Statistical Analysis Using High Dimensional Neuroimaging Data

Danielle J. Harvey
UC Davis

Acknowledgements

- Funded in part by Grant R13AG030995-01A1 from the National Institute on Aging

- The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

Outline

- Typical neuroimaging analytic strategies
- ADNI neuroimaging data
- Challenges
- Image data as outcomes
- Image data as predictors
- Longitudinal models

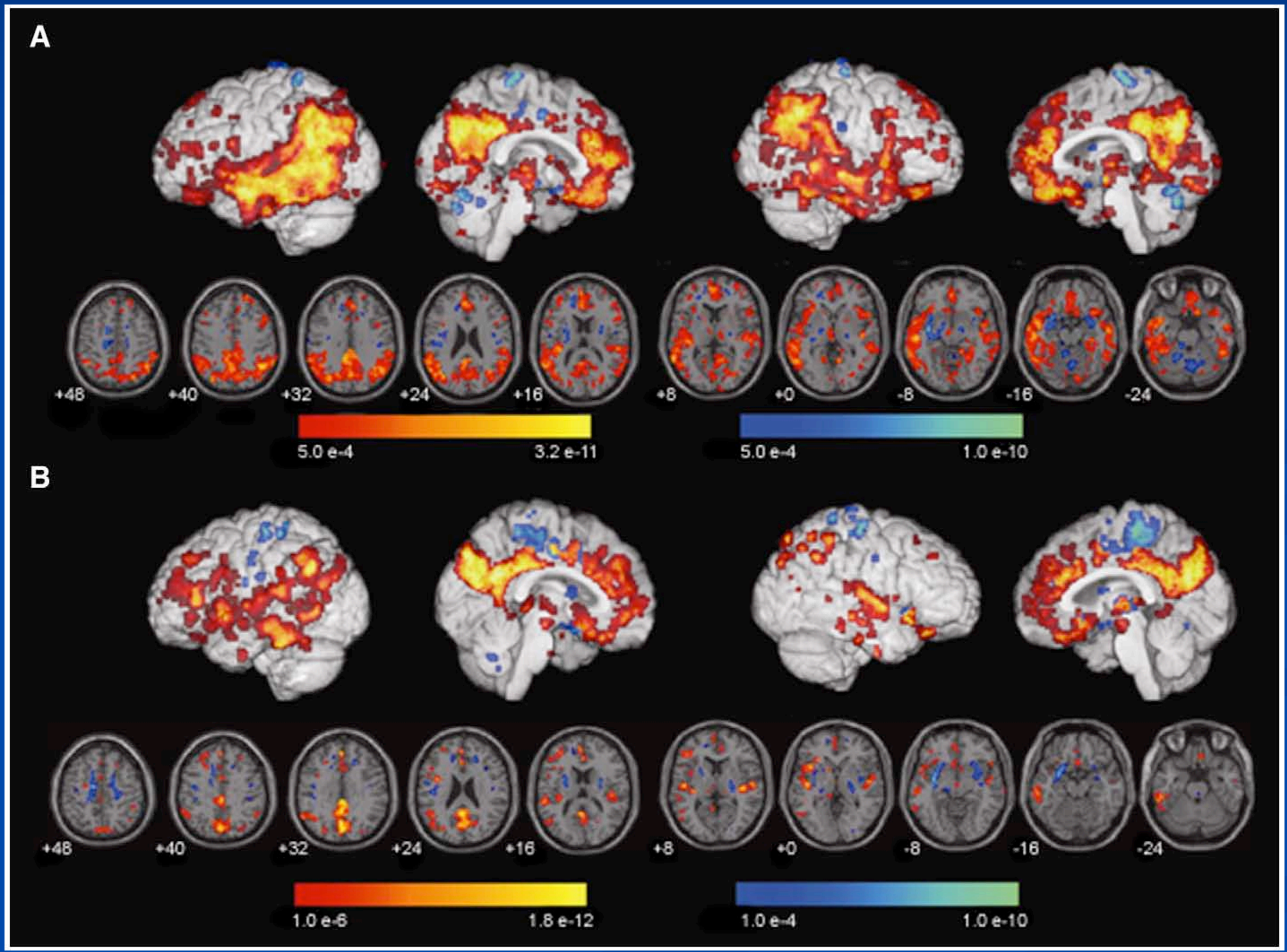
Typical Analytic Strategies

- One-number summaries derived from images
  - Regional volumes (MRI)
  - Regional glucose metabolism (FDG-PET)
  - Regional standard uptake value ratio (PiB)
  - Can be used as outcomes or predictors
- Voxel-based methods
  - Statistical Parametric Mapping (SPM)

SPM

- Statistical analyses done at every voxel (hundreds of thousands of them)
- General linear model framework
  - Simplest context: t-test at every voxel
- Multiple comparison adjustment to identify significant clusters of voxels
  - Gaussian random field theory
  - False discovery rate

