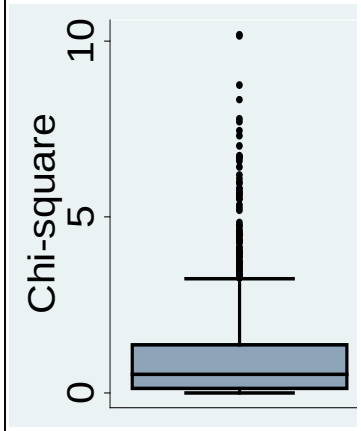
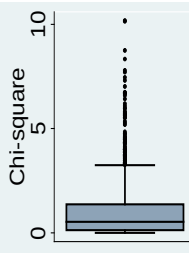
Repeated Measures, Lecture 2, Part 1

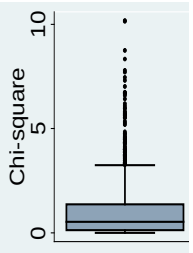
Charles E. McCulloch,
Division of Biostatistics,
Dept of Epidemiology and Biostatistics,
UCSF





Outline

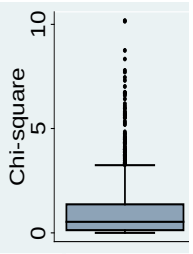
1. Two overarching analysis approaches to accommodating correlation: Generalized estimating equations (GEE) and mixed models.
2. Longitudinal designs
3. Lecture 2 covers numerical outcomes, so the analogs of linear regression.
4. Part 1: GEE (`xtgee`) methods
5. Summary



Analyses for correlated data

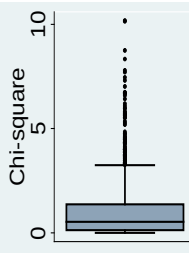
- Generalized estimating equations (GEEs)
  - Stata `xtgee`
- Mixed effects (ME) models
  - Stata `meglm`
  - Or specialized routines for different outcome types
    - Numeric (`mixed`)
    - Binary (`melogit, meprobit`)
    - Count (`mepoisson`)
    - Ordered categorical (`meologit, meoprobit`)
- *Interpretations are virtually the same for corresponding correlated and non-correlated analyses (linear/linear, logit/logit)*

Analyses for correlated data

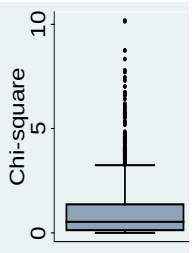


- Generalized estimating equations (GEEs)
  - Only works for one level of clustering (e.g., patients).
  - Requires largish number of clusters (ideally > 50) and works best when the number of clusters >> number of observations per cluster.
  - Does not require modeling of the correlation when using “robust” standard errors.
  - No overall measure of model fit like likelihood or R^2.
  - Less robust than mixed effects models to possible bias due to missing data or dropout.

Analyses for correlated data



- Mixed models
  - Can handle a wider variety of data structures. For example times within patients within doctors or compartments within knees within time within subjects.
  - Requires modeling the correlation.
  - Gives additional information about the correlation.
  - Based on likelihood fits, so can conduct likelihood ratio tests to compare two model fits.
  - More robust to potential bias due to missing data and dropout.

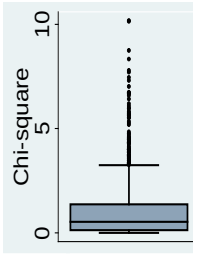


Consider first longitudinal data

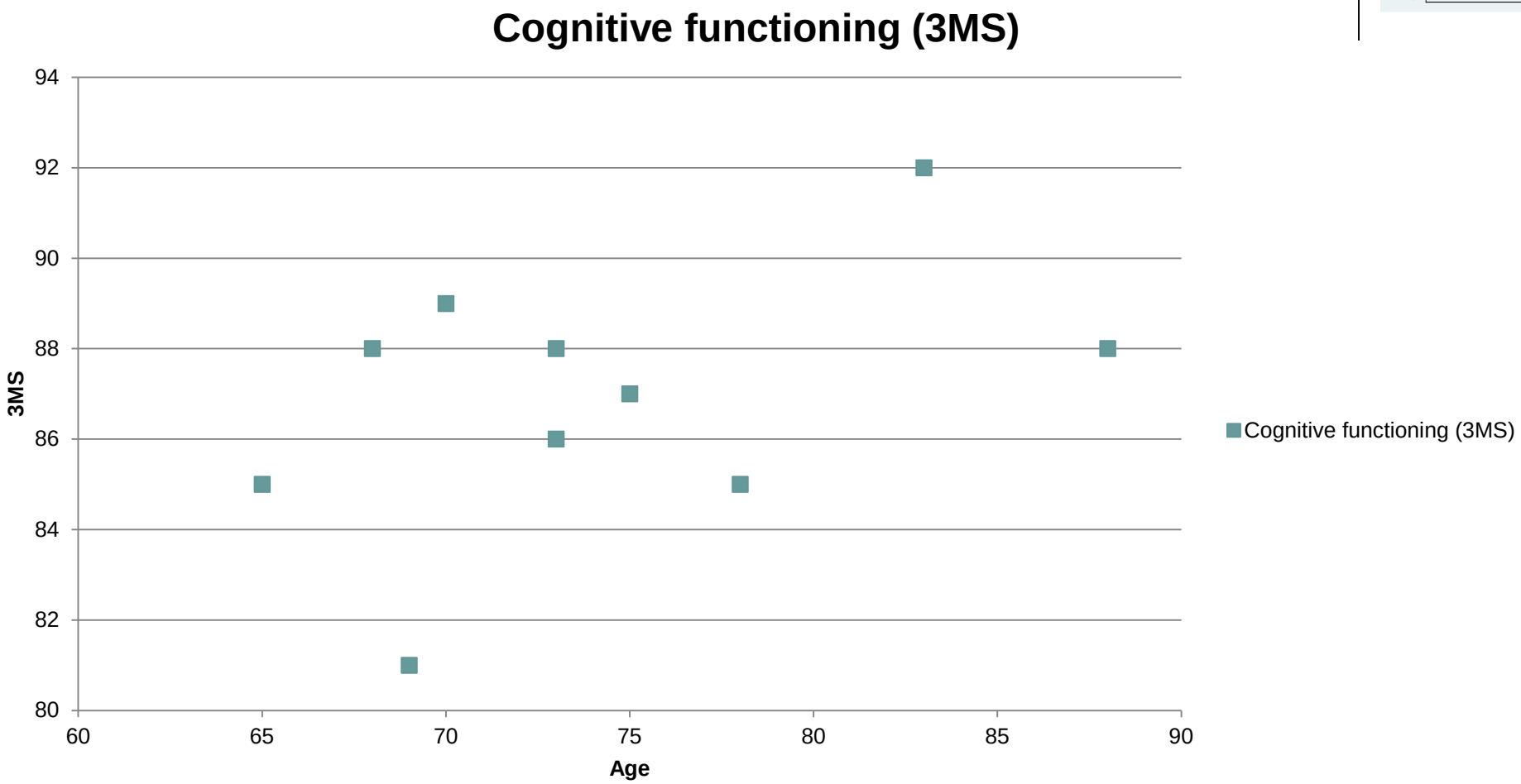
Longitudinal design: Individuals (interpreted broadly) are measured repeatedly over time.

