Dynamic Structural Equation Modeling of Intensive Longitudinal Data Using Multilevel Time Series Analysis in Mplus Version 8 Part 7

Mårten Schultzberg

Bengt Muthén, Tihomir Asparouhov & Ellen Hamaker

Workshop at Johns Hopkins University, August 17 - 18, 2017

Outline

Example: For a group of individuals that has quit smoking, smoke urge is measured several times a day for 30 days. The difference in weight before quitting and at the end of the study, was measured.

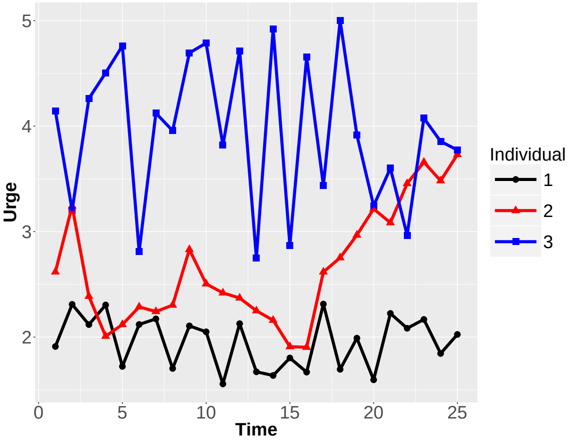
- Can different weight gains be explained by differences in the urge process?
- What "kind" of urge process is associated with large weight gains?
- Can the differences in urge process be explained by other covariates such as age, gender or health?

This example will be used throughout the presentation, also for the generated data to make the results easier to follow.

What features of the urge process might differ between individuals?

- Random mean
- Random autocorrelation
- Random residual variance

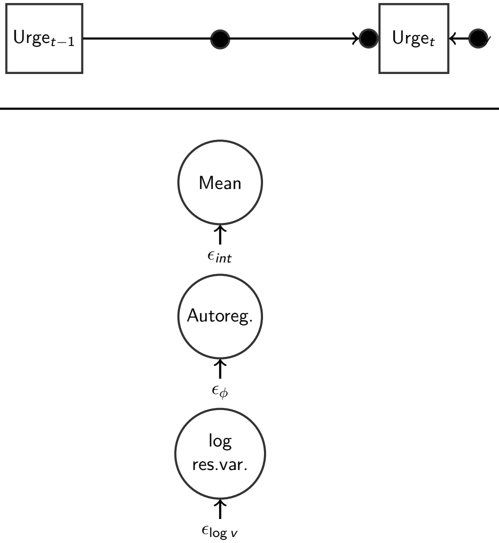
Random=individual specific



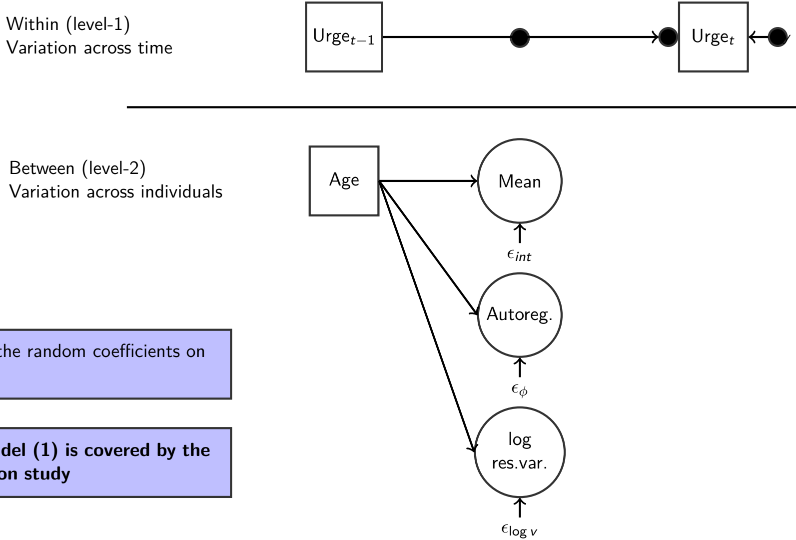
- Can the random coefficients be predicted by some covariate such as age, gender, health?
- Can the random coefficients be used as predictors for e.g. weight changes?
- With the new DSEM modelling framework these three random coefficients can be utilized as **both independent and dependent variables** (Asparouhov et al., 2017).

