Some Issues in Generalized Linear Modeling

Alan Agresti

Abstract This chapter discusses cautions, questions, challenges, and proposals regarding five issues that arise in generalized linear modeling. With primary emphasis on categorical data, we summarize (1) bias that can occur in using ordinary linear models with ordinal response variables, (2) a new proposal about simple ways to interpret effects in generalized linear models that use nonlinear link functions, (3) problems with using Wald significance tests and confidence intervals, (4) a question about the behavior of residuals for generalized linear models, and (5) a new approach in using generalized estimating equations (GEE) methods for marginal multinomial models.

Key words: GEE methods; multinomial models; ordinal models; residuals; Wald inference

1 Introduction

This chapter discusses several issues about generalized linear models that I believe deserve more attention, either in terms of additional research or greater awareness of already existing literature. I discuss these issues, with primary emphasis on categorical data, in the style of cautions, questions, challenges, and proposals. I became aware of these issues in recent years while writing a book on linear and generalized linear models (Agresti 2015) and while revising two books on categorical data analysis (Agresti 2010, 2013).

Section 2 explains the floor and ceiling bias that can occur with modeling ordinal response variables by assigning scores to the outcome categories and using ordinary linear models. Section 3 proposes a simple way to interpret effects in generalized linear models that use nonlinear link functions, by comparing groups using a probability summary about the higher response. Section 4 summarizes problems with

Alan Agresti
University of Florida, Gainesville, Florida 32605, USA, e-mail: aa@stat.ufl.edu

using Wald significance tests and confidence intervals in modeling binary response variables and suggests related research for other types of response variables. Section 5 raises questions about the behavior of ordinary residuals for generalized linear models, and argues that a standardized residual is more relevant than the popular Pearson residual. Section 6 summarizes some awkward aspects of standard generalized estimating equations (GEE) methods for marginal multinomial models and presents a recently proposed approach that is now available with R software.

Most of this chapter has the style of a tutorial or survey paper. But it is hoped that the material is relevant for a conference that has general consideration of topics related to linearity and modeling.

2 Bias in Ordinary Linear Modeling of Ordinal Responses

Ordinal categorical response variables are common in many disciplines, especially the social sciences with sample survey data. An example is one's report of political ideology, selected from the categories (very liberal, slightly liberal, moderate, slightly conservative, very conservative). Many ordinal variables have a rather subjective outcome choice, such as in medical assessments of patient quality of life (excellent, good, fair, poor) or amount of pain (none, little, considerable, severe).

With ordinal response variables, many methodologists assign monotone scores to the ordered outcome categories and then apply ordinary regression methods. That is, they use least squares to estimate parameters in a linear model for the mean response for the chosen scores. This is sometimes problematic because of requiring the choice of scores, which can be quite unclear when the categories are highly subjective. Are categories such as (excellent, good, fair, poor) equally distant, and if not, how does one decide on relative distances? But here we discuss a less known but perhaps more worrisome problem, that *floor effects* or *ceiling effects* due to the boundedness of the discrete ordinal scale can result in seriously biased estimates of effect magnitudes.

We illustrate this potential problem using an example with an assumed connection between an observed ordinal variable and an underlying continuous latent variable. For an ordinal response variable y, it is often realistic to assume the existence of an underlying continuous latent variable y^* that we would ideally observe if we could measure the response in a more refined manner. For instance, for variables such as political ideology or quality of life, there is nothing sacred about a particular choice of categories, and it's easy to imagine increasing the number of categories until the variable becomes essentially continuous. Our example uses the simple normal linear latent variable model for observation i,

$$y_i^* = 20.0 + 0.6x_i - 40z_i + \varepsilon_i$$

in which we take $x_i \sim$ uniform(0, 100), $P(z_i = 0) = P(z_i = 1) = 0.50$ independent of x_i, and $\varepsilon_i \sim N(0, 10^2)$. We randomly generate $n = 100$ observations from this model

and focus on the issue of comparing the two groups represented by the values of z in terms of the effect of the covariate x.

