Data Mining and Business Intelligence

Lecture 10: Models with Regressors

Jing Peng
University of Connecticut

4/2/20

Agenda

- Feedback for live streaming
- Assignment 2 common mistakes
  - clustering results not useful to increase engagements
  - use rules with target=0
  - variable with negative coefficient
  - insights based on brand names
- Assignment 3 & Project (WebEx recording + remote control) & Exam
- ESM and ARIMA Recap
- Regressors and Events
- ARIMA Notations
- Transfer function
- Seasonal ARIMA models

Simple Exponential Smoothing Predictions

$\widehat{Y}_1 = Y_0$ (starting value)

$$\widehat{Y}_2 = \omega Y_1 + (1 - \omega)\widehat{Y}_1$$

...

$$\widehat{Y}_{t+1} = \omega Y_t + (1 - \omega)\widehat{Y}_t$$

- The starting value Y_0 is often taken to be the mean of the first *n* observations. SAS TSFS uses *n*=6.
- Does not work well when there is a trend

Double Exponential (Holt) Smoothing

$$L_t = \omega Y_t + (1-\omega)(L_{t-1} + T_{t-1}) \;\; 0 \le \omega \le 1$$

Level equation

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1} \;\; 0 \le \gamma \le 1$$

Trend equation

$\hat{Y}_{t+m} = L_t + mT_t$ (m-period-ahead forecast)

Prediction equation

- Smoothing for both level and trend
- TSFS uses Double Exponential Smoothing to refer to a simpler model with only one smoothing parameter (see slide 22)
- Damped-trend smoothing: a third weight on T_{t-1}

Fore more details
https://onlinecourses.science.psu.edu/stat501/node/363
http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

Winters Method — Additive

$$L_t = \omega(Y_t - S_{t-p}) + (1-\omega)(L_{t-1} + T_{t-1})$$

Level equation

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$$

Trend equation

$$S_t = \delta(Y_t - L_t) + (1-\delta)S_{t-p}$$

Seasonality equation

$Y_{t+m} = L_t + mT_t + S_t$ (m-period-ahead forecast)

Prediction equation

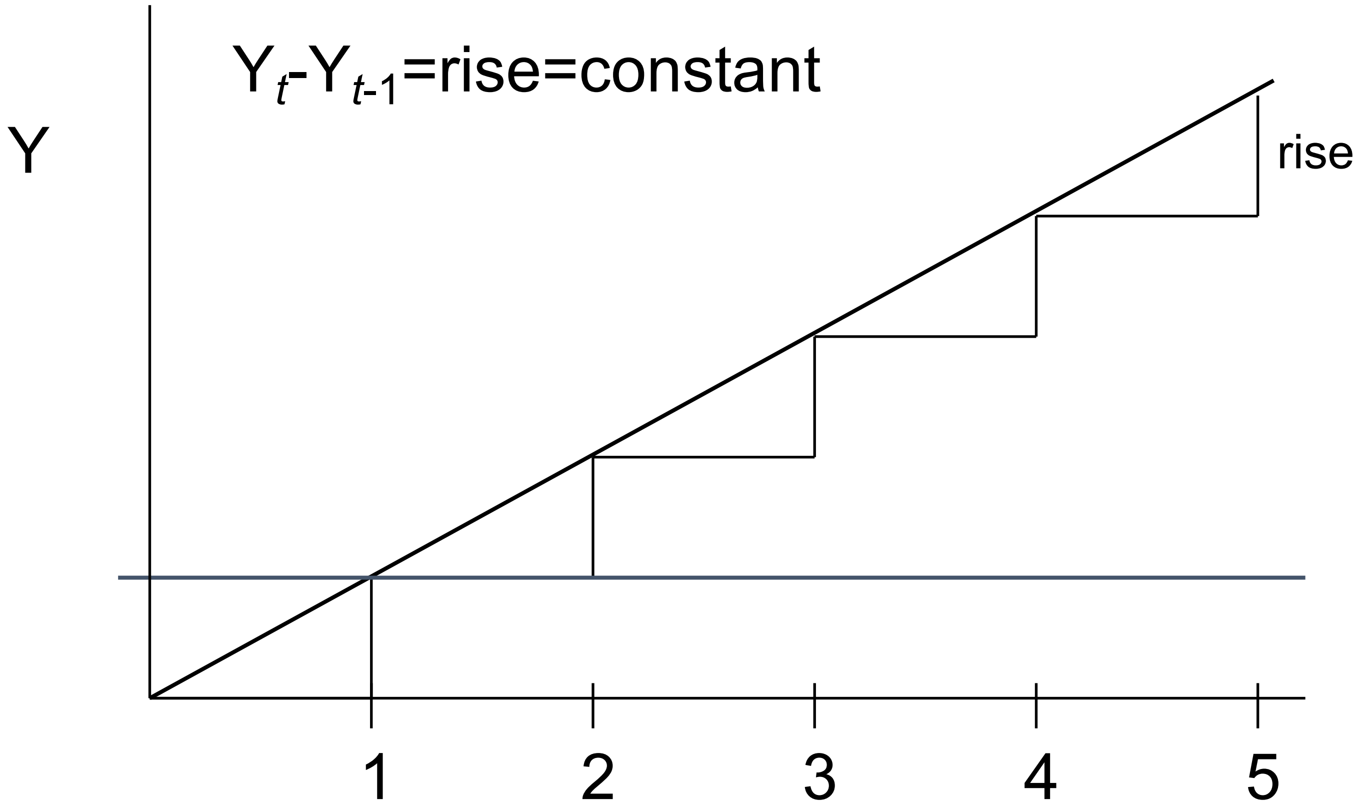
p is the period of seasonality

Two Types of Trend (Seasonality)

- **Deterministic**: a mathematical function of **time**
  - Linear, quadratic, logarithmic, exponential (e.g., $Y_t = \alpha t + \varepsilon_t$)
  - How to model: mathematical functions

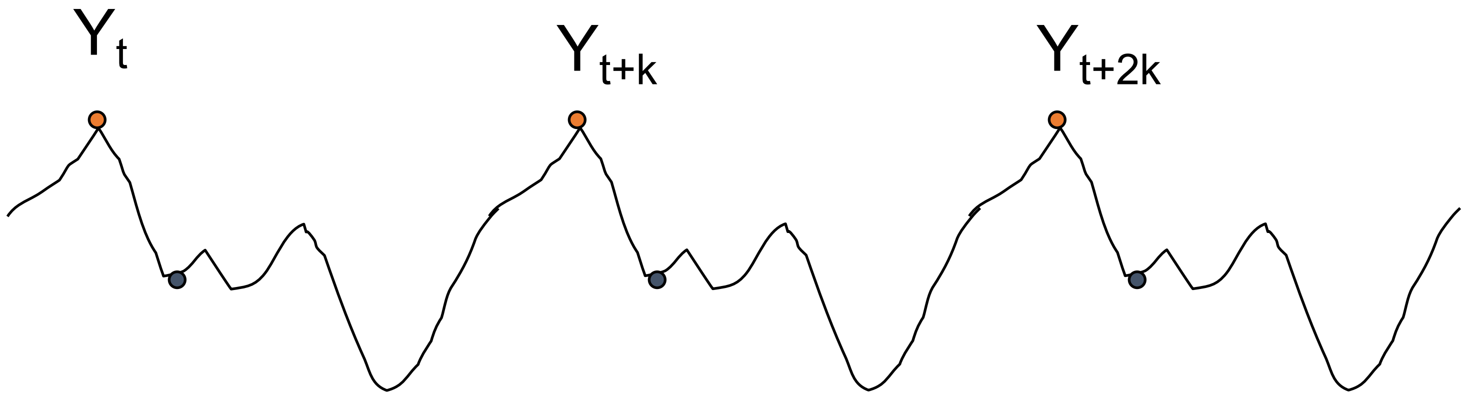
- **Stochastic**: future values depend on **past values** plus error
  - e.g., Random walk with drift ($Y_t = \theta + Y_{t-1} + \varepsilon_t$)
  - How to model: first (seasonal) difference

Frist Difference on Straight Line



Seasonal Difference

- Can account for both stochastic and deterministic seasonality



$\Delta_k = 0$

Takeaway on Deterministic vs. Stochastic Trend

- Stochastic trend increases the variance, whereas deterministic trend changes the mean instead of the variance

- Deterministic trend component cannot address stochastic trend

- First difference cannot address nonlinear deterministic trend

- A time series may exhibit both deterministic and stochastic trend, which may require a combination of first difference and deterministic trend component

$$Y_t = \theta + \alpha t + Y_{t-1} + \varepsilon_t$$

Which of the Following Can Help Diagnose Trend and Seasonality?