Within (level-1)
Variation across time

Between (level-2)
Variation across individuals



All individuals are assumed to have different means , autoregressive coefficients and residual variances

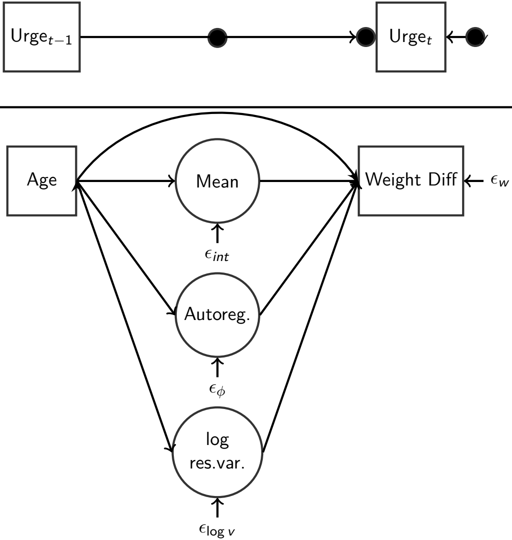


Regress the random coefficients on Age.

This model (1) is covered by the simulation study

Within (level-1)
Variation across time

Between (level-2)
Variation across individuals



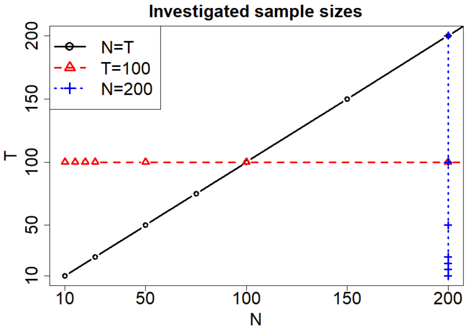
Regress the difference in weight on the random coefficients and Age.

This model (2) is covered by the simulation study

Consider three processes with different number of individuals (N) and number of time points (T):

- Self reported smoking urge:
  - Cheap and easy to collect with high frequency for many individuals with smart phone applications → Large N and T
- Blood sugar level:
  - Gives measure every minute → Large T
  - Expensive device limits number of participants → Small N
- Cognitive ability measured by psychologist:
  - Expensive to include many participants → Small N
  - Expensive to measure at many time points → Small T

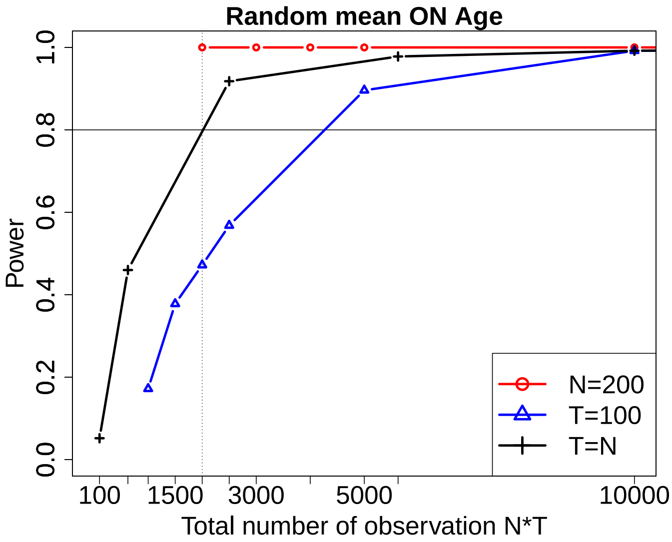
- What sizes of N and T are needed for good estimation of the DSEM models?
- Can large N compensate for small T and vice versa?
- The study considered three cases:
  1. N=200, combined with different T
  2. T=100, combined with different N
  3. N=T

