{\beta+1}\right)^{-1}\right]d\lambda
\end{aligned}$$

The integrand is the kernel of a Gamma density with shape parameter $\alpha + y$ and scale parameter $\beta/(\beta + 1)$.

To get a Gamma density and make the integral equal 1, we divide the integrand by $\Gamma(\alpha+y) \times [\beta/(\beta+1)]^{\alpha+y}$ and of course to keep $h(y)$ from changing, we must multiply the factor outside the integral by that quantity. So

$$
\begin{aligned}
h(y) &= \frac{\Gamma(\alpha+y)[\beta/(\beta+1)]^{\alpha+y}}{\Gamma(y+1)\Gamma(\alpha)\beta^{\alpha}} \\
&= \frac{\Gamma(\alpha+y)}{\Gamma(y+1)\Gamma(\alpha)}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(\frac{\beta}{1+\beta}\right)^{y}.
\end{aligned}
$$

Recall that the Gamma distribution for λ has mean $\alpha\beta$ and variance $\alpha\beta^2$.
$E(y) = E[E(y|\lambda)] = E[\lambda] = \alpha\beta = \mu$ and
$Var(Y) = E[Var(Y|\lambda)] + Var[E(Y|\lambda)] = E[\lambda] + Var[\lambda] = \alpha\beta + \alpha\beta^2 = \mu + \mu\beta = \mu(1+\beta)$

4.3.4 Negative binomial regression

Let $k = \alpha$ and $\mu = \alpha\beta$, then $Y \sim \text{negbin}(k, \mu)$ with

$$p(y) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y \text{ for } y = 0, 1, 2, 3, \ldots.$$

Then $E(Y) = \mu$ and $\text{var}(Y) = \mu + \mu^2/k$.

- A sampling model that includes another parameter allows some separation between the mean and variance.
- The negative binomial distribution is a discrete probability distribution of the number of successes in a sequence of Bernoulli trials before a specified (non-random) number k of failures occurs.
- The index k^{-1} is called a *dispersion* parameter. As $k \to \infty$ the Poisson distribution is obtained.
- Here, the variance *increases* with the mean; is that appropriate for the crab data? Book looks at crab data on p. 127.
- Another modeling approach: adding a random effect for each crab, coming up toward the end of the semester.

4.4 Mean & Variance for GLMs

- A two parameter exponential family includes a *dispersion parameter* ϕ:

$$f(y_i|\theta_i,\phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\}.$$

- When ϕ is known, it simplifies to $f(y_i|\theta_i) = a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)]$, the nature one parameter exponential family.
- This includes binomial, negative binomial, Poisson, normal, and many others.
- Let $L_i = \log f(y_i;\theta_i,\phi)$. This is the contribution of the i^{th} observation to the likelihood in terms of θ_i and ϕ.
- Then $L(\boldsymbol{\theta},\phi) = \sum_{i=1}^N \log f(y_i;\theta_i,\phi) = \sum_{i=1}^N L_i$, where $L_i = [y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)$.

- Then some work gives us
$$\mu_i = E(Y_i) = b'(\theta_i) \text{ and } \mathrm{var}(Y_i) = b''(\theta_i)a(\phi).$$
- The model imposes $\mu_i = b'(\theta_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$. The N-dimensional $\boldsymbol{\mu}$, or equivalently $\boldsymbol{\theta}$, is reduced to the p-dimensional $\boldsymbol{\beta}$ (and ϕ in a 2-parameter family). Then
$$L(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{N} \left[\frac{y_i (b')^{-1}(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})) - b((b')^{-1}(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})))}{a(\phi)} + c(y_i, \phi) \right],$$
as $\theta_i = (b')^{-1}(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))$.
- The MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are found by taking first derivatives of this, setting equal to zero, and solving (pp. 132-136). Things simplify when using the canonical link.

- The asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ is the inverse of the fisher information matrix, $\text{cov}(\hat{\boldsymbol{\beta}})$. This is a function of the unknown $\boldsymbol{\beta}$ and ϕ, and in practice we just plug in the MLE values $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ yielding $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$.
- Section 4.4.3 shows how Poisson and binomial GLMs fit into the general exponential family form and specifies corresponding $b(\theta_i)$, $a(\phi)$, and $c(y_i, \phi)$.
- Section 4.4.9 carries out computations leading to $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$ in the Poisson regression model with a log link.

