Regression Analysis of Group Testing Samples

Minge Xie
Department of Statistics
Rutgers University
Piscataway, NJ 08854
Telephone: 732-445-2546
Fax: 732-445-3428
E-mail: mxie@stat.rutgers.edu

Short Title: Regression of Group Testing Samples

Dr. Xie’s research is partly supported by NSF grant DMS 9803273
Regression Analysis of Group Testing Samples

Summary

This paper develops a general regression methodology that relates the group testing responses to individual covariate information. It can be used to study samples from a group testing procedure and to deal with a wide range of regression problems. A detailed illustration of the methodology is provided for a group testing procedure proposed by Gastwirth and Hannick. To demonstrate the utility of the method, simulation studies are performed on an HIV antibody testing data set published by Nusscher et al.
1. Introduction

In a study of a large population with a small percentage of positive members, pooling samples of individuals together and testing them as a batch can considerably increase efficiency. This method, known as group testing, was first introduced by Dorfman\(^3\) as a cost effective procedure to detect syphilis in US army recruits during the second World War. Under Dorfman’s procedure, individual recruits’ blood samples were placed in pools and the pools of mixed blood samples were tested for syphilis antigen. Those individuals in pools free of syphilis antigen were declared free of syphilis. Individuals in the positive pools had their blood samples tested individually to identify the recruits with syphilis. Since then, the group testing method and its variations have been widely applied to blood screening, HIV and AIDS testing, drug use testing, chemical compound screening, among others. A large number of applications have been to identify the positive individuals, and many others use the group testing method to estimate the percentage of positive members in the population (the positive rate or prevalence) or other parameters of interest; see, for example, Chen and Swallow\(^4\), Gastwirth and Hammick\(^1\), Gastwirth and Johnson\(^5\), Hardwick, Page, and Stout\(^6\), among others. Much of the statistical focus has been to study the efficiency and to find optimal group testing schemes under various applications. Some recent publications include Yao and Hwang\(^7\), Hughes-Oliver and Swallow\(^8\), Phatarfod and Sudbury\(^9\), Litvak, Tu and Pagano\(^10\), Brookmeyer\(^11\), among others.

We develop in this paper a general regression method that is applicable to most samples from a group testing experiment. Often covariate information on individual members is available, or can be obtained in the experiment. So one can use a regression technique to analyze the relationship between the responses and the available covariate variables. For example, in blood testing cases, the persons who provide blood samples in the experiment are often asked to fill out a form to provide their age, gender, general health information, etc. (Nusbacher
et al.\textsuperscript{2} and Gunson and Rawlison\textsuperscript{12}). Also, in industrial application on chemical compound screening at the early stage of drug discovery, many descriptors of their molecular structures can be computed. It is of interest to find out whether such information may be related to the testing responses or not. So far, regression methodology has not been considered in the analysis of group testing samples. As the number of observations in applications of group testing is often very large, recent advances in computation technologies has made the study of such data sets by binary regression models possible.

Depending on the group testing scheme used in each application, the amount of available information from the responses varies. Sometimes it is possible that each individual member's response can be implied or known from the available group testing observations, particularly when the testing error can be ignored and the scheme is designed to identify all positive members. In such a case, binary regression methods are directly applicable to the individual outcomes. In many other situations, individual responses are not known and can not be fully determined from the available testing results. The methodology developed in this paper is especially targeting the latter case. The earlier situation can be regarded as a special case; see Section 2.

The rest of the paper is organized as follows. Section 2 presents a class of latent binary regression models for samples from a group testing experiment, and develops a general EM procedure to fit the models. Section 3 applies the methodology developed in Section 2 to a group testing scheme suggested by Gastwirth and Hammick\textsuperscript{1}, and simulation studies are performed on the Canadian blood donors individual anti-HIV screening data published by Nusbacher et al.\textsuperscript{2}. The analysis illustrates that a proper designed group testing procedure can save the cost substantially with only minor loss of accuracy. Section 4 provides further comments.

2. Regression Analysis of Group Testing Samples
The group testing method is often used when there is only a small proportion of positive individuals among a large population of size $N$. In a group testing experiment, samples from individual members are combined together into, say, $n$ pools and the entire pool is tested. Often, the individuals or the subsets of them for many pools (depending on the particular testing scheme adopted) will be selected and their samples will be further tested. If the testing scheme adopted is a sequential one, such combining and testing procedure will be repeated with successively smaller and smaller pools.

Without loss of generality, we label the large collection of $N$ individuals as $1, 2, \ldots, N$. For the $i$th individual, let $y_i$ be the indicator of whether it is positive or not. Suppose in total there are, say, $m$ tests (often, $m \geq n$) performed on $m$ sets of individuals, $g_1, g_2, \ldots, g_m$, where the sets correspond to the pools or the subsets of the pools formed in a group testing scheme. Conceivably a set could contain only one individual member so that a test is performed for this individual. Denote the $m$ testing results as $t_1, t_2, \ldots, t_m$, corresponding to sets $g_1, g_2, \ldots, g_m$ respectively. The testing results $t$'s are of binary type with 1 indicating positive and 0 otherwise. We write $T = \{t_1, t_2, \ldots, t_m\}$ and $G = \{g_1, g_2, \ldots, g_m\}$.