Key features: Can analyze change over time within an individual.

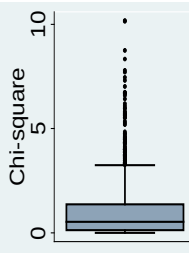
GEE methods are a very common approach for longitudinal data when there are large numbers of individuals and not so many repeated measurements.



Measuring change



Modified Mini-mental State Exam



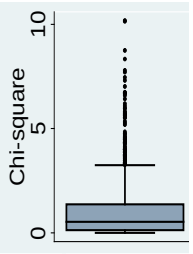
Numerical score from 0 to 100

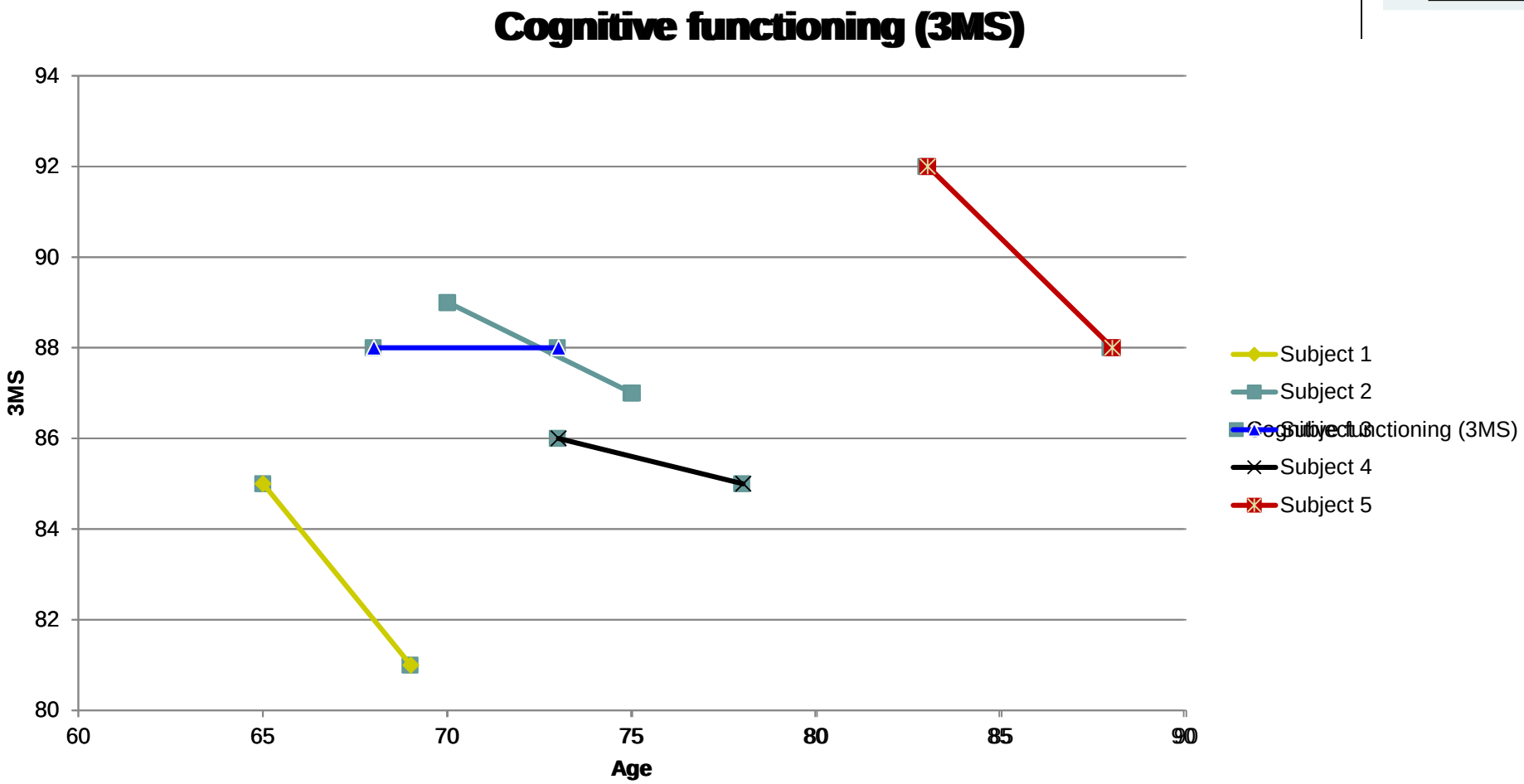
Tell patient “I'd like to test your memory; say these words: boat, cucumber, wire” (up to 3 points for 3 correct answers)

