

GSERM - St. Gallen 2025

Analyzing Panel Data

June 16, 2025

Logistics:

- Instructor: Prof. Christopher Zorn
 - Email: zorn@psu.edu
 - Phone: +1-803-553-4077
 - Bluesky / Instagram / etc.: [@prisonrodeo](https://bluesky.social/@prisonrodeo)
- Class: June 16-20, 2025, 09:15-15:15 CET.
- 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-2025>).

PrisonRodeo / GSERM-Panel-2025

Type to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

GSERM-Panel-2025 Public

Unpin Watch 0 Fork 0 Star 1

main 1 Branch 0 Tags Go to file Code

	PrisonRodeo	Useful R Resources
		7949fc2 · 2 minutes ago
	Code	Code
	Data	WDI Description
	Misc. Materials	Miscellaneous Materials
	Readings	Readings
	.gitattributes	Initial commit
	.gitignore	Create .gitignore
	GSERM-2025-Useful-R-Resources...	Useful R Resources
	GSERM-Panel-WDI-Description-20...	WDI Description

About

Analyzing Panel Data - GSERM 2025

Readme Activity 1 star 0 watching 0 forks

Releases

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

Packages

No packages published

Evaluation at GSERM isn't easy... the plan:

- One “homework exercise”
 - Practical exercise – “real” data analysis and discussion
 - Assigned Tuesday (June 17); due Friday (June 20)
 - Worth 300 possible points
- Final Examination
 - Multiple essay-style questions + “real” data analysis
 - Some choice of questions to answer
 - Assigned Friday (June 14)
 - Due either Friday, June 20, 2025 (“in-class” alternative) or Friday, June 27, 2025 (“take-home” alternative)
 - Worth 700 possible points
- Total course = 1000 possible points
- Grades assessed on Swiss (1 - 6) scale

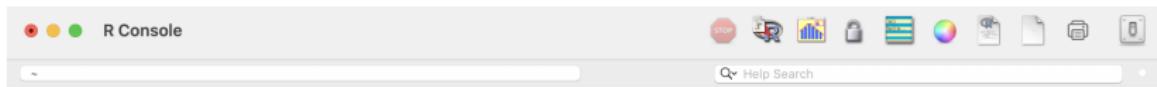
R

- All examples, plots, etc. are generated using R
- Current version (as of last Friday) is 4.5.1
- Desktop: Be sure to get the [RStudio / Posit IDE...](#)
- Alternatively: Can be run in a browser, using [Posit Cloud](#)
- The course Github repo contains a bit of [introductory code](#) for people who may never have used R, and a list of [R resources](#).
- A few of the primary packages we'll use include:
 - `plm`
 - `lme4`
 - `gee`

See the [econometrics R task view](#) for more.

Stata

- Current version is v. 19
- Mostly use the `-xt-` series of commands (for “cross-sectional time series”)



```
R version 4.4.0 (2024-04-24) -- "Puppy Cup"  
Copyright (C) 2024 The R Foundation for Statistical Computing  
Platform: aarch64-apple-darwin20
```

```
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 1.80 (8376) aarch64-apple-darwin20]
```

```
[History restored from /Users/cuz10/.Rapp.history]
```

```
> |
```

RStudio / Posit

The screenshot shows the RStudio IDE interface. The top navigation bar includes tabs for Environment, History, Connections, and Tutorial, along with a Project dropdown set to (None). The main workspace on the right displays the message "Environment is empty". Below this are tabs for Files, Plots, Packages, Help, Viewer, and Presentation, with sub-options like Zoom and Export. The bottom-left pane is the R Console, showing the startup message for R version 4.4.0, followed by standard R license and copyright information. The status bar at the bottom indicates the R version is 4.4.0 and the current working directory is ~/.

R version 4.4.0 (2024-04-24) -- "Puppy Cup"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20

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.

RStudio (annotated)

This screenshot shows the RStudio interface with several annotations:

- Source window (left):** A text editor where you type your R code. Annotations include:
 - A red circle highlights the "Save" button in the toolbar.
 - A red circle highlights the "Run" button in the toolbar.
 - The text "Click here to save your source code. Save often!" is displayed above the toolbar.
 - The text "This is the 'Source' window." is displayed below the toolbar.
 - A list of bullet points:
 - 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").
 - The text "You'll spend most of your time working here."- Environment window (top right):** Shows the Global Environment tab. Annotations include:
 - The text "Highlight text in the Source window, then click this button to 'run' the code."
 - The text "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"
 - The text "There's also a 'History' tab above; switching to that will show what has transpired in the Console window recently."
- Console window (bottom left):** Shows the command line interface. Annotations include:
 - A red circle highlights the "Working Directory" dropdown.
 - The text "This is the 'working directory.' Anything you save will be saved here, unless you tell the program to save it somewhere else."
 - The text "Console ~/Dropbox (Personal)/" is shown in the dropdown.
 - The text "Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help."
 - The text "Type 'q()' to quit R."
 - The text "This is the 'Console.' When you run the code in the Source window, the results that aren't graphics appear here."

- 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) ($N = \text{large}, T = 25$)
- 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, $T \approx 60$)
- 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

Typical: “long”:

id	t	Y	X	...
1	1	250	3.4	...
1	2	290	3.3	...
⋮	⋮	⋮	⋮	...
2	1	160	4.7	...
2	2	150	4.9	...
⋮	⋮	⋮	⋮	...

Sometimes: “wide”:

id	Y1	Y2	...	X1	X2	...
1	250	290	...	3.4	3.3	...
2	160	1250	...	4.7	4.9	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Introduction to Panel Variation: A Tiny (Fake) Example

Firm	Year	Sector	Inflation	FemaleCEO	NPM
1	2021	IT	0.6	0	21.6
1	2022	IT	2.8	0	19.9
1	2023	IT	2.1	0	23.1
1	2024	IT	1.1	0	22.0
2	2021	Energy	0.6	0	13.7
2	2022	Energy	2.8	0	12.1
2	2023	Energy	2.1	0	12.3
2	2024	Energy	1.1	0	13.0
3	2021	Retail	0.6	1	5.1
3	2022	Retail	2.8	1	4.9
3	2023	Retail	2.1	1	4.8
3	2024	Retail	1.1	0	5.0

Aggregation (means)

Cross-Sectional:

Firm	Year	Sector	Inflation	FemaleCEO	NPM
1	2022	IT	1.65	0.00	21.65
2	2022	Energy	1.65	0.00	12.78
3	2022	Retail	1.65	0.75	4.95

Temporal:

Year	Firm	Inflation	FemaleCEO	NPM
2021	2	0.6	0.333	13.5
2022	2	2.8	0.333	12.3
2023	2	2.1	0.333	13.4
2024	2	1.1	0.000	13.3

Aggregation:

- *Always* loses information
- Occasionally forces *arbitrary decisions* / transformations
- Sometimes *distorts* relationships

If you have variation in multiple dimensions, use it.

Two-Way Variation

Two “dimensions” of variation:

- Cross-sectional variation: how each unit is – on average – different from other units – a/k/a **between-unit** variation
- Temporal variation: how each measurement / time point is – on average – different from other time points on average – a/k/a/ **within-unit** variation

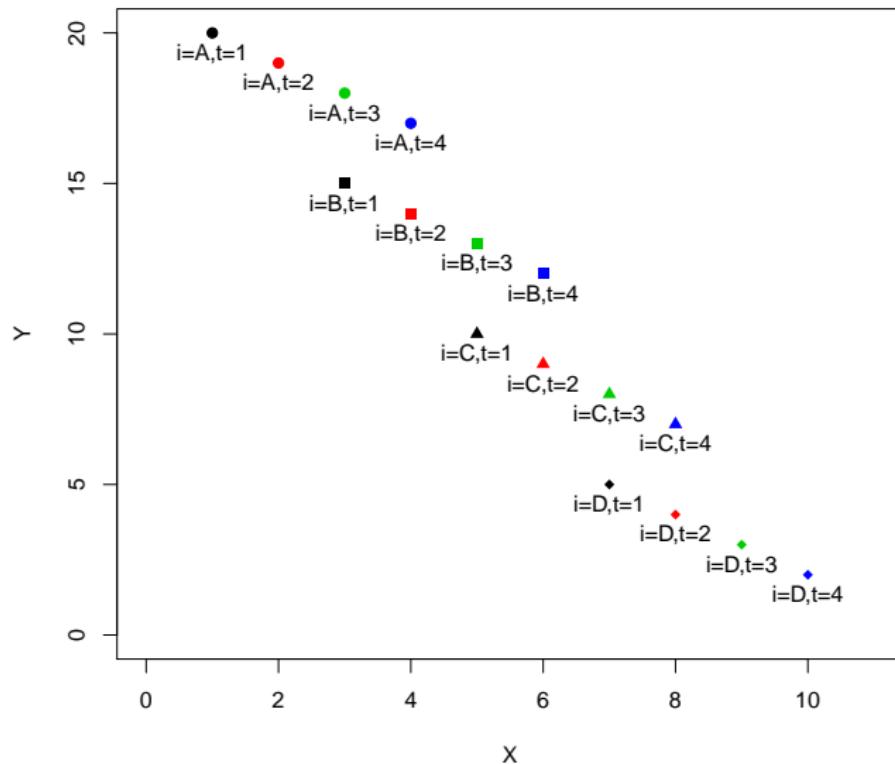
In panel data, a random variable may:

