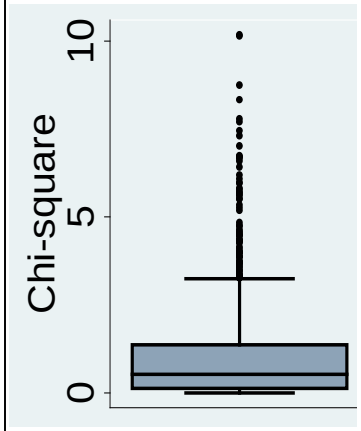
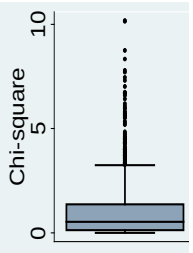
Repeated Measures
Lecture 1, Part 2

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

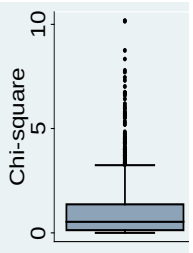




Last time

- Hierarchical data structures are common.
- They lead to correlated data.
- Ignoring the correlation can be a serious error.
- Can’t easily predict the directionality of the error.

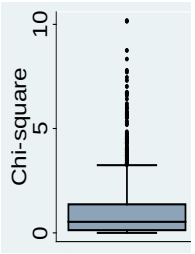
Today: Nuts and bolts in starting a repeated measures analysis.



Outline

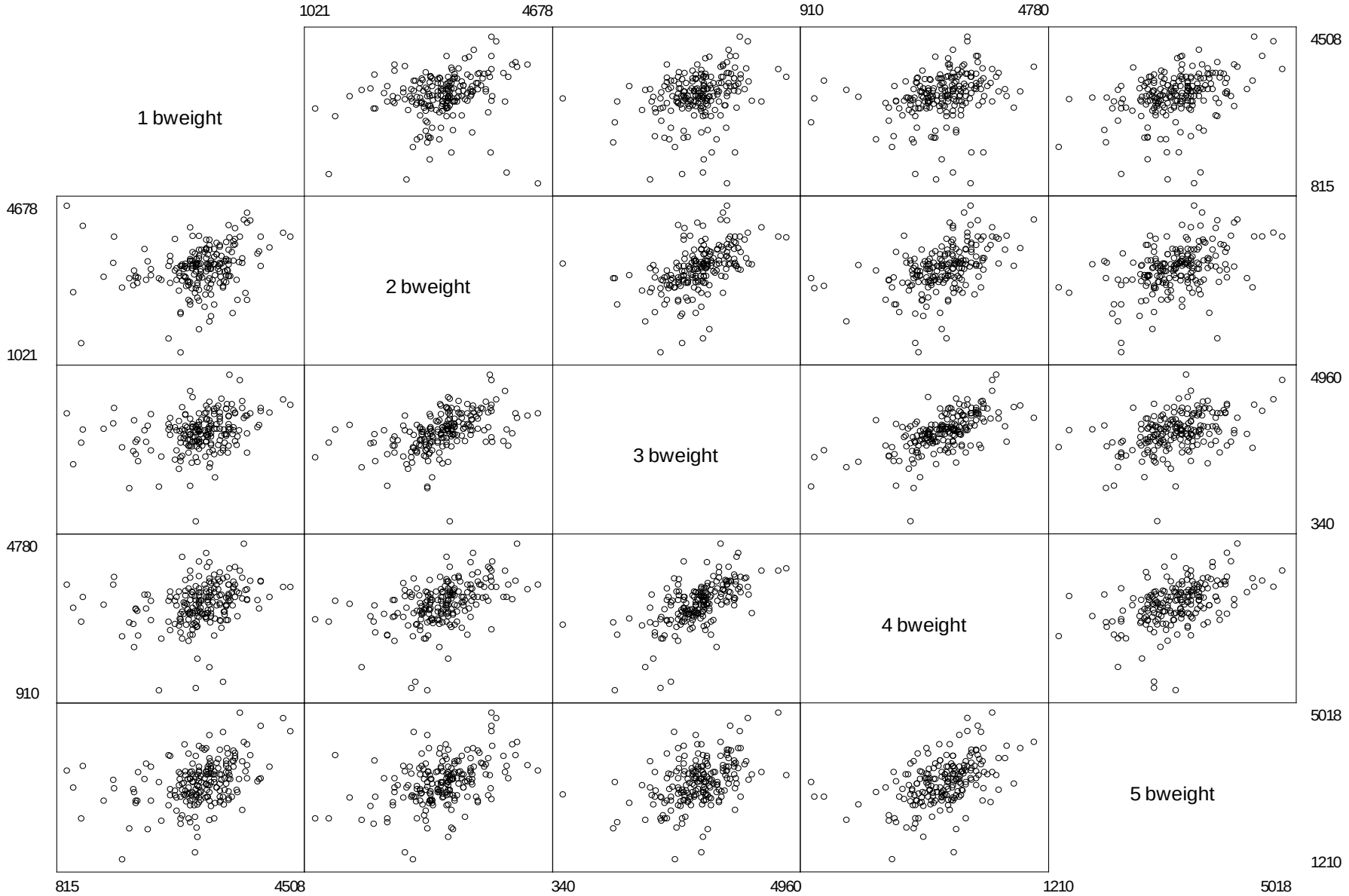
1. Correlation structures
2. Long and wide data formats
3. Descriptive methods for longitudinal data
4. Summary

