

Repeated Measures Data

Clustered/Repeated

- Data is sampled in some unit
- Within that unit multiple measures
 - clinic: people
 - household: family member
 - patients: measured values
- Clinic/Households: clustered data
- Repeated measures: longitudinal data

Data is Different

- Used to assumption of independence -- comes from an idea of sampling units
- Multiple measures per unit, likely correlated
- Need to keep account of this correlation
- May even need to model this correlation
simpler for clustered than longitudinal

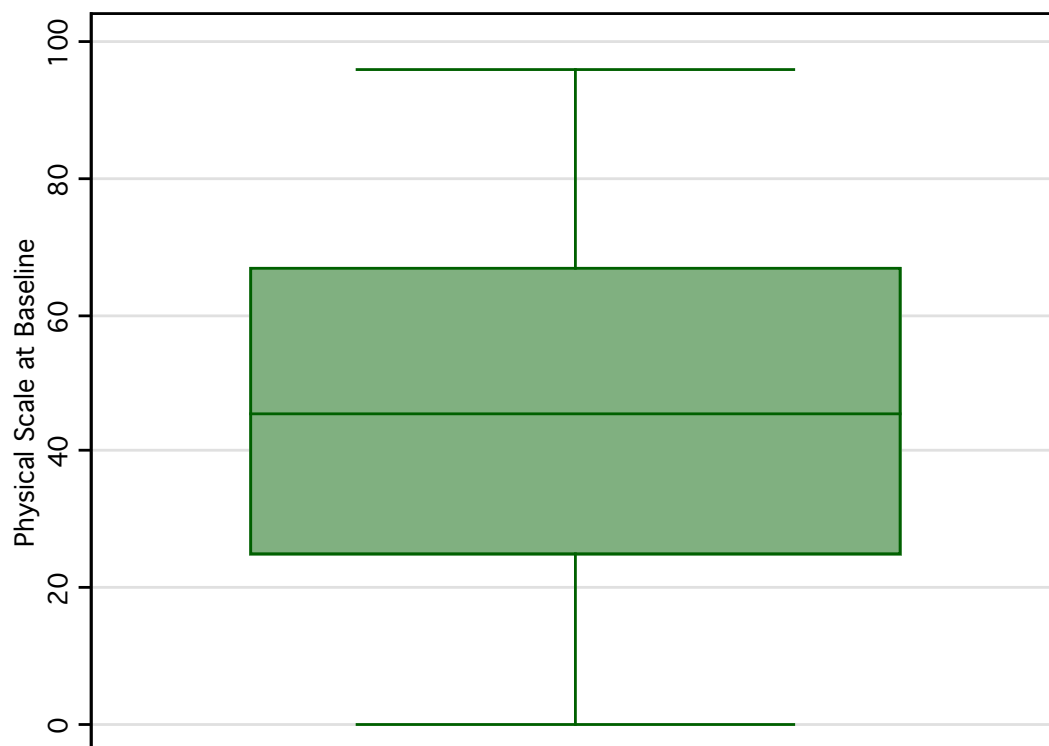
Longitudinal Data

- Measure a series of factors in a group of subjects over a defined follow-up period
- Example: AIDS Care Study in Uganda
- HIV+ pts followed after initiation of anti-retroviral therapy (ARV)
- Outcomes: CD4, viral load, scales of function and well-being

Sample Question

- Outcome: scale of physical well-being ranging from 0-100
- How does it change after ARV initiation?
- Is the change affected by baseline CD4 count? Is it affected by baseline depression?

