Longitudinal Data Analysis

Danielle Harvey, Ph.D.
July 12, 2017

This seminar is jointly supported by the following NIH-funded centers:

UCDAVIS
CLINICAL AND TRANSLATIONAL
SCIENCE CENTER

UCDAVIS
MIND INSTITUTE

UCDAVIS
COMPREHENSIVE
CANCER CENTER

UCDAVIS
ENVIRONMENTAL HEALTH
SCIENCES CENTER

We are video recording this seminar so please hold questions until the end.

Thanks

Seminar Objectives

- Understand what statistical methods to use to analyze repeated measures data
- Be able to conduct simple analyses of repeated measures data using SAS

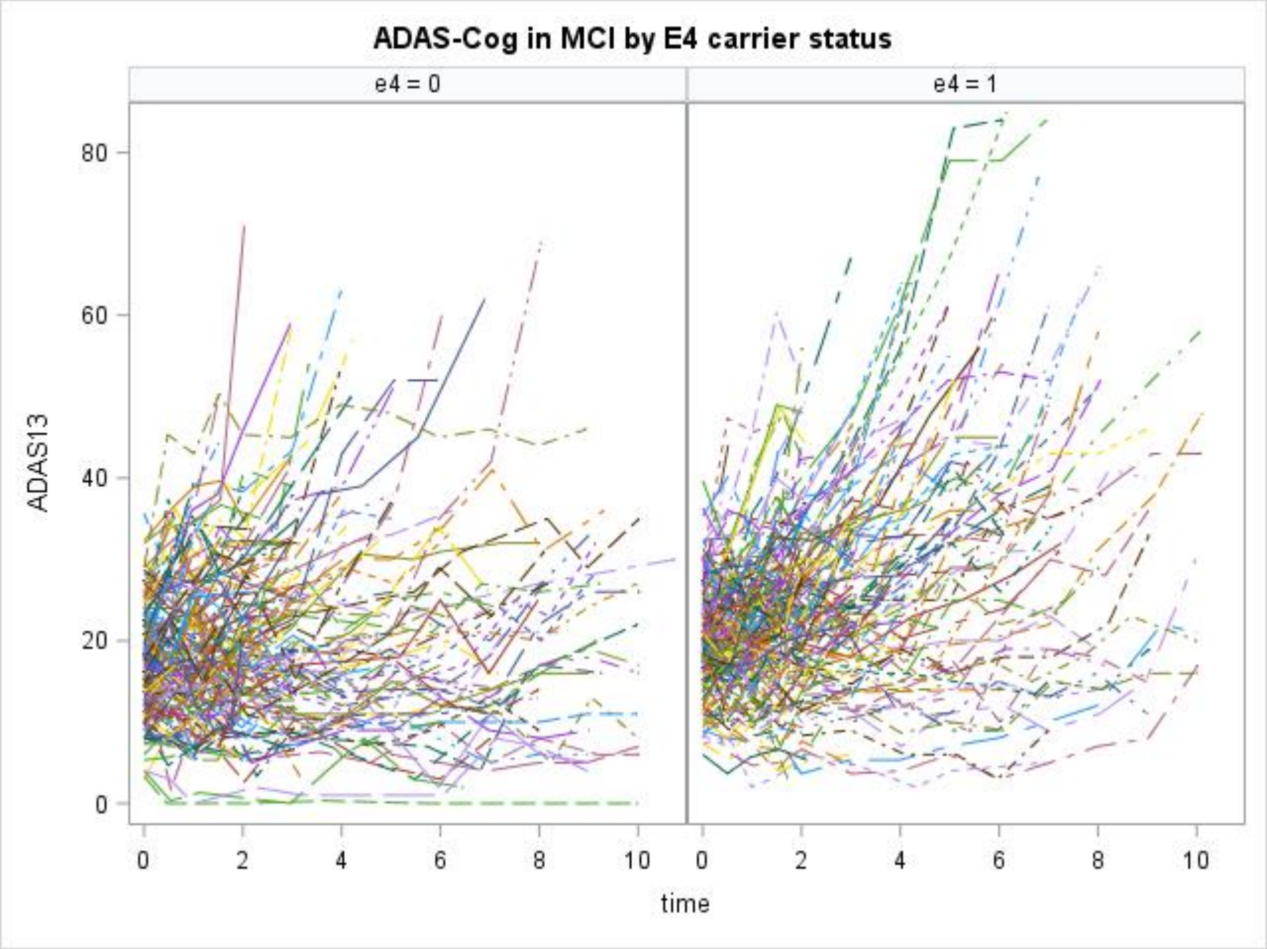
Background

- **Prospective Studies**
  - Follow individuals over time
  - Repeat assessments on the same individual
  - Questions of interest are often about change over time and variables associated with change
  - Observations from the same individual are correlated
  - Linear regression and ANOVA not appropriate

Example: Alzheimer's Disease Neuroimaging Initiative (ADNI)

- Longitudinal study of dementia
- Ongoing since 2004
- Enrolled older individuals with normal cognition, mild cognitive impairment (MCI) or mild dementia
- Seen every 6 months for ~ 2 years, then annual follow-ups
- Clinical eval, neuropsych testing, neuroimaging at each visit
- CSF samples annually
- http://adni.loni.usc.edu/

Spaghetti Plots of ADNI data



Standard Methods for Longitudinal Data Analysis

- **Repeated Measures ANOVA**
  - Extension of ANOVA to correlated data
  - Extension of paired t-test to more than 2 observations per person
  - Continuous outcome with categorical predictors
- **Mixed Effects Regression**
  - Extension of linear regression to correlated data
  - Continuous outcome with continuous or categorical predictors