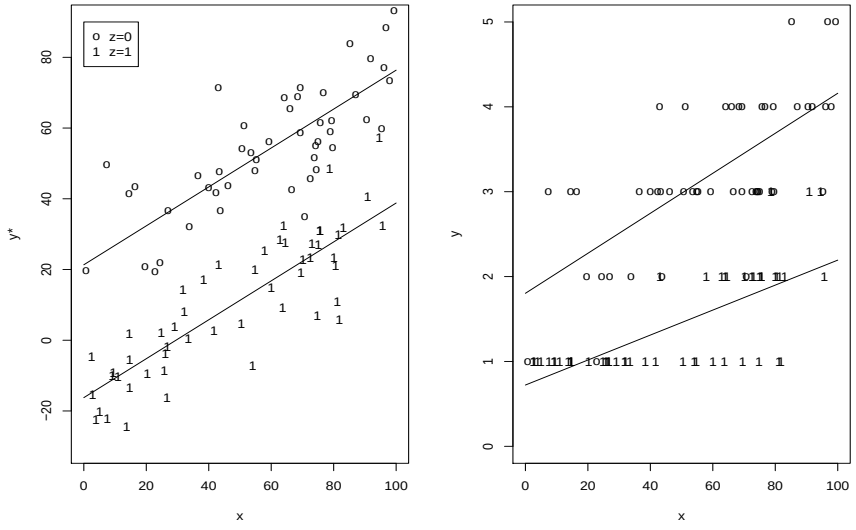
Now, suppose that the variable we actually observe for subject i is directly related to this latent variable by

$$y_i = 1 \text{ if } y_i^* \leq 20, \;\; y_i = 2 \text{ if } 20 < y_i^* \leq 40, \;\; y_i = 3 \text{ if } 40 < y_i^* \leq 60,$$

$$y_i = 4 \text{ if } 60 < y_i^* \leq 80, \quad y_i = 5 \text{ if } y_i^* > 80.$$

That is, cutpoints chop up the continuous scale for $y*$, yielding five ordered categories with corresponding values for the observed y. Figure 1 shows the connection between the observed variable y and the latent variable y^*. The first scatterplot in the figure shows the 100 observations on y^* and x, each data point labelled by the category for z. The plot also shows the regression lines that generated the data.

Fig. 1 Ordered categorical data (in second panel) for which ordinary regression suggests interaction, because of a floor effect, but ordinal modeling does not. The data were generated (in first panel) from a normal main-effects regression model with continuous (x) and binary (z) explanatory variables. When the continuous response y^* is categorized and y is measured as (1, 2, 3, 4, 5), the observations labelled '1' for the category of z have a linear x effect with only half the slope of the observations labelled '0' for the category of z.



Now, for the observed data, suppose we fit the linear model

$$y_i = \alpha + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i \cdot z_i) + \varepsilon_i$$

using the scores (1,2,3,4,5), to study the effect of x for the two groups and to analyze whether interaction occurs between x and z in their effects on y. The right panel of the figure shows the result, plotting the least squares fit. For the observed response, the slope of the line is about twice as high when $z = 0$ as when $z = 1$. Why? When $x_i < 50$ with $z_i = 1$, $P(y_i^* \leq 20) = P(y_i = 1)$ is relatively high. As x gets lower, the underlying value y^* can continue to tend to get lower, but the observed ordinal response cannot fall below 1, resulting in a floor effect. This interaction effect is caused by the observations when $z = 1$ tending to fall in category $y = 1$ whenever x takes a relatively low value.

For the observed data, the interaction is statistically and practically significant. Analyzing the data with an ordinary linear model, we would conclude that an effect exists that actually does not. Such spurious effects would not occur if we instead fitted a proper ordinal model, such as the *cumulative logit model*

$$\text{logit}[P(y_i \leq j)] = \alpha_j + \beta_1 x_i + \beta_2 z_i$$

or the *cumulative probit model*

$$\Phi^{-1}[P(y_i \leq j)] = \alpha_j + \beta_1 x_i + \beta_2 z_i$$

with Φ being the standard normal cdf, with $j = 1,2,3,4$ for the four cumulative probabilities. In fact, we'll note in the next section that such models are implied for this latent variable model. The models account for the ordinality by using cumulative probabilities for y without needing to assign scores and assume linearity on that scale.

We are not suggesting that it is always inappropriate to use ordinary linear models with ordinal response variables. With several outcome categories and observations spread among them without high concentrations in boundary categories, such a model can be adequate. Using such a model can be helpful for relatively unsophisticated methodologists who may be comfortable with linear modeling but not with models that imply more complex effect summaries, such as odds ratios. However, in using this strategy, one should be aware of the potential bias that can result.