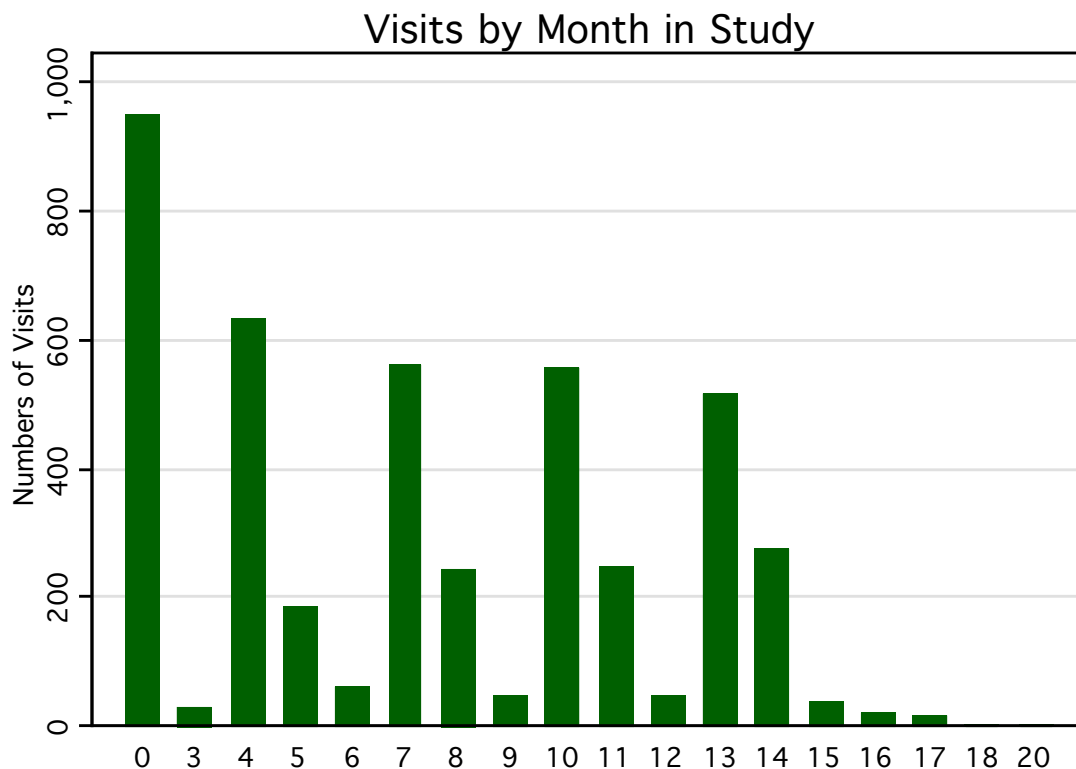
Boxplot of Baseline Physical Scale



Features of Longitudinal Data

- Varying follow-up periods
- Measures not always precisely timed
- Correlation within subject over time

I will focus primarily on continuous outcomes
most issues are largely similar for binary or count
responses



Correlated Measures Across Time*

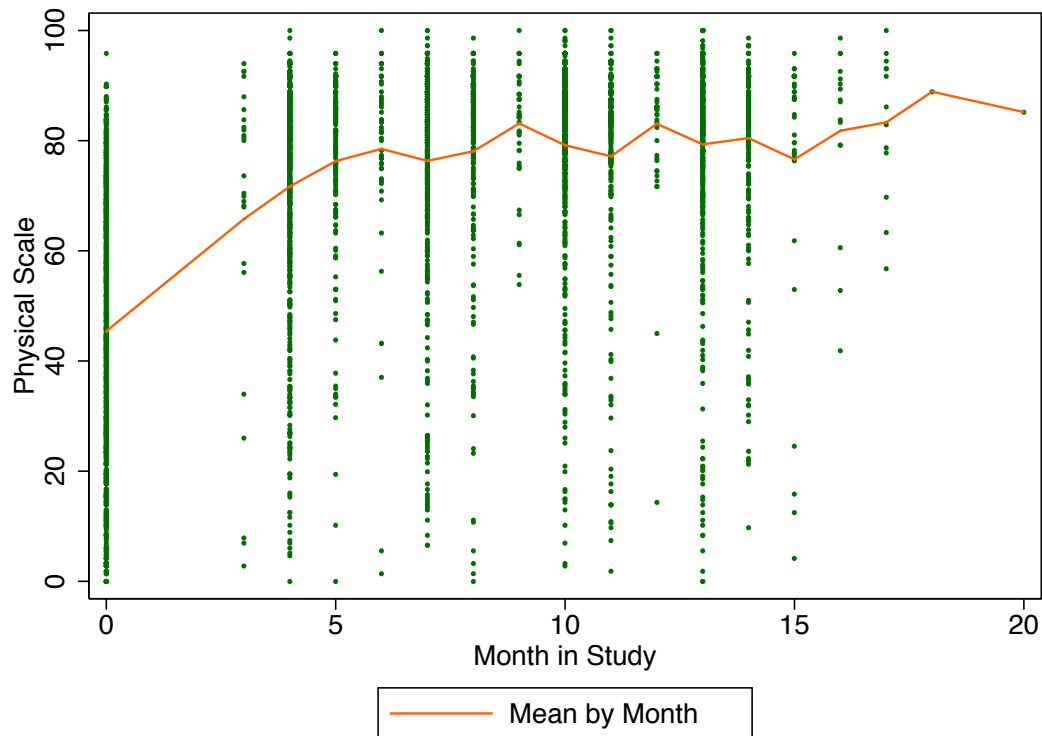
	physic~0	physic~1	physic~2	physic~3	physic~4
physical0	1.0000				
physical1	0.3776	1.0000			
physical2	0.3063	0.5167	1.0000		
physical3	0.2283	0.4159	0.5757	1.0000	
physical4	0.2257	0.3218	0.3573	0.4851	1.0000