Basics: Data Structure

- **Wide format**
  - One row per person
  - Multiple outcomes are given as separate variables
  - Typical format for repeated measures ANOVA
- **Long format**
  - One row per observation
  - Multiple rows per person
  - Need individual ID number to link observations from the same person
  - Preferred format for most repeated longitudinal analysis techniques

Basics: Wide Format Data

RID	E4	ADAS13_bl	ADAS13_m06	ADAS13_m12
4	0	21.33	25.33	22
41	1	28.33	25.67	27
54	0	32.33	36.33	39
57	1	19.67	24	41

Basics: Long Format Data

RID	E4	Time	ADAS13
4	0	0	21.33
4	0	0.5	25.33
4	0	1	22
41	1	0	28.33
41	1	0.5	25.67
41	1	1	27
54	0	0	32.33
54	0	0.5	36.33
54	0	1	39
57	1	0	19.67
57	1	0.5	24
57	1	1	41

Basics: Terminology

- **Between-person factors/effects**
  - Variables that change between people
  - Example: sex, baseline age, E4 carrier status
- **Within-person factors/effects**
  - Variables that change within person
  - Example: time
- **Often interested in both between- and within- person factors as well as interactions between the two**

Repeated Measures ANOVA

- **Generally assumes balanced design (no missing data)**
- **Null hypothesis: means are all equal**
- **Alternative hypothesis: at least two means are different**
- **Assumptions**
  - Similar to ANOVA (normality of residuals, constant variance across groups)
  - Added assumption: sphericity (variances of differences between all possible pairs of within-level conditions are the same)

Repeated Measures ANOVA in SAS

No univariate models for each outcome (meaningless for repeated measures analysis)

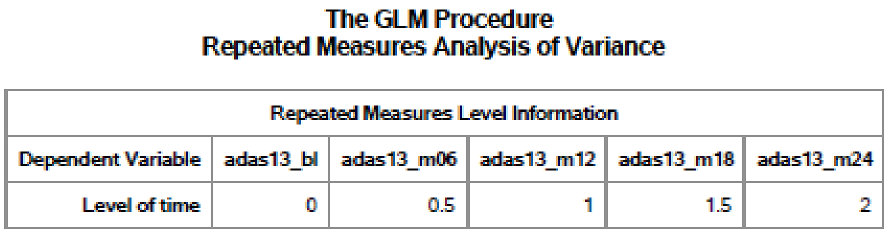
Requests tests of sphericity

```
proc glm data=adni_wide;
     class e4;
     model adas_bl--adas_m24 = e4/nouni;
     repeated time 5 (0 0.5 1 1.5 2)/printe;
run;
```

5 outcome assessments

Levels of time (in years)

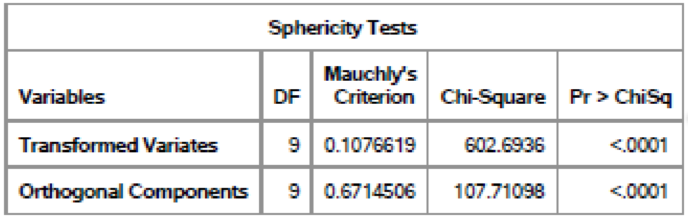
SAS Output for Proc GLM
Some Initial Checks



The GLM Procedure
Repeated Measures Analysis of Variance

Repeated Measures Level Information					
Dependent Variable	adas13_bl	adas13_m06	adas13_m12	adas13_m18	adas13_m24
Level of time	0	0.5	1	1.5	2

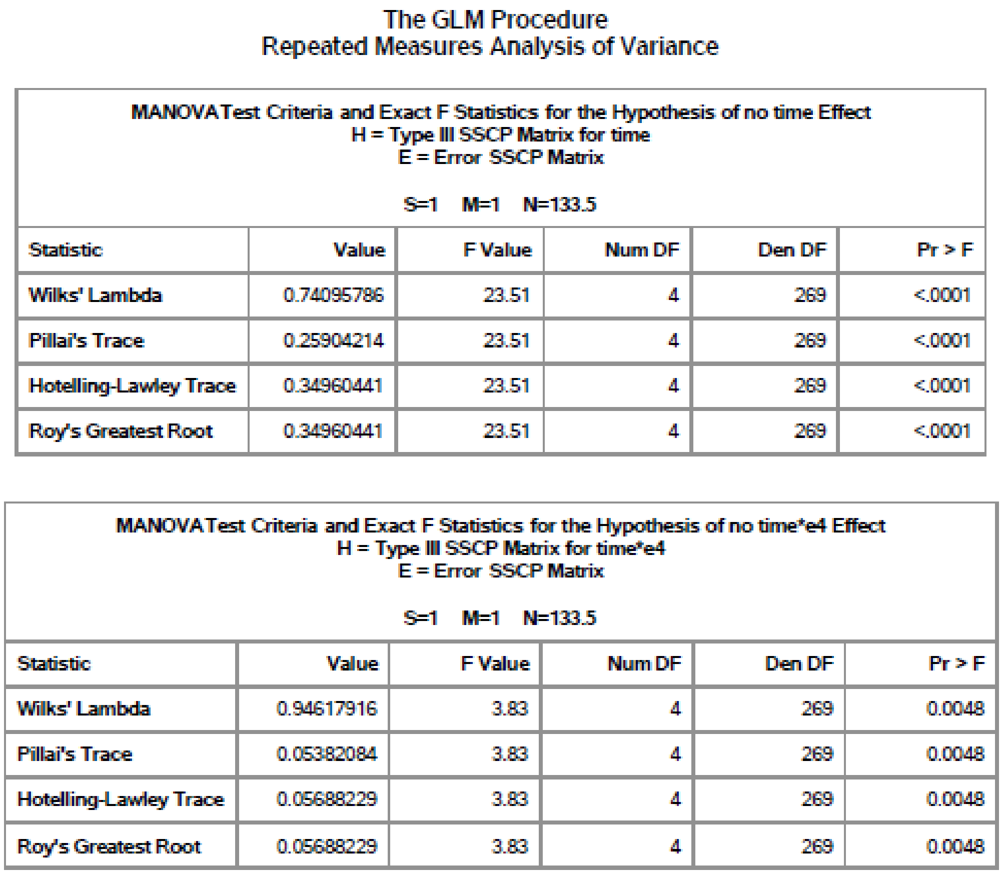
Make sure your levels of time match up with your outcomes



