WELCOME TO THE 2023 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT

NDACAN

Children's Bureau

An Office of the Administration for Children & Families

NDACAN SUMMER TRAINING SERIES SCHEDULE 2023

July 5 — Introduction to NDACAN and the Administrative Data Series

July 12 — New Data Acquisition: CCOULD Data

July 19 — Causal Inference Using Administrative Data

July 26 — Evaluating and Dealing with Missing Data in R

August 2 — Time Series Analysis in Stata

August 9 — Data Visualization in R

SESSION AGENDA

- Overview of time series analysis
- Examples of time series analysis using NDACAN data
- Demonstration of time series analysis in Stata

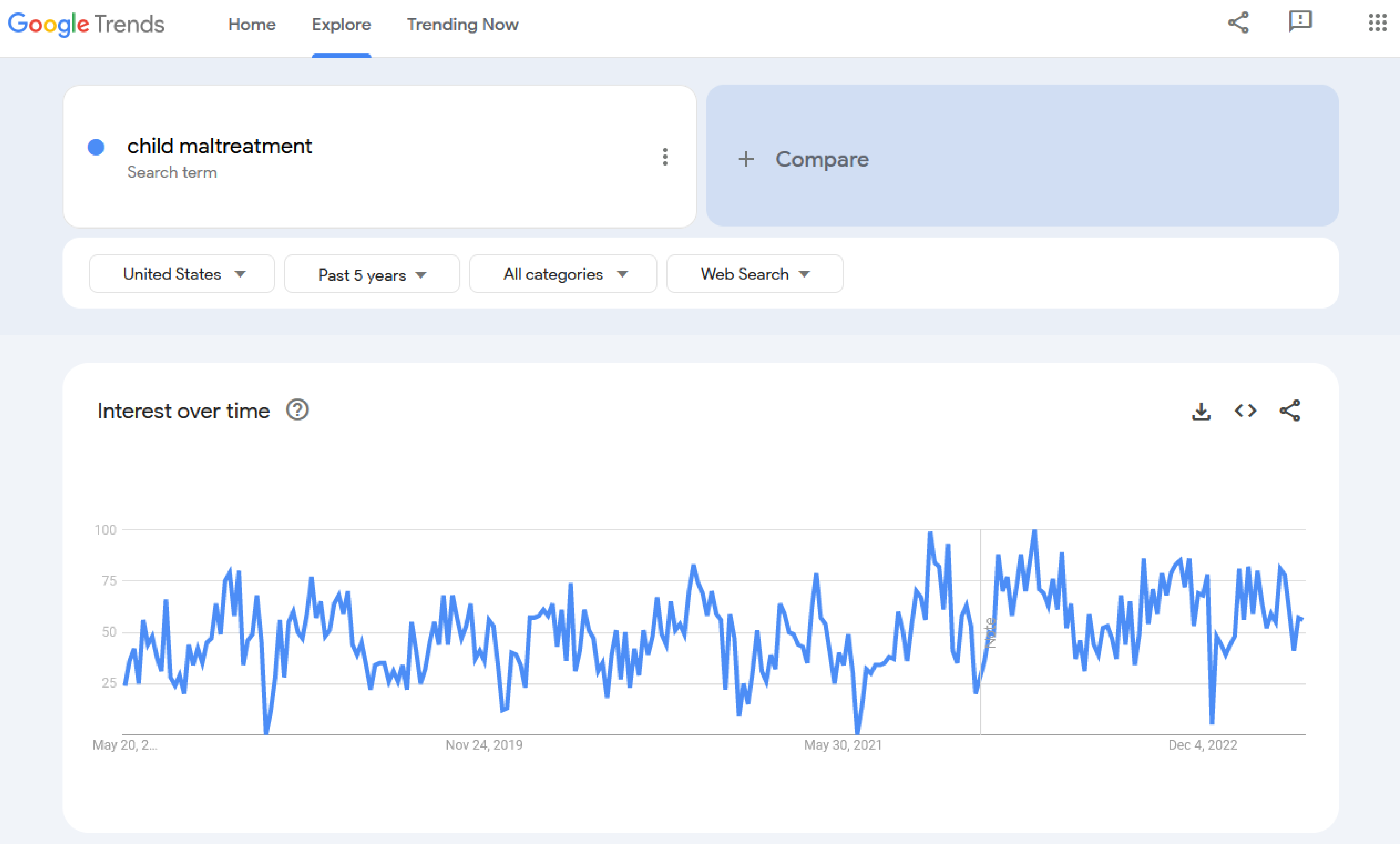
OVERVIEW OF TIME SERIES ANALYSIS

WHAT IS TIME SERIES ANALYSIS?

