

GSERM - St. Gallen 2022

Analyzing Panel Data

June 6, 2022

- Instructor: Prof. [Christopher Zorn](mailto:zorn@psu.edu) (zorn@psu.edu).
- Class: June 6-10, 2022, 13:00-18:00 CET, at the [University of St. Gallen](#) (via [Zoom](#)).
- The course outline / syllabus is [here](#).
- More important: The syllabus, slides, readings, code, data, etc. are all available on the course [github repo](#) (viewable at <https://github.com/PrisonRodeo/GSERM-Panel-2022>).
- Class sessions will be recorded; links to those recordings will be available on [CANVAS](#).

github.com/PrisonRodeo/GSERM-Panel-2022

Search or jump to... Pull requests Issues Marketplace Explore

PrisonRodeo / GSERM-Panel-2022 Public

Unpin Unwatch 1 Fork 0 Star 2

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags

Go to file Add file Code

PrisonRodeo Useful R Resources be4c6f8 2 days ago 15 commits

Code	Code	5 days ago
Data	Data	5 days ago
Readings	Readings	5 days ago
Slides	Slides	2 days ago
.gitattributes	Initial commit	5 days ago
GSERM-2022-Useful-R-Resources....	Useful R Resources	2 days ago
GSERM-Remote-Software-2022.pdf	Virtual Machine Instructions	5 days ago
GSERM-StGallen-Analyzing-Panel-...	Syllabus	2 days ago
README.md	Create README.md	5 days ago

About

This is the Github repository for the summer 2022 version of GSERM's "Analyzing Panel Data."

Readme 2 stars 1 watching 0 forks

Releases

No releases published
[Create a new release](#)

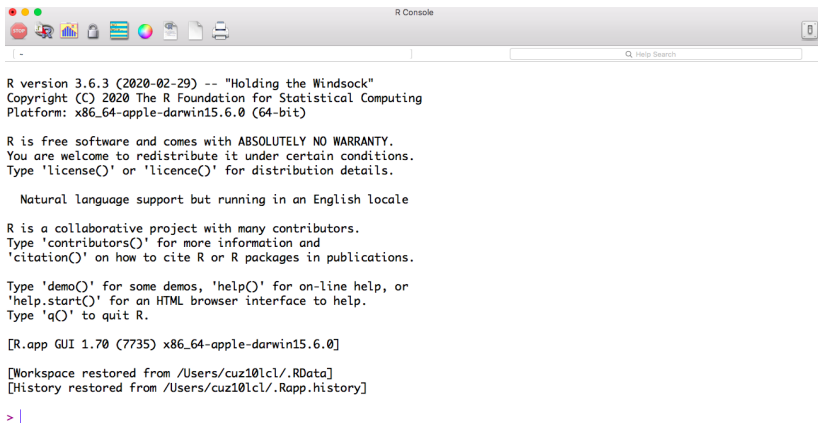
Packages

R

- All examples, plots, etc.
- Current version is 4.2.0
- Also be sure to get the [RStudio](#) IDE...
- The Github repo contains a bit of [introductory code](#) for people who may never have used R, and a list of [R resources](#).
- Primary packages we'll use (see the [econometrics](#) task view for more):
 - `plm`
 - `lme4`
 - `gee`

Stata

- Current version is 17
- Mostly use the `-xt-` series of commands (for “cross-sectional time series”)



```
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

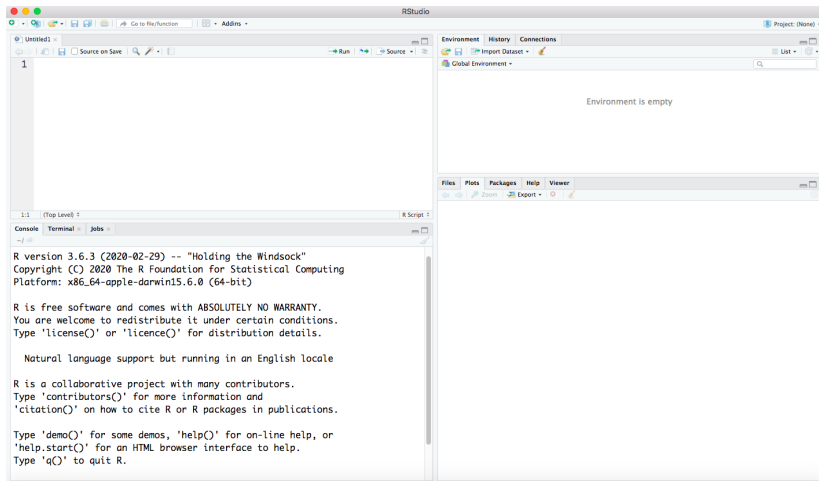
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.70 (7735) x86_64-apple-darwin15.6.0]

[Workspace restored from /Users/cuz10lcl/.RData]
[History restored from /Users/cuz10lcl/.Rapp.history]

> |
```



RStudio (annotated)

Source window:

- Click here to save your source code. Save often!
- Source on Save
- Run

This is the "Source" window.

- It's the place where you'll type the code that will then be sent to R.
- It's basically a text editor. You can open text files of any kind here if you want.
- Files that appear here end in (and should be saved with) the extension ".R" (as in "MyCode.R").

You'll spend most of your time working here.

Environment window:

This is the "Environment" window. It is where you can find all the various "objects" that you create, grouped by object type (data frames, lists, graphs, etc.). Environment is empty

There's also a "History" tab above; switching to that will show what has transpired in the Console window recently.

