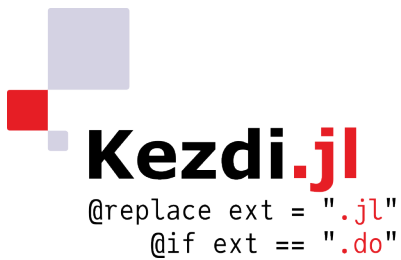


Kezdi.jl: Bridging Stata and Julia for economists

Miklós Koren

JuliaCon 2024



Who am I?

Economist + wannabe software developer

Language	Since
Stata	1997
Python	2003
Julia	2015

Data Editor at the *Review of Economic Studies*

What do economists do?

Evidence from 357 replication packages

They don't use Julia

They use multiple languages

What is Stata?

Best of Stata

```
use "trade.dta"  
replace distance = 5 if distance < 5  
generate log_trade = log(trade)  
generate log_distance = log(distance)  
regress log_trade log_distance, robust
```

Vs

```
df[df.distance < 5].distance = 5  
df.log_distance = log(df.distance)
```


What commands do economists use?

Tradeoffs in user interface

k standards $\rightarrow k + 1$ standards

I don't think if you at all

What do users want?

- 1 convenience
- 2 correctness

The Production Possibilities Frontier

- Rust: static typing, memory safe
- ChatGPT: convenient, may or may not be correct

Good example: row-level `if`

```
replace distance = 5 if distance < 5
```

Bad example: implicit variable abbreviation:

```
summarize gdp_per_capita
```

```
regress gdp population
```

Features of Kezdi.jl

Command syntax is \approx exactly like in Stata

```
@use "trade.dta"  
@replace distance = 5 @if distance < 5  
@generate log_trade = log(trade)  
@generate log_distance = log(distance)  
@regress log_trade log_distance, robust
```

Notes

- 1 Commands are macros
- 2 Variable names refer to column names in the *default* DataFrame
- 3 Function calls are vectorized automatically
- 4 Options are given with , option

Every command can operate on a subset of rows

```
@keep @if !ismissing(distance)
@replace distance = 5 @if distance < 5
@regress log_trade log_distance @if exporter_country != importer_country, 1
```

Notes

Handling missing values

Given a DataFrame

Row	x
1	1
2	2
3	missing
4	4

can you guess the output of

```
@collapse mean_x = mean(x)
@keep @if x < 3
```

Proper data structures

User-defined functions

Roadmap

Data wrangling

Programming convenience

Statistics

Acknowledgements