Sphericity Tests				
Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	9	0.1076619	602.6936	<.0001
Orthogonal Components	9	0.6714506	107.71098	<.0001

Results of sphericity tests: $p<0.05$ generally indicates violation of sphericity assumption

SAS Output – Within person Multivariate tests



The GLM Procedure
Repeated Measures Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no time Effect
H = Type III SSCP Matrix for time
E = Error SSCP Matrix

S=1 M=1 N=133.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.74095786	23.51	4	269	<.0001
Pillai's Trace	0.25904214	23.51	4	269	<.0001
Hotelling-Lawley Trace	0.34960441	23.51	4	269	<.0001
Roy's Greatest Root	0.34960441	23.51	4	269	<.0001

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no time*e4 Effect
H = Type III SSCP Matrix for time*e4
E = Error SSCP Matrix

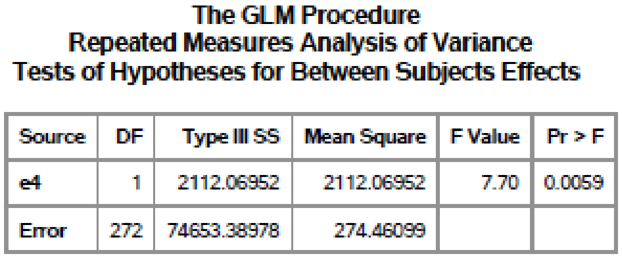
S=1 M=1 N=133.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.94617916	3.83	4	269	0.0048
Pillai's Trace	0.05382084	3.83	4	269	0.0048
Hotelling-Lawley Trace	0.05688229	3.83	4	269	0.0048
Roy's Greatest Root	0.05688229	3.83	4	269	0.0048

Time is significant

Time*E4 is significant

SAS output – Between-person effect and Univariate within-person tests

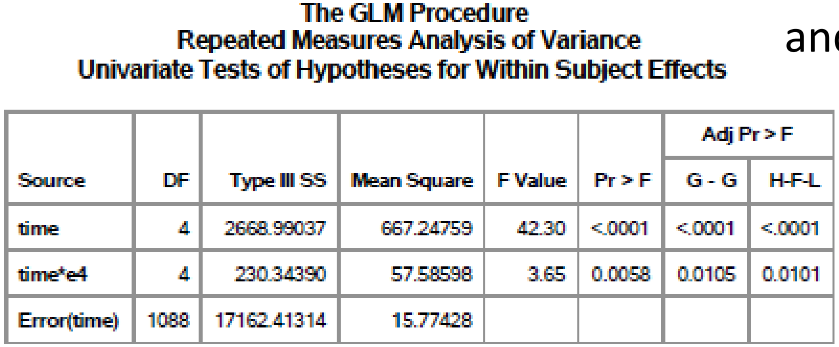


The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
e4	1	2112.06952	2112.06952	7.70	0.0059
Error	272	74653.38978	274.46099		

E4 is significant

Good idea to compare results from multivariate and univariate tests



The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

						Adj Pr > F	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	G - G	H-F-L
time	4	2668.99037	667.24759	42.30	<.0001	<.0001	<.0001
time*e4	4	230.34390	57.58598	3.65	0.0058	0.0105	0.0101
Error(time)	1088	17162.41314	15.77428				

Univariate tests of within person effects (matches output of proc mixed to be shown later)

Adjusted p-values account for violation of sphericity (Huynh-Feldt-Lecoutre (H-F-L) is generally preferred over Greenhouse-Geisser (G-G))

Mixed Effects Regression (Mixed Model): Notation

- Let Y_{ij} = outcome for i^{th} person, j^{th} measurement
- Let Y be a vector of all outcomes for all subjects
- X is a matrix of independent variables (such as E4 carrier or time)
- Z is a matrix associated with random effects

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 associated with the outcome
  - With longitudinal data, some coefficients (of time and interactions with time) will also tell us how variables are associated with change in the outcome
- **γ are the "random effects", $\gamma \sim N(0,\Sigma)$**
- **ε are the errors, $\varepsilon \sim N(0,R)$**
  - simple example: $R = \sigma^2$

Random Effects

- **Why use them?**
  - Not everybody responds the same way (even people with similar demographic and clinical information respond differently)
  - Want to allow for random differences in baseline level and possibly rate of change that remain unexplained by the covariates

Random Effects Cont.