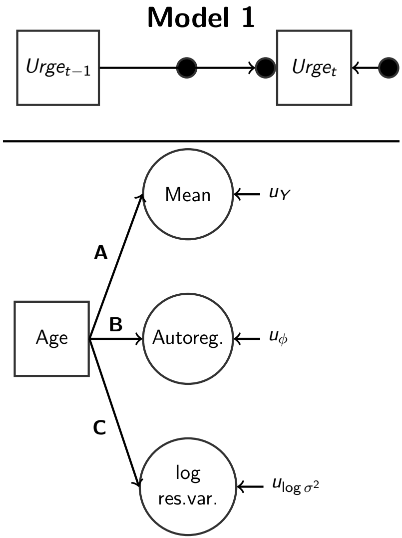


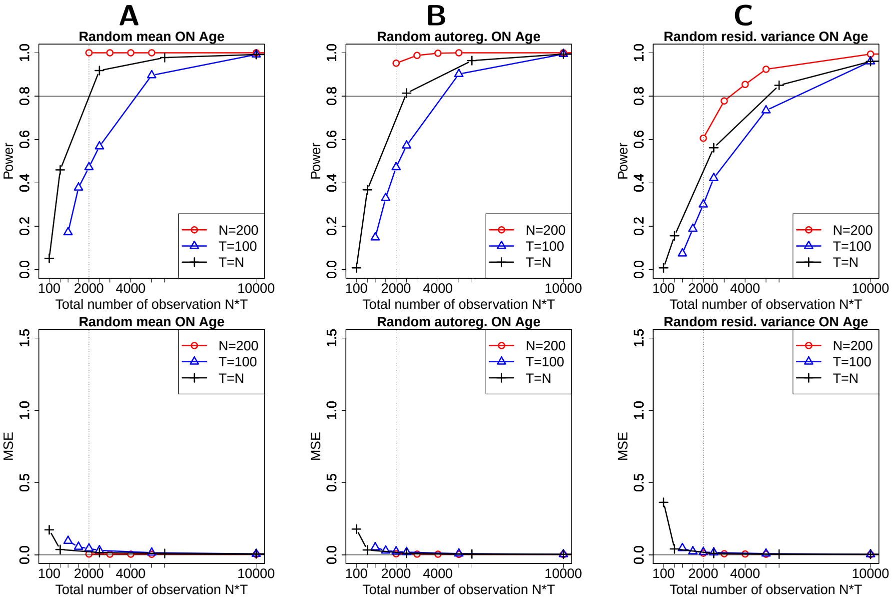
Total number of observations, N*T

- T=10, N=200 → 10*200=2000
- T=100, N=20 → 100*20=2000
- T=45, N=45 → 45*45=2025

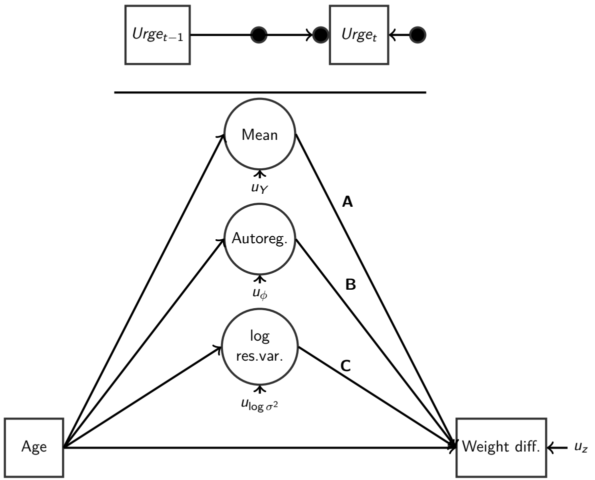
Using the total number of observations as a common scale simplifies the comparison of different N and T allocations