If the testing methods used in a screening process are perfectly accurate, then the testing result $t_j$ is determined by the individuals in the group $g_j$, assuming

$$ t_j = 1(\sum_{i \in g_j} y_i \geq 0), $$

where $1(C)$ is an indicator function equal to 1 if the statement $C$ is true, and equal to 0 otherwise.

However, the testing methods are usually not 100% accurate. Two concepts, sensitivity and specificity, are introduced in the literature to specify the accuracy of a testing method. Sensitivity, denoted as $\eta$, is referred to as the probability of a positive individual being tested positive. Specificity, denoted as $\theta$, is the probability of a negative sample being tested negative. See, for instance, Gastwirth\textsuperscript{13} or
Brookmeyer and Gail\textsuperscript{14}. We have $0 < \eta \leq 1$ and $0 < \theta \leq 1$. Under this situation, (1) becomes

$$t_j = W_j 1(\sum_{i \in g_j} y_{ik} > 0) + (1 - V_j) 1(\sum_{i \in g_j} y_{ik} = 0),$$

(2)

where $W_j$ and $V_j$ are independent Bernoulli random variables equal to 1 with probability $\eta$ and $\theta$ respectively.

When covariate information on individual subjects is available, we should relate the group testing outcome to it. Denote by $x_i$, a $p \times 1$ vector, as the $p$ covariate variables of the $i$th individual, for $i = 1, 2, \ldots, N$. We assume that the individual response follows a latent regression model,

$$P(y_i = 1) = f(x_i; \beta)$$

(3)

where $f$ is a known function and $\beta$ is the unknown parameter vector. In general, $f(x_i; \beta)$ can be a generalized linear regression model, a non-linear model, or even a non-parametric function. If $f(x_i; \beta) = H(x_i^T \beta)$ where $H$ is a known inverse link function, then (3) is a standard binary regression model in the framework of the generalized linear models (McCullagh and Nelder\textsuperscript{15}).

As mentioned in Introduction, sometimes it may be possible to determine all the $y_i$ values from the testing results $t_1, \ldots, t_m$. In this case, one can just use a standard software package to fit model (3) with $y_i$ values. However, in many other cases, the $y_i$ values cannot be determined from $t_j$'s, either due to imperfectly accurate testing method (for example, Gastwirth and Johnson\textsuperscript{5} and Litvak et al.\textsuperscript{10}) or because the goal of a group testing experiment is not to identify positive individuals (for example, Gastwirth and Hammick\textsuperscript{1} and Chen and Swallow\textsuperscript{4}) or due to violations of pooling assumptions (Xie et al.\textsuperscript{16}). In these cases, other regression techniques need to be developed.

For some testing schemes, explicit formula of the likelihood function for $t_1, \ldots, t_m$ may not be available, and different testing schemes lead to different forms of the
likelihood functions. Direct maximization of the likelihood function could be a
tedious task. We present next a general EM algorithm that works for a wide range
of samples obtained in group testing.

Under the latent regression model (3), the likelihood function of \( y_1, \ldots, y_N \) is
very simple,

\[
L(\beta | y_1, \ldots, y_N) = \prod_{i=1}^{N} f_i^{y_i} (1 - f_i)^{1-y_i}
\]

(4)

where \( f_i = f(x_i; \beta) \). We treat \( y_1, \ldots, y_N \) as the complete responses and \( t_1, \ldots, t_n \)
as the observed responses. From standard EM derivation procedures, we have
the following EM algorithm to obtain parameter estimates. This EM algorithm is
applicable to the situations when the individual outcomes can be determined from
available observations; only one EM iteration is needed in this case, .

Step 1. Select a starting point \( \beta^{(0)} \) of \( \beta \).

Step 2. (E-step) For a given \( \beta^{(k)} \), \( k = 0, 1, 2, \ldots \), calculate

\[
w_i^{(k)} = E(y_i | t_1, \ldots, t_m, \beta^{(k)}) = P(y_i = 1 | t_1, \ldots, t_m, \beta^{(k)}), \quad i = 1, \ldots, N.
\]

(5)

Step 3. (M-step) Given \( w_1^{(k)}, \ldots, w_N^{(k)} \) for a fixed \( k = 0, 1, 2, \ldots \), update the
parameter estimate at the \( (k+1) \)th iteration, \( \beta^{(k+1)} \), by maximizing the following
function of \( \beta \), where \( \ell(\beta | y_1, \ldots, y_N) = \log \{ L(\beta | y_1, \ldots, y_N) \} \),

\[
E \left\{ \ell(\beta | y_1^{(k)}, \ldots, y_N^{(k)}) | t_1, \ldots, t_m \right\} = \sum_{i=1}^{N} \left\{ w_i^{(k)} \log \left( \frac{f_i}{1 - f_i} \right) + \log(1 - f_i) \right\}.
\]

(6)

Step 4. Repeat steps 2 and 3 until \( \| \beta^{(k+1)} - \beta^{(k)} \| \) is very small; that is, until
the algorithm numerically converges.

