Some Perspectives about Generalized Linear Modeling

Alan Agresti

Professor Emeritus

Department of Statistics

University of Florida, USA

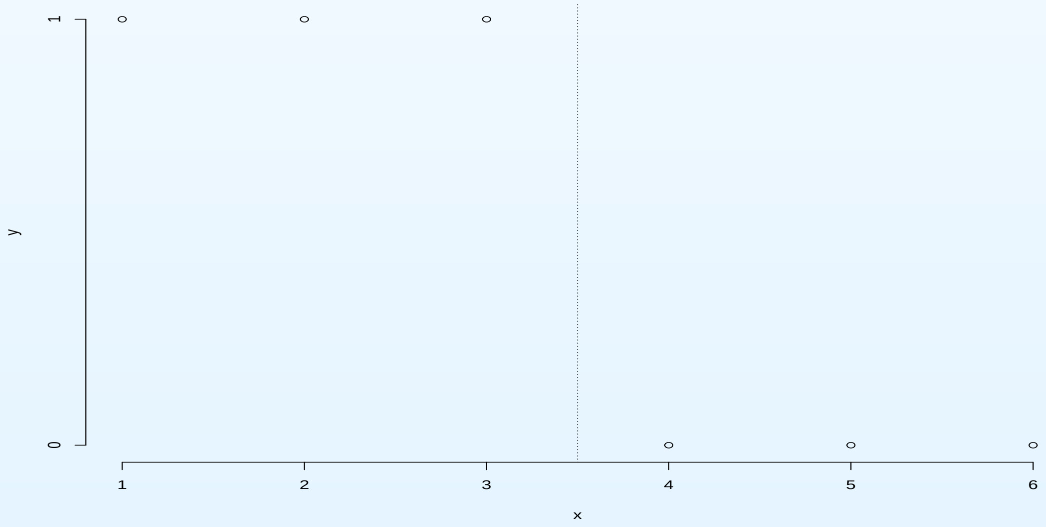
Outline

- With primary emphasis on categorical data, my talk presents *cautions*, *questions*, and *challenges* regarding:
  (1) Wald inference (tests and CI's) with binary responses
  (2) Ordinary linear modeling of ordinal responses
  (3) Behavior and choice of residuals for GLMs
  (4) "Objective" Bayesian inference for high-dimensional data
  (5) Modeling nonnegative responses
  (6) GEE for marginal multinomial models
- Talk has style of tutorial/overview rather than new research, but topics relevant (I hope!) in a conference on *perspectives about linear statistical inference*.
- Motivation: These topics drew my interest while recently writing a book,
  Foundations of Linear and Generalized Linear Models
  (to be published by Wiley 2015).

(1) Wald inference with Large Effects for Binary Data

Infinite maximum likelihood (ML) $\hat{\beta}$ in binary regression models occur when *complete separation* occurs in the space of explanatory variables (Albert and Anderson 1984).

Example: $y = 1$ at $x = 1, 2, 3$, and $y = 0$ at $x = 4, 5, 6$



Infinite Logistic Regression Effects and R Output

```
---------------------------------------------------------------------
> x <- c(1,2,3,4,5,6);  y <- c(1,1,1,0,0,0)  # complete separation
> 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 estimate is
x             -47.2   115264.4       0        1  # actually -infinity

Number of Fisher Scoring iterations: 25

> logLik(fit)
’log Lik.’ -1.107576e-10 (df=2) # maximized log-likelihood = 0
---------------------------------------------------------------------
```

Large Effects: Wald Inference Can Perform Poorly

Wald methods:

- Test H_0: $\beta = 0$ with $z = \hat{\beta}/(SE)$ (or $z^2 \sim \chi^2$, $df = 1$)
- Confidence interval (95%): $\hat{\beta} \pm 1.96(SE)$

where SE is unrestricted ML estimate of standard error.

As $|\beta|$ in a logistic regression model increases (for fixed n), Fisher information decreases so quickly that SE grows faster than β (Hauck and Donner 1977).

Example: $y \sim$ binomial(n, π) distribution, model logit$(\pi) = \beta$. Test H_0: $\beta = 0$ (i.e., $\pi = 0.50$).

$\hat{\beta} =$ logit$(\hat{\pi})$ with $\hat{\pi} = y/n$ has asymptotic var. $[n\pi(1-\pi)]^{-1}$.

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

Large Effects: Wald Inference Can Perform Poorly

Suppose $n = 25$.