- **Time series data** are a series of data points indexed in **time order** (i.e. **sequenced**)
- **Time series analysis** comprises methods for extracting **statistics** and other meaningful information from time-ordered data
- **Time series forecasting** entails the use of a statistical **model** to **predict** future (unobserved) data points based on patterns of past (observed) data
- **Regression analysis** tests the **relationship** between **multiple** time series

WHY SHOULD I USE TIME SERIES ANALYSIS?

- Trends in your variable of interest are **serially correlated**
- You would like to **visualize** noisy trends in your variable of interest
- You are interested in **forecasting** future values of your variable of interest



child maltreatment

Search term

United States ▾ Past 5 years ▾ All categories ▾ Web Search ▾

Interest over time

UNIVARIATE AUTOREGRESSION

$$\boldsymbol{abuse_t = \alpha_0 + \alpha_1 abuse_{t-1} + \cdots + \alpha_k abuse_{t-k} + \varepsilon_t}$$

VECTOR AUTOREGRESSION (VAR)

$$\begin{bmatrix} abuse_t \\ neglect_t \end{bmatrix} = a_0 + A_1 \begin{bmatrix} abuse_{t-1} \\ neglect_{t-1} \end{bmatrix} + \cdots + A_k \begin{bmatrix} abuse_{t-k} \\ neglect_{t-k} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

WHAT DO I NEED FOR TIME SERIES ANALYSIS?

- A relatively **large sample** of sequenced observations
- Observations that are measured at **regular intervals**
- Dedicated **methods**

EXAMPLES OF TIME SERIES ANALYSIS USING NDACAN DATA

WHAT KINDS OF QUESTIONS CAN TIME SERIES ANALYSIS ANSWER? (NCANDS)

- *How seasonal are screened-in reports of maltreatment?*
  - Filter state/county trends for cyclicality
- *How does the rate of confirmed maltreatment in a time interval depend on the rate in the previous interval?*
  - Use autoregression model (ARIMA, ARFIMA) for state/county trends
- *What do we expect future rates of confirmed neglect to be?*
  - Build forecast model of future state/county trends
- *How are trends in confirmed physical abuse and confirmed neglect related?*
  - Use vector autoregression model (VAR) for multiple trends

DEMONSTRATION OF TIME SERIES ANALYSIS IN STATA

Link to Stata do file:
https://drive.google.com/file/d/1kzZl6JmID_gEv8zzdConrgCQJ3X9JRou/view?usp=sharing

Link to Presentation Slides:
https://docs.google.com/presentation/d/1b9MPcKcD7_Unfo0IYYIH-rhUoiIP5oSi/edit?usp=sharing&ouid=114322564655947637684&rtpof=true&sd=true

HELPFUL RESOURCES FOR TIME SERIES ANALYSIS IN STATA

- Stata reference manual on time series
  - https://www.stata.com/manuals13/tstimeseries.pdf
- Dr. Torres-Reyna's slides on time series analysis in Stata
  - https://www.princeton.edu/~otorres/TS101.pdf
- Becketti's *Introduction to Time Series Using Stata, Revised Edition*
  - https://www.stata-press.com/books/introduction-to-time-series-using-stata/

DEMONSTRATION IN STATA

The program, written in Stata, is included in the downloadable files for the slides and the transcript.

STATA DO FILE CODE, PAGE 1 OF 12

```
*************************************
*************************************
* NDACAN SUMMER TRAINING SERIES
* AUGUST 2, 2023
* TIMES SERIES (TS) ANALYSIS IN STATA
*************************************
*************************************

*********
* LINKS *
*********

* Stata .do file:
* Powerpoint slides:

*********************
* HELPFUL RESOURCES *
*********************

* Stata tutorial series (beginner):
https://www.youtube.com/playlist?list=PLN5IskQdgXWIEVJe6t9urIMoJVHdifFuR
* Stata reference manual: https://www.stata.com/manuals/ts.pdf
* Juan D'Amico's tutorial series (intermediate):
https://youtube.com/playlist?list=PLsZ8kVwX52ZEFZsVViYs60If7idJuKKUO
```