3 Interpreting Effects in GLMs with Nonlinear Link Function

For a $n \times 1$ vector $\boldsymbol{y}$ of response observations with $\boldsymbol{\mu} = E(\boldsymbol{y})$, consider a generalized linear model

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$$

for link function g and model matrix $\boldsymbol{X}$ with a set of explanatory variables. For many standard link functions, the interpretation of $\boldsymbol{\beta}$ is difficult for non-statisticians and for methodologists who are mainly familiar with ordinary linear models.

For instance, suppose y is ordinal as we considered in the previous section. Let c denote the number of outcome categories for y. For observation i, let x_{ik} denote the

value of explanatory variable k. Consider the *cumulative link model*

$$\text{link}[P(y_i \le j)] = \alpha_j + \sum_k \beta_k x_{ik}, \quad j = 1, \ldots, c-1,$$

for links such as the logit, probit, or complementary log-log,. For the probit link (i.e., the inverse of the standard normal cdf), β_k represents the change in $\Phi^{-1}[P(y_i \le j)]$ for a 1-unit increase in x_k, adjusting for the other explanatory variables. This is a rather obscure interpretation, as very few people can make sense of effects on the scale of an inverse of a cdf.

One way used to interpret such effects relies more on using means for underlying latent variable models. For the observed ordinal response y and for an underlying continuous response y^*, suppose we assume that $y_i^* = \boldsymbol{\beta}^T \boldsymbol{x}_i + \varepsilon_i$, where ε_i has some parametric cdf G with mean 0. Suppose that there are thresholds (cutpoints) $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_c = \infty$ such that

$$y_i = j \text{ if } \alpha_{j-1} < y_i^* \le \alpha_j.$$

Then, at a fixed value $\boldsymbol{x}$,

$$P(y_i \le j) = P(y_i^* \le \alpha_j) = P(y_i^* - \boldsymbol{\beta}^T \boldsymbol{x}_i \le \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)$$

$$= P(\varepsilon_i \le \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) = G(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i).$$

This implies the model

$$G^{-1}[P(y_i \le j \mid \boldsymbol{x}_i)] = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i$$

with G^{-1} as the link function. In particular, one obtains the cumulative probit model when G is the standard normal cdf Φ; then Φ^{-1} is the probit link. (In practice, whether we use $+$ or $-$ for the coefficient of the linear predictor $\boldsymbol{\beta}^T \boldsymbol{x}_i$ merely affects the sign of the estimates, and varies among the common software packages.) Thus, the cumulative probit model fits well when an ordinary normal linear model holds for an underlying continuous response variable. For this model, β_k has the interpretation that a 1-unit increase in x_k corresponds to a change in $E(y^*)$ of β_k standard deviations, adjusting for the other explanatory variables (Anderson and Philips 1981, McKelvey and Zavoina 1975). But this interpretation can still be rather obscure for non-methodologists who do not think of effects in terms of multiples of standard deviations. Moreover, the latent variable model may not be appropriate in some applications.

We suggest next a simpler interpretation, proposed by Agresti and Kateri (2016). We formulate it in terms of a summary for comparing two groups, adjusting for the other explanatory variables. Let z be an indicator variable for the two groups. At any potential setting $(x_1, \ldots, x_p)$ of p explanatory variables, let y_1^* and y_2^* denote independent latent variables when $z = 1$ and when $z = 0$, respectively. For the latent variable model that generates the cumulative probit model

$$\Phi^{-1}[P(y \le j)] = \alpha_j - \beta z - \beta_1 x_1 - \cdots - \beta_p x_p,$$

the difference between the conditional means of y_1^* and y_2^* is β, and

$$P(y_1^* > y_2^*) = P[(y_1^* - y_2^*) > 0]$$

$$= P\left[\frac{(y_1^* - y_2^*) - \beta}{\sqrt{2}} > \frac{-\beta}{\sqrt{2}}\right] = 1 - \Phi(-\beta/\sqrt{2}) = \Phi(\beta/\sqrt{2}).$$

That is, $P(y_1^* > y_2^*) = \Phi(\beta/\sqrt{2})$ at any setting of the p explanatory variables. Differences between the normal conditional means for the two groups of $\beta = (0, 0.5, 1, 2, 3)$ standard deviations correspond to $P(y_1^* > y_2^*)$ values of (0.50, 0.64, 0.76, 0.92, 0.98).

