Group Testing Model Estimation and Inference

Christopher R. Bilder

University of Nebraska Lincoln
Department of Statistics

Background

Contact information

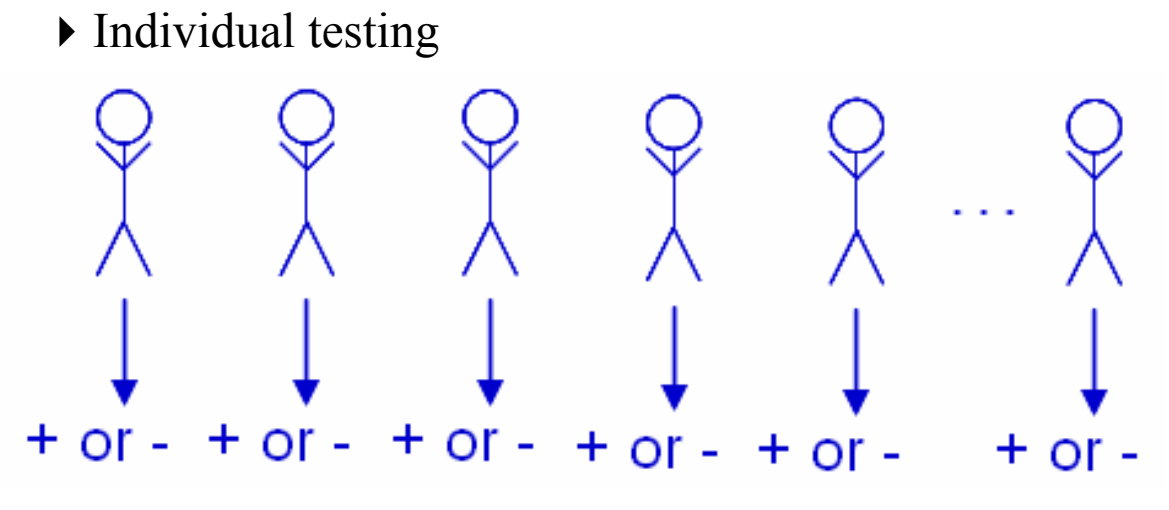
Christopher R. Bilder
Department of Statistics
University of Nebraska-Lincoln
chris@chrisbilder.com
www.chrisbilder.com

Abstract

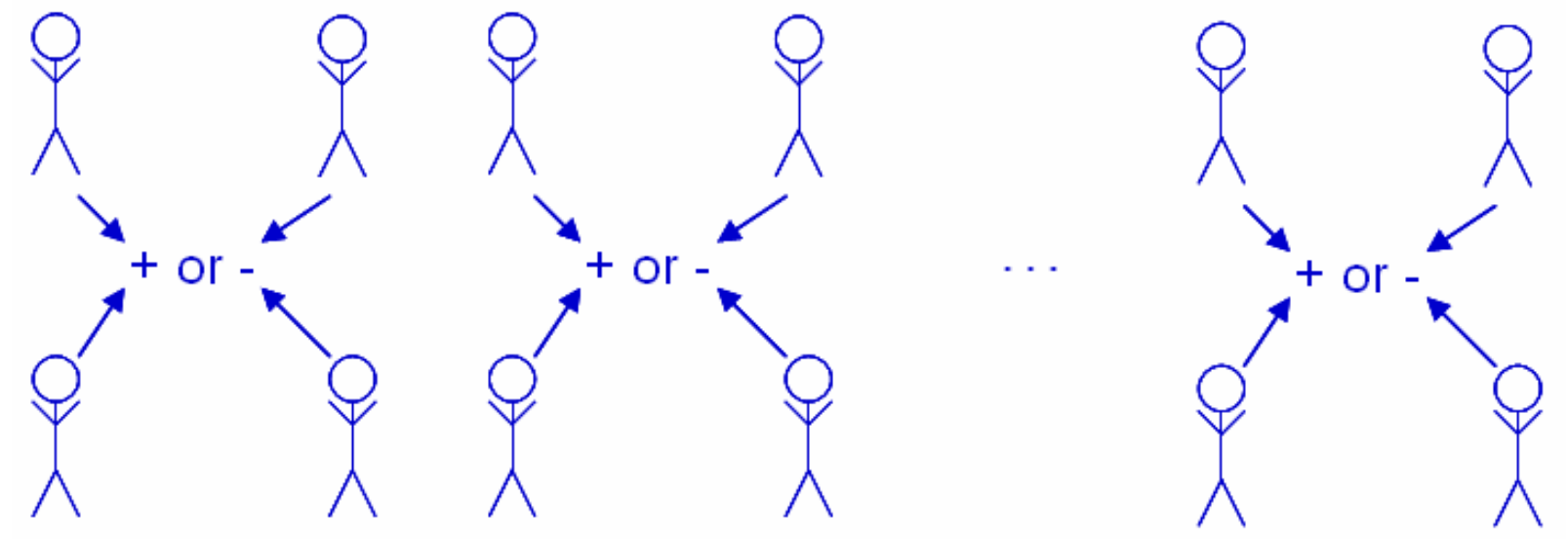
Group testing has long been used to estimate a trait prevalence, p, in situations where the prevalence is small in order to reduce time and cost or to make infeasible individual experiments feasible by grouping. Most of the statistical research in group testing has focused on estimating a single prevalence p for a homogenous population. Recently, Vansteelandt et al. (2000) and Xie (2001) have proposed models to incorporate covariates to estimate p for a heterogeneous population. The purpose here is to further examine these modeling methods through a set of comparisons between individual and group testing models. First, the relative efficiency of model parameter estimates is investigated under a number of grouping strategies. Second, agreement between model parameter estimates is examined to determine how well estimates coincide. Third, the effect of group size on model estimation is examined. Overall recommendations are given in order to show the benefits and sacrifices to using group testing models.