SPM example (FDG-PET)



Chen K, et al. NeuroImage (2010)

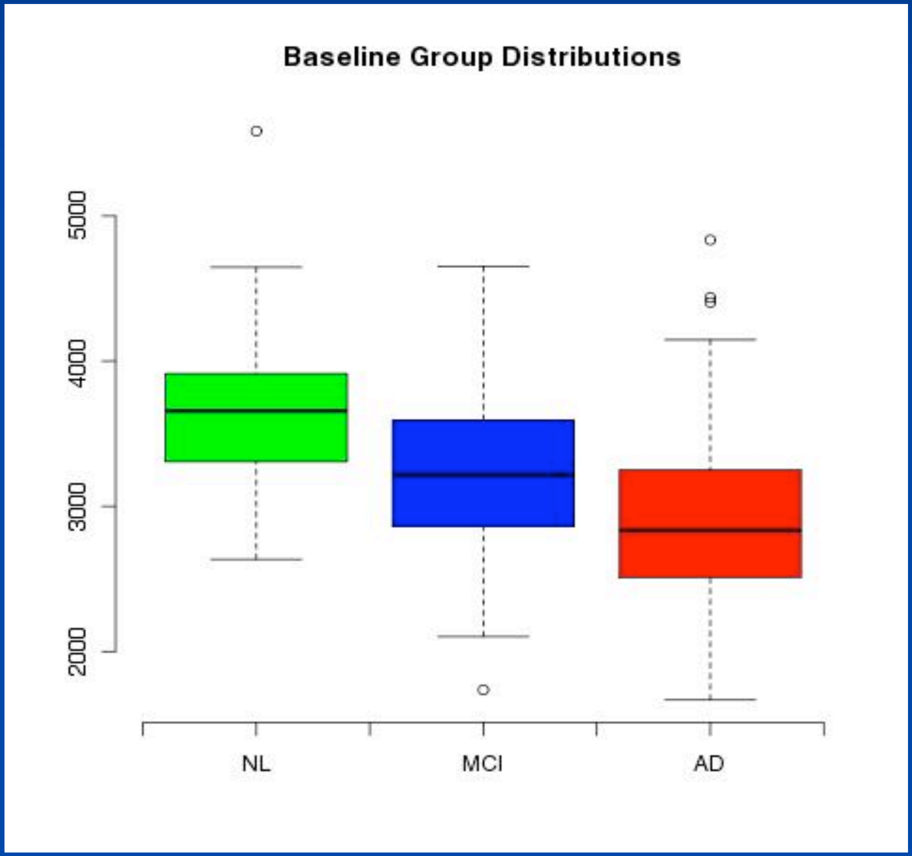
ADNI Neuroimaging Data

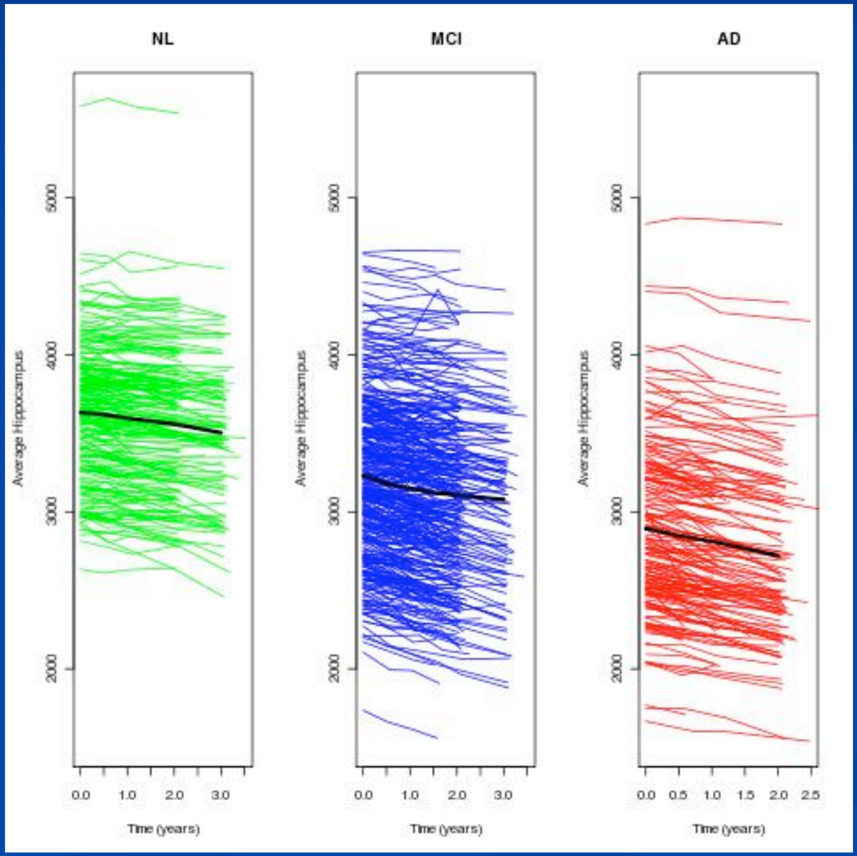
- Structural MRI
  - Regional volumes, cortical thicknesses (Anders Dale, UCSD)
  - White matter hyperintensity volume and stroke information (Charles DeCarli, UCD)
  - Boundary shift integral, regional volumes (Nick Fox, UCL)
  - FreeSurfer data (Norbert Schuff, UCSF)
  - SNT hippocampus (limited data, Norbert Schuff, UCSF)
  - TBM summaries (Paul Thompson, UCLA)
    - Currently not available

