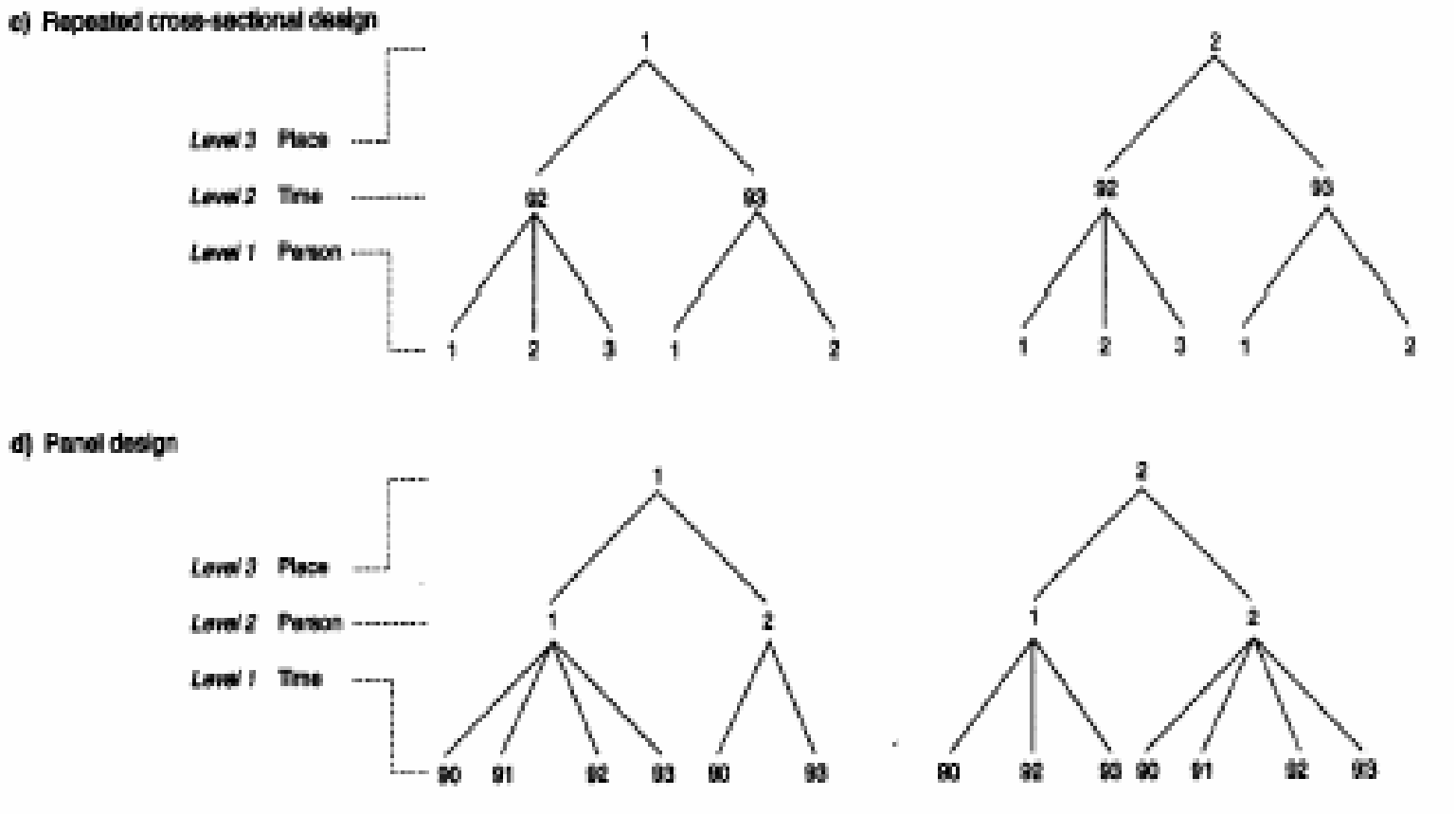
Statistical Methods for Multilevel Analysis

Xiao-Hua Andrew Zhou, PhD
Co-investigator and biostatistician, NACC
Professor
Department of Biostatistics, University of Washington
Director of Biostatistics Unit, HSR&D VA Seattle Medical Center