- 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

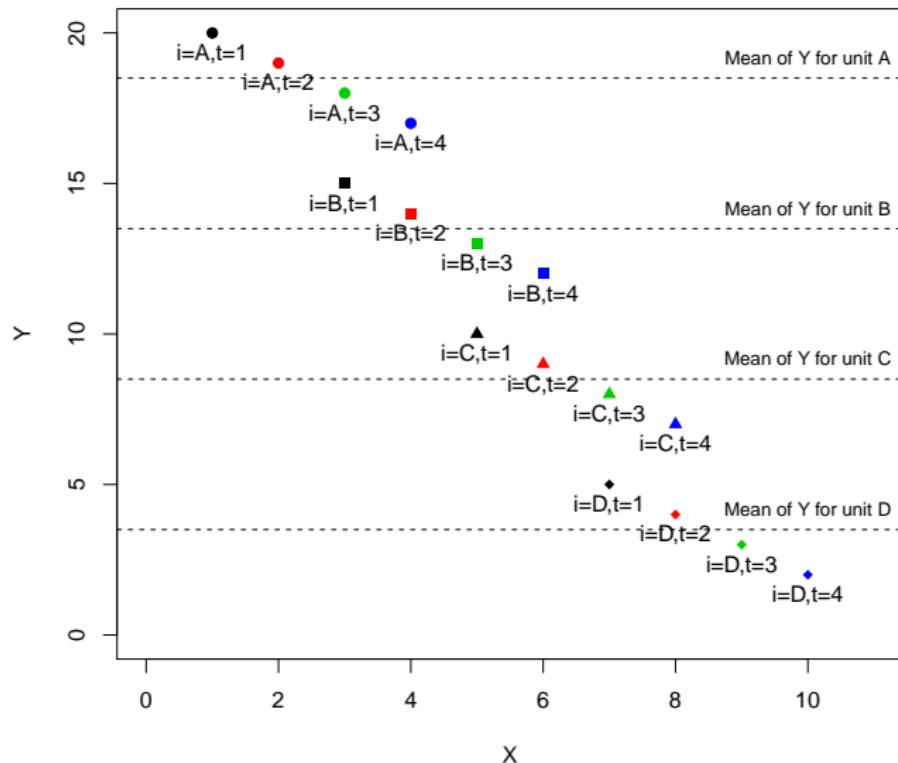
For Y_{it} , a variable that varies over both units and time:

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

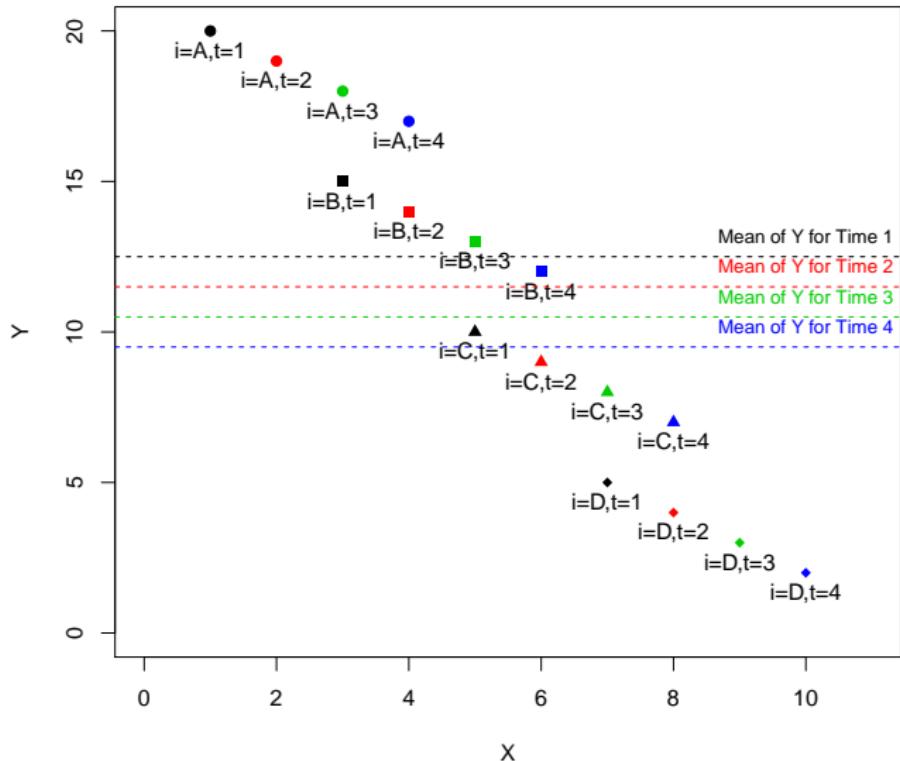
Dimensions of Variation



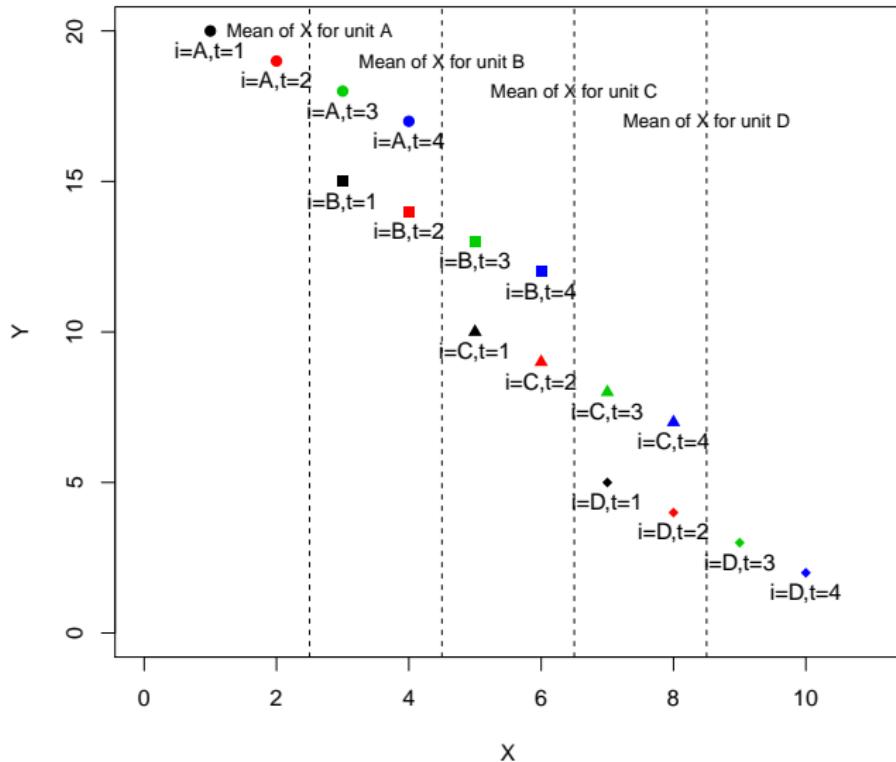
Dimensions of Variation: Y



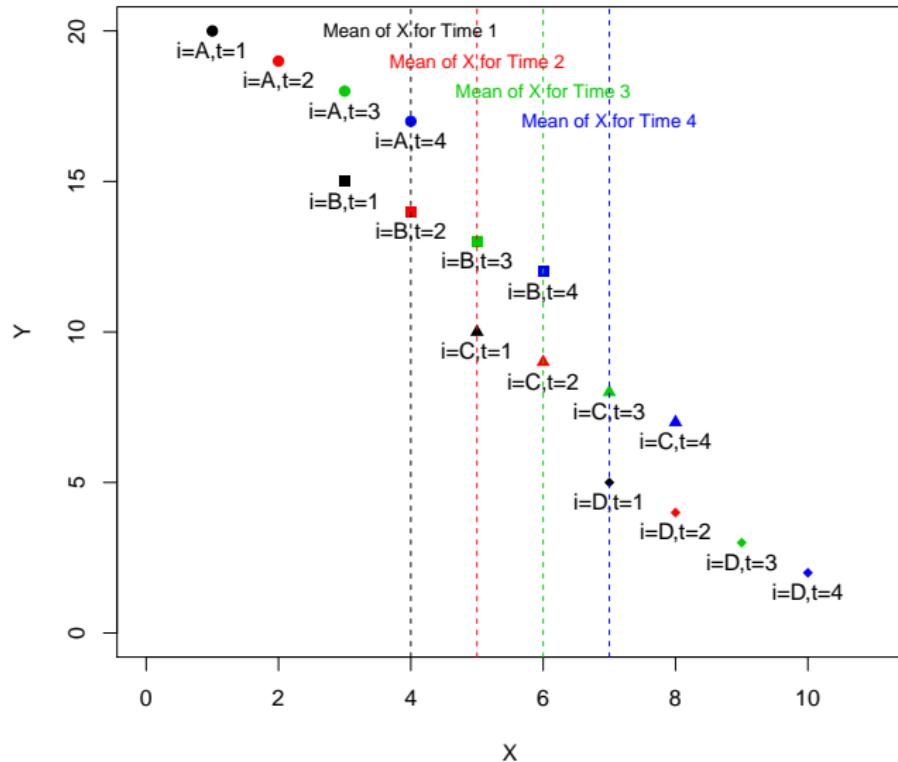
Dimensions of Variation: Y



Dimensions of Variation: X



Dimensions of Variation: X



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-Time Point Variation

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, Y)