ADNI data cont.

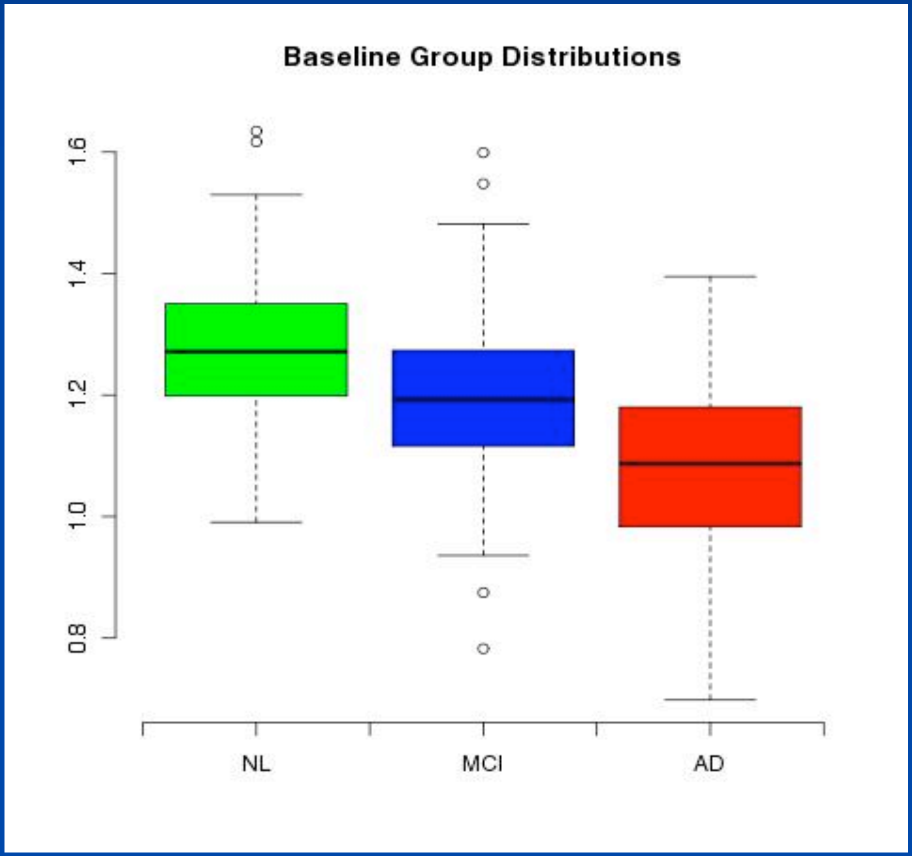
- FDG-PET
  - Stereotactic Surface Projection (SSP) summaries (Norman Foster, Utah)
  - Regional glucose metabolism (William Jagust, UCB)
  - SPM summaries, hypometabolic convergence index (Eric Reiman, Arizona)
- PiB (available for ~100 subjects)
  - Regional standard uptake value ratio (SUVR) (Chet Mathis, Pittsburgh)