4.5.1 Deviance and GOF

- For now assume we're able to get $\hat{\beta}$. Anyway, we *are* able to, in SAS or R!
- Recall that the saturated model estimates the N μ_is with the N y_is, providing perfect fit. This model does not reduce data, provide a means for prediction for arbitrary covariate values **x**, allow for meaningful hypotheses to be tested, etc.
- However, we can use the saturated model to check the fit of a "real" GLM.
- If, essentially, the number of distinct covariate vectors remains fixed but N increases then $G^2 = -2\log\mathcal{L}(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}), \hat{\phi}_r; \mathbf{y}) - \log\mathcal{L}(\mathbf{y}, \hat{\phi}_f; \mathbf{y})]$ is the LRT statistic for testing $H_0 : g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$ relative to the alternative that the means $\boldsymbol{\mu}$ are unstructured.

- In Poisson and binomial regression models $a(\phi) = 1$, i.e. there is no dispersion parameter, and this LRT statistic is equal to the model deviance as described last time *for grouped data*.
- When there is a dispersion parameter ϕ (e.g. normal, negative binomial, or gamma regression models), -2 times the difference in saturated and reduced models log-likelihood is D/ϕ in most models, called the scaled deviance; see top, p. 137.
- The scaled deviance has an approximate chi-squared distribution when the reduced model holds.

4.5.4 LR model comparison

- In Poisson and binomial regression, $D_r = -2[L_r - L_s]$ where D is deviance, L_r is log-likelihood evaluated at $\hat{\boldsymbol{\beta}}$ for the GLM, and L_s is log-likelihood evaluated at $\hat{\mu}_i = y_i$ under the saturated model.
- Say we add a few more predictors to the model so the dimension of $\boldsymbol{\beta}$ goes from p to $p + q$. Compute the deviance from the smaller model (D_r) and the larger model (D_f). Then $D_r - D_f$ is the likelihood ratio test statistic for testing H_0 : smaller model holds, and is asymptotically χ^2_q when H_0 is true. The larger this difference, the more evidence there is that the new predictors significantly improve model fit (and hence significantly reduce model deviance).
- Often data are not grouped; in this case it's safer to use $L(\boldsymbol{\beta}; \mathbf{y})$ directly from the output!

4.5.6 Residuals for GLMs

- Residuals indicate where model fit is inadequate.
- The deviance residual d_i is defined in such a way that $\sum_{i=1}^{N} d_i^2 = D$, see p. 141.
- The Pearson residual is given by $e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{var}}(Y_i)}}$. These have variance < 1.
- The standardized Pearson residuals r_i properly standardize the residual to have variance one and in large samples are $N(0,1)$ if the model holds. This means reasonably large n_i for binomial data and reasonably large counts for Poisson data. So residuals $|r_i| > 3$ show rather extreme lack of fit for $(\mathbf{x}_i, Y_i)$ according to the model.
- Residuals can be plotted versus predictors or against the linear predictor $\hat{\eta}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ to assess systematic departures from model assumptions.
- Note: $X^2 = \sum_{i=1}^{N} e_i^2 \overset{\bullet}{\sim} \chi^2_{s-p}$ when $H_0 : g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta}$ is true.

Newton-Raphson Method in One Dimension

Say we want to find where $f(x) = 0$ for differentiable $f(x)$. Let x_0 be such that $f(x_0) = 0$. Taylor's theorem tells us

$$f(x_0) \approx f(x) + f'(x)(x_0 - x).$$

Plugging in $f(x_0) = 0$ and solving for x_0 we get $\hat{x}_0 = x - \frac{f(x)}{f'(x)}$. Starting at an x near x_0, $\hat{x}_0$ should be closer to x_0 than x was. Let's iterate this idea t times:

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}.$$

Eventually, if things go right, $x^{(t)}$ should be close to x_0.

If $\mathbf{f}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^p$, the idea works the same, but in vector/matrix terms. Start with an initial guess $\mathbf{x}^{(0)}$ and iterate

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - [D\mathbf{f}(\mathbf{x}^{(t)})]^{-1}\mathbf{f}(\mathbf{x}^{(t)}).$$

If things are "done right," then this should converge to $\mathbf{x}_0$ such that $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$.
We are interested in solving $DL(\boldsymbol{\beta}) = \mathbf{0}$ (the score, or likelihood equations!) where

$$DL(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p} \end{bmatrix} \text{ and } D^2L(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1^2} & \cdots & \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p^2} \end{bmatrix}.$$

So for us, we start with $\boldsymbol{\beta}^{(0)}$ (maybe through a MOM or least squares estimate) and iterate

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [D^2 L(\boldsymbol{\beta}^{(t)})]^{-1} DL(\boldsymbol{\beta}^{(t)}).$$

This is (4.45) on p. 143 disguised.
The process is typically stopped when $|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}| < \epsilon$.

