Published on *STAT 504* (https://onlinecourses.science.psu.edu/stat504)

7.2.2 - Overdispersion

Overdispersion is an important concept in the analysis of discrete data. Many a time data admit more variability than expected under the assumed distribution. The greater variability than predicted by the generalized linear model random component reflects overdispersion. Overdispersion occurs because the mean and variance components of a GLM are related and depends on the same parameter that is being predicted through the independent vector.

There is no such thing as overdispersion in ordinary linear regression. In a linear regression model

$$y_i \sim N(x_i^T \beta, \sigma^2)$$

the variance σ^2 is estimated independently of the mean function $x_i^T \beta$.

With discrete response variables, however, the possibility for overdispersion exists because the commonly used distributions specify particular relationships between the variance and the mean; we will see the same holds for Poisson.

For the binomial response, if $y_i \sim Bin(n_i, \pi_i)$, the mean is $\mu_i = n_i \pi_i$ and the variance is $\mu_i(n_i - \mu_i) / n_i$.

- ***Overdispersion*** means that the data show evidence that the variance of the response y_i is greater than $\mu_i(n_i - \mu_i) / n_i$.
- ***Underdispersion*** is also theoretically possible, but rare in practice. McCullagh and Nelder (1989) say that overdispersion is the rule rather than the exception.

In the context of logistic regression, overdispersion occurs when the discrepancies between the observed responses y_i and their predicted values $\hat{\mu}_i = n_i \hat{\pi}_i$ are larger than what the binomial model would predict. Overdispersion arises when the n_i Bernoulli trials that are summarized in a line of the dataset are

- not identically distributed (i.e. the success probabilities vary from one trial to the next), or
- not independent (i.e. the outcome of one trial influences the outcomes of other trials).

In practice, it is impossible to distinguish non-identically distributed trials from non-independence; the two phenomena are intertwined.

Issue: If overdispersion is present in a dataset, the estimated standard errors and test statistics the overall goodness-of-fit will be distorted and adjustments must be made. When a logistic model fitted to n binomial proportions is satisfactory, the residual deviance has an approximate χ^2 distribution with $(n - p)$ degrees of freedom, where p is the number of unknown parameters in the fitted model. Since the expected value of a χ^2 distribution is equal to its degree of freedom, it follows that the residual deviance for a well-fitting model should be approximately equal to its degrees of freedom. Equivalently, we may say that the mean deviance (deviance/df) should be

close to one. Similarly, if the variance of the data is greater than that under under binomial sampling, the residual mean deviance is likely to be greater than 1.

The problem of overdispersion may also be confounded with the problem of omitted covariates. If we have included all the available covariates related to y_i in our model and it still does not fit, it could be because our regression function $x_i^T\beta$ is incomplete. Or it could be due to overdispersion. Unless we collect more data, we cannot do anything about omitted covariates. *But we can adjust for overdispersion.*

Adjusting for Overdispersion

The most popular method for adjusting for overdispersion comes from the theory of quasilikelihood. Quasilikelihood has come to play a very important role in modern statistics. It is the foundation of many methods that are thought to be "robust" (e.g. Generalized Estimating Equations (GEE) for longitudinal data) because they do not require specification of a full parametric model. For more details see Agresti (2007, Sec 9.2) or Agresti (2013, Sec 12.2).

In the quasilikelihood approach, we must first specify the "mean function" which determines how $\mu_i = E(y_i)$ is related to the covariates. In the context of logistic regression, the mean function is

$$\mu_i = n_i \text{expit}(x_i^T\beta),$$

which implies

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta$$

Then we must specify the "variance function," which determines the relationship between the variance of the response variable and its mean. For a binomial model, the variance function is $\mu_i(n_i - \mu_i) / n_i$.

But to account for overdispersion, we will include another factor σ^2 called the "scale parameter," so that

$$V(y_i) = \sigma^2\mu_i(n_i - \mu_i)/n_i.$$

- If $\sigma^2 \neq 1$ then the model is not binomial; $\sigma^2 > 1$ is called "overdispersion" and $\sigma^2 < 1$ is called "underdispersion."
- If σ^2 were known, we could obtain a consistent, asymptotically normal and efficient estimate for β by a quasi-scoring procedure, sometimes called "estimating equations." For the variance function shown above, the quasi-scoring procedure reduces to the Fisher scoring algorithm that we mentioned as a way to iteratively find ML estimates.

Note that no matter what σ^2 is assumed to be, we get the same estimate for β. Therefore, this method for overdispersion does not change the estimate for β at all. However, the estimated covariance for $\hat{\beta}$ changes from

$$\hat{V}(\hat{\beta}) = (X^TWX)^{-1}$$

to

$$\hat{V}(\hat{\beta}) = \sigma^2 (X^T W X)^{-1}$$

That is, the estimated standard errors must be multiplied by the factor $\sigma = \sqrt{\sigma^2}$.

How do we estimate σ^2?

McCullagh and Nelder (1989) recommend

$$\hat{\sigma}^2 = X^2/(N-P)$$

where X^2 is the usual Pearson goodness-of-fit statistic, N is the number of sample cases (number of rows in the dataset we are modeling), and p is the number of parameters.