* Grouped by quarter in the study

Modeling Longitudinal Data

- Two parts
- Mean Structure: What factors affect the average value (time? predictors? both?)
- Correlation structure: How does correlation change over time?
- First step.... explore the data to inform modeling

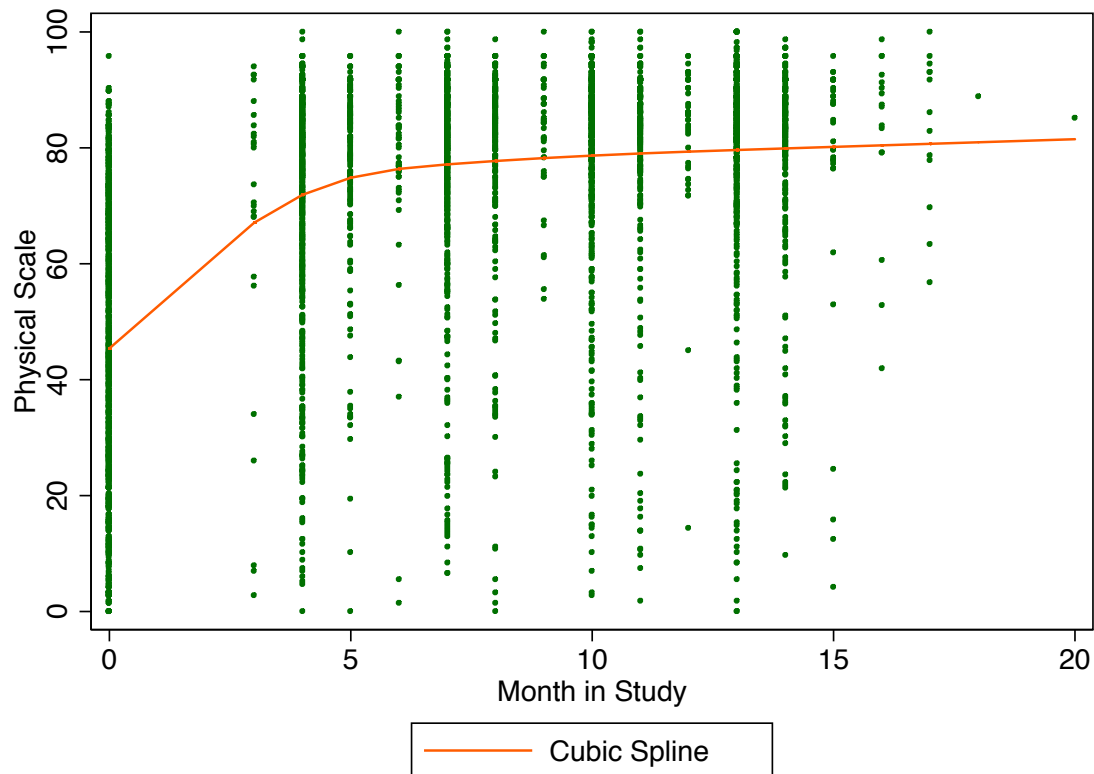
Mean of Physical Scale by Month