Examples of Multilevel (Hierarchical) Data

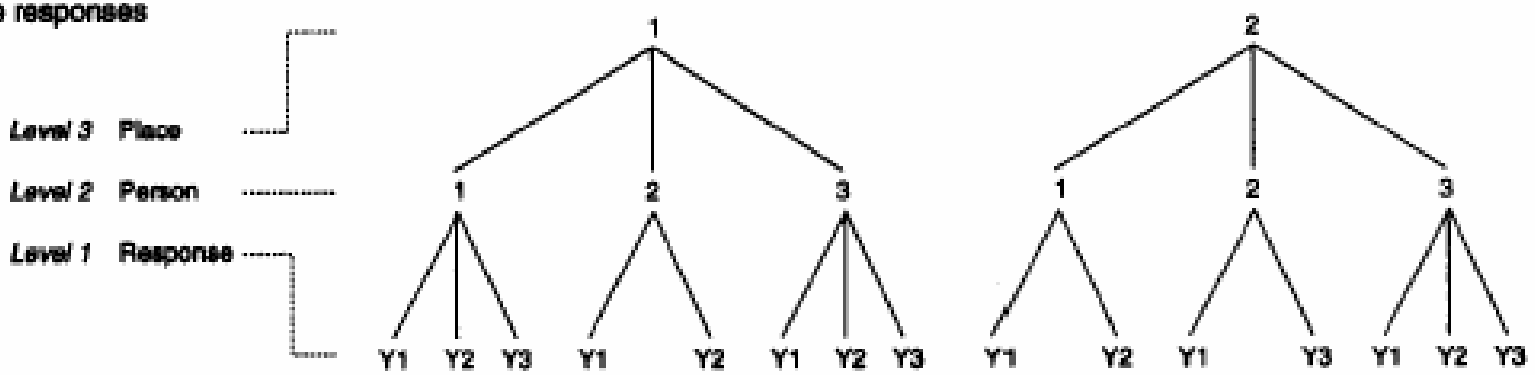
- Individual-family-neighborhood
- Students-classroom-school-district
- ADC subjects – ADC Centers
- Other types, multiple outcomes nested within individual

Data Structure Typology

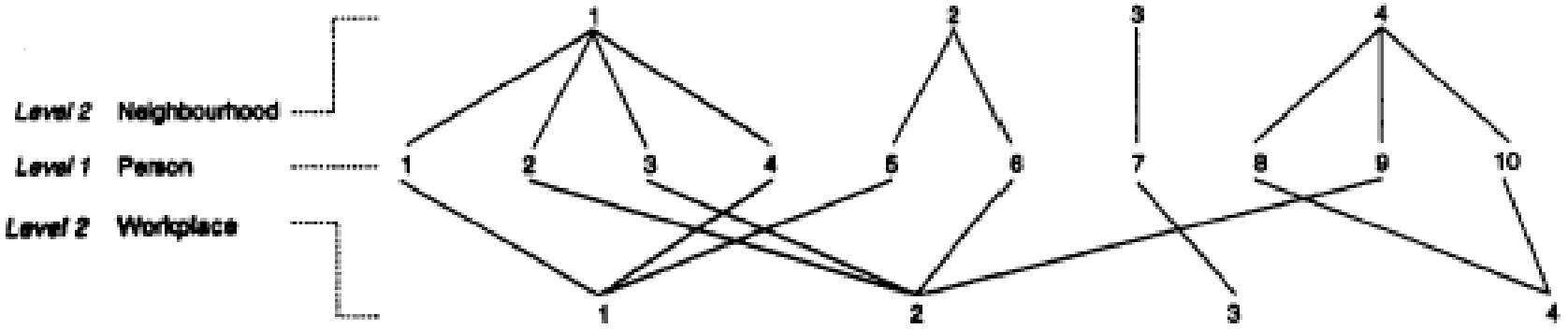


Data Structure Typology

e) Multivariate responses



f) Cross-classified structure



Hierarchical Nature of Data

- Subjects – Centers
- Subject's characteristics, e.g. demographics, clinical diagnosis, etc
- Center's characteristics, e.g. target population, individual objectives

Uniform Data Set

- Standardized clinical examination, testing, diagnostic evaluation and data collection protocol with annual longitudinal follow-up.

Research Questions

- Determine the association between MMSE and age among demented subjects at their first UDS visit. (Linear regression example)

- Determine the association between independence and age among demented subjects at their first UDS visit. (Logistic regression example)

Research Questions, Cont.

