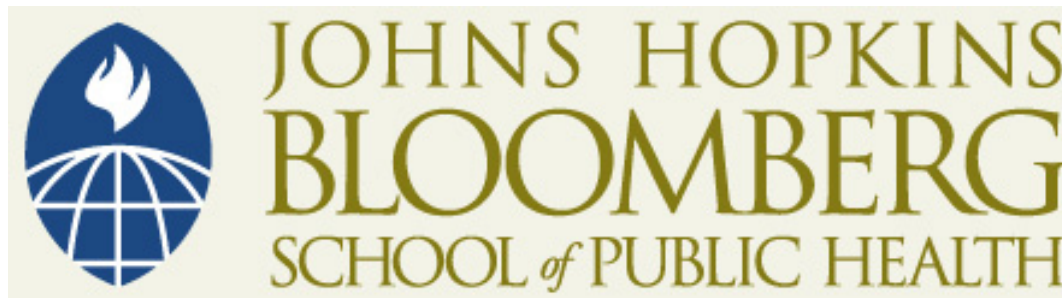


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2007, The Johns Hopkins University and Qian-Li Xue. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

# Item Regression: Multivariate Regression Models

Qian-Li Xue PhD

Assistant Professor of Medicine, Biostatistics,  
Epidemiology

Johns Hopkins Medical Institutions

Statistics for Psychosocial Research II: Structural  
Models

# General Idea

- Y's are all measuring the same thing or similar things.
- Want to summarize the association between an X and all of the Y's.
- BUT! We are not making the STRONG assumption that there is latent variable accounting for the correlation between the Y's.
- First: Make model that allows each  $Y_i$  to be associated with X
- Next: Summarize/Marginalize over associations
- Sort of like ATS
  - But wait! I thought ATS was “bad” relative to SAA!
  - Not if you don't want to make the assumption of a latent variable!
  - More later.....



## Example: Vision Impairment in the Elderly

- Salisbury Eye Evaluation (SEE, West et al. 1997).
  - Community dwelling elderly population
  - N = 1643 individuals who drive at night
- Want to examine which aspects of vision (X's) (e.g. visual acuity, contrast sensitivity) affect **performance of activities** that require seeing at a distance (Y's).

# Variables of Interest

- Y's: Difficulty....
  - reading signs at night
  - reading signs during day
  - seeing steps in dim light
  - seeing steps in day light
  - watching TV
- X's:
  - “Psychophysical” vision measures
    - visual acuity
    - contrast sensitivity
    - glare sensitivity
    - stereopsis (depth perception)
    - central vision field
  - Potential confounders
    - age
    - sex
    - race
    - education
    - MMSE
    - GHQ
    - # of reported comorbidities

# Technically.....

- The Y's are binary, and we are using logistic regression.
- To simplify notation, I refer to the outcomes as “Y” but in theory, they are “logit(Y).”

- Assume  $N$  individuals,  $k$  outcomes ( $Y$ 's),  $p$  predictors ( $X$ 's).
- For individual  $i$ :

$$Y_{i1} = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \dots + \beta_{1p}x_{ip}$$

$$Y_{i2} = \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2} + \dots + \beta_{2p}x_{ip}$$

⋮

$$Y_{ik} = \beta_{k0} + \beta_{k1}x_{i1} + \beta_{k2}x_{i2} + \dots + \beta_{kp}x_{ip}$$

- What is the same and what is different across equations here?
- We are fitting  $k$  regressions and estimating  $k^*(p+1)$  coefficients

# Good or Bad approach?

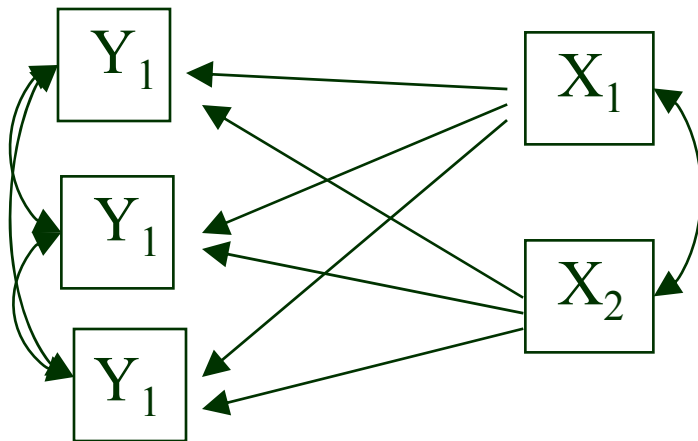
- Not accounting for correlations between Y's from the same individual:
  - e.g. may see that  $X \rightarrow Y_1$ , but really  $X \rightarrow Y_2$  and  $Y_1$  is correlated with  $Y_2$ .
- Simply: not summarizing!
- Alternative: Fit one “grand” model.
  - Can decide if same coefficient is appropriate across Y's or not.
  - Accounting for correlation among responses within individuals.