$\hat{\pi} = \frac{24}{25}$ stronger evidence against H_0: $\pi = 0.50$ than $\hat{\pi} = \frac{23}{25}$.

Wald statistic = 9.7 when $\hat{\pi} = 24/25$
Wald statistic = 11.0 when $\hat{\pi} = 23/25$.

For comparison, likelihood-ratio statistics are 26.3 and 20.7.

- Note: For large or infinite effects, likelihood-ratio (LR) tests and LR test-based confidence intervals remain valid.
- With infinite ML estimates, can smooth data and produce finite estimates using (1) Bayesian approach, (2) penalized likelihood with aim of reducing bias (Firth, 1993), which corresponds to Bayesian posterior mode with Jeffreys prior.
- Extensions of poor Wald behavior to other GLMs?

ex. Poor Wald performance: CI for binomial parameter

Wald methods for categorical data perform poorly for probabilities near 0 or 1

(for proportions, differences, odds ratio, relative risk).

e.g., for binomial(n, π), 95% Wald confidence interval for π is

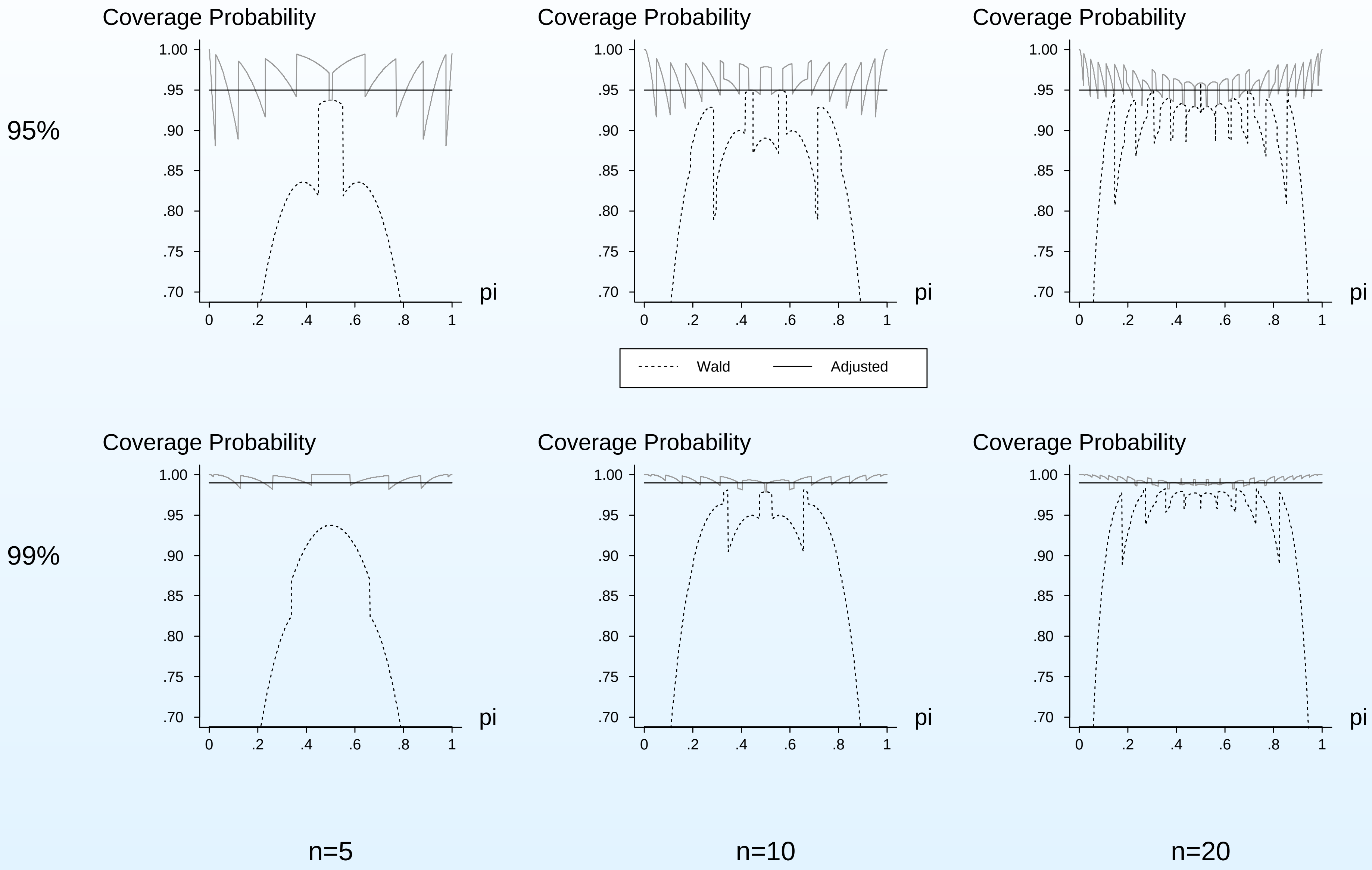
$$\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}.$$

