W271 Live Session 11: Analysis of Panel Data 1

Devesh Tiwari

4/4/2017

Main topics covered in Week 12 (Async Unit 11)

```
- Introduction to panel data
- Using OLS regression model on panel data
- Exploratory panel data analysis
- Unobserved effect models
- Pooled OLS models
- First-Difference models
- Distributed Lag models
```

Readings

W2016: Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach.* 6th edition. Cengage Learning

```
- Ch. 13 (skip 13.4)
- [package plm](https://cran.r-project.org/web/packages/plm/plm.pdf)
- [plm vignettes](https://cran.r-project.org/web/packages/plm/vignettes/plm.pdf)
```

Agenda for this week's live session:

1. Questions about ARIMA and SARIMA models
2. Overview of the next unit: Panel data methods
3. EDA of panel and grouped data

Individual-level regression, pooled crossing regression, and first-difference regression

Some start-up codes:

```r
#sessionInfo()

# Insert the function to *tidy up* the code when they are printed out
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

# Set Numeric Value Display
# See reference from https://stat.ethz.ch/R-manual/R-devel/library/base/html/options.html
options(digits=2) # Set the printed number of digits to 2. Note: It is a suggestion only. Default is 7.
#options("scipen" = 10)

# Set memory limit
memory.limit(50000000)
```

```
## Warning: 'memory.limit()' is Windows-specific
## [1] Inf
```

```
# Clean up the workspace before we begin
rm(list = ls())

# Set working directory
wd <- "~/Documents/Projects/MIDS/Summer 2017/live_sessions/week11/"
setwd(wd)

# Load libraries
library(car)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
library(ggplot2)
library(lattice)
library(plm)
```

```
##
## Attaching package: 'plm'

## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

1. Exploratory data analysis of panel data

- *Estmated Time: Breakout session 25 minutes*
- *Estmated Time: Class discussion 20 minutes*

In this example, we use five waves of data from "National Youth Survey". Each year, participants in the age of 11, 12, 13, 14, 15 filled out a nine-item instrument designed to assess their tolerance of deviant behiavor such as cheat on tests, sell hard drugs, etc. Response to each item is provided in a 4-point scale (1 = very wrong, 2 = wrong, 3 = a little bit wrong, and 4 = not wrong at all). At each occasion (i.e. wave), the outcome, *TOL*, is computed as the respondent's average across the nine responses. Two potential explanatory variables in this dataset include *male*, which is equal to 1 if the respondent is *male*, and *exposure*, which is a respondent's estimated proportion of their close friends who were involved in each of the nine activities, measuring on a 5-point scale (0 = none, 5 = all).

Task 1: - Import the data as data.frame. - Examine the basic structure of each of the datasets. - Print the person-level dataset. Discuss it. - Also, answer "how many male and female in the dataset?" - Construct the correlation matrix of the variables tol11 - tol15. Discuss what you observe in this matrix. Do the corrleation values in the matrix make sense? Why? Why not? - Print the person-period level dataset. Discuss it. - How is the person-period level dataset different from that of the person-level dataset? - Does the person-level dataset contain exactly the same information as that in the person-period-level dataset as far as *tolerance*, *male*, and *exposure* variables are concern? Why? Why not? - Conduct a throughout EDA using ther person-period level dataset

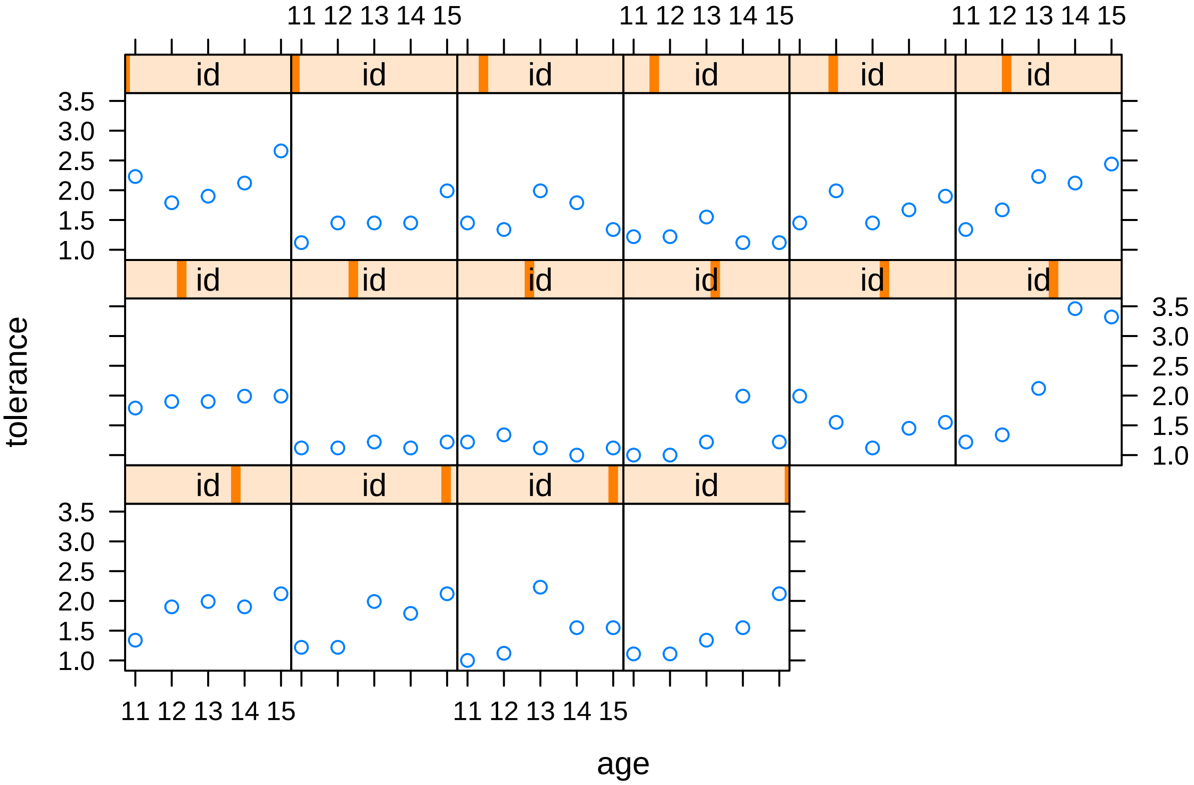
```r
# YOUR CODE HERE
df <- read.table("tol_person.txt", sep = ",", header = T)


