Group testing models with unknown link function

Dewei Wang
Karunarathna B. Kulasekera
Colin M. Gallagher
Christopher S. McMahan

Follow this and additional works at: https://tigerprints.clemson.edu/grads_symposium

Recommended Citation
Wang, Dewei; Kulasekera, Karunarathna B.; Gallagher, Colin M.; and McMahan, Christopher S., "Group testing models with unknown link function" (2013). Graduate Research and Discovery Symposium (GRADS). 79.
https://tigerprints.clemson.edu/grads_symposium/79

This Poster is brought to you for free and open access by the Research and Innovation Month at Tiger Prints. It has been accepted for inclusion in Graduate Research and Discovery Symposium (GRADS) by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.
Introduction

Let us consider the problem of screening a large number of individuals for an infectious disease. Traditionally, a specimen (e.g., blood, urine, plasma, etc.) is collected from each of the individuals and is subsequently tested for the presence of the infection:

Due to the large number of individuals, this process can be both expensive (with respect to testing cost) and time consuming.

Group testing, also known as pooled testing, was first proposed by Dorfman in 1943 as a method for reducing the cost associated with screening World War II soldiers for syphilis. In general, group testing involves testing pooled specimens formed from amalgamated specimens collected from individuals, rather than testing each specimen separately:

Information that we gain from this pooled testing process are:

If a group tests negative, then we may conclude that all contributing individuals are negative.

If a group tests positive, then we may conclude that at least one positive individual is in the pool.

When testing for low prevalence diseases, pooling specimens has become a common method of increasing screening efficiency. In practice, group testing strategies have been successfully applied in a variety of areas, including genetics, bioterrorism detection, and drug discovery.

Statistical research in group testing are branched into two major areas:

Classification: How to design an efficient decoding algorithm so that one can diagnose all individuals as either positive or negative with minimum cost. The basic idea here is to retest individuals’ specimen in positive pools.

Estimation: How to estimate individuals’ disease risks (the probability of being infected) by use pool response data only.

Methodology

The methodology we investigated could be summarized as in the following figure. We first estimate the dependence of the link function \( \phi \) on the coefficient \( \beta \) to obtain the final estimator of \( \beta \).

Numerical Analysis

We consider the following models:

1. \( \phi(x; \beta) = \exp(1 + x^2 \beta) / \exp(1 + x^2 \beta) \)
2. \( \phi(x; \beta) = \exp(x^2 \beta) / \exp(x^2 \beta) \)
3. \( \phi(x; \beta) = 1 / (x^2 + \beta) \)
4. \( \phi(x; \beta) = \exp(\log(2) + 500 x^2 \beta / \exp(\log(2) + 500 \beta) \)

The above data generating process was repeated 200 times, for each model and setting, and our methodology was applied to each. The following table summarizes the behavior of the 200 resulting estimates.

Hepatitis B data

We analyzed a real hepatitis B data set from NHANES 2009-2010. The is data set consists of 6533 individual observations. Each observation has six variables:

1. \( y \) is binary, indicating the presence (\( y = 1 \)) or absence (\( y = 0 \)) of the antibodies to the hepatitis B core antigen in the patient’s serum or plasma.
2. \( x_1 \) is age (continuous).
3. \( x_2 \) is the alanine aminotransferase level (continuous).
4. \( x_3 \) is the cholesterol level (continuous).
5. \( x_4 \) is the alanineaminotransferase level (continuous).
6. \( x_5 \) is ethnicity (discrete).

We considered group sizes \( n_j = n = 2, 5, 10 \). After randomly grouping individuals, we artificially generate the pooled testing response by

\[ Y_j = \mathbb{1}\{X_j \geq \tau \} \]

Then estimates are computed by our methodology. This procedure is repeated 200 times. The pattern of these 200 estimates are summarized in the following figure.

Conclusion

We have proposed a new method for modeling data collected from a group testing scheme which has becomes a standard procedure for screening a large number of individuals for infectious diseases. Numerical investigation and a real data analysis have demonstrated the performance of our estimators under practical settings. We also extend our method to cover the cases of imperfect testing and missing covariate information. If you are interested in this work, you are very welcome to contact me at dwang@clemson.edu.