Console window:

This is the "working directory." Anything you save will be saved here, unless you tell the program to save it somewhere else.

Console - /Dropbox (Personal) /

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

This is the "Console." When you run the code in the Source window, the results that aren't graphics appear here.

Files window:

This is a window that shows various other things. Those things are tabbed above ^ and include:

- Plots (graphs) that you have created
- Packages that are loaded
- Help results (obtained by typing "?XXX" in the Console window, e.g. "?table").

- Panel data: Data comprising repeated observations over time on a set of cross-sectional units.
- Terminology:
 - “Unit” / “Units” / “Units of observation” / “Panels” = Things we observe repeatedly
 - “Observations” = Each (one) measurement of a unit
 - “Time points” = When each observation on a unit is made
 - $i \in \{1 \dots N\}$ indexes units
 - $t \in \{1 \dots T\}$ or $\{1 \dots T_i\}$ indexes observations / time points
 - If $T_i = T \forall i$ then we have **“balanced”** panels / units
 - Balanced panels also imply $N_t = N \forall t$
 - NT = Total number of observations (if balanced)
- “Panel” \neq “Time Series”
- “Panel” \neq “Multilevel” / “Hierarchical” / etc.

$N \gg T \rightarrow$ “panel” data...

- (American) National Election Study panel studies ($N \approx 2000$, $T = 3$)
- Swiss Household Panel (FORS) ($N = \text{large}$, $T = 21$)
- Often:
 - Cross-sectional units are a sample from a population
 - T is (relatively) fixed

$T \gg N$ or $T \approx N \rightarrow$ “time-series cross-sectional” (“TSCS”) data

- National OECD data ($N = 20$ original members, $T \approx 70$)
- Often:
 - N is an entire population, and is (relatively) fixed
 - Asymptotics are in T

$N = 1 \rightarrow$ “time series” data

$T = 1 \rightarrow$ “cross-sectional” data

Panel Data Structure + Organization

id	t	Y	X_1	...
1	1	250	3.4	...
1	2	290	3.3	...
\vdots	\vdots	\vdots	\vdots	...
2	1	160	4.7	...
2	2	150	4.9	...
\vdots	\vdots	\vdots	\vdots	...

Introduction to Panel Variation: A Tiny (Fake) Example

	ID	Year	Female	Pres	GOP	Approve
1	1	2014	1	obama	0	4
2	1	2016	1	obama	0	5
3	1	2018	1	trump	0	2
4	1	2020	1	trump	0	1
5	2	2014	0	obama	1	2
6	2	2016	0	obama	1	1
7	2	2018	0	trump	1	4
8	2	2020	0	trump	1	3
9	3	2014	0	obama	1	2
10	3	2016	0	obama	1	2
11	3	2018	0	trump	1	4
12	3	2020	0	trump	0	1

Aggregation (means)

Cross-Sectional:

	ID	Year	Female	Pres	GOP	Approve
1	1	2017	1	?	0.00	3.00
2	2	2017	0	?	1.00	2.50
3	3	2017	0	?	0.75	2.25

Temporal:

	Year	Female	Pres	GOP	Approve
1	2014	0.333	obama	0.667	2.67
2	2016	0.333	obama	0.667	2.67
3	2018	0.333	trump	0.667	3.33
4	2020	0.333	trump	0.333	1.67

Aggregation:

- Always loses information
- Sometimes distorts relationships
- Occasionally forces arbitrary decisions

If you have variation in multiple dimensions, use it.

Two-Way Variation

Two “dimensions” of variation:

- Unit-Level Variation (how each unit is – on average – different from other units on average) – a/k/a **between-unit** variation
- Time-Level Variation (how each measurement / time point is – on average – different from other time points on average) – a/k/a/**within-unit** variation

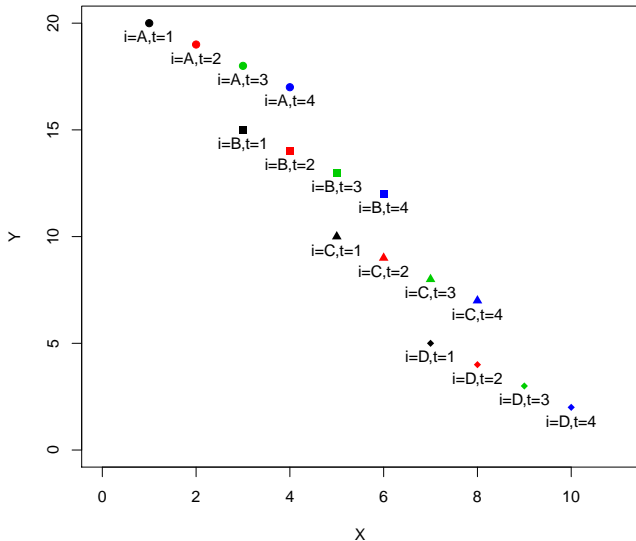
A random variable may:

- Have only between-unit variation (i.e., lack *temporal variation*)
- Have only within-unit variation (i.e., lack *cross-sectional variation*)
- Have *both* within- and between-unit variation