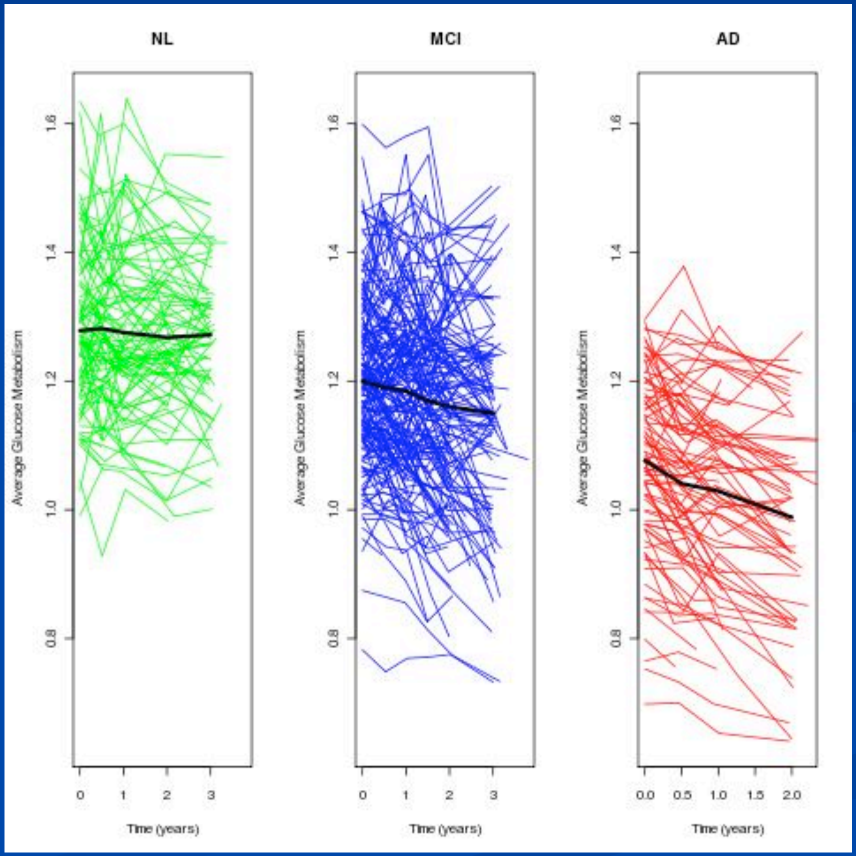
Hippocampal Volume (UCSD)





Average glucose metabolism (Jagust)





Challenges: ADNI image data

- Some measures obtained for every image, while others measure change directly
- Some measures created using a training set to derive the measure
  - Derived measure applied to independent test set
- Several methods used to obtain same measure

More Challenges

- Large number of variables
  - May be highly correlated
- Not all measures available for everyone
  - By design
    - Not everyone has FDG-PET, MRI, and CSF
  - Not all images processed by all labs
- Lab-specific quality control
  - May be outliers in the data due to failing QC for a lab's particular processing method
  - Some labs included a QC variable to flag values considered less reliable

Image Data as an Outcome

- Voxel-based methods
- Numeric summary from each image at each assessment
  - Longitudinal models (repeated measures, random effects models)
- Numeric summary of change utilizing information from two images at a time
  - Linear regression models

Image data as predictors

- High dimensional data
  - Lots of variables
  - Potentially highly correlated
- Need strategies for dimension reduction or handling correlated variables
  - Pick key variables based on underlying hypotheses
  - Principal components analysis (PCA) or factor analysis
  - Cluster analysis
  - Ridge regression
  - Partial least squares regression

PCA/Factor Analysis

- Goal: reduce a set of potentially correlated variables into a smaller set of uncorrelated components
- Identifies underlying structure of the data that explains the most information
- Considered an exploratory technique
  - Dependent on specific data used

PCA/Factor Analysis cont.

- Assumption
  - Data are multivariate normal
- Results
  - Identified components with loadings (weights) of specific variables
    - Generate component scores for each individual
    - Can be used as predictors in models

PCA/Factor analysis cont.

- Limitations
  - Data-dependent
    - Results may change if use different data
  - May be difficult to interpret
    - Components are linear combinations of variables
    - May not be obvious what each component represents

Cluster analysis

- Identifies groups of “similar” observations
- Unsupervised learning
- Many different approaches
  - Define distance metric for assessing similarity
  - Can start with one cluster and split the observations up into a set of clusters
  - Or can start with each observation as its own cluster and then join clusters together
- Results
  - Cluster membership
  - Can be used as predictors of an outcome not used in defining the clusters

Cluster Analysis Cont.

- Challenges/Limitations
  - Specifying the number of clusters
  - Cluster membership may be "fuzzy" at the boundaries
    - Different cluster algorithms may assign individuals to other clusters
  - Data dependent
    - May be difficult to generalize results

Ridge Regression

- Approach to regression that handles highly correlated predictors
- Why would we be interested in using multiple correlated variables in the same model
  - Determine if we can get a better prediction by using, for example, all information we get from an MRI
  - Not interested in independent contribution of each variable

Ridge Regression cont.

- Includes a penalty term in the estimation of the parameters
  - Essentially shrinks the estimates closer to zero
  - If the penalty term is 0, results are usual LS estimates
- No longer yields unbiased estimates of the coefficients
  - But the variance of the estimates may be smaller

Ridge Regression cont.

