

Group testing regression models with fixed and random effects

Joshua M. Tebbs*
Associate Professor
Department of Statistics
University of South Carolina

November 23, 2009

* This is collaborative research with Peng Chen (Takeda Pharmaceuticals) and Christopher R. Bilder ([University of Nebraska](#)). This work is funded by [National Institutes of Health](#) Grant R01-AI067373.

Chen, P., Tebbs, J., and Bilder, C. (2009). Group testing regression models with fixed and random effects. *Biometrics*, in press.

OUTLINE

1. Nebraska data example
2. Motivation for pooling
3. Regression models with mixed effects
4. Tests for homogeneity among individuals
5. Nebraska data
6. Discussion

1. Nebraska data example

- Nebraska Health and Human Services collects **chlamydia** and **gonorrhea** testing data (binary response); Region VII of IPP
- 30,000 individual tests/year
- Covariates recorded:
 1. **age, gender, other demographic information**
 2. **number of sexual partners, STD symptoms**
 3. **cervical friability/urethritis status**
 4. **testing site**
- Researchers are interested in using **pooled testing** for estimation and identification of positive subjects

2. Motivation for pooling: Group testing

- **General premise:** tests are performed on “pools” of individuals (e.g., urine, swabs, blood, etc.)
- The basic model assumes a binary pool response
 - **positive pool:** at least one individual in the pool is positive
 - **negative pool:** all individuals in the pool are negative
- Common in blood/infectious disease screening
- Idea applied to **other areas**; e.g., drug discovery, genetics, entomology, etc.
 - Unifying reason for use: Reduce testing costs
- **Estimation** versus classification

2. (A very) brief review of group testing research

- A large majority of the research (up until around 2000) has assumed that the population of individuals being screened is **homogenous** with common

p = probability of individual positivity

This assumption flies in the face of reality!!

- Individual covariate information is (almost always) available
 - Estimation: Compute covariate adjusted estimates of risk
 - Identification: Use covariates to improve retesting efficiency
- **Vansteelandt et al. (2000)** and **Xie (2001)** developed regression approaches to handle pooled response data from group testing
 - Both consider **fixed effects models** and treat individuals as **independent**

3. Regression models with mixed effects

- Our interest is in **modelling** the probability of a single infection (e.g., chlamydia) based on pooled testing results
 - We are not interested in identification today
- Individuals are **randomly assigned** to pools within site
- Let $Y_{ijk} = 1$ if the k th individual in the j th pool at site i is positive, $Y_{ijk} = 0$ otherwise, for $i = 1, 2, \dots, l$, $j = 1, 2, \dots, n_i$, $k = 1, 2, \dots, c_{ij}$
 - l = number of sites
 - n_i = number of pools in site i
 - c_{ij} = pool size for pool j in site i
- We treat clinic site effects as **random**

3. Regression models with mixed effects

- Site-specific $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_l \sim \text{iid } \mathcal{N}_q(\mathbf{0}, \mathbf{D})$, where $\mathbf{D} \equiv \mathbf{D}(\boldsymbol{\varphi})$ and $\boldsymbol{\varphi}$ is an $m \times 1$ vector of **variance components**
- **Model:** Relate the latent status Y_{ijk} to the covariates through

$$\text{pr}(Y_{ijk} = 1 | \mathbf{u}_i) = g^{-1}(\boldsymbol{\beta}' \mathbf{x}_{ijk} + \mathbf{u}_i' \mathbf{z}_{ijk})$$

- $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects parameters
 - \mathbf{x}_{ijk} is a $p \times 1$ covariate vector associated with the fixed effects
 - \mathbf{z}_{ijk} is a $q \times 1$ covariate vector associated with the site-specific random effects
 - $g(\cdot)$ is a known, monotonic, differentiable link function
- **Difficulty:** Y_{ijk} is not observed, unless $c_{ij} = 1!!$