In practice, the probit link is used much less than the logit link, especially in biostatistics. The cumulative logit model

$$\text{logit}[P(y \le j)] = \alpha_j - \beta z - \beta_1 x_1 - \cdots - \beta_p x_p$$

is implied when an underlying latent variable has a logistic distribution. The derivation just shown for a normal latent variable does not have an exact analog for a logistic latent variable, as the difference between two independent standard logistic random variables does not have a logistic distribution. However, the distribution can be very closely approximated by the logistic with mean 0 and double the variance. This generates the approximation

$$P(y_1^* > y_2^*) \approx \frac{\exp(\beta/\sqrt{2})}{[1 + \exp(\beta/\sqrt{2})]}$$

in terms of the β parameter from the cumulative logit model. This approximation is adequate for practical application. Because of the very close similarity of logit and probit models, another good approximation is to fit also the cumulative probit model and use the exact expression $P(y_1^* > y_2^*) = \Phi(\beta/\sqrt{2})$ for the β parameter from that model.

In practice, it is often sensible to assume a latent variable distribution that, unlike the normal and the logistic, is skewed and has a long tail. For the ordinal model with log-log link, the underlying latent variable has the extreme-value (Gumbel) distribution. The difference between two independent random variables of this type has the standard logistic distribution. So, in this case,

$$P(y_1^* > y_2^*) = \frac{\exp(\beta)}{[1 + \exp(\beta)]}$$

in terms of the β parameter for the cumulative link model with log-log link.

For any of these cumulative link models, ordinary confidence intervals for the β coefficient of the indicator variable induce confidence intervals for $P(y_1^* > y_2^*)$. The measures and the related inferences are presented by Agresti and Kateri (2016). They also proposed analogous measures for the observed ordinal scale that do not

require a latent variable connection, and they have available R functions for confidence intervals for the measures.

Such probability-based measures may be especially helpful for practitioners who cannot easily interpret odds ratios and other measures that result from nonlinear link functions. For instance, for a medical researcher, reading that at fixed values for the explanatory variables, the estimated probability the response to drug ($z = 1$) is better than response to placebo ($z = 0$) is 0.72 probably has greater meaning than reading that (1) the estimated cumulative odds for drug is $\exp(\hat{\beta}) = 2.7$ times the estimated cumulative odds for placebo (i.e., the interpretation for the cumulative logit model), or (2) that the estimated cumulative probits differ by $\hat{\beta} = 0.8$ or an underlying mean for drug is $\hat{\beta} = 0.8$ standard deviations better than for placebo (i.e., the interpretation for the cumulative probit model), or (3) that the estimated probability that the response for drug is worse than a particular outcome category is the power $\exp(\hat{\beta}) = 1.7$ of the estimated probability that response for placebo is worse than that category (i.e., interpretation of cumulative link model with complementary log-log link).

This type of probability measure for comparing groups is also relevant for ordinary normal linear models. With constant error variance σ^2 and potential response outcomes (y_1, y_2) for two groups at some setting of explanatory variables, the corresponding measure is

$$P(y_1 > y_2) = \Phi\left(\frac{\beta}{\sqrt{2}\sigma}\right),$$

for coefficient β of the indicator variable for the two groups. This would seem to be a useful summary in many applications. For two groups with no explanatory variables, a related popular effect size measure is $\beta/\sigma = (\mu_1 - \mu_2)/\sigma$ (Hedges and Olkin 1985). One can derive a confidence interval for β/σ in linear models with explanatory variables using the noncentral t distribution, and such an interval induces one for $P(y_1 > y_2)$. For details and an example using R, see Agresti and Kateri (2016).

4 Wald Inference when Effects for Binary Data are Large

To introduce this topic, we start with a toy example that illustrates an awkward aspect that often occurs in using logistic regression, namely that at least one of the maximum likelihood (ML) estimates of model parameters is infinite. This happens when *complete separation* or *quasi-complete separation* occurs in the space of the explanatory variables (Albert and Anderson 1984).