Correlation structures

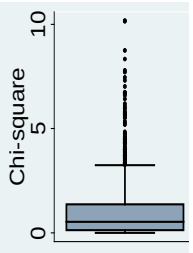


With continuous, balanced data we can plot the measurements that occur at different time points (or are repeated measurements of different types). The “Georgia babies” dataset follows successive birthweights of infants to mothers (each of whom had five children) from vital statistics in Georgia. We are interested in whether birthweight increases with birth order and mothers’ age. In lab we will generate the following plots.

Georgia Babies



Georgia Babies

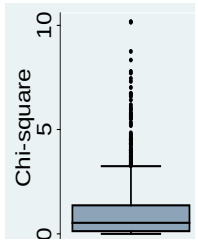


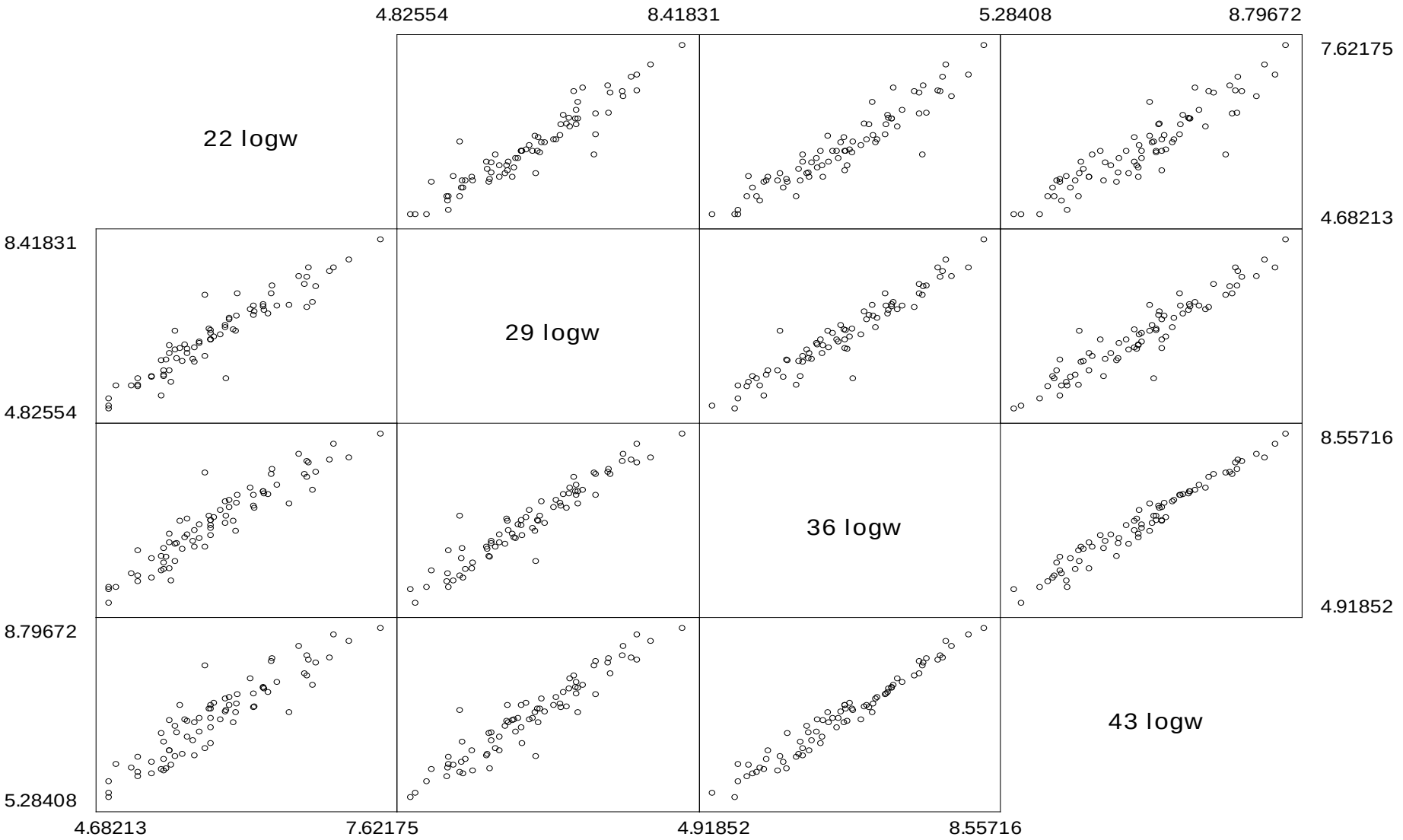
- Another common summary is the correlation matrix. Here is the correlation matrix for the Georgia babies data set:

```
. pwcorr  bweight1 bweight2 bweight3 bweight4 bweight5
             | bweight1 bweight2 bweight3 bweight4 bweight5
-------------+------------------------------------------
    bweight1 |   1.0000
    bweight2 |   0.2282   1.0000
    bweight3 |   0.2950   0.4833   1.0000
    bweight4 |   0.2578   0.4676   0.6185   1.0000
    bweight5 |   0.3810   0.4261   0.4233   0.4642   1.0000
```