What is group testing?

- Used when testing an item for a trait
- Example: Testing blood for the presence or absence of a disease
  - Individual testing



    - Problems: Cost and time
  - Group testing



    - If the GROUP sample is negative, then all I people in the group do not have the disease
    - If the GROUP sample is positive, then at least ONE of the I people in the group have the disease
    - Cost and time savings!
    - Strategy works well when prevalence of the trait is small
- Many other examples of group testing
  - Disease transmission by an insect vector to a plant (Swallow, 1985)
  - Drug-discovery experiments (Xie et al. 2001; Zhu, Hughes-Oliver, and Young, 2001)

Purpose

- Until recently, no one had used covariates in a regression setting to help estimate the probability an individual item is positive for a trait
- Vansteelandt et al. (2000)
  - Use maximum likelihood estimation to estimate parameters for a model in the form of a generalized linear model
  - Estimation done directly on the group responses
  - Shows smallest variance estimators occur when covariates are most alike within a group
- Xie (2001)
  - Use maximum likelihood estimation to estimate parameters for a model in the form of a generalized linear model
  - Estimation done on the unobservable individual responses through using the EM algorithm
- Since maximum likelihood estimation is used for both, the Vansteelandt et al. (2000) fitting method will be used here only
- **Purpose:**
  - Compare individual and group testing models
  - Examine bias and efficiency of model parameter estimates
  - Assess agreement between model parameter estimates
  - Investigate the effect of group size
  - Analyze the effects of three grouping strategies

Notation

- Individual responses
  - $Y_{ik} = 1$ if the i^{th} item in the k^{th} group has the trait (positive)
    $Y_{ik} = 0$ otherwise (negative) for $i = 1, \ldots, I_k$ and $k = 1, \ldots, K$
  - $p_{ik} = P(Y_{ik} = 1)$
  - Y_{ik} are independent Bernoulli(p_{ik}) random variables
- Group responses
  - $Z_k = 1$ denotes a positive response and
    $Z_k = 0$ denotes a negative response for the k^{th} group
  - $\theta_k = P(Z_k = 1) = 1 - \prod_{i=1}^{I_k}(1 - p_{ik})$
  - Z_k are independent Bernoulli(θ_k) random variables
- Individual and group relationship
  - $Z_k = 1$ if and only if $\sum_{i=1}^{I_k} Y_{ik} > 0$
    $Z_k = 0$ if and only if $\sum_{i=1}^{I_k} Y_{ik} = 0$
  - Y_{ik}'s are "observed" when $Z_k = 0$ and there are no measurement errors; Y_{ik}'s are unobservable otherwise
- Model
  - $\mathbf{x}_{ik} = (x_{ik1}, x_{ik2}, \ldots, x_{ikp})'$ is a vector of covariates for the i^{th} subject in the k^{th} group
  - $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the corresponding vector of model parameters
  - $\log[p_{ik}/(1-p_{ik})] = \boldsymbol{\beta}'\mathbf{x}_{ik}$
  - Other link functions could be used as well