Much worse than interval inverting score test of H_0: $\pi = \pi_0$ with test statistic

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

or approximation to score CI that adds 2 ‘successes’ and 2 ‘failures’ before forming Wald CI (Agresti and Coull 1998).

ex. Coverage prob’s of Wald, “add 2+2” adjusted CI



(2) Ordinal Data: Bias in Using Linear Model with OLS

With ordinal categorical responses,

e.g.,

patient quality of life (excellent, good, fair, poor)

pain (none, little, considerable, severe)

political philosophy (liberal, moderate, conservative)

many researchers (especially in social sciences) assign scores to ordered categories and use ordinary regression methods, estimating parameters using least squares.

Ordinal Data: Bias in Using Linear Model with OLS

Example of bias: Floor effect

For ordinal y, it's often realistic to assume underlying continuous latent variable y^*.

Suppose

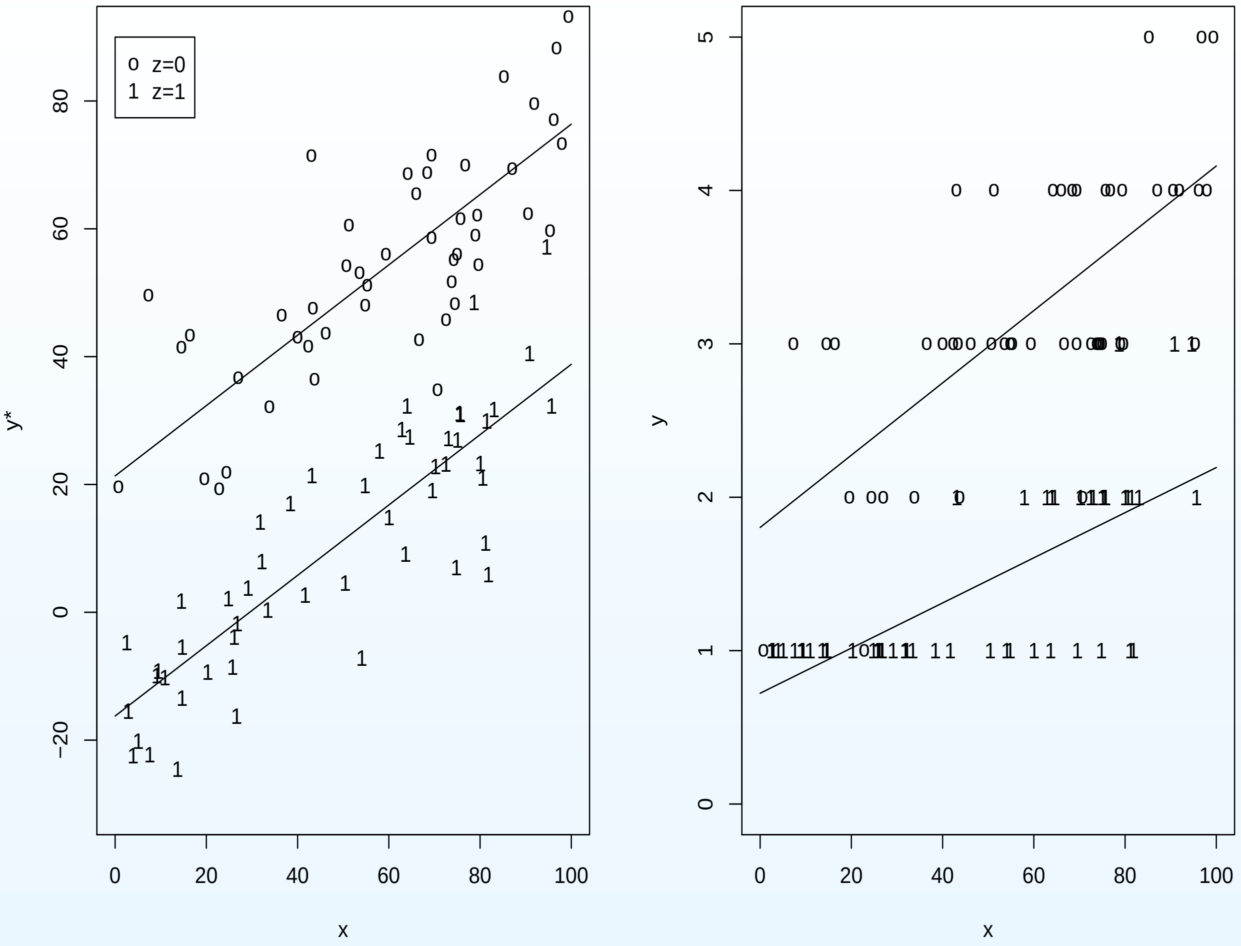
$$y^* = 20.0 + 0.6x - 40z + \epsilon$$

$x \sim$ uniform(0, 100), $P(z=0) = P(z=1) = 0.50$, $\epsilon \sim N(0, 10^2)$.

For random sample of $n = 100$, suppose

$y = 1$ if $y^* \leq 20$, $y = 2$ if $20 < y^* \leq 40$, $y = 3$ if $40 < y^* \leq 60$,

$y = 4$ if $60 < y^* \leq 80$, $y = 5$ if $y^* > 80$.


o z=0
1 z=1
y*
y
x
x
-20
0
20
40
60
80
0
1
2
3
4
5
0
20
40
60
80
100

Floor effect with ordinal data

Fit model $y = \alpha + \beta_1 x + \beta_2 z + \beta_3(x \cdot z) + \epsilon$

to investigate effects and possible interaction.

When $x < 50$ with $z = 1$, high $P(y^* \leq 20) = P(y = 1)$.

Because of floor effect, slope of least squares line when $z = 1$ only half of when $z = 0$. Interaction is statistically and practically significant.

Such spurious effects would not occur with a true ordinal model, such as *cumulative logit model*

$$\text{logit}[P(y \leq j)] = \alpha_j + \beta_1 x + \beta_2 z$$

or *cumulative probit model* (implied by $\epsilon \sim N(0, \sigma^2)$)

$$\Phi^{-1}[P(y \leq j)] = \alpha_j + \beta_1 x + \beta_2 z.$$

(3) 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 a GLM

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

for link function g, model matrix $\boldsymbol{X}$ with p explanatory variables. Maximum likelihood (ML) fitted values $\hat{\boldsymbol{\mu}}$.