- Independent effects of subject-level and Center-level factors.
- Quantification of Center-to-Center variability and the degree which it can be explained by subject-level and Center-level factors

Multilevel (Hierarchical) Models

A hierarchical model analysis will treat the Centers as random effects and will parse out the amount of total variation in the outcome that is attributable to this level of the hierarchy.

An example using two-level linear model on Center

- A study of the relationship between a single subject-level predictor variable (say, age) and one subject-level outcome variable (MMSE) in 26 Centers randomly drawn from the entire population of Centers.

The Age-MMSE relationship in one Center

- Our regression model would be:

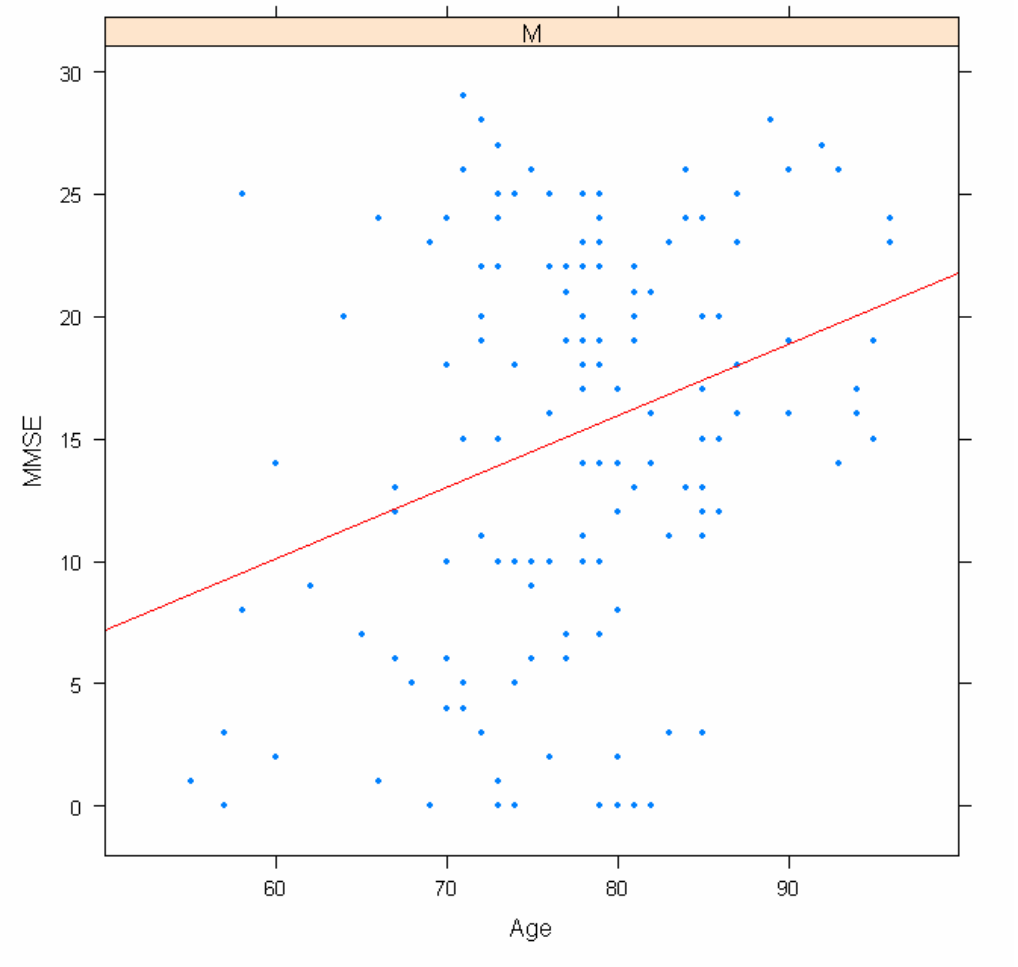
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

Y_i is the MMSE score for demented subject i,

x_i is the AGE for demented subject i,

β_0 is the intercept and β_1 is the slope.