Estimation

- Simplifications for rest of presentation
  - One covariate, x_{ik}
  - No measurement errors (sensitivity = specificity = 1)
  - Equal group sizes ($I_1 = I_2 = \ldots = I_K = I$)
- Maximum likelihood estimation
  - Likelihood function: $L = \prod_{k=1}^{K} \theta_k^{z_k}(1-\theta_k)^{1-z_k} = \prod_{k=1}^{K}\left[1 - \prod_{i=1}^{I}(1-p_{ik})\right]^{z_k}\left[\prod_{i=1}^{I}(1-p_{ik})\right]^{1-z_k}$
  - Maximizing L with respect to $\boldsymbol{\beta}$ yields the maximum likelihood estimator, $\hat{\boldsymbol{\beta}}$
- Asymptotic variance of $\hat{\beta}_1$, $As\hat{V}ar(\hat{\beta}_1) =$

$$\frac{\sum_{k=1}^{K}\frac{1-\theta_k}{\theta_k}\left[\sum_{i=1}^{I}p_{ik}\right]^2}{\left(\sum_{k=1}^{K}\frac{1-\theta_k}{\theta_k}\left[\sum_{i=1}^{I}p_{ik}\right]^2\right)\left(\sum_{k=1}^{K}\frac{1-\theta_k}{\theta_k}\left[\sum_{i=1}^{I}p_{ik}x_{ik}\right]^2\right)-\left(\sum_{k=1}^{K}\frac{1-\theta_k}{\theta_k}\left[\sum_{i=1}^{I}p_{ik}\right]\left[\sum_{i=1}^{I}p_{ik}x_{ik}\right]\right)^2}$$

- For individual testing, the standard asymptotic variance for $\hat{\beta}_1$, $As\hat{V}ar(\hat{\beta}_1) =$

$$\frac{\sum_{k=1}^{K}\sum_{i=1}^{I}p_{ik}(1-p_{ik})}{\left[\sum_{k=1}^{K}\sum_{i=1}^{I}p_{ik}(1-p_{ik})\right]\left[\sum_{k=1}^{K}\sum_{i=1}^{I}p_{ik}(1-p_{ik})x_{ik}^2\right]-\left[\sum_{k=1}^{K}\sum_{i=1}^{I}p_{ik}(1-p_{ik})x_{ik}\right]^2}$$