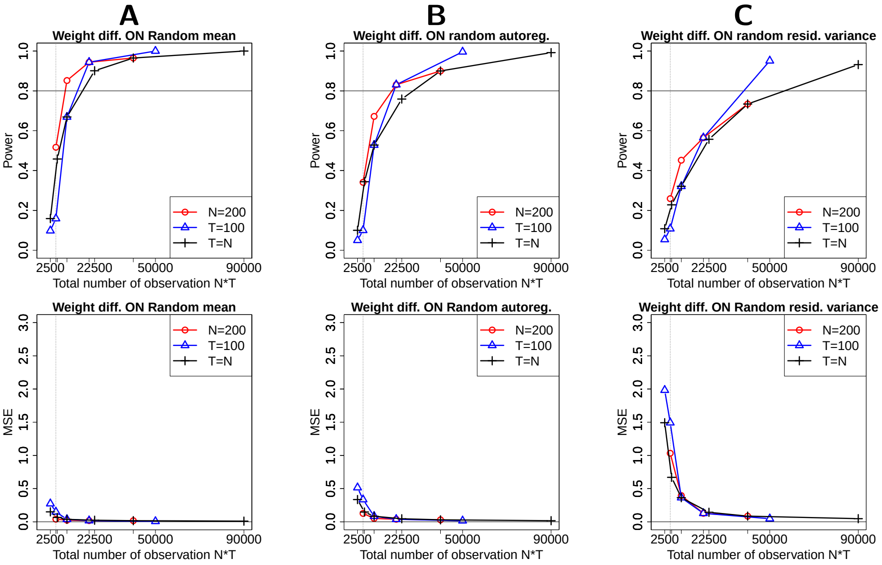

Random mean ON Age
Power
0.0
0.2
0.4
0.6
0.8
1.0
100
1500
3000
5000
10000
Total number of observation N*T
N=200
T=100
T=N

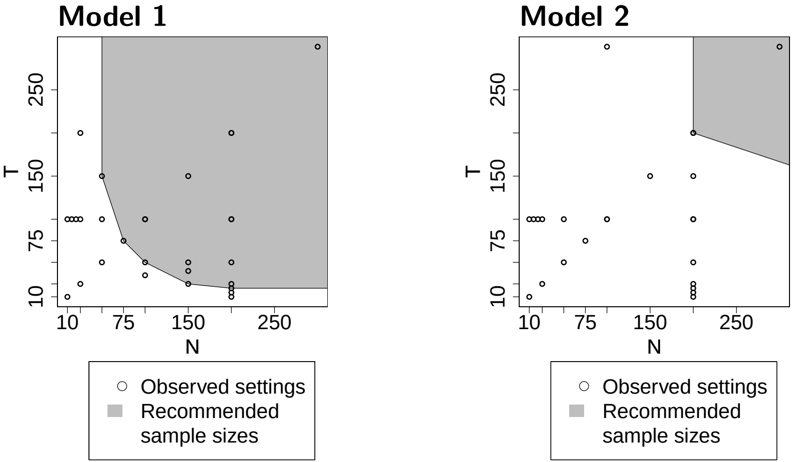

Model 1
$Urge_{t-1}$
$Urge_t$
Mean
u_Y
A
Age
B
Autoreg.
u_ϕ
C
log
res.var.
$u_{\log\sigma^2}$


A
B
C
Random mean ON Age
Random autoreg. ON Age
Random resid. variance ON Age
Power
MSE
N=200
T=100
T=N
100
2000
4000
10000
Total number of observation N*T
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5

Model 2


$Urge_{t-1}$
$Urge_t$
Mean
u_Y
A
Autoreg.
u_ϕ
B
log
res.var.
$u_{\log \sigma^2}$
C
Age
Weight diff.
u_z


A
Weight diff. ON Random mean
Power
N=200
T=100
T=N
Total number of observation N*T
B
Weight diff. ON random autoreg.
C
Weight diff. ON random resid. variance
Weight diff. ON Random mean
MSE
Weight diff. ON Random autoreg.
Weight diff. ON Random resid. variance


Model 1
Model 2
T
N
10
75
150
250
Observed settings
Recommended sample sizes

Conclusion:

- Large N can compensate for small T better than vice versa
- The regressions of the random coefficients (Model 1) have lower sample size demands than regressions using the random coefficients as predictors (Model 2)
- Random mean has lower demands than random autoregressive coefficient and residual variance (Model 1 and 2)
- For Model 1 I would recommend T larger than 25, and N larger than 200.
- For Model 2 I would recommend T and N larger than 150. This model is reasonable only if the effect sizes of the slopes of the regression on the random coefficients are at least moderately large.
- In the paper we consider 9 model variations.

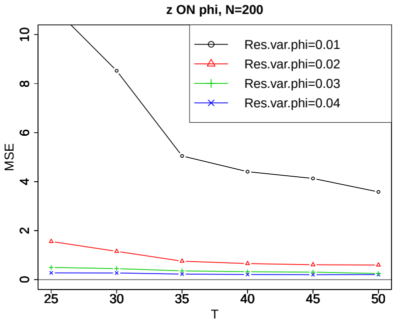
General setup:

- The usual Monte Carlo function in Mplus Version 8 was used (Muthén and Muthén, 2017).
- Batches of simulations was ran and summarised with `MplusAutomation` in R.
- Three (sometimes four) parallel Mplus applications running on each computer.
  - The computer supports 8 threads
  - Each simulation uses 2 threads by the `processors=2;` option

- `biter=(5000);`
  - Convergence
  - Coverage
  - Power
- Autoregressive parameters > 1
  - Non-stationary - Exploding time series
  - Mplus detection limit - Helps, but only for large T?
  - The `Results = res.dat;` option allows for manual checks.

