The importance of the “Means Model” in Mx for modeling regression and association

Dorret Boomsma, Nick Martin

Boulder 2008

This session

- Introduction
- Example 1: regression analysis: how well do sex and age predict Autism Quotient? (AQ)
- Example 2: regression analysis: how well do behavior problems predict Autism Quotient? (AQ)
- Example 3: regression analysis: how well do SNPs in the vitamin D receptor predict body height? (i.e. *genetic association* test)

Means testing, regression analysis etc. for clustered data

- If Ss are unrelated any statistical package can be used for regression analysis, tests of mean differences, estimation of variance.
- If data come from related Ss (e.g. twins) we need to model the covariance structure between Ss to obtain the correct answer.

MX

- Mx allows us to model means and covariance structures (for dependent variables)
- Input must be “raw data” (Full Information Maximum Likelihood - FIML)
- In addition, the user can specify “definition variables” (these are the predictors in a regression equation (= independent variables))
- The independent variables are not modeled in the covariance matrix

The likelihood of the i^{th} family

$$(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}_i-\mu_i)'\Sigma^{-1}(\mathbf{x}_i-\mu_i)\right)$$

- **x** is vector of observed values (dependent variables) for twin1, twin2 etc
- **μ** is vector of expected values, given observed independent variables (predictors, regressors) such as age, sex, genotype etc.
- **Σ** is the variance covariance matrix of residual values after the regression effects on the expected values have been removed

Testing assumptions (1) (see Monday afternoon)

- Are means (and variances) same for twin1 & twin2 (birth order effects)?
- Are means same for MZ & DZ?
- Are means same for DZ-SS & DZOS (by sex) ? (intrauterine effects – or postnatal)
- **Are means same for men and women?**

Testing assumptions (2)

- Use Option Mx%P= outputfilename to check for outliers at all stages (see manual for a description of the output file)
- Use Sarah Medland's ViewDist Java applet for ease; See: Medland et al.: ViewPoint and ViewDist: utilities for rapid graphing of linkage distributions and identification of outliers. Behav Genet, jan 2006

Testing assumptions (3)

- Remember that bivariate outliers may not be univariate outliers (e.g. MZ cotwins +/- 1.5 sd)
- Remember that outlier status depends on transformation – do I clean before I transform, or vice versa?
- Worth spending time to get this right!

Individual differences

- Our ultimate goal is to be able to measure all causal variables so the residual variance approaches zero – except for measurement error.

- Until that time we have to continue to model variance components in terms of A, C – and E (latent (=unmeasured) constructs).

- However, if causal variables are also influenced by genes, we want to use multivariate modeling (and not correct the dependent variable)

"Means model" – or preferably model for expected individual values

- $X_i = M + B*P_i + e_i$
  - M = grand mean
  - B = regression
  - P = predictor(s)
  - e = residual term
- i stands for individual (M and B are invariant over individuals)
- but how do I read in the predictor variables?

Importance of getting the means model right (1)

- Age regression can look like C in twin model – check for linear, quadratic and even cubic regression on age – plenty of degrees of freedom –
- Check for different age regression in males and females – age*sex, age2*sex (hint: create these definition variables in SPSS)

Importance of getting the means model right (2)

- If pooling data from 2 sexes, sex differences in means can create C
- Best to model grand mean (female) + male deviation – identification problem
- BUT correcting for age, sex effects on means does not mean that residual variance components are necessarily homogeneous between groups – need GxE modeling (this afternoon)

Mx script for age/sex correction

- Script = sat_mzdz_regres sex age.mx
- Data file = Comb_AQ_YSR.dat

- Dependent variable is a quantitative Autism Score (AQ)
- Data were collected on 18 year old twins and 1 extra sibling

Mx script for age/sex correction

- Saturated model for twin-sibling data (2 twins and 1 sibling)
- Test for similarity in means between twins and sibs
- Test for similarity in variances between twins and sibs
- Regression on age and sex

Definition variables: age and sex

Definition variables can**not** be missing, even if dependent variable is missing in FIML

- if dependent variable is missing, supply a valid dummy value (doesn't matter which value, as long as it is not the same as the missing code for the dependent variable!)
- if dependent variable is not missing, supply e.g. the population mean for the definition variable, or the co-twin's value – i.e. impute with care!

Mx script for age/sex correction

Saturated model for twin data (2 twins and 1 sib) with regression on age and sex:

Estimates:

- Mean for twins, sibs (2)
- Variance for twins, sibs (2)
- Covariance for MZ twins, DZ twins and twin-sib (3)
- Regression age, sex (2)

assignment

- Test for equality of twin and sib means and variances
- Regression: test for significance of age and sex regression
- When age & sex are included what is the total variance of (the residual) AQ?
- When age & sex are NOT included what is the total variance of AQ?
- In each model (with/without) age and sex as covariates, what are the MZ, DZ and sib correlations?

- if difficult look at :
- Complete sat_mzdz_regres sex age.mx

Output: means and variances

- **Mean twins = 105.96**
- **Mean sibs = 106.27**
- **Constrained to be equal; m = 105.85**

- **Variance twins = 106.92**
- **Variance sibs = 115.63**
- **Constrained to be equal; V = 108.82**

Test: -2LL = 3248.698 (9 parameters)
-2LL = 3248.983 (7 parameters)

Mx output

Regression (full model):

- **MATRIX P**

- **AGE -0.1102**
- **SEX -3.1879**

Regression (2nd model):

- **MATRIX P**

- **AGE -0.0983**
- **SEX -3.2651**

Mx output

- **What are the MZ, DZ and twin-sib covariances?**

- **MATRIX G (MZ)**
- **47.9526**

- **MATRIX H (DZ)**
- **37.5642**

- **MATRIX I (sib-twin)**
- **25.4203**

Mx output

- **What are the MZ, DZ and twin-sib correlations?**

- **MZ 0.4407**
- **DZ 0.3452**
- **Sib-twin 0.2336**

Same dataset: Multiple regression with age/sex and 8 additional predictors (CBCL scales) of AQ

- Anxious/Depressed,
- Withdrawn Behavior,
- Somatic Complaints,
- Social Problems,
- Thought Problems,
- Attention Problems,
- Aggressive Behavior,
- Rule-Breaking Behavior.