In the E-step, the conditional expectation (5) can be reduced to the expecta-
tion of \( y_i \), conditional on \( \beta^{(k)} \) and only those \( t_j \)'s whose corresponding pool \( g_j \)
contains individual \( i \), i.e. \( \{ t_j | i \in g_j \} \). Although the formulation depends on the
particular group testing scheme adopted, in many situations \( w_i^{(k)} \) can be explicit-
ly formulated. An example is provided in the next section for the group testing
scheme suggested by Gastwirth and Hammick\(^1\). The maximization (6) in the M-
step (step 3) is exactly the maximization of the log-likelihood function of a binary
regression model (3) with true observations replaced by \(w_i^{(k)}\). Standard software
for binary regression can be used for this step.

The above EM algorithm does not provide the variance-covariance matrix
calculation for the parameter estimators. To obtain an estimator of the variance-
covariance matrix, we use the missing information principle and the Louis's method
(see, e.g. Tanner\(^1\), Sections 4.4.2, 4.4.3). In particular, the negative information
matrix is, from (4),

\[
H_n \overset{d}{=} -E \left\{ \frac{\partial^2}{\partial \beta^2} \ell(\beta | t_1, \ldots, t_m) \right\} = -E \left\{ \frac{\partial^2}{\partial \beta^2} \ell(\beta | y_1, \ldots, y_N | t_1, \ldots, t_m) \right\} - \text{Var} \left\{ \frac{\partial}{\partial \beta} \ell(\beta | y_1, \ldots, y_N | t_1, \ldots, t_m) \right\} = - \sum_{i=1}^N \left\{ w_i \frac{\partial^2}{\partial \beta^2} \left[ \log \left( \frac{f_i}{1 - f_i} \right) \right] + \frac{\partial^2}{\partial \beta^2} \left[ \log(1 - f_i) \right] \right\} - \left\{ \sum_{i=1}^N \sum_{i'=1}^N (w_{ii'} - w_i w_{i'}) C_{ii'} \right\},
\]

where \(w_i = E(y_i | t_1, \ldots, t_m)\), \(w_{ii'} = E(y_i y_{i'} | t_1, \ldots, t_m)\), and \(C_{ii'} = \frac{\partial}{\partial \beta} \left[ \log \left( \frac{f_i}{1 - f_i} \right) \right] \).
The variance covariance matrix of \(\hat{\beta}\) then can be estimated by \(\hat{H}_n^{-1}\), where \(\hat{H}_n\)
has the same form as \(H_n\) but with \(\beta\) replaced by \(\hat{\beta}\). See the next section for
an illustration on how the formula can be further simplified for the group testing
scheme suggested by Gastwirth and Hammick\(^1\).

The \(w_i\) and \(w_{ii'}\) formula in (5) and (7) might not be explicit in some cases.
For example, in the square array group testing scheme considered by Xie et al.\(^{16}\)
the formula for \(w_i\) does not have an explicit form. For such cases, one can use the
Gibbs sampling technique to obtain the \(w_i\)'s. This is because the fully conditional
distribution of each \(y_i\) given observations \(t_1, \ldots, t_m\) and \(\{y_s, \text{ for } s \neq i\}\) can always
be explicitly obtained from the model assumptions (1) and (3) or (2) and (3). Cycling
through the individuals by simulating Gibbs samples \(y_i^t\) from \(f(y_i | y_s, s \neq i, t_1, \ldots, t_m, \beta)\),
we can generate a Monte Carlo Markov chain from which the
joint distribution of \((y_1^*, \ldots, y_N^*)\) converges to \(f(y_1, \ldots, y_N|t_1, \ldots, t_m, \beta)\) after a large number of cycles. Repeating the process \(K\) times, we can get \(K\) sets of Gibbs samples \((y_1^*, \ldots, y_N^*)\). Each \(w_i\) can be estimated by \(\sum y_i^* / K\), where the summation is over the \(K\) sets of Gibbs samples. We do not need to calculate the \(w_{i,t}\)'s in this approach. The observed information matrix can be estimated by (Tanner\textsuperscript{18}, section 4.4.4.)

\[
H_n = -\frac{1}{K} \sum \frac{\partial^2}{\partial \beta^2} \ell(\beta | y_1^*, \ldots, y_N^*) - \left[ \frac{1}{K} \sum \left\{ \frac{\partial}{\partial \beta} \ell(\beta | y_1^*, \ldots, y_N^*) \right\}^2 - \left\{ \frac{1}{K} \sum \frac{\partial}{\partial \eta} \ell(\beta | y_1^*, \ldots, y_N^*) \right\}^2 \right]
\]

where the summations are over the set of \(K\) Gibbs samples \((y_1^*, \ldots, y_N^*)\) in the final round of the EM algorithm. This simulation based Gibbs sampling approach is applicable widely to almost all cases. However, from my experience this approach can be slow, considering the huge population size in the group testing problems and the large number of cycles needed in the Gibbs simulations. If explicit forms for \(w_i\) and \(w_{i,t}\) are available, we prefer to use them in the EM-algorithm and the variance-covariance computation.