- Newton-Raphson uses $D^2 L(\boldsymbol{\beta})$ as is, with the $\mathbf{y}$ plugged in.
- Fisher scoring instead uses $E\{D^2 L(\boldsymbol{\beta})\}$, with expectation taken over $\mathbf{Y}$, which is *not* a function of the observed $\mathbf{y}$, but harder to get.
- The latter approach is harder to implement, but conveniently yields $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) \approx [-E\{D^2 L(\boldsymbol{\beta})\}]^{-1}$ evaluated at $\hat{\boldsymbol{\beta}}$ when the process is done.

- The MLE $\boldsymbol{\beta}$ satisfies:

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

  where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ and $v(\mu_i) = \text{var}(Y_i)$, a function of μ_i.
- These are the partial derivatives of the log-likelihood function set to zero, also called the *score* equations.
- In exponential families, a given $\mu_i = E(Y_i)$ and $v(\mu_i) = \text{var}(Y_i)$ uniquely determines the distribution. For example, if we say $E(Y_i) = \mu_i$ and $\text{var}(Y_i) = v(\mu_i) = \mu_i$, and that Y_i is a distribution in the exponential family, then Y_i *has to be* Poisson.

- For Poisson data, we know $v(\mu_i) = \mu_i$; for Binomial data $(E(Y_i) = \mu_i = n_i\pi_i)$, we have $v(\pi_i) = n_i\pi_i(1 - \pi_i)$.
- If we add a dispersion parameter ϕ and declare that $v(\mu_i) = \phi\mu_i$ (Poisson) or $v(\pi_i) = \phi n_i\pi_i(1 - \pi_i)$ (binomial), the resulting family may not be exponential, or not even unique, but the score equations on the previous slide *remain the same*.
- So $\hat{\boldsymbol{\beta}}$ does not change.
- What does change is the estimate $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$. This estimate is the same as from the original model (where $v(\mu_i) = \mu_i$ or $v(\pi_i) = n_i\pi_i(1 - \pi_i)$ for Poisson or Binomial respectively) except multiplied by ϕ. Therefore regression effect standard errors are simply multiplied by $\sqrt{\hat{\phi}}$ where $\hat{\phi}$ is an estimate of ϕ.

- Let $X^2 = \sum_{i=1}^N (y_i - \hat{\mu}_i)^2/\hat{\mu}_i$ for Poisson and $X^2 = \sum_{i=1}^N (y_i - n_i\hat{\pi}_i)^2/[n_i\hat{\pi}_i(1-\hat{\pi}_i)]$ for binomial, the Pearson statistic for assessing (original) model fit.
- ϕ is not in the score equations, however, $X^2/\phi \overset{\bullet}{\sim} \chi^2_{s-p}$ (when the dispersion model is true) where s is the number of unique covariate vectors in $\{\mathbf{x}_i\}$. Since $E(\chi^2_{df}) = df$, a MOM estimate of ϕ is $\hat{\phi} = X^2/(s-p)$. When data are grouped, $s = N$.
- The adjusted estimate is $\widehat{\text{cov}}_a(\hat{\beta}) = \hat{\phi}\ \widehat{\text{cov}}(\hat{\beta})$. When $\hat{\phi} > 1$, which happens with overdispersed data, standard errors get properly inflated.
- This is an easy, *ad hoc* fix to overdispersion, but commonly done and useful.
- SAS does everything automatically when you specify SCALE=PEARSON in the MODEL statement of GENMOD. Also: SCALE=DEVIANCE works similarly.

To recall, SAS code for crab data without dispersion parameter:

```
proc genmod; model satell = width / dist=poisson link=identity;
```

Output:

```
                                 The GENMOD Procedure
                                  Model Information
                         Criteria For Assessing Goodness Of Fit
              Criterion                       DF          Value            Value/DF
              Deviance                       171       557.7083              3.2615
              Scaled Deviance                171       557.7083              3.2615
              Pearson Chi-Square             171       542.4854              3.1724
              Scaled Pearson X2              171       542.4854              3.1724
              Log Likelihood                             73.5314
              Full Log Likelihood                      -456.5030
              AIC (smaller is better)                   917.0060
              AICC (smaller is better)                  917.0766
              BIC (smaller is better)                   923.3126

         Algorithm converged.

                     Analysis Of Maximum Likelihood Parameter Estimates
                               Standard      Wald 95% Confidence          Wald
Parameter      DF    Estimate      Error           Limits             Chi-Square   Pr > ChiSq
Intercept       1    -11.5321   1.510400    -14.4924    -8.57173          58.29       <.0001
width           1    .5494968   .0592926    .4332855    .6657082          85.89       <.0001
Scale           0    1.000000   .0000000    1.000000    1.000000
```

SAS code for crab data with dispersion parameter:

proc genmod; model satell = width / dist=poisson link=identity scale=pearson;