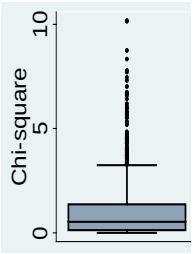
- How do we read this? Why isn't there anything in the top right hand corner?

Here is another example, giving the log weights (why log?) of mice for several weeks of measurement, mostly reflecting gain in tumor weight





Tumor/weight

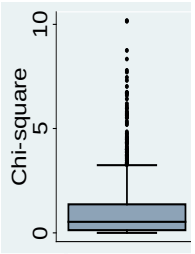


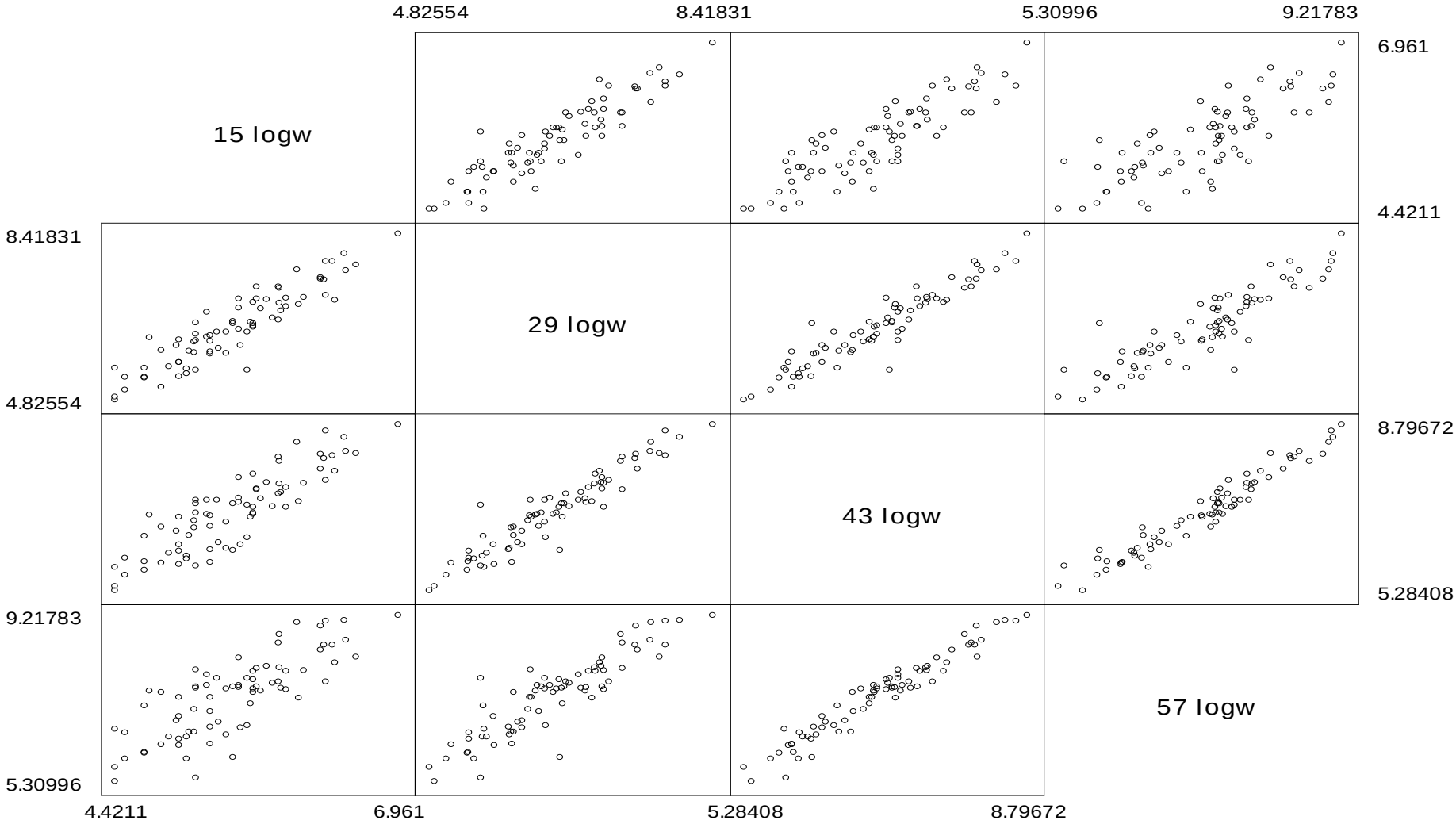
- And here is the corresponding correlation matrix:

```
pwcorr  logw22 logw29 logw36 logw43
             |   logw22   logw29   logw36   logw43
-------------+------------------------------------
      logw22 |   1.0000
      logw29 |   0.9414   1.0000
      logw36 |   0.9400   0.9568   1.0000
      logw43 |   0.9190   0.9466   0.9803   1.0000
```

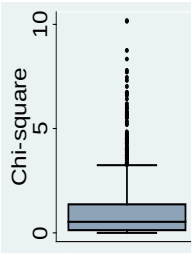
Tumor/weight

Here is a different collection of weeks for the same dataset. What does this suggest?





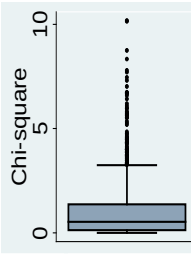
Tumor/weight



Here is the correlation matrix for that set of weeks:

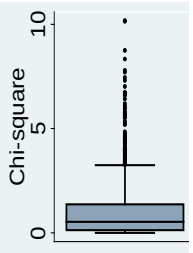
```
pwcorr  logw15 logw29 logw43 logw57

             |   logw15   logw29   logw43   logw57
-------------+------------------------------------
      logw15 |   1.0000
      logw29 |   0.9145   1.0000
      logw43 |   0.8713   0.9466   1.0000
      logw57 |   0.7937   0.8952   0.9692   1.0000
```

Correlation structures



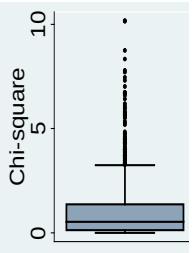
- The Georgia babies and tumor data sets are tidy because each “subject” has the same collection of observations (five observations for each mom and a tumor weight for each week). This is called “balanced” data.
- The Korff et al, back pain example is an example of unbalanced data, both because the sample sizes are unequal and because the “case-mix” is unequal. Because we don’t have a variable on which to reasonably order the observations (like parity for the Georgia babies data or time for the tumor data), there is not a reasonable plot we can make. But why are the data correlated in the back pain example?