“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
```

Cross-sectional regression:

$$\underset{N \times 1}{\mathbf{Y}_i} = \underset{N \times K}{\mathbf{X}_i} \underset{K \times 1}{\boldsymbol{\beta}} + \underset{N \times 1}{\mathbf{u}_i}$$

...requires all the usual OLS assumptions:

- $E(\mathbf{u} = 0)$
- $E(\mathbf{u}\mathbf{u}') = \sigma^2 \mathbf{I}$
- $\text{Cov}(\mathbf{X}, \mathbf{u}) = 0$
- $\text{Rank}(\mathbf{X}) = K$

In addition, we usually assume:

- $\beta_i = \boldsymbol{\beta} \forall i$

Regression with Panel Data

Key point: For the model

$$\mathbf{Y}_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{u}_{it}$$

...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:

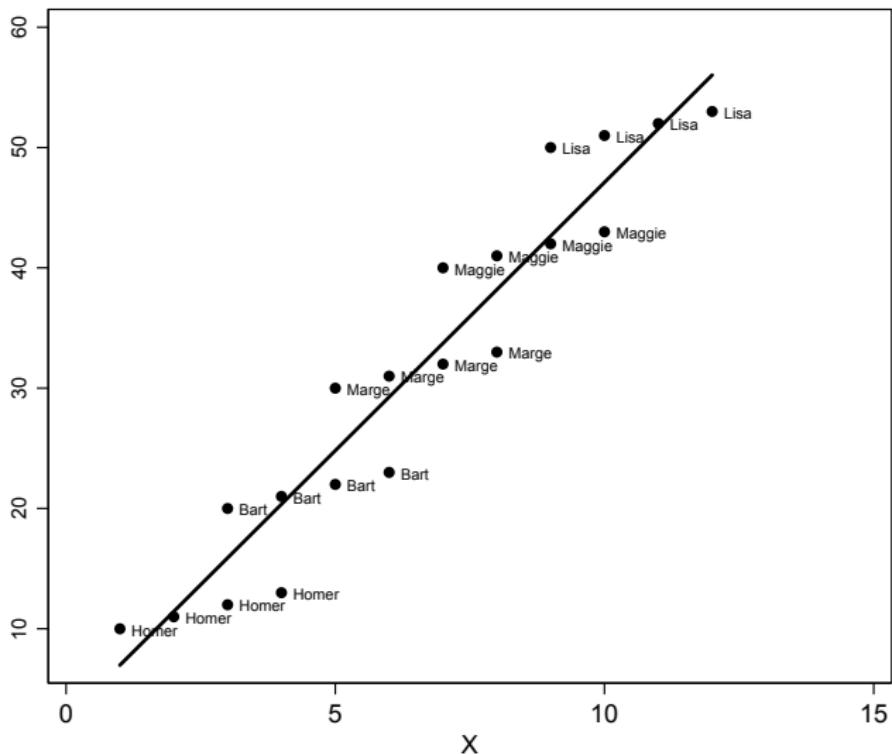
$$Y_{it} = \beta_{0t} + \beta_1 X_{it} + u_{it} \quad (2)$$

Both:

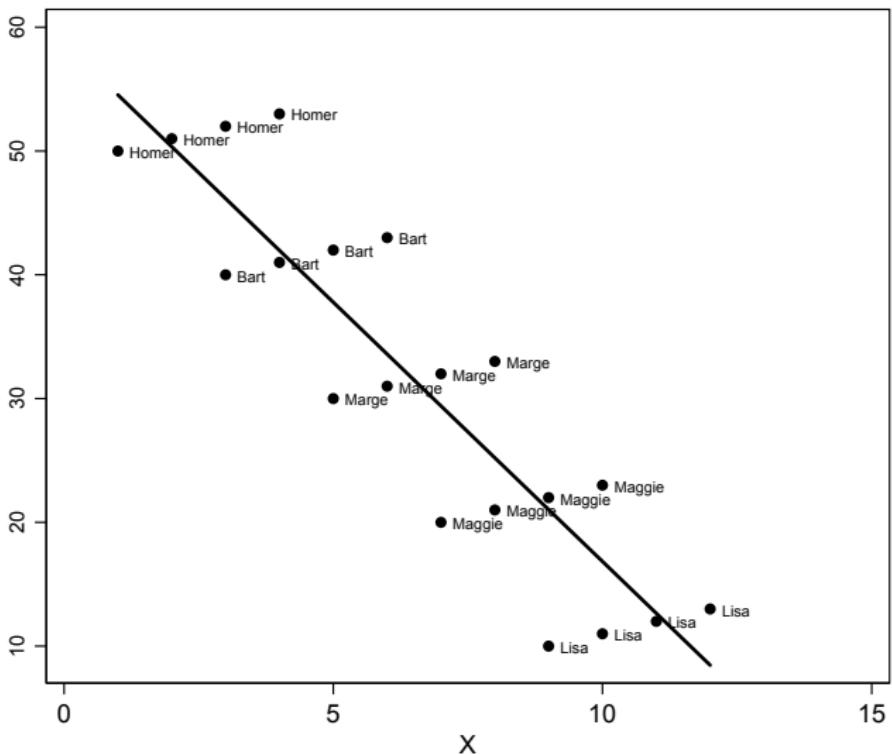
$$Y_{it} = \beta_{0it} + \beta_1 X_{it} + u_{it} \quad (3)$$

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:

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

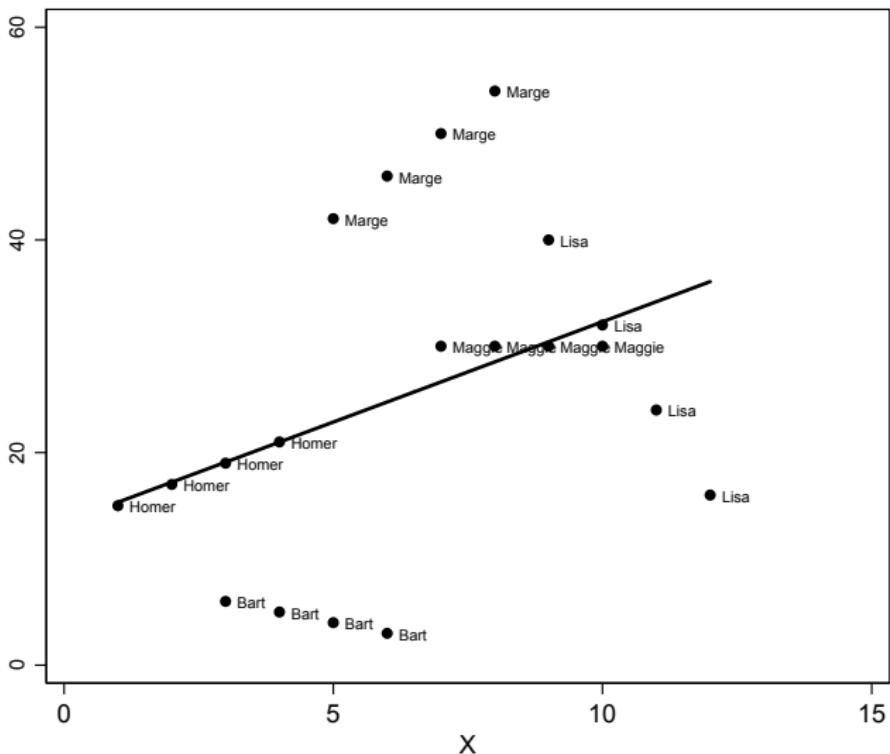
Time-point-specific slopes + intercepts:

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

Both...:

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

Varying Slopes + Intercepts



The Error Term...

Usual OLS assumption:

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

or, equivalently:

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

implies:

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

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

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

Pooling = combining (repeated) observations on different units, and/or observations on different time points, into a single data frame.

Why should we pool data?:

- Adds data (\rightarrow increases *precision*)
- Enhances *generalizability*

Every panel dataset requires that we make decisions about pooling.

Fitting (say) 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

Pooling A s and B s gives:

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

What is the expectation?

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

We can assess whether $\hat{\beta}_A = \hat{\beta}_B$ via:

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

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)

Our “Running Example” Data: WDI, 1960-2024

The World Development Indicators

- Cross-national country-level panel data
- $N = 215$ countries, $T = 65$ years (1960-2024) + missingness
- Variable types:
 - Geography: land area, arable land
 - Population indicators
 - Demographics: Birth rates, life expectancy, etc.
 - Economics: GDP, inflation, trade, FDI, etc.
 - Governments: expenditures, policies, etc.
- Full descriptions are listed in the Github repo [here](#).

Summary Statistics: WDI Data

	vars	n	mean	sd	skew	kurtosis	se
ISO3	1	13975	NaN	NA	NA	NA	NA
Year	2	13975	1992.00	18.76	0.00	-1.20	0.16
Region	3	13975	NaN	NA	NA	NA	NA
country	4	13975	NaN	NA	NA	NA	NA
iso3c	5	13975	NaN	NA	NA	NA	NA
LandArea	6	12068	596744.44	1643440.14	5.45	34.69	14960.16
ArablePercent	7	11662	13.40	13.67	1.47	1.97	0.13
Population	8	13730	25346115.55	105429954.52	9.73	105.72	899764.29
PopGrowth	9	13513	1.75	1.81	0.73	21.38	0.02
RuralPopulation	10	13696	48.11	25.74	-0.11	-1.00	0.22
UrbanPopulation	11	13696	51.89	25.74	0.11	-1.00	0.22
BirthRatePer1K	12	13730	27.67	13.17	0.24	-1.24	0.11
FertilityRate	13	13728	3.82	1.99	0.43	-1.18	0.02
PrimarySchoolAge	14	11120	6.13	0.61	-0.04	0.11	0.01
LifeExpectancy	15	13726	65.17	11.25	-0.76	0.18	0.10
AgeDepRatioOld	16	13730	10.76	7.16	1.79	5.00	0.06
CO2Emissions	17	10962	5.05	11.40	8.70	102.13	0.11
GDP	18	11040	240200887836.82	1130067880812.37	11.48	160.11	10755237244.82
GDPPerCapita	19	11045	12332.27	19272.78	3.11	14.10	183.38
GDPPerCapGrowth	20	10970	1.89	6.49	1.68	41.63	0.06
Inflation	21	8882	22.90	320.90	54.09	3625.88	3.40
TotalTrade	22	8777	77.93	53.26	3.03	18.39	0.57
Exports	23	8777	36.19	28.54	2.99	16.80	0.30
Imports	24	8786	41.75	27.29	2.54	13.97	0.29
FDIIn	25	8969	5.53	45.10	15.13	541.91	0.48
AgriEmployment	26	6134	28.11	23.70	0.66	-0.67	0.30
NetAidReceived	27	9043	506951242.00	997064633.65	8.32	157.34	10484966.48
MobileCellSubscriptions	28	10212	36.32	51.76	1.29	1.14	0.51
NaturalResourceRents	29	9211	6.85	11.06	2.60	8.04	0.12
MilitaryExpenditures	30	7733	2.71	3.18	9.43	239.85	0.04
GovtExpenditures	31	8421	16.34	8.27	3.63	32.81	0.09
PublicEdExpend	32	5177	4.35	2.14	6.89	166.21	0.03
PublicHealthExpend	33	4346	3.29	2.37	1.33	3.08	0.04
HIVDeaths	34	4656	6473.06	18922.25	5.78	45.97	277.31
WomenBusLawIndex	35	10152	59.85	18.74	0.02	-0.58	0.19
PaidParentalLeave	36	10152	0.11	0.31	2.50	4.27	0.00
PostColdWar	37	13975	0.54	0.50	-0.15	-1.98	0.00

Missing Data

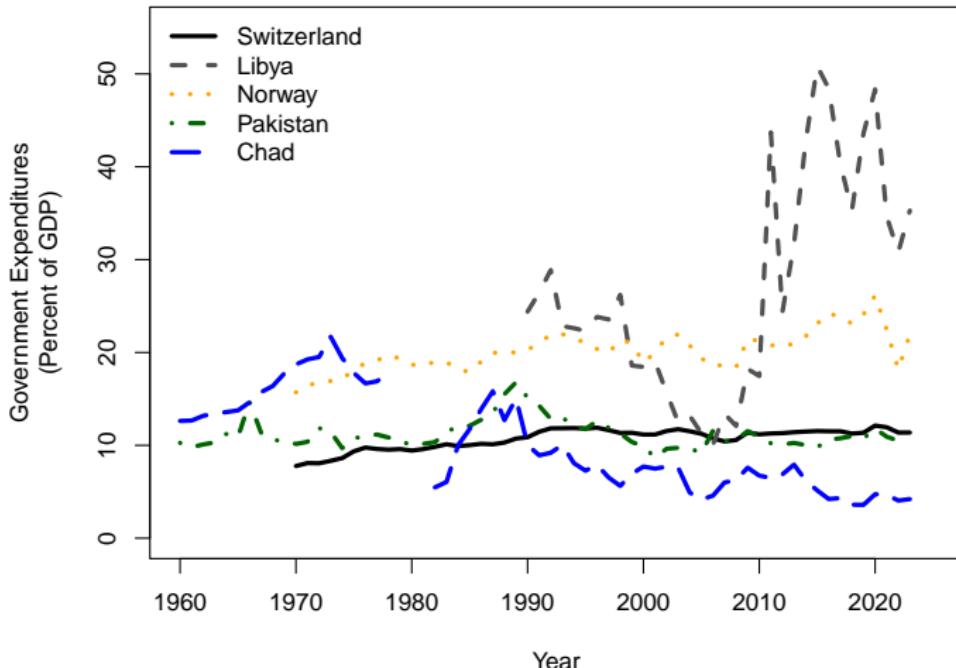
```
> library(naniar)  
> vis_miss(wdi)
```



Exploring Variation: GovtExpenditures

General Government Final Consumption Expenditure (% of GDP)

"General government final consumption expenditure (formerly general government consumption) includes all government current expenditures for purchases of goods and services (including compensation of employees). It also includes most expenditures on national defense and security, but excludes government military expenditures that are part of government capital formation."



Panel Summary Statistics: GovtExpenditures

Summary: Government Expenditures

Variable	Dim	Mean	SD	Min	Max	Observations
GovtExpenditures	overall	16.338	8.27	0	147.735	$N = 8421$
	between		9.721	2.26	79.279	$N = 186$
	within		4.364	-34.745	84.794	$T = 45.274$

This means:

- The overall average of GovtExpenditures is 16.3
- The overall standard deviation of GovtExpenditures is 8.3
- The between-country standard deviation (standard deviation of the 185 country-level means) of GovtExpenditures is 9.7
- The within-country standard deviation of GovtExpenditures is 4.4
- GovtExpenditures is *cross-sectionally dominated*

Exploring Variation: Inflation

Inflation, Consumer Prices (Annual %)

"Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used."

Summary: (logged) Inflation¹

Variable	Dim	Mean	SD	Min	Max	Observations
lnInflation	overall	1.661	1.293	-6.909	10.076	$NT = 8342$
	between		0.779	-1.2	3.978	$N = 191$
	within		1.101	-6.559	8.854	$T = 43.675$

The point: Inflation is *temporally dominated*.

¹Note: Logging inflation leads to missing values in the (relatively rare) instances where deflation occurs.

Exploring Variation: MidEast

Create a variable that = 1 if the country is in the Middle East, 0 otherwise:

```
> wdi$MidEast<-ifelse(wdi$Region=="Middle East & North Africa",1,0)
```

Summary: Mid-East

Variable	Dim	Mean	SD	Min	Max	Observations
MidEast	overall	0.098	0.297	0	1	$NT = 13975$
	between		0.298	0	1	$N = 215$
	within		0	0.098	0.098	$T = 65$

Note:

- Approximately 9.8 percent of all the country-years in the data are for countries defined as being in the Middle East
- MidEast has *no* within-country variation, so...
- ...it has the same value for a particular country for every year in which that country is in the data.

Exploring Variation: PostColdWar

The PostColdWar variable is coded:

- 0 if that observation occurred \leq 1989
- 1 if that observation occurred \geq 1990

Summary: Post-Cold War

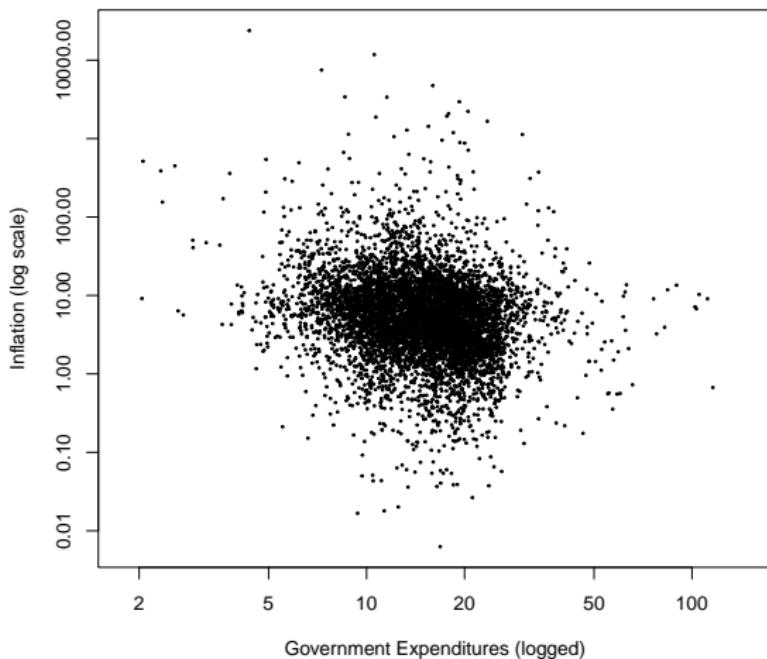
Variable	Dim	Mean	SD	Min	Max	Observations
PostColdWar	overall	0.538	0.499	0	1	$NT = 13975$
	between		0	0.538	0.538	$N = 215$
	within		0.499	0	1	$T = 65$

Note:

- Approximately 54 percent of all the country-years in the data occur after the end of the Cold War
- PostColdWar has *no* between-country variation, so...
- ...in a given year, it has the same value for every country that is in the data that year.

Regression!

Q: What, if anything, is the (contemporaneous) relationship between (logged) inflation rates and (logged) government spending?



Inflation: Regression

Regression of lnInflation on (logged) GovtExpenditures: all / pooled data:

```
> Inf.fit<-lm(lnInflation~lnGovtExp,data=wdi)
> summary(Inf.fit)
```

Call:

```
lm(formula = lnInflation ~ lnGovtExp, data = wdi)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.808	-0.689	-0.005	0.714	7.711

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.2578	0.1042	31.3	<2e-16 ***
lnGovtExp	-0.6060	0.0383	-15.8	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

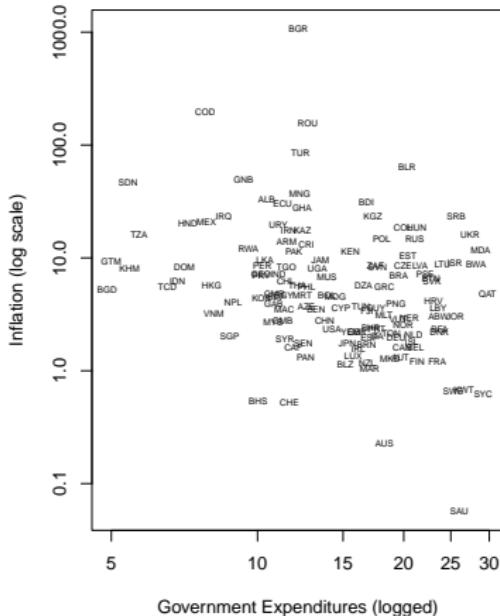
Residual standard error: 1.26 on 6583 degrees of freedom

(7390 observations deleted due to missingness)