Tell patient “Begin with 100 and count backwards by 7” (up to 5 points for five correct answers)

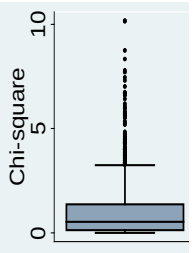
Tell patient “Can you name the three objects I named before?” (up to 3 points for 3 correct answers)

Measuring change





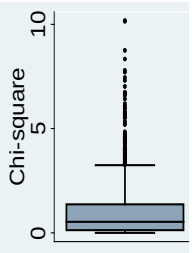
Analyzing change with longitudinal data



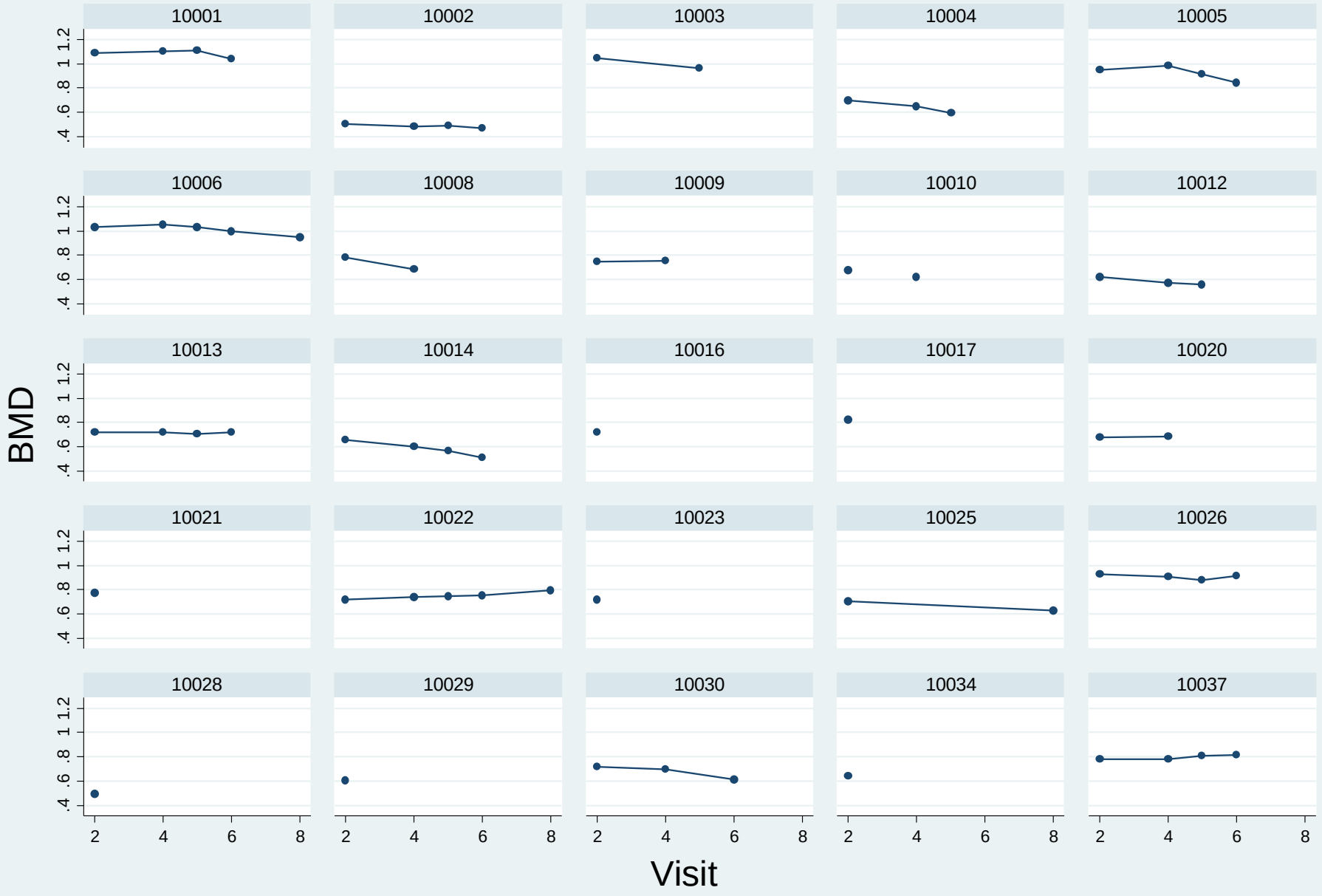
Example (SOF): It is well known that women who are overweight and obese have higher bone density. However, is the *change* in bone mineral density related to whether a woman is obese at baseline? We will define obese as BMI > 30.

How should we get started?