- **Way to think about them**
  - Bins with numbers in them
  - Every person draws a number from each bin and carries those numbers with them
  - Predicted outcome based on "fixed effects" adjusted according to a person's random numbers
  - Similar to residuals (ε are residuals for each observation, while γ are residuals for person level data)

Random Effects Cont.

- **Accounts for correlation in observations**
- **Correlation structures**
  - Compound symmetry (common within-individual correlation)
    - Most common structure for repeated measures at the same visit
  - Autoregressive (AR)
    - Each assessment most strongly correlated with previous one
  - Unstructured (most flexible)

Assumptions of Model

- **Linearity**
- **Homoscedasticity (constant variance)**
- **Errors are normally distributed**
- **Random effects are normally distributed**
- **Typically assume Missing at Random (MAR)**
  - Missingness is statistically unrelated to the variable itself
  - May be related to other variables in data set

Determining best covariance structure

- Can compare models fit with different covariance structures
- Compare AIC and pick model with the smallest AIC
- Only valid when maximum likelihood is the method of estimation (in SAS, you must change the method, since the default is something different)
- We’ll see more in the example

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 change varies by one unit change in an independent variable
- **Covariance parameters**
  - Measure of between-person variability (random effects)
  - Measure of within-person variability (residual variance)

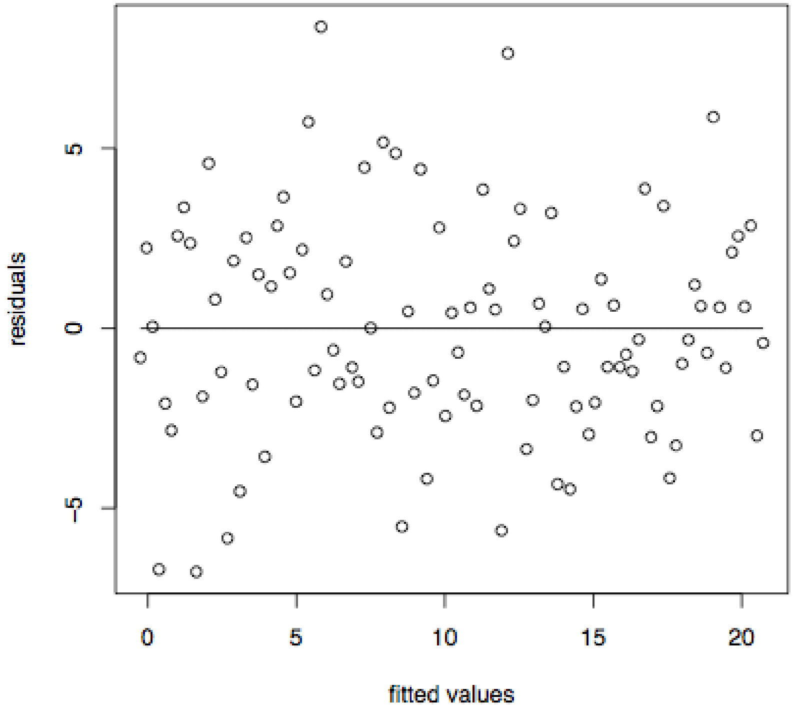
Graphical Tools for Checking Assumptions

