Modeling of Temporal Data

David Carlson

April 18, 2022

What is Temporal Data and Modeling?

- Repeated measurements over time

What is Temporal Data and Modeling?

- Repeated measurements over time
- Usually one unit of observation, aggregated data, or no need to account for unit heterogeneity

What is Temporal Data and Modeling?

- Repeated measurements over time
- Usually one unit of observation, aggregated data, or no need to account for unit heterogeneity
- The reading goes over typical procedures, but most assume balanced / equally spaced observations

What is Temporal Data and Modeling?

- Repeated measurements over time
- Usually one unit of observation, aggregated data, or no need to account for unit heterogeneity
- The reading goes over typical procedures, but most assume balanced / equally spaced observations
- What are some basic issues that may arise when inferring from temporal data?

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change?

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change?

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? . . . Correlation

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? ... Correlation ... Teasing out causality is particularly difficult with TS
- How did a shock or treatment affect an outcome or a relationship?

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? ... Correlation ... Teasing out causality is particularly difficult with TS
- How did a shock or treatment affect an outcome or a relationship?
  - Regression discontinuity (analyze either side of some cut-off)

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? ... Correlation ... Teasing out causality is particularly difficult with TS
- How did a shock or treatment affect an outcome or a relationship?
  - Regression discontinuity (analyze either side of some cut-off)
  - Change point model (can be agnostic about shift timing or test a hypothesis about specific time)

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? ... Correlation ... Teasing out causality is particularly difficult with TS
- How did a shock or treatment affect an outcome or a relationship?
  - Regression discontinuity (analyze either side of some cut-off)
  - Change point model (can be agnostic about shift timing or test a hypothesis about specific time)
  - Synthetic control and extensions (more in TSCS)

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? . . . Correlation . . . Teasing out causality is particularly difficult with TS
- How did a shock or treatment affect an outcome or a relationship?
  - Regression discontinuity (analyze either side of some cut-off)
  - Change point model (can be agnostic about shift timing or test a hypothesis about specific time)
  - Synthetic control and extensions (more in TSCS)
- Forecasting / prediction

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? ... Correlation ... Teasing out causality is particularly difficult with TS
- How did a shock or treatment affect an outcome or a relationship?
  - Regression discontinuity (analyze either side of some cut-off)
  - Change point model (can be agnostic about shift timing or test a hypothesis about specific time)
  - Synthetic control and extensions (more in TSCS)
- Forecasting / prediction
- Analyzing trends

Goals of Temporal Modeling

- Could be same as before: As x changes, how does y change? ... Correlation ... Teasing out causality is particularly difficult with TS
- How did a shock or treatment affect an outcome or a relationship?
  - Regression discontinuity (analyze either side of some cut-off)
  - Change point model (can be agnostic about shift timing or test a hypothesis about specific time)
  - Synthetic control and extensions (more in TSCS)
- Forecasting / prediction
- Analyzing trends
- Survival, time until failure / success, growth rates

Pervasive Issues

- (Non-)Stationarity

Pervasive Issues

- (Non-)Stationarity
- Auto-correlation

Pervasive Issues

- (Non-)Stationarity
- Auto-correlation
- Covariation of vars without any causality

Pervasive Issues

- (Non-)Stationarity
- Auto-correlation
- Covariation of vars without any causality
- Confounders (especially *unobserved*) particularly problematic

Pervasive Issues

- (Non-)Stationarity
- Auto-correlation
- Covariation of vars without any causality
- Confounders (especially *unobserved*) particularly problematic
- Seasonality

Pervasive Issues

- (Non-)Stationarity
- Auto-correlation
- Covariation of vars without any causality
- Confounders (especially *unobserved*) particularly problematic
- Seasonality
- Non-ergodic (process that changes erratically at an inconsistent rate)

Pervasive Issues

- (Non-)Stationarity
- Auto-correlation
- Covariation of vars without any causality
- Confounders (especially *unobserved*) particularly problematic
- Seasonality
- Non-ergodic (process that changes erratically at an inconsistent rate)
- Unevenly spaced obs

Stationarity

- Most approaches require stationarity

Stationarity

- Most approaches require stationarity
- 3 criterion:

Stationarity

- Most approaches require stationarity
- 3 criterion:
  - Mean of series is constant (not a function of time)

Stationarity

- Most approaches require stationarity
- 3 criterion:
  - Mean of series is constant (not a function of time)
  - Variance of series is not function of time (homoskedasticity)

Stationarity

- Most approaches require stationarity
- 3 criterion:
  - Mean of series is constant (not a function of time)
  - Variance of series is not function of time (homoskedasticity)
  - Covariance of ith term and $(i + m)$th term not a function of time

Stationarity

- Most approaches require stationarity
- 3 criterion:
  - Mean of series is constant (not a function of time)
  - Variance of series is not function of time (homoskedasticity)
  - Covariance of ith term and $(i + m)$th term not a function of time
- In cases where the stationary criterion are violated, stationarize the time series and then try stochastic models to predict this time series

Stationarity

- Most approaches require stationarity
- 3 criterion:
  - Mean of series is constant (not a function of time)
  - Variance of series is not function of time (homoskedasticity)
  - Covariance of ith term and $(i + m)$th term not a function of time
- In cases where the stationary criterion are violated, stationarize the time series and then try stochastic models to predict this time series
  - Detrending, differencing, transformation, etc.