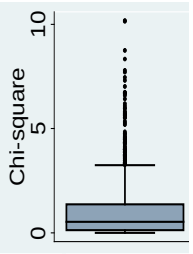
Example: SOF basics

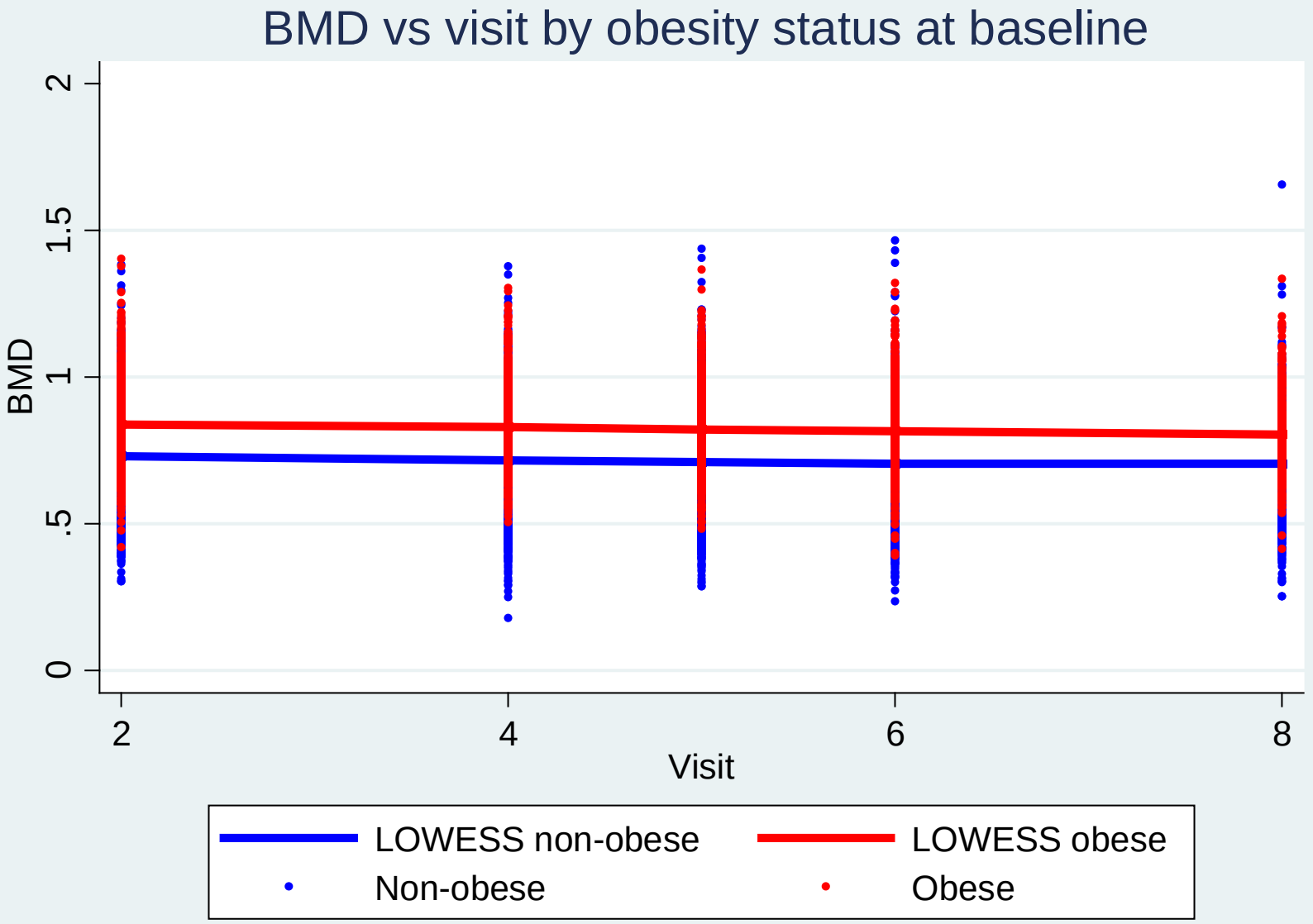



```
. xtset id visit
       panel variable:  id (unbalanced)
        time variable:  visit, 2 to 8, but with gaps
                delta:  1 unit
. xtdes
      id:  10001, 10002, ..., 42486          n =       8590
   visit:  2, 4, ..., 8                      T =          5
           Delta(visit) = 1 unit      Span(visit)  = 7 periods
           (id*visit uniquely identifies each observation)

Distribution of T_i:   min  5%  25%  50%  75%  95%  max
                         1  1    2    3    5    5    5
     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
     2275     26.48   26.48 |  1.111.1
     1734     20.19   46.67 |  1.111..
     1696     19.74   66.41 |  1.....
      926     10.78   77.19 |  1.11...
….
       99      1.15   90.94 |  1.1.1..
       89      1.04   91.98 |  1..11.1
      689      8.02  100.00 | (other patterns)
 ---------------------------+---------
     8590    100.00         |  X.XXX.X
```

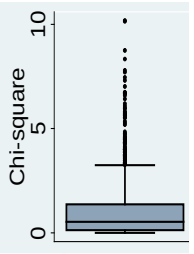

10001
10002
10003
10004
10005
10006
10008
10009
10010
10012
10013
10014
10016
10017
10020
10021
10022
10023
10025
10026
10028
10029
10030
10034
10037
BMD
.4 .6 .8 1 1.2
2 4 6 8
Visit
Graphs by ID

