Human or Cylon?

Group testing on the Battlestar Galactica

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

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

Battlestar Galactica

- Statistics and Battlestar Galactica
- The story so far…
  - Video

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

Battlestar Galactica

- Statistics and Battlestar Galactica
- The story so far…
  - Video
- Cylons
  - Centurion
  - Humanoid form (new)
- How can you distinguish a human from a Cylon?

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

Battlestar Galactica

- Dr. Gaius Baltar
  - Asked to develop a Cylon detector
    - Season 1's *Bastille Day* episode
  - # of Cylons in fleet is expected to be small
  - 47,905 individuals to test!

Battlestar Galactica

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Dr. Gaius Baltar (continued)
  - Season 1's *Tigh me up and Tigh me down*

  - Video

Battlestar Galactica

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

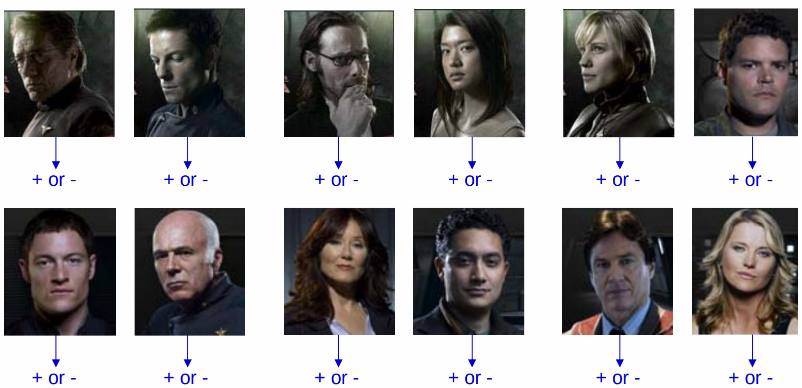
- Dr. Gaius Baltar (continued)
  - Season 1's *Tigh me up and Tigh me down*

  - Video
  - (47,905 blood tests)$*$(11 hours each) = 21,956 days

Battlestar Galactica

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Individual testing



- Problems:
  - Time
  - Limited resources

Battlestar Galactica

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Group testing

- If a GROUP is negative, then all 4 individuals are not Cylons
- If the GROUP is positive, then at least ONE of the 4 individuals is a Cylon
  - "Retesting" can be done to determine who is a Cylon

Battlestar Galactica

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Group testing (continued)
  - Time savings
  - Save resources
  - Strategy works well when prevalence of a "trait" is small
    - If prevalence is large, all groups may test positive

Other examples

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Screening blood donations
  - American Red Cross uses groups of size 16
  - HIV, Hepatitis B, Hepatitis C, …
  - Screen about 6 million a year
    - Source: Roger Dodd, Executive Director of Blood Services R & D at ARC
    - See Dodd et al. (*Transfusion*, 2002)
- Drug discovery experiments
  - Screen hundreds of thousands of chemical compounds to look for potentially good ones
  - Remlinger et al. (*Technometrics*, 2006)

Other examples

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Multiple vector transfer design experiments
  - Estimate probability an insect vector transfers a pathogen to a plant
  - Swallow (*Phytopathology*, 1985, 1987)

- Veterinary
  - Bovine viral diarrhea in cattle (Peck, *Beef*, 2006)
  - Avian pneumovirus (APV) in turkeys (Maherchandani et al., *J. Veterinary Diagnostic Investigation*, 2004)
- Public health studies
  - Prevalence of HCV (Liu et al., *Transfusion*, 1997)
  - Prevalence of HIV (Verstraeten et al., *Trop. Med. & International Health*, 2000)

Notation

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Individual responses
  - Y_{ik} = 1 if the i^{th} item in the k^{th} group has the "trait" (positive) and
    Y_{ik} = 0 otherwise (negative) for i=1, …, I and k=1, …, K
  - Y_{ik} are independent Bernoulli(p) random variables
    - $p = P(Y_{ik} = 1)$
    - Homogenous population
    - p can be thought of as the "individual probability" or "prevalence in a population"
  - Y_{ik}'s are not directly observed (at least initially)