Random Walk

- An intuitively non-stationary process (but often thought to be stationary)

Random Walk

- An intuitively non-stationary process (but often thought to be stationary)
- Imagine a cat moving randomly on a chess board

Random Walk

- An intuitively non-stationary process (but often thought to be stationary)
- Imagine a cat moving randomly on a chess board
- Next position is only a function of the previous position plus noise (because it is a *random* walk)

Random Walk

- An intuitively non-stationary process (but often thought to be stationary)
- Imagine a cat moving randomly on a chess board
- Next position is only a function of the previous position plus noise (because it is a *random* walk)
- In words: You want to predict the position of the cat with time. How accurate will you be? Of course you will become more and more inaccurate as the position of the cat changes. At $t = 0$ you exactly know where the cat is. Next time, it can only move to 8 squares and hence your probability dips to $\frac{1}{8}$ instead of 1 and it keeps on going down.

Random Walk Formulation

$$X(t) = X(t-1) + e(t) \tag{1}$$

$$X(t) = X(0) + \sum_{l=1}^{t} e(l) \tag{2}$$

$$E[X(t)] = E[X(0)] + \sum_{l=1}^{t} E[e(l)] \tag{3}$$

$$= E[X(0)] \rightarrow \text{constant} \tag{4}$$

$$Var[X(t)] = Var[X(0)] + \sum_{l=1}^{t} Var[e(l)] \tag{5}$$

$$= t \times Var(\mathbf{e}) \rightarrow \text{time dependent} \tag{6}$$

Make it Stationary with ρ

$$X(t) = \rho X(t-1) + e(t) \tag{7}$$

$$E[X(t)] = \rho \times E[X(t-1)] \tag{8}$$

$$X(t) - X(t-1) = (\rho - 1)X(t-1) + e(t) \tag{9}$$

- What if $\rho = 1$? No force can pull the X down in the next step $\rightarrow$ non-stationary
- We test if $(\rho - 1)$ is significantly different than zero or not
- If the null hypothesis gets rejected, we have a stationary time series (Dickey Fuller test)

Stationarize the Series

- Detrending — remove the trend component from the time series (substract out of the mean function an estimated trend)

Stationarize the Series

- Detrending — remove the trend component from the time series (substract out of the mean function an estimated trend)
- Differencing — model the differences of the terms and not the actual term; AR(I)MA

Stationarize the Series

- Detrending — remove the trend component from the time series (substract out of the mean function an estimated trend)
- Differencing — model the differences of the terms and not the actual term; AR(I)MA
- Seasonality — can be incorporated into ARIMA, can use seasonal controls, can detrend at different temporal level, can model separately, ...

Stationarize the Series

- Detrending — remove the trend component from the time series (substract out of the mean function an estimated trend)
- Differencing — model the differences of the terms and not the actual term; AR(I)MA
- Seasonality — can be incorporated into ARIMA, can use seasonal controls, can detrend at different temporal level, can model separately, ...
- While very useful for prediction / forecasting, very difficult to make substantive claims (e.g., causality or unconfounded correlation)

Inference: Regression Discontinuity

- RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned

Inference: Regression Discontinuity

- RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned
- Commonly used for TS analyses (RDiT), but:

Inference: Regression Discontinuity

- RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned
- Commonly used for TS analyses (RDiT), but:
  - Many applications lack cross-sectional variation and are estimated using observations far from the temporal threshold. This common empirical practice is hard to square with the assumptions of a cross-sectional RD, which is conceptualized for an estimation bandwidth shrinking even as the sample size increases

Inference: Regression Discontinuity

- RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned
- Commonly used for TS analyses (RDiT), but:
  - Many applications lack cross-sectional variation and are estimated using observations far from the temporal threshold. This common empirical practice is hard to square with the assumptions of a cross-sectional RD, which is conceptualized for an estimation bandwidth shrinking even as the sample size increases
  - Estimates may be biased if the time-series properties of the data are ignored (for instance in the presence of an autoregressive process), or more generally if short-run and long-run effects differ