3. Regression models with mixed effects

- **Observed data:** $T_{ij} = 1$ if the j th pool at site i tests positive, $T_{ij} = 0$ otherwise, for $i = 1, 2, \dots, l$ and $j = 1, 2, \dots, n_i$
- We assume assay tests have the following characteristics:
 - sensitivity = γ_1 , specificity = γ_2
 - γ_1 and γ_2 are constants, close to 1, independent of c_{ij}
- We also assume that
 - individuals from different sites are independent
 - individuals from the same site are conditionally independent (given \mathbf{u}_i)
- The probability that the j th pool in the i th site tests **positive** is

$$\text{pr}(T_{ij} = 1 | \mathbf{u}_i) = \gamma_1 + (1 - \gamma_1 - \gamma_2) \prod_{k=1}^{c_{ij}} \{1 - g^{-1}(\boldsymbol{\beta}' \mathbf{x}_{ijk} + \mathbf{u}_i' \mathbf{z}_{ijk})\}$$

3. Regression models with mixed effects

- The log-likelihood of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\varphi}')'$ based on $\mathbf{T} = (T_{ij})$ is

$$l(\boldsymbol{\theta}|\mathbf{T}) = \sum_{i=1}^l \log \left\{ \int_{\mathcal{R}^q} \prod_{j=1}^{n_i} f_{ij}(c_{ij}, \mathbf{x}_{ij}, \mathbf{u}_i, \boldsymbol{\beta}) f(\mathbf{u}_i|\boldsymbol{\varphi}) d\mathbf{u}_i \right\},$$

where

$$f_{ij}(c_{ij}, \mathbf{x}_{ij}, \mathbf{u}_i, \boldsymbol{\beta}) = \begin{cases} \gamma_1 + (1 - \gamma_1 - \gamma_2) \prod_{k=1}^{c_{ij}} \{1 - g^{-1}(\boldsymbol{\beta}' \mathbf{x}_{ijk} + \mathbf{u}_i' \mathbf{z}_{ijk})\}, & T_{ij} = 1 \\ 1 - \gamma_1 - (1 - \gamma_1 - \gamma_2) \prod_{k=1}^{c_{ij}} \{1 - g^{-1}(\boldsymbol{\beta}' \mathbf{x}_{ijk} + \mathbf{u}_i' \mathbf{z}_{ijk})\}, & T_{ij} = 0, \end{cases}$$

and

$$f(\mathbf{u}_i|\boldsymbol{\varphi}) = (2\pi)^{-q/2} |\mathbf{D}|^{-1/2} \exp(-\mathbf{u}_i' \mathbf{D}^{-1} \mathbf{u}_i / 2)$$

- Goal:** Compute the MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\varphi}})'$

3. Regression models with mixed effects: Estimation

- We investigated **two methods** to compute $\hat{\boldsymbol{\theta}}$
 - Guass-Hermite quadrature with Newton Raphson (**good for q small**)
 - Monte-Carlo EM algorithm (**more general**)
- The **complete data log-likelihood** (for MCEM) can be written as

$$l_c(\boldsymbol{\theta}|\mathbf{T}, \mathbf{u}) = I_1 + I_2 + c_0,$$

where

$$I_1 = \sum_{i=1}^l \sum_{j=1}^{n_i} \log f_{ij}(c_{ij}, \mathbf{x}_{ij}, \mathbf{u}_i, \boldsymbol{\beta})$$

$$I_2 = -\frac{l}{2} \log |\mathbf{D}| - \frac{1}{2} \sum_{i=1}^l \mathbf{u}_i' \mathbf{D}^{-1} \mathbf{u}_i$$

and c_0 is **free** of $\boldsymbol{\theta}$

3. Regression models with mixed effects: MCEM details

1. **(E-step)**: For a given $b = 0, 1, 2, \dots$, approximate $E(I_1|\mathbf{T})$ and $E(I_2|\mathbf{T})$ by

$$\widehat{I}_1^{(b)} = \frac{1}{M} \sum_{h=1}^M \sum_{i=1}^l \sum_{j=1}^{n_i} \log f_{ij}(c_{ij}, \mathbf{x}_{ij}, \mathbf{u}_i^{(h)}, \boldsymbol{\beta})$$

$$\widehat{I}_2^{(b)} = -\frac{l}{2} \log |\mathbf{D}| - \frac{1}{2M} \sum_{h=1}^M \sum_{i=1}^l \mathbf{u}_i^{(h)'} \mathbf{D}^{-1} \mathbf{u}_i^{(h)},$$

respectively, where $\mathbf{u}_i^{(h)}$, $h = 1, 2, \dots, M$, are M draws from the conditional distribution $f(\mathbf{u}_i|\mathbf{T}; \boldsymbol{\theta}^{(b)})$, $i = 1, 2, \dots, l$, using the **MH algorithm**