- **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

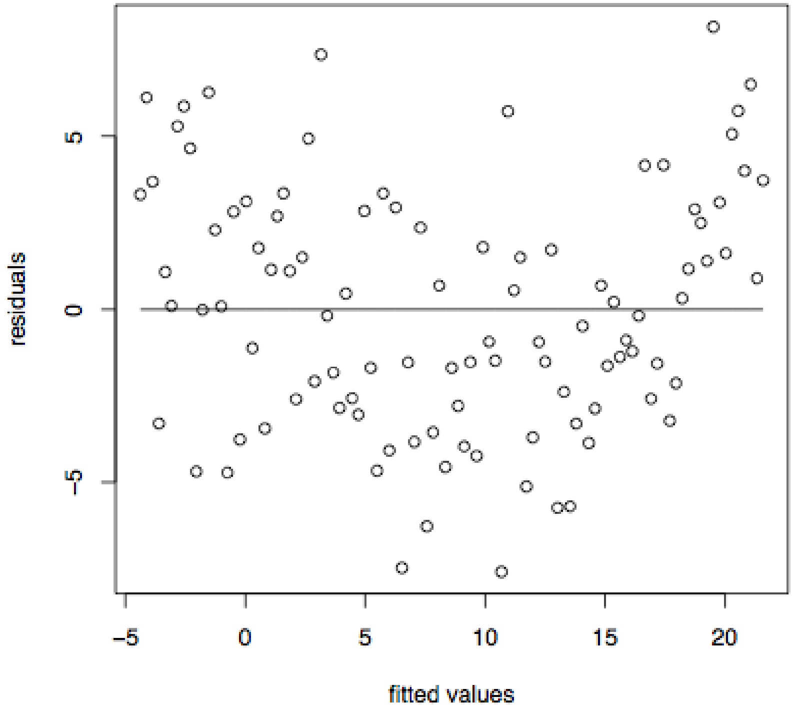
Ideal Residual Plot

- “cloud” of points
- no pattern
- evenly distributed about zero



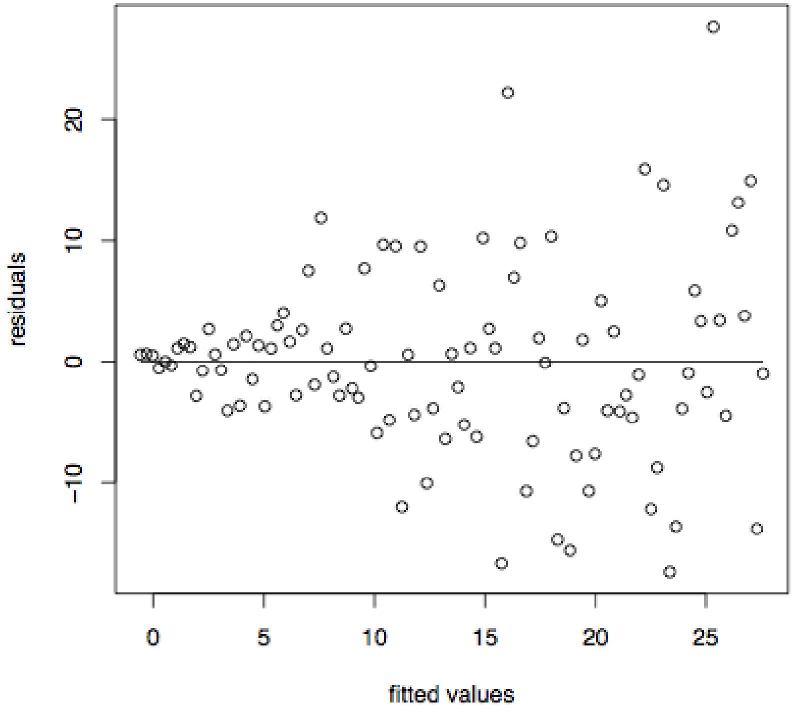
Non-linear relationship

- **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**

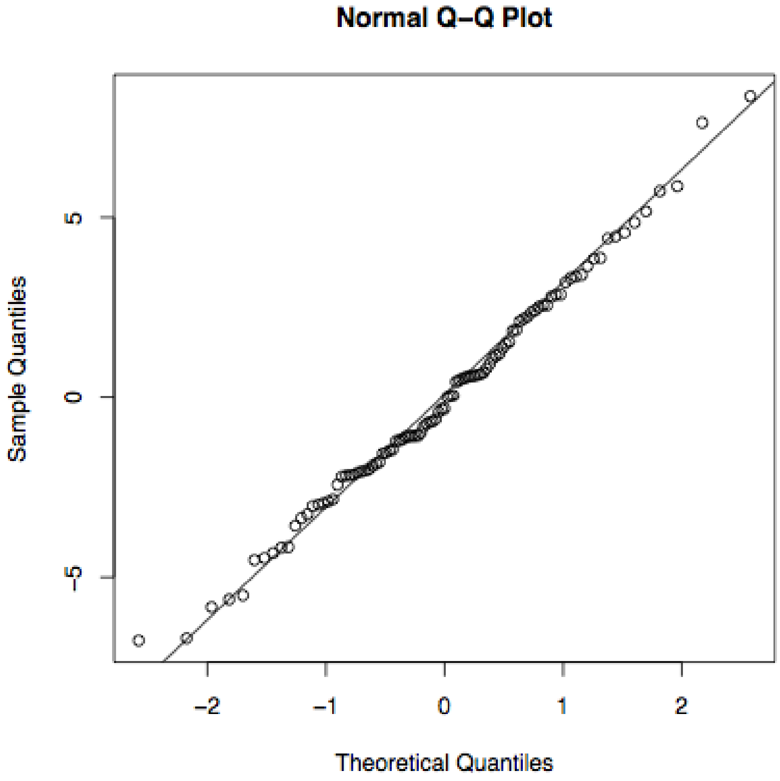


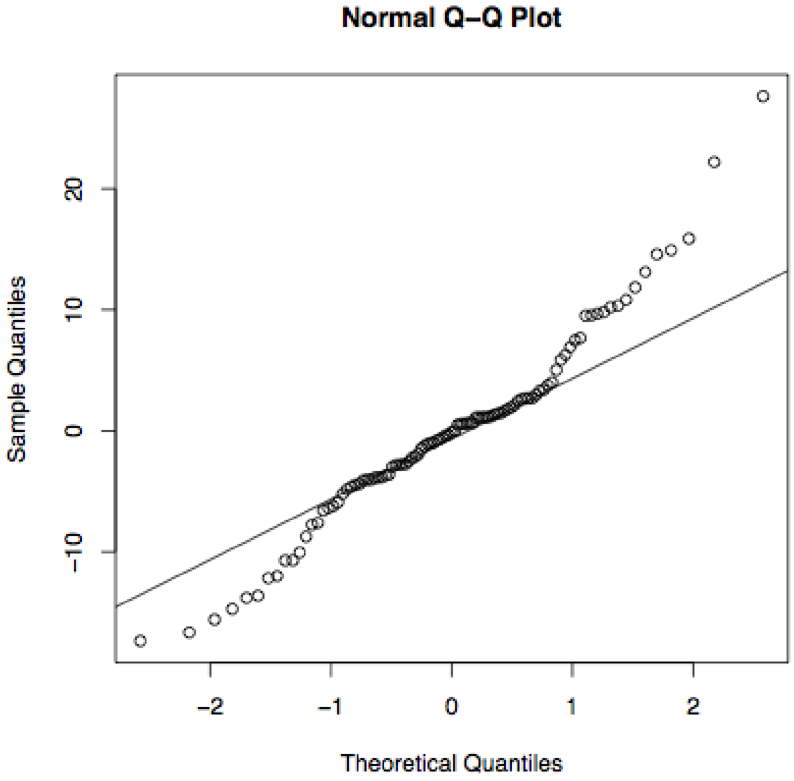
Non-constant variance

- 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)