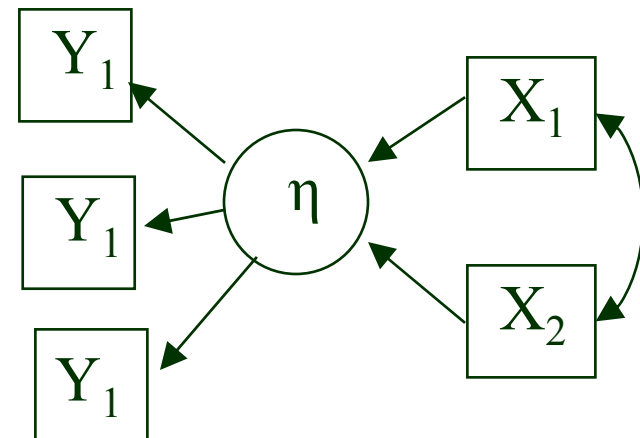
# Analyze THEN Summarize, OR Analyze AND Summarize?

- Includes all of the outcomes (Y's) in the same model
- But, there is not an explicit assumption of a latent variable (LV).
- Includes correlation among outcomes
  - Do not assume that Y's are indep. given a latent variable
  - Avoid LV approach and allows Y's to be directly correlated

“Multivariate” Model



Latent Variable Approach



# Why Multivariate Approach?

- Latent variable approach makes stronger assumptions
- Assumes underlying construct for which Y's are “symptoms”
- Multivariate model is more exploratory
- Based on findings from MV model, we may adopt latent variable approach.

## Data Setup for Individuals 1 and 2

	item (Y)	ID	Visual Acuity	Age
Person 1	y11	1	x11	x12
	y12	1	x11	x12
	y13	1	x11	x12
	y14	1	x11	x12
	y15	1	x11	x12
Person 2	y21	2	x21	x22
	y22	2	x21	x22
	y23	2	x21	x22
	y24	2	x21	x22
	y25	2	x21	x22

We have a “block” for each individual instead of a “row” like we are used to seeing. Stack the “blocks” together to get the whole dataset.

What if we entered this in standard logistic regression model?

# Model Interpretation

$$Y_{ij} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i1} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i2} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i3} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i4} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i5} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

# Additional Parameters....

Dummy variables for different Y

item (Y)	ID	Visual Acuity	Age	I(item=2)	I(item=3)	I(item=4)	I(item=5)
y11	1	x11	x12	0	0	0	0
y12	1	x11	x12	1	0	0	0
y13	1	x11	x12	0	1	0	0
y14	1	x11	x12	0	0	1	0
y15	1	x11	x12	0	0	0	1
y21	2	x21	x22	0	0	0	0
y22	2	x21	x22	1	0	0	0
y23	2	x21	x22	0	1	0	0
y24	2	x21	x22	0	0	1	0
y25	2	x21	x22	0	0	0	1

Now what does regression model look like?

What are the interpretations of the coefficients?

# Model Interpretation

$$Y_{ij} = \beta_0 + \beta_1 va_i + \beta_2 age_i + \alpha_2 I(j = 2) + \\ \alpha_3 I(j = 3) + \alpha_4 I(j = 4) + \alpha_5 I(j = 5)$$

$$Y_{i1} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i2} = \beta_0 + \alpha_2 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i3} = \beta_0 + \alpha_3 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i4} = \beta_0 + \alpha_4 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i5} = \beta_0 + \alpha_5 + \beta_1 va_i + \beta_2 age_i$$

# Parameter Interpretation

- $\beta_0$  = intercept (i.e. log odds) for item 1
- $\alpha_2$  = difference between intercept for item 1 and for item 2.
- $\beta_0 + \alpha_2$  = intercept for item 2
- $\beta_1$  = expected difference in risk of difficulty in any item for a one unit change in visual acuity (i.e.  $\exp(\beta_1)$  is log odds ratio).
- Intuitively, how does this model differ than previous one (i.e. one without  $\alpha$  terms)?
  - Each item has its own intercept
  - Accounts for differences in prevalences among outcome items
  - Still assumes that age and visual acuity all have same association with outcomes.

Is that enough parameters?

What if the association between visual acuity is  
NOT the same  
for reading signs at night and for watching TV?