“Ordinary linear model”: identity link $\boldsymbol{\mu} = \boldsymbol{X\beta}$, and $\boldsymbol{V} = \sigma^2\boldsymbol{I}$.

- Ordinary linear model exploits orthogonal decomposition

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

- With GLMs, $\hat{\boldsymbol{\mu}}$ and $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ are not orthogonal when depart from identity link and constant variance.

Correlation(GLM Fitted Values, Residuals) Approx. 0?

- Conventional wisdom: As n increases, $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ becomes asymptotically uncorrelated with $\hat{\boldsymbol{\mu}}$.

- For large n, if $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ approximately uncorrelated with $\hat{\boldsymbol{\mu}}$, then $\boldsymbol{V} \approx \text{var}(\hat{\boldsymbol{\mu}}) + \text{var}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, and

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

  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

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

Correlation(GLM Fitted Values, Residuals) Approx. 0?

- 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 corr$(\boldsymbol{y} - \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}})$ close enough to 0 that we can ignore it?

- It seems we need to consider two types of asymptotics: Traditional $n \to \infty$, and alternative with n fixed and asymptotics applying to individual components, such as binomial indices and Poisson expected counts in contingency table.

- For the alternative, (*small-dispersion asymptotics*, Jørgensen 1987), individual y_i asymptotically normal, $(\boldsymbol{y} - \boldsymbol{\mu})$ and $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ jointly have asymptotic normal distribution, as does their difference.

Correlation(GLM Fitted Values, Residuals) Approx. 0?

- Lovison (2014): If $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ and $\hat{\boldsymbol{\mu}}$ were not approximately 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 contradicts the ML estimator $\hat{\boldsymbol{\mu}}$ being asymptotically efficient.

