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
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 $i$:

  \[
  Y_{i1} = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \ldots + \beta_{1p}x_{ip}
  \]

  \[
  Y_{i2} = \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2} + \ldots + \beta_{2p}x_{ip}
  \]

  \[
  \vdots \]

  \[
  Y_{ik} = \beta_{k0} + \beta_{k1}x_{i1} + \beta_{k2}x_{i2} + \ldots + \beta_{kp}x_{ip}
  \]

- 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

<table>
<thead>
<tr>
<th>Item (Y)</th>
<th>ID</th>
<th>Visual Acuity</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>y11</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
</tr>
<tr>
<td>y12</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
</tr>
<tr>
<td>y13</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
</tr>
<tr>
<td>y14</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
</tr>
<tr>
<td>y15</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
</tr>
<tr>
<td>y21</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
</tr>
<tr>
<td>y22</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
</tr>
<tr>
<td>y23</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
</tr>
<tr>
<td>y24</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
</tr>
<tr>
<td>y25</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
</tr>
</tbody>
</table>

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

\[ Y_{ij} = \beta_0 + \beta_1 v_{a_i} + \beta_2 a_{ge_i} \]

\[ Y_{i1} = \beta_0 + \beta_1 v_{a_i} + \beta_2 a_{ge_i} \]
\[ Y_{i2} = \beta_0 + \beta_1 v_{a_i} + \beta_2 a_{ge_i} \]
\[ Y_{i3} = \beta_0 + \beta_1 v_{a_i} + \beta_2 a_{ge_i} \]
\[ Y_{i4} = \beta_0 + \beta_1 v_{a_i} + \beta_2 a_{ge_i} \]
\[ Y_{i5} = \beta_0 + \beta_1 v_{a_i} + \beta_2 a_{ge_i} \]
## Additional Parameters

Now what does regression model look like?
What are the interpretations of the coefficients?

<table>
<thead>
<tr>
<th>item (Y)</th>
<th>ID</th>
<th>Visual Acuity</th>
<th>Age</th>
<th>l(item=2)</th>
<th>l(item=3)</th>
<th>l(item=4)</th>
<th>l(item=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>y11</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y12</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y13</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y14</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>y15</td>
<td>1</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>y21</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y22</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y23</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y24</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>y25</td>
<td>2</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Model Interpretation

\[ Y_{ij} = \beta_0 + \beta_1(va_i + \beta_2age_i + \alpha_2I(j = 2) + \alpha_3I(j = 3) + \alpha_4I(j = 4) + \alpha_5I(j = 5) \]

\[ Y_{i1} = \beta_0 + \beta_1(va_i + \beta_2age_i) \]
\[ Y_{i2} = \beta_0 + \alpha_2 + \beta_1(va_i + \beta_2age_i) \]
\[ Y_{i3} = \beta_0 + \alpha_3 + \beta_1(va_i + \beta_2age_i) \]
\[ Y_{i4} = \beta_0 + \alpha_4 + \beta_1(va_i + \beta_2age_i) \]
\[ Y_{i5} = \beta_0 + \alpha_5 + \beta_1(va_i + \beta_2age_i) \]
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)

<table>
<thead>
<tr>
<th>item (Y)</th>
<th>Visual Acuity</th>
<th>Age</th>
<th>I2</th>
<th>I3</th>
<th>I4</th>
<th>I5</th>
<th>va*I2</th>
<th>va*I3</th>
<th>va*I4</th>
<th>va*I5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y11</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y12</td>
<td>x11</td>
<td>x12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y13</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y14</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x11</td>
<td>0</td>
</tr>
<tr>
<td>y15</td>
<td>x11</td>
<td>x12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x11</td>
</tr>
<tr>
<td>y21</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y22</td>
<td>x21</td>
<td>x22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x21</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y23</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y24</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x21</td>
<td>0</td>
</tr>
<tr>
<td>y25</td>
<td>x21</td>
<td>x22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>x21</td>
</tr>
</tbody>
</table>

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

- $\beta_1 + \delta_2 = \text{expected difference in risk in item 2 for a one unit change in visual acuity.}$
- 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

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Estimate</th>
<th>Robust SE</th>
<th>Model SE</th>
<th>Robust Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Visual acuity ($\beta_1$)</td>
<td>-4.10</td>
<td>0.28</td>
<td>0.27</td>
<td>-14.7</td>
</tr>
<tr>
<td>Age ($\beta_2$)</td>
<td>-0.03</td>
<td>0.008</td>
<td>0.008</td>
<td>-3.5</td>
</tr>
<tr>
<td>$I_2$ ($\alpha_2$)</td>
<td>-1.47</td>
<td>0.06</td>
<td>0.06</td>
<td>-24.5</td>
</tr>
<tr>
<td>$I_3$ ($\alpha_3$)</td>
<td>0.74</td>
<td>0.12</td>
<td>0.13</td>
<td>6.0</td>
</tr>
<tr>
<td>$I_4$ ($\alpha_4$)</td>
<td>-0.21</td>
<td>0.07</td>
<td>0.07</td>
<td>-3.1</td>
</tr>
<tr>
<td>$I_5$ ($\alpha_5$)</td>
<td>0.85</td>
<td>0.18</td>
<td>0.17</td>
<td>4.7</td>
</tr>
<tr>
<td>$I_2^{*va}$ ($\delta_2$)</td>
<td>0.66</td>
<td>0.21</td>
<td>0.27</td>
<td>3.2</td>
</tr>
<tr>
<td>$I_3^{*va}$ ($\delta_3$)</td>
<td>2.25</td>
<td>0.32</td>
<td>0.29</td>
<td>7.1</td>
</tr>
<tr>
<td>$I_4^{*va}$ ($\delta_4$)</td>
<td>2.10</td>
<td>0.31</td>
<td>0.27</td>
<td>6.8</td>
</tr>
<tr>
<td>$I_5^{*va}$ ($\delta_5$)</td>
<td>0.59</td>
<td>0.30</td>
<td>0.28</td>
<td>2.0</td>
</tr>
</tbody>
</table>
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.
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.
Suppose $Y_i, i = 1,...,N$ 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.

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

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

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

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.
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

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

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

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