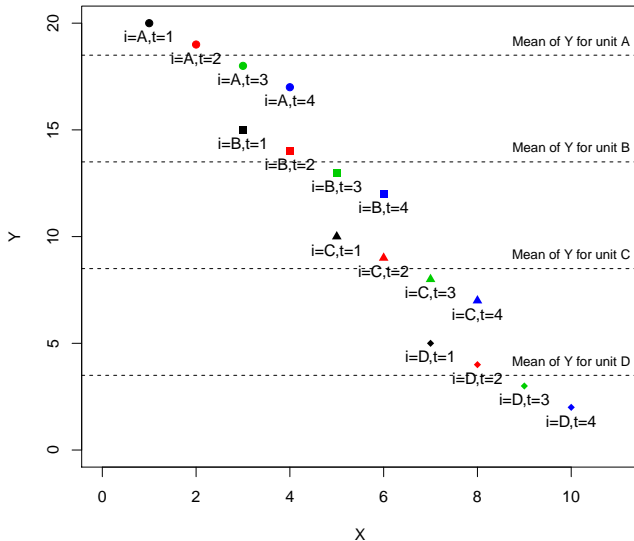
For Y_{it} , a variable that varies over both units and time:

- $\bar{Y}_i = \frac{1}{T_i} \sum_{t=1}^T Y_{it}$ is the over-time mean of Y for unit i ,
- $\bar{Y}_t = \frac{1}{N_t} \sum_{i=1}^N Y_{it}$ is the across-unit mean of Y at time t , and
- $\bar{Y} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T Y_{it}$ is the grand mean of Y .

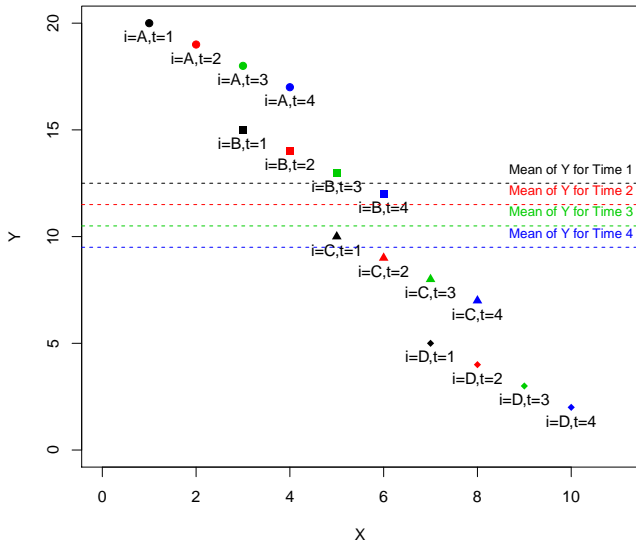
Dimensions of Variation



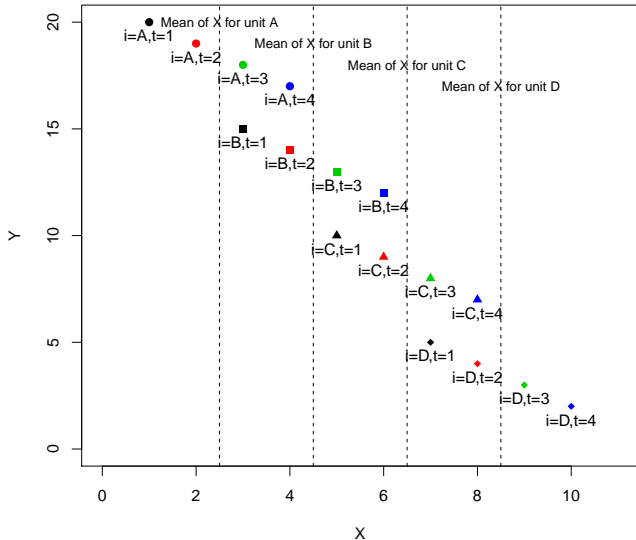
Dimensions of Variation



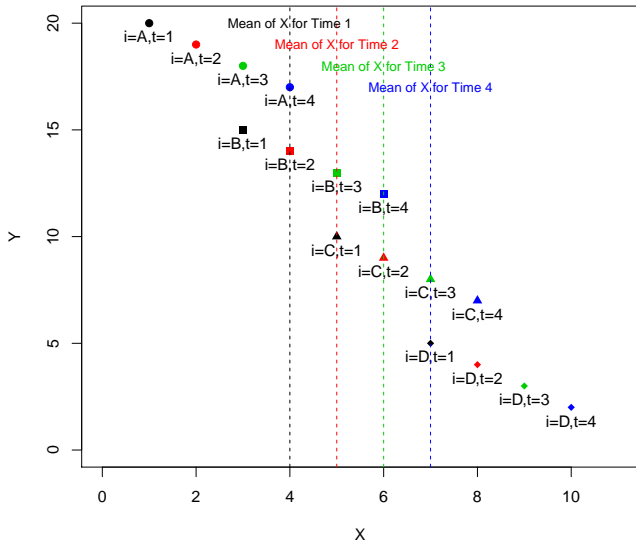
Dimensions of Variation



Dimensions of Variation



Dimensions of Variation



Within- and Between-Unit Variation

The *within-unit* mean of Y is:

$$\bar{Y}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} Y_{it}$$

That means that we can write:

$$Y_{it} = \bar{Y}_i + (Y_{it} - \bar{Y}_i).$$

That is, the *total* variation in Y_{it} can be decomposed into:

- The *between-unit* variation in the \bar{Y}_i s, and
- The *within-unit* variation around \bar{Y}_i (that is, $Y_{it} - \bar{Y}_i$).

Within- and Between-Unit Variation (continued)

Note that (while unusual) one could do a similar decomposition vis-à-vis time:

$$Y_{it} = \bar{Y}_t + (Y_{it} - \bar{Y}_t).$$

That is, the *total* variation in Y_{it} can be decomposed into:

- The *temporal* variation in the \bar{Y}_t s, and
- The *within-time-point* variation around \bar{Y}_t (that is, $Y_{it} - \bar{Y}_t$).