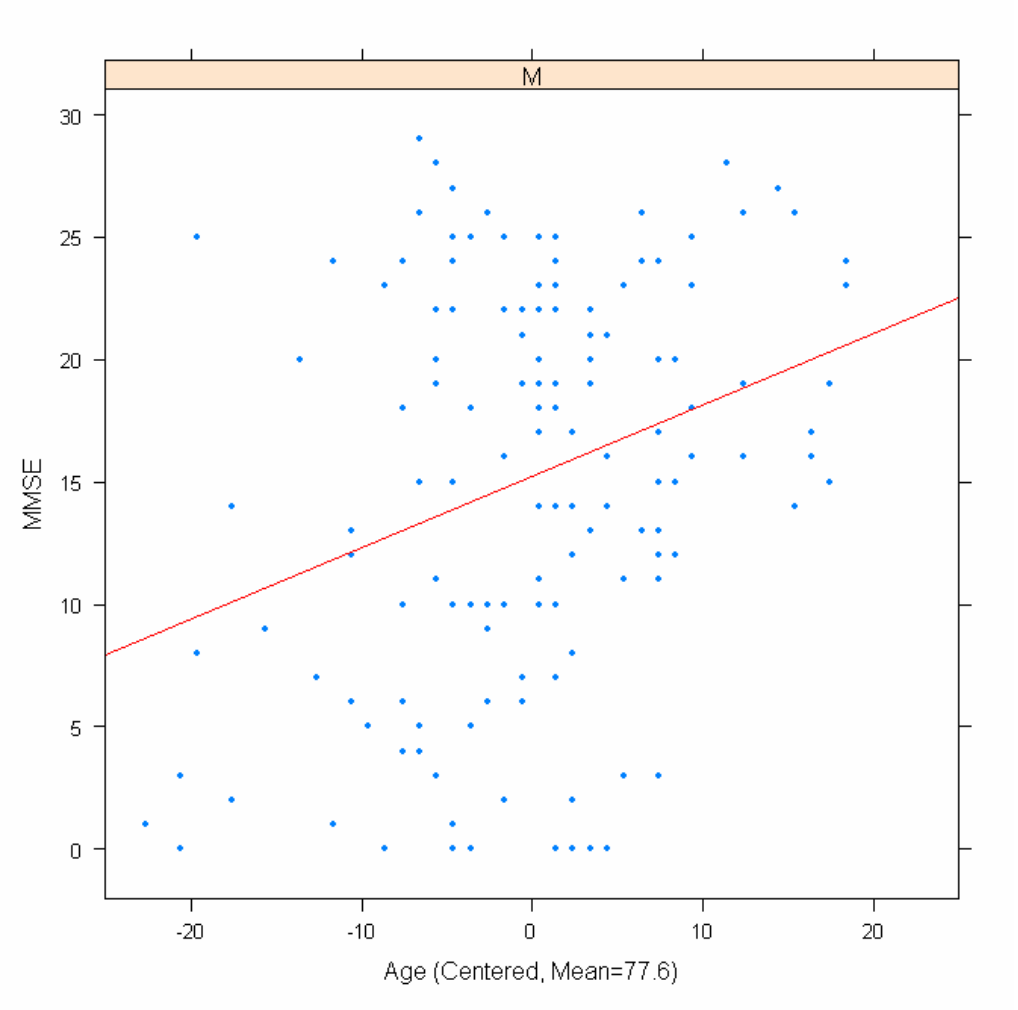
Scatterplot between Age and MMSE in one Center (n=137)



Centering in covariates

- β_0 is defined as the expected MMSE of a demented subject whose AGE is zero.
- It may be helpful to scale the independent variable, X, so that the intercept will be meaningful.
- We center AGE by subtracting the mean AGE from each score.

Scatterplot between Age (Centered) and MMSE



The Age-MMSE relationship in two Centers

- Our regression models would be:

$$Y_{i1} = \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \quad \varepsilon_{i1} \sim N\left(0, \sigma^2\right)$$

Y_{i1} is the MMSE score for demented subject *i* in Center 1,

x_{i1} is the AGE for demented subject *i* in Center 1,

β_{01} is the intercept and β_{11} is the slope for Center 1.