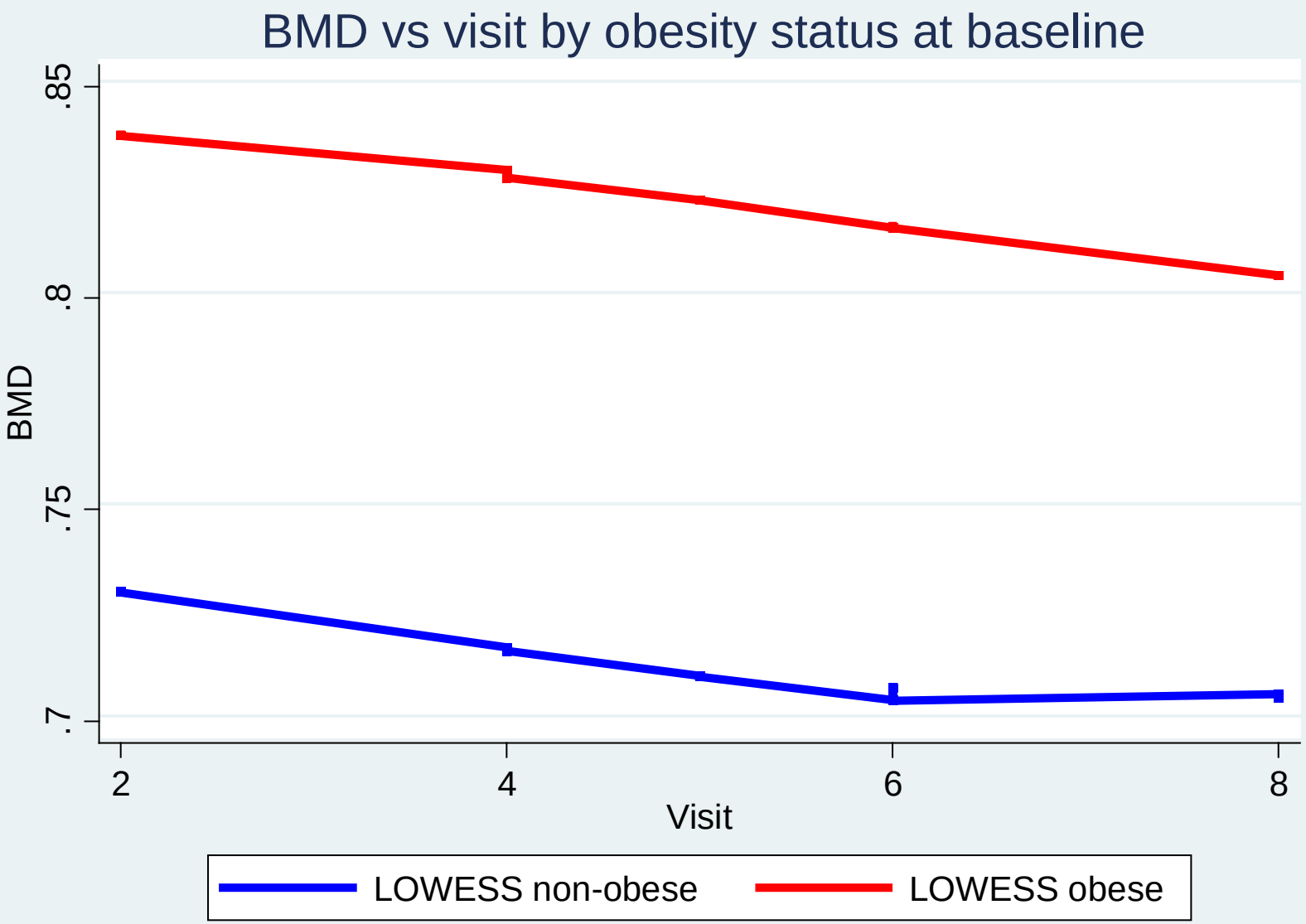
Example: BMD/Obesity



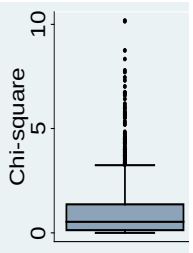


Example: BMD/Obesity



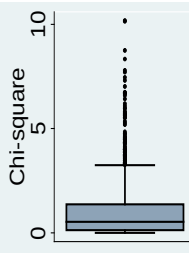


Analyzing change with longitudinal data



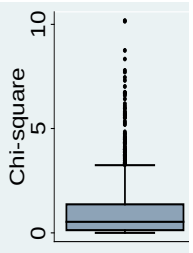
- Including a variable for *time* (or *visit*) describes the change over time.
- Inclusion of *time* (or *visit*) interactions with baseline predictors allows analysis of whether baseline predictors are associated with change over time.
- Inclusion of a time-varying predictor (e.g., BMI at sequential visits) allows analysis of whether change in that predictor is associated with change in the outcome.

How to handle “time” in a longitudinal analysis



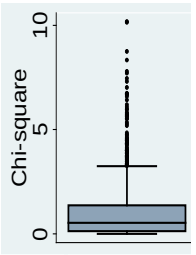
- We often try to model the shape of the change in the outcome over time.
- There is a natural “ladder” of handling a time predictor like visit, moving from simpler (to model *and* interpret) and more restrictive to more flexible: linear, quadratic, spline (flexible smooth fit), categorical.
- Moving up the “ladder” is a simple way to test adequacy of the simpler model.

Example: BMD/Obesity



- Want to characterize the change over time and see if it is the same or different between the obese and non-obese.
- Plot suggests that the trend might not be linear over time.
- So we need to check.