- Results
  - Parameter estimates for a range of values of the penalty term
- Challenges/Limitations
  - Determining the best value of the penalty term
  - Typical inference procedures (hypothesis tests) do not work in this setting

Partial Least Squares

- Useful when you have many predictors (relative to the number of observations)
- Predictors may be highly correlated
- Goal: identify a few “factors” that explain most of the variability in the outcome
- May fit model in a training set and then see how well the model predicts the outcome in an independent test set

Longitudinal Models

Random Effects Models - Notation

- Let Y_{ij} = outcome for i^{th} person at the j^{th} time point
- Let Y be a vector of all outcomes for all subjects
- X is a matrix of independent variables (such as age, ApoE4 status, and time)
- Z is a matrix associated with random effects (typically includes a column of 1s and time)

Mixed Model Formulation

- $Y = X\beta + Z\gamma + \varepsilon$
- β are the "fixed effect" parameters
  - Similar to the coefficients in a regression model
  - Coefficients tell us how variables are related to baseline (or overall) level and change over time in the outcome
- γ are the "random effects", $\gamma \sim N(0,\Sigma)$
- ε are the errors, $\varepsilon \sim N(0,\sigma^2)$

Random Effects

- Why use them?
  - Not everybody responds the same way (even people with similar demographic and biomarker levels respond differently)
  - Want to allow for random differences in baseline level and rate of change that remain unexplained by the covariates
  - Accounts for between-person variability in level and change

Assumptions of Model

- Linearity
- Homoscedasticity (constant variance)
- Errors are normally distributed
- Random effects are normally distributed
- Typically assume MAR

Interpretation of parameter estimates

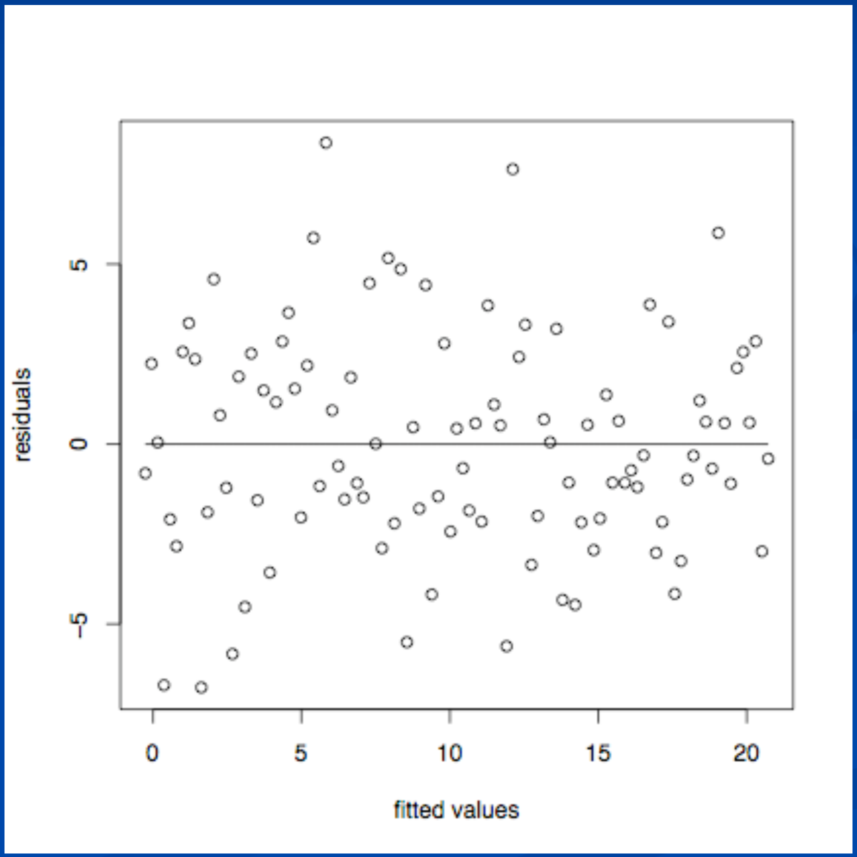
- Main effects
  - Continuous variable: average association of one unit change in the independent variable with the baseline level of the outcome
  - Categorical variable: how baseline level of outcome compares to "reference" category
- Time
  - Average annual change in the outcome for "reference individual"
- Interactions with time
  - How annual change varies by one unit change in an independent variable
- Covariance parameters

Graphical Tools for Checking Assumptions

- Scatter plot
  - Plot one variable against another one (such as random slope vs. random intercept)
  - E.g. Residual plot
    - Scatter plot of residuals vs. fitted values or a particular independent variable
- Quantile-Quantile plot (QQ plot)
  - Plots quantiles of the data against quantiles from a specific distribution (normal distribution for us)