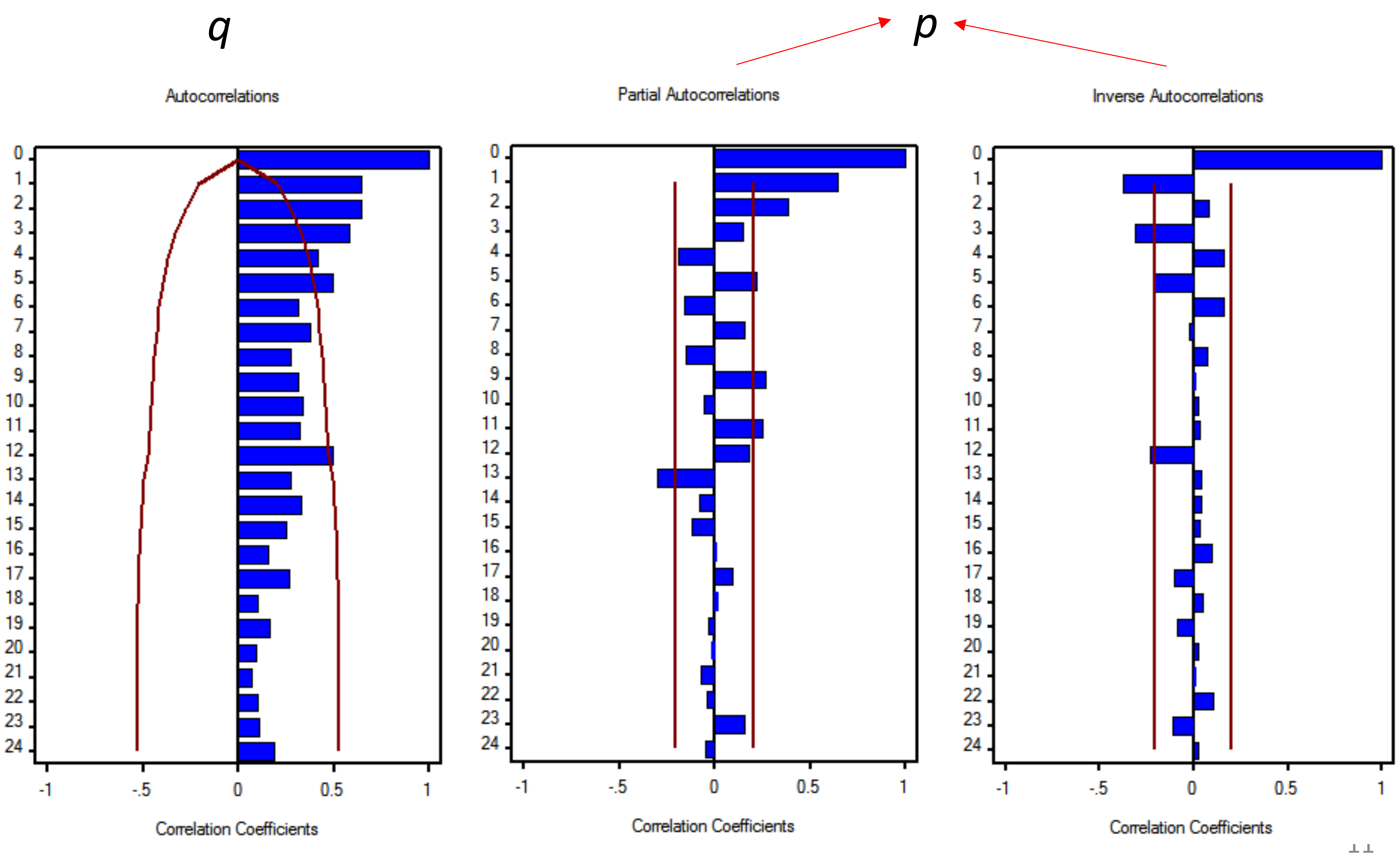
- Time series plot

- Autocorrelation functions

- Unit/Seasonal root test

- White noise test

Identifying Orders of ARMA model



ARIMA(p,d,q) Model Selection

1. Assumes series is stationary. If not, apply first difference first

2. Find q such that ACF(q) falls outside confidence limits and ACF(k) falls inside confidence limits for all $k>q$.

3. Find p such that PACF(p) / IACF(p) falls outside confidence limits and PACF(k) / IACF(k) falls inside confidence limits for all $k>p$.

ARMA(p,d,q) Model Selection

4 Determine all ordered pairs (j,k) such that $0 \leq j \leq p$ and $0 \leq k \leq q$.

5 For each ordered pair (j,k) found in step 4, fit an ARIMA(j,d,k) model.

6 For all of the models fit in step 5, select the model with the smallest values of RMSE on the holdout sample or AIC or SBC on the fit sample.

Regressors

ARIMA Models with Regressors

- Simplest example with a regressor
  - $Y_t = \beta X_t + Z_t$
  - Z_t is an ARIMA error term

Two Types of Regressors

- **Ordinary** regressor: a variable that has a concurrent influence on the target variable
  - X at times before t is uncorrelated with Y at time t

- **Dynamic** regressor: a variable that influences the target variable at current and past values
  - X at times before t can be correlated with Y at time t
  - A dynamic regressor is often specified as a function of an ordinary regressor (transfer function)

- Example: $Y_t = \alpha X_t + \beta Z_{t-1} + \phi_1 Y_{t-1} + \varepsilon_t$

Some Special Regressors

- Time (linear trend, quadratic trend, etc.)
- Seasonal dummies
- Event variables

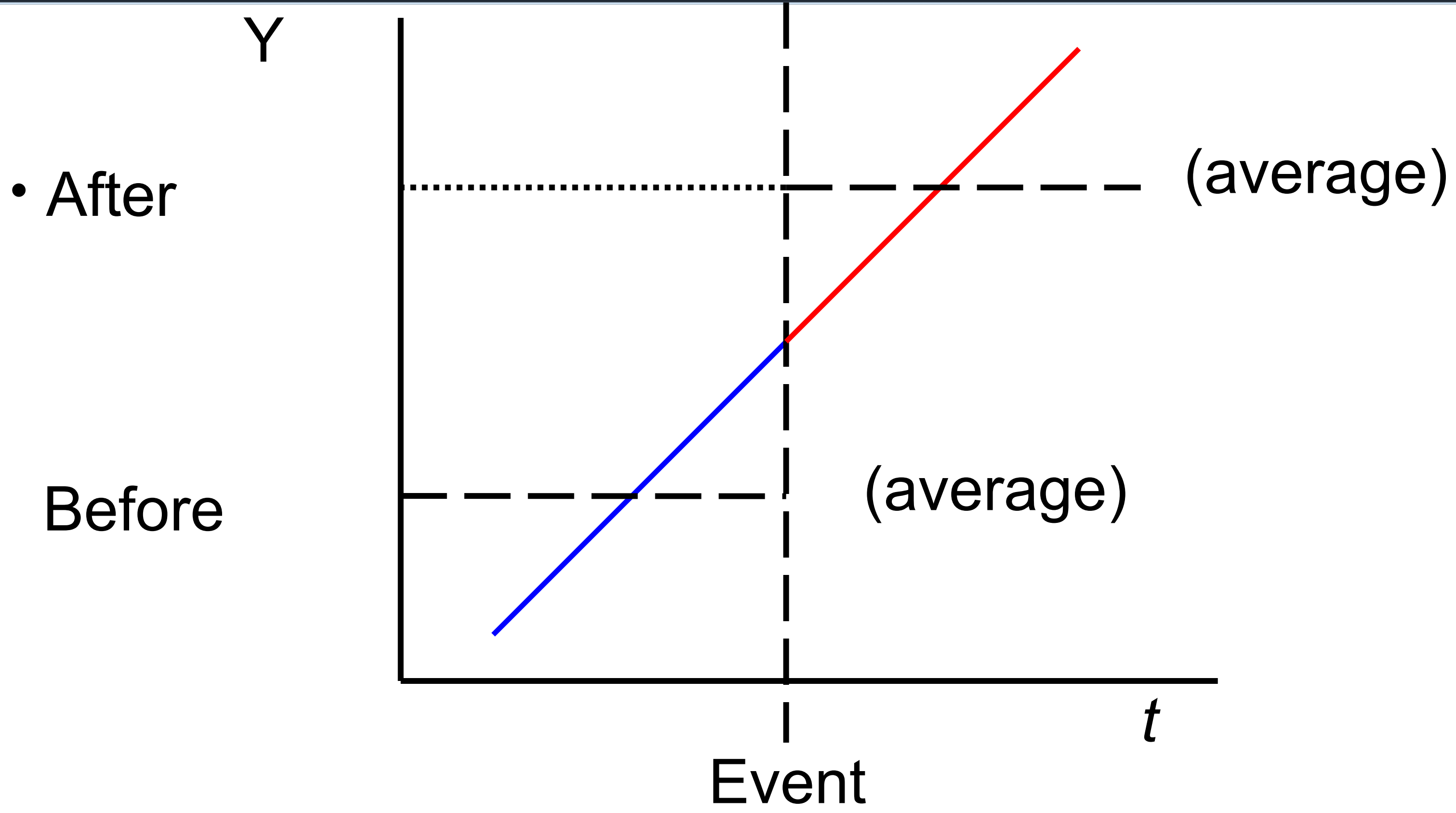
Events

Events (Intervention Analysis)

- An *event* is anything that changes the underlying process that generates time series data, such as
  - Changes in level
  - Changes in trend

- The analysis of events includes two activities:
  - Exploration to identify the functional form of the effect of the event
  - Inference to determine if the event has a statistically significant effect