Example: BMD/Obesity

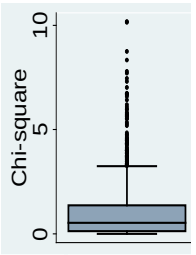



```
. xtgee totbmd visit##obese, i(id) robust

GEE population-averaged model                   Number of obs      =     26829
Group variable:                        id       Number of groups   =      8468
Link:                            identity       Obs per group: min =         1
Family:                          Gaussian                      avg =       3.2
Correlation:                 exchangeable                      max =         5
                                                Wald chi2(9)       =   3733.60
Scale parameter:                 .0169246       Prob > chi2        =    0.0000

                                          (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
      totbmd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |
          4  |  -.0164933   .0004961   -33.25   0.000    -.0174655    -.015521
          5  |  -.0270842   .0006331   -42.78   0.000     -.028325   -.0258434
etc.
```

Summary information about hierarchy

Overall model test (rarely useful)

If there are interactions between two variables in the model, the “main effect” of each of those variables is the effect when the other is at its reference value.



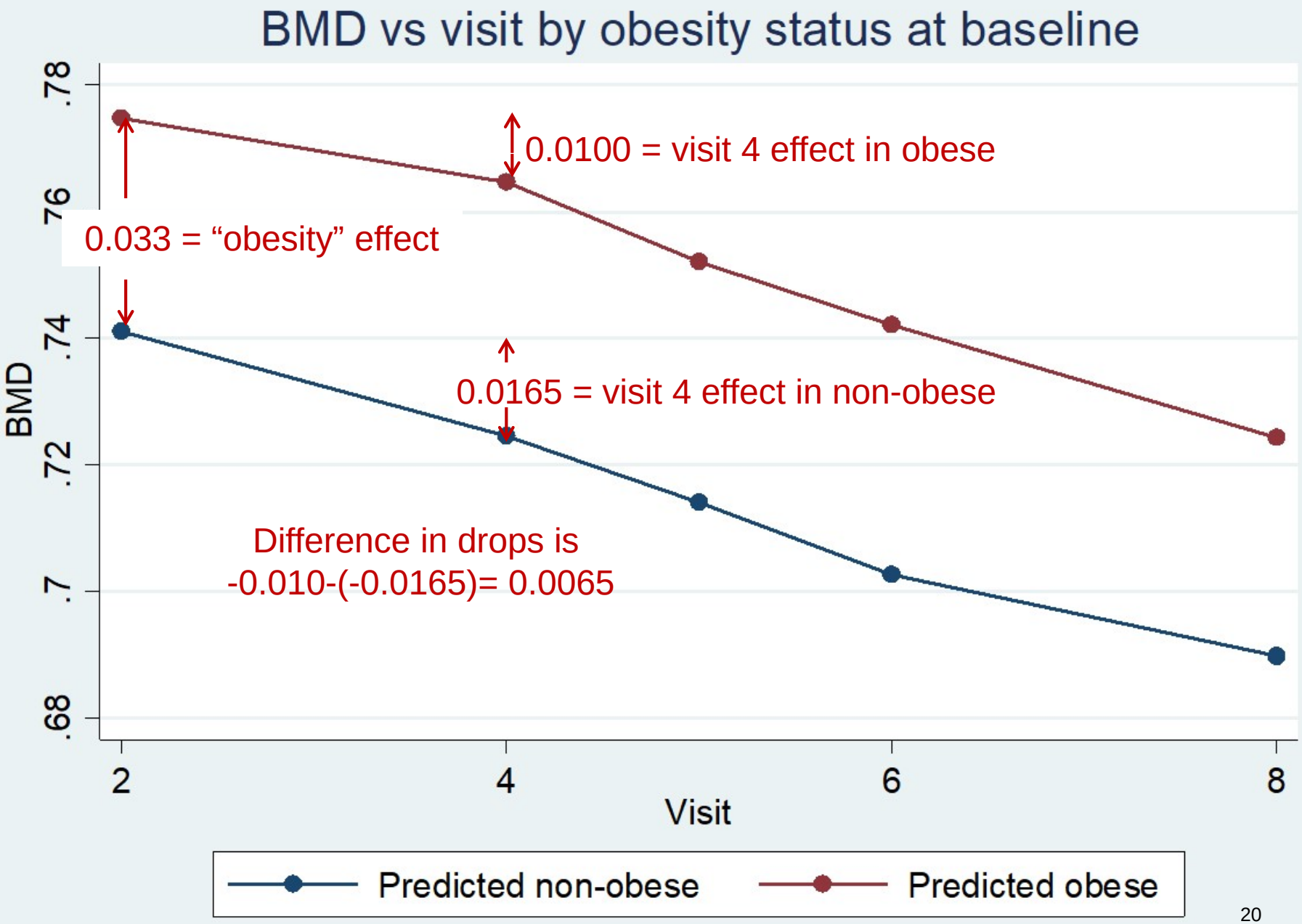
totbmd	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
visit						
4	-.0164933	.0004961	-33.25	0.000	-.0174655	-.015521
5	-.0270842	.0006331	-42.78	0.000	-.028325	-.0258434
6	-.0383912	.0007898	-48.			432
8	-.0513558	.001284	-40.			391
1.obese	.033644	.0018641	18.			976
visit#obese						
4 1	.0064715	.0014952	4.			402
5 1	.0044571	.0017778	2.			415
6 1	.0057885	.0021825	2.			61
8 1	.0008468	.0032882	0.2			16
_cons	.7410788	.0013913	532.6			57

On average, BMD drops by .0165 from

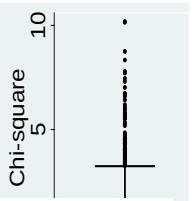
On average, BMD

On average, BMD is higher by 0.033 for the

The difference in the drop for obese v. non-obese between visits 2 and 4 is 0.0064