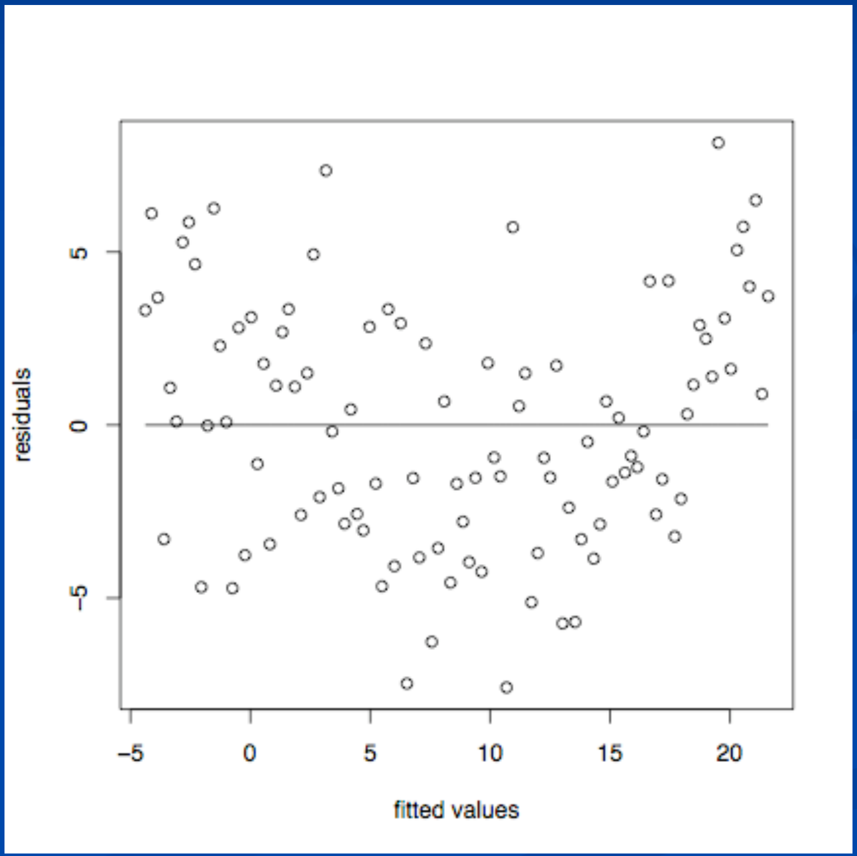
Residual Plot

Ideal Residual Plot
- "cloud" of points
- no pattern
- evenly distributed about zero



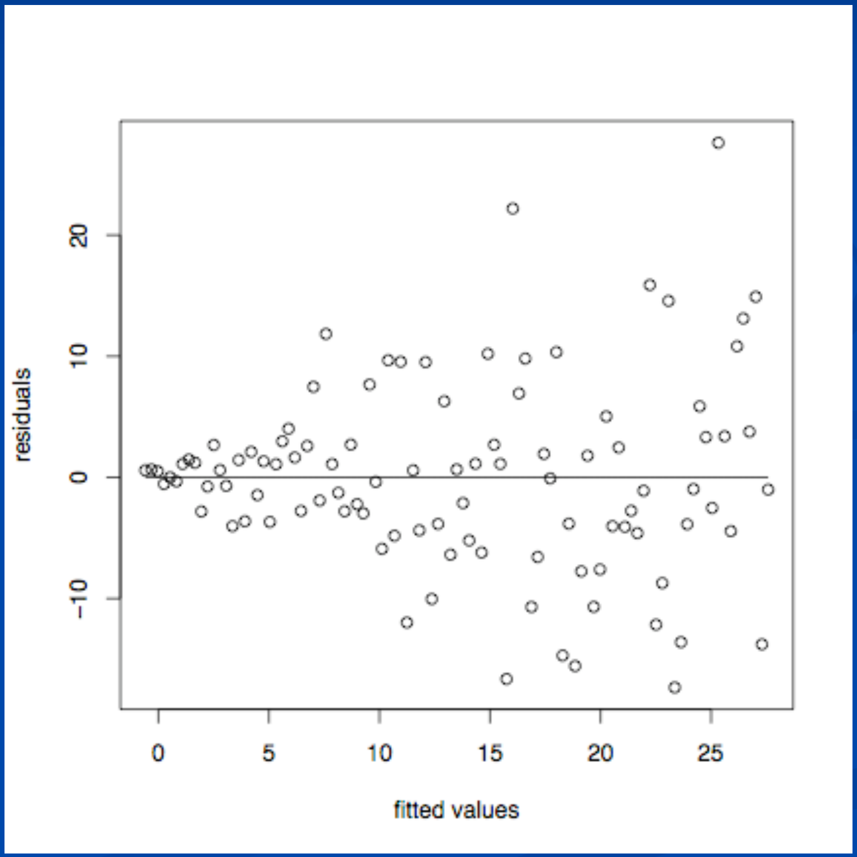
Non-linear relationship

- Residual plot shows a non-linear pattern (in this case, a quadratic pattern)
- Best to determine which independent variable has this relationship then include the square of that variable into the model