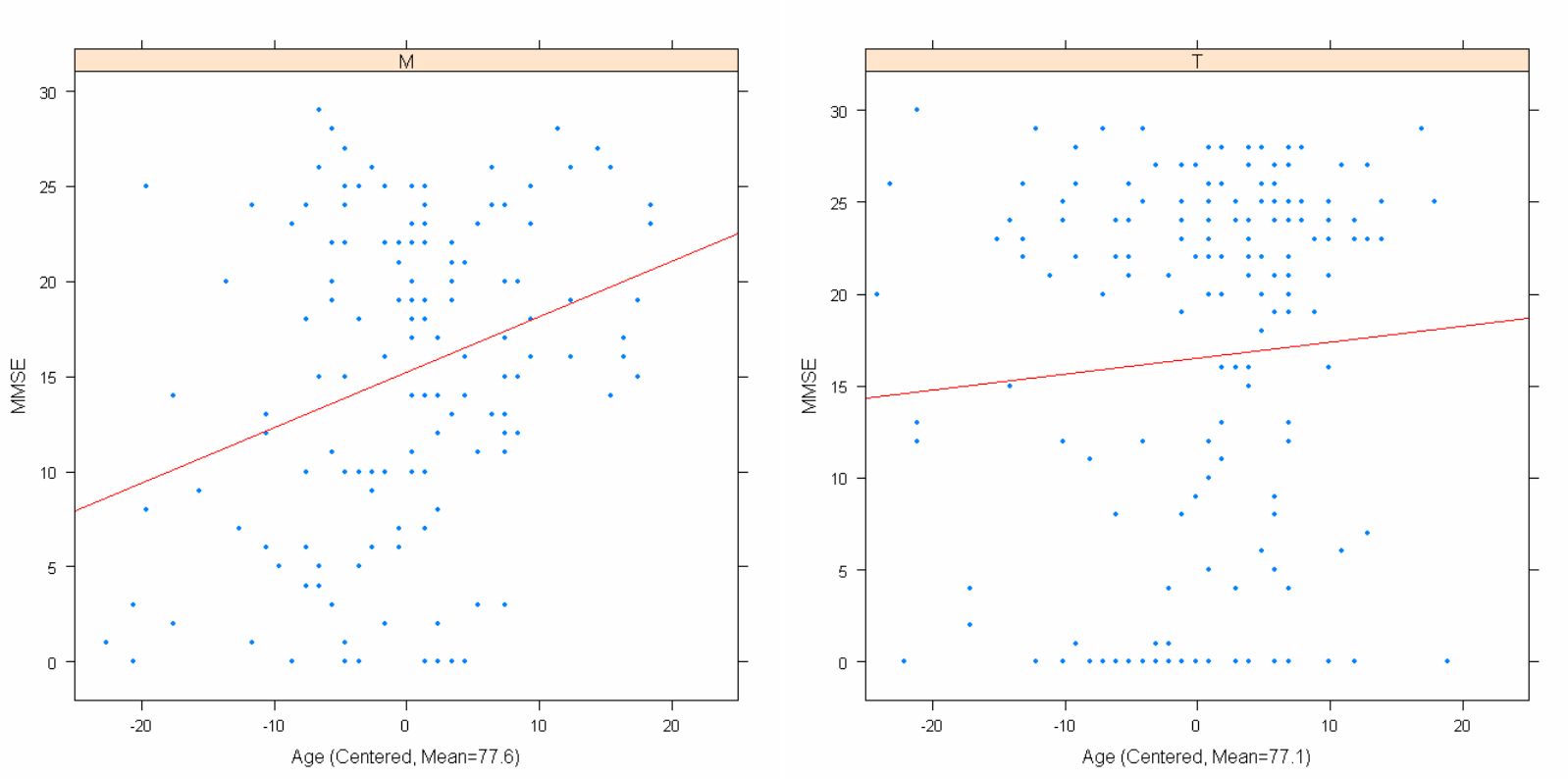
$$Y_{i2} = \beta_{02} + \beta_{12} x_{i2} + \varepsilon_{i2}, \quad \varepsilon_{i2} \sim N\left(0, \sigma^2\right)$$

Y_{i2} is the MMSE score for demented subject *i* in Center 2,

x_{i2} is the AGE for demented subject *i* in Center 2,

β_{02} is the intercept and β_{12} is the slope for Center 2.