For six observations, suppose that $y = 1$ at $x = 1, 2, 3$, and $y = 0$ at $x = 4, 5, 6$, a simple example of complete separation. When we use R to fit the ordinary logistic regression model, $\text{logit}[P(y = 1)] = \alpha + \beta x$, we obtain:

```
------------------------------------------------------------
> x <- c(1,2,3,4,5,6);  y <- c(1,1,1,0,0,0)
> fit <- glm(y ~ x, family = binomial(link = logit))
```

```
> summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   165.3   407521.4       0        1
x             -47.2   115264.4       0        1

Number of Fisher Scoring iterations: 25
------------------------------------------------------------
```

After 25 iterations, Fisher scoring converges, as the log-likelihood function is essentially flat at that stage. The fit looks nearly identical to a step function that takes value 1 below $x = 3.5$ and takes value 0 above $x = 3.5$. The maximized log-likelihood value is essentially 0, reflecting the basically perfect fit. Although in fact $\hat{\beta} = -\infty$, R reports $\hat{\beta} = -47.2$. R also reports a huge standard error, reflecting that the unrestricted ML estimate of the standard error (SE) is based on the Fisher information, which summarizes the curvature of the log-likelihood function at $\hat{\beta}$. A perhaps surprising consequence is that $z = \hat{\beta}/SE = 0$, yielding a P-value of 1.0 when we use this ratio as a test statistic for testing H_0: $\beta = 0$. By contrast, the model fit gives evidence of a potentially very strong effect. By contrast, the likelihood-ratio test statistic equals 8.32 with $df = 1$ and yields P-value = 0.004.

The statistic $z = \hat{\beta}/(SE)$ is an example of a *Wald test*. This approach uses the fact that a ML estimator has an asymptotic normal distribution by testing H_0: $\beta = 0$ with $z = \hat{\beta}/(SE)$, or else treating z^2 as an approximate chi-squared random variable with, $df = 1$. The corresponding confidence interval has the form $\hat{\beta} \pm z(SE)$ for the appropriate standard normal percentile z, for instance with $z = 1.96$ for 95% confidence. A classic result shown by Hauck and Donner (1977) is that as $|\beta|$ in a logistic regression model increases (for fixed n), the Fisher information decreases so quickly that SE grows faster than β. The result is poor performance of Wald methods when effects are large.

The poor performance of Wald methods shows up even in very simple contexts, such as a single binomial response variable without any explanatory variables. For a binomial random variable y based on n independent trials with parameter π, in the context of logistic regression the model is $\text{logit}(\pi) = \beta$. To test H_0: $\beta = 0$ (i.e., $\pi = 0.50$), $\hat{\beta} = \text{logit}(\hat{\pi})$ with $\hat{\pi} = y/n$ has asymptotic variance $[n\pi(1-\pi)]^{-1}$. The Wald chi-squared statistic is

$$(\hat{\beta}/SE)^2 = [\text{logit}(\hat{\pi})]^2[n\hat{\pi}(1-\hat{\pi})].$$

Now, suppose $n = 25$. For testing H_0: $\pi = 0.50$, $\hat{\pi} = \frac{24}{25}$ is stronger evidence against H_0 than $\hat{\pi} = \frac{23}{25}$. Yet the Wald statistic equals 9.7 when $\hat{\pi} = 24/25$ and equals 11.0 when $\hat{\pi} = 23/25$. By comparison, the likelihood-ratio statistic takes values 26.3 and 20.7.

With large or infinite effects, likelihood-ratio (LR) tests and test-based confidence intervals remain valid and behave well because of the concavity of the log-likelihood function. For example, when $\hat{\beta} = -\infty$, a confidence interval consists of a range of plausible values from $-\infty$ to some finite upper bound. With infinite ML

estimates, one can alternatively smooth the data and produce finite estimates and finite endpoints of intervals using a Bayesian approach. Or, one can use a penalized likelihood approach with the aim of reducing bias (Firth, 1993), which corresponds to using a Bayesian posterior mode with Jeffreys prior to generate a point estimate.

The poor performance of the Wald test implies poor performance also of corresponding confidence intervals. This has been shown for a variety of measures for categorical data, such as proportions, differences of proportions, odds ratio, and relative risk, particularly when probabilities are near 0 or 1. For a summary of these and various other cases involving categorical data, see Agresti (2011). For example, again for a single binomial parameter π, the 95% Wald confidence interval for π is

$$\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}.$$

In terms of achieving close to the nominal coverage probability, this interval performs much worse than the interval based on inverting a likelihood-ratio test or inverting the score test of H_0: $\pi = \pi_0$, which has test statistic

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

It also behaves much more poorly than a simple approximation to the score confidence interval that (in the 95% case) adds 2 "successes" and 2 "failures" before forming the Wald confidence interval (Agresti and Coull 1998).

