

Regression models for group testing data

Joshua M. Tebbs

University of South Carolina
Department of Statistics

October 25, 2011

This work is funded by the National Institutes of Health (R01-AI067373).

- 1 Nebraska IPP
- 2 Group testing
- 3 Regression models
- 4 Informative retesting
- 5 Multiple traits
- 6 Illustration
- 7 Discussion

Nebraska IPP

- Nebraska Health and Human Services collects **chlamydia** and **gonorrhea** testing data
- Region VII of IPP
- 30,000 individual tests/year
- Covariates recorded:
 - **age, gender, race**
 - **number of sexual partners, STD symptoms**
 - **clinical observations (PID, urethritis, etc.)**
 - **testing site, reason for visit**
- Researchers are interested in using **group testing** for estimation and identification of positive subjects

Group testing

- **General premise:** tests are performed on “pools” of individuals (e.g., urine, swabs, blood, etc.)
 - **Positive pool:** at least one individual in the pool is positive
 - **Negative pool:** all individuals in the pool are negative
- Group testing has been applied to many areas
 - blood screening
 - drug discovery
 - genetics
 - food safety
 - animal/plant disease applications
- Compelling alternative to testing subjects individually
- **Estimation** versus **classification**

Notation

- Suppose N individuals are drawn from a population and each individual is assigned to one of K pools
- Define

$$\tilde{Y}_{ik} = \begin{cases} 1, & \textit{ith individual in the kth pool is truly positive} \\ 0, & \textit{otherwise,} \end{cases}$$

for $i = 1, 2, \dots, I_k$ and $k = 1, 2, \dots, K$

- The \tilde{Y}_{ik} (latent) statuses are assumed independent
- We do get to observe

$$\mathbf{x}_{ik} = (1, x_{ik1}, x_{ik2}, \dots, x_{ikp})',$$

a $(p + 1) \times 1$ vector of covariates for each individual

Notation

- Define

$$\tilde{Z}_k = \begin{cases} 1, & \text{kth pool is truly positive} \\ 0, & \text{otherwise} \end{cases}$$

$$Z_k = \begin{cases} 1, & \text{kth pool tests positive} \\ 0, & \text{otherwise} \end{cases}$$

- The **sensitivity** and the **specificity** of a diagnostic test are

$$\gamma_1 = \text{pr}(Z_k = 1 | \tilde{Z}_k = 1) \quad \text{and} \quad \gamma_2 = \text{pr}(Z_k = 0 | \tilde{Z}_k = 0)$$

- It is easy to show that

$$\text{pr}(Z_k = 1) = \gamma_1 + \gamma_2 \prod_{i=1}^{I_k} (1 - p_{ik}),$$

where $\gamma_{12} = 1 - \gamma_1 - \gamma_2$ and $p_{ik} = \text{pr}(\tilde{Y}_{ik} = 1)$

Assumptions

- We assume that γ_1 and γ_2
 - are known
 - do not depend on the covariates
 - do not depend on I_k (pool size)
- Empirical evidence supports last assumption:
 - Litvak et al. (1994), $I_k = 15$, HIV ELISA tests
 - Pilcher et al. (2005), $I_k = 90$, HIV NAT tests
 - Kacena et al. (1998a, 1998b), $I_k = 10$, C/G NAT tests

Observed data likelihood

- For pool k , suppose we get to see the datum

$$(Z_k, \mathbf{x}'_{1k}, \mathbf{x}'_{2k}, \dots, \mathbf{x}'_{l_k k})'$$

- The **likelihood function** of $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ is

$$L(\boldsymbol{\beta}|\mathbf{z}) = \prod_{k=1}^K \left\{ \gamma_1 + \gamma_{12} \prod_{i=1}^{l_k} (1 - p_{ik}) \right\}^{z_k} \left\{ 1 - \gamma_1 - \gamma_{12} \prod_{i=1}^{l_k} (1 - p_{ik}) \right\}^{1-z_k},$$

where $\mathbf{Z} = (Z_1, Z_2, \dots, Z_K)'$ and $p_{ik} = h^{-1}(\mathbf{x}'_{ik}\boldsymbol{\beta})$

- $L(\boldsymbol{\beta}|\mathbf{z})$ can be maximized **numerically**
- Farrington (1992)** and **Vansteelandt et al. (2000)**

Regression literature

- **Xie (2001)**: Proposes a flexible EM algorithm approach
- **Bilder and Tebbs (2009)**: Investigates how pool compositions affect estimates
- **Chen et al. (2009)**: Proposes GOF techniques
- **Chen et al. (2009)**: Extends model to include random effects
- **Huang and Tebbs (2009)**: Allows mismeasured covariates
- **Delaigle and Meister (2011)**: Proposes a nonparametric modeling approach
- **McMahan et al. (2012)**: Allows γ_1 and γ_2 to depend on pool size and covariates