2. **(M-step)**: Maximize $\widehat{I}_1^{(b)}$ with respect to $\boldsymbol{\beta}$, obtaining a new estimate $\boldsymbol{\beta}^{(b+1)}$; maximize $\widehat{I}_2^{(b)}$ with respect to $\boldsymbol{\varphi}$, obtaining a new estimate $\boldsymbol{\varphi}^{(b+1)}$
3. Repeat E and M steps until $\|\boldsymbol{\beta}^{(b+1)} - \boldsymbol{\beta}^{(b)}\|$ and $\|\boldsymbol{\varphi}^{(b+1)} - \boldsymbol{\varphi}^{(b)}\|$ are small

3. Covariance matrix estimation

- Information matrix $I(\boldsymbol{\theta})$ can be **estimated** using
 - negative Hessian at the last iteration of NR (quadratures)
 - missing information principle (MCEM)
- The observed information matrix $I(\boldsymbol{\theta})$ can be written

$$\begin{aligned}
 I(\boldsymbol{\theta}) &= -E \left\{ \frac{\partial^2 l_c(\boldsymbol{\theta} | \mathbf{T}, \mathbf{u})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \middle| \mathbf{T} \right\} - \text{cov} \left\{ \frac{\partial \log l_c(\boldsymbol{\theta} | \mathbf{T}, \mathbf{u})}{\partial \boldsymbol{\theta}} \middle| \mathbf{T} \right\} \\
 &= -E_{\mathbf{u} | \mathbf{T}} \left[\begin{array}{cc} \frac{\partial^2 I_1}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} + \frac{\partial I_1}{\partial \boldsymbol{\beta}} \frac{\partial I_1}{\partial \boldsymbol{\beta}'} & \frac{\partial I_1}{\partial \boldsymbol{\beta}} \frac{\partial I_2}{\partial \boldsymbol{\varphi}'} \\ \frac{\partial I_2}{\partial \boldsymbol{\varphi}} \frac{\partial I_1}{\partial \boldsymbol{\beta}'} & \frac{\partial^2 I_2}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}'} + \frac{\partial I_2}{\partial \boldsymbol{\varphi}} \frac{\partial I_2}{\partial \boldsymbol{\varphi}'} \end{array} \right]
 \end{aligned}$$

- $I(\boldsymbol{\theta})$ can be **estimated** using random samples $\mathbf{u}_i^{(h)}$ from the conditional density $P(\mathbf{u} | \mathbf{T}; \hat{\boldsymbol{\theta}})$, generated using the MH algorithm
- **Wald inference** is available; valid for l large

3. Simulation evidence

- True model:

$$\text{logit}\{\text{pr}(Y_{ijk} = 1|u_i)\} = \beta_0 + \beta_1 x_{ijk} + u_i,$$

where

- $u_i \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 = 1$
 - $\beta = (\beta_0, \beta_1)' = (-5, 1)'$
 - $x_{ijk} \sim \mathcal{N}(0, 0.64)$
 - $\gamma_1 = \gamma_2 = 1$ (perfect testing)
- These settings provide a mean prevalence of about 1.4 percent (range: 0.1 to 5.0 percent)
 - Simulation results based on $B = 500$ simulated data sets
 - Results from quadratures presented only; MCEM results available in paper