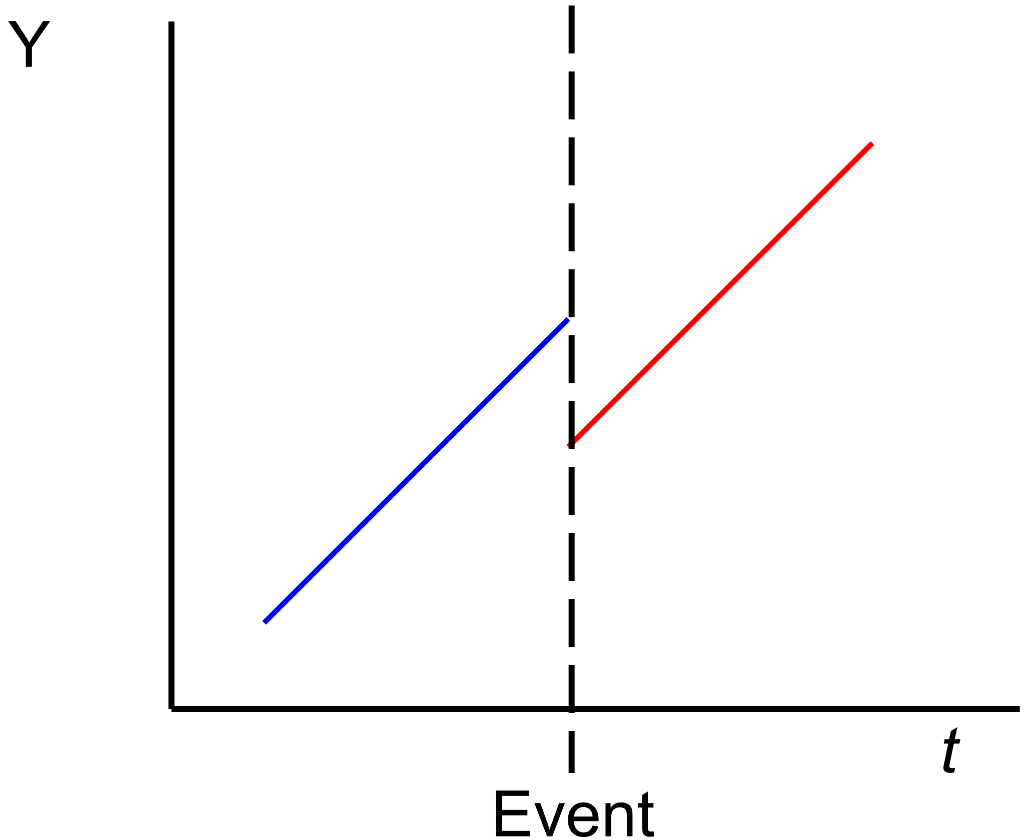
Changes in Level and Trend for Events



False Inference: The event causes the result to increase because AVERAGE(after) > AVERAGE(before).

Valid Inference: The event has no effect on the results.

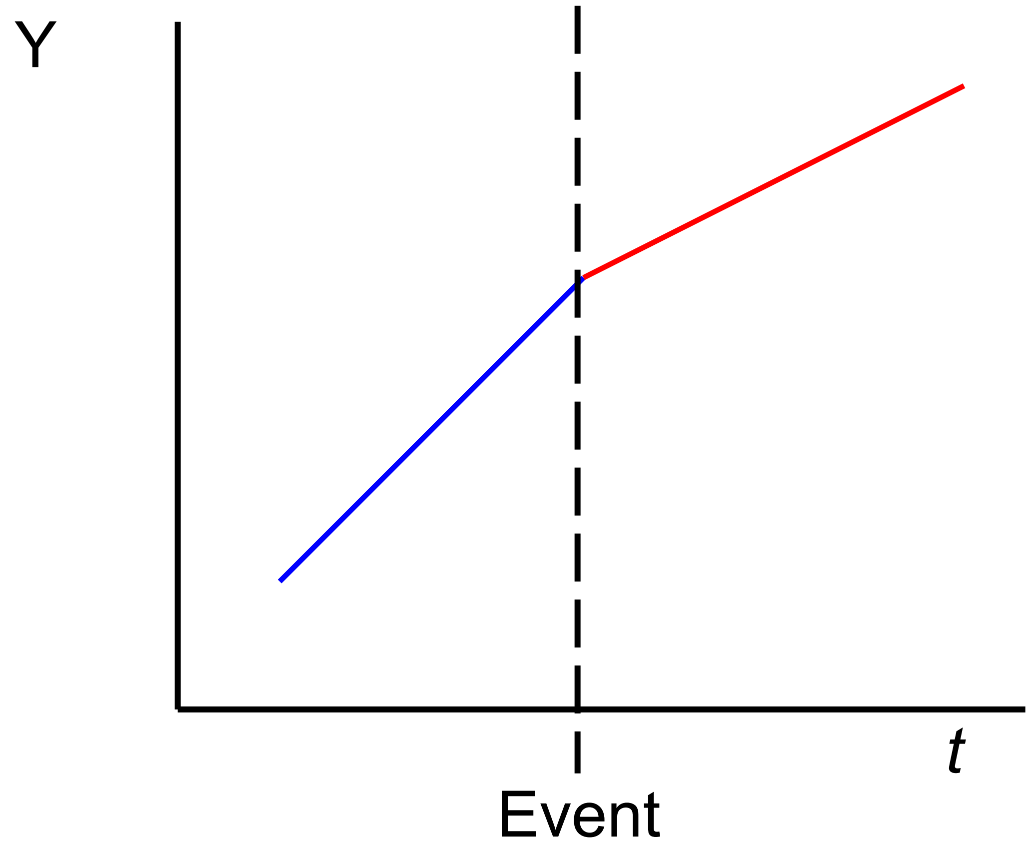
Changes in Level and Trend for Events



Valid Inference: The event causes a change in level.

continued...

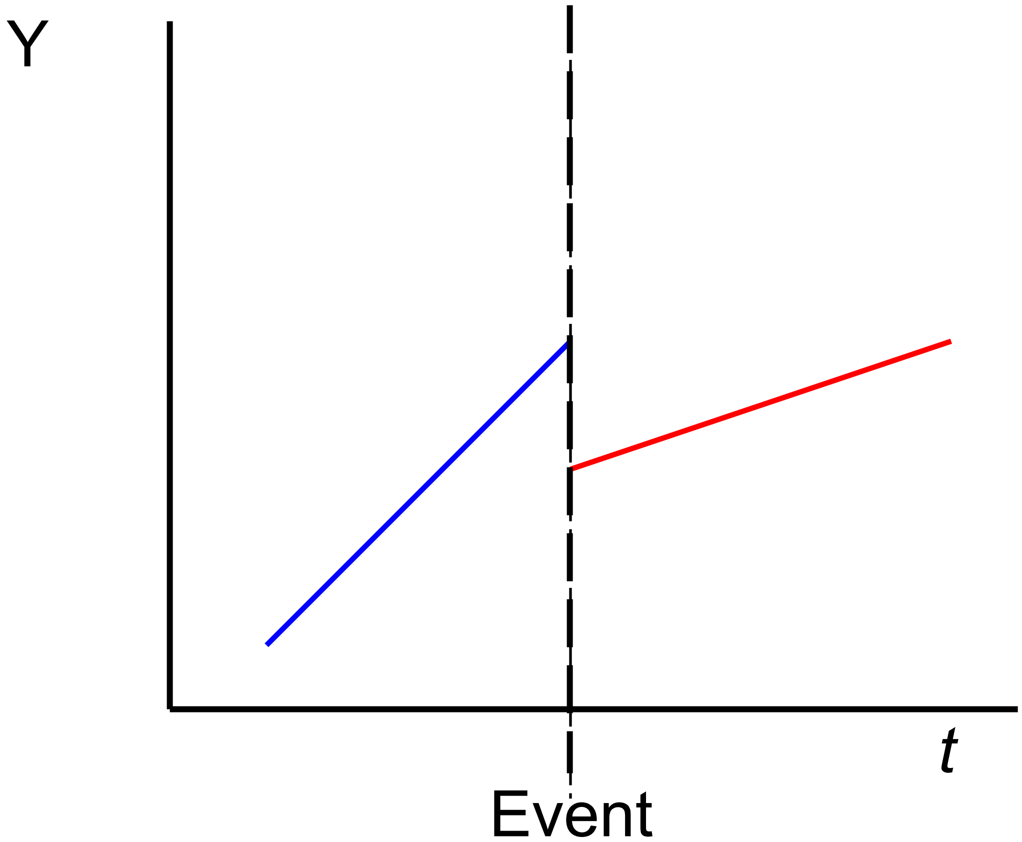
Changes in Level and Trend for Events



Valid Inference: The event causes a change in the slope of the trend line.

continued...

Changes in Level and Trend for Events



Valid Inference: The event causes a change in the level and the slope.

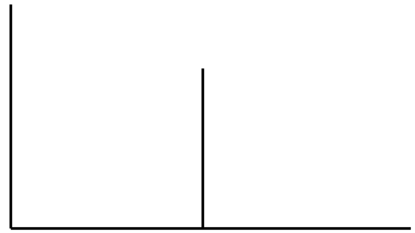
continued...

How to Model Events

- The impact of an event can be captured by an event variable
- We need to construct different types of event variables for different types of events

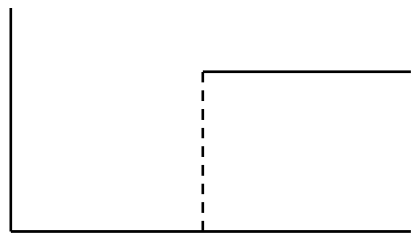
Primary Event Variables

Point/Pulse



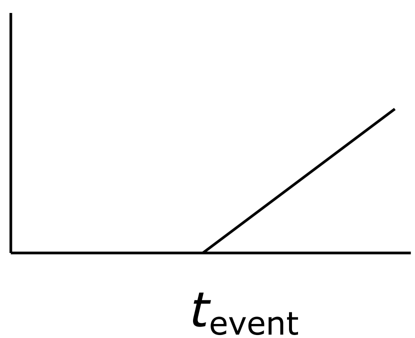
$$J_t = \begin{cases} 0 & \text{for } t \neq t_{\text{event}} \\ 1 & \text{for } t = t_{\text{event}} \end{cases}$$

Step



$$I_t = \begin{cases} 0 & \text{for } t < t_{\text{event}} \\ 1 & \text{for } t \geq t_{\text{event}} \end{cases}$$