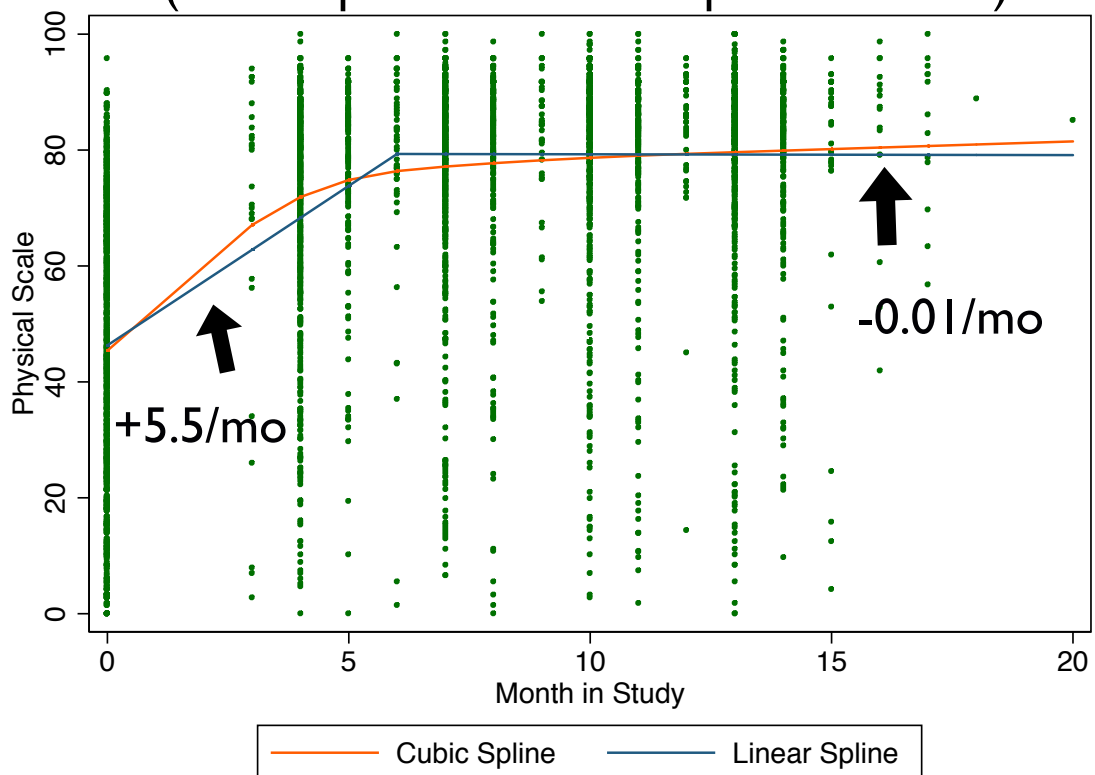
Cubic Splines

- Exploratory technique
- Fits a smooth curve through a cloud of points
- Flexible: doesn't have to make a straight line

Mean of Physical Scale by Month
(Cubic Spline Fit)



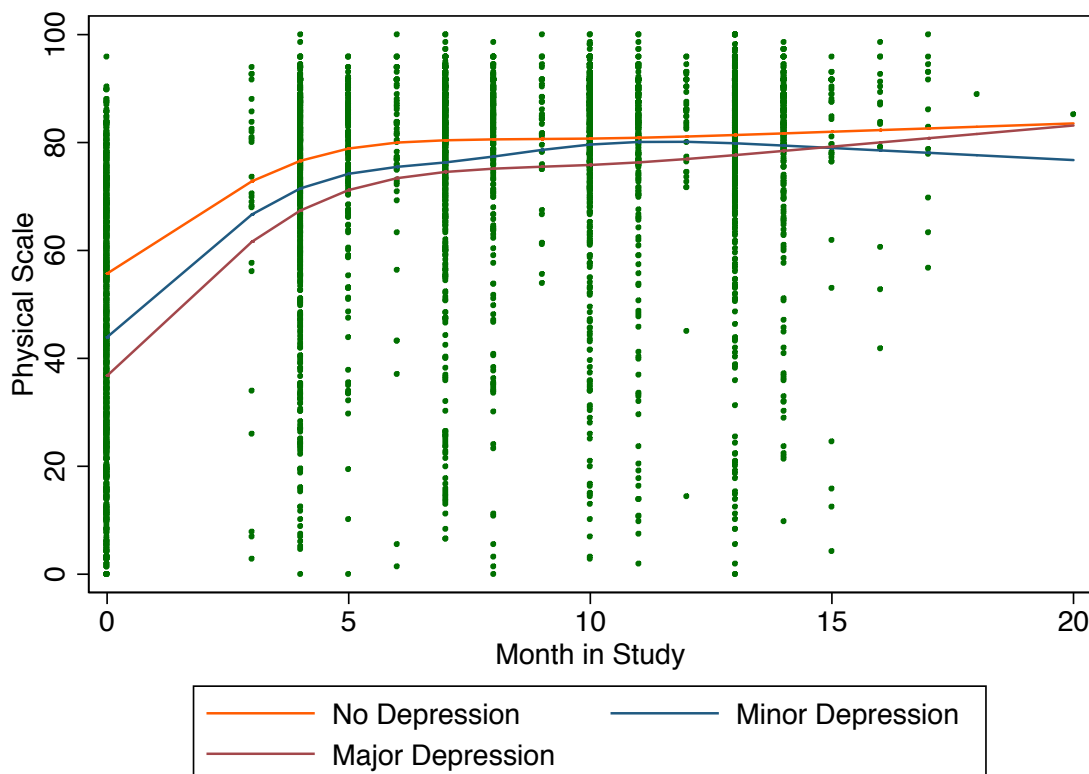
Mean of Physical Scale by Month
(Cubic Spline and Linear Spline at 6mos.)