3. Simulation evidence

| n | c | | $\hat{\beta}_0$ ($\beta_0 = -5$) | | $\hat{\beta}_1$ ($\beta_1 = 1$) | | $\hat{\sigma}$ ($\sigma = 1$) | |
|-----|-----|---------|------------------------------------|-------------|-----------------------------------|-------------|---------------------------------|-------------|
| | | | IND | POOL | IND | POOL | IND | POOL |
| 100 | 2 | Mean | -5.04 | -5.06 | 1.02 | 1.02 | 0.91 | 0.92 |
| | | Std dev | 0.48 | 0.49 | 0.26 | 0.32 | 0.37 | 0.36 |
| | | Cov | | 0.95 | | 0.96 | | 0.97 |
| 40 | 5 | Mean | -5.04 | -5.07 | 1.02 | 0.97 | 0.91 | 0.93 |
| | | Std dev | 0.48 | 0.54 | 0.26 | 0.52 | 0.37 | 0.36 |
| | | Cov | | 0.94 | | 0.95 | | 0.97 |
| 20 | 10 | Mean | -5.04 | -5.11 | 1.02 | 0.89 | 0.91 | 0.94 |
| | | Std dev | 0.48 | 0.59 | 0.26 | 0.74 | 0.37 | 0.37 |
| | | Cov | | 0.96 | | 0.91 | | 0.98 |

- $N = 2000$ individuals from $l = 10$ sites; $n =$ number of pools per site; $c =$ pool size. Margin of error for Wald coverage (Cov) is 0.03.

4. Tests for homogeneity among individuals

- **Goal:** Test whether individual random effects are present using the available pooled responses
- **Simplest case:** $q = 1$ and $\mathbf{z}_{ijk} = 1$, for all i, j , and k

$$H_0 : \sigma^2 = 0 \text{ versus } H_1 : \sigma^2 > 0,$$

where $\sigma^2 = \text{var}(u_i)$

- Sufficient for Nebraska IPP data
- Generalizations for $q > 1$ are possible
- The test of H_0 versus H_1 has been described as “nonstandard” (Self and Liang, 1987); see also Molenberghs and Verbeke (2007)

4. Tests for homogeneity: LRT

- **LRT** statistic:

$$T_{LR} = 2 \ln \left\{ \frac{\max_{H_0 \cup H_1} l(\boldsymbol{\beta}, \sigma^2)}{\max_{H_0} l(\boldsymbol{\beta}, \sigma^2)} \right\},$$

where $l(\boldsymbol{\beta}, \sigma^2)$ is the log-likelihood with $\boldsymbol{\varphi} = \sigma^2$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$

- Under $H_0 : \sigma^2 = 0$,

$$T_{LR} \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2,$$

where χ_0^2 is a **point mass** distribution at 0 and χ_1^2 denotes the χ^2 distribution with 1 degree of freedom

- **Large** values of T_{LR} are evidence against H_0

4. Tests for homogeneity: Score

- **Score** statistic:

$$T_S = \begin{cases} \{S(\hat{\beta})\}^2 / \widehat{\text{var}}\{S(\hat{\beta})\}, & S(\hat{\beta}) > 0 \\ 0, & S(\hat{\beta}) \leq 0 \end{cases}$$

- $S(\beta)$ = **score function**
 - $\hat{\beta}$ is the **MLE** computed under $H_0 : \sigma^2 = 0$
 - $\widehat{\text{var}}\{S(\hat{\beta})\}$ is an **estimate** of $\text{var}\{S(\hat{\beta})\}$
 - Full details given in the paper and online appendices
- Under $H_0 : \sigma^2 = 0$,

$$T_S \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$$

- **Large** values of T_S are evidence against H_0

4. Simulation evidence

- True model:

$$\text{logit}\{\text{pr}(Y_{ijk} = 1|u_i)\} = \beta_0 + \beta_1 x_{ijk} + u_i,$$

where

- $u_i \sim \mathcal{N}(0, \sigma^2)$
 - $\boldsymbol{\beta} = (\beta_0, \beta_1)' = (-4, 1)'$
 - $x_{ijk} \sim \mathcal{N}(0, 0.64)$
 - $\gamma_1 = \gamma_2 = 1$ (perfect testing)
- Simulation results based on $B = 500$ simulated data sets
 - Margin of error for size estimate $\hat{\alpha}$ is 0.03