Residual variance of the random autoregressive coefficient:

- If the model is saturated, like Model 2, the quality of the estimation of the slope of Z on the random autoreg. is determined by the residual variance of the autoreg.
- Since the variation in the autoreg. might be small to begin with this might be a problem



The `MplusAutomation` package for R:

- Create many similar syntax files:
  - Simulations with different sample sizes
  - Excluding different parts of a sample
- Running batches of input files
- Extract output from output files

The simulation study of sample sizes needed for DSEM

- 9 Models, around 25 settings for each model, different effect sizes→300 runs.
- 2 computers, 3 mplus applications running on each computer.
- 3 R-sessions on each computer using `MplusAutomation` managed all runs during days and nights.

Example of running on autoregressive model for *each individual in a sample*. The setup consists of

- An R-script (.R)
- A Text file (.txt)
- A Data file (.dat)

R-studio syntax

R_auto_ar1.R*

Source on Save

```r
library(MplusAutomation)
#Set the folder containing mplus.txt as working
createModels("mplus.txt")

#Set the folder containing input files as working directory
runModels()


#Set the folder containing output files as working directory
outputs<-extractModelParameters()

```

mplus - Anteckningar

Arkiv Redigera Format Visa Hjälp

```
[[init]]
iterators = individual;
individual = seq(1:202);
filename = "individual-[[individual]]-AR1.inp";
outputDirectory = "E:\Dropbox\Bengt\Testing\AR(1)_for_all_n\EXAMPLE";
[[/init]]

TITLE:
    AR for one individual at the time.

DATA:
    file = example.dat;

VARIABLE:
    NAMES = id Y;
    USEV = Y;
    USEOBS= (id == [[individual]]);
    LAGGED = Y(1);

ANALYSIS:
ESTIMATOR=BAYES;

MODEL:
Y ON Y&1;
```

R-studio syntax continued

```
ar<-numeric(202)  #vector of length #outputs
#i=list (one per output)
a=1 #1=unstandardized, 2=standardized (if available)
b=3 #which column: param name,point estimate, posterior sd,
c=1 #which parameter: AR-coff, intercept, resid.var.

for(i in 1:202){

  ar[i]<-as.numeric((outputs[[i]][[a]][b])[c,])
}
summary(ar)

```

18:53 (Top Level)

Console E:/Dropbox/Bengt/Testing/AR(1)_for_all_n/EXAMPLE/

```
> summary(ar)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.3160  0.0460  0.2565  0.2309  0.3887  0.8660 
```

Creating and running batches of Mplus input files in R:

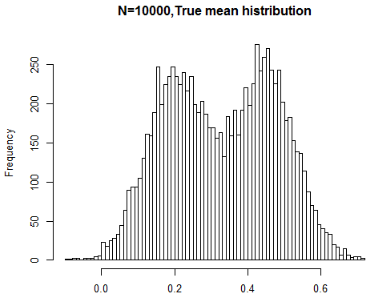
1. Create a folder and save all the following files in this folder
2. Create a .txt-file containing:
   2.1 What part (iterator) of the syntax you want to vary and what values the iterator should take
   2.2 What each inout file should be named and where to save it
   2.3 The Mplus syntax for one case with the varying part specified accordingly
3. Create an R-script containing:
   3.1 Load/install the `MplusAutomation` package
   3.2 `CreateModels()`. Creates the input files from the txt-file
   3.3 `runModels()`. Runs the created input files
   3.4 `extractModelParameters`. Extracts the information from the output files

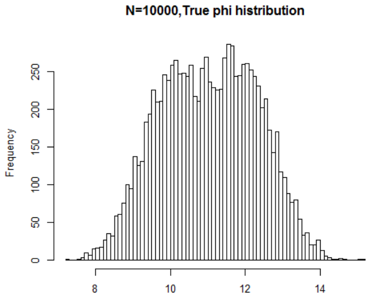
The distributional assumption of the random coefficients:

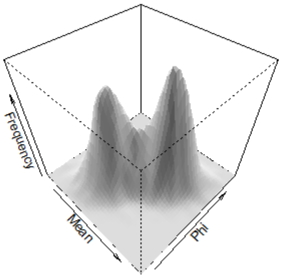
- A model assumption regarding the distribution of the random coefficients
- A random coefficient in Bayesian two-level time series models in Mplus is assumed to follow a Normal distribution

What if the random coefficient in the population is **not** Normal?