An important question that could be addressed in future research is whether the poor Wald performance for binary data holds also for various other generalized linear models for other types of data, with nonlinear link functions. For some theoretical work in this direction, see Brown, Cai, and DasGupta (2003).

5 Behavior of Residuals for GLM Fits

For a $n \times 1$ vector $\boldsymbol{y}$ of response observations with $\boldsymbol{\mu} = E(\boldsymbol{y})$, $\boldsymbol{V} = \text{var}(\boldsymbol{y})$, consider an arbitrary generalized linear model

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$$

with link function g and model matrix $\boldsymbol{X}$. Denote the maximum likelihood fitted values by $\hat{\boldsymbol{\mu}}$.

The ordinary linear model uses identity link $\boldsymbol{\mu} = X\boldsymbol{\beta}$, and assumes $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. For that model, standard results exploit the orthogonal decomposition

$$\boldsymbol{y} = \hat{\boldsymbol{\mu}} + (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \quad \text{(i.e., data = fit + residual)}.$$

With generalized linear models, $\hat{\boldsymbol{\mu}}$ and $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ are not orthogonal when we depart from identity link and constant variance. Then, Pythagoras's Theorem does

not apply, because maximizing the likelihood does not correspond to minimizing $\|\boldsymbol{y}-\hat{\boldsymbol{\mu}}\|$. With a nonlinear link function, although the space of linear predictor values $\boldsymbol{\eta}$ that satisfy a particular model is a linear vector space, the corresponding set of $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$ values is not.

Despite the lack of orthogonality, conventional wisdom seems to be that as n increases, $(\boldsymbol{y}-\hat{\boldsymbol{\mu}})$ is asymptotically uncorrelated with $\hat{\boldsymbol{\mu}}$. If this truly holds, then one can obtain an asymptotic covariance matrix for the residuals, because then

$$\boldsymbol{V} = \text{var}(\boldsymbol{y}) \approx \text{var}(\hat{\boldsymbol{\mu}}) + \text{var}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}).$$

It then follows from standard results using the delta method (e.g., see Agresti 2015, p. 136) that

$$\text{var}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}) \approx \boldsymbol{V}^{1/2}[\boldsymbol{I}-\boldsymbol{H}]\boldsymbol{V}^{1/2},$$

where $\boldsymbol{H}$ is a generalized hat matrix

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$$

incorporating a diagonal weight matrix

$$\boldsymbol{W} = \text{diag}\{(\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i)\}.$$

But why, and under what conditions, is $(\boldsymbol{y}-\hat{\boldsymbol{\mu}})$ asymptotically uncorrelated with $\hat{\boldsymbol{\mu}}$? And for small-to-moderate n, is $\text{corr}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}})$ close enough to 0 that we can safely ignore it? When I was recently writing a book on generalized linear models (Agresti 2015), I was surprised not to find literature about this. It seems that we should consider two types of asymptotics: Traditional asymptotics with $n \to \infty$, and the alternative with n fixed and asymptotics applying to individual components, such as binomial indices and Poisson expected counts in a contingency table. For the alternative (called *small-dispersion asymptotics* by Jørgensen 1987), with individual y_i asymptotically normal, $(\boldsymbol{y}-\boldsymbol{\mu})$ and $(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})$ jointly have an asymptotic normal distribution, as does their difference.

When I asked several statisticians if they knew of the existence of a general result about residuals and fitted values being asymptotically uncorrelated, G. Lovison gave me a heuristic solution. In Lovison (2014), he discussed this point in an article that dealt with analogs of linear model results for generalized linear models. He argued that if $(\boldsymbol{y}-\hat{\boldsymbol{\mu}})$ and $\hat{\boldsymbol{\mu}}$ were not asymptotically uncorrelated, one could construct an asymptotically unbiased and more efficient estimator of $\boldsymbol{\mu}$ using $\hat{\boldsymbol{\mu}}^* = [\hat{\boldsymbol{\mu}} + \boldsymbol{L}(\boldsymbol{y}-\hat{\boldsymbol{\mu}})]$ for a matrix $\boldsymbol{L}$. But this would then contradict the ML estimator $\hat{\boldsymbol{\mu}}$ being asymptotically efficient. This argument is sort of an asymptotic version for ML estimators of one in the Gauss–Markov Theorem that unbiased estimators other than least squares estimator have difference from that estimator that is uncorrelated with it. The Lovison argument is heuristic, not distinguishing between the two possible types of asymptotics, and there still seems to be scope for a formal proof of the general result.