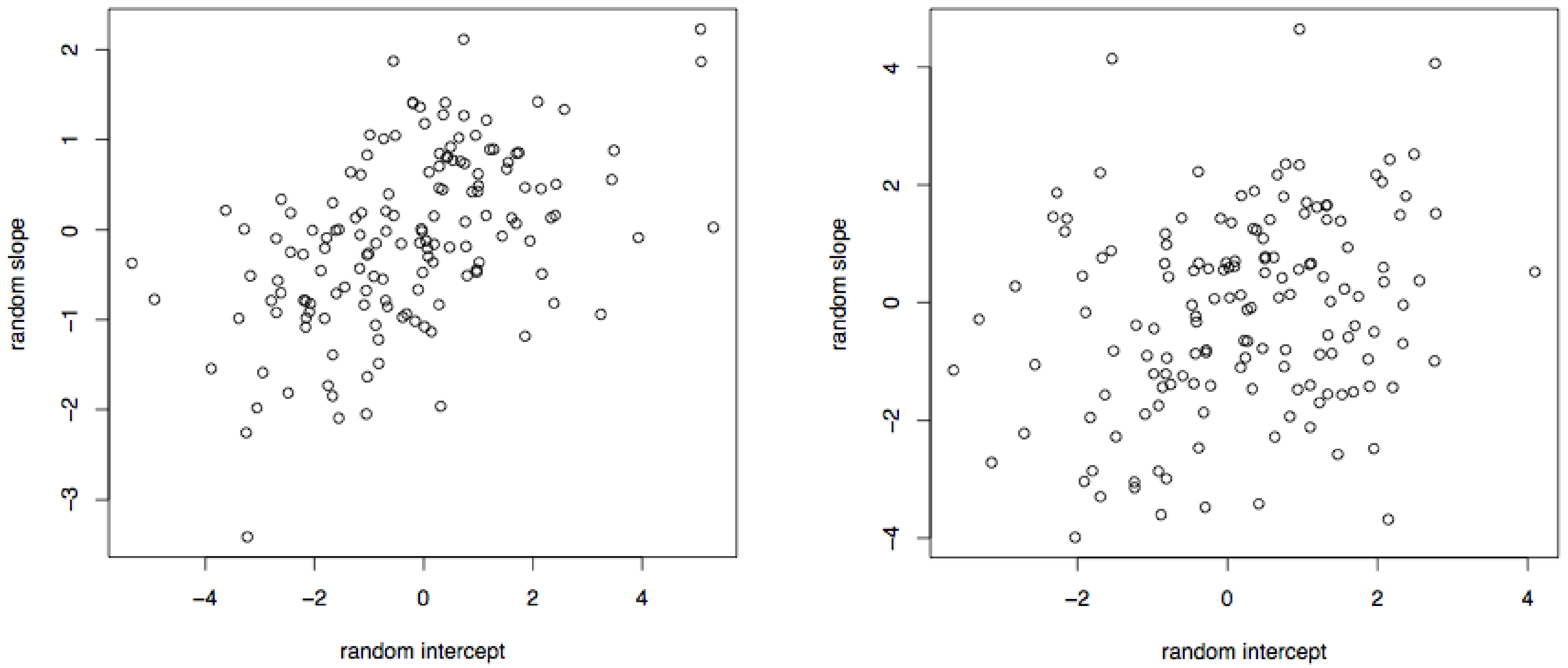


QQ-Plot


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



Mixed Effects Models in SAS

Specifies between-person covariance structure (unstructured here)

Options: reml (default), ml, mivque0

Random intercept and slope

```
proc mixed data=adni method=reml;
    class rid e4(ref=‘0’);
    model adas13=e4 time e4*time/s;
    random int time/sub=rid type=un g;
    repeated /sub=rid type=cs r;
run;
```

Requests estimates

ID variable

Specifies within-person covariance structure (compound symmetry)

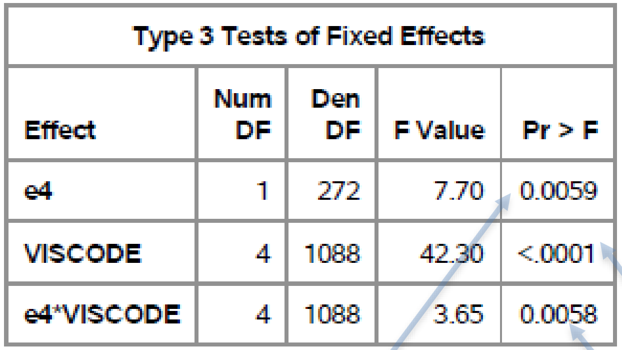
Data Analysis Example: ADNI Standard Repeated Measures ANOVA (similar to earlier results)

```
proc mixed data=adni plots=all;
   class rid e4(ref=‘0’) viscode(ref=‘bl’);
   model adas13=e4 viscode e4*viscode/s;
   repeated viscode/sub=rid type=cs r;
run;
```

(only uses a repeated statement)