Non-normal distributed random coefficients - MC study



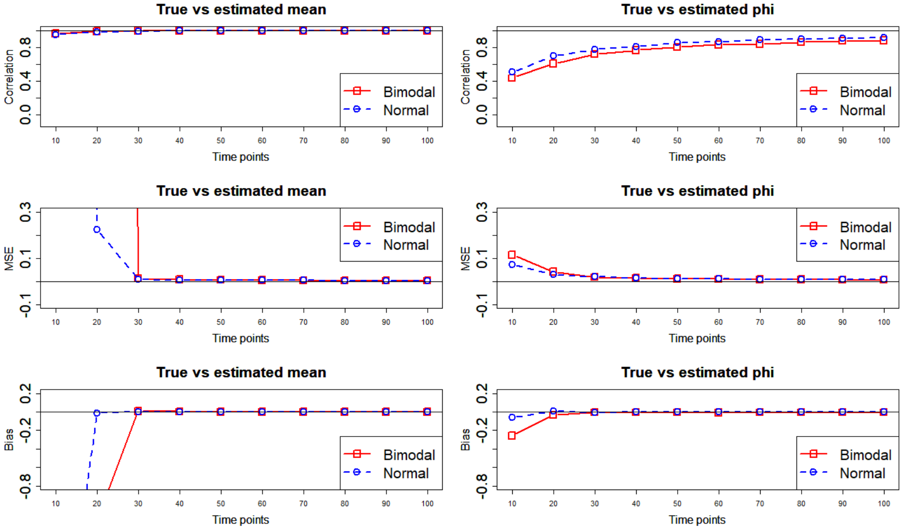




MonteCarlo simulation

- Generate multilevel time series (AR1) data with random coefficients according to previous slide
- Save each individual's true value
- Run the DSEM model in Mplus
- Save the factor score for each individual in Mplus
- Compare to the true value
- Replicate 500 times
- Change the number of individuals and time points per individual

N=10


True vs estimated mean
True vs estimated phi
Correlation
MSE
Bias
Time points
Bimodal
Normal
0.0
0.4
0.8
-0.1
0.1
0.3
-0.8
-0.2
0.2
10
20
30
40
50
60
70
80
90
100

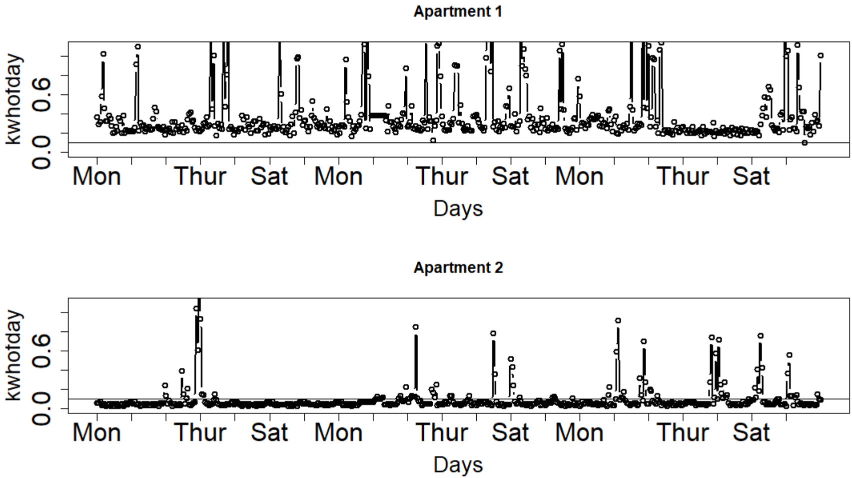
Conclusions about non-normal random coefficients:

- If T larger than 30 data seem to overrule the assumption
- For small T the estimation quality is better fir normal data
- For T larger than 30, the estimation quality is similar for non-normal and normal data
- The distributional assumption can be understood as a prior, however cannot be specified by the user.