Ramp



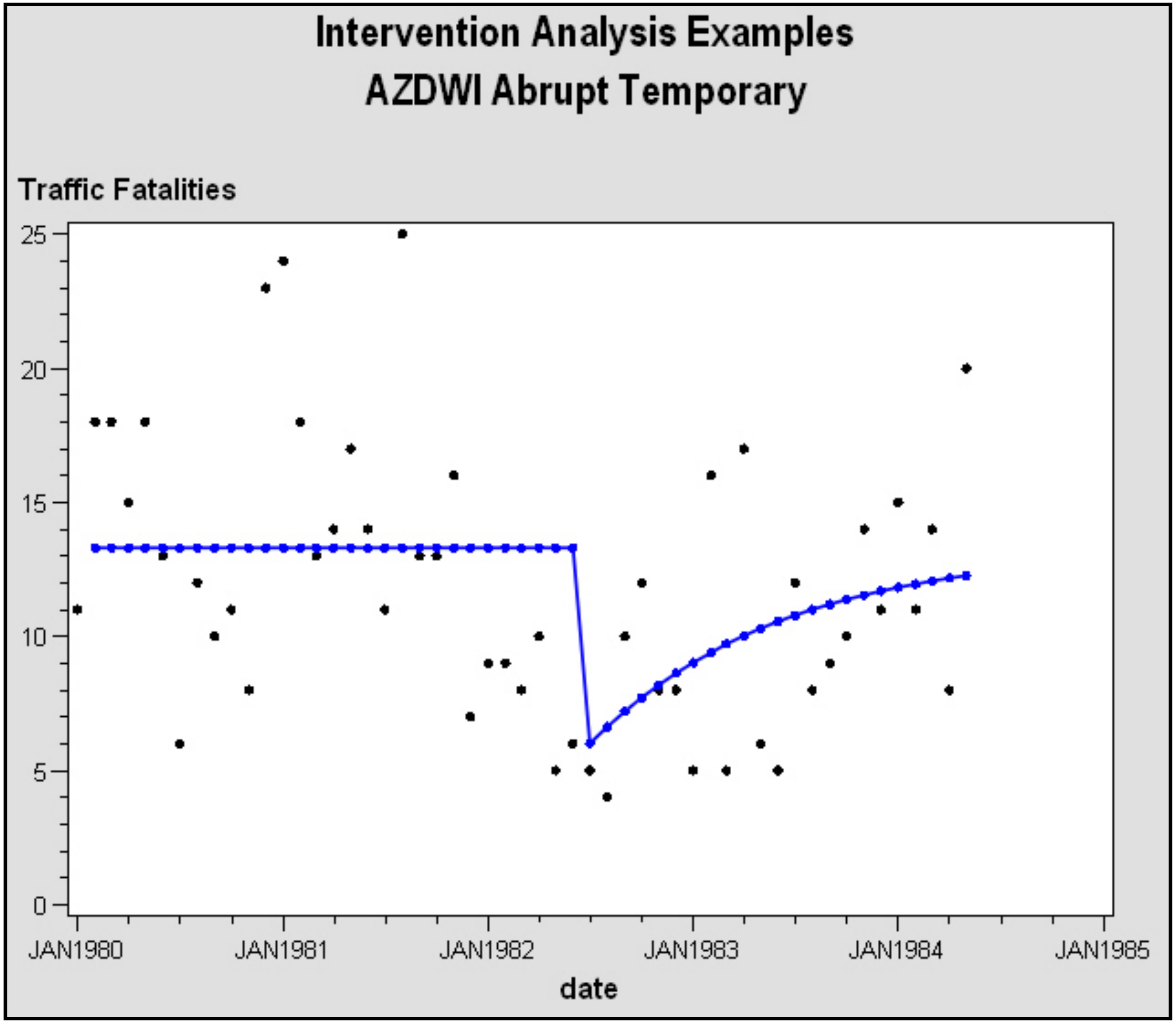
$$R_t = \begin{cases} 0 & \text{for } t < t_{\text{event}} \\ t - t_{\text{event}} & \text{for } t \geq t_{\text{event}} \end{cases}$$

Other Types of Changes

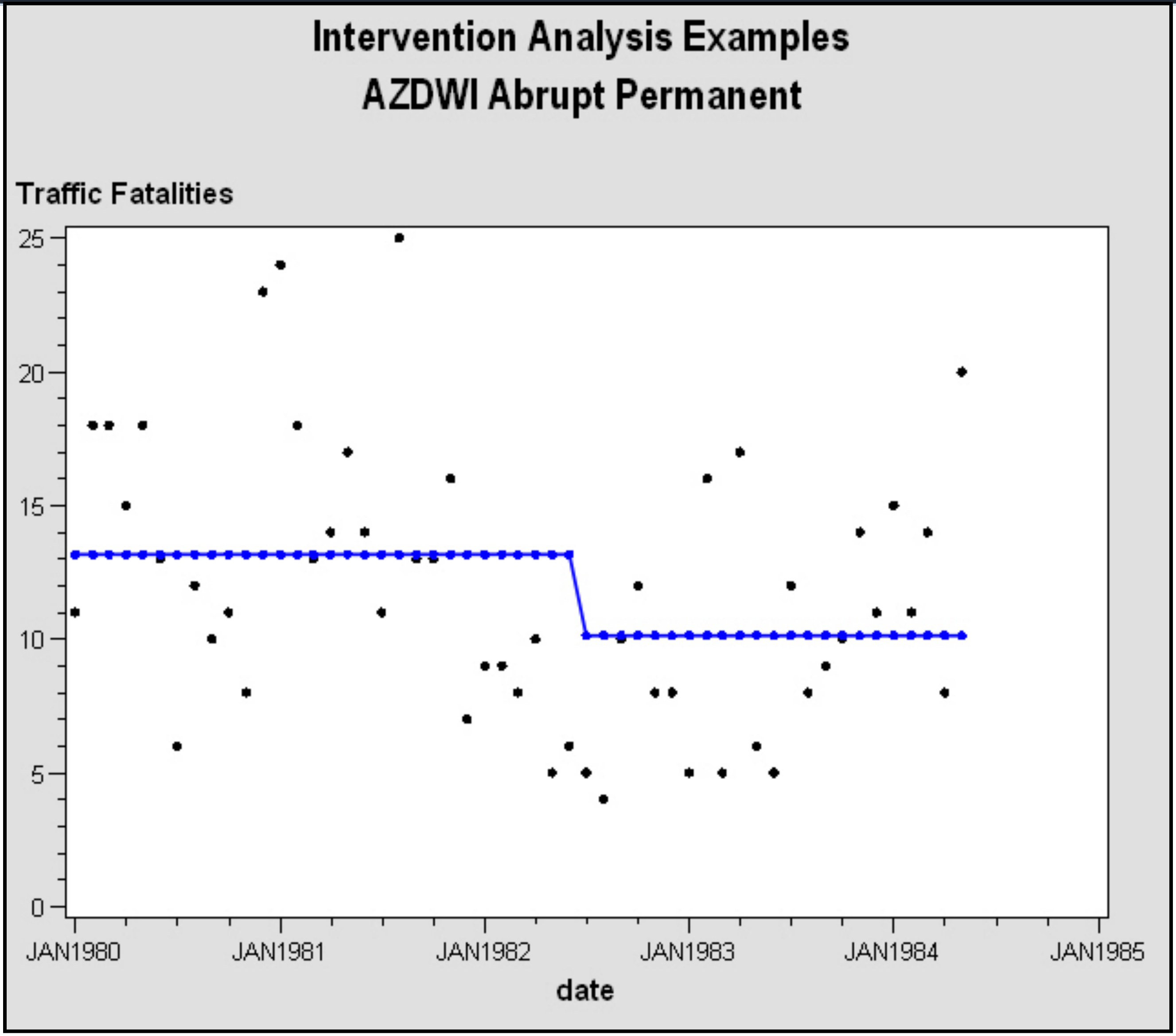
- The primary event variables can only capture very restrictive changes

- Given that the effect of an event typically varies over time, the change resulting from an event can be a lot more complicated