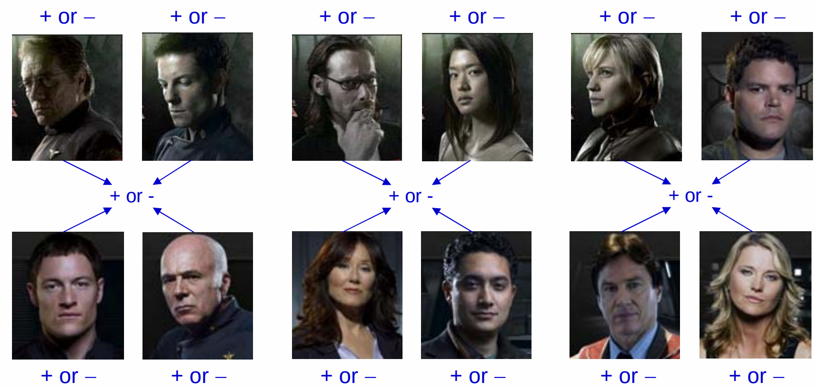
Notation

- Group responses
  - $Z_k = 1$ denotes a positive response
    $Z_k = 0$ denotes a negative response for the k^{th} group
  - Z_k are independent Bernoulli(θ) random variables
    - $\theta = P(Z_k = 1)$
- Individual and group response relationship
  - $Z_k = 1$ if and only if $\sum_{i=1}^{I} Y_{ik} > 0$
    $Z_k = 0$ if and only if $\sum_{i=1}^{I} Y_{ik} = 0$

Notation

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

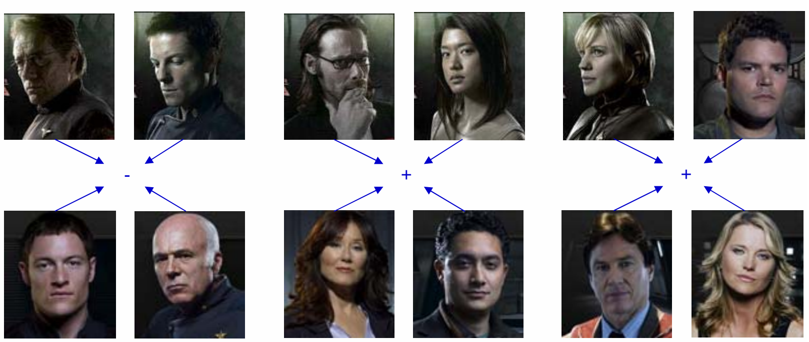
- Example random variables



Notation

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

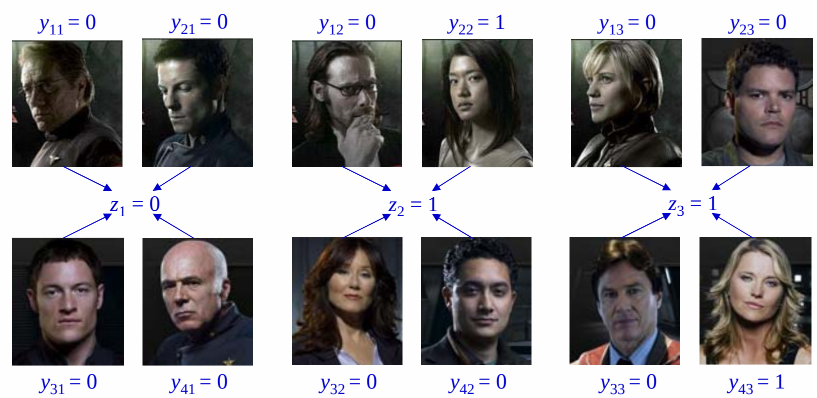
- Example observed values



Notation

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Example observed values



Purpose

- Prevalence of a trait in a population (estimation problem)
- Which items are positive (identification problem)

Estimate *p*

- How can we estimate $p = P(Y_{ik} = 1)$?
  - We observe information about the groups, not individuals!
  - $\theta = 1 - P(Y_{ik} = 0, \forall i) = 1 - (1 - p)^I$
  - Then $p = 1 - (1 - \theta)^{1/I}$
  - MLE for p: $\hat{p} = 1 - \left(1 - \sum_{k=1}^{K} z_k / K\right)^{1/I}$
- Unequal group sizes
  - Likelihood function

$$L(p) = \prod_{k=1}^{K} \theta_k^{z_k}(1-\theta_k)^{1-z_k} = \prod_{k=1}^{K}\left[1-(1-p)^{I_k}\right]^{z_k}(1-p)^{I_k(1-z_k)}$$

where

θ_k = positive probability for group k

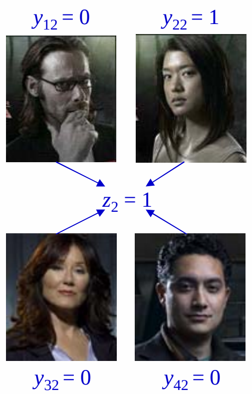
I_k = size of group k

Testing error

- What if there is testing error?
  - Can incorporate sensitivity (η) and specificity (δ)
  - $\theta_k = \eta + (1 - \delta - \eta)(1 - p)^{I_k}$