Depression and Physical Functioning

- Used baseline CESD instrument
> 27, 16 to 26, <16
- Depression: major, minor, none
- How does baseline depression affect physical function after ARV treatment?
- Explore this question before heavy modeling

Mean of Physical Scale by Month by Baseline Depression (Cubic Spline)



Information for Modeling the Mean*

- The effect of depression varies with time
(the effect of time varies by depression)
interaction between time and depression
- The trajectories are non-linear
ouch!
need to treat time as a factor (grouping) or
use a flexible method (e.g., splines)

*no matter what technique you use

Very Simple Analysis before treatment

```
reg phys i.depress if month==0
```

Source	SS	df	MS	Number of obs = 913		
Model	59947.1204	2	29973.5602	F(2, 910) = 58.06		
Residual	469754.145	910	516.213346	Prob > F = 0.0000		
Total	529701.265	912	580.812791	R-squared = 0.1132		
				Adj R-squared = 0.1112		
				Root MSE = 22.72		

physical	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
depressed						
1	-11.87678	1.90932	-6.22	0.000	-15.62396	-8.1296
2	-18.9506	1.770853	-10.70	0.000	-22.42603	-15.47516
_cons	55.75984	1.270105	43.90	0.000	53.26716	58.25251

Very Simple Analysis

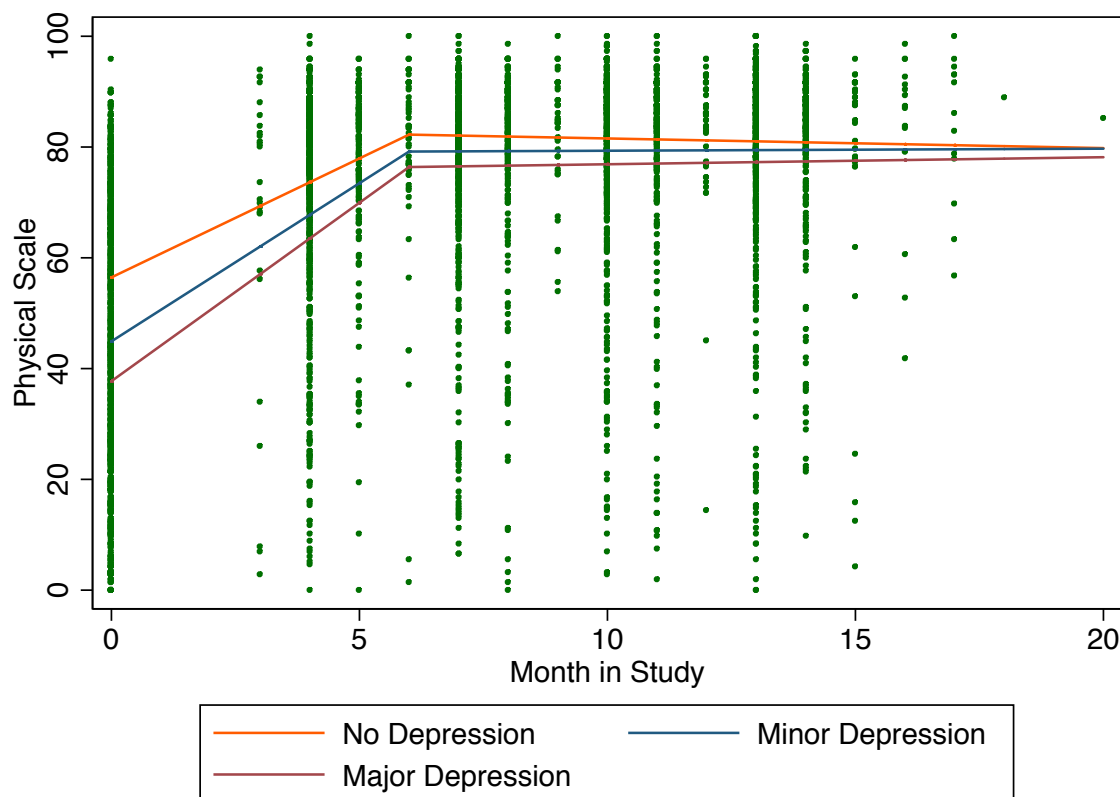
4 months after treatment

```
. reg phys i.depress if month==4
```