Output:

```
                                 The GENMOD Procedure
                          Criteria For Assessing Goodness Of Fit
               Criterion                         DF         Value          Value/DF
               Deviance                         171      557.7083            3.2615
               Scaled Deviance                  171      175.7985            1.0281
               Pearson Chi-Square               171      542.4854            3.1724
               Scaled Pearson X2                171      171.0000            1.0000
               Log Likelihood                             23.1783
               Full Log Likelihood                      -143.8970
               AIC (smaller is better)                   291.7939
               AICC (smaller is better)                  291.8645
               BIC (smaller is better)                   298.1005
          Algorithm converged.

                     Analysis Of Maximum Likelihood Parameter Estimates
                                 Standard      Wald 95% Confidence       Wald
Parameter     DF      Estimate        Error          Limits           Chi-Square   Pr > ChiSq
Intercept      1      -11.5321     2.690221    -16.8048    -6.25932       18.38       <.0001
width          1      .5494968     .1056079    .3425092    .7564845       27.07       <.0001
Scale          0      1.781132     .0000000    1.781132    1.781132
NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Squared/DOF.
```

Note that $\hat{\beta}$ is the same with or without the dispersion parameter. What changes are se($\hat{\beta}_j$).

- This approach to handling overdispersion works well when the mean structure is well modeled. Otherwise, what does $\hat{\phi}$ really estimate?
- This was a lot of information thrown at you very quickly. Meant to introduce notation and be an overview of things to come.
- We will slow down and investigate specific models in more detail.
- Be careful distinguishing s from N! In the saturated model, s is the number of distinct *categories* that data fall into. However, SAS takes the *df* for deviance to be the number of records N regardless. Data should be grouped in as few groups as possible when checking for dispersion. See 4.5.3, pp. 137-138.

4.7.4 Teratology example

- Female rats given one of four treatments: placebo, weekly iron supplement, days 7 & 10, days 0 & 7. See p. 152 for the data. The number dead y_{ij} out of litter size n_{ij} was recorded where $i = 1, 2, 3, 4$ is the treatment group, and $j = 1, \ldots, m_i$ is the number of litters in group i (31, 12, 5, 10).
- Let π_i denote the probability of death in group i. The model is simply $Y_{ij} \sim \text{bin}(n_{ij}, \pi_i)$.
- The sum of two independent binomials with the same probability is also binomial. So according to the *model*, there really is only four observations:

i	y_{i+}	n_{i+}
1	248	327
2	12	118
3	2	58
4	5	104

- The idea behind this example is that there is litter-to-litter variability and so the data are really a mixture of binomial distributions and overdispersion might be present.
- If we *do not* group the data, then

$$X^2 = \sum_{i=1}^{4} \sum_{j=1}^{m_i} \frac{(y_{ij} - n_{ij}\hat{\pi}_i)^2}{n_{ij}\hat{\pi}_i(1-\hat{\pi}_i)}.$$

  This has an approximate χ^2_{58-4} distribution when we think of litter sizes $n_{ij} \to \infty$. Then $\hat{\phi} = 2.86$ and there's evidence of overdispersion.
- According to the model, the groupings are *arbitrary*; we cannot tell the difference between litters. If we group the data then

$$X^2 = \sum_{i=1}^{4} \frac{(y_{i+} - n_{i+}\hat{\pi}_i)^2}{n_{i+}\hat{\pi}_i(1-\hat{\pi}_i)} = 0.$$

- The problem is that the latter model being fit *is the saturated model*. There is no way to check for overdispersion using the grouped data.
- However, according to the model, the groupings according to data recorded in terms on litters are arbitrary. We know this isn't the case, but *the model* cannot tell the difference between litters, only treatments.
- We are using information on litters to assess overdispersion, but not explicitly including this information in a real probability model, but rather through ϕ. (Better than ignoring the possibility entirely!)
- A possibly better approach is to include a separate term for each litter!

  $$Y_{ij} \sim \text{bin}(n_{ij}, \mu_{ij}), \ \text{logit}(\mu_{ij}) = \pi_i + \gamma_{ij},$$

  where $\gamma_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. This *random effects model* explicitly includes litter-to-litter heterogeneity in the model. The γ_{ij} serve as a proxy to unmeasured, latent genetic differences among litters.

Comments:

- Which approach is better, estimating ϕ and inflating the se's for $\hat{\pi}_i$ or the random effects model?
- What assumptions under the random effects model might be violated? What strengths does it have?
- What assumptions using $v(\pi_i) = \phi n_i \pi_i (1 - \pi_i)$ might be violated? How does this affect the model? Can you see a potentially bigger problem here in using an estimate $\hat{\phi}$?
- How would I analyze these data? With a random effects model, then examine $\hat{\gamma}_{ij}$ to check the normality assumption.