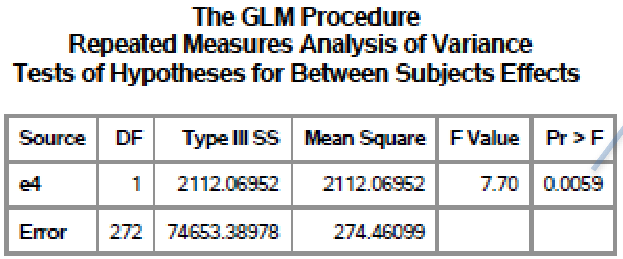
Repeated Measures ANOVA output

proc mixed:



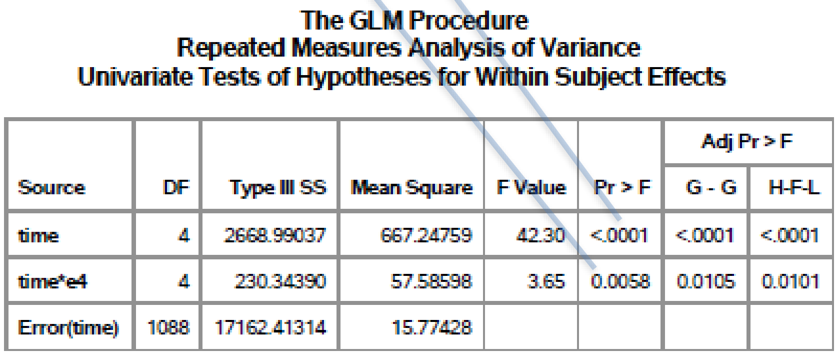
Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
e4	1	272	7.70	0.0059
VISCODE	4	1088	42.30	<.0001
e4*VISCODE	4	1088	3.65	0.0058

proc glm:



The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
e4	1	2112.06952	2112.06952	7.70	0.0059
Error	272	74653.38978	274.46099		



The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

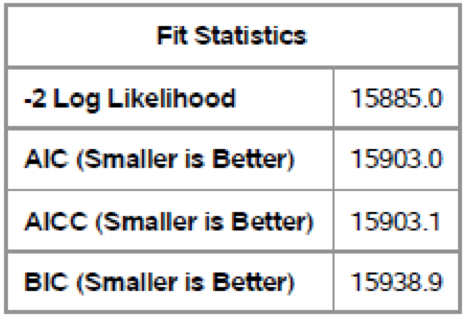
						Adj Pr > F	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	G - G	H-F-L
time	4	2668.99037	667.24759	42.30	<.0001	<.0001	<.0001
time*e4	4	230.34390	57.58598	3.65	0.0058	0.0105	0.0101
Error(time)	1088	17162.41314	15.77428				

Data Analysis (Continuous time)

- Now want to use all available data, even if individuals are missing some visits
- Use time since baseline as a continuous time measure (to further account for differences in when specific visits happened)

Picking Covariance Structure

```
proc mixed data=adni method=ML;
   class rid e4(ref=‘0’);
   model adas13=e4 time e4*time/s;
   random int time/sub=rid type=un g;
   repeated /sub=rid type=ar(1) r;
run;
```



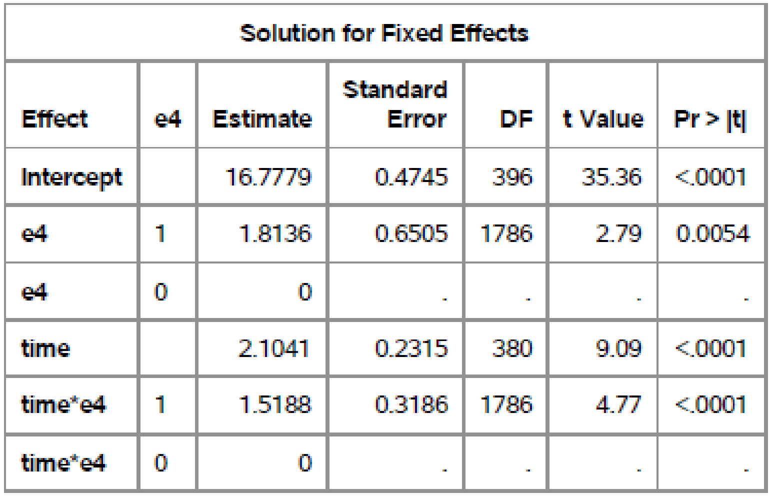
Fit Statistics	
-2 Log Likelihood	15885.0
AIC (Smaller is Better)	15903.0
AICC (Smaller is Better)	15903.1
BIC (Smaller is Better)	15938.9

Random Int	Random Slope	Repeated Statement	G-structure	R-structure	AIC
Y	N	N	CS	-	17315.2
Y	Y	N	CS	-	16095.4
Y	Y	N	AR(1)	-	16095.4
Y	Y	N	UN	-	15952.3
Y	Y	Y	UN	CS	15954.3
Y	Y	Y	UN	AR(1)	15903.0

Mixed Model Output

At time=0 (study start), E4 non-carriers have an ADAS13 score of 16.8 on average

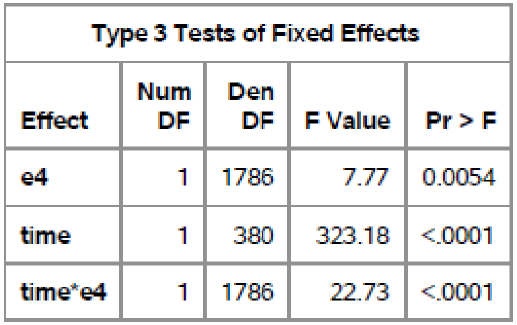
E4 carriers start 1.8 points higher



Solution for Fixed Effects						
Effect	e4	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		16.7779	0.4745	396	35.36	<.0001
e4	1	1.8136	0.6505	1786	2.79	0.0054
e4	0	0	.	.	.	.
time		2.1041	0.2315	380	9.09	<.0001
time*e4	1	1.5188	0.3186	1786	4.77	<.0001
time*e4	0	0	.	.	.	.