- Argument is an asymptotic version for ML estimators of one in Gauss–Markov Theorem that unbiased estimators other than least squares estimator have difference from that estimator that is uncorrelated with it.

- Lovison shows *weighted* version of adjusted responses that has approximately constant variance has orthogonality of fitted values and residuals. On original scale, such residual is “Pearson residual” $e_i = (y_i - \hat{\mu}_i)/\sqrt{v(\hat{\mu}_i)}$ for variance function v.

Pearson residuals vs. standardized residuals

For contingency tables, Pearson residual is popular, because of decomposition of Pearson chi-squared (and corresponding decomposition of deviance provides *deviance residual*).

e.g., for Poisson counts $\{y_i\}$,

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

Editorial comment: Preferable to use *standardized residual*

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

for “leverage” $\hat{h}_{ii}$ from the estimated hat matrix $\widehat{\boldsymbol{H}}$.

Pearson residuals vs. standardized residuals

- For small dispersion asymptotics, r_i (but not Pearson e_i) is asymptotically *standard normal* when model holds.
- Appropriately recognizes redundancies in data.
  e.g., for *independence model* (Poisson or multinomial) for 2×2 table of counts $\{y_{ij}\}$ with fitted values

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

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

  In 2×2 tables, $df = 1$ (all four $|y_{ij} - \hat{\mu}_{ij}|$ identical).
  Yet, all four Pearson residuals can take different values.

  $r_{11} = -r_{12} = -r_{21} = r_{22}$ and any $r_{ij}^2 = X^2$.

(4) Bayesian methods in large dimensions

- For “objective Bayesian” approach, how to select prior distributions when model has very large number of parameters p?
- Even with very diffuse prior, estimated posterior effect may depend strongly on choice of prior.

Example: multinomial data with p outcome categories, $p >> n$.

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$ = multinomial trial for observation i.
$y_{ij} = 1$ when outcome in category j, $y_{ij} = 0$ otherwise.

Then $\sum_j y_{ij} = 1$.

Let $n_j = \sum_i y_{ij}$ = total number of observations in category j.

Here, for simplicity, consider model without explanatory variables, let $\pi_j = P(y_{ij} = 1)$.

Bayesian methods for multinomial with large p

- Consider *Dirichlet* prior, proportional to $\prod_{j=1}^{p} \pi_j^{\alpha_j - 1}$.
- Posterior density is Dirichlet with parameters $\{n_j + \alpha_j\}$, posterior mean of π_j is $(n_j + \alpha_j)/(n + \sum_k \alpha_k)$.

  $\{\alpha_j = 1\}$, uniform prior distribution over probability simplex, smooths toward equi-probability model.

Example: Suppose $n = 100$ but $p = 1000$. π_j has
prior mean = 0.001,
posterior mean = $(n_j + 1)/(n + p) = (n_j + 1)/1100$.

When $n_j = 1$, $n_j/n = 0.010$, posterior mean = 0.0018,
shrinking sample proportion n_j/n toward 0.001.

What if $n_j = 100$, $n_j/n = 1.0$? Posterior mean = 0.092.
Prior is diffuse but has very strong impact on results.

Bayesian methods for multinomial with large p

Berger (2013): Prior should have marginal posteriors close to posterior we'd obtain in single-parameter case.

e.g., aim for posterior of π_j to be approximately beta dist. with parameters $n_j + 1$ and $n - n_j + 1$, which we'd obtain with uniform prior for binomial with parameter π_j.

So, use Dirichlet hyperparameters $\{\alpha_j = 2/p\}$ instead of $\{\alpha_j = 1\}$, yielding posterior mean for π_j of $(n_j + 2/p)/(n + 2)$.

Example: With $n = 100$ observations in $p = 1000$ cells, this is 0.0098 when $n_j = 1$ and is 0.980 when $n_j = 100$.

But suppose $n = 2$, of which $n_j = 1$. Posterior mean = 0.25. Shrink $n_j/n = 1/2$ based on only $n = 2$ so little toward 0.001?

(5) Models for *Nonnegative* y that Merit More Use

1. *Continuous* responses

When $y \geq 0$, standard deviation of y often grows proportionally to mean.