Correlation structures



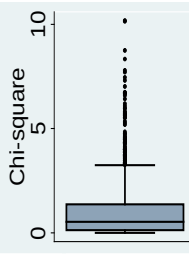
- Back to the tumor data: With 10 weeks of tumor data, there is the correlation of week 1 with week 2, week 1 with week 3, …, week 9 with week 10 for 45 unique correlations in all.
- Rather than having to estimate a separate correlation between each pair of times, we often use a simpler correlation "structure," both for ease of model specification and for statistical efficiency.

Correlation structures



- Common correlation structures used in STATA are:
  1. Exchangeable (all correlations equal).
  2. Autoregressive (correlations closer in time are more highly correlated, but drop off to zero as the difference in time increases).
  3. Unstructured (no assumptions made – estimate a separate correlation for each pair of time points).
  4. Independent (all correlations zero).
- In an AR(1) structure, if the correlation of adjacent time points is, say, 0.8, then the correlation of observations two time points apart is $0.8^2 = 0.64$.

Correlation structures



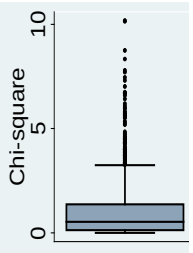
Which correlation structures do you feel best describe the examples we've considered?

Georgia babies

Tumor weights weeks 22 through 43

Tumor weights weeks 15 through 57

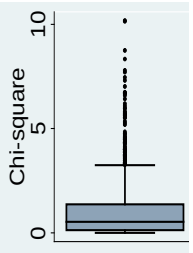
Back pain



Data layouts for longitudinal/clustered data

For longitudinal analyses: "long format"

id	visit	gender	bmi	totbmd
9000296	2	M	29.8	0.82
9000296	4	M	29.8	0.8
9000296	6	M	29.4	0.8
9000296	8	M	29.4	0.79
…				
9000798	2	F	32.4	0.4
9000798	4	F	32.3	0.41
9000798	6	F	32.3	0.39
9000798	8	F	32.5	0.39



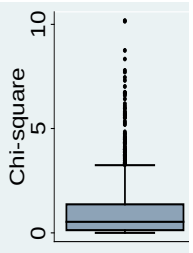
Data layouts for longitudinal/clustered data

For change score analyses: "wide format"

id	v02bmi	v04bmi	v06bmi
9000296	29.8	29.4	29.1
9000798	32.4	32.3	32.5

In lab: how to convert data back and forth between the two data formats.

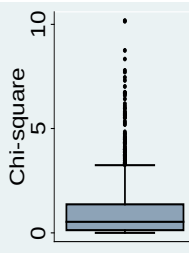
Example: HERS



Heart and Estrogen/Progestin Study (HERS - Hulley, *et al*, 1998, *JAMA*), was a randomized trial with long-term, yearly followup. pptid is the participant ID and nvisit is the visit number, with 0 being baseline:

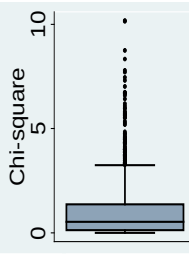
```
. xtset pptid nvisit
      panel variable:  pptid (unbalanced)
       time variable:  nvisit, 0 to 5, but with gaps
               delta:  1 unit
```

Descriptive methods: pattern of data



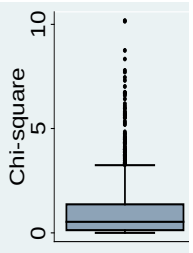

```
. xtdescribe
   pptid:  1, 2, ..., 2763                                 n =        2763
  nvisit:  0, 1, ..., 5                                    T =           6
           Delta(nvisit) = 1 unit
           Span(nvisit)  = 6 periods
           (pptid*nvisit uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       2       4         5         5       6       6

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
     1469     53.17   53.17 |  11111.
      531     19.22   72.39 |  1111..
      327     11.83   84.22 |  111111
      119      4.31   88.53 |  111...
      106      3.84   92.36 |  1.....
       87      3.15   95.51 |  11....
       40      1.45   96.96 |  111.1.
       22      0.80   97.76 |  1111.1
       11      0.40   98.15 |  11.11.
       51      1.85  100.00 | (other patterns)
 ---------------------------+---------
     2763    100.00         |  XXXXXX
```