Scatterplot between Age (Centered) and MMSE

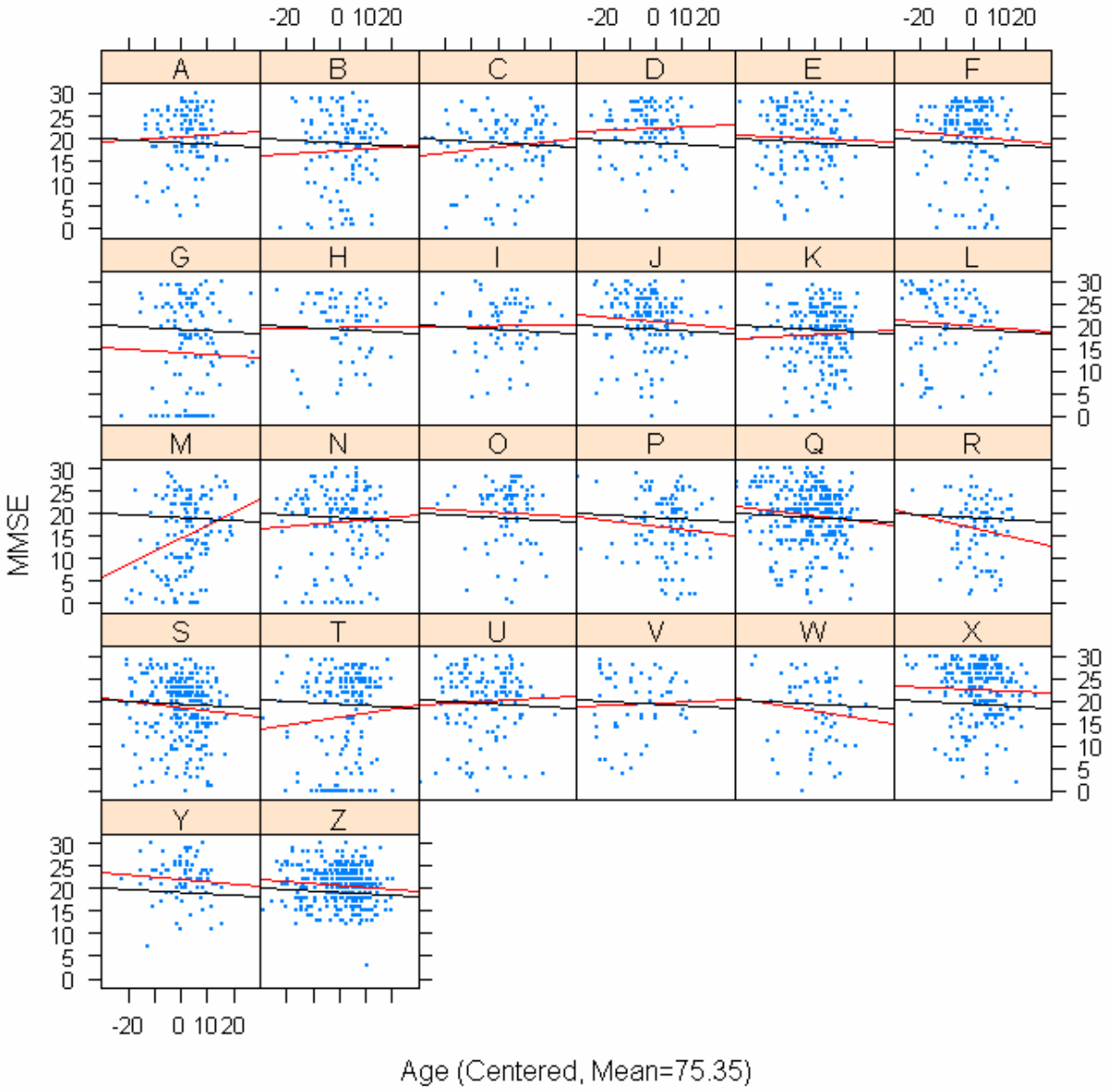


Scatterplot Comments

- The two lines indicate that Center "M" and Center "T" differ in two ways.
  - (1) Center "T" has higher mean MMSE than Center "M" ($\beta_{01} > \beta_{02}$)
  - (2) AGE is less predictive of MMSE in Center "T" than Center "M" ($\beta_{12} < \beta_{11}$)

The Age-MMSE relationship in J Centers (2-level Variance Component)

A trellis plot of each Center is provided next, where red lines denote Center specific fitted lines and black lines are what would result if Center was ignored.


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
MMSE
Age (Centered, Mean=75.35)

The Age-MMSE relationship in J Centers (2-level Variance Component)

- Our regression models would be:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

Y_{ij} is the MMSE score for demented subject i in Center j,

x_{ij} is the AGE for demented subject i in Center j,

β_{0j} is the intercept and β_{1j} is the slope for Center j.