Example

Setup

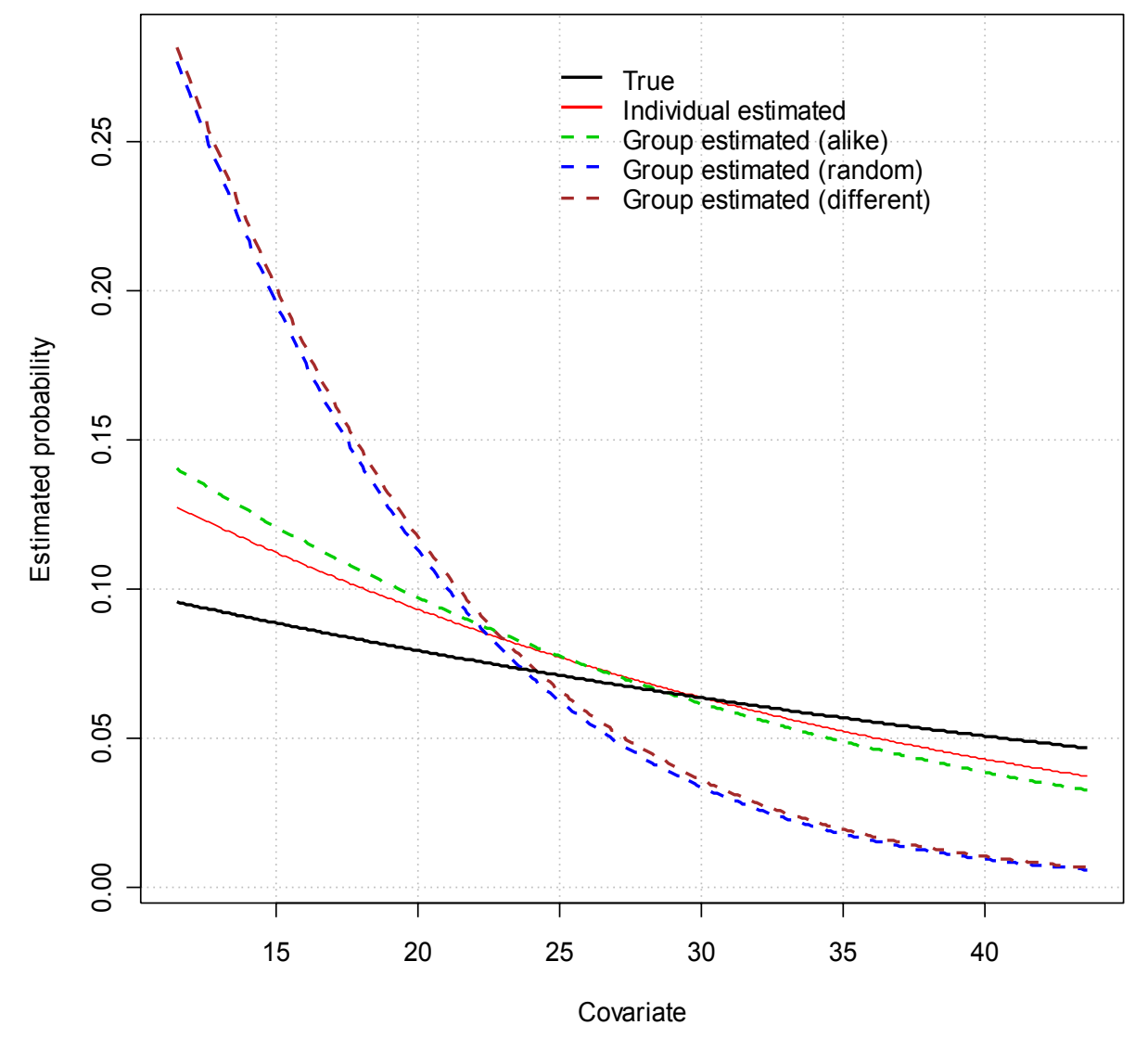
- Motivated from example in Vansteelandt et al. (2000)
  - Examines the covariate specific prevalence of HIV among pregnant women in an area of Kenya
  - One covariate of interest is age
- Model: $\log[p_{ik}/(1-p_{ik})] = \beta_0 + \beta_1 x_{ik}$
- Simulate data from model fitted to the individual observations in paper
  - β_0 = -1.97 and β_1 = -0.024
  - Generate x_{ik} from Gamma(20.95, 1.16) since it provides a good fit to the observed age distribution
  - $I = 7$ subjects per group
  - $K = 100$ groups
  - Overall sample size is $I{*}K = 700$
- Generate the Y_{ik} individual responses from Bernoulli distribution with parameter $p_{ik} = \exp(\beta_0 + \beta_1 x_{ik})/[1 + \exp(\beta_0 + \beta_1 x_{ik})]$
  - Groups are formed from these individual responses
  - Thus, both individual and group responses are available!
  - Example simulated data

k Group	i Item	Y_{ik} Individual response	Z_k Group response	x_{ik} Covariate
1	1	0	1	11.55
1	2	0	1	12.14
1	3	1	1	12.56
1	4	0	1	12.79
1	5	0	1	12.88
1	6	0	1	13.28
1	7	0	1	13.88
		⋮		
100	1	0	0	39.65
100	2	0	0	39.77
100	3	0	0	39.91
100	4	0	0	39.92
100	5	0	0	40.55
100	6	0	0	40.71
100	7	0	0	43.62

Grouping strategies

- Alike - Subjects with similar covariates are put into groups (sort by covariate, then assign to successive groups)
- Random - Subjects are randomly put into groups (emulates chronological if there is no response dependence over time)
- Different - Subjects with covariates as different as possible are put into groups (emulates worse case scenario)