Inference: Regression Discontinuity

- RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned
- Commonly used for TS analyses (RDiT), but:
  - Many applications lack cross-sectional variation and are estimated using observations far from the temporal threshold. This common empirical practice is hard to square with the assumptions of a cross-sectional RD, which is conceptualized for an estimation bandwidth shrinking even as the sample size increases
  - Estimates may be biased if the time-series properties of the data are ignored (for instance in the presence of an autoregressive process), or more generally if short-run and long-run effects differ
  - Tests for sorting or bunching near the threshold are often irrelevant, making the framework closer to an event study than a regression discontinuity design

Inference: Regression Discontinuity

- RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned
- Commonly used for TS analyses (RDiT), but:
  - Many applications lack cross-sectional variation and are estimated using observations far from the temporal threshold. This common empirical practice is hard to square with the assumptions of a cross-sectional RD, which is conceptualized for an estimation bandwidth shrinking even as the sample size increases
  - Estimates may be biased if the time-series properties of the data are ignored (for instance in the presence of an autoregressive process), or more generally if short-run and long-run effects differ
  - Tests for sorting or bunching near the threshold are often irrelevant, making the framework closer to an event study than a regression discontinuity design
- More in TSCS, where it is more applicable

Change-Point Models

- Change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes

Change-Point Models

- Change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes
- Concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes

Change-Point Models

- Change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes
- Concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes
- Identify whether, and if so when, the behavior of the series changes significantly

Change-Point Models

- Change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes
- Concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes
- Identify whether, and if so when, the behavior of the series changes significantly
- We will focus on offline (post hoc) analyses for hypothesis testing

Change-Point Models

- Change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes
- Concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes
- Identify whether, and if so when, the behavior of the series changes significantly
- We will focus on offline (post hoc) analyses for hypothesis testing
- For hypothesized break-point $\rightarrow$ segmented regression (piecewise or broken-stick)

Change-Point Models

- Change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes
- Concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes
- Identify whether, and if so when, the behavior of the series changes significantly
- We will focus on offline (post hoc) analyses for hypothesis testing
- For hypothesized break-point $\rightarrow$ segmented regression (piecewise or broken-stick)
  - Can simply fit OLS on either side and test for differences in estimates

Change-Point Models

- Change detection or change point detection tries to identify times when the probability distribution of a stochastic process or time series changes
- Concerns both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes
- Identify whether, and if so when, the behavior of the series changes significantly
- We will focus on offline (post hoc) analyses for hypothesis testing
- For hypothesized break-point $\rightarrow$ segmented regression (piecewise or broken-stick)
  - Can simply fit OLS on either side and test for differences in estimates
  - Often better (though harder to convey) to use non-parametric approaches

Growth Models

- Many types of growth models

Growth Models

- Many types of growth models
- Generally fall into two categories: Population dynamics in demography, and economic growth

Growth Models

- Many types of growth models
- Generally fall into two categories: Population dynamics in demography, and economic growth
- Can apply to any question regarding dynamic observations

Growth Models

- Many types of growth models
- Generally fall into two categories: Population dynamics in demography, and economic growth
- Can apply to any question regarding dynamic observations
- Again, have to decide between parametric and non-parametric

Growth Models

- Many types of growth models
- Generally fall into two categories: Population dynamics in demography, and economic growth
- Can apply to any question regarding dynamic observations
- Again, have to decide between parametric and non-parametric
- May wish to look at *stages* of growth

Survival Models

- Analyzing the expected duration of time until one event occurs

Survival Models

- Analyzing the expected duration of time until one event occurs
- Reliability analysis, duration analysis, event history analysis

Survival Models

- Analyzing the expected duration of time until one event occurs
- Reliability analysis, duration analysis, event history analysis
- Death or failure is considered an event in the survival analysis literature — traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken

Survival Models

- Analyzing the expected duration of time until one event occurs
- Reliability analysis, duration analysis, event history analysis
- Death or failure is considered an event in the survival analysis literature — traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken
- Recurring event or repeated event models relax that assumption

Survival Analysis Uses

- To describe the survival times of members of a group

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups
  - Log-rank test

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups
  - Log-rank test
- To describe the effect of categorical or quantitative variables on survival

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups
  - Log-rank test
- To describe the effect of categorical or quantitative variables on survival
  - Cox proportional hazards regression

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups
  - Log-rank test
- To describe the effect of categorical or quantitative variables on survival
  - Cox proportional hazards regression
  - Parametric survival models

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups
  - Log-rank test
- To describe the effect of categorical or quantitative variables on survival
  - Cox proportional hazards regression
  - Parametric survival models
  - Survival trees

Survival Analysis Uses

- To describe the survival times of members of a group
  - Life tables
  - Kaplan-Meier curves
  - Survival function
  - Hazard function
- To compare the survival times of two or more groups
  - Log-rank test
- To describe the effect of categorical or quantitative variables on survival
  - Cox proportional hazards regression
  - Parametric survival models
  - Survival trees
  - Survival random forests