In a similar fashion, we can also calculate the within- and between-unit variability (e.g., the standard deviations) of the constituent variables \bar{Y}_i and $(Y_{it} - \bar{Y}_i)$...

Variation (“Toy” Data from Above)

“Total” Variation:

```
> with(toy, describe(Y))
  vars  n mean  sd median trimmed mad min max range skew kurtosis  se
X1     1 16  11 5.9     11      11 7.4  2  20   18    0    -1.5 1.5
```

“Between” Variation:

```
> Ymeans <- ddply(toy,.(ID),summarise,
+               Y=mean(Y))
> with(Ymeans, describe(Y)) # between-unit variation
  vars n mean  sd median trimmed mad min max range skew kurtosis  se
X1     1 4  11 6.5     11      11 7.4 3.5  18   15    0    -2.1 3.2
```

“Within” Variation:

```
> toy <- ddply(toy,.(ID), mutate,
+             Ymean=mean(Y))
> toy$within <- with(toy, Y-Ymean)
> with(toy, describe(within)) # within-unit variation
  vars  n mean  sd median trimmed mad  min max range skew kurtosis  se
X1     1 16   0 1.1     0      0 1.5 -1.5 1.5    3    0    -1.6 0.29
```

Model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

...makes all the usual OLS assumptions, plus

- $\beta_{0i} = \beta_0 \forall i$
- $\beta_{1i} = \beta_1 \forall i$

For the model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

...the same is true.

Variable Intercepts

Unit-specific intercepts:

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + u_{it} \quad (1)$$

Time-point-specific intercepts:

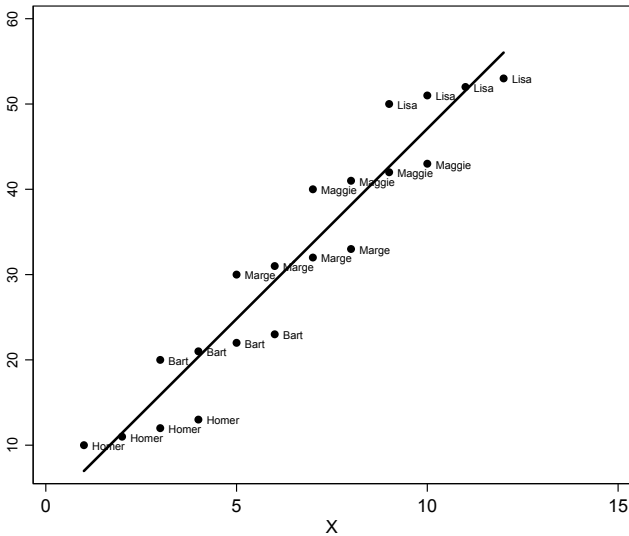
$$Y_{it} = \beta_{0t} + \beta_1 X_{it} + u_{it} \quad (2)$$

Both:

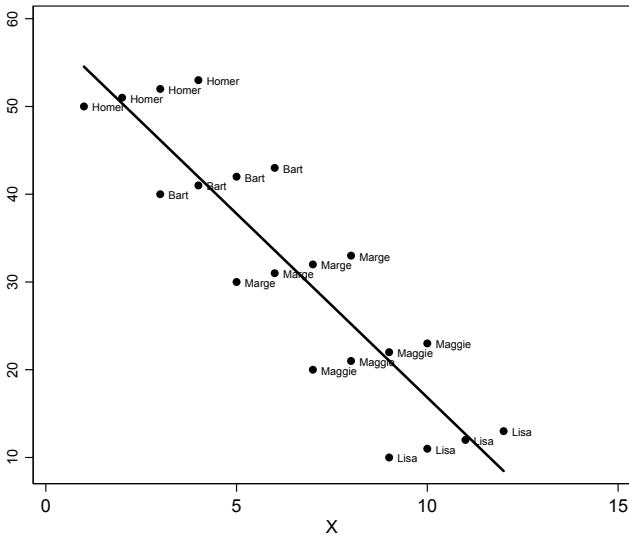
$$Y_{it} = \beta_{0it} + \beta_1 X_{it} + u_{it} \quad (3)$$

Note: Equation 3 is not identified (as written)!

Varying Intercepts



Varying Intercepts



Varying Slopes (+ Intercepts)

Unit-specific slopes:

$$Y_{it} = \beta_0 + \beta_{1i}X_{it} + u_{it} \quad (4)$$

(...one can also have time-point specific slopes, or both – again, the last of those is not identified as written.)

Unit-specific slopes + intercepts:

$$Y_{it} = \beta_{0i} + \beta_{1i}X_{it} + u_{it} \quad (5)$$

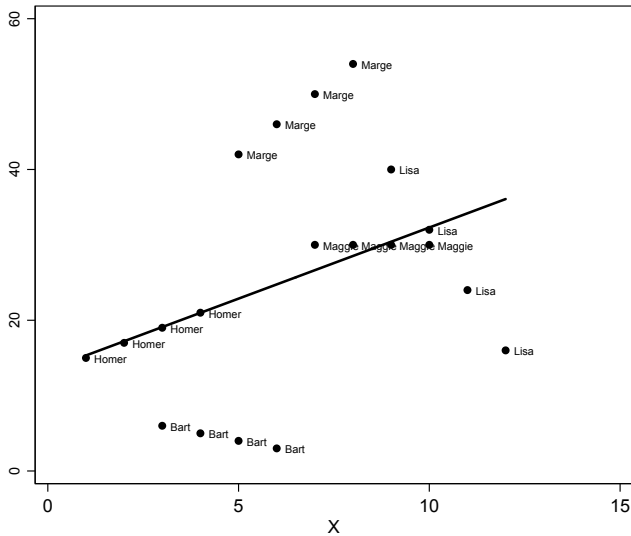
Time-point-specific-specific slopes + intercepts:

$$Y_{it} = \beta_{0t} + \beta_{1t}X_{it} + u_{it} \quad (6)$$

Both...:

$$Y_{it} = \beta_{0it} + \beta_{1it}X_{it} + u_{it} \quad (7)$$

Varying Slopes + Intercepts



The Error Term...