Descriptive methods: variation between vs within individuals

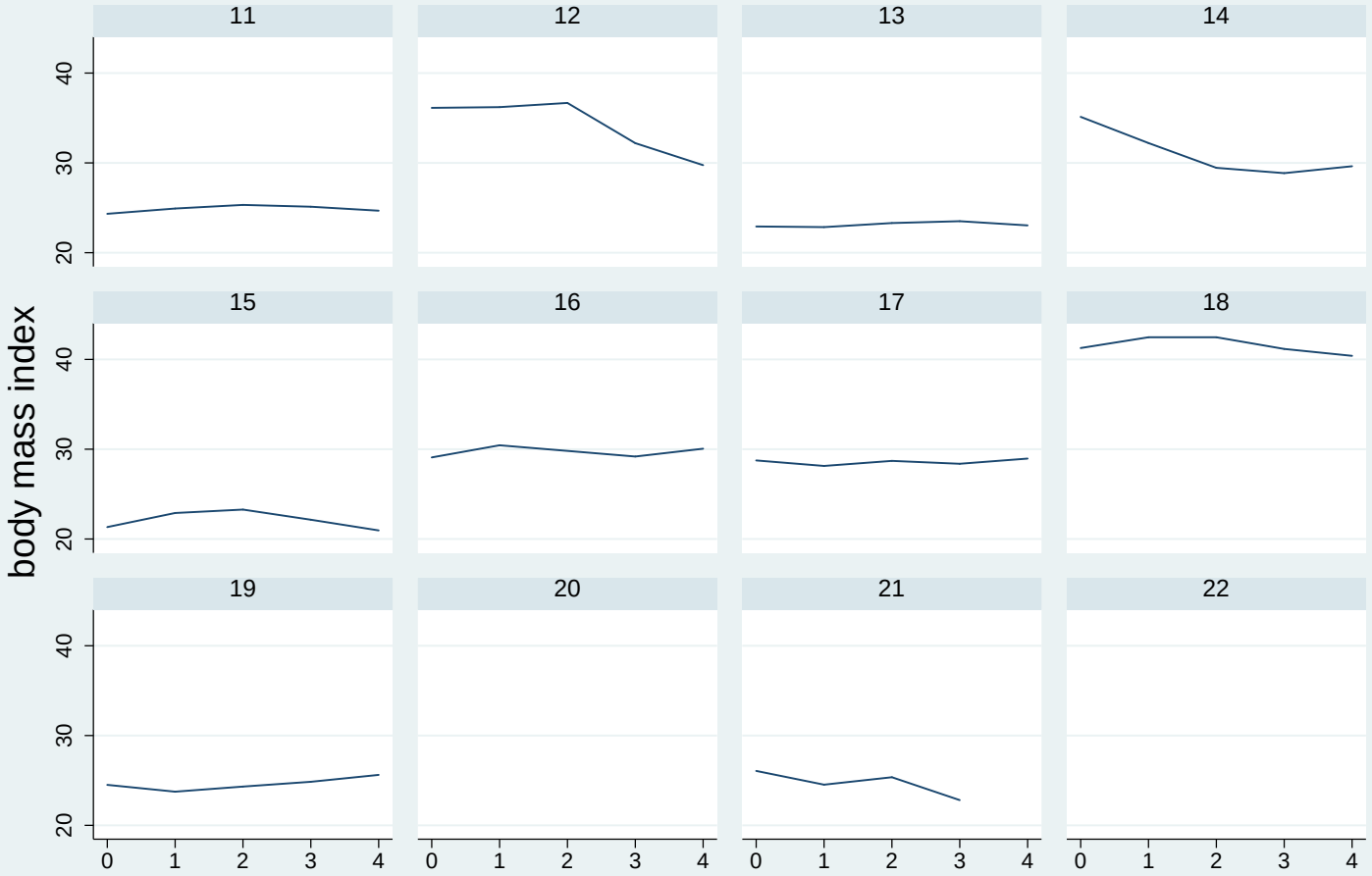



```
. xtsum bmi glucose white htnmeds

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
bmi      overall |  28.52184   5.565558      14.73      57.56 |     N =   12258
         between |             5.506242      15.21   54.89667 |     n =    2761
         within  |             1.174147   20.50784   38.41184 | T-bar =  4.4397
                 |                                           |
glucose  overall |  113.4589   41.18884          4        478 |     N =    7608
         between |             36.56199         53        311 |     n =    2763
         within  |             20.33938  -59.20781   322.7922 | T-bar = 2.75353
                 |                                           |
white    overall |  .8936408   .3083091          0          1 |     N =   12533
         between |             .3153701          0          1 |     n =    2760
         within  |                   0   .8936408   .8936408 | T-bar = 4.54094
                 |                                           |
htnmeds  overall |  .8341568   .3719546          0          1 |     N =   12548
         between |              .324773          0          1 |     n =    2763
         within  |             .1811389   .0008235    1.66749 | T-bar = 4.54144
```