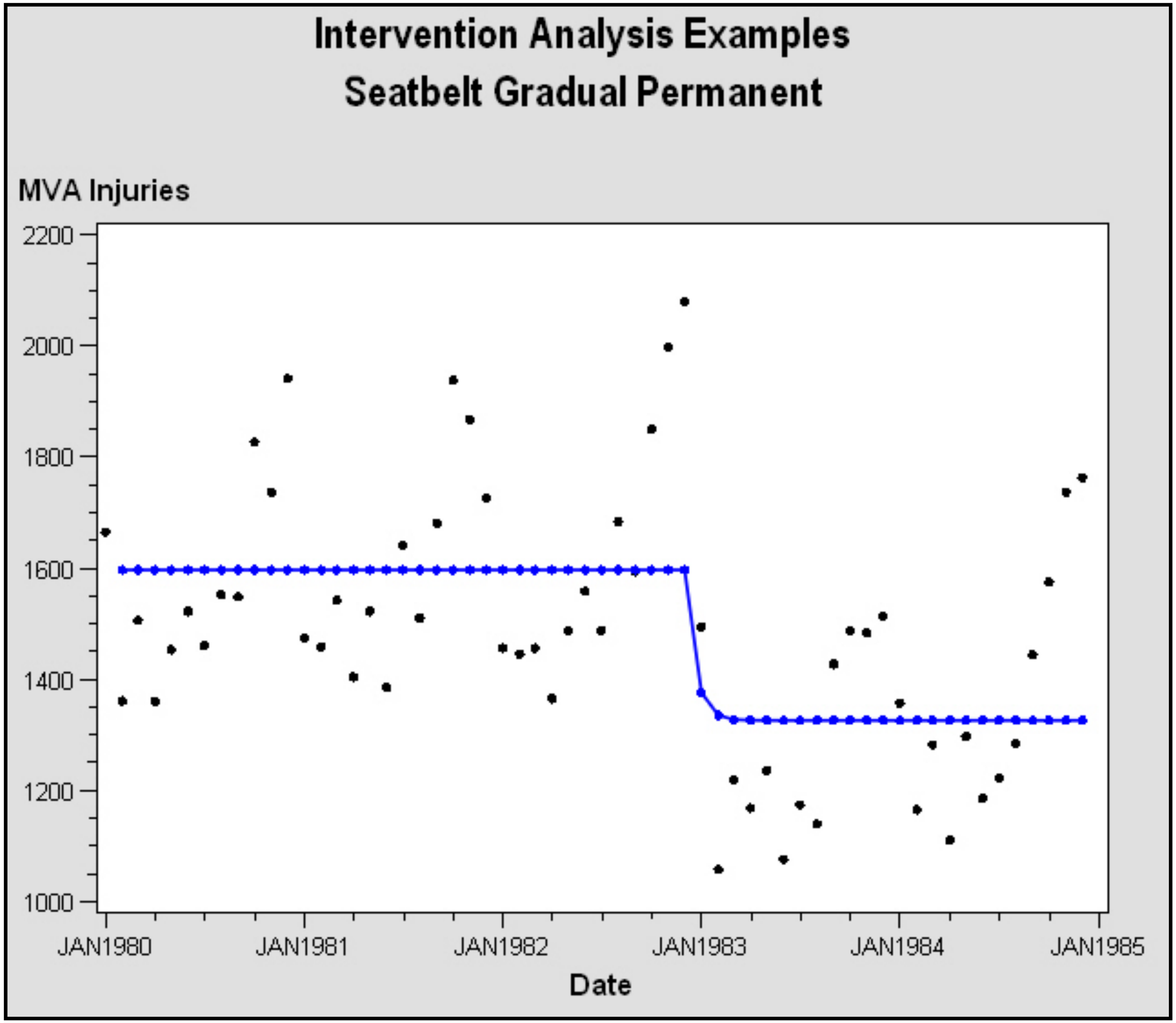
Abrupt, Temporary Effect



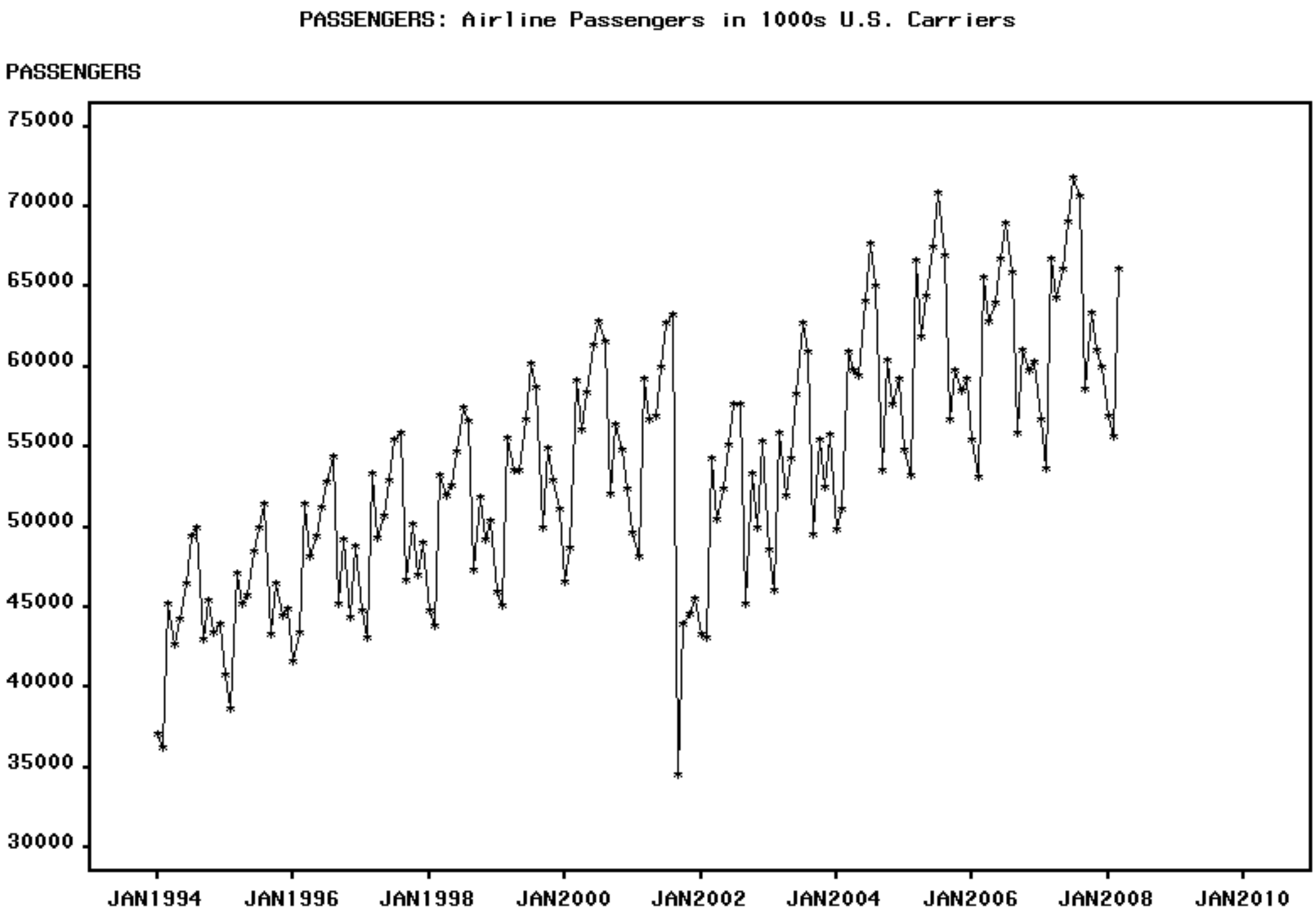
Abrupt, Permanent Effect



Gradual, Permanent Effect



What Type of Effect is this?



Transfer Function

Transfer Function

- A function that provides the mathematical relationship between a regressor (including event variable) and the target variable.

- Transfer functions allow us to account for the time varying effect of a regressor (including event variable)

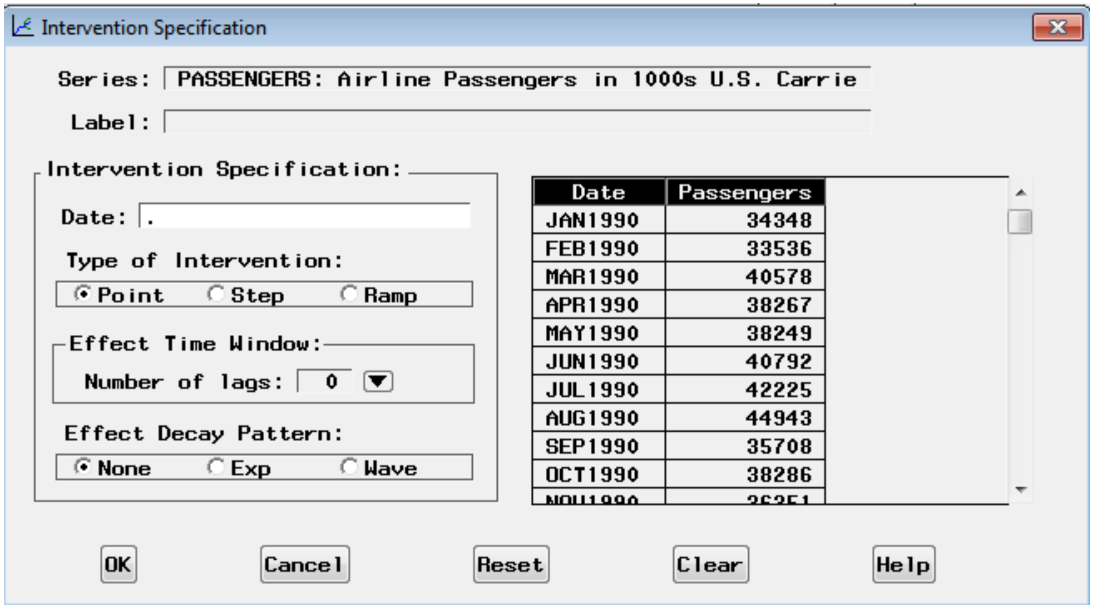
Coefficients ⇨ Transfer Functions

Ordinary Regression	Regression with Transfer Function
$$Y_t = \omega_0 X_t + Z_t$$	$$Y_t = \frac{\omega(B)}{\delta(B)} X_t + Z_t$$

Transfer Functions

- Z_t is an ARMA error term
- $\omega(B) = \omega_0 + \omega_1 B + \omega_2 B^2 + \cdots + \omega_m B^m$
- $\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_n B^n$

Transfer Function for Events

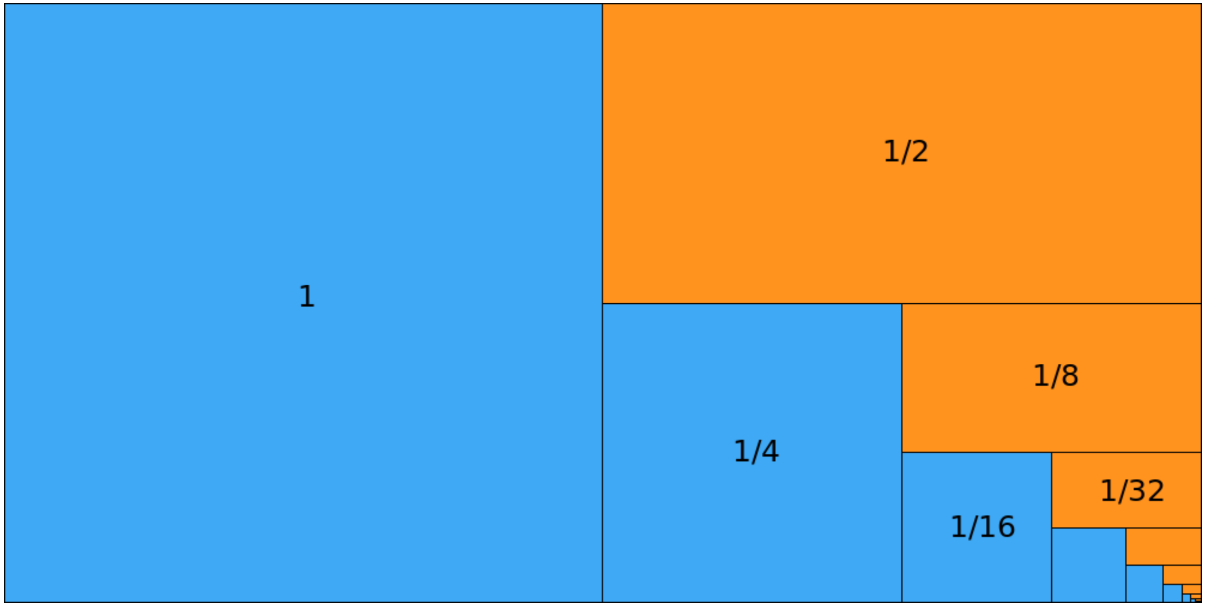


- Effect Time Window: number of lags to include
  - Specifies the order of the **numerator** $\omega(B)$
  - Typically set to 0 for event variables
- Effect Decay Pattern: how effect of the event decays
  - Specifies the order of the **denominator** $\delta(B)$
  - Exp: the event has sustaining effect that decays exponentially
  - Wave: the event has sustaining effect that decays like a wave