# Is that enough parameters?

Interaction between VA and I(Y)

item (Y)	Visual Acuity	Age	I2	I3	I4	I5	va*I2	va*I3	va*I4	va*I5
y11	x11	x12	0	0	0	0	0	0	0	0
y12	x11	x12	1	0	0	0	x11	0	0	0
y13	x11	x12	0	1	0	0	0	x11	0	0
y14	x11	x12	0	0	1	0	0	0	x11	0
y15	x11	x12	0	0	0	1	0	0	0	x11
y21	x21	x22	0	0	0	0	0	0	0	0
y22	x21	x22	1	0	0	0	x21	0	0	0
y23	x21	x22	0	1	0	0	0	x21	0	0
y24	x21	x22	0	0	1	0	0	0	x21	0
y25	x21	x22	0	0	0	1	0	0	0	x21

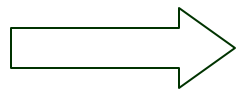
NOW how are regression parameters interpreted?

Note: I2 = I(item=2); va = Visual Acuity

# Model Interpretation

$$Y_{ij} = \beta_0 + \beta_1 va_i + \beta_2 age_i + \sum_{k=2}^5 \alpha_k I(j = k) +$$

$$\sum_{k=2}^5 \delta_k I(j = k) \times va_i$$



$$Y_{i1} = \beta_0 + \beta_1 va_i + \beta_2 age_i$$

$$Y_{i2} = \beta_0 + \alpha_2 + (\beta_1 + \delta_2) va_i + \beta_2 age_i$$

$$Y_{i3} = \beta_0 + \alpha_3 + (\beta_1 + \delta_3) va_i + \beta_2 age_i$$

$$Y_{i4} = \beta_0 + \alpha_4 + (\beta_1 + \delta_4) va_i + \beta_2 age_i$$

$$Y_{i5} = \beta_0 + \alpha_5 + (\beta_1 + \delta_5) va_i + \beta_2 age_i$$

# Parameter Interpretation

- $\beta_0$  = intercept for item 1
- $\alpha_2$  = difference between intercept for item 1 and for item 2.
- $\beta_1$  = expected change in risk in item 1 for a one unit change in visual acuity.
- $\delta_2$  = difference between expected change in risk in item 2 for a unit change in visual acuity and expected change in risk in item 1.
- $\beta_1 + \delta_2$  = expected difference in risk in item 2 for a one unit change in visual acuity.

# Parameter Interpretation

- $\beta_1 + \delta_2 =$  expected difference in risk in item 2 for a one unit change in visual acuity.
- The  $\delta$  terms allow for the association between visual acuity and each of the outcomes to be different.
- We can test whether or not all the  $\delta$  terms are equal to zero or not.
- If they are equal to zero, that implies.....



# Logistic Regression: Vision example

Covariate	Estimate	Robust SE	Model SE	Robust Z
Intercept ( $\beta_0$ )	----	-----	----	-----
Visual acuity ( $\beta_1$ )	-4.10	0.28	0.27	-14.7
Age ( $\beta_2$ )	-0.03	0.008	0.008	-3.5
I2 ( $\alpha_2$ )	-1.47	0.06	0.06	-24.5
I3( $\alpha_3$ )	0.74	0.12	0.13	6.0
I4( $\alpha_4$ )	-0.21	0.07	0.07	-3.1
I5( $\alpha_5$ )	0.85	0.18	0.17	4.7
I2*va ( $\delta_2$ )	0.66	0.21	0.27	3.2
I3*va ( $\delta_3$ )	2.25	0.32	0.29	7.1
I4*va ( $\delta_4$ )	2.10	0.31	0.27	6.8
I5*va ( $\delta_5$ )	0.59	0.30	0.28	2.0

So far...same logistic and linear regression type stuff.  
The difference:

- We need to deal with the associations!
- **Items from the same individual are NOT independent**
- Vision example: Odds Ratio between items is 7.69! We can't ignore that!
- We incorporate an “association” model into the model we already have (the “mean” model).
- Consider an adjustment:
  - mean model: used for inference
  - association model: adjustment so that test statistics are valid.

# Accounting for Correlations Within Individuals

- “Marginal Models”
  - parameters are the same as if you analyzed separately for each item, but measures of precision are more appropriate
  - describes population average relationship between responses and covariates as opposed to subject-specific.
  - We average (or marginalize) over the items in our case.

# Fitting Approach #1

## Post-hoc adjustment

Idea: Ignoring violation of independence invalidates standard errors, but not the slope coefficients.

So: We fit the model “näively” and then adjust the standard errors to correctly account for the association afterwards.

Problem with this? Its outdated! We have better ways of dealing with this presently.