STATA DO FILE CODE, PAGE 2 OF 12

```
*********
* SETUP *
*********

* Let's set up our workspace.
clear // clear any data in memory
set more off // avoid having to click 'more' all the time
set seed 1013 // always set a seed for any random processes
cd "C:\Users\aroehrkasse\Box\Presentations\-NDACAN\2023_summer_series" // set your working directory

****************************************
* SET UP DATA FOR TIME SERIES ANALYSIS *
****************************************

* Let's read in some example data, specifically,
* an anonymized 1% sample of several variables
* from the 2017 NCANDS Child File.
use "data\ts_example.dta", clear // read dta file

* Let's examine the first observation.
list in f
list in f, nol

* Now let's clean our data.
* First let's create a state FIPS code variable.
gen stfips = round(rptfips/1000,1)
list in f, nol

* Next, for TS analysis,
* let's reformat the report date variable
* into a monthly format that Stata recognizes as such.
```

STATA DO FILE CODE, PAGE 3 OF 12

```
* First tell Stata that our report date variable is a date.
gen date = date(rptdt, "YMD")
format date %td
list in f, nol

* Then convert this date into a year-month variable.
gen datem = mofd(date)
format datem %tm
list in f, nol

* Finally, let's create a binary variable that is
* 1 if there is confirmed abuse, and
* 0 if there is confirmed neglect but not confirmed abuse.
gen abuse = 0
replace abuse = 1 if           chmal1 == 1 & mal1lev <= 2 | /// // physical abuse
                                                          chmal1 == 4 & mal1lev <= 2 | /// // sexual abuse
                                                          chmal1 == 5 & mal1lev <= 2 | /// // psych/emo
maltreatment
                                                          chmal1 == 7 & mal1lev <= 2 | /// // sex traficking
                                                          chmal1 == 8 & mal1lev <= 2 | /// // other
                                                          chmal2 == 1 & mal2lev <= 2 | ///
                                                          chmal2 == 4 & mal2lev <= 2 | ///
                                                          chmal2 == 5 & mal2lev <= 2 | ///
                                                          chmal2 == 7 & mal2lev <= 2 | ///
                                                          chmal2 == 8 & mal2lev <= 2 | ///
                                                          chmal3 == 1 & mal3lev <= 2 | ///
                                                          chmal3 == 4 & mal3lev <= 2 | ///
                                                          chmal3 == 5 & mal3lev <= 2 | ///
                                                          chmal3 == 7 & mal3lev <= 2 | ///
                                                          chmal3 == 8 & mal3lev <= 2 | ///
                                                          chmal4 == 1 & mal4lev <= 2 | ///
                                                          chmal4 == 4 & mal4lev <= 2 | ///
                                                          chmal4 == 5 & mal4lev <= 2 | ///
                                                          chmal4 == 7 & mal4lev <= 2 | ///
                                                          chmal4 == 8 & mal4lev <= 2
```

STATA DO FILE CODE, PAGE 4 OF 12

```
gen neglect = 0
replace neglect = 1 if     chmal1 == 2 & mal1lev <= 2 | /// // neglect
                                                              chmal1 == 3 & mal1lev <=
2 | /// // medical neglect
                                                              chmal2 == 2 & mal2lev <=
2 | ///
                                                              chmal2 == 3 & mal2lev <=
2 | ///
                                                              chmal3 == 2 & mal3lev <=
2 | ///
                                                              chmal3 == 3 & mal3lev <=
2 | ///
                                                              chmal4 == 2 & mal4lev <=
2 | ///
                                                              chmal4 == 3 & mal4lev <=
2

keep if abuse == 1 | neglect == 1

* Let's keep only the variables we need.
* Note that after the previous command, if abuse = 0, neglect = 1.
keep abuse datem stfips
list in f/10, nol
```

STATA DO FILE CODE, PAGE 5 OF 12

```
* And finally let's collapse our data into counts of reports by month.
* Note that half-months will be combined.
gen n = 1
collapse (count) n, by(abuse stfips datem)
order stfips abuse datem n
sort stfips abuse datem
list in f/10

* Now let's read in pre-processed count data for FY 2012-2021.
* Note that small counts are arbitrarily inflated to prevent disclosure risk.
use "data\ts.dta", clear // read dta file

* Let's merge it to a utility file that contains
* state FIPS codes and state abbreviations (ab).
merge m:1 stfips using "data\statecodes.dta"
drop if _merge < 3
drop _merge
list in f/3

* And let's label our state FIPS variable and its values.
* (this requires installation of labutil package).
* ssc install labutil // uncomment this to install
label var stfips "State"
labmask stfips, values(ab)

* And now let's tell Stata that our data are time-series data so that we can run
* specialized TS commands. Note that the optional first term is our panel variable,
* and the required second term is our time variable.
tsset stfips datem, m
```