Classification problem

- Retest individuals in positive pools! [How?](#)
- **Non-informative** retesting procedures view the infection statuses as **iid** random variables with common prevalence p
- **Informative retesting**: Use covariate information to help decide how (and in what order) individuals are retested
 - Informative **Array** (*Biometrics*)
 - Informative **Dorfman** (*Biometrics*)
 - Informative **Halving** (*JRSS-C*)
 - Informative **Sterrett** (*JASA*)
- Dozens of interesting aspect to this problem!!
 - Informative “back-end” screening

Reference

Zhang, B., Bilder, C., and Tebbs, J. (2011+). Marginal regression models for multiple-disease group testing data. *Biometrics*, under review.

Multiple traits

- In most screening environments, individual specimens are tested for multiple infections—**simultaneously**
 - often, using the same assay test
- Examples
 - Infertility and Prevention Project (C/G)
 - American Red Cross (HIV, Hepatitis B/C)
 - German Red Cross (HIV, Hepatitis B/C)
- All available group testing regression methodology is for **one** infection (trait)
 - Joint models are needed

Multiple traits

- **Goal:** Extend group testing regression models to handle multiple infections

Infection	Individual				Response for
	1	2	...	l	Group k
1	\tilde{Y}_{11k}	\tilde{Y}_{21k}	...	\tilde{Y}_{l1k}	Z_{1k}
2	\tilde{Y}_{12k}	\tilde{Y}_{22k}	...	\tilde{Y}_{l2k}	Z_{2k}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
J	\tilde{Y}_{1Jk}	\tilde{Y}_{2Jk}	...	\tilde{Y}_{lJk}	Z_{Jk}

- Z_{jk} are **Bernoulli** random variables with mean

$$\theta_{jk} \equiv \text{pr}(Z_{jk} = 1) = \gamma_1^{(j)} + \gamma_{12}^{(j)} \prod_{i=1}^{l_k} (1 - p_{ijk})$$

Regression model

- **Model:**

$$p_{ijk} = \text{pr}(\tilde{Y}_{ijk} = 1) = h^{-1}(\beta_j' \mathbf{x}_{ik}),$$

for $i = 1, 2, \dots, I_k$, $j = 1, 2, \dots, J$, and $k = 1, 2, \dots, K$

- Is this possible using only group responses

$$\mathbf{z}_k = \begin{pmatrix} Z_{1k} \\ Z_{2k} \\ \vdots \\ Z_{Jk} \end{pmatrix} ??$$

- **YES!** The key is to relate $\text{cov}(Z_{jk}, Z_{j'k})$ to $\text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$
- We develop marginal regression models for **unobserved** correlated binary responses

Expected-Solution algorithm

- Elashoff and Ryan (2004) propose ES algorithm
 - a general technique to solve estimating equations in the presence of missing data
 - We view individual statuses \tilde{Y}_{ijk} as “missing”

GEE 101

- **Hypothetical:** Suppose we get to see the \tilde{Y}_{ijk} 's

- $\mathbf{R}(\boldsymbol{\alpha}) = J \times J$ working correlation matrix
- $\mathbf{V}_{ik} = \mathbf{B}_{ik}^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{B}_{ik}^{1/2}$, where

$$\mathbf{B}_{ik} = \text{diag}[p_{i1k}(1 - p_{i1k}), p_{i2k}(1 - p_{i2k}), \dots, p_{iJk}(1 - p_{iJk})]$$

- Note that $\text{cov}(\tilde{\mathbf{Y}}_{ik}) = \mathbf{V}_{ik}$ if $\mathbf{R}(\boldsymbol{\alpha})$ is correct
- Write out **estimating equations** in terms of individual responses

$$\boldsymbol{\Psi}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_k \sum_i \mathbf{D}'_{ik} \mathbf{V}_{ik}^{-1} (\tilde{\mathbf{y}}_{ik} - \mathbf{p}_{ik}) = \mathbf{0},$$

where $\mathbf{D}_{ik} = (\partial/\partial\boldsymbol{\beta})\mathbf{p}_{ik}$ and

$$p_{ijk} = \text{pr}(\tilde{Y}_{ijk} = 1) = h^{-1}(\boldsymbol{\beta}'_j \mathbf{x}_{ik})$$

Applying ES algorithm

- Analogous to EM, replace individual responses \tilde{y}_{ijk} with

$$\omega_{ijk} = E(\tilde{Y}_{ijk} | Z_{jk} = z_{jk})$$

- Closed form expressions are available
- Rewrite estimating equations as

$$\Psi^*(\beta, \alpha) = \sum_k \sum_i \mathbf{D}'_{ik} \mathbf{V}_{ik}^{-1} (\omega_{ik} - \mathbf{p}_{ik}) = \mathbf{0}$$