3. Application to Gastwirth - Hammick Group Testing Method and Numerical Examples

In this section we study a group testing scheme proposed by Gastwirth and Hammick\textsuperscript{1} for estimating the prevalence of a rare disease. In Section 3.1, the estimating procedure developed in Section 2 is applied to the samples obtained from the Gastwirth - Hammick group testing scheme. In Section 3.2, we present numerical examples through simulations on the individual blood donor data published by Nusbacher et al.\textsuperscript{2}.

3.1. Regression Analysis of Samples from Gastwirth - Hammick Group Testing Method

9
Gastwirth and Hammick\(^1\) proposed a group testing scheme to “estimate the prevalence of a disease or trait possessing a possible stigma”, while at the same time the method preserves “individual anonymity”. Under the Gastwirth - Hammick group testing scheme, individual samples to be tested are batched into pools. Each pool is first given a screening test, then those classified as positives are given a confirmatory test. Typically the screening test is cheap but not extremely accurate while the confirmatory test is near perfect but substantially more expensive. Gastwirth and Hammick\(^1\) noted that in blood testing practice for screening HIV positives, the commonly performed screening test is the ELISA kit and the standard confirmatory test is the Western Blot (WB) analysis.

Assume, without loss of generality, there are total \(N = nk\) individuals and their samples are placed into \(n\) pools of size \(k\). According to the Gastwirth - Hammick group testing scheme, screening tests are applied to these \(n\) pools, and then for those pools tested positive confirmatory tests are applied. Suppose the binary type screening testing responses are denoted by \(t_{1}^{(s)}, t_{2}^{(s)}, \ldots, t_{n}^{(s)}\); they correspond to pools \(g_{1}, g_{2}, \ldots, g_{n}\) respectively. Let \(r = \sum_{j=1}^{n} t_{j}^{(s)}\). \(r\) is the number of the positive outcomes among the \(n\) screening test results \(\{t_{1}^{(s)}, \ldots, t_{n}^{(s)}\}\). Denote the positive outcomes of screening tests by \(t_{j_{1}}^{(s)}, t_{j_{2}}^{(s)}, \ldots, t_{j_{r}}^{(s)}\), and they correspond to the pools \(g_{j_{1}}, g_{j_{2}}, \ldots, g_{j_{r}}\). Write the \(r\) confirmatory testing results as \(t_{j_{1}}^{(c)}, \ldots, t_{j_{r}}^{(c)}\). They also correspond to pools \(g_{j_{1}}, g_{j_{2}}, \ldots, g_{j_{r}}\). In the notation of Section 2, we have testing outcomes \(T = \{t_{1}^{(s)}, \ldots, t_{n}^{(s)}, t_{j_{1}}^{(c)}, \ldots, t_{j_{r}}^{(c)}\}\) from pools \(G = \{g_{1}, \ldots, g_{n}, g_{j_{1}}, \ldots, g_{j_{r}}\}\) and \(m = n + r\).

We denote \((\eta^{(s)}, \theta^{(s)})\) and \((\eta^{(c)}, \theta^{(c)})\) as the sensitivity and specificity of the screening and confirmatory testing method respectively. For the screening tests, (2) is

\[
t_{j}^{(s)} = W_{j}^{(s)} 1(\sum_{i \in g_{j}} y_{i} > 0) + (1 - V_{j}^{(s)}) 1(\sum_{i \in g_{j}} y_{i} = 0)
\]
and for the confirmatory testing method,
\[
\ell_{j\ell}^{(c)} = W_{\ell}^{(c)}1_{(\sum_{i\in g_{j\ell}} y_i > 0)} + (1 - V_{\ell}^{(c)})1_{(\sum_{i\in g_{j\ell}} y_i = 0)}
\]
where \(W\)’s and \(V\)’s are the Bernoulli random variables.

If the covariate information on each individual is available, we can perform regression analysis. Denote \(x_i\) as the \(p \times 1\) vector of the covariate variables for the \(i\)th individual, and \(\beta\) is the unknown regression parameter. To simplify our discussion, we consider in the rest of paper the most commonly used binary regression model,
\[
\text{logit}\{P(y_i = 1)\} = x_i^T \beta,
\]
where \(\text{logit}\{t\} = \log(t/(1 - t))\) is the logit function.