STATA DO FILE CODE, PAGE 6 OF 12

```
* Oops! Because our data are long (i.e. "n" counts both abuse and neglect),
* our panel data isn't identified. So let's reshape.
reshape wide n, i(stfips ab datem) j(abuse)
rename n0 neglect
rename n1 abuse

* And try TS setting our data again.
tsset stfips datem, m

***********************
* VISUALIZING TS DATA *
***********************

* Let's say we want to visualize some trends in our data, but they're noisy.
* Let's first visualize raw data on abuse across a few states.
* If we want to visualize the same time series across multiple panels,
* it can actually be easier to use Stata's xt commands,
* for panel data. These mostly work with tsset data, but you may have to xtset.
xtline abuse if stfips < 9, ///
xlabel(, angle(vertical)) ylabel(, angle(horizontal)) xtitle("Time") ytitle("Confirmed
abuse reports")
```

STATA DO FILE CODE, PAGE 7 OF 12

```
* Note that counts seem very low in early/late months. This is because many reports
* are lagged in their submission to NDACAN relative to the report date.
* For this reason, it is EXTREMELEY important to censor your data appropriately.
* My rule of thumb is you can only analyze one fewer fiscal year than submission year.
* We're using the 2012-2021 Child Files (submission year),
* so we'll censor to FY2012-2020 (fiscal year).
drop if datem < tm(2011m9) | datem > tm(2020m8)
xtline abuse if stfips < 9, ///
xlabel(, angle(vertical)) ylabel(, angle(horizontal)) xtitle("Time") ytitle("Confirmed abuse reports")

* Our data look kinda noisy. What if we want to plot a smoother line?
* We can do this using Stata's moving-average capability.
tssmooth ma abuse_ma1 = abuse, window(1 1 1)
twoway (tsline abuse abuse_ma*) if stfips == 6, ///
ylabel(, angle(horizontal)) xtitle("Time") ytitle("Confirmed abuse reports") legend(order(1 "Raw data" 2 “3-mo. moving avg."))

* Or we can compute a weighted moving average, where nearer observations count more.
tssmooth ma abuse_ma2 = abuse, weights(1/6 <7> 6/2)
twoway (tsline abuse abuse_ma*) if stfips == 6, ///
ylabel(, angle(horizontal)) xtitle("Time") ytitle("Confirmed abuse reports") legend(order(1 "Raw data" 2 "3-mo. moving avg." 3 "12-mo. weighted moving avg."))
```

STATA DO FILE CODE, PAGE 8 OF 12

```
**************************
* TIME-SERIES OPERATIONS *
* ************************


* Stata also makes it very easy to calculate common time-series quantities of interest.

* Let's say we want to know the one-month lead of a variable,
* Stata has a specific syntax for this.
list stfips datem abuse F1.abuse in f/10

* We can do the same for lags.
list stfips datem neglect L2.neglect in f/10

* Let's say we want to calculate the difference in values
* across time periods (in our case, months).
* We again use Stata's special TS syntax.
list stfips datem abuse D1.abuse in f/10

* Note that d2 is NOT a two-period difference, but rather
* a second-order difference.
list stfips datem abuse D1.abuse D2.abuse in f/20

* Let's say we want to know the 12-month change,
* i.e. the seasonal difference. Here we use different syntax.
list stfips datem abuse S12.abuse in f/20
```

STATA DO FILE CODE, PAGE 9 OF 12

```
* Let's visualize this seasonal difference,
* or year-over-year monthly change.
gen abuse_s12 = S12.abuse
xtline abuse_s12 if stfips < 9, ///
xlabel(, angle(vertical)) ylabel(, angle(horizontal)) xtitle("Time") ytitle("12mo change in confirmed abuse reports")


******************************
* UNIVARIATE AUTOREGRESSION *
* ****************************

* Let's say we want a statistical model that captures the properties of our maltreatment trends.

* To keep things simple, let's just focus on CA from here on out.
keep if stfips == 6
tsset datem, m

* Time series models generally require that the variable of interest is stationary,
* basically meaning that it is independent of time.

* Are abuse trends in CA stationary? Simply examining, it appears not.
tsline abuse, ///
xlabel(, angle(vertical)) ylabel(, angle(horizontal)) xtitle("Time") ytitle("Confirmed abuse reports")

* However, formal tests reject the null hypothesis that the abuse trend
* has a unit root (i.e. is not stationary). That double negative is tricky:
* in other words, they seem to indicate that the process is stationary.
dfuller abuse, trend regress
pperron abuse, trend regress
```