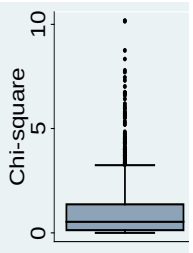

BMD vs visit by obesity status at baseline
0.0100 = visit 4 effect in obese
0.033 = “obesity” effect
0.0165 = visit 4 effect in non-obese
Difference in drops is
-0.010-(-0.0165)= 0.0065
BMD
.68
.7
.72
.74
.76
.78
2
4
6
8
Visit
Predicted non-obese
Predicted obese

Example: If you prefer tables




```
. margins visit#obese

Adjusted predictions                            Number of obs     =     26,829
Model VCE    : Robust

Expression   : Linear prediction, predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 visit#obese |
       2 0   |   .7410788   .0013913   532.65   0.000     .7383519    .7438057
       2 1   |   .7747228   .0021152   366.27   0.000     .7705771    .7788684
       4 0   |   .7245855   .0014573   497.22   0.000     .7217293    .7274418
       4 1   |    .764701   .0019559   390.97   0.000     .7608676    .7685345
       5 0   |   .7139946   .0015048   474.47   0.000     .7110452    .7169441
       5 1   |   .7520958   .0019877   378.37   0.000     .7481999    .7559916
       6 0   |   .7026876   .0015825   444.04   0.000      .699586    .7057892
       6 1   |   .7421201   .0021336   347.82   0.000     .7379382    .7463019
       8 0   |    .689723   .0018483   373.17   0.000     .6861005    .6933456
       8 1   |   .7242138   .0030273   239.22   0.000     .7182803    .7301473
------------------------------------------------------------------------------
. *Drop in obese group from visit 2 to 4 is from .7747228 to .764701
. di .764701-.7747228
-.0100218
```

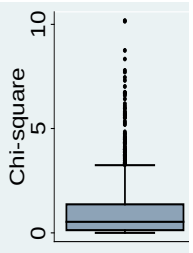
Back to the science: BMD/Obesity




```
. testparm visit#obese

 ( 1)  4.visit#1.obese = 0
 ( 2)  5.visit#1.obese = 0
 ( 3)  6.visit#1.obese = 0
 ( 4)  8.visit#1.obese = 0

           chi2(  4) =   22.20
         Prob > chi2 =    0.0002
```

The rate of change in BMD during the study is different for the obese and non-obese. The obese change less (since all the interaction coefficients are > 0).

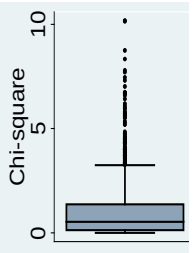
Back to the science:
Checking the linearity assumption




```
. xtgee totbmd c.visit##obese visit##obese, i(id) robust

. testparm visit#obese
 ( 1)  4.visit#1.obese = 0
 ( 2)  5.visit#1.obese = 0
 ( 3)  6.visit#1.obese = 0
           chi2(  3) =   19.02
         Prob > chi2 =    0.0003
```

The categorical interaction is statistically significant after fitting linear, therefore linear inadequate. Stick with categorical visit variable.

Example: BMD/BMI (time varying predictor)



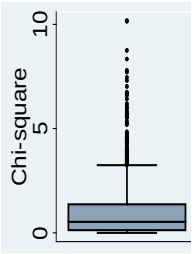
- Including a time-varying predictor automatically models a change over time

$$BMD_t = b0+b1*BMI_t \text{ implies}$$

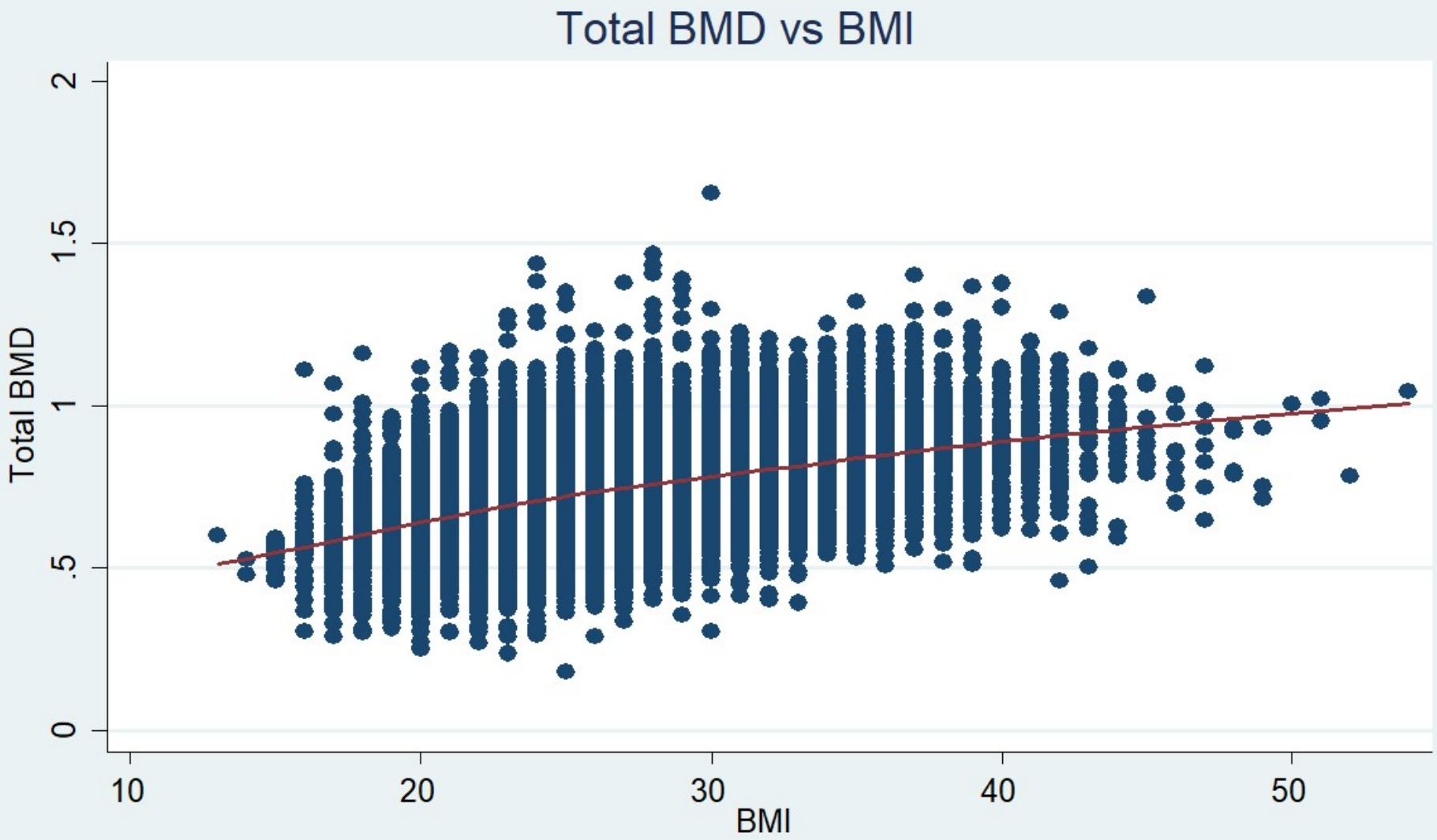
$$BMD_t - BMD_{t-1} = b1*(BMI_t - BMI_{t-1})$$