Under (8), the log-likelihood for the latent individuals is
\[
\ell(\beta | y_1, \ldots, y_N) = \sum_{i=1}^{N} \left\{ y_i x_i^T \beta - \log(1 + e^{x_i^T \beta}) \right\}.
\]
The EM algorithm developed in Section 2 is directly applicable under the current setting. First, we notice that in the \(M\)-step we need to solve estimating equation
\[
\sum_{i=1}^{n} (w_{i} - p_{i})x_i = 0,
\]
where \(p_{i} = e^{x_i^T \beta} / (1 + e^{x_i^T \beta})\). This can be handled by any software that fits logistic regressions. Second, the \(w_{i}\)’s in the \(E\)-step can be evaluated explicitly. Suppose individual \(i\) belongs to the group \(g_{j}\). There are only three possible group testing responses: (i) \(t_{j}^{(s)} = 0\), (ii) \(t_{j}^{(s)} = 1, t_{j}^{(c)} = 0\), and (iii) \(t_{j}^{(s)} = t_{j}^{(c)} = 1\). Bayes’ formula and direct computation yield,
\[
w_{i} = \frac{(1 - \eta^{(s)})p_{i}}{\eta^{(s)}(1 - \eta^{(c)})p_{i}} 1(t_{j}^{(s)} = 0) + \frac{(1 - \eta^{(s)})(1 - \eta^{(c)})p_{i}}{\eta^{(s)}(1 - \eta^{(c)})(1 - \eta^{(c)})p_{i}} 1(t_{j}^{(s)} = 1, t_{j}^{(c)} = 0) + \frac{\eta^{(s)}(1 - \eta^{(c)})p_{i}}{\eta^{(s)}(1 - \eta^{(c)})p_{i}} 1(t_{j}^{(s)} = t_{j}^{(c)} = 1).
\]
One assumption, which was later adopted in Gastwirth and Hamnick\(^1\) and is also adopted later in Section 3 of this paper, is that the confirmatory test is perfectly accurate, i.e., \(\eta^{(c)} = \theta^{(c)} = 1\). In this case,

\[
    w_i = \frac{(1 - \eta^{(s)})p_i}{(1 - \eta^{(s)})[1 - \prod_{i' \in g_j} (1 - p_{i'})] + \theta^{(s)}[\prod_{i' \in g_j} (1 - p_{i'})]} \cdot (\delta_j^{(s)} = 0)
    + \frac{p_i}{1 - \prod_{i' \in g_j} (1 - p_{i'})} \cdot (\delta_j^{(s)} = \delta_j^{(c)} = 1).
\]

In a Gastwirth - Hamnick group testing method, the information matrix of the observations has an explicit form as well. Note that, in the current setting, \(f_i = p_i = e^{X_i^T \beta} / (1 + e^{X_i^T \beta})\), \(\frac{\partial^2}{\partial \beta^2} \log(\frac{f_i}{1 - f_i}) = 0\), \(\frac{\partial^2}{\partial \beta^2} \log(1 - f_i) = -\sum_{i=1}^N p_i (1 - p_i) x_i x_i^T\) and \(C_i = x_i\). Therefore, from (7),

\[
H_n = \sum_{i=1}^N p_i (1 - p_i) x_i x_i^T - \left\{ \sum_{j=1}^n \sum_{i \neq i', j \in g_j} (w_{i,i'} - w_{i} w_{i'}) x_i x_i^T \right\}.
\]

The only thing left is to calculate \(w_{i,i'}\)’s for \(i\) and \(i'\) in \(g_j\). Again by the Bayes formula, we have

\[
w_{i,i'} = P(y_i = 1, y_{i'} = 1 | t_j^{(s)}, t_j^{(s)} \text{ or } t_j^{(c)} \text{ and } t_j^{(s)} = 1) = \begin{cases} w_i, & \text{if } i = i' \text{;} \\ w_i p_{i'}, & \text{if } i \neq i'. \end{cases}
\]

Notice that the conditional distribution of \(t_j^{(s)}\) and \(t_j^{(c)}\) given \(y_i = 1\) and \(y_{i'} = 1\) is the same as the conditional distribution of \(t_j^{(s)}\) and \(t_j^{(c)}\) given \(y_i = 1\) only.

3.2. Simulation Studies on Canadian Blood Donor Individual AIDS Testing Data

In order to evaluate the effectiveness of informing prospective blood donors about HIV high-risk behavior patterns, Nusbacher et al.\(^2\) published a screening study which was carried out in the Center Ontario region in 1985-1986. In the study, all volunteer donors were given a questionnaire to fill out after registration, routine medical screening and hemoglobin determination. The questionnaire, which was completed in a private area, described behaviors associated with a high
AIDS risk. Each donor was given a self-exclusion option by checking a box to designate his or her blood to be “used for laboratory purposes only” if he (or she) feels he (or she) belongs to the high-risk groups, or to be “used for transfusion into patients” if he (or she) does not feel he (or she) belongs to the high-risk groups. At the same time, a control group was formed from the “transfusion purposes” group by matching for age, gender, and donor clinic site with donors who were in “laboratory purposes only” group. Units of blood from donors in the “laboratory purposes only” and control groups were tested for anti-HIV both by an enzyme-linked immunoassay (EIA) (Abbott Laboratories, Chicago, IL) and by the accurate but much more expensive Western blot (WB) analysis. The specimens of all other donors in the “transfusion purposes” group were first tested for anti-HIV by EIA, and then for those EIA positive donors their blood samples were tested by WB analysis.