# Related Example: Drinks per Week

Suppose  $Y_i$ ,  $i = 1, \dots, N$  are independent but each is sample mean of  $n_i$  responses with equal variances,  $\sigma^2$ . (e.g. drinks per week, averaged over 2 weeks).

Results from “usual” SLR, where  $y$  is drinks per week and  $x$  family support.

$$\widehat{se}(\beta_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}}$$

But, it is true (due to the averaging of  $y$ ) that the actual s.e. is

$$\widehat{se}(\beta_1) = \sqrt{\frac{\sigma^2 \sum_{i=1}^N [(x_i - \bar{x})^2 / n_i]}{\left[ \sum_{i=1}^N (x_i - \bar{x})^2 \right]^2}}$$

This is a valid analysis: We first fit the SLR and then correct the standard error of the slope.

# Fitting Approach #2

- Marginal Model (GEE or ML)
  - approach #1 is okay, but not as good as simultaneously estimating the mean model and the association model (i.e. we can iterate between the two, and update estimates each time).
  - We estimate regression coefficients using a procedure that accounts for lack of independence, and specifically the correlation structure that you specify.
  - Correlation structure is estimated as part of the model.

## Related Example Revisited: Drinks per Week

- If  $Y_1$  is based on 2 observations (i.e. 2 weeks), and  $Y_2$  is based on 20 observations (i.e. 20 weeks), we want to account for that.
- We want to “weight” individuals with more observations more heavily because they have more “precision” in their estimate of  $Y$ .
- Results: Weight is proportional  $\sqrt{n_i}$  .
- Resulting regression is better by accounting for this in the estimation procedure.



## Fitting Approach #2 (continued)

- Here we use the within unit correlation to compute the weights.
- GEE solution: “working correlation”
- If specified structure is good, the regression coefficients are very good.
- If specified structure is bad, coefficients and standard errors are still valid, but not as good.
- **ROBUST PROCEDURE**

# Fitting this for the Vision example

Approach 1: too complex to be feasible. Need to know all of the associations and adjust many estimates.

Approach 2: account for correlation in estimation procedure

In STATA:

## Logistic model:

```
xtgee y va age i2 i3 i4 i5 va2 va3 va4 va5, i(id)  
      link(logit) corr(exchangeable) robust
```

```
xi: xtgee y i.item*va age, i(id) link(logit) robust  
(default corr is exc)
```

## Linear model:

```
xtgee y x, i(id) corr(exchangeable) robust
```

# Problem with Approach #1

- Often correlation structure is more complex (our example was very simple compared to most situations)
- Post-hoc adjustments won't always work because estimating the correlation structure is not as simple.
- In general, people don't use approach #1 especially because many stats packages can handle the adjustments currently (Stata, Splus, R, SAS)

# How do I know the correlation structure?

- You don't usually.
- Approaches commonly used for multivariate outcome
  - Exchangeable:
    - ❖ individuals items are all equally correlated with each other.
    - ❖ Simple and intuitive, easy to estimate and describe.
    - ❖ Could be a bad assumption
  - Unstructured:
    - ❖ uses empirical estimates from data.
    - ❖ Less prone to model mis-specification
    - ❖ less powerful approach.

# Summarizing Findings

- (1) Constrain equal slopes across items
- (2) Constrain slopes that should be constrained, and allow others to vary
- (3) Detailed summary discussion that covers everything
- (4) Complicated: joint tests/CI's for groups of items





## Multiple Regression Results:

Odds Ratio between items estimated to be 8.69

Vision Variable	Item	Item O.R.	95% CI for OR
Visual acuity	1	0.427	(0.36,0.51)
	2	0.515	(0.43,0.62)
	3	0.863	(0.71,1.06)
	4	0.817	(0.68,0.99)
	5	0.514	(0.43,0.62)
Best contrast sensitivity	1-5	1.477	(1.26,1.73)
Diff in contrast sensitivity	1-5	0.696	(0.58,0.84)
Log(steropsis)	1-5	0.904	(0.86,0.95)
Best central vision field	1-5	0.902	(0.83,0.98)

1 = day signs; 2 = night signs; 3 = day steps; 4 = dim steps; 5 = TV

West SK, Munoz B, Rubin GS, Schein OD, Bandeen-Roche K, Zeger S, German S, Fried LP. Function and visual impairment in a population-based study of older adults. The SEE project. Salisbury Eye Evaluation. Invest Ophthalmol Vis Sci. 1997 Jan;38(1):72-82.

# Closing Remarks

- Model specification is still important here!
  - Mean model
  - Correlation structure
    - ❖ GEE, random effects
- Get robust estimates if possible (GEE)
- Fitting methods:
  - Stata: `xtgee`
  - WinBugs hierarchical model