Interestingly, in his article, Lovison shows that a weighted version of adjusted responses that has approximately constant variance has orthogonality of fitted values and residuals. On the original scale, such a residual is the "Pearson residual" $e_i = (y_i - \hat{\mu}_i)/\sqrt{v(\hat{\mu}_i)}$ for variance function v evaluated at the model fit. For contingency tables, the Pearson residual is popular, because it results from the decomposition of the Pearson chi-squared statistic. For example, with Poisson counts $\{y_i\}$, the Pearson statistic satisfies

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = \sum_i e_i^2 \quad \text{with} \quad e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

As an editorial comment, however, I believe it is strongly preferable to use *standardized residuals* rather than Pearson residuals. The standardized residual is

$$r_i = \frac{y_i - \hat{\mu}_i}{\text{std. error}(y_i - \hat{\mu}_i)} = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)(1 - \hat{h}_{ii})}} = \frac{e_i}{\sqrt{1 - \hat{h}_{ii}}}$$

for "leverage" $\hat{h}_{ii}$ from the estimated hat matrix $\widehat{\boldsymbol{H}}$. For small-dispersion asymptotics, r_i is asymptotically standard normal when the model holds. This is not true of the Pearson residual e_i, because the denominator ignores the fact that $\hat{\mu}_i$ is random. The standardized residual appropriately recognizes redundancies in data. For example, for the independence model assuming Poisson or multinomial sampling for a 2×2 table of counts $\{y_{ij}\}$, the fitted values are

$$\{\hat{\mu}_{ij} = np_{i+}p_{+j}\} \text{ for } p_{i+} = (\sum_j y_{ij})/n,\ p_{+j} = (\sum_i y_{ij})/n,$$

and so these two forms of residual then have expressions

$$e_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}, \quad r_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}.$$

For 2×2 tables, $df = 1$, reflecting that all four $|y_{ij} - \hat{\mu}_{ij}|$ are identical, so it seems sensible to have a single value for lack of fit. Yet, all four Pearson residuals can take different values. By contrast, $r_{11} = -r_{12} = -r_{21} = r_{22}$ and each $r_{ij}^2 = X^2$.

6 Improved Marginal Modeling of Multinomial Data

The final topic we consider deals with analyzing correlated observations using marginal models. Suppose that each subject has a cluster of correlated observations $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})^T$, such as in a longitudinal study or an experiment with repeated measures. (The dimension T could vary by cluster, but for simplicity our notation uses a common value.) For each y_{it} marginally, we assume a model $g(\mu_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta}$.

For discrete data, ML for such a model is awkward because of the lack of a simple multivariate distribution that is characterized by pairwise correlations. For $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$ and $\text{var}(\boldsymbol{y}_i) = \boldsymbol{V}_i$, it is common in practice to use estimates that are solutions of *generalized estimating equations* (GEE),

$$\sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}.$$

with $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$. The GEE provide a multivariate generalization of quasi-likelihood methods, generalizing likelihood equations for univariate response without specifying a full multivariate distribution. Such an approach is useful when one's primary interest is modeling the marginal distribution of each y_{it} in terms of explanatory variabless, rather than modeling dependence among $(y_{i1}, y_{i2}, \ldots, y_{iT})$.

In the estimating equations, in $\boldsymbol{V}_i$ GEE methods assume a "working" correlation structure (e.g., exchangeable, autoregressive) for $\boldsymbol{y}_i$. The resulting estimate of $\boldsymbol{\beta}$ is consistent even if the correlation structure is misspecified, when the marginal model is correct. However, standard errors are not appropriate. The method uses empirical robust estimates of the standard errors that are valid even when the correlation structure is misspecified, based on a "sandwich" covariance matrix. The GEE method was originally specified by Liang and Zeger (1986) for univariate y_{it} (e.g., binomial, Poisson), but extensions exist for multinomial models with $c > 2$ response categories. This has mainly been for ordinal responses, as in Lipsitz et al. (1994).