Table 1 lists the testing results reported by Nusbacher et al.\textsuperscript{2}, who used only the WB results (i.e., values at the third and forth rows of Table 1) in their analysis. Note that, if the WB testing method is assumed to be perfectly accurate with 100% sensitivity and specificity (i.e., $\eta_{WB} = \theta_{WB} = 1$), then Table 1 implies that the specificity of the EIA testing method $\theta_{EIA}$ is about $\{(627 + 625) - (14 + 3)\}/\{(627 + 625) - (12 + 1)\} \approx .996$ and the sensitivity of the EIA testing method $\eta_{EIA}$ is about $\{11 + 1\}/\{(11 + 1) + 1\} \approx .923$; see also Gastwirth and Hammick\textsuperscript{1}. The “NA” in Table 1 is most likely to be $14/0.923 - 14 \approx 1.16$. Based on Table 1, it was concluded that the WB positive rates (odds) of any two of the three groups are significantly different (Nusbacher et al.\textsuperscript{2}).

We consider two simulation studies in this section under the setting of Nusbacher et al.’s\textsuperscript{2} study. In the first simulation study, we demonstrate that, with only minor loss of accuracy, one can save up to more than 90% of tests by using a group testing scheme modified from the method of Gastwirth and Hammick\textsuperscript{1}. The simulation study uses the data set in Table 1 where “NA” is replaced with 1. In
the second simulation study, we consider a regression model with two continuous
covariate variables and illustrate how one can handle the the group testing samples
using the method proposed in this paper. As in Gastwirth and Hammick\textsuperscript{1}, we will
adopt an assumption that the effect of dilution on the sensitivity and specificity
from batching \( k \) samples together is negligible. The pool size \( k \) is assumed to be
fixed at 10 as well.

Let's consider two grouping testing strategies in the first simulation study. The first strategy is a direct application of the Gastwirth - Hammick testing scheme
without considering the information on the different risk groups. In this case, the
total \( N = 94,496 + 627 + 625 = 95,748 \) individual blood samples are randomly
batched into 9,575 pools of size 10, and the mixed pool blood samples are tested
for anti-HIV by the EIA test; for those pools tested positive, confirmatory tests
are performed by the WB analysis method. The second group testing strategy
also uses the idea of Gastwirth and Hammick\textsuperscript{1}, however it forms pools within
each of the three risk groups: the “transfusion purposes”, “laboratory purposes
only” and control groups. Individuals from the same risk group are randomly
placed into pools of size 10. The total \( 94,496/10 = 9,450 \) pools of blood sam-
ple in the “transfusion purposes” group are screened by the EIA test; then the
EIA positive pools are tested again by the WB analysis for confirmation. For the
627/10 + 625/10 = 126 pools of blood samples in the “laboratory purposes only”
and controls groups, both the screening EIA test and the confirmatory WB analy-
sis, as in Nusbacher et al.’s experiment, are performed regardless of the screening
EIA testing results.

Based on the pooling assumption (2) and the two strategies described above,
we simulate two group testing data sets. Assume the individual testing information
is not available, we now analyze the two simulated group testing data sets. Note
that the simulated data sets from the second strategy can be analyzed by the
standard method (for example, Gastwirth and Hammick\textsuperscript{1}) if we separate the group
testing data set into three separated data sets, one for each risk group. However, in the group testing data set from the first strategy, the individuals from the different risk groups are mixed in the pools. We need to consider it as a regression problem. Denote \( \beta \) of (8) to be \((\beta_0, \beta_1, \beta_2)^T\), and \(x_i^T = (1, 0, 0), (1, 1, 0)\) and \((1, 0, 1)\) if the \(i\)th individual respectively belongs to the "transfusion", "laboratory purposes only" and control groups. The individual responses (unobserved) in the above simulated group testing data sets can be modeled by model (8). Under this parameterization, the odds of HIV positive for the three risk groups are \(e^{\beta_0}\), \(e^{\beta_0+\beta_1}\) and \(e^{\beta_0+\beta_2}\) respectively. We use the regression techniques developed in this paper to analyze the group testing data sets from the both strategies. Our EM algorithm was programmed in Splus (Math Software Inc.) and the numerical convergence criterion in the algorithm is \(|\beta_i^{(k+1)} - \beta_i^{(k)}| < \epsilon \times |\beta_i^{(k+1)}| + \epsilon\) for \(i = 0, 1, 2\), where the tolerance level \(\epsilon\) is 0.0001.