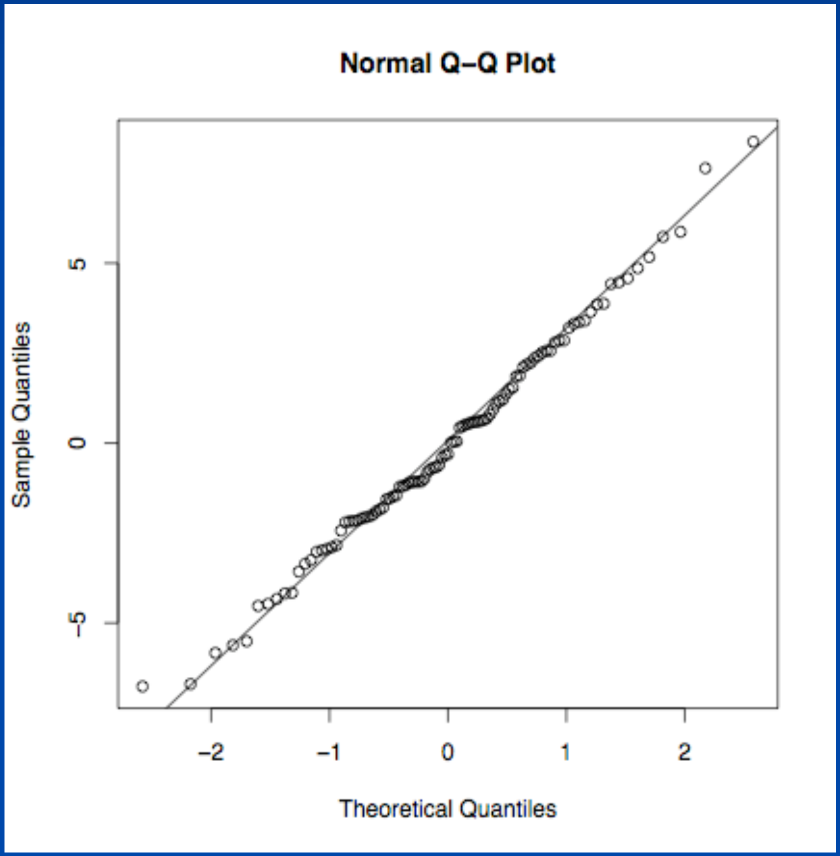


Non-constant variance

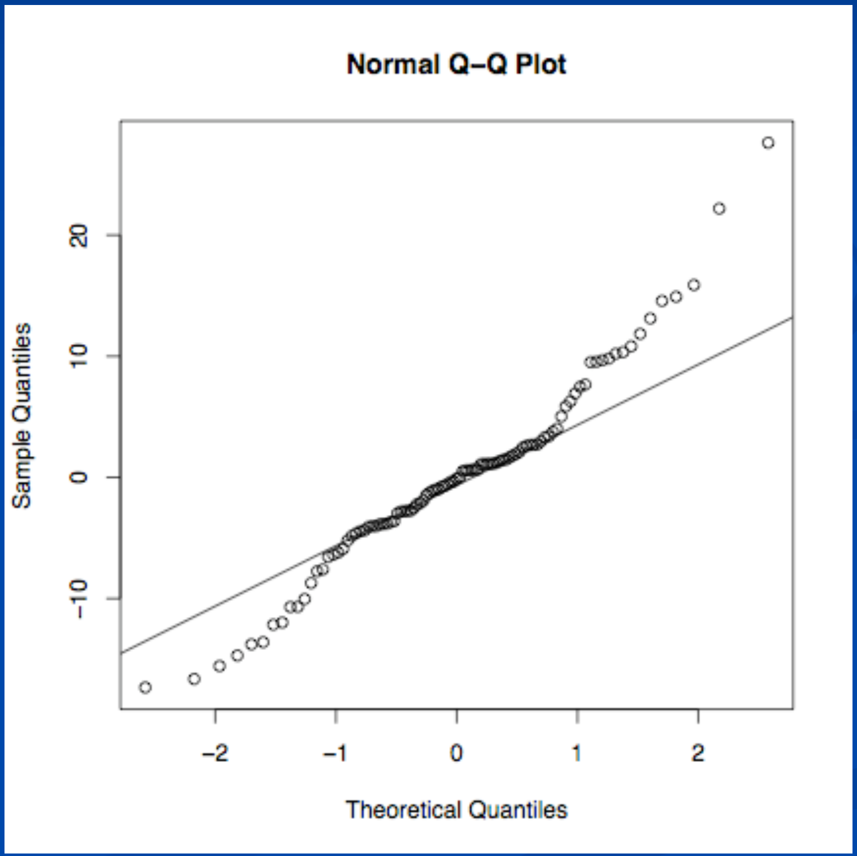
- Residual plot exhibits a "funnel-like" pattern
- Residuals are further from the zero line as you move along the fitted values
- Typically suggests transforming the outcome variable (ln transform is most common)