In the multinomial context, let y_{ijt} = 1 if subject i makes response j for observation t. Then, for each pair (s,t) of times, one chooses a working $\text{corr}(y_{ijs}, y_{ikt})$, such as exchangeable (= ρ_{jk} for all s,t). However, Touloumis et al. (2013) showed that certain correlation patterns do not correspond to a legitimate joint multinomial distribution, especially with large c. They argued that it is more sensible to model the covariance based on structure for local odds ratios, both for ordinal and nominal responses. In the binary case, this was suggested by Lipsitz et al. 1991. Structure specified in terms of local odds ratios using adjacent rows and adjacent columns is compatible with all possible multinomial joint distributions and their margins, and it can be used both with ordinal and nominal response variables.

Specifically, for any $s < t$, one supposes that the marginal $P(y_{ias} = 1, y_{ibt} = 1)$ has expected frequencies

$$\log \mu_{ab}^{(st)} = \lambda^{(st)} + \lambda_a^{(s)} + \lambda_b^{(t)} + \beta^{(st)} u_a u_b,$$

for some set of scores $\{u_j\}$. This is a special case of the *linear-by-linear association* loglinear model, in which row and column scores are identical. For this model, the local log odds ratios satisfy

$$\log \left[\frac{\mu_{ab}^{(st)} \mu_{a+1,b+1}^{(st)}}{\mu_{a,b+1}^{(st)} \mu_{a+1,b}^{(st)}} \right] = \beta^{(st)} (u_{a+1} - u_a)(u_{b+1} - u_b).$$

For an ordinal response variable, one takes $\{u_a\}$ to be fixed, monotone scores. For example, scores $\{u_a = a\}$ imply a uniform local log odds ratio that is merely $\beta^{(st)}$ (the so-called *uniform association model*). Exchangeable structure for the T responses then uses the same $\beta^{(st)}$ for each s, t. For a nominal response variable, one treats $\{u_a\}$ as parameters. This pairwise association structure is then a special case of Goodman's (1979) *RC model* and relates to Anderson's (1984) *stereotype model*.

With this multinomial GEE approach, Touloumis et al. noted strong efficiency gains over an independence working structure for studies with strong correlation and time-varying covariates. Touloumis has implemented ordinal and nominal local odds ratio structures with his recently developed *multgee* R package. See

`http://cran.r-project.org/web/packages/multgee/multgee.pdf`.

This package seems to have convergence problems and improper results much less often than existing R multinomial GEE routines. Also, other existing GEE multinomial packages in R do not handle nominal responses.

References

Agresti, A. (2010) *Analysis of Ordinal Categorical Data*, 2nd ed. Wiley.

Agresti, A. (2011) Score and pseudo score confidence intervals for categorical data analysis. *Statistics in Biopharmaceutical Research*, **3**: 163-172.

Agresti, A. (2013) *Categorical Data Analysis*, 3rd ed. Wiley.

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*. Wiley.

Agresti, A., and Coull, B. A. 1998. Approximate is better than exact for interval estimation of binomial parameters. *American Statistician* **52**: 119–126.

Agresti, A., and Kateri, M. (2016) Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics*, **73**: 214–219.

Albert, A., and Anderson, J. A. 1984. On the existence of maximum likelihood estimates in logistic models. *Biometrika* **71**: 1–10.

Anderson, J. A. 1984. Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B* **46**: 1–30.

Anderson, J. A., and Philips, P. R. (1981) Regression, discrimination, and measurement models for ordered categorical variables. *Applied Statistics*, **30**, 22–31.

Brown, L., Cai, T., and DasGupta, A. (2003) Interval estimation in exponential families. *Statistica Sinica* **13**: 19–49.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* **80**: 27–38.

Hedges, L. V., and Olkin, I. (1985) *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Hauck, W. W., and Donner, A. 1977. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* **72**: 851–853.

Jørgensen, B. 1987. Exponential dispersion models. *Journal of the Royal Statistical Society, Series B* **49**: 127–162.

Liang, K. Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.

Lipsitz, S. R., Kim, K., and Zhao, L. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* **13**: 1149–1163.

Lipsitz, S., Laird, N., and Harrington, D. 1991. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**: 153–160.

Lovison, G. 2014. A note on adjusted responses, fitted values and residuals in generalized linear models. *Statistical Modelling* **14**: 337–359.

McKelvey, R. D., and Zavoina, W. 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* **4**: 103–120.

Touloumis, A., Agresti, A., and Kateri, M. 2013. GEE for multinomial responses using a local odds ratios parameterization. *Biometrics* **69**:633–640.