STATA DO FILE CODE, PAGE 10 OF 12

```
* If the process isn't stationary (which it usually isn't),
* we can often model the first difference, which usually is.
*This difference is also of policy interest: will abuse go up or down this month?
* For illustration, let's model this difference.
tsline D1.abuse, ///
xlabel(, angle(vertical)) ylabel(, angle(horizontal)) xtitle("Time") ytitle("1mo change in confirmed abuse reports")

*The most popular time-series model is the
* autoregressive integrated moving average (ARIMA) model.
*This model combines analysis of autoregressive and moving-average processes.
* Parametric ARIMA models require us to specify how we want to model these processes.
* How should we choose these parameters? It's more of an art than a science,
* though new versions of Stata include model selection features (arimasoc).

* First, moving-average processes are fundamentally about autocorrelation.
*What does the autocorrelation of our first difference look like?
*We can use a correlogram to see.
ac D1.abuse
*The fact that the first lag is outside the confidence interval
* tells us that 1 is a good starting point for our moving-average parameter ("q").

* Second, autoregressive processes are fundamentally about partial autocorrelation.
pac D1.abuse
*The fact that the first four lags are outside the confidence interval
* tells us that 4 is a good starting point for our autoregressive parameter ("p").
```

STATA DO FILE CODE, PAGE 11 OF 12

```
* Our final parameter in the ARIMA model is the integrated (difference) order ("d"), which will be 1.

* Let's fit our model using the (p,d,q) syntax!
arima abuse, arima(4,1,1)
* Note that the above could also be written as the following:
* arima D1.abuse, ar(4) ma(1)

* Recall from our correlogram that we had a noticable 12-month lagged autocorrelation.
* This is seasonality! We can adjust for this using a helpful option in Stata.
arima abuse, arima(4,1,1) sarima(1,1,1,12)

***************
* FORECASTING *
***************

* So what!? Well, learning about time-series processes can help us predict the future,
* based solely on the pattern of trends in the outcome variable.

        * To forecast, we would first want to do some diagnostics (beyond today's scope).
        predict error, resid
        summarize error
        tsline error, yline(-22.08081) // Are residuals tightly grouped around the mean (good)?
        wntestq error // Do we fail to reject the null hypothesis that our process is white noise (good)?
        estat aroots // Are the roots inside the circle (good)?
        * If we meet these criteria, we have a good candidate model for forecasting!
```

STATA DO FILE CODE, PAGE 12 OF 12

```
* Let's create some empty cells to forecast into.
tsappend, add(36)

* And predict values using our SARIMA model.
predict abuse_f, y dynamic(m(2020m9))

* Let's get confidence intervals for our forecasting
predict abuse_fv, mse dynamic(m(2020m9))
generate ub = abuse_f + 1.96*sqrt(abuse_fv)
generate lb = abuse_f - 1.96*sqrt(abuse_fv)

* And finally, plot our forecast against the real data.
twoway          (rarea ub lb datem if datem >= tm(2020m8), fcolor(blue%25)) ///
                                (tsline abuse) ///
                                (tsline abuse_f if datem >= tm(2020m8)), ///
                                xlabel(, angle(vertical)) ylabel(, angle(horizontal)) xtitle("Time") ///
                                ytitle("Confirmed abuse reports") legend(order(2 "Actual" 3 "Forecast" 1 "95% CI" ))

******************
* GOING FURTHER *
******************

* NDACAN data users further interested in time-series analysis will likely benefit from exploring:

                * 1.Vector autoregression models
                * 2. Panel data models
                * 3. State-space models
```

QUESTIONS?

ALEX ROEHRKASSE
ASSISTANT PROFESSOR,
BUTLER UNIVERSITY

AROEHRKASSE@BUTLER.EDU

NEXT WEEK…

August 9, 2023

Presenter:

Sarah Sernaker, NDACAN

Topic:

Data Visualization in R