Source	SS	df	MS	Number of obs =	826
Model	1266.98873	2	633.494367	F(2, 823) =	2.10
Residual	247973.601	823	301.304497	Prob > F =	0.1228
Total	249240.59	825	302.109806	R-squared =	0.0051
				Adj R-squared =	0.0027
				Root MSE =	17.358

physical	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
depressed					
1	-.7459929	1.523518	-0.49	0.625	-3.736432 2.244446
2	-2.829356	1.422033	-1.99	0.047	-5.620594 -.0381176
_cons	81.12603	1.005529	80.68	0.000	79.15232 83.09973

Mean of Physical Scale by Month by Baseline Depression (Spline Fit)



Spline Output

Not Depressed

physical	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
before 6mos.	4.296064	.2365165	18.16	0.000	3.832137	4.759991
after 6mos.	-.1736788	.1886747	-0.92	0.357	-.5437642	.1964066
baseline	56.46386	.9602246	58.80	0.000	54.58038	58.34734

Mean Score at Baseline = 56.5

First 6 months, average change of +4.3 per month

After 6 months, average change of -0.17 per month

Spline Output

Mildly Depressed

physical	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
before 6mos.	5.711129	.2902105	19.68	0.000	5.141758	6.2805
after 6mos.	.037069	.2326473	0.16	0.873	-.4193671	.4935052
baseline	44.91511	1.175242	38.22	0.000	42.60937	47.22084

Mean Score at Baseline = 44.9

First 6 months, average change of +5.7 per month

After 6 months, average change of +0.04 per month

Spline Output

Major Depression

physical	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
before 6mos.	6.442615	.2775006	23.22	0.000	5.898299	6.986931
after 6mos.	.1269116	.2258695	0.56	0.574	-.3161302	.5699534
baseline	37.72776	1.113948	33.87	0.000	35.54276	39.91277

Mean Score at Baseline = 37.7

First 6 months, average change of +6.4 per month

After 6 months, average change of +0.13 per month

One Problem

- That analysis didn't take into account the correlation

	physic~0	physic~1	physic~2	physic~3	physic~4
physical0	1.0000				
physical1	0.3776	1.0000			
physical2	0.3063	0.5167	1.0000		
physical3	0.2283	0.4159	0.5757	1.0000	
physical4	0.2257	0.3218	0.3573	0.4851	1.0000

Modeling Correlation

- Correlation=0: **independence**
- Correlations all equal: **exchangeable**
- Correlations diminish steady as farther apart in time: **AR(1)**
- Need some sense of correlation in one's data

Next Step

- A bonefide longitudinal model
- Two major options:
 - Mixed effects models
 - Generalized estimating equations

Two major modeling strategies

- **Random effects:** requires explicit model for correlation, needs to be correct for full validity
- **GEE:** required to nominate model for correlation, doesn't need to be right to be valid if close, standard errors will be smaller

Other differences arise if data is binary or counts

For a GEE model

- Need: “link”. This describes how predictors related to outcome. If data continuous, link is almost always “identity”. If binary, almost always “logit”
- Need distribution family: if cts, usually “Normal”; if binary, should be binomial

For a GEE model, cont.