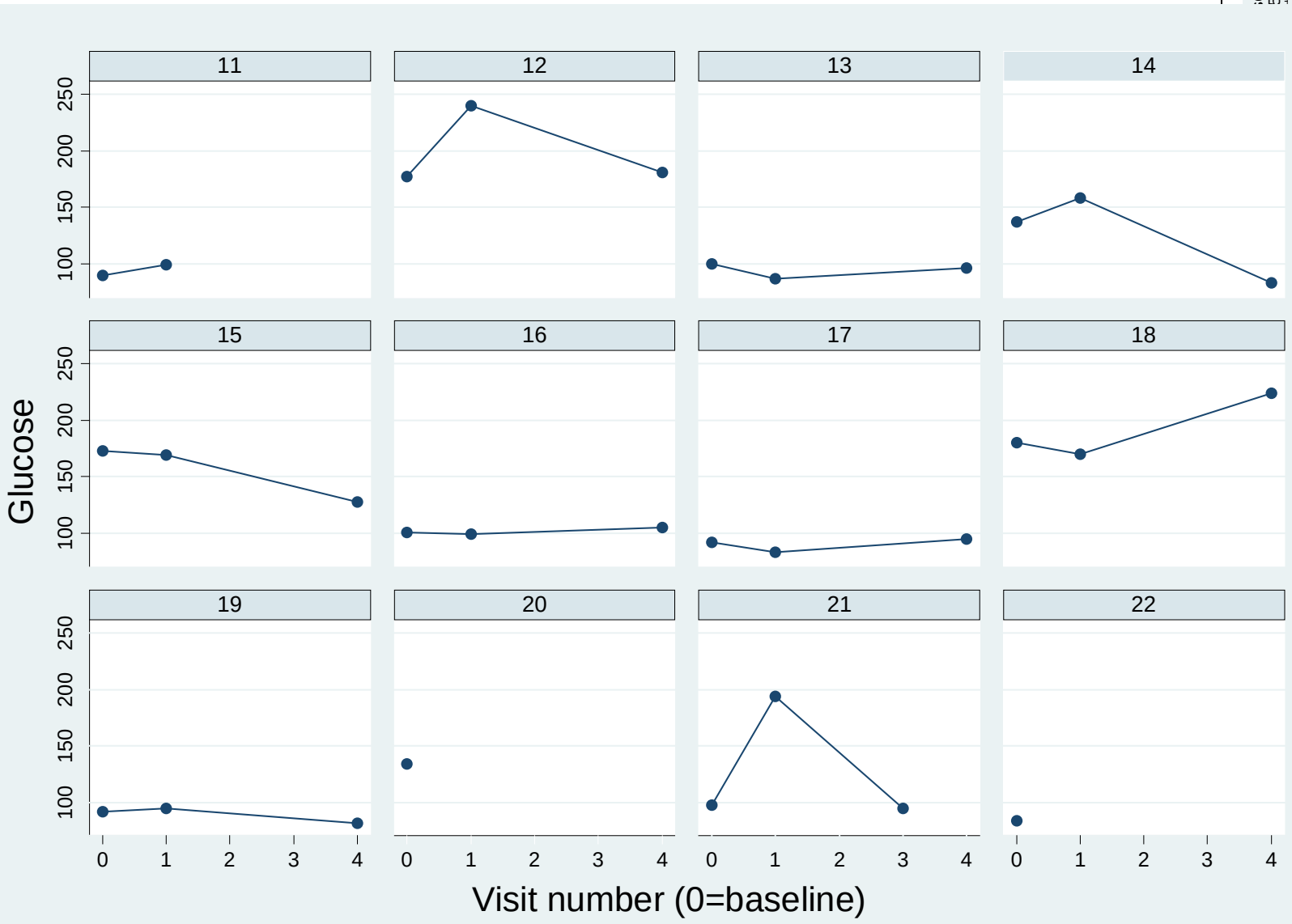
Descriptive methods: graphing variation over time