Usual OLS assumption:

$$u_{it} \sim \text{i.i.d.} N(0, \sigma^2) \forall i, t$$

or, equivalently:

$$\mathbf{u}\mathbf{u}' \sim \sigma^2 \mathbf{I}$$

implies:

$$\text{Var}(u_{it}) = \text{Var}(u_{jt}) \forall i \neq j \text{ (i.e., no cross-unit heteroscedasticity)}$$

$$\text{Var}(u_{it}) = \text{Var}(u_{is}) \forall t \neq s \text{ (i.e., no temporal heteroscedasticity)}$$

$$\text{Cov}(u_{it}, u_{js}) = 0 \forall i \neq j, \forall t \neq s \text{ (i.e., no auto- or spatial correlation)}$$

Pooling pros:

- Adds data (\rightarrow precision)
- Enhances generalizability

BUT: fitting the model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

Implies

- that the process governing the relationship between X and Y is exactly the same for each i ,
- that the process governing the relationship between X and Y is the same for all t ,
- that the process governing the us is the same $\forall i$ and t as well.

Q: When can we “pool” data on different units?

“Partial” Pooling (Bartels 1996)

Two regimes:

$$Y_A = \beta'_A \mathbf{X}_A + u_A$$

$$Y_B = \beta'_B \mathbf{X}_B + u_B$$

with $\sigma_A^2 = \sigma_B^2$, and $\text{Cov}(u_A, u_B) = 0$.

Estimators:

$$\hat{\beta}_{A,B} = (\mathbf{X}'_{A,B} \mathbf{X}_{A,B})^{-1} \mathbf{X}'_{A,B} Y_{A,B}$$

and

$$\widehat{\text{Var}}(\hat{\beta}_{A,B}) = \hat{\sigma}_{A,B}^2 (\mathbf{X}'_{A,B} \mathbf{X}_{A,B})^{-1},$$

A Pooled Estimator

$$\begin{aligned}\hat{\beta}_P &= (\mathbf{X}'_A \mathbf{X}_A + \mathbf{X}'_B \mathbf{X}_B)^{-1} (\mathbf{X}'_A Y_A + \mathbf{X}'_B Y_B) \\ &= (\mathbf{X}'_A \mathbf{X}_A + \mathbf{X}'_B \mathbf{X}_B)^{-1} [\beta_A (\mathbf{X}'_A \mathbf{X}_A) + \beta_B (\mathbf{X}'_B \mathbf{X}_B)],\end{aligned}$$

$$\begin{aligned}E(\hat{\beta}_P) &= \beta_A + (\mathbf{X}'_A \mathbf{X}_A + \mathbf{X}'_B \mathbf{X}_B)^{-1} \mathbf{X}'_B \mathbf{X}_B (\beta_B - \beta_A) \\ &= \beta_B + (\mathbf{X}'_A \mathbf{X}_A + \mathbf{X}'_B \mathbf{X}_B)^{-1} \mathbf{X}'_A \mathbf{X}_A (\beta_A - \beta_B)\end{aligned}$$

Pooling: A Test

$$F = \frac{\frac{\hat{\mathbf{u}}_P' \hat{\mathbf{u}}_P - (\hat{\mathbf{u}}_A' \hat{\mathbf{u}}_A + \hat{\mathbf{u}}_B' \hat{\mathbf{u}}_B)}{K}}{\frac{(\hat{\mathbf{u}}_A' \hat{\mathbf{u}}_A + \hat{\mathbf{u}}_B' \hat{\mathbf{u}}_B)}{(N_A + N_B - 2K)}} \sim F_{[K, (N_A + N_B - 2K)]}$$

Bartels suggests:

$$\hat{\beta}_{\lambda} = (\lambda^2 \mathbf{X}'_A \mathbf{X}_A + \mathbf{X}'_B \mathbf{X}_B)^{-1} (\lambda^2 \mathbf{X}'_A Y_A + \mathbf{X}'_B Y_B)$$

with $\lambda \in [0, 1]$:

- $\lambda = 0 \rightarrow$ separate estimators for $\hat{\beta}_A$ and $\hat{\beta}_B$,
- $\lambda = 1 \rightarrow$ “fully pooled” estimator $\hat{\beta}_P$,
- $0 < \lambda < 1 \rightarrow$ a regression where data in regime A are given some “partial” weighting in their contribution towards an estimate of β .

“(R)oughly speaking, it makes sense to pool disparate observations if the underlying parameters governing those observations are sufficiently similar, but not otherwise.”