- Select correlation structure: independence, exchangeable, AR(1)
- Pick one closest to data or independence is fine
- Request “robust” standard errors ensures validity if your guess was wrong

For this data

- Link: identity, Family: normal
- Use the mean structure given before interaction between depression & time
- Nominated correlation: independence
- Spline with a “knot” at 6 months

Stata Command & Output: Depressed Pts.

```
xtgee phys predictors if dep==1, family(normal)
link(identity) i(idno) t(mo) corr(indep) robust
```

Note coefficients have not changed,
only the std. errs., Z's & CIs

(Std. Err. adjusted for clustering on idno)

physical	Semirobust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
before 6mos.	6.442615	.242425	26.58	0.000	5.967471	6.917759
after 6mos.	.1269116	.1827468	0.69	0.487	-.2312656	.4850889
baseline	37.72776	1.301259	28.99	0.000	35.17734	40.27819

Better Model

```
xtgee phys c.mosp1##dep c.mosp2##dep, family(normal) link(identity) i(idno) t(mo) corr(indep) robust
```

(Std. Err. adjusted for clustering on idno)

physical	Semirobust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
mosp1	4.296064	.2354781	18.24	0.000	3.834535	4.757593
depressed						
1	-11.54875	1.80214	-6.41	0.000	-15.08088	-8.01662
2	-18.73609	1.730903	-10.82	0.000	-22.1286	-15.34358
depressed#						
c.mosp1						
1	1.415065	.3813247	3.71	0.000	.6676825	2.162448
2	2.146551	.3378082	6.35	0.000	1.484459	2.808643
mosp2	-.1736788	.1563033	-1.11	0.266	-.4800277	.1326701
depressed#						
c.mosp2						
1	.2107479	.2547885	0.83	0.408	-.2886284	.7101242
2	.3005905	.2403482	1.25	0.211	-.1704833	.7716642
_cons	56.46386	1.142711	49.41	0.000	54.22418	58.70353

Some Thoughts

- Two levels of modeling for longitudinal data: means and correlations
- Means can be hard to model and require careful exploration
- GEE is a useful technique and allows you to avoid needing to model correlation
- That is all it does for you...but that's good

Mixed Modeling

- More parametric approach
- Explicitly models the correlation
- Modeling correlation requires knowledge of type of models

Correlation

Exchangeable Correlation

	physic~0	physic~1	physic~2	physic~3	physic~4
physical0	1.0000				
physical1	0.4000	1.0000			
physical2	0.4000	0.4000	1.0000		
physical3	0.4000	0.4000	0.4000	1.0000	
physical4	0.4000	0.4000	0.4000	0.4000	1.0000

Correlation is same across any 2 timepoints

Single correlation parameter -- $\rho = 0.40$

Not usually plausible for longitudinal data

Random effect is a way to induce this

Correlation

Autoregressive I - (AR1) Correlation

	physic~0	physic~1	physic~2	physic~3	physic~4
physical0	1.0000				
physical1	0.5000	1.0000			
physical2	0.2500	0.5000	1.0000		
physical3	0.1250	0.2500	0.5000	1.0000	
physical4	0.0625	0.1250	0.2500	0.5000	1.0000

Correlation for timepoints adjacent = 0.50

timepoints 1 apart = 0.50×0.50

timepoints 2 apart = $0.50 \times 0.50 \times 0.50$

More plausible for longitudinal data -- if
measures at equally spaced timepoints

Correlation

Unstructured Correlation

	physic~0	physic~1	physic~2	physic~3	physic~4
physical0	1.0000				
physical1	0.3776	1.0000			
physical2	0.3063	0.5167	1.0000		
physical3	0.2283	0.4159	0.5757	1.0000	
physical4	0.2257	0.3218	0.3573	0.4851	1.0000