Multiple R-squared: 0.0366, Adjusted R-squared: 0.0364

F-statistic: 250 on 1 and 6583 DF, p-value: <2e-16

Inflation Regression: 1997 only



Inflation and spending, 1997:

```
> Inf.fit97<-lm(lnInflation~lnGovtExp,data=wdi97)
> summary(Inf.fit97)

Call:
lm(formula = lnInflation ~ lnGovtExp, data = wdi97)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.930	-0.758	-0.207	0.741	5.125

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.312	0.723	5.97	0.000000021 ***
lnGovtExp	-0.990	0.267	-3.71	0.00031 ***

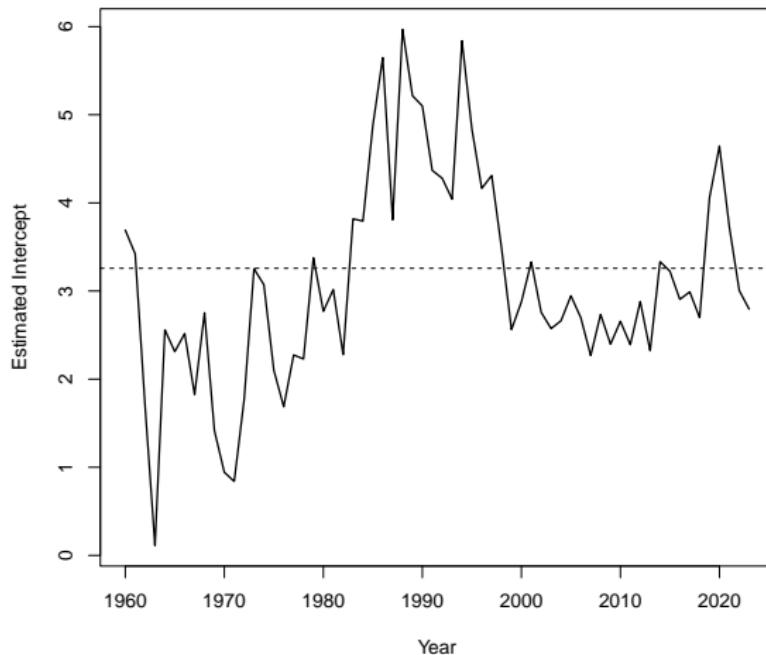
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 1.25 on 131 degrees of freedom
(82 observations deleted due to missingness)

Multiple R-squared: 0.095, Adjusted R-squared: 0.0881
F-statistic: 13.7 on 1 and 131 DF, p-value: 0.000308

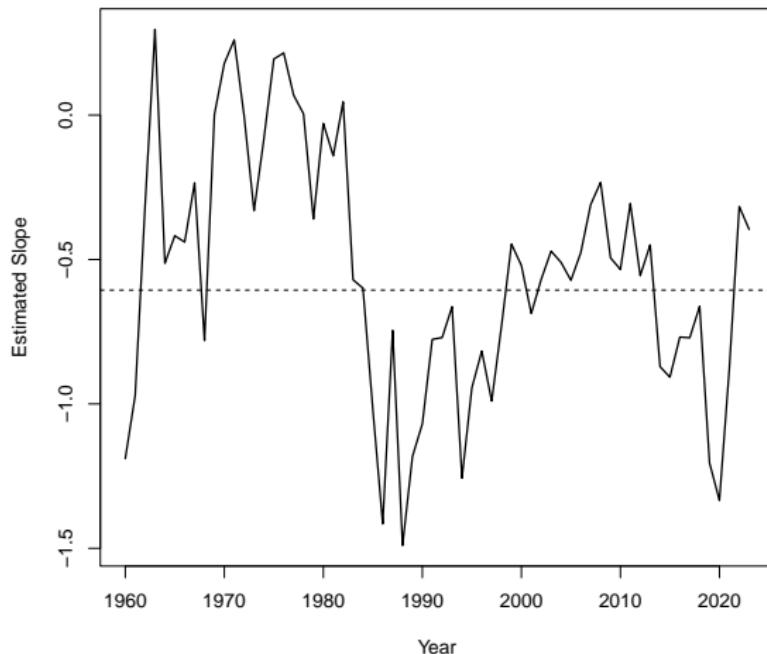
Intercepts For Every Year, 1960-2024

Estimated intercepts:



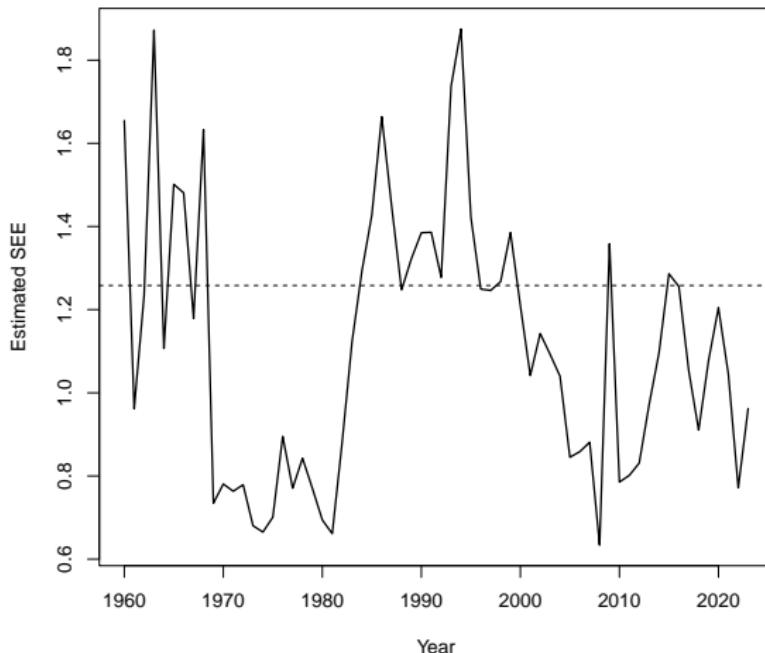
Slopes For Every Year, 1960-2023

Estimated slopes:

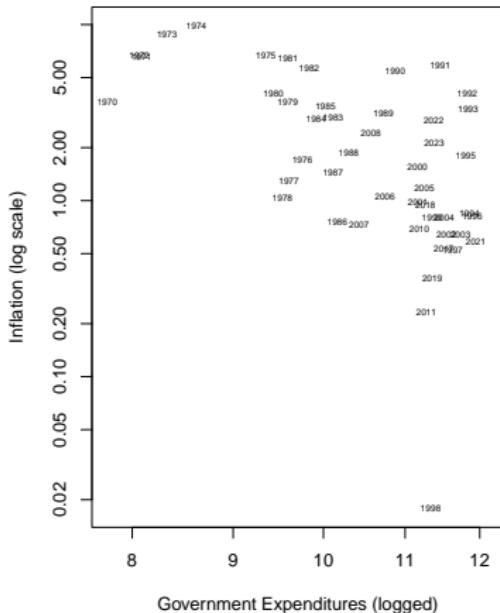


SEEs ($\hat{\sigma}$) For Every Year, 1960-2023

Estimated SEEs:

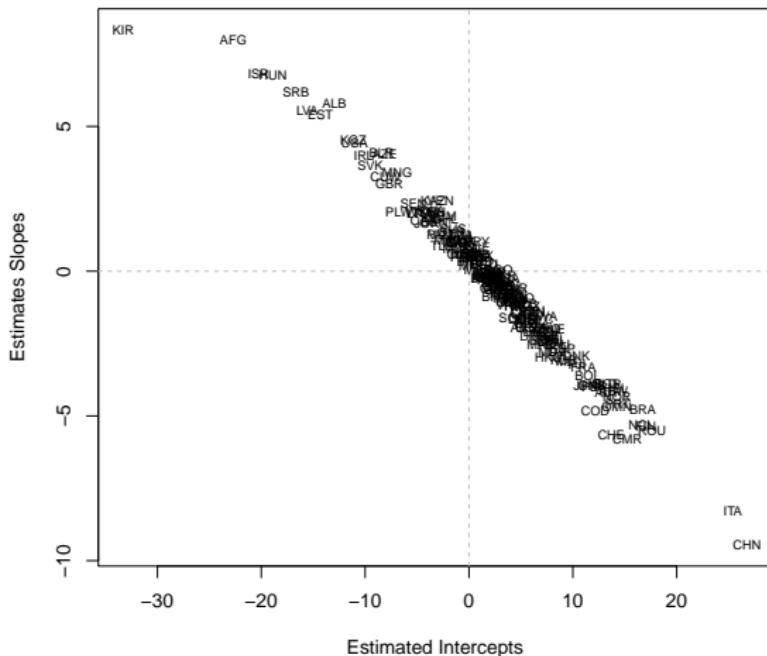


Time Series Regression: Switzerland



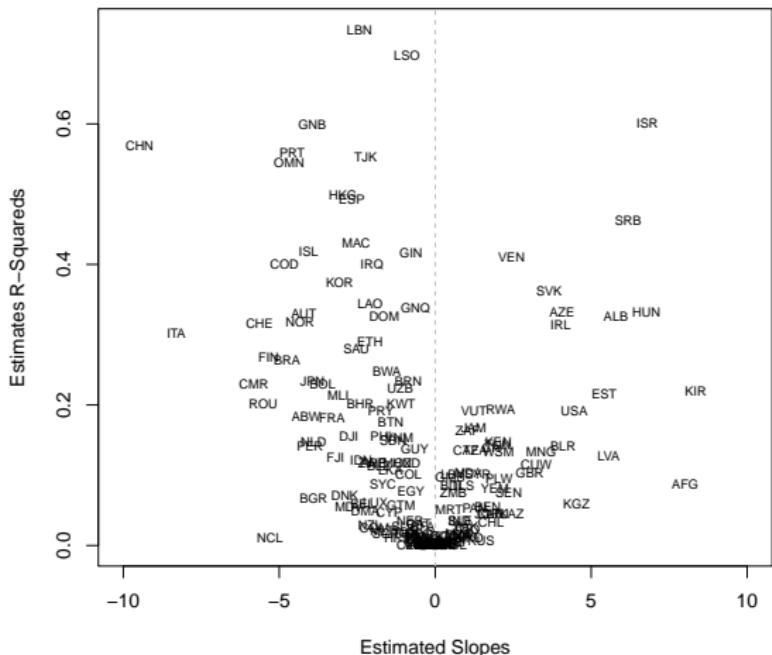
Country-Level Regressions: Intercepts & Slopes

Estimated intercepts and slopes:



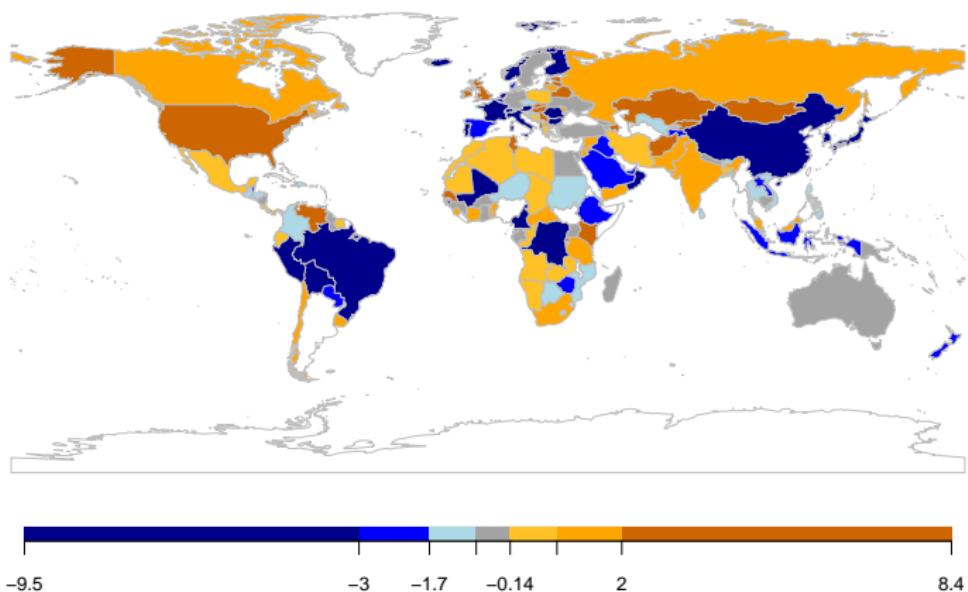
Country-Level Regressions: Slopes & R^2 s

Estimated slopes and R^2 s:



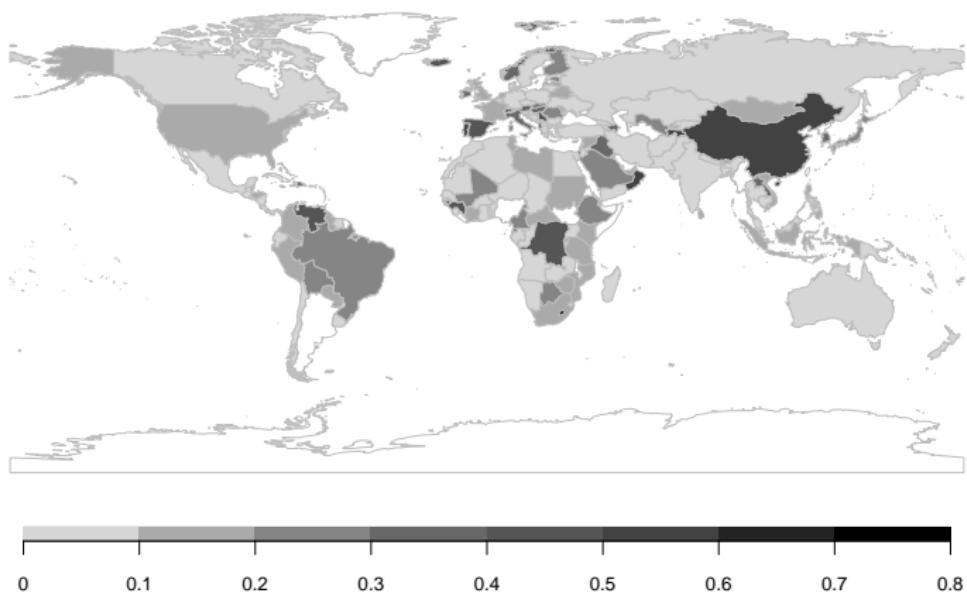
Maps! Country-Level Slopes ($\hat{\beta}_1$ s)

Estimated Slopes



Maps! Country-Level R^2 s

Estimated R-Squareds



Visualization: ExPanDaR

An interactive tool for exploring panel data...

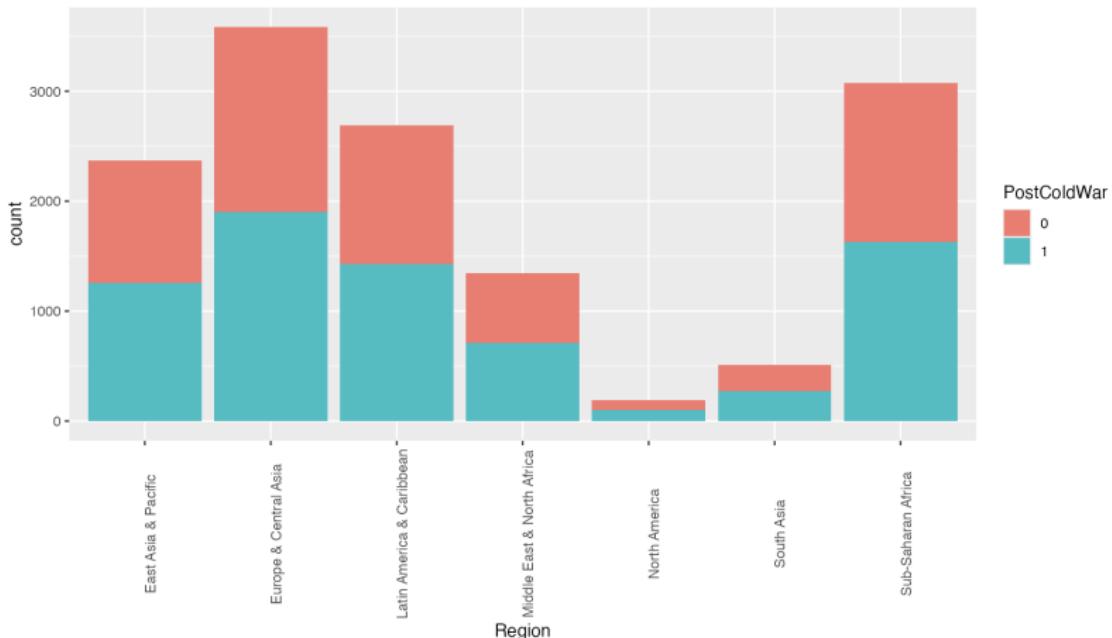
- Creator: Joachim Gassen (Department of Accounting,
Humboldt-Universität zu Berlin)
- Built upon / consistent with `ggplot` / `tidyverse`
- Requires installing the ExPanDaR package
- Calling
 - > `ExPanD()`

...opens the Shiny app, and asks for a (pre-formatted) data frame
(typically in CSV format)
- More information is here: <https://github.com/trr266/ExPanDaR>

Some examples...

ExPanDaR: Summaries

Counts by factors:



ExPanDaR: More Summaries

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](#)