- Often sensible and convenient to assume that the intercept and slope have a bivariate normal distribution across the population of Centers.

$$E(\beta_{0j}) = \gamma_0, Var(\beta_{0j}) = \tau_{00}, E(\beta_{1j}) = \gamma_1,$$

$$Var(\beta_{1j}) = \tau_{11}, Cov(\beta_{0j}, \beta_{1j}) = \tau_{01}$$

Interpretation

- γ_0: the average Center mean for the population of Centers
- τ_{00}: the population variance among the Center means
- γ_1: the average Age-MMSE slope for the population of Centers
- τ_{11}:the population variance among the slopes
- τ_{01}: the population covariance between slopes and intercepts

Estimation methods

- It is not possible to estimate the parameters of these regression models directly because the outcomes (β_{0j}, β_{1j}) are not observed.
- However, the data contain information needed for this estimation.

Estimation methods, Cont.

- Combining models in two stages, we obtain

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{ij} - \bar{X}) + \gamma_{11} W_j (X_{ij} - \bar{X}) + \eta_{ij},$$

$$\eta_{ij} = u_{0j} + u_{1j}(\mathrm{X_{ij}} - \bar{\mathrm{X}}) + \varepsilon_{ij}.$$

Estimation methods, Cont.

- The overall linear regression model is not the typical linear model assumed in standard ordinary least squares (OLS).
- Efficient estimation and accurate hypothesis testing based on OLS require that the random errors are independent, normally distributed, and have constant variance.
- In contrast, random errors in our overall model are dependent within each Center and also have non-constant variances.

Estimation methods, Cont.

- The variance of random errors has the following complicated form:

$$Var(\eta_{ij}) = \tau_{00} + \tau_{11}(\mathrm{X_{ij}} - \bar{\mathrm{X}})^2 + \sigma^2.$$

Estimation methods, Cont

- Three types of parameters to estimate to be estimated:
- Fixed effects (γ_{00},γ_{01},γ_{10},γ_{11})
- Random level-1 coefficients (β_{0j},β_{1j})
- Variance-covariance components (σ^2,τ_{00},τ_{11},τ_{01})

Three common estimation methods

- Maximum likelihood (ML) method is a general estimation procedure, which produces estimates for the population parameters that maximize the probability of the observing the data given the model.
- Iterative generalized least squares (IGLS) and Restricted Iterative generalized least squares.
- Bayesian method

ML method

- Two different likelihood functions:

1. Full Maximum Likelihood (FML) – both the regression coefficients and the variance components are included in the likelihood function.
2. Restricted Maximum Likelihood (RML) – only the variance components are included in the likelihood function, and the regression coefficients are estimated in a second estimation step.

Comparison of these two methods

- FML is more efficient and can provide estimates for both variance components and fixed effect parameters. But, FML may produce biased estimates for variance components.
- RML can provide less biases estimates for the variance components and is equivalent to ANOVA estimates, which are optimal, if the groups are balanced.
- FML still continues to be used because (1) its computation is generally easier, and (2) it is easier to compare two models that differ in the fixed parameters using the likelihood-based tests. However, with RML, only differences in the random part can be compared with likelihood-based tests