No model for the correlation -- totally free

Can take long time to converge in Stata

Or fail to converge -- requires lots of parameters

Mixed Model -- AR(1)

xtmixed phys c.mosp1##dep c.mosp2##dep || idno:, residual(ar 1, t(mo))

physical	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
mosp1	4.336866	.225644	19.22	0.000	3.894612 4.77912
depressed					
1	-11.51941	1.639709	-7.03	0.000	-14.73318 -8.305645
2	-18.62879	1.52096	-12.25	0.000	-21.60982 -15.64776
depressed#					
c.mosp1					
1	1.314409	.3404492	3.86	0.000	.6471413 1.981678
2	1.969806	.3181863	6.19	0.000	1.346172 2.593439
mosp2	-.1822822	.1858448	-0.98	0.327	-.5465314 .181967
depressed#					
c.mosp2					
1	.2369709	.2808057	0.84	0.399	-.3133982 .7873399
2	.413691	.2635603	1.57	0.117	-.1028776 .9302596
_cons	56.16341	1.090842	51.49	0.000	54.0254 58.30142

Mixed Model -- AR(1)

```
xtmixed phys c.mosp1##dep c.mosp2##dep || idno:, residual(ar 1, t(mo))
```

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
idno: Identity					
	sd(_cons)	9.622091	.6090749	8.499408	10.89307
Residual: AR(1)					
	rho	.7073725	.0188279	.6685106	.7423818
	sd(e)	17.21154	.3314007	16.57411	17.87348
LR test vs. linear regression:		chi2(2) =	816.77	Prob > chi2 = 0.0000	

Correlation 1 month apart = 0.71

Correlation 2 months apart = $0.71 * 0.71 = 0.50$

Correlation 3 months apart = $0.71^3 = 0.35$

Mixed Model -- Exch

```
xtmixed phys c.mosp1##dep c.mosp2##dep || idno:,
```

physical	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
mosp1	4.296664	.2105279	20.41	0.000	3.884037	4.709291
depressed						
1	-11.56244	1.60663	-7.20	0.000	-14.71137	-8.413502
2	-18.55899	1.490839	-12.45	0.000	-21.48098	-15.637
depressed#						
c.mosp1						
1	1.351961	.3176372	4.26	0.000	.7294035	1.974518
2	1.964092	.2966764	6.62	0.000	1.382617	2.545567
mosp2	-.2061081	.1689812	-1.22	0.223	-.5373052	.125089
depressed#						
c.mosp2						
1	.2122151	.2552857	0.83	0.406	-.2881356	.7125658
2	.3355955	.2396816	1.40	0.161	-.1341719	.8053629
_cons	56.41977	1.068939	52.78	0.000	54.32469	58.51485

Mixed Model -- Exch

```
xtmixed phys c.mosp1##dep c.mosp2##dep || idno;
```

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
idno: Identity					
	sd(_cons)	11.73131	.3963987	10.97955	12.53454
	sd(Residual)	15.65877	.1920512	15.28685	16.03975

$$\text{Correlation} = 11.7^2 / (11.7^2 + 15.7^2) = 0.36$$

Random Effects

- Easy way to specify a correlation model
- Like a unknown shared predictor
 - for a person -- person level covariate
 - for a clinic -- clinic level covariate
- Induces a covariance structure

Example of Random Effects

- Suppose people selected within clinics and clinic is treated as random effect
- `xtmixed phys predictors || clinic: || idno:, residual(ar 1, t(mo))`
- Random effects for clinic and AR1 for time

GEE v. Mixed Models

- Longitudinal data models have two levels
 - model for mean structure
 - model for correlation structure
- GEE and Mixed models differ in the second part
- GEE allows you to pay little attention to the correlation structure

GEE

- With GEE you nominate a correlation structure
- GEE valid if correlation is wrong you can even assume independence
- Recovers validity from “robust” standard error
- If # clusters < 10 , robust SE is not great -- need to be cautious

Mixed Models

- Normal based model -- more parametric
- Has a formal model for dependence -- should be approximately correct
- Can handle multiple layers of clustering *very complicated structures*
- Can answer questions about clustering
- Behaves better in certain missing data situations (discussed in coming weeks)

Bottom Line

- If you have >1 level of clustering
- Have questions which involve clustering pretty rare
- Desire subject specific interpretation not common
- Then, you should seek about mixed models
- Modeling correlation and specifying it is not easy

Remember

- Pay careful attention to modeling means interactions with time are not uncommon
- Effects often depend on time in a complex way
- Choose those interactions judiciously
- GEE doesn't really mitigate these issues *still have to model mean carefully*