Click here to delete all user defined variables and to restore the original variable set of the analysis sample.

[Restore Sample](#)

	Analysis Set	Base Set	N	Mean	Std. dev.	Min.	25 %	Median	75 %	Max.
V1	13,760	6,880.500	3,972.314	1.000	3,440.750	6,880.500	10,320.250	13,760.000		
LandArea	11,941	605,302.933	1,639,812.915	2.027	10,230.000	107,160.000	472,710.000	16,389,950.000		
ArablePercent	11,542	13.345	13.606	0.043	2.889	9.303	19.002		73.389	
Population	13,515	25,109,276.976	104,750,539.243	2,646.000	504,576.000	4,294,396.000	13,872,632.500	1,417,173,173.000		
PopGrowth	13,298	1.754	1.783	-27.722	0.690	1.702	2.651		20.473	
RuralPopulation	13,482	48,283	25,740	0.000	27,394	49,083	69,380		97,923	
UrbanPopulation	13,482	51.717	25,740	2.077	30,620	50,917	72,606		100,000	
BirthRatePer1K	12,937	28.018	13.084	5.000	16,100	26,604	39,778		58,121	
FertilityRate	12,779	3.906	2.000	0.772	2.080	3.418	5.791		8.864	
PrimarySchoolAge	10,896	6.196	0.614	4.000	6.000	6.000	7.000		8.000	
LifeExpectancy	12,766	64.827	11.291	11.995	57.540	67.467	72,941		85.498	
AgeDepRatioOld	13,515	10.698	7.037	0.200	5.905	7.826	14.082		70.360	
CO2Emissions	5,920	4,241	5.446	0.000	0.579	2.284	6.184		47.657	
GDP	10,099	250,284,546,944.349	1,140,901,242,824.164	21,562,114.297	3,698,198,246.005	15,734,874,525.098	102,516,943,295.367	20,926,835,051,000.000		
Inflation	8,547	23.496	327.098	-17.640	2.134	4.858	10.182		23,773.132	
TotalTrade	8,622	78,378	53.993	0.021	44.566	67.412	97.594		863.195	
Exports	8,622	36.511	28.891	0.005	18.169	29.339	46.489		433.836	
Imports	8,631	41.877	27.655	0.016	24.106	35.224	52.702		429.359	

ExPanDaR: Distributions

Histograms:

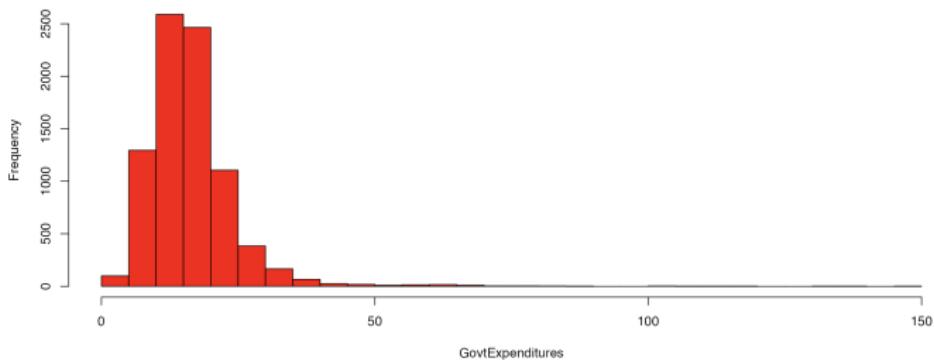
Histogram

Select variable to display

GovtExpenditures ▾

Suggested number of cells

5 50 250



Outlier Detection:

Extreme Observations	ISO3	Year	Inflation
<input type="button" value="Select variable to sort data by"/>	COD	1994	23,773.1
<input type="button" value="Inflation"/>	BOL	1985	11,749.6
<input type="button" value="Select period to subset to"/>	PER	1990	7,481.7
<input type="button" value="All"/>	UKR	1993	4,734.9
	AGO	1996	4,145.1

	TCD	1986	-13.1
	GNQ	1987	-13.2
	IRQ	1996	-16.1
	LSO	2009	-16.9
	GNQ	1986	-17.6