One simulated data set

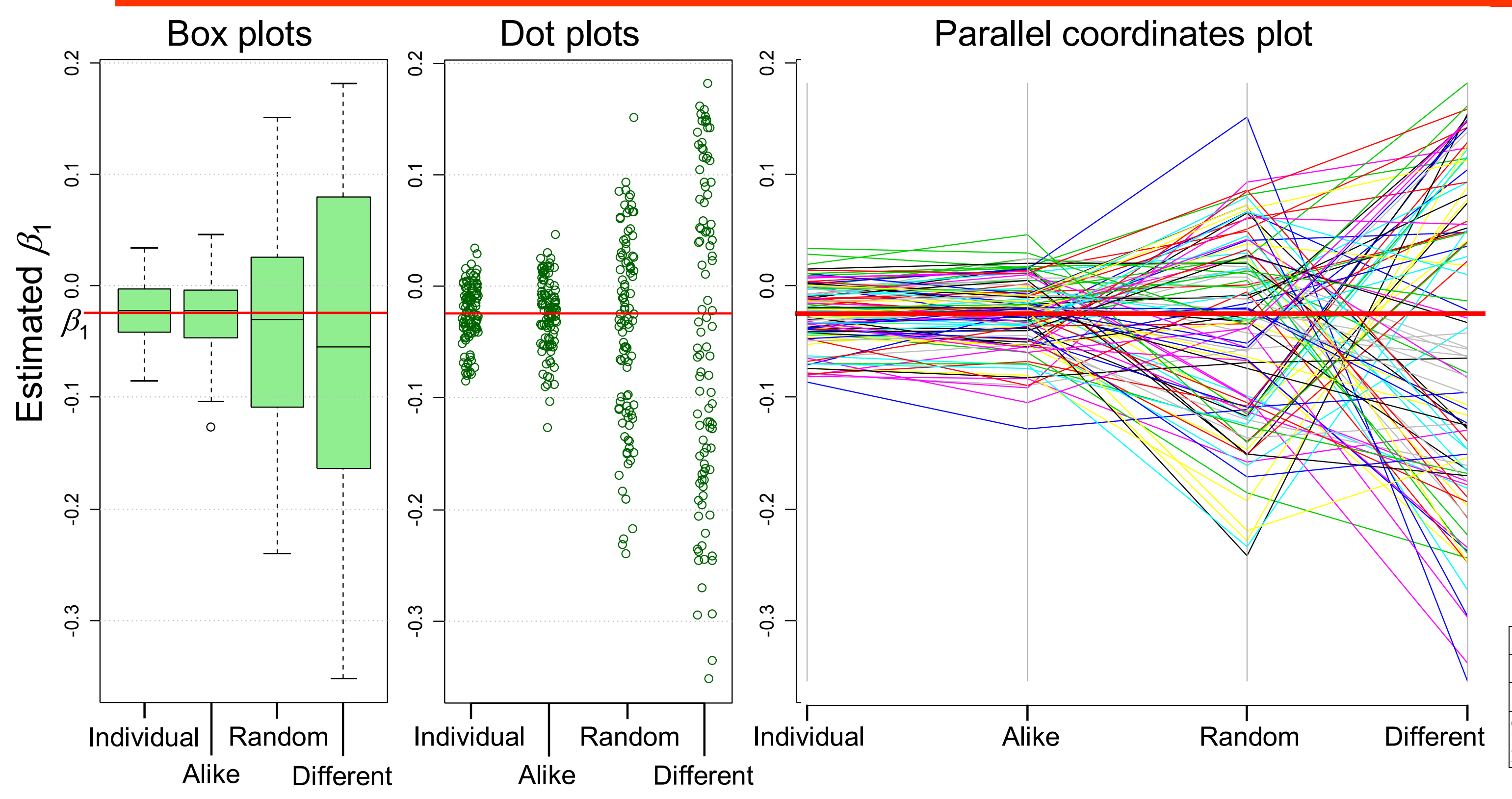


- Summary

	Individual	Grouping Alike	Random	Different
$\hat{\beta}_0$	-1.4452	-1.2423	0.5431	0.5357
$\hat{\beta}_1$	-0.0415	-0.0493	-0.1301	-0.1275
$As\hat{V}ar(\hat{\beta}_1)$	0.00079	0.00114	0.00676	0.02925
Relative efficiency		0.70	0.12	0.03

  - True values: β_0 = -1.97 and β_1 = -0.024
  - Relative efficiency = (Individual Var.) / (Group Var.)
  - Remember that 7 times more tests are done using individual testing!