Recap: Infinite Geometric Series

- Suppose the absolute value of the common ratio r is less than 1, the sum of an infinite geometric series can be written as

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$$

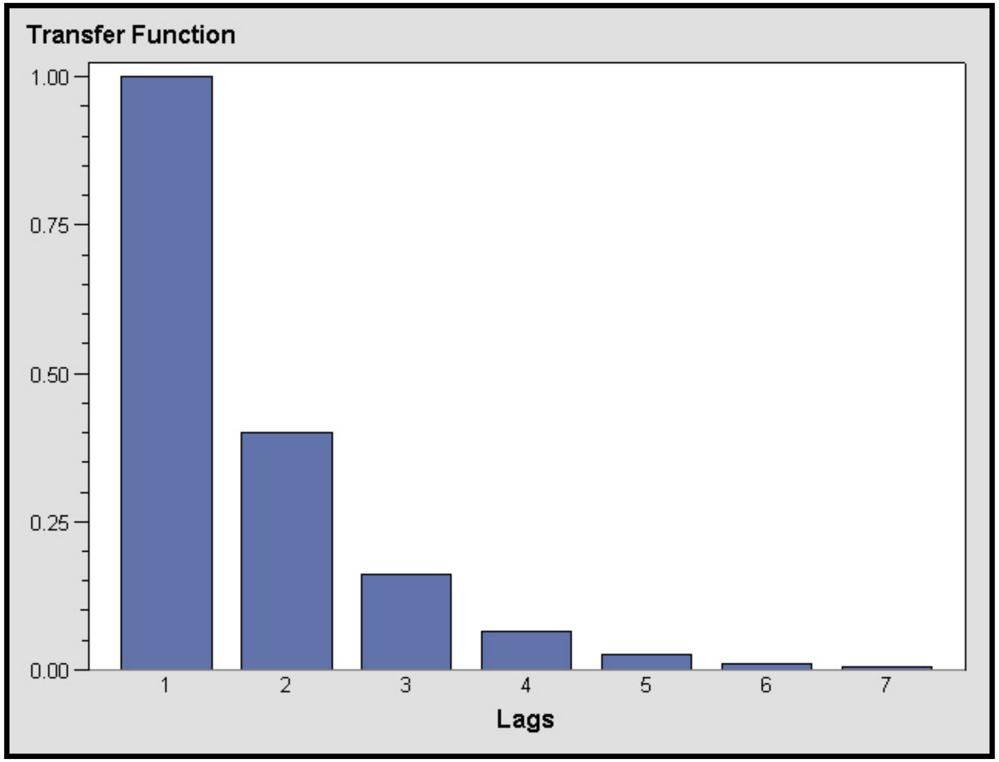


https://en.wikipedia.org/wiki/Geometric_series

Exponential Decay (Infinite Memory)

$$\frac{\omega(B)}{\delta(B)}X_t = \frac{\omega_0}{1-\delta_1 B}X_t \;= \omega_0(1+\delta_1 B+\delta_1^2 B^2+\cdots)X_t$$

Order of numerator is 0, order of denominator is 1



$$\omega_0 = 1, \delta_1 = 0.4$$

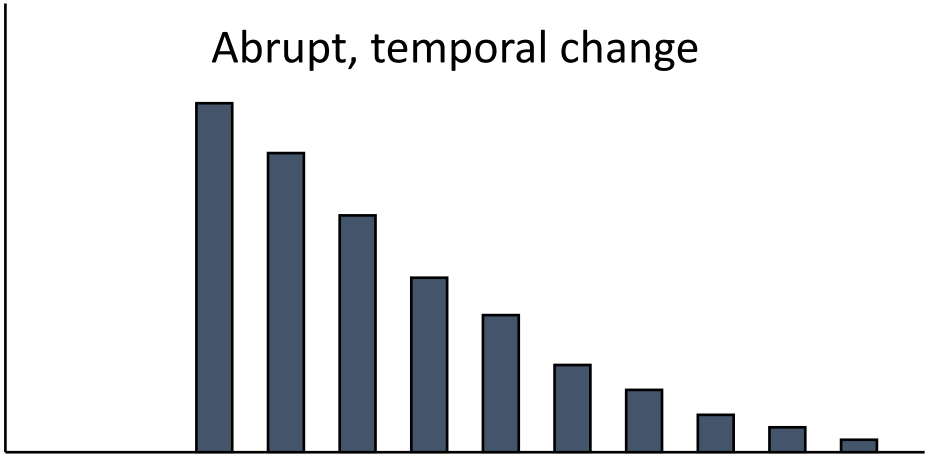
Exponential Decay: Point vs. Step Event

- Point: the event variable X_t is nonzero only on the event day T
  - Effect of X_T on Y_{T+m}:
  - Effect of X_{T+1} on Y_{T+m}:
  - Effect of X_{T+m} on Y_{T+m}:

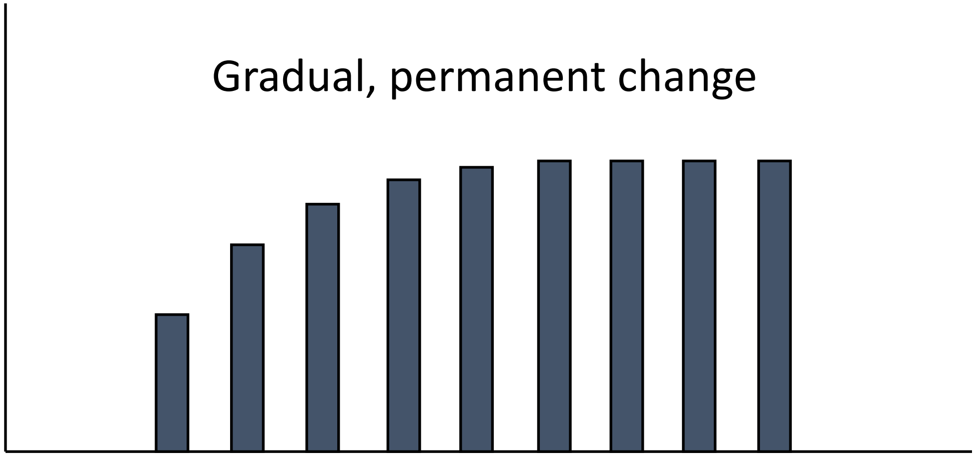
- Step: the event variable X_t is nonzero starting from the event day T
  - Effect of X_T on Y_{T+m}:
  - Effect of X_{T+1} on Y_{T+m}:
  - Effect of X_{T+m} on Y_{T+m}:

Exponential Decay: Point vs. Step Event

- Point with exponential decay

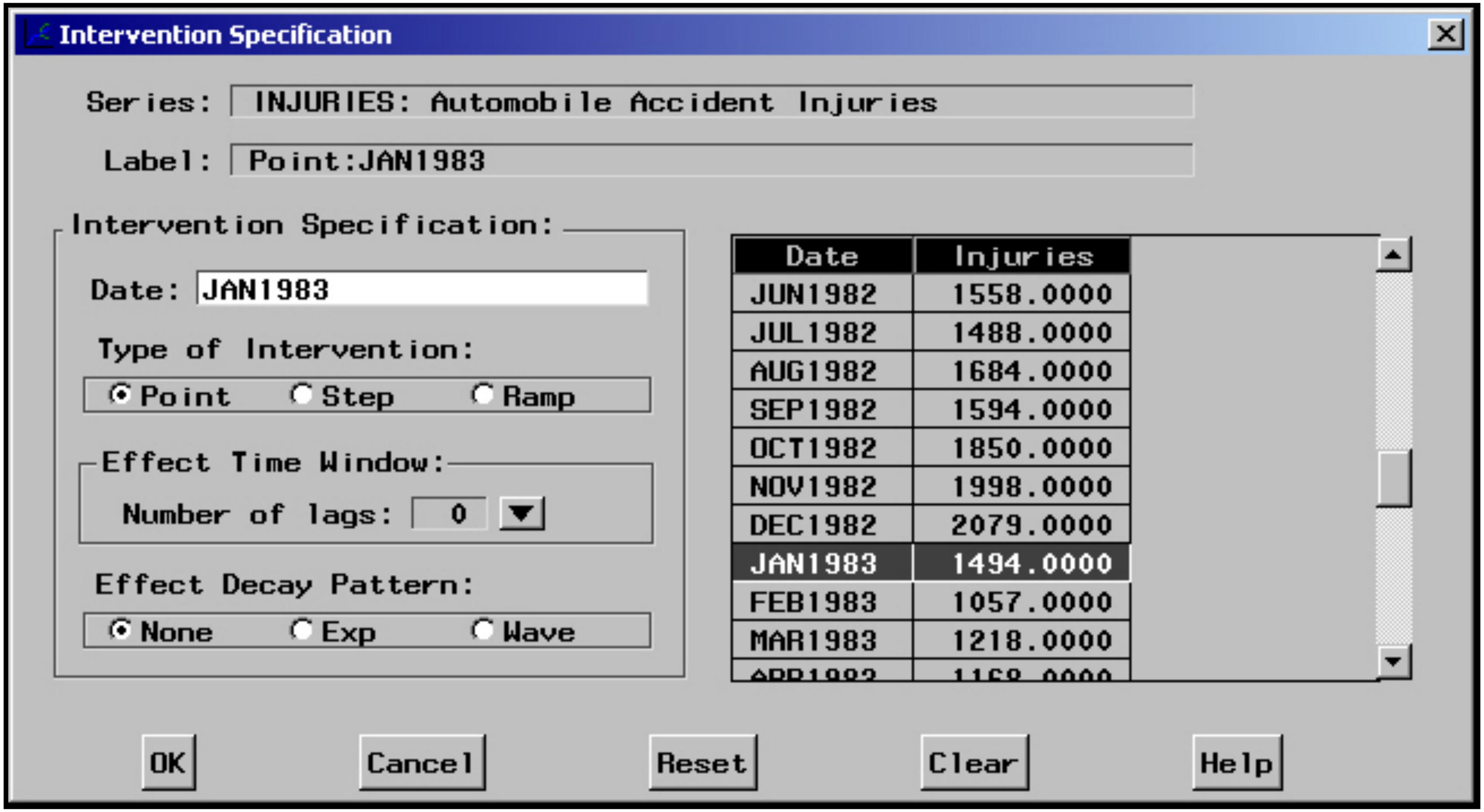


- Step with exponential decay

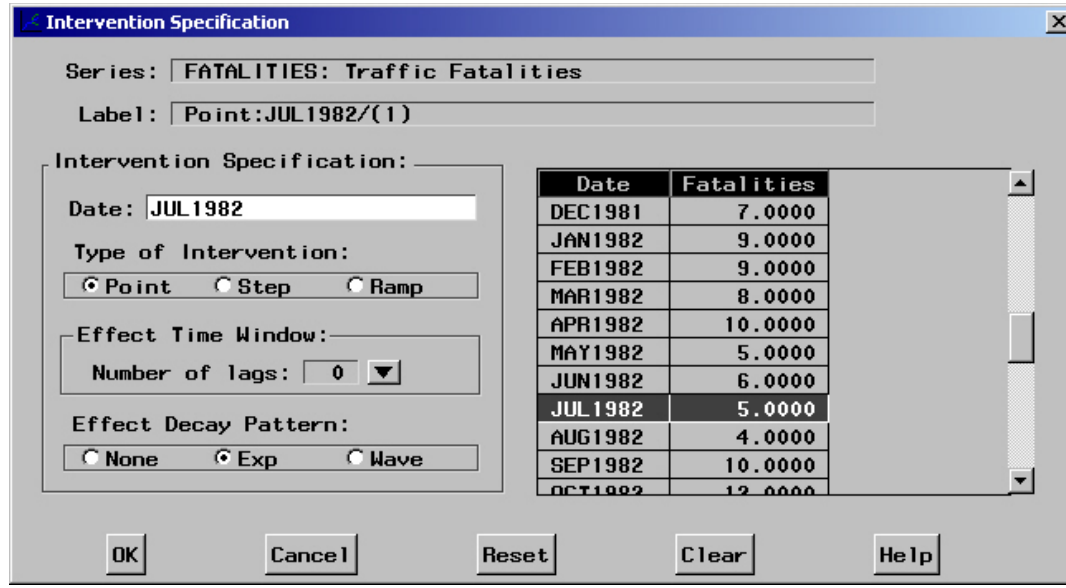


Transfer Functions for Events

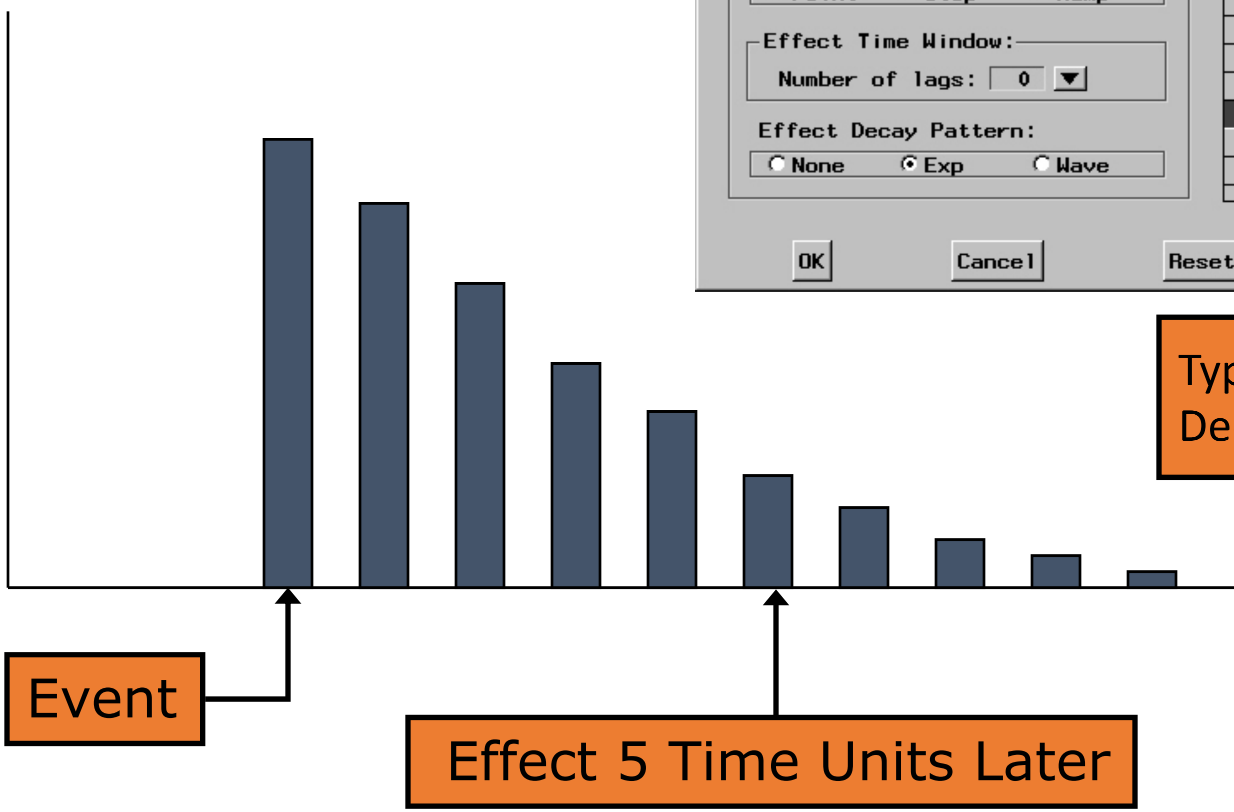
TSFS Intervention Specification Window



Abrupt, Temporary Effect

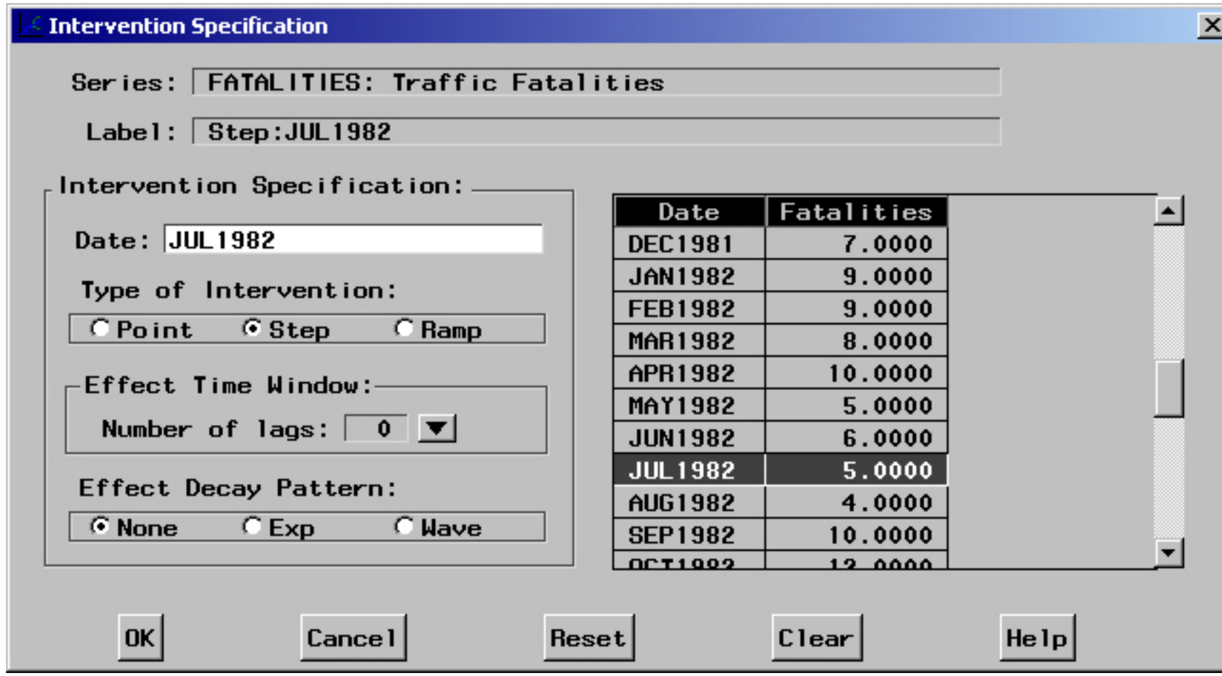


Type: Point
Decay Pattern: Exp

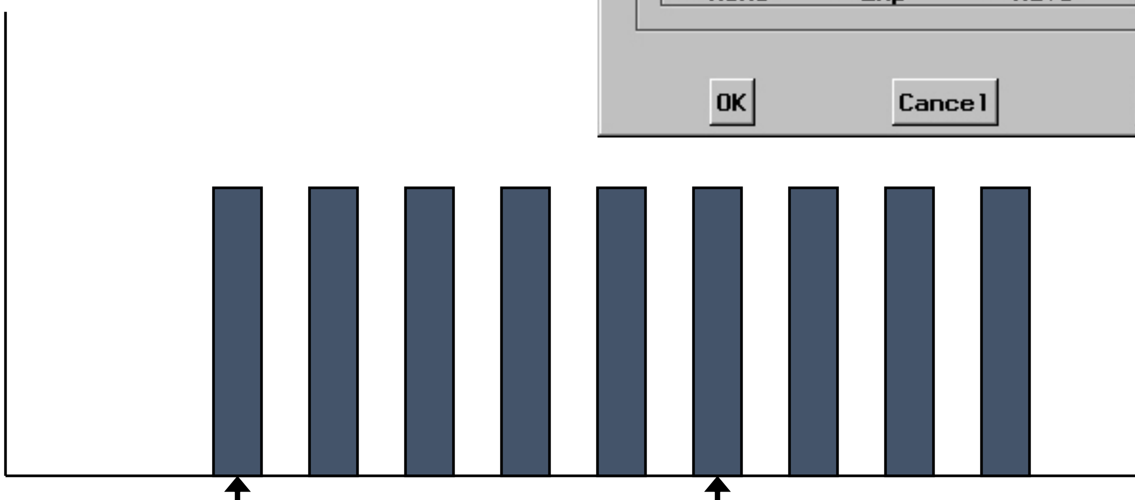