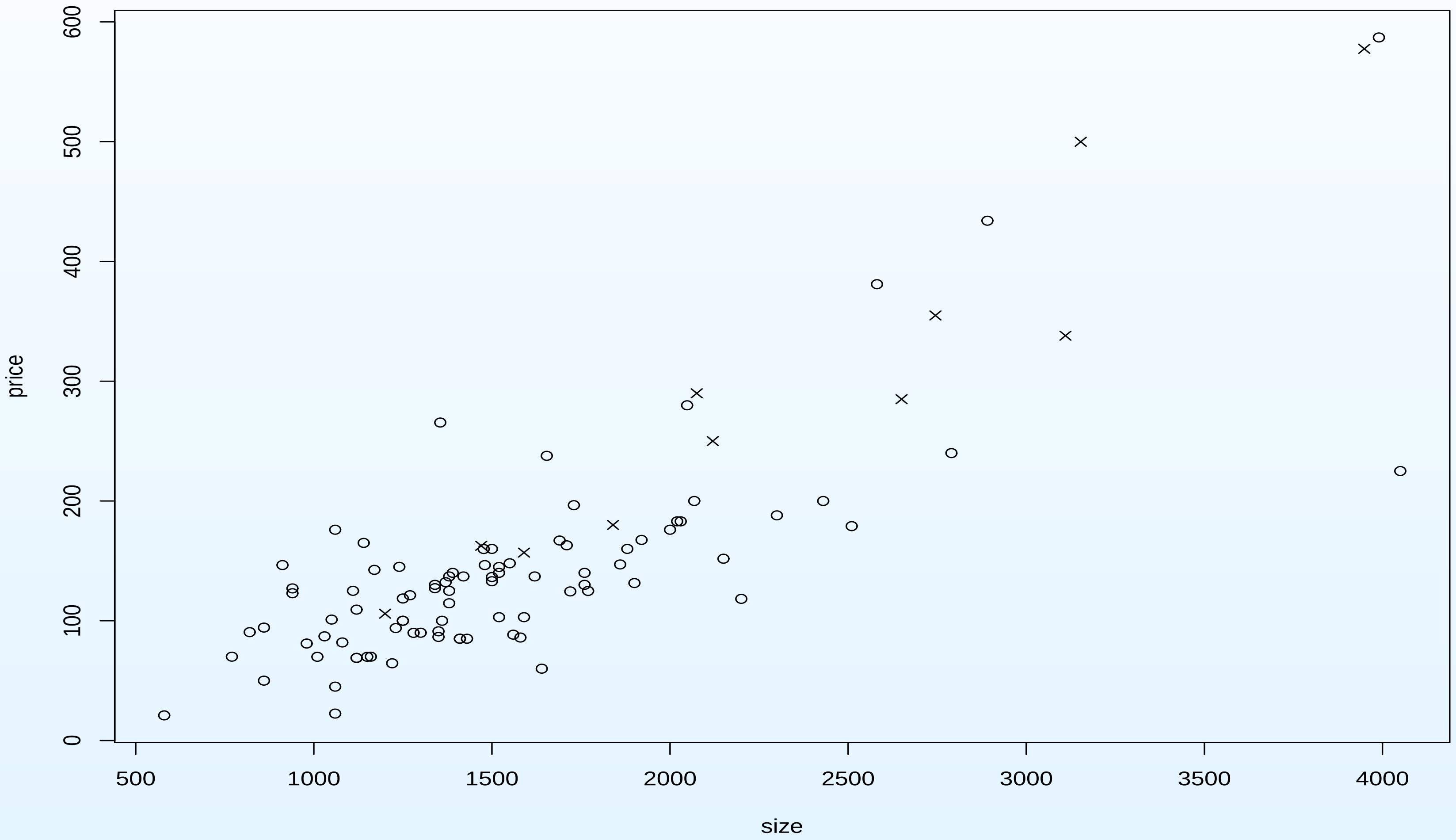
Why not use GLM assuming gamma distribution for response?

Alternative to log-normal model, with advantages:

- With log link, modeling $\log[E(y)]$ rather than $E[\log(y)]$.
- Sometimes identity link is adequate.

Example: Modeling house selling prices (thousands of dollars) in terms of size of house (square feet) and whether new

Scatterplot (× = new, o = not new)



Some summaries of model fits

Standardized residuals for outlier:
Normal: no interaction -4.2, interaction -3.8
Gamma: no interaction -1.5, interaction -1.5

Cook's distance for outlier:
Normal: no interaction 1.3, interaction 1.0
Gamma: no interaction 0.03, interaction 0.02

AIC values:
Normal: no interaction 1086.1, interaction 1079.9
Gamma: no interaction 1049.5, interaction 1047.9

Normal models have $\hat{\sigma} \approx 52$ for all μ.