4. Simulation evidence

| N | l | n | c | | $\sigma = 0$ | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
|------|-----|-----|-----|----------|--------------|------|------|------|------|------|
| | | 50 | 2 | T_{LR} | 0.03 | 0.10 | 0.30 | 0.55 | 0.78 | 0.89 |
| | | | | T_S | 0.04 | 0.13 | 0.36 | 0.63 | 0.83 | 0.92 |
| 2000 | 20 | 20 | 5 | T_{LR} | 0.04 | 0.08 | 0.24 | 0.45 | 0.75 | 0.88 |
| | | | | T_S | 0.05 | 0.12 | 0.28 | 0.51 | 0.80 | 0.88 |
| | | 10 | 10 | T_{LR} | 0.03 | 0.08 | 0.19 | 0.38 | 0.61 | 0.74 |
| | | | | T_S | 0.05 | 0.11 | 0.26 | 0.45 | 0.69 | 0.80 |

- l = number of sites
- n = number of pools/site
- c = pool size

5. Nebraska data

- The state of Nebraska takes part in the nationwide IPP through its Sexually Transmitted Diseases and Infertility Control Program
- $l = 78$ clinic sites throughout the state
- Urine or swab (cervical or male urethra) specimens and covariate information are collected on each individual
- Use individual testing results from 2006Q1 \implies 6,138 subjects; number of subjects per site varies from 1 to 540
- We consider the model:

$$\begin{aligned} \text{logit}\{\text{pr}(Y_{ijk} = 1|u_i)\} &= \beta_0 + \beta_1 \text{Age}_{ijk} + \beta_2 \text{Gender}_{ijk} \\ &\quad + \beta_3 \text{Urethritis}_{ijk} + \beta_4 \text{Symptoms}_{ijk} + u_i, \end{aligned}$$

where $u_i \sim \mathcal{N}(0, \sigma^2)$, $\gamma_1 = 0.95$, and $\gamma_2 = 0.98$

5. Nebraska data

| Chlamydia | | | | | | | |
|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|--------|
| c | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\sigma}$ | Reject |
| 1 | −2.45(0.26) | −0.05(0.01) | 0.68(0.14) | 0.87(0.25) | 0.46(0.12) | 0.46(0.09) | — |
| 2 | −2.46(0.35) | −0.05(0.01) | 0.65(0.18) | 0.86(0.31) | 0.50(0.16) | 0.46(0.09) | 1.00 |
| 5 | −2.65(0.44) | −0.03(0.02) | 0.47(0.24) | 0.93(0.46) | 0.57(0.24) | 0.39(0.10) | 0.88 |
| 8 | −2.89(0.45) | −0.01(0.02) | 0.31(0.26) | 0.96(0.60) | 0.58(0.29) | 0.32(0.11) | 0.77 |
| Gonorrhea | | | | | | | |
| c | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\sigma}$ | Reject |
| 1 | −4.88(0.50) | −0.02(0.01) | 0.25(0.26) | 1.57(0.34) | 0.95(0.23) | 0.99(0.24) | — |
| 2 | −5.42(0.62) | −0.01(0.02) | 0.48(0.34) | 1.39(0.39) | 1.04(0.30) | 0.90(0.27) | 1.00 |
| 5 | −6.21(0.73) | 0.00(0.02) | 0.79(0.45) | 1.20(0.51) | 1.19(0.42) | 0.83(0.27) | 1.00 |
| 8 | −6.62(0.81) | 0.01(0.03) | 0.93(0.49) | 1.01(0.65) | 1.42(0.50) | 0.83(0.27) | 0.96 |

- Pooled estimates based on $B = 100$ simulated data sets
- Pool size c not necessarily equal within site (unbalanced data)
- Score test for homogeneity

6. Discussion

- We have proposed new regression models to incorporate **random effects** with pooled binary responses
- We assumed individuals are assigned to pools (within site) **at random**
 - Should an individual's covariate **(\mathbf{x}, \mathbf{z})** play a role?
 - What are the **gains**?
- Our model specifies that individuals are pooled **within site**
 - may not be practical in some applications
 - would expect estimates/power to be affected **drastically** if individuals from different sites were pooled together
- **Joint regression models** for multiple infections