There are many different arrangements of positive units in the pools of group testing. In order to cover all possible arrangements that can occur with a good chance, we repeat our simulation a large number of times (1000 times). The side by side boxplots in Figure 1 summarizes the \(\beta\) estimates of the 1000 simulations under the two pooling strategies. The solid horizon line is the \(\beta\) estimates using all individual observations; the two dashed horizon lines are corresponding to the values of the estimates plus and minus 1.96 times their standard errors. We can see that most of the estimates are around the solid lines. But the estimates of \(\beta_2\) in the first strategy are very spread. In fact, a small portion of the estimates of \(\beta_2\) are extremely big negative numbers, they were not even printed in Figure 1. Further examination of these simulated data sets reveals that this is caused by mis-identifying the positive group that contains the sole positive unit in the control group. In this case, \(\beta_2\) is \(-\infty\) or \(e^{\beta_0+\beta_2} = 0\). Under the second strategy, we do not have such a problem and the algorithm converges for all 1000 simulations. Indeed, the estimates from the second pooling strategy are all very close to the
estimates using the individual testing results. From the model fitting results (not shown here), we reject the hypothesis $H_0: \beta_l = 0$ vs $H_1: \beta_l \neq 0$, for $l = 1, 2$, and $H_0: \beta_1 - \beta_2 = 0$ vs $H_1: \beta_1 - \beta_2 \neq 0$ in all 1000 simulations under the second strategy. Again we conclude that the positive rates (odds) of any two of the three groups are significantly different, just as Nusbacher et al.\textsuperscript{2} who used the individual responses did.

Table 2 compares the total number of tests required for the group testing methods with the number of tests done in the Nusbacher et al.'s\textsuperscript{2} experiment. The number of EIA tests needed is only one tenth of the number of EIA tests needed by Nusbacher et al\textsuperscript{2}. The expected numbers of times that a WB test is needed under the two strategies are 28 and 141 respectively. They are only 1.6\% and 8.5\% of the number required in the Nusbacher's experiment. The group testing method of the second strategy is an efficient alternative to the individual screening method.

The above simulation study only concerns a very simple (regression) setting with three groups (of samples). It would be more interesting if we can consider the regular regression problems with continuous covariates. Follow the suggestion of a referee, we next perform a simulation study to illustrate our regression method under the setting of Nusbacher et al.'s\textsuperscript{2} study. Suppose in the experiment by Nusbacher et al.\textsuperscript{2}, there are two continuous covariate variables $z = (z_1, z_2)^T$ measured at the individual level. In particular, we assume that the model for each individual is,

$$\logit\{P(y_i = 1)\} = x_i^T \beta + z_i^T \gamma$$  \hspace{1cm} (10)

where $x_i$ is the covariate of the risk groups (as defined before in this section) and $z_i = (z_{i1}, z_{i2})^T$ is a $2 \times 1$ continuous covariate vector. In our simulation study, $z_{i1}$ is simulated from $N(0.6, 1)$ if the $i$th individual is from the "laboratory purposes only" group and from $N(0.4, 1)$ otherwise. Also, we let $z_{i2} = \log(z_{i1}^*)$
where $z_{i1}$ follows an exponential distribution with mean equal to 0.4. We set \( \beta = (-11.5, 5.0, 3.0)^T \) and \( \gamma = (1.6, 2)^T \) and simulate individual responses from (10); we choose these parameter values so that the prevalence of each risk group will be comparable to those listed in Table 1. Based on the second strategy discussed above, we simulate pools and group testing responses. Assume that $x_i$’s and $z_i$’s are given. We are interested to see whether the estimates of the regression parameters based on the group testing responses (assuming the individual responses $y_i$’s are not available) are similar to those based on the individual responses. To cover potential cases that can occur with a significant probability, we repeat the study 250 times.

Table 3 lists model fitting results using the first five data sets in the simulation study. The parameter estimates using the individual testing responses are compared with those using the group testing responses. The numbers in the parenthesis are the standard errors of the corresponding estimates. It can be seen that in these cases there is little difference between the estimates obtained from the group testing samples and from the individual observations. Figure 3 uses side by side boxplots to compare the parameter estimates of all 250 simulated data sets. The solid horizon lines correspond to the true parameter values used in the simulations. In both cases, the shapes of the boxplots are similar and the estimates are all around the true parameter values, indicating the unbiasedness of the estimators. As expected, the estimates from the group testing responses are slightly variant than those from using individual observations. The group testing method provides similar conclusions to those obtained from testing all individuals with only a minor loss of accuracy.

4. Conclusion

In this paper, we developed a latent regression analysis method for samples from group testing. The methodology developed is quite general and can be applied to most of the group testing schemes. It is especially useful in situations
when single individual responses cannot be inferred from the observations. The
illustrative examples indicate that the group testing method provides an alterna-
tive to the conventional and much more expensive individual testing method. A
strength of our latent regression method is that it is very flexible and covers many
different regression models.

5. Acknowledgement

The author wishes to thank Dr. Kay Tatsuoka for his discussions at the
early-stage development of this paper. He is grateful to the reviewers for their
constructive suggestions.