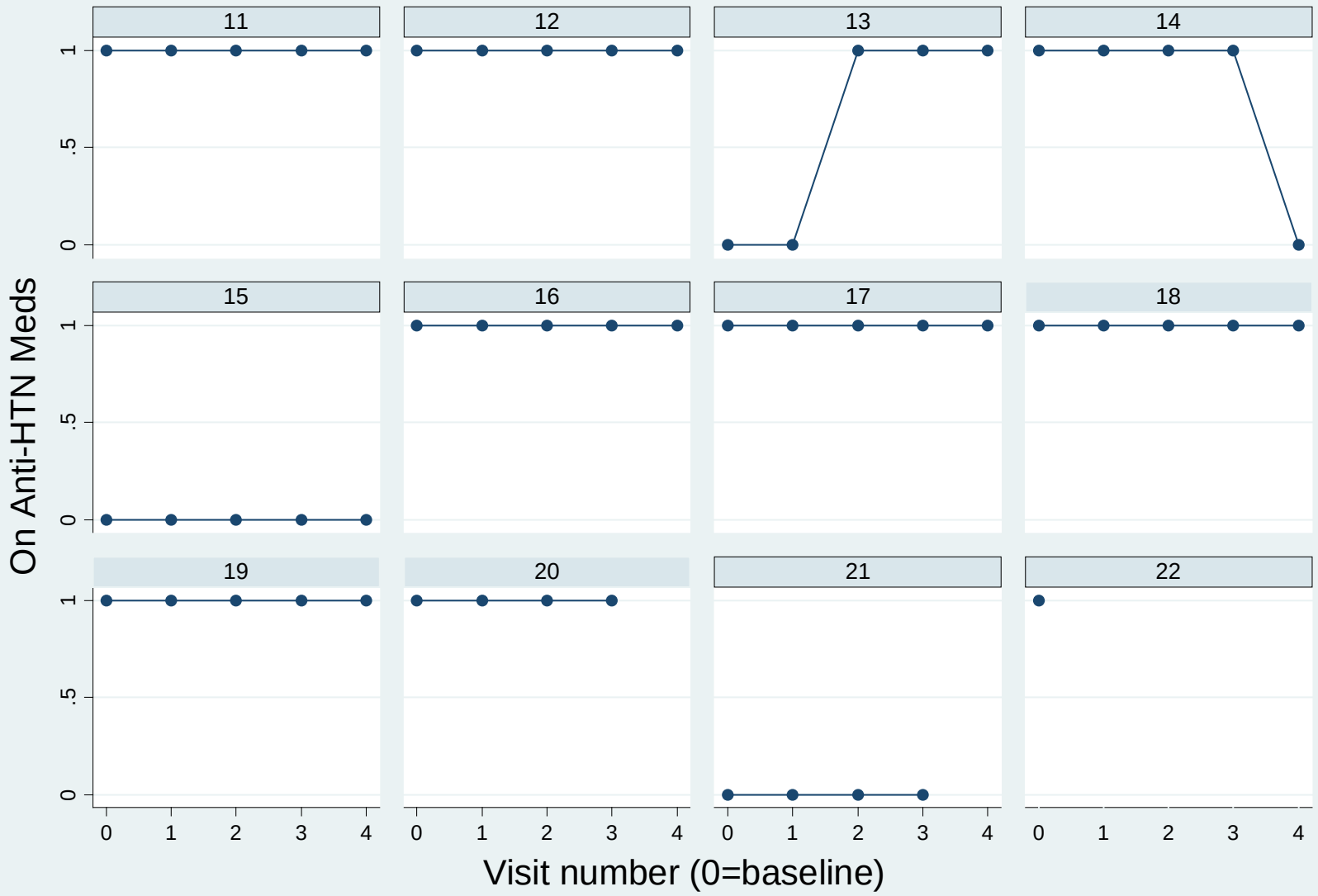

```
xtline bmi if pptid>10 & pptid<23
```



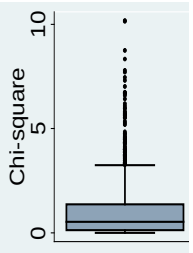
Descriptive methods: graphing



Descriptive methods: graphing



Summary – Part 2



- We will often need to specify a correlation "structure" when using correlated data methods.
- Repeated measures analyses need data in "long" format, but "wide" format is more convenient for certain calculations.
- Stata has convenient, built-in functions that generate descriptive statistics to better understand hierarchical data. Use them before embarking on formal statistical analyses.