- We now turn our attention to estimating $\mathbf{R}(\alpha)$

Correlation estimation

- **Main result:** If group responses are conditionally independent,

$$\text{cov}(Z_{jk}, Z_{j'k}) = \gamma_{12}^{(j)} \gamma_{12}^{(j')} \prod_{i=1}^{I_k} \{\text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) g_1 + g_2\} + g_3,$$

where g_1 , g_2 , and g_3 are all functions of the p_{ijk} 's

- **Main point:** It is possible to estimate $\text{cov}(Z_{jk}, Z_{j'k})$ with the observed group responses
 - Can create a system of $S = \dim(\alpha)$ equations to estimate each component of $\mathbf{R}(\alpha)$
 - Easy in theory; not in practice
 - $\text{cov}(Z_{jk}, Z_{j'k})$ is a I_k degree polynomial of $\text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$
 - We find (very good) approximate solutions

ES algorithm

- 1 Select initial value $\beta^{(0)}$
- 2 **E-Step:** For a given $\beta^{(b)}$, $b = 0, 1, 2, \dots$, calculate each

$$\omega_{ijk} = E(\tilde{Y}_{ijk} | Z_{jk} = z_{jk}; \beta^{(b)}),$$

for $i = 1, 2, \dots, I_k$, $j = 1, 2, \dots, J$, and $k = 1, 2, \dots, K$

- 3 **S-Step:** Calculate $\hat{\alpha}(\beta^{(b)})$ and solve

$$\Psi^*[\beta, \hat{\alpha}(\beta^{(b)})] = \sum_k \sum_i \mathbf{D}'_{ik} \mathbf{V}_{ik}^{-1} (\omega_{ik}^{(b)} - \mathbf{p}_{ik}) = \mathbf{0}$$

for β to get the updated estimate $\beta^{(b+1)}$

- 4 Iterate between E and S until convergence

Simulation results: Summary

- ES solution $\hat{\beta}$ enjoys usual “nice” large-sample properties
- Can use methods of ER (2004) to estimate $\text{cov}(\hat{\beta})$
 - Wald inference available; valid for large K
- **Main findings:**
 - Estimates generally on target
 - $\text{cov}(\hat{\beta})$ estimated well
 - Wald confidence intervals operate at nominal level

Nebraska IPP

- 14,530 females screened for C/G in 2009
 - individual testing used on swab specimens
 - C: $\gamma_1^{(1)} = 0.928$ and $\gamma_2^{(1)} = 0.960$
 - G: $\gamma_1^{(2)} = 0.966$ and $\gamma_2^{(2)} = 0.980$
- We (artificially) assign individuals to pools of size $l_k = 5$ based on the specimen arrival date
- Covariates recorded:
 - age, race
 - information on sexual partners
 - PID, cervicitis
 - contact to STD, symptoms

Nebraska IPP

Effect	Infection	ES algorithm	Individual
Age	G	-0.03(0.02)	-0.04(0.01)
	C	-0.11(0.02)	-0.09(0.01)
Symptoms	G	1.09(0.38)	0.93(0.18)
	C	0.39(0.18)	0.29(0.08)
PID	G	0.28(0.96)	1.16(0.52)
	C	0.79(0.63)	0.40(0.39)
Cervicitis	G	0.29(0.35)	0.55(0.20)
	C	0.53(0.20)	0.59(0.11)
Multiple Partners	G	1.17(0.31)	1.05(0.17)
	C	0.28(0.22)	0.47(0.10)
STD Contact	G	1.38(0.29)	1.17(0.18)
	C	0.59(0.21)	0.93(0.10)

Table: Nebraska IPP data. Exchangeable correlation: $\hat{\alpha} = 0.27$.
 Estimated standard errors are in parentheses.

Nebraska IPP

- It is possible to test

$$H_0 : \beta_{r1} = \beta_{r2} \quad \text{versus} \quad H_1 : \beta_{r1} \neq \beta_{r2},$$

for the **rth** regression parameter

- In other words, we can test for a **common parameter** across infections
- This is not possible without a joint model
- The following covariates have large (Wald) p-values
 - 1 PID
 - 2 Cervicitis
- We also fit a smaller model where the parameter for each of these is **shared** across infections

Discussion

- We are currently working on multiple infection **informative screening** (classification) protocols
 - Complete identification
 - Purely negative identification
- We are also developing a general **Bayesian modeling** framework for group testing data to include
 - multiple traits
 - random effects (e.g., site effects)
 - covariate measurement error
 - information from retesting individuals in positive groups

THANK YOU!