Mx script for age/sex correction and regression of 8 additional predictors (CBCL scales)

- **modify the age/sex correction script to include an additional 8 predictors**
- OR: sat_mzdz_regres .mx

- (can use the same input file)

Output age/sex correction and 8 additional predictors (CBCL scales)

- **AGE -0.1567**
- **SEX -4.3844**
- **ATT 0.0189**
- **ANX 0.0149**
- **AGG 0.0229**
- **SOM 0.3240**
- **DEL -0.7590**
- **THO 0.6941**
- **SOC 1.2001**
- **WIT 1.4749**

What is now the variance of AQ?

Output age/sex correction and 8 additional predictors (CBCL scales)

- **AGE -0.1567**
- **SEX -4.3844**
- **ATT 0.0189**
- **ANX 0.0149**
- **AGG 0.0229**
- **SOM 0.3240**
- **DEL -0.7590**
- **THO 0.6941**
- **SOC 1.2001**
- **WIT 1.4749**

Variance of AQ = 82.0361

Backward stepwise regression

- All predictors are entered in the regression equation.
- The predictor explaining the least variance in AQ scores (based on b^2 * Var (predictor) is dropped from the model.
- This procedure is repeated until the significance of each syndrome scale is tested.
- CBCL scales that were significant predictors of AQ scores can be included in multivariate genetic analyses

See Hoekstra et al: Twin Research Human Genetics, 2007

Genetic and Environmental Covariation Between Autistic Traits and Behavioral Problems

Rosa A. Hoekstra,[1] Meike Bartels,[1] James J. Hudziak,[2] Toos C. E. M. Van Beijsterveldt,[1] and Dorret I. Boomsma[1]

[1]Department of Biological Psychology, VU University, Amsterdam, the Netherlands
[2]Department of Psychiatry, University of Vermont College of Medicine, Burlington, United States of America

Our objective was to examine the overlap between autistic traits and other behavioral problems in a general population sample, and explore the extent to which this overlap is due to genetic or environmental factors. Youth Self Report (YSR) data were collected in a general population sample of 424 twin pairs at 18 years of age, and their nontwin siblings. In 197 of these twin families, self-report ratings on the Autism-spectrum Quotient (AQ) were collected. Stepwise backward regression analyses revealed that of all 8 YSR syndrome scales, the Withdrawn Behavior (WB) and Social Problems (SOC) scale were the most important predictors of AQ scores, and together with sex, explained 23% of

Landa et al., 1992; Piven et al., 1997). These findings suggest that the same genetic variants that affect the risk for autism may influence the expression of a 'broader autism phenotype' in relatives of autistic probands (Piven et al., 1997; Spiker et al., 2002). Rather than treating autism as a distinct disorder, recent twin and family studies incorporated a dimensional approach to study the etiology of autistic traits and showed that genetic effects also explain a substantial proportion of the variance in autistic traits in the general population (Constantino & Todd, 2000; Constantino & Todd, 2003; Hoekstra et al., 2007b; Ronald et al., 2005; Ronald et al., 2006).

Download: www.tweelingenregister.org

Saving residuals

- If running linkage on 800 markers (or GWA on 500k SNPs) it is wasteful to estimate invariant fixed effects (age, sex, batch effects etc) for every marker
- Mx allows you to save residuals after baseline run and then use these as input variables for batch runs
- Option saveres

Including genotypes in the means model

- Allelic model (2 alleles), Genotypic model (3 genotypes (0,1,2 alleles))
- For SNPs, one allelic deviation, 2 genotype deviations (dominance)
- For microsatellites with k alleles, k-1 deviations, k(k-1)/2 – 1 deviations!
- Missing genotypes?

Including genotypes in the means model

- Phenotype = height
- Predictors: sex and **birth cohort** (not age)
- Predictors: 4 SNPs in vitD receptor
- Coding: 0,1,2 (N of alleles)

- Script: vitD_mzdz_regres.mx
- Data: vitdata.dat

- OR: modify one of the existing scripts

Including genotypes in the means model

- Only 4 SNPs as main effects; if all interactions are included; this is equivalent to haplotype analysis

- Parameters (11):
- grand mean (1)
- sex, cohort & SNP regression (6)
- variance (1)
- MZ, DZ, Sib correlation (3)

Output: MZ, DZ and Sib correlation

- MZ 0.9382
- DZ 0.4264
- Sib 0.4096

output

- MATRIX J (mean)
- 169.2414

- MATRIX P (regression)

- SEX -13.3969
- COH 0.2001
- SNP1 0.1259
- SNP2 -0.6854
- SNP3 0.4106
- SNP4 0.6152

- Test which SNPs are significant: is there evidence for genetic association?

Output: test of the 3 SNPs

- Your model has 11 estimated parameters
- -2 times log-likelihood = **2712.577**

- Your model has 7 estimated parameters
- -2 times log-likelihood = **2713.359**