- Bartels (1996)


Exploring Variation: A Running Example

The U.S. Supreme Court, 1946-2020

- Court has nine “justices” at any time
- Appointed by the President, confirmed by the Senate (simple majority vote)
- Serve for life / good behavior
- One is appointed as the “Chief Justice” (a sitting justice may be elevated to that position)
- Sit in annual “terms” (October through June/July); decide 80-150 cases per term
- Cases are appealed from lower federal and state supreme courts
- Simple majority decision rule (“five”)
- Nearly all decisions have a ideological (left / right) valence (“liberal” vs. “conservative”)

The Supreme Court Database

Washington University Law

THE SUPREME COURT DATABASE				
	ABOUT	DATA	ANALYSIS	DOCUMENTATION
<p>The Supreme Court Database is the definitive source for researchers, students, journalists, and citizens interested in the U.S. Supreme Court. The Database contains over two hundred pieces of information about each case decided by the Court between the 1791 and 2019 terms. Examples include the identity of the court whose decision the Supreme Court reviewed, the parties to the suit, the legal</p>	<p>MODERN Database</p> <p>2020 Release 01</p> <p>released September 21, 2020</p> <p>includes terms 1946 - 2019</p>	<p>LEGACY Database</p> <p>SCDB Legacy 06</p> <p>released September 29, 2019</p> <p>includes terms 1791 - 1945</p>	<p>Are you interested in a particular legal or political issue? Do you seek information about the current Court or about a particular year? Perhaps you are interested in the votes of the Justices in cases about religion, commerce, or another area of the law. The analysis tools allow you to select and summarize cases from the Modern or Legacy Database based on your needs.</p>	<p>Getting Started</p> <p>SCDB Web 101</p> <p>Are you new to the Supreme Court Database? Wondering how to start doing your online analysis? The SCDB Web 101 series can get you underway on the quick.</p> <p>View the 101 Lessons</p> <p>Looking for the Codebook? We have an online and downloadable version. Access</p>

<http://scdb.wustl.edu/>

Supreme Court Panel Data

Structure: One observation per justice (i) per term (t)

Important variables:

- **justice**: A justice (unit) ID variable [range: 78-116]
- **term**: A term (time) variable [range: 1946-2019]
- **LiberalPct**: The percentage of cases in that term in which that justice voted in a politically left / “liberal” direction
- **MajPct**: The percentage of cases in that term in which that justice voted with the majority in a case
- **Ideology**: A variable measuring the justice's (pre-confirmation) political ideology [range: 0 (most conservative) - 1 (most liberal)]*
- **Qualifications**: A measure of the justice's qualifications prior to his/her appointment [range: 0 (least qualified) - 1 (most qualified)]*
- **President**: The name of the president who appointed that justice*
- **YearApptd**: The year that justice was appointed*
- **NCases**: The number of cases the Court decided *during that term***
- **ChiefJustice**: The identity of the Chief Justice *during that term***

* indicates variables that are non-time-varying (that is, that have only between-unit variation)

** indicates variables that are non-unit-varying (that is, that have only within-unit variation)

Summary Statistics

```
> summary(SCData)
```

term	justice	justiceName	LiberalPct	MajPct
Min. :1946	Min. : 78.0	Length:672	Min. :16.7	Min. : 46.7
1st Qu.:1964	1st Qu.: 91.0	Class :character	1st Qu.:38.4	1st Qu.: 76.2
Median :1982	Median :100.0	Mode :character	Median :49.5	Median : 82.8
Mean :1982	Mean : 98.1		Mean :51.9	Mean : 81.7
3rd Qu.:2001	3rd Qu.:106.0		3rd Qu.:65.7	3rd Qu.: 88.0
Max. :2019	Max. :116.0		Max. :87.7	Max. :100.0

Order	Nominee	SenateVote	Ideology	Qualifications
Min. : 1.0	Length:672	Length:672	Min. :0.000	Min. :0.125
1st Qu.:15.0	Class :character	Class :character	1st Qu.:0.160	1st Qu.:0.750
Median :27.0	Mode :character	Mode :character	Median :0.488	Median :0.885
Mean :24.5			Mean :0.488	Mean :0.802
3rd Qu.:36.0			3rd Qu.:0.750	3rd Qu.:0.978
Max. :47.0			Max. :1.000	Max. :1.000

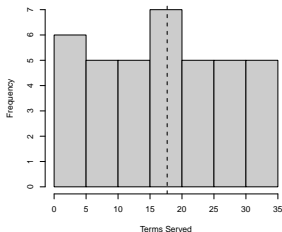
President	YearApptd	NCases	ChiefJustice
Length:672	Min. :1937	Min. : 76	Length:672
Class :character	1st Qu.:1955	1st Qu.: 96	Class :character
Mode :character	Median :1970	Median :141	Mode :character
	Mean :1970	Mean :142	
	3rd Qu.:1988	3rd Qu.:182	
	Max. :2018	Max. :258	

Some Basics

How many justices are in the data?

```
> length(unique(SCData$justice))  
[1] 38
```

How many terms do justices typically serve?

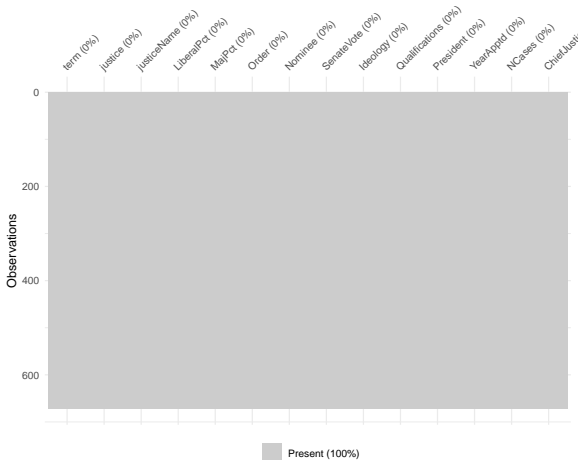


So we have:

- $N = 38$ units (justices)
- $\bar{T} = 17.7$ time periods (terms) of data per justice, on average [range: 2-35], for a total of
- $NT = 672$ justice-terms in the data

Missing Data

```
> library(naniar)
> vis_miss(SCData)
```



Variation: LiberalPct

```
> # Total variation:
>
> with(SCData, describe(LiberalPct))
  vars   n mean   sd median trimmed mad  min  max range skew kurtosis   se
X1     1 672 51.9 16.2  49.5    51.5  19 16.7 87.7   71 0.19      -1 0.63

> # Between-Justice variation:
>
> LibMeans <- ddply(SCData,.(justice),summarise,
+                   MeanLibPct=mean(LiberalPct))
> with(LibMeans, describe(MeanLibPct))
  vars   n mean   sd median trimmed mad  min  max range skew kurtosis   se
X1     1 38 52.5 14.8  48.3    52.2 16.5 29.9 77.2  47.2 0.31      -1.29 2.39

> # Within-Justice variation:
>
> SCData <- ddply(SCData,.(justice), mutate,
+                 LibMean=mean(LiberalPct))
> SCData$LibWithin <- with(SCData, LiberalPct-LibMean)
> with(SCData, describe(LibWithin))
  vars   n mean   sd median trimmed mad  min  max range skew kurtosis   se
X1     1 672   0 7.36  -0.16  -0.02 7.05 -30.8 32.8  63.6 0.04      1.03 0.28
```

Variation: Ideology

```
> # Total variation:
>
> with(SCData, describe(Ideology))
  vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1     1 672 0.49 0.32   0.49   0.48 0.39   0   1     1 0.09    -1.3 0.01

> # Between-Justice variation:
>
> IdeoMeans <- ddply(SCData,.(justice),summarise,
+                   MeanIdeo=mean(Ideology))
> with(IdeoMeans, describe(MeanIdeo))
  vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1     1  38 0.54 0.33   0.58   0.54 0.43   0   1     1 -0.11    -1.46 0.05

> # Within-Justice variation (hint - there is none):
>
> SCData <- ddply(SCData,.(justice), mutate,
+               IdeoMean=mean(Ideology))
> SCData$IdeoWithin <- with(SCData, Ideology-IdeoMean)
> with(SCData, describe(IdeoWithin))
  vars   n mean sd median trimmed mad min max range skew kurtosis se
X1     1 672   0  0       0       0  0  0  0  0     0  NaN    NaN  0
```

Variation: NCases

```
> # Total variation:
>
> with(SCData, describe(NCases))
  vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1     1 672 142 44.7   141     141 60.8  76 258   182 0.18   -1.06 1.73

> # Between-Term variation:
>
> NCMeans <- ddply(SCData,.(term),summarise,
+                   MeanNCases=mean(NCases))
> with(NCMeans, describe(MeanNCases))
  vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1     1  74 142 45.1   142     141 60.8  76 258   182 0.17   -1.11 5.24

> # Within-Term variation (none):
>
> SCData <- ddply(SCData,.(term), mutate,
+                 NCMean=mean(NCases))
> SCData$NCWithin <- with(SCData, NCases-NCMean)
> with(SCData, describe(NCWithin))
  vars   n mean sd median trimmed mad min max range skew kurtosis se
X1     1 672   0  0       0       0   0   0   0   0   0   NaN   NaN   0
```

An interactive tool for exploring panel data...

- Creator: Joachim Gassen (Department of Accounting, Humboldt-Universität zu Berlin)
- Built upon / consistent with `ggplot` / `tidyverse`
- Requires installing the ExPanDaR package
- Calling

```
> ExPanD()
```

...opens the Shiny app, and asks for a (pre-formatted) data frame (typically in CSV format)

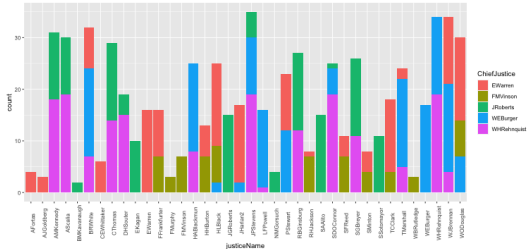
- More information is here:

<https://joachim-gassen.github.io/ExPanDaR/>

Some examples...

ExPanDaR: Summaries

Counts by factors:



Summary statistics:

Descriptive Statistics

Hover over variable names with mouse to see variable definitions.

Select Tab to choose the analysis set of variables or the base set of variables (to define new variables).

Click here to delete selected variables from the analysis sample.

Delete Variables

Analysis Set

Base Set

Variable	N	Mean	Std. dev.	Min.	25 %	Median	75 %	Max.
V1	672	351.475	208.836	1.000	168.750	336.500	538.250	707.000
LiberalPct	672	51.856	16.213	16.667	38.446	49.473	65.713	87.662
MajPct	672	81.714	8.827	46.667	76.193	82.828	87.964	100.000
Order	672	24.531	12.784	1.000	15.000	27.000	36.000	47.000
Ideology	672	0.488	0.320	0.000	0.160	0.487	0.750	1.000
Qualifications	672	0.802	0.245	0.125	0.750	0.885	0.978	1.000
YearApptd	672	1,970.324	20.975	1,837.000	1,955.000	1,970.000	1,988.000	2,018.000

Histograms:

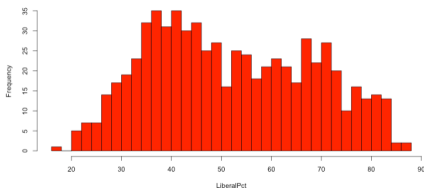
Histogram

Select variable to display

LiberalPct

Suggested number of cells

5 10 250



Outlier Detection:

Extreme Observations

Select variable to sort data by

LiberalPct

Select period to subset to

All

justice	term	LiberalPct
81	1958	87.7
81	1956	87.0
81	1971	84.7
90	1963	84.1
81	1955	83.7
...
102	1979	21.3
108	2003	21.3
102	1998	20.2
108	1998	20.2
115	2016	16.7

ExPanDaR: More Distributions

Bar Charts of Means (by factors):

By Group Bar Chart

Select variable to
display

LiberalPct

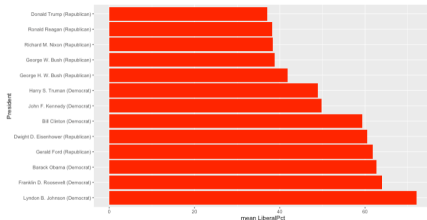
Select variable to
group by

President

Select statistic to
display

Mean

☒ Sort by statistic



Violin Plots:

By Group Violin Chart

Select variable to
display

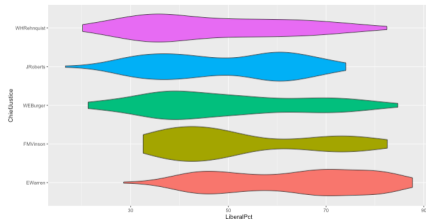
LiberalPct

Select variable to
group by

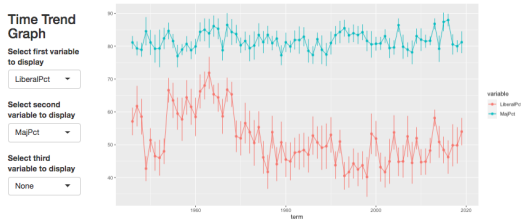
ChiefJustice

☒ Sort by group
means

(Note: Consider
treating your outliers
if this graph looks
odd)



General Trends + Variation:



Trends in Quantiles:



ExPanDaR: More Trends

Trends By Group:

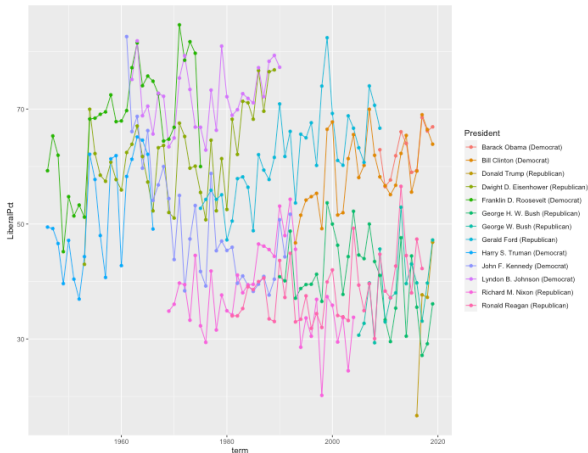
By Group Time Trend Graph

Select variable to
display

LiberalPct ▾

Select variable to
group by

President ▾



Fancy Scatterplots:

Scatter Plot

Select the x
variable to display

Ideology

Select the y
variable to display

LiberalPct

Select the variable
to be reflected by
dot size

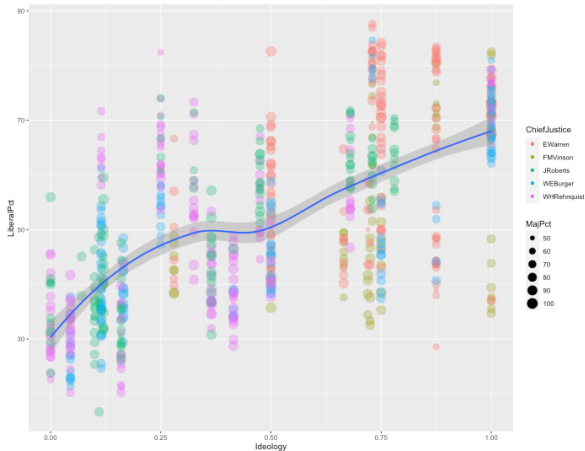
MajPct

Select the variable
to be reflected by
color

ChiefJustice

☐ Sample 1,000
observations to
display if number
of observations is
higher

☒ Display smoother



ExPanDaR: Correlations / Regression

Bivariate Correlations:

Correlation Plot

This plot visualizes sample correlations (Pearson above, Spearman below diagonal). Reports correlations for all continuous variables. Hover over ellipse to get rho, P-Value and n.



Regression (OLS) Analysis:

Regression Analysis

Select the dependent variable

LiberalPct

Select independent variable(s)

Ideology

MajPct

Select a categorical variable as the first fixed effect

None

Dependent variable:	
LiberalPct	
Ideology	31.200*** (1.490)
MajPct	-0.287*** (0.054)
Constant	60.100*** (4.650)
Estimator	ols
Fixed effects	None
Std. errors clustered	No
Observations	672
R ²	0.445
Adjusted R ²	0.443

Note: *** p<0.01; ** p<0.05; * p<0.1

- Tuesday, 7 June: One- and Two-Way “Unit Effects” Models (fixed, “random,” etc.)
- Wednesday, 8 June: Dynamics in Panel Data
- Thursday, 9 June: Panel Data and Causal Inference
- Friday, 10 June: Models for Discrete Responses