continued...

Abrupt, Permanent Effect



Type: Step
Decay Pattern: None

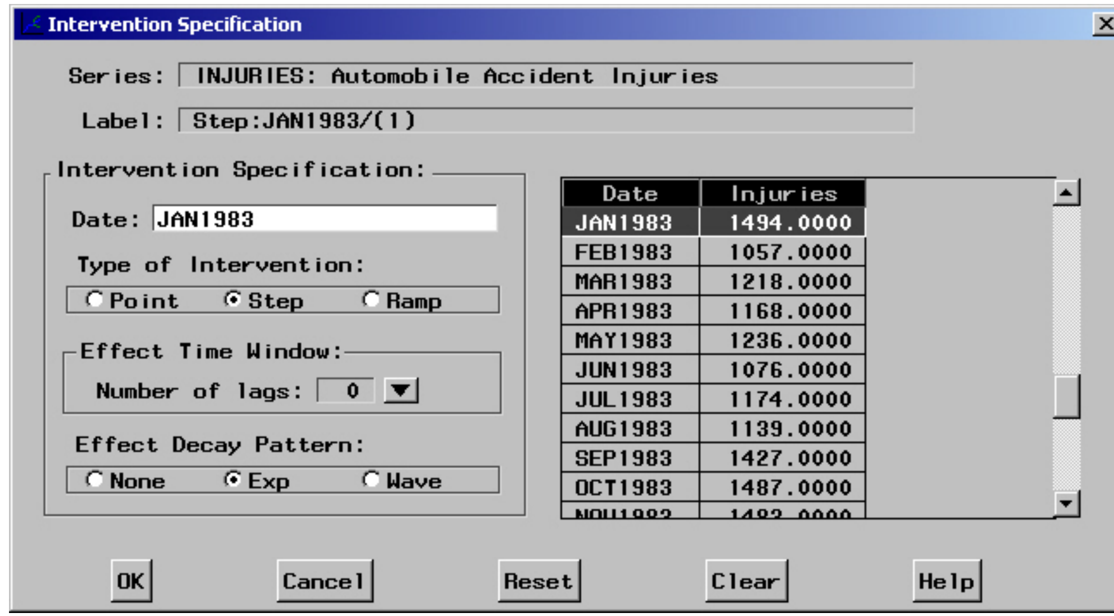


Event

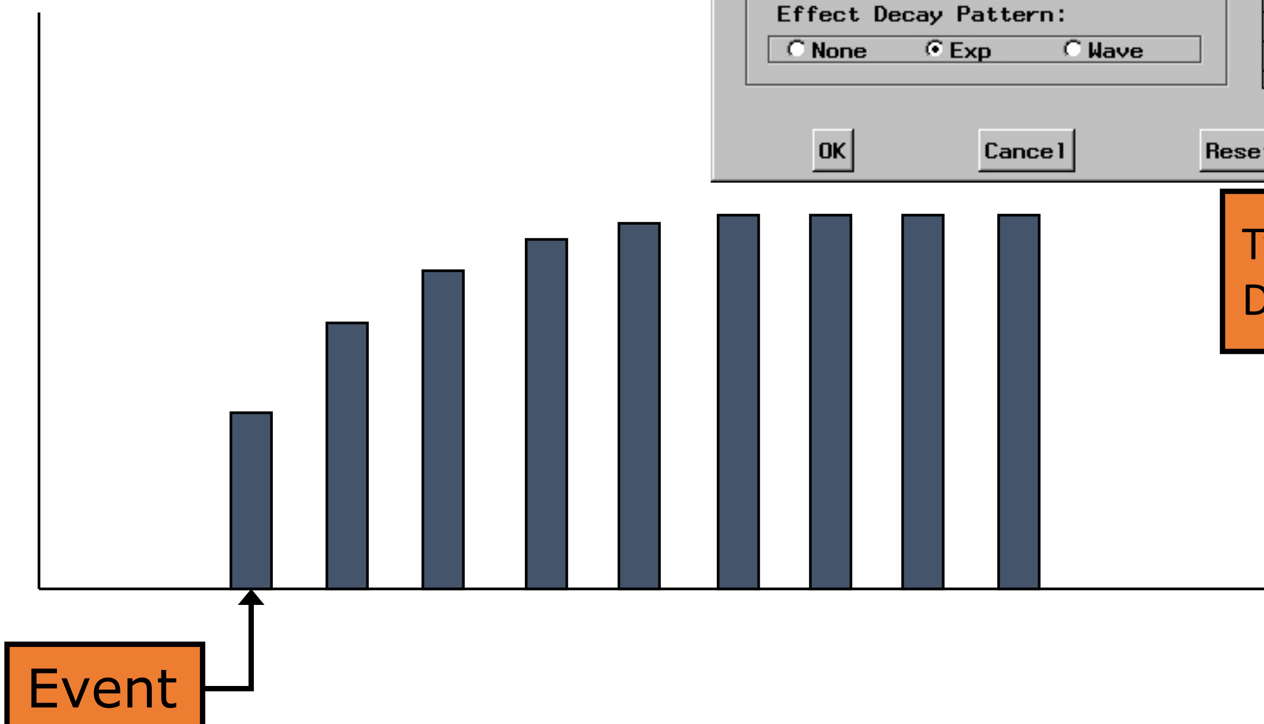
Effect 5 Time Units Later

continued...

Gradual, Permanent Effect



Type: Step
Decay Pattern: Exp



continued...

Demo

Estimate a seasonal ARIMA model with events for Airline data

Chapter 5 p24-55

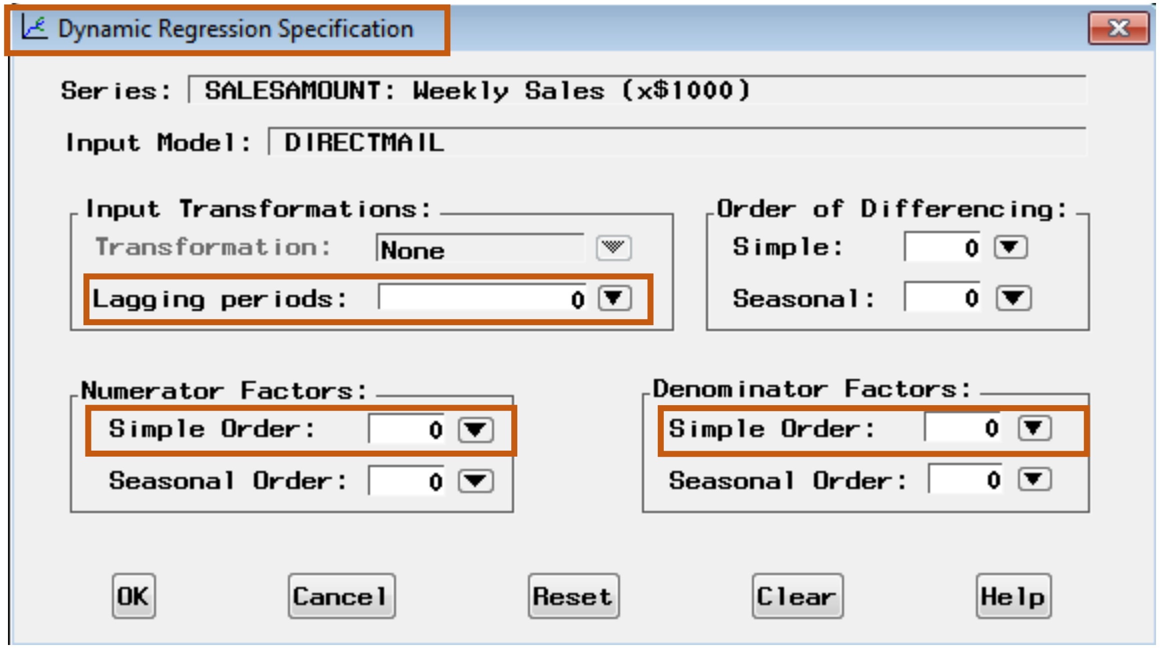
- Examine event type after