df2 <- read.table("tol_person_period.txt", sep = ",", header = T)
```

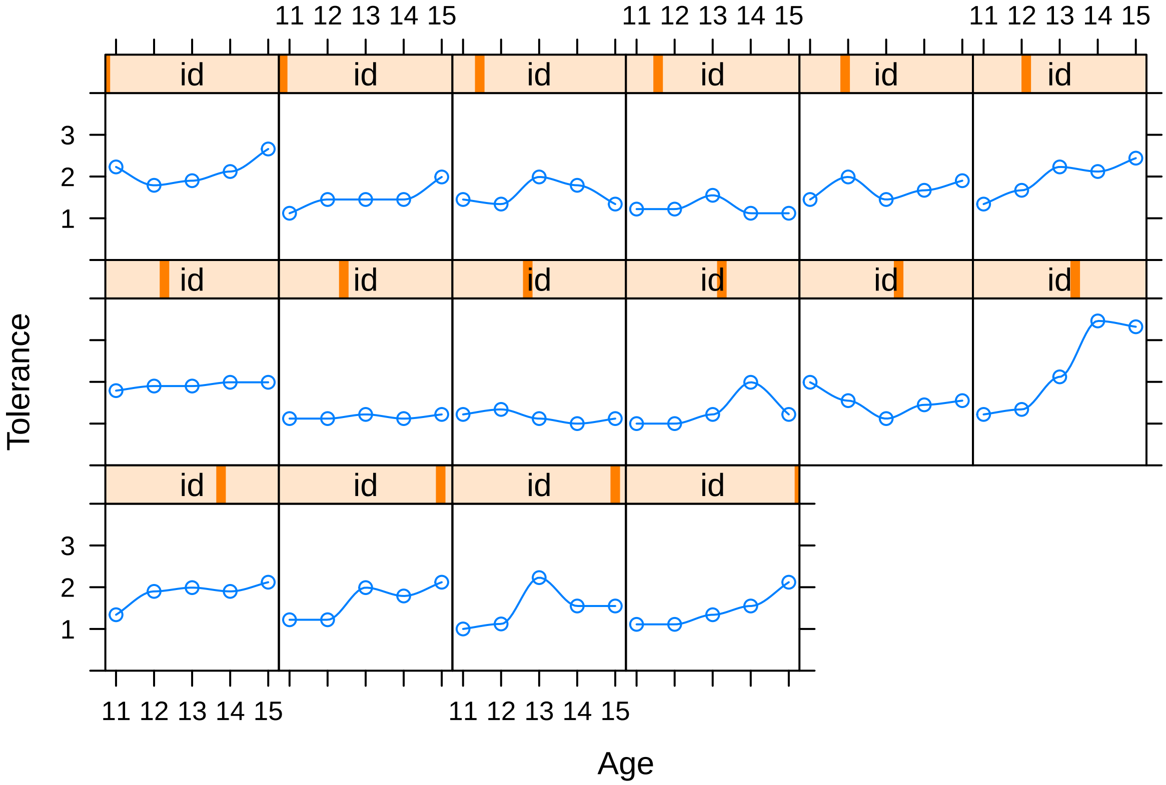
2. Growth-curve Analysis (something not covered in the book and the async)

- *Estmated Time: Instructor's teaching 10 minutes*

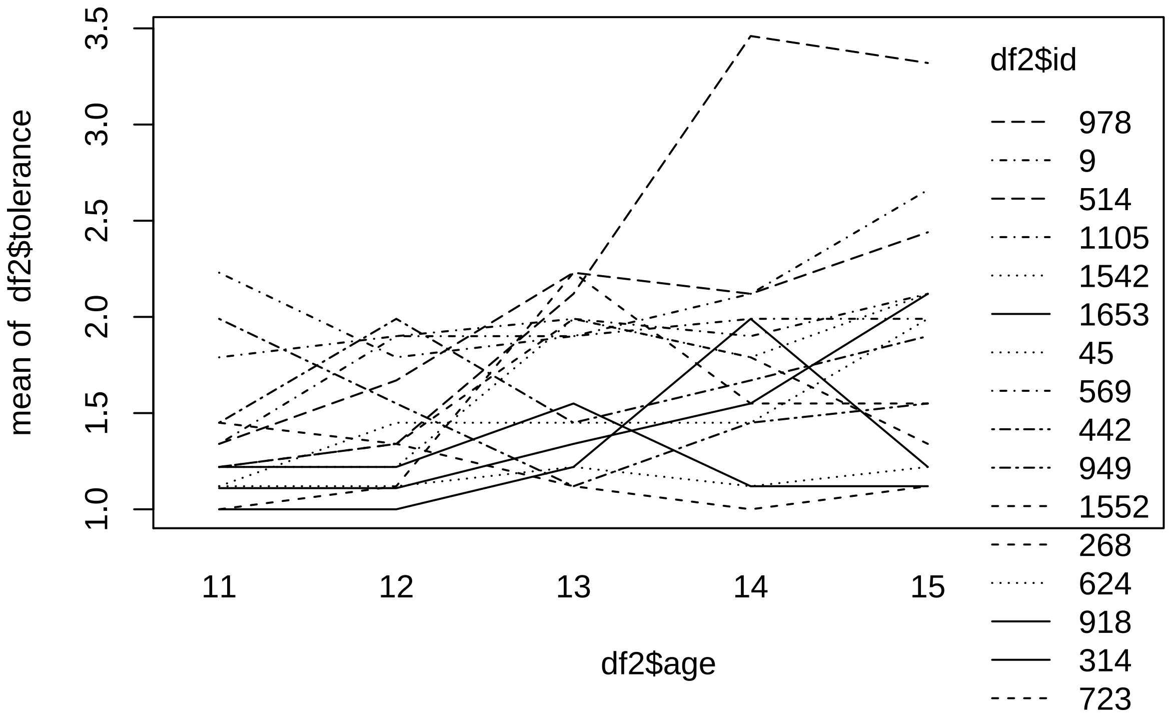
```r
xyplot(tolerance ~ age | id, data = df2, as.table = T)
```



```
# Smooth nonparametric trajectories superimposed on empirical
# growth plots.
xyplot(tolerance ~ age | id, data = df2, prepanel = function(x,
    y) prepanel.loess(x, y, family = "gaussian"), xlab = "Age",
    ylab = "Tolerance", panel = function(x, y) {
        panel.xyplot(x, y)
        panel.loess(x, y, family = "gaussian")
    }, ylim = c(0, 4), as.table = T)
```



```
# plot of the raw data
interaction.plot(df2$age, df2$id, df2$tolerance)
```



3. Build regression models to answer "Does being exposed to friends having deviant behavior increase one's tolerance of deviant behavior?"

- *Estmated Time: Breakout session 20 minutes*
- *Estmated Time: Class discussion 15 minutes*

Task 1: Estimate individual-level regression models. Remember that potential explanatory variables include *time*, *male*, and *exposure*. Can you use all of these variables in the regression models? Why? Why not? Does this regression help us answer the question? Why? Why not?

Task 2: Answer the question using a pooled-OLS appoach. Interpret the model results.

Task 3: Answer the question using a first-difference approach. Interpret the model results. Conduct EDA on both your dependent variable and explanatory variables first.

```
# YOUR CODE HERE
```