Linear Regression Model Results

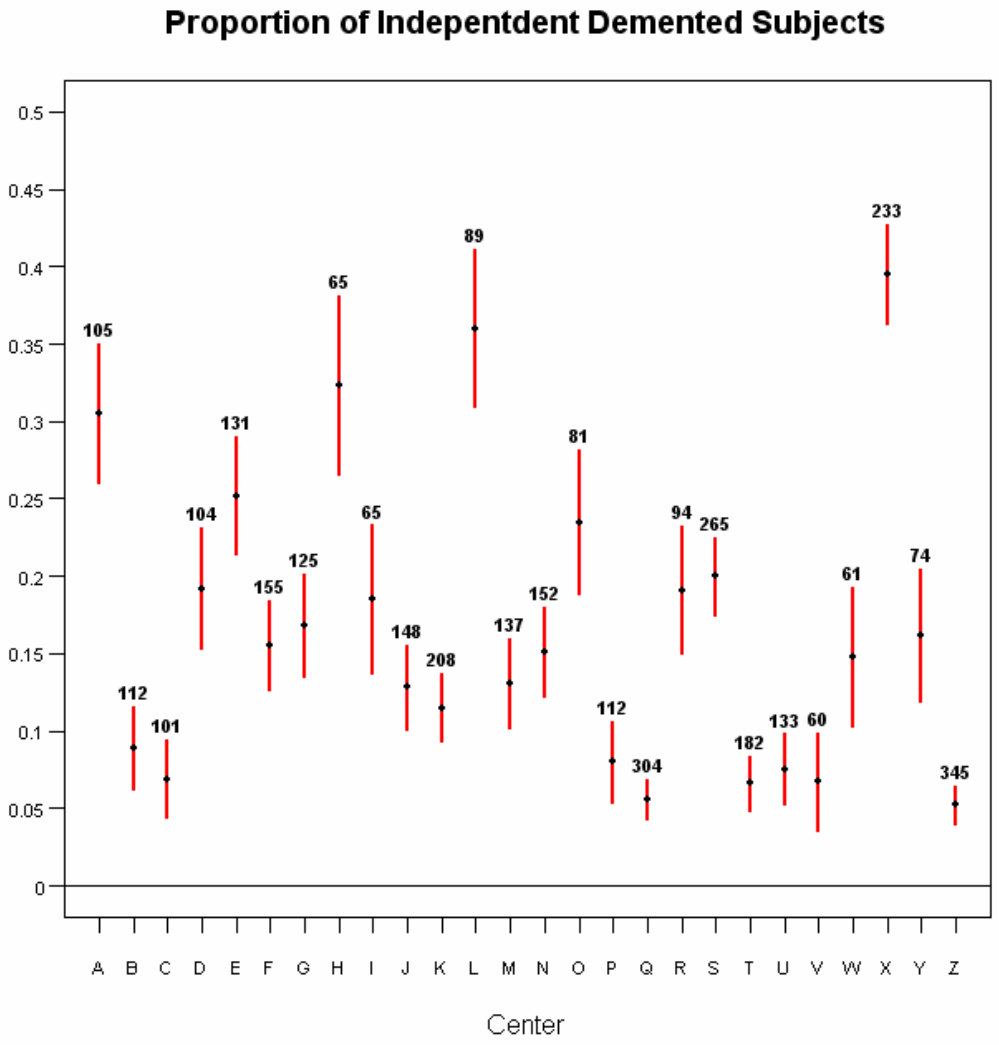
	Estimate	Std. Error	P-Value
Ignoring Center			
- Intercept	19.172	0.120	<0.001
- Slope	-0.032	0.012	0.009
Modeling Center			
(FML) Fixed Effects			
- Intercept	19.084	0.405	<0.001
- Slope	-0.009	0.012	0.459
Random Effects	**Std. Dev.**		
- Center	2.002		
- Residual	7.007		
(RML) Fixed Effects			
- Intercept	19.084	0.413	<0.001
- Slope	-0.009	0.012	0.459
Random Effects	**Std. Dev.**		
- Center	1.959		
- Residual	7.007		

Two-level binary response models for Independence

- Let Y_{ij} be the binary response variable for whether demented subject i is independent (or requires any assistance) in Center j.

- X_{ij} is the age of subject i in Center j.

Dotplot Between Independence and Center



Two-level logistic regression

$$\text{logit}(\Pr(Y_{ij} = 1)) = \beta_{0j} + \beta_1 X_{ij}$$

$$\beta_{0j} = \beta_0 + u_j$$

$$u_j \sim N\left(0, \sigma_u^2\right)$$

Estimation methods

Several estimation methods for multi-level logistic regression models:

- A quasi-likelihood approach (PQL)
- Laplacian approximation (Lap.)
- Adaptive Gaussian Quadrature (AGD)
- Bayesian approach with MCMC methods.

Age (Centered) – Independence

$$\text{logit}(\Pr(Independenence_{ij} = 1)) = \beta_{0j} + \beta_1 Age_{ij} + e_{ij}$$

$$Independenence = 1 \text{ if completely independent, 0 otherwise}$$

$$\beta_{0j} = \beta_0 + u_j$$

$$u_j \sim N\left(0, \sigma_u^2\right)$$

$$e_{ij} \sim N\left(0, \sigma_e^2\right)$$

Logistic Regression Model Results

	Estimate	Std. Error	P-Value
Ignoring Center			
- Intercept	-1.697	0.0460	<0.001
- Slope	-0.018	0.0045	<0.001
Modeling Center			
(Lap.) Fixed Effects			
- Intercept	-1.771	0.1399	<0.001
- Slope	-0.022	0.0049	<0.001
Random Effects	**Std. Dev.**		
- Center	0.659		
- Residual	0.362		
(PQL) Fixed Effects			
- Intercept	-1.748	0.1407	<0.001
- Slope	-0.022	0.0049	<0.001
Random Effects	**Std. Dev.**		
- Center	0.664		
- Residual	0.362		