Identification

- Dorfman (*Annals of Mathematical Statistics*, 1943)
  - Retest all items in a positive group
  - Often credited for the very first use of group testing
- Sterrett (*Annals of Mathematical Statistics*, 1957)
  - Individually retest until first positive is found
  - Re-group remaining items
    - If group is negative, STOP
    - If group is positive, repeat
  - Expected number retests is smaller than Dorfman
- Gupta and Malina (*Statistics in Medicine*, 1999) provides a summary



Infertility Prevention Program

- U.S. national program funded by Centers for Disease Control and Prevention
  - Assess and reduce prevalence of chlamydia and gonorrhea
- Nebraska
  - Swab or urine specimens are sent to the Nebraska Public Health Laboratory at U. of Nebraska Medical Center
  - NATs
  - About 30,000 individual tests done per year
- Group testing!

Infertility Prevention Program

- Lindan et al. (*J. Clinical Microbiology*, 2005)
  - Estimates that 12% of the laboratories in the U.S. are already using group testing
  - Group testing has allowed "laboratories to achieve a significant increase in specimen loads."
- Quarter #1 of 2006, chlamydia testing
  - Urine specimens – 1,384 total
  - Ignore sensitivity and specificity here
  - Individual data: $\hat{p}$ = 111/1,384 = 0.0802
  - Group testing:
    - Randomly put known individual responses into groups of size I = 2
    - $\hat{p} = 1-(1-\sum_{k=1}^{K} z_k / K)^{1/I} = 1-(1-105/692)^{1/2} = 0.0790$

Infertility Prevention Program

- Quarter #1 of 2006 (continued)
  - Individual data: 111/1,384 = 0.0802
  - Group testing:

	Group size				
	1	2	3	5	10
$\hat{p}$	0.0802	0.0790	0.0791	0.0776	0.0843
Dorfman		902	765	737	949
Sterrett		902	728	653	744

  - Approximate cost per test
    - $16 for urine
    - $11 for swab

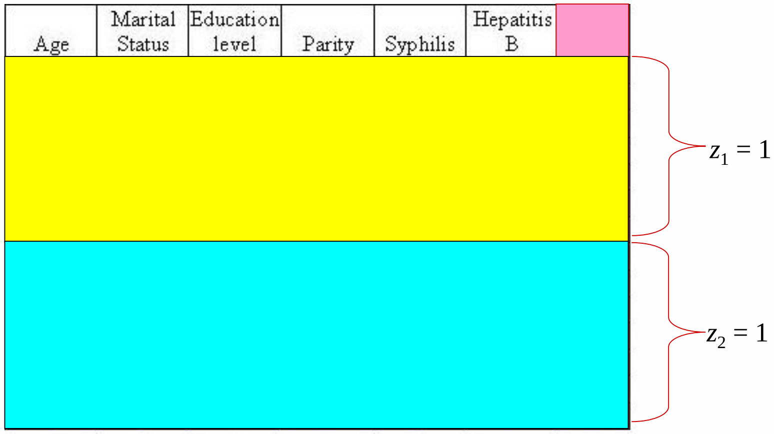
Heterogonous populations

- Individual responses
  - Y_{ik} are independent Bernoulli(p_{ik}) random variables
  - $p_{ik} = P(Y_{ik} = 1)$ for item i in group k
- Group responses
  - Z_k are independent Bernoulli(θ_k) random variables
  - $\theta_k = P(Z_k = 1)$ for group k
- Covariates
  - $x_{ik1}, x_{ik2}, \ldots, x_{ikp}$ for the i^{th} item in the k^{th} group
  - Incorporate factors which influence trait status
  - Not really done until recently in group testing!

Kenyan pregnant women study

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Part of the data from Vansteelandt et al. (*Biometrics*, 2000)



Heterogonous populations

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Model
  - logit(p_{ik}) = $\beta_0 + \beta_1 x_{ik1} + \ldots + \beta_p x_{ikp}$
- Estimation of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$
  - Note that Y_{ik} are not directly observed
  - Vansteelandt et al. (*Biometrics*, 2000)
    - Likelihood function is written in terms of the Z_k

$$L = \prod_{k=1}^{K} \theta_k^{z_k} (1-\theta_k)^{1-z_k}$$
$$= \prod_{k=1}^{K} \left[1 - \prod_{i=1}^{I_k} (1-p_{ik})\right]^{z_k} \left[\prod_{i=1}^{I_k} (1-p_{ik})\right]^{1-z_k}$$