Can basic information make residents in a house use their electronic devices optimally?

Joint work with the Cognition lab and the department of industrial engineering and management at Uppsala university.

- 60 apartments in a house built 2 years ago
- The house have solar panels on the roof
- The residents are unaware of the panels
- The residents are not using the electricity from the panels optimally
- The energy consumption is measured every hour for each apartment
- N=60, T=4000 (7 months)
- In addition daily energy consumption is available from the time point when the residents moved in (1 year)


Apartment 1
kwhofday
0.0
0.6
Mon
Thur
Sat
Mon
Thur
Sat
Mon
Thur
Sat
Days
Apartment 2
kwhofday
0.0
0.6
Mon
Thur
Sat
Mon
Thur
Sat
Mon
Thur
Sat
Days

Features of data:

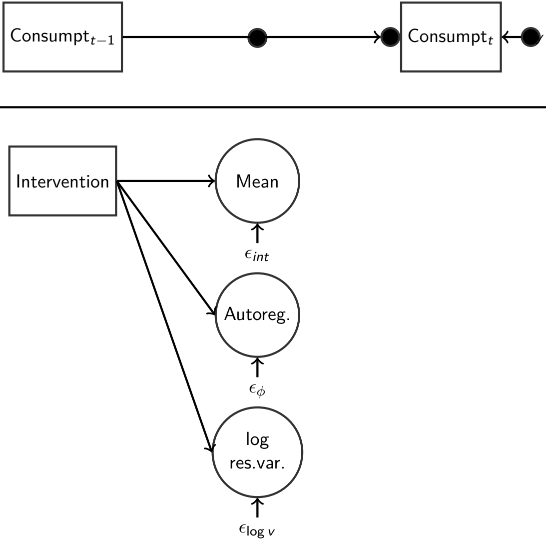
- Strong weekday effects
- Cycles over 24 hours, weeks, months, years
- Depends on the sunrise and sunset (at least in Sweden)
- Large variation across apartments

The research question is: Can general information make people change their electricity consumption behaviour?

An intervention in the form of information about how to use the timer functions on electronic devices is rolled out.

Within (level-1)
Variation across time

Between (level-2)
Variation across individuals



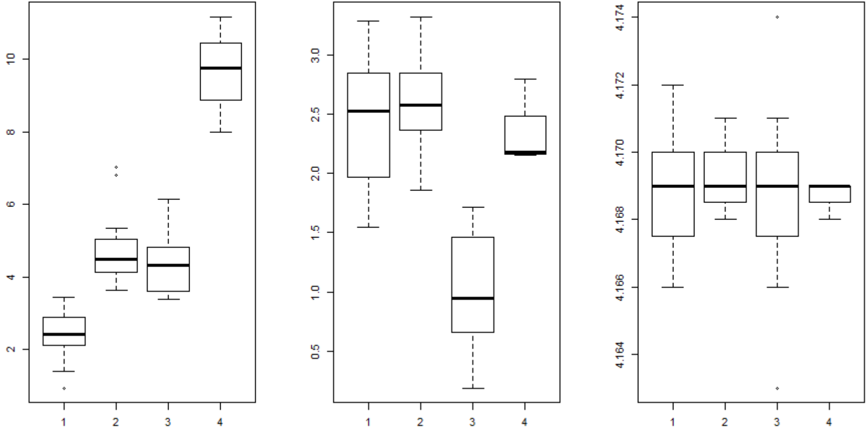
Regress the random coefficients on the treatment indicator.

Has the intervention significantly changed the mean of any random coefficient?

N is small, efficiency is crucial:

- Cluster on values of random coefficients, randomize within clusters. Only on data before treatment.
- Matching based on the values of the random coefficients.
- A step wedge design
- Use the daily data to obtain informative priors for e.g. effects of weekdays and other cycles.
- Test for intervention effect with a DSEM model.

4 clusters: Means of random **Mean, Autoreg. and Residual var**.



The study will be finished during the fall, intervention starts soon.

Thank you!

Asparouhov, T., Hamaker, E., and Muthen, B. (2017). Dynamic Structural Equation Models. *Unpublished*, pages 1–56.

Muthén, L. K. and Muthén, B. O. (2017). *Mplus User ' s Guide 8*.