QQ-Plot

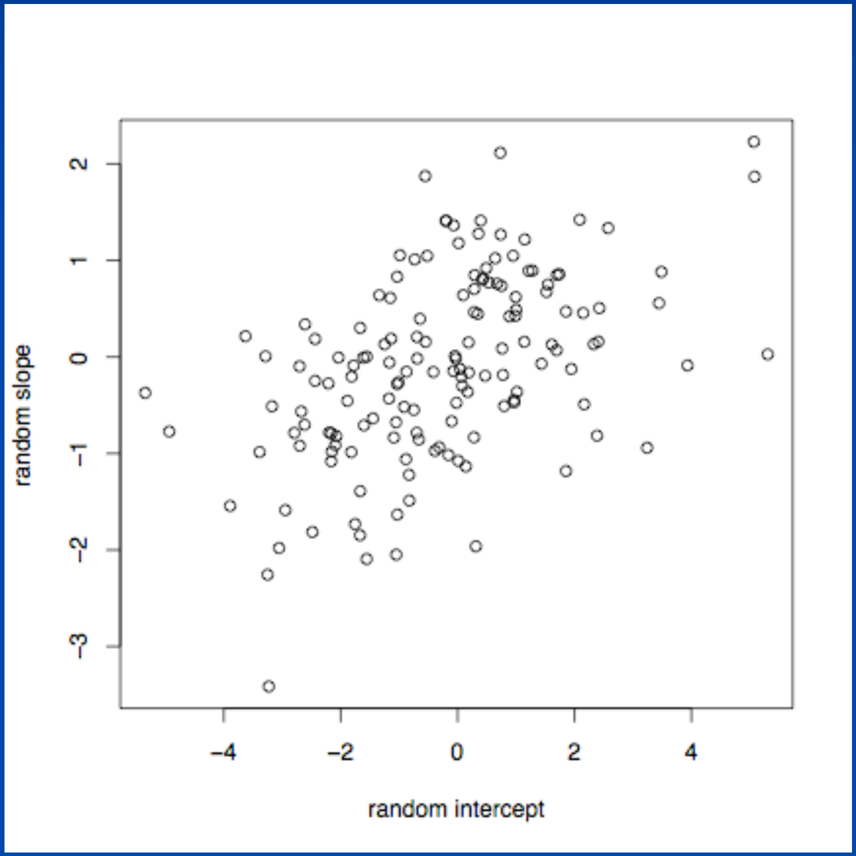


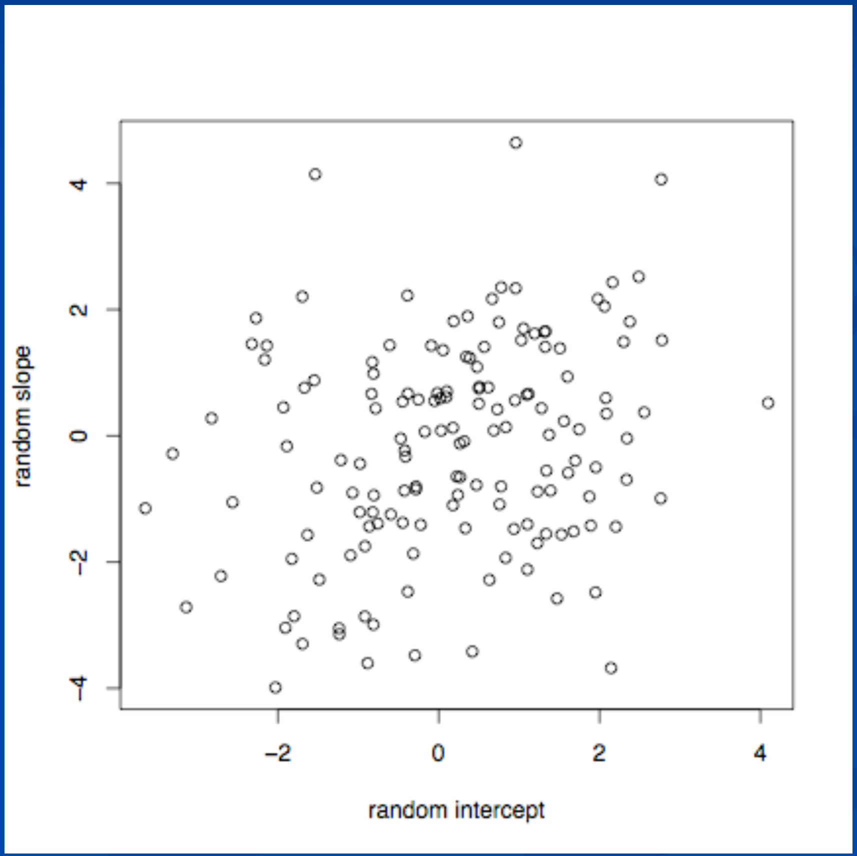
Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−5
0
5
−2
−1
0
1
2



Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−10
0
10
20
−2
−1
0
1
2

Scatter plot of random effects





Conclusions

- A lot of imaging data available in ADNI
- Many challenges/complications with the data
- Strategies for reducing the number of variables to use in analyses
- Introduction to longitudinal models