References

1. Gastwirth JL, Hammick PA. Estimation of prevalence of a rare disease, pre-
serving anonymity of subjects by group testing: Application to estimating the

of a confidential method of excluding blood donors exposed to human immuno-

3. Dorfman, R. The detection of defective members of large populations. Annals
of Mathematical Sciences 1943; 14:436-440.

4. Chen CL and Swallow WH. Using group testing to estimate a proportion, and

5. Gastwirth JL, Johnson WO. Screening with cost-effective quality control: Poten-
tial applications to HIV and drug testing. Journal of the American Statistical
Association 1994; 89:972-981.

6. Hardwick J, Page C, and Stout Q. Sequentially deciding between two experi-
ments for estimating a common success probability. Journal of the American


Table 1
Anti-HIV testing among Canada blood donors

<table>
<thead>
<tr>
<th></th>
<th>Transfusion</th>
<th>Laboratory</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number tested</td>
<td>94,496</td>
<td>627</td>
<td>625*</td>
</tr>
<tr>
<td>EIA+</td>
<td>405</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>EIA+/WB+</td>
<td>14</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>EIA−/WB+</td>
<td>NA</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

* Appropriate controls were not available in two instances.

Table 2
Expected number of tests in group testing method
vs. the test numbers in Nusbacher et al.

<table>
<thead>
<tr>
<th></th>
<th>Nusbacher et al.</th>
<th>Strategy 1</th>
<th>Strategy 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIA test</td>
<td>95,748</td>
<td>9,575</td>
<td>9,576</td>
</tr>
<tr>
<td>WB analysis</td>
<td>1,657</td>
<td>28</td>
<td>141</td>
</tr>
</tbody>
</table>
Table 3
Comparison of Parameter Estimates in the First Five Simulated Data Sets
(Individual level responses vs. Group testing responses)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Responses Used in Regression Calculation</th>
<th>Parameter Estimates</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual level</td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\gamma_1$</td>
</tr>
<tr>
<td>1st</td>
<td>Individual level</td>
<td>-11.521</td>
<td>4.984</td>
<td>3.195</td>
<td>1.718</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.603)</td>
<td>(0.416)</td>
<td>(0.772)</td>
<td>(0.209)</td>
</tr>
<tr>
<td></td>
<td>Group testing</td>
<td>-11.558</td>
<td>4.945</td>
<td>3.141</td>
<td>1.746</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.527)</td>
<td>(0.406)</td>
<td>(0.762)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>2nd</td>
<td>Individual level</td>
<td>-12.351</td>
<td>4.795</td>
<td>3.869</td>
<td>1.976</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.743)</td>
<td>(0.493)</td>
<td>(0.738)</td>
<td>(0.251)</td>
</tr>
<tr>
<td></td>
<td>Group testing</td>
<td>-12.358</td>
<td>4.860</td>
<td>3.995</td>
<td>1.843</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.695)</td>
<td>(0.502)</td>
<td>(0.731)</td>
<td>(0.292)</td>
</tr>
<tr>
<td>3rd</td>
<td>Individual level</td>
<td>-11.908</td>
<td>5.352</td>
<td>2.493</td>
<td>1.767</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.582)</td>
<td>(0.451)</td>
<td>(1.098)</td>
<td>(0.258)</td>
</tr>
<tr>
<td></td>
<td>Group testing</td>
<td>-11.819</td>
<td>5.315</td>
<td>2.431</td>
<td>1.839</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.582)</td>
<td>(0.451)</td>
<td>(1.098)</td>
<td>(0.258)</td>
</tr>
<tr>
<td>4th</td>
<td>Individual level</td>
<td>-11.722</td>
<td>4.828</td>
<td>3.223</td>
<td>1.791</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.630)</td>
<td>(0.467)</td>
<td>(0.682)</td>
<td>(0.224)</td>
</tr>
<tr>
<td></td>
<td>Group testing</td>
<td>-11.594</td>
<td>4.917</td>
<td>3.213</td>
<td>1.705</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.558)</td>
<td>(0.440)</td>
<td>(0.685)</td>
<td>(0.249)</td>
</tr>
<tr>
<td>5th</td>
<td>Individual level</td>
<td>-10.974</td>
<td>5.475</td>
<td>2.805</td>
<td>1.473</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.553)</td>
<td>(0.411)</td>
<td>(1.062)</td>
<td>(0.221)</td>
</tr>
<tr>
<td></td>
<td>Group testing</td>
<td>-10.475</td>
<td>5.563</td>
<td>2.823</td>
<td>1.219</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.471)</td>
<td>(0.438)</td>
<td>(1.069)</td>
<td>(0.256)</td>
</tr>
</tbody>
</table>

1. The individual responses are simulated from model (10), and the group testing data sets were obtained by strategy 2 from the individual responses.
2. The values in the parenthesis are the standard errors of the corresponding estimates.