- If the model holds, then $X^2/(N - p)$ is a consistent estimate for σ^2 in the asymptotic sequence $N \to \infty$ for fixed n_i's.
- The deviance-based estimate $G^2/(N - p)$ does not have this consistency property and should not be used.

This is a reasonable way to estimate σ^2 if the mean model $\mu_i = g(x_i^T \beta)$ holds. But if important covariates are omitted, then X^2 tends to grow and the estimate for σ^2 can be too large. For this reason, we will estimate σ^2 under a **maximal model**, a model that includes all of the covariates we wish to consider.

The best way to estimate σ^2 is to identify a rich model for μ_i and designate it to be the most complicated one that we are willing to consider. For example, if we have a large pool of potential covariates, we may take the maximal model to be the model that has every covariate included as a main effect. Or, if we have a smaller number of potential covariates, we decide to include all main effects along with two-way and perhaps even three-way interactions. But we must omit at least a few higher-order interactions, otherwise we will end up with a model that is saturated.

In an overdispersed model, we must also adjust our test statistics. The statistics X^2 and G^2 are adjusted by dividing them by σ^2. That is, tests of nested models are carried out by comparing differences in the scaled Pearson statistic, $\Delta X^2/\sigma^2$, or the scaled deviance, $\Delta G^2/\sigma^2$ to a chisquare distribution with Δdf degrees of freedom.

If the data are overdispersed — that is, if

$$V(y_i) \approx \sigma^2 n_i \pi_i (1 - \pi_i)$$

for a scale factor $\sigma^2 > 1$, then the residual plot may still resemble a horizontal band, but many of the residuals will tend to fall outside the ± 3 limits. In this case, the denominator of the Pearson residual will tend to understate the true variance of the y_i, making the residuals larger. If the plot looks like a horizontal band but X^2 and G^2 indicate lack of fit, an adjustment for overdispersion might be warranted. A warning about this, however: If the residuals tend to be too large, it doesn't necessarily mean that overdispersion is the cause. Large residuals may also be caused by *omitted covariates*. If some important covariates are omitted from x_i, then the true π_i's will depart from what your model predicts, causing the *numerator* of the Pearson residual to be larger than usual.

That is, apparent overdispersion could also be an indication that your mean model needs additional covariates. If these additional covariates are not available in the dataset, however, then there's not much we can do about it; we may need to attribute it to overdispersion.

Note, there is **no overdispersion for ungrouped data.** McCullagh and Nelder (1989) point out that overdispersion is not possible if $n_i=1$. If y_i only takes values *0* and *1*, the it must be distributed as Bernoulli(π_i) and its variance must be $\pi_i(1 - \pi_i)$. There is no other distribution with support {0,1}. Therefore, with ungrouped data, we should always assume ***scale=1*** and not try to estimate a scale parameter and adjust for overdispersion.

Summary of Adjusting for Overdispersion in the Binary Logistic Regression

The usual way to correct for overdispersion in a logit model is to assume that:

$$E(y_i) = n_i\pi_i$$
$$V(y_i) = \sigma^2 n_i\pi_i(1-\pi_i)$$

where σ^2 is a scale parameter.

Under this modification, the Fisher-scoring procedure for estimating β does not change, but its estimated covariance matrix becomes $\sigma^2(X^TWX)^{-1}$—that is, the usual standard errors are multiplied by the square root of σ^2.

This will make the confidence intervals wider.

Let's get back to our example and refit the model, making an adjustment for overdispersion.

- Using SAS
- Using R

If you are using `glm()` in R, and want to refit the model adjusting for overdispersion one way of doing it is to use `summary.glm()` function. For example fit the model using `glm()` and save the object as RESULT. By default dispersion is equal to 1. This will perform the adjustment. It will not change the estimated coefficients β_j, but it will adjust the standard errors.

Estimate from the MAXIMAL model dispersion value as X^2/df. Then you can call

```
summary(RESULT, dispersion=4.08,correlation=TRUE,symbolic.cor =
TRUE).
```

This should give you the same model but with adjusted covariance matrix, that is adjusted standard errors (SEs) for your beta's (estimated logistic regression coefficients) and also changed z-values. Notice it will not adjust overall fit statistics. For that try the package "dispmod" (see assay.R).

R output after adjusting for overdispersion:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -15.834      4.923  -3.216 0.001298 **
logconc        5.578      1.680   3.319 0.000902 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 4.08)

    Null deviance: 83.631  on 8  degrees of freedom
Residual deviance: 29.346  on 7  degrees of freedom
AIC: 62.886
```

There are other corrections that we could make. If we were constructing an analysis-of-deviance table, we would want to divide G^2 and X^2 by $\hat{\sigma}^2$ and use these scaled versions for comparing nested models. Moreover, in reporting residuals, it would be appropriate to modify the Pearson residuals to

$$r_i^* = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{\hat{\sigma}^2 n_i\hat{\pi}_i(1-\hat{\pi}_i)}};$$

that is, we should divide the Pearson residuals (or the deviance residuals, for that matter) by $\sqrt{\hat{\sigma}^2}$.

Source URL: https://onlinecourses.science.psu.edu/stat504/node/162

Links:
[1] http://www.dynamicdrive.com