Bar Charts of Means (by factors):

By Group Bar Chart

Select variable to display

Inflation

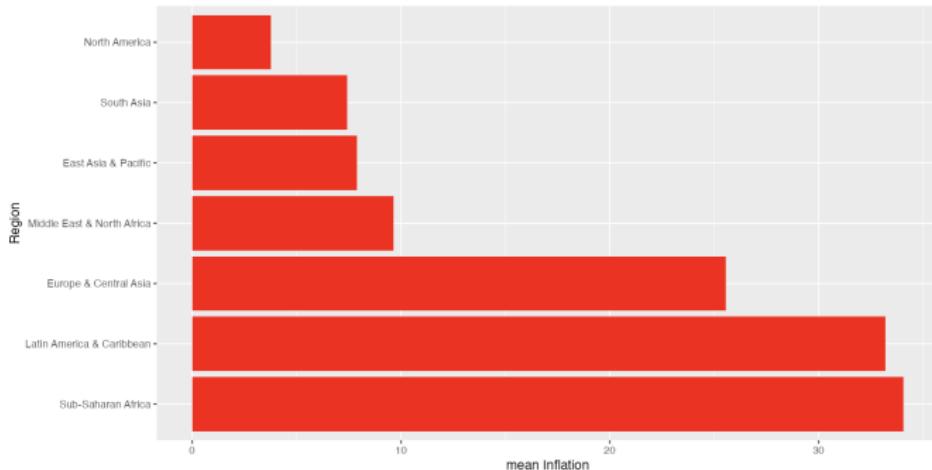
Select variable to group by

Region

Select statistic to display

Mean

Sort by statistic



ExPanDaR: More Plots

Violin Plots:

By Group Violin Chart

Select variable to display

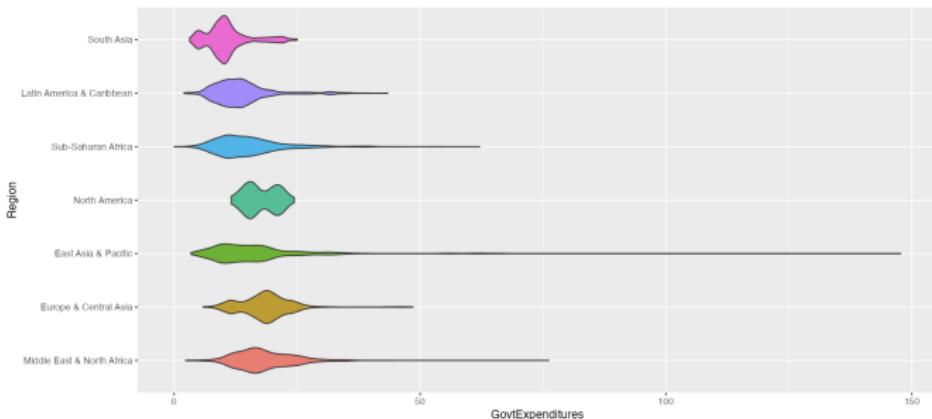
GovtExpenditures▼

Select variable to group by

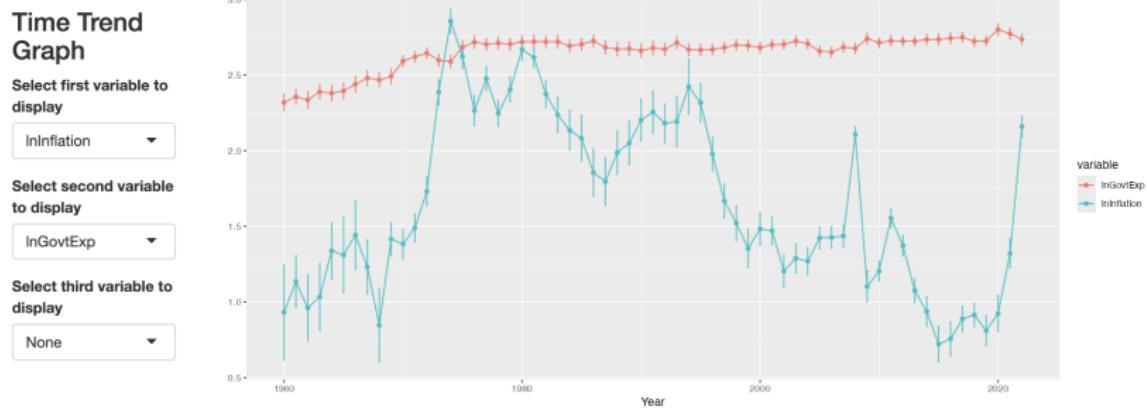
Region ▾

Sort by group means

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



General Trends + Variation:



ExPanDaR: More Trends

Trends in Quantiles:

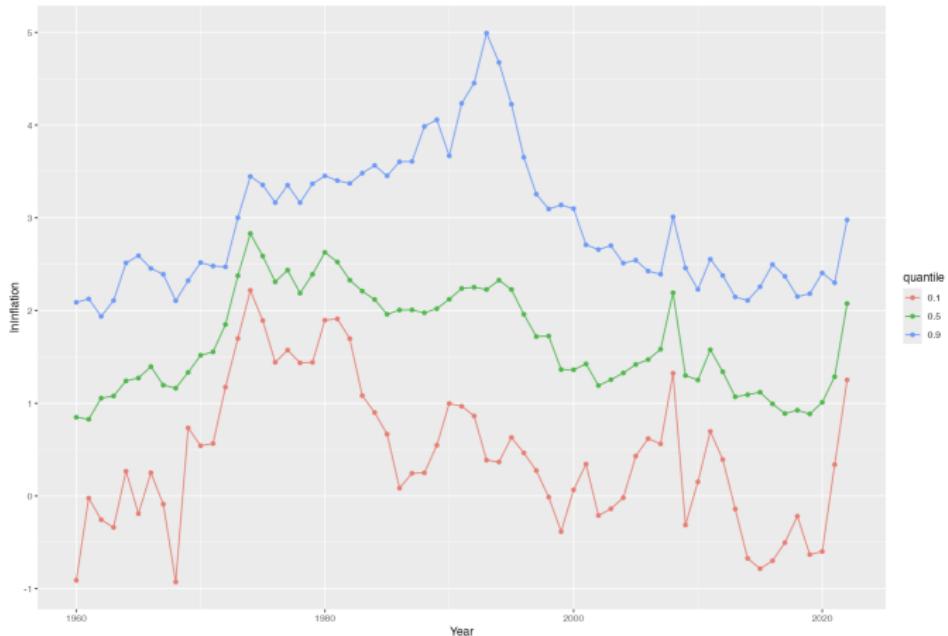
Quantile Time Trend Graph

Select variable to display

Inflation

Quantiles to show:

- Min
- 1 %
- 5 %
- 10 %
- 25 %
- 50 %
- 75 %
- 90 %
- 95 %
- 99 %
- Max



ExPanDaR: Even More Trends

Trends By Group:

By Group Time Trend Graph

Select variable to display

GovtExpenditures▼

Select variable to group by

Region ▼



ExPanDaR: Scatterplots

Fancy Scatterplots:

Scatter Plot

Select the x variable to display

InGovtExp

Select the y variable to display

InInflation

Select the variable to be reflected by dot size

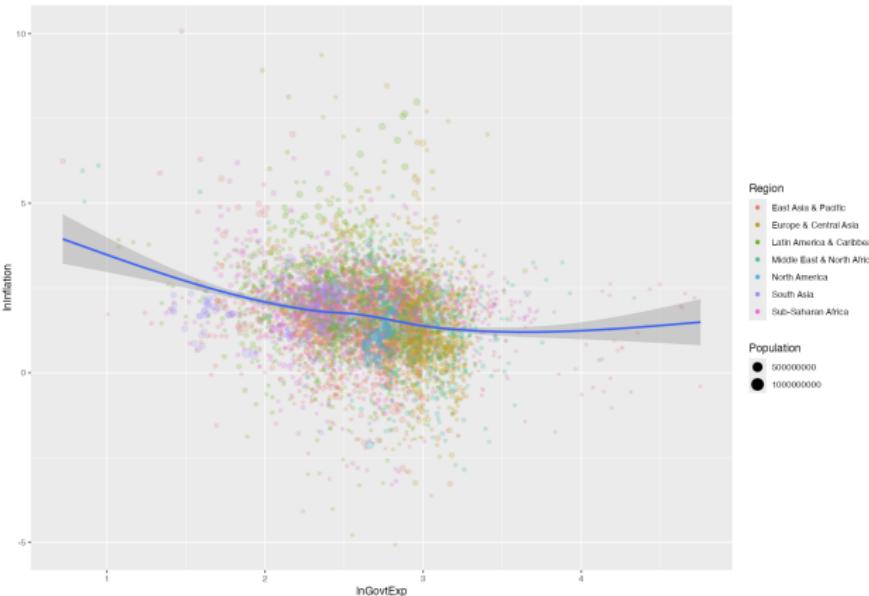
Population

Select the variable to be reflected by color

Region

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

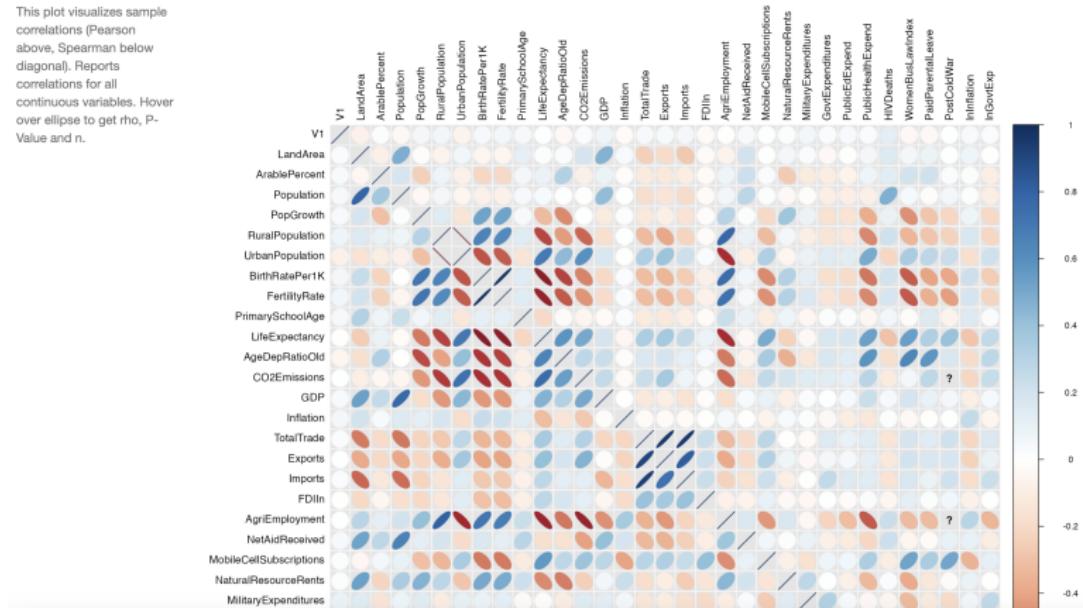
Display smoother



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

InInflation ▾

Select independent variable(s)

InGovtExp

PopGrowth

PostColdWar

TotalTrade

FDIIn

Select a categorical variable as the first fixed effect

None ▾

<i>Dependent variable:</i>	
	InInflation
InGovtExp	-0.647*** (0.040)
PopGrowth	-0.001 (0.010)
PostColdWar	-0.652*** (0.035)
TotalTrade	-0.004*** (0.0003)
FDIIn	-0.002** (0.001)
Constant	4.160*** (0.115)
Estimator	ols
Fixed effects	None
Std. errors clustered	No
Observations	5,901
R ²	0.143
Adjusted R ²	0.142

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

The Course Plan

- Tuesday, June 17: One- and Two-Way “Unit Effects” Models (fixed, “random,” etc.)
- Wednesday, June 18: Dynamics in Panel Data
- Thursday, June 19: Panel Data and Causal Inference
- Friday, June 20: Models for Discrete Responses (+ Exam)