So change in BMD is predicted by change in BMI.

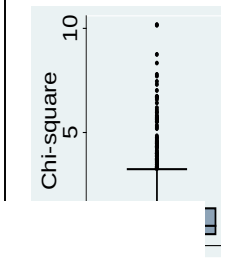
Ex 2: BMD/BMI (time varying predictor)



Does BMI predict total BMD? LOWESS plot

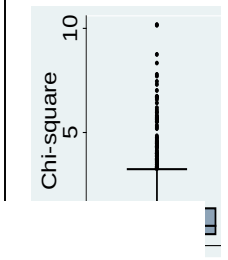


Example 2: xtgee




```
. xtgee totbmd bmi, i(id) robust

GEE population-averaged model                   Number of obs     =     26,829
Group variable:                        id       Number of groups  =      8,468
Link:                            identity       Obs per group:
Family:                          Gaussian                     min =          1
Correlation:                 exchangeable                     avg =        3.2
                                                              max =          5
                                                Wald chi2(1)      =    1639.11
Scale parameter:                 .0146148       Prob > chi2       =     0.0000

                                          (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
      totbmd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .0091548   .0002261    40.49   0.000     .0087116     .009598
       _cons |   .4890844   .0060705    80.57   0.000     .4771865    .5009823
------------------------------------------------------------------------------
```

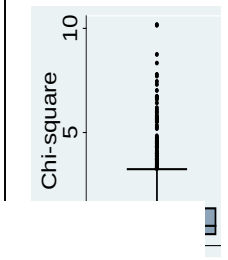
Example 2: Standardize




```
. egen bmi_std=std(bmi)
. egen totbmd_std=std(totbmd)

. xtgee totbmd_std bmi_std, i(id) robust
                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
  totbmd_std |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bmi_std |   .3218237    .007949    40.49   0.000     .3062439    .3374035
       _cons |  -.0317306   .0096224    -3.30   0.001    -.0505903    -.012871
------------------------------------------------------------------------------
```

- A one standard deviation change in BMI is associated with a .3 standard deviation change in total BMD.
- So a small to moderate sized effect using the usual rules of thumb (0.2 = small, 0.5 = medium 0.8 = large).

Example 2: Check linearity

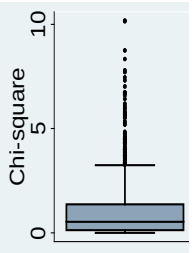


Assess linearity in BMI

```
. xtgee totbmd_std c.bmi_std##c.bmi_std, i(id) robust
--------------------+----------------------------------------------------------------
            bmi_std |   .3362406   .0077491    43.39   0.000     .3210527    .3514284
                    |
c.bmi_std#c.bmi_std |  -.0179974   .0041768    -4.31   0.000    -.0261838   -.0098111
                    |
              _cons |  -.0131305   .0103411    -1.27   0.204    -.0333986    .0071377
--------------------------------------------------------------------------------------
```

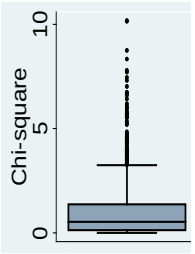
Since a quadratic term is statistically significant, the linearity in BMI assumption is not correct. Is it an important deviation? A plot of the predicted values versus BMI shows that the quadratic effect is minimal.

Longitudinal analysis strategy



- Assess if a predictor is time-varying or time-invariant (use `xtsum` if you don't know)
- Time-varying (like BMI) – just include it as a predictor.
- Time-invariant (like obesity at baseline) – include it in the model as well as the interaction of that variable with the time variable. A longitudinal relationship is tested by testing the interaction.

Summary



- Approximately normally distributed outcomes can be handled with mixed models (`mixed`) or generalized estimating equations (`xtgee`).
- Mixed models have the advantage of handling multiple levels of clustering and more explicit modeling of sources of variability and correlation. Better with missing data.
- Generalized estimating equations (when using the robust option) makes fewer assumptions.