Gamma models have $\hat{\sigma} \approx 0.33\hat{\mu}$

Models for *Nonnegative* y that Merit More Use

Count responses:

- Poisson models usually fail, because of overdispersion ($v(\mu) = \mu$, mode = integer part of μ).
- Negative binomial (NB) GLM gives much more flexibility ($v(\mu) = \mu + \mu^2/k$, mode can be 0 for any μ).
- But count data are often *zero-inflated*:

  Examples are counts of activities for which many subjects necessarily report 0, such as number of times during some period of going to gym, having an alcoholic drink, smoking marijuana, having sexual intercourse.

Zero-Inflated Models

- NB sometimes adequate for zero-inflation, but fits poorly when data strongly bimodal.
- Zero-inflated Poisson (ZIP, Lambert 1992) and zero-inflated negative binomial (ZINB) provide mixture of ordinary count data with one that places all its mass at 0.
- ZIP often does not allow sufficient dispersion for count-data component. ZINB gives considerable flexibility.

ZINB: Simultaneous logistic model for $P(y = 0)$ and NB loglinear model for mean response

Example: Modeling horseshoe crab counts at `www.stat.ufl.edu/~aa/cda/crabs.pdf`

Challenge: Random effects in ZINB (Min and Agresti 2005) and overall summaries of effects

(6) Improved Marginal Modeling of Multinomial Data

Data: Each subject has cluster of correlated observations
$\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})^T$
(e.g., repeated measures or longitudinal data)

For each y_{it} marginally, $g(\mu_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta}$.

For discrete data, ML awkward because of lack of simple multivariate dist. characterized by pairwise correlations.

For $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$ and var$(\boldsymbol{y}_i) = \boldsymbol{V}_i$, can 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 $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$.

Generalized estimating equations (GEE) approach

The GEE provide multivariate generalization of quasi-likelihood methods, generalizing likelihood equations for univariate response without specifying full multivariate distribution.

Steps of GEE Methodology:

- Assume marginal model for each component of $\boldsymbol{\mu}_i$.
- In $\boldsymbol{V}_i$, assume "working" correlation structure (e.g., exchangeable, autoregressive) for $\boldsymbol{y}_i$.
- Estimate of $\boldsymbol{\beta}$ consistent even if correlation structure misspecified (if marginal model correct).
- Method uses "empirical" robust estimates of std. errors that are valid even if correlation structure misspecified, based on "sandwich" covariance matrix.

GEE for correlated *multinomial* observations

- GEE method originally specified (Liang and Zeger 1986) for univariate y_{it} (e.g., binomial, Poisson).
- Extensions exist for multinomial (mainly ordinal) models with $c > 2$ response categories; e.g., Lipsitz et al. (1994).

  Let y_{ijt} = 1 if subject i makes response j for observation t. Then, for each pair (s, t) of times, choose working corr(y_{ijs}, y_{ikt}), such as exchangeable (= ρ_{jk} all s, t).
- Touloumis, Agresti, and Kateri (2013): Certain correlation patterns do not correspond to legitimate joint multinomial distribution, especially with large c.
- More sensible to model covariance based on structure for “local” odds ratios, for *ordinal* and *nominal* responses.

Multinomial GEE using working local odds ratios

For any $s < t$, suppose marginal $P(y_{ias} = 1, y_{ibt} = 1)$ has expected frequencies

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

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

- For *ordinal* response, $\{u_a\}$ are fixed, monotone scores. e.g., with $\{u_a = a\}$, common local log odds ratio = $\beta^{(st)}$. Exchangeable structure uses same $\beta^{(st)}$ for each s, t.
- For *nominal* response, treat $\{u_a\}$ as parameters, and this structure is special case of Goodman's (1979) *RC model*.

Implementing nominal / ordinal multinomial GEE

- Strong efficiency gains over independence working structure for studies with strong correlation and time-varying covariates.

- Touloumis (2013) has implemented ordinal and nominal local odds ratio structures with *multgee* R package.

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

  Has convergence problems much less often than existing R multinomial GEE routines.

(In R, I've had varying success using *ordgee*, *repolr* for ordinal data. Nominal data?)

See also recent review of software for GEE for ordinal data by Noorae, Molenberghs, and van den Heuvel (*CSDA*, 2014).

Thanks to LinStat scientific committee for the invitation.