100 simulated data sets



- Notice that the largest $\hat{\beta}_1$ for Random corresponds to the smallest for Different!
- Pearson correlation between $\hat{\beta}_1$ values on the same data sets

		Individual	Grouping Alike	Random
	Individual			
Grouping	Alike	0.85		
	Random	0.33	0.24	
	Different	-0.05	-0.09	-0.13

- Summary of $\hat{\beta}_1$ values with β_1 = -0.024
  - Alike grouping strategy results in only a little more variability compared to individual testing
  - Random and different grouping strategies result in much more variability compared to individual testing

	Individual	Grouping Alike	Random	Different
Mean	-0.0247	-0.0253	-0.0391	-0.0472
Median	-0.0217	-0.0224	-0.0298	-0.0550
Variance	0.0007	0.0010	0.0071	0.0197
95% C.I. for mean	(-0.0301, -0.0193)	(-0.0316, -0.0189)	(-0.0558, -0.0223)	(-0.0750, -0.0194)

Simulations

Settings

- Model
  - β_0 = -2 and β_1 = 0.6931 for $\log[p_{ik}/(1-p_{ik})] = \beta_0 + \beta_1 x_{ik}$
  - x_{ik} sampled from Uniform(-7.079, 1.663)
  - Thus, $0.001 < p_{ik} < 0.3$
  - Average value of p_{ik} is 0.02
- $b = 1, \ldots, 500$ simulated data sets for each setting of I and K
- R's *glm*() function used to fit model to individual responses
- R's *optim*() function used to fit models to group responses
- Additional simulations for different β_0, β_1, I, K, and x_{ik} distribution settings were performed with similar results

Fixed sample size ($I{*}K$) comparisons

- Percent bias = $\left[\left(\sum_{b=1}^{500}\hat{\beta}_{1,b}/500\right) - \beta_1\right]/\beta_1 * 100\%$

I*K = 200

		I 1	2	5	10	20
Grouping	Alike		10.8%	20.3%	30.7%	9.1%
	Random		9.8%	14.0%	34.1%	69.2%
	Different		7.9%	9.6%	20.6%	121.3%
	Individual	9.9%				

For example, Alike is biased by 10.8% when 100 groups of size 2 are formed for $I{*}K$=200

I*K = 500

		1	2	5	10	20
Grouping	Alike		3.5%	5.5%	13.9%	29.7%
	Random		3.6%	2.8%	8.3%	38.9%
	Different		2.8%	7.8%	20.4%	42.4%
	Individual	3.4%				

I*K = 1000

		1	2	5	10	20
Grouping	Alike		1.4%	2.5%	5.8%	15.7%
	Random		1.5%	3.5%	5.5%	17.6%
	Different		0.7%	3.0%	9.3%	37.7%
	Individual	1.1%				

- Relative efficiency = $(1/500)\sum_{b=1}^{500} As\hat{V}ar(\hat{\beta}_{1,b}^{Individual}) \Big/ As\hat{V}ar(\hat{\beta}_{1,b}^{Group})$
- Pearson correlation between $\hat{\beta}_1^{Individual}$ and $\hat{\beta}_1^{Group}$

I*K = 200

		Relative efficiency I 2	5	10	20	Correlation I 2	5	10	20
Grouping	Alike	0.87	0.76	0.57	0.27	0.97	0.79	0.56	0.30
	Random	0.64	0.29	0.12	0.04	0.85	0.56	0.32	0.15
	Different	0.46	0.08	0.02	0.01	0.71	0.30	0.09	0.08

I*K = 500

		2	5	10	20	2	5	10	20
Grouping	Alike	0.93	0.78	0.59	0.33	0.97	0.89	0.65	0.47
	Random	0.68	0.30	0.12	0.04	0.85	0.60	0.35	0.21
	Different	0.49	0.08	0.02	0.00	0.73	0.37	0.12	0.00