  - Xie (*Statistics in Medicine*, 2001)
    - Likelihood function is written in terms of the Y_{ik}
    - EM algorithm used

Forming groups

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Alike
  - Individuals with "similar" covariates are put into pools
  - Smallest variability in parameter estimates
  - How implement?
    - One covariate: Sort by covariate, then assign successive individuals to pools
    - Multiple covariates: ?
  - Usually requires one to have all individual testing specimens up front and available for testing at the same time

Forming groups

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Random
  - Individuals are assigned to pools at random
  - Emulates chronological if no dependence over time
- Different
  - Pool individuals with covariates as different as possible
  - Emulates "worse case scenario" (?)

Forming groups

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Simulate data from model fitted to the individual observations in Vansteelandt et al. (*Biometrics*, 2000)
  - $\text{logit}(p_{ik}) = \beta_0 + \beta_1 x_{ik} = -1.97 - 0.024x_{ik}$
  - Simulate the individual and group responses
    - $I = 7$ subjects per group
    - $K = 100$ groups
    - Overall sample size is $I*K = 700$

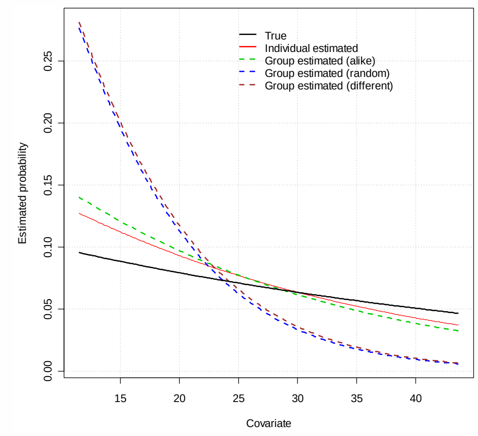
Forming groups

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- One simulated data set
- Relative efficiency

$$\widehat{RE}(\hat{\beta}_1) = \frac{\widehat{AsVar}(\hat{\beta}_1^{Individual})}{\widehat{AsVar}(\hat{\beta}_1^{Group})}$$

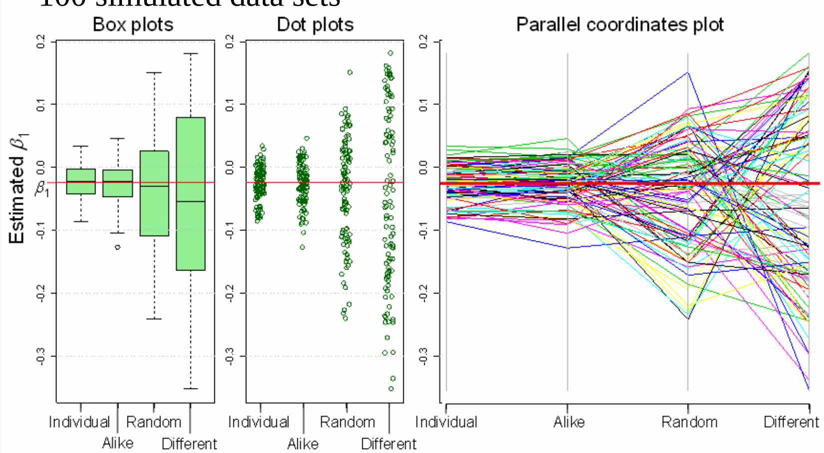
	$\widehat{RE}(\hat{\beta}_1)$
Alike	0.71
Random	0.12
Different	0.03



Forming groups

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- 100 simulated data sets



- Pearson correlations:

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

Forming groups

Outline
- BSG
- Basics
- Estimation
- Identification
- Covariates
- NIH grant

- Last slide examined a fixed $I*K$
- What if we fix the number of groups (tests), K, instead?
  - Settings
    - $\text{logit}(p_{ik}) = -2 + 0.6931x_{ik}$
    - $x_{ik} \sim \text{Uniform}(-70.079, 1.663)$
    - $0.001 < p_{ik} < 0.3$
    - Average value of p_{ik} is 0.02
    - 500 simulated data sets for each simulation
  - Relative efficiency:

	I		
$K = 500$	2	5	10
Alike	2.20	4.62	6.72
Random	1.61	1.79	1.50
Different	1.16	0.51	0.22

NIH Grant

- Content removed

NIH Grant

- Content removed

NIH Grant

- Content removed

NIH Grant

- Content removed

Human or Cylon?

Group testing on the Battlestar Galactica

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