Non-carriers are increasing at 2.1 points per year

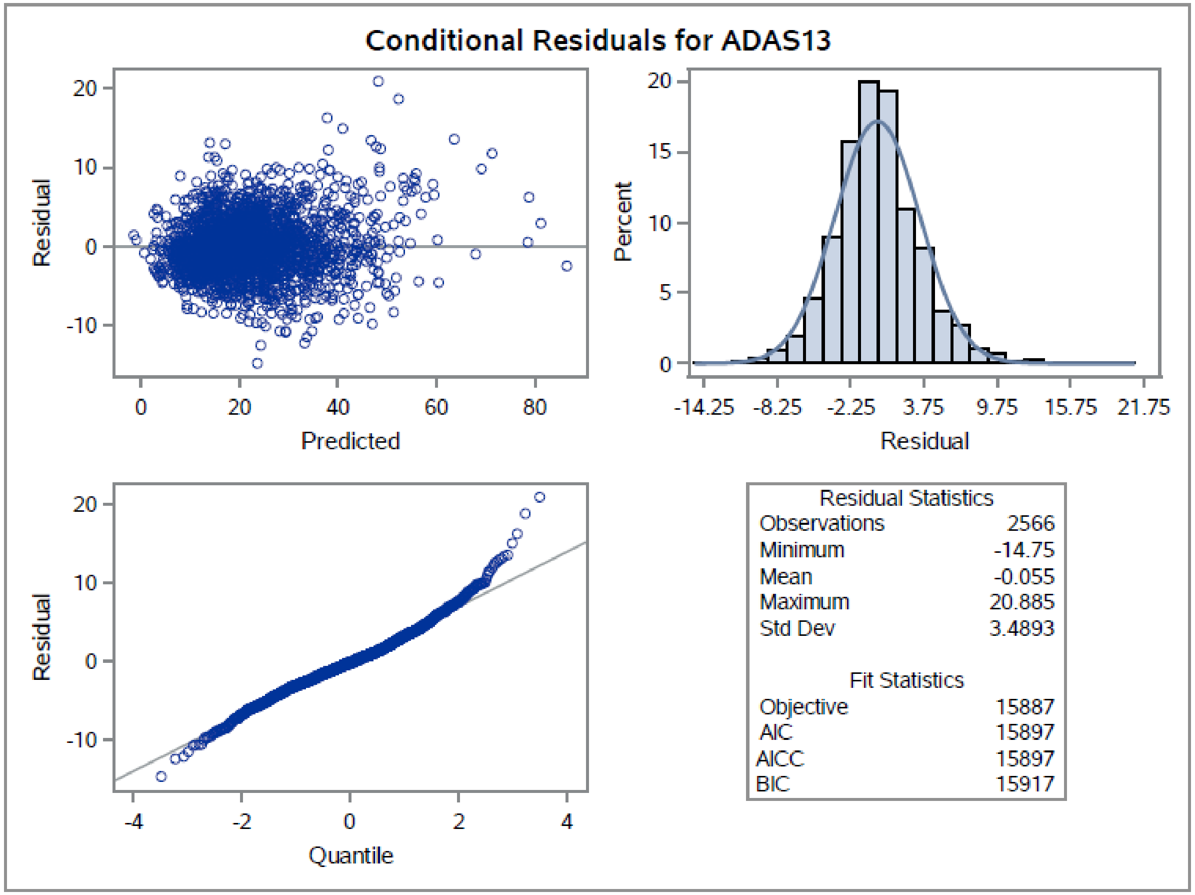
E4 carriers are increasing an additional 1.5 points per year (annual increase is 2.1+1.5=3.6)



Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
e4	1	1786	7.77	0.0054
time	1	380	323.18	<.0001
time*e4	1	1786	22.73	<.0001

Overall test of significance for each term in the model

Some Diagnostics



Residual Statistics	
Observations	2566
Minimum	-14.75
Mean	-0.055
Maximum	20.885
Std Dev	3.4893

Fit Statistics	
Objective	15887
AIC	15897
AICC	15897
BIC	15917

Advanced topics

- **Non-normal data**
  - Generalized Estimating Equations (GEE)
  - Repeated measures models for binary, ordinal, and count data
- **Time-varying covariates**
- **Simultaneous growth models (modeling two types of longitudinal outcomes together)**
  - Allows you to directly compare associations of specific independent variables with the different outcomes
  - Allows you to estimate the correlation between change in the two processes

Summary

- Longitudinal studies often result in repeated assessments on individuals
- Repeated measures ANOVA and mixed effects regression models are main strategies for analysis
- Mixed models can be more flexible than standard repeated measures ANOVA models
- SAS can fit both types of models

Help is Available

- **CTSC Biostatistics Office Hours**
  - Every Tuesday from 12 – 1:30 in Sacramento
  - Sign-up through the CTSC Biostatistics Website
- **EHS Biostatistics Office Hours**
  - Every Monday from 2-4 in Davis
- **Request Biostatistics Consultations**
  - CTSC - www.ucdmc.ucdavis.edu/ctsc/
  - MIND IDDRC - www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
  - Cancer Center and EHS Center