I*K = 1000

		2	5	10	20	2	5	10	20
Grouping	Alike	0.94	0.79	0.59	0.33	0.97	0.88	0.72	0.46
	Random	0.69	0.31	0.13	0.04	0.82	0.53	0.36	0.17
	Different	0.50	0.09	0.02	0.00	0.73	0.33	0.16	0.09

Fixed number of tests (K) comparisons

- Percent bias = $\left[\left(\sum_{b=1}^{500}\hat{\beta}_{1,b}/500\right) - \beta_1\right]/\beta_1 * 100\%$

K = 100

		I 1	2	5	10	20	30	40
Grouping	Alike		10.8%	5.5%	5.8%	6.5%	5.7%	6.5%
	Random		9.8%	2.8%	5.5%	6.3%	12.0%	22.0%
	Different		7.9%	7.8%	9.3%	11.9%	26.0%	102.9%
	Individual	29.4%						

For example, Alike is biased by 10.8% when 100 groups of size 2 are formed for $I{*}K$=200

K = 200

		1	2	5	10
Grouping	Alike		5.1%	2.5%	2.2%
	Random		4.2%	3.5%	4.4%
	Different		4.0%	3.0%	4.4%
	Individual	9.9%			

K = 500

		1	2	5	10
Grouping	Alike		1.4%	1.0%	0.4%
	Random		1.5%	1.3%	0.3%
	Different		0.7%	4.0%	3.6%
	Individual	3.4%			

- Relative efficiency = $\left[\sum_{b=1}^{500} As\hat{V}ar(\hat{\beta}_{1,b}^{Individual})\right]\Big/\left[\sum_{b=1}^{500} As\hat{V}ar(\hat{\beta}_{1,b}^{Group})\right]$

K = 100

		I 2	5	10	20	30	40
Grouping	Alike	8.04	17.10	25.23	27.61	25.89	24.47
	Random	5.84	6.45	5.39	3.55	2.40	1.63
	Different	4.19	1.83	0.81	0.30	0.15	0.08

K = 200

		2	5	10
Grouping	Alike	2.79	5.87	8.58
	Random	2.02	2.26	1.89
	Different	1.46	0.64	0.28

K = 500

		2	5	10
Grouping	Alike	2.20	4.62	6.72
	Random	1.61	1.79	1.50
	Different	1.16	0.51	0.22

Conclusions

- $\hat{\beta}_1$ is biased for finite samples
  - Bias increases with group size for fixed $I{*}K$ here
  - Bias is smaller for group testing than individual testing with K fixed
- Relative efficiency
  - For the same $I{*}K$, individual testing is more efficient
    - Remember that less tests are done with group testing!
  - When K is fixed, group testing is more efficient (except for Different)
  - Alike is the most efficient of the grouping methods
- Pearson correlation between individual and grouping methods
  - Correlation decreases as group size increases
  - Depending on the group size, Random and Different grouping can produce quite different $\hat{\beta}_1$ values than found for individual testing!
- Which is the more fair comparison - fixed $I{*}K$ or fixed K?
  - If tests are expensive and individual items are cheap to obtain, fixed K is better to compare
  - If individual items are expensive to obtain, fixed $I{*}K$ is better to compare
- Is the Alike grouping strategy realistic?
  - Only if ALL individual samples are available at once since groups are formed by covariate
    - Example: All samples are available at the same time in Thorburn et al. (2001) when assessing hepatitis prevalence in Glasgow, Scotland
    - More than one covariate makes Alike grouping more difficult
  - Often, Alike is not realistic due to limited "shelf-life" for item samples
  - As a compromise, some individual items could be constructed in homogenous groups by covariates as the samples are received
- How should group size(s) be chosen?
  - Vansteelandt et al. (2000) suggests one way if all individual samples are available at once
  - Without this information, group size should be chosen based upon the possible range of θ_k by avoiding values close to 0 or 1
- Convergence of parameter estimates
  - Complete separation problems - this happens most often with Alike due to how the groups are formed
  - Low trait prevalence means small number of $Y_{ik} = 1$ for individual testing
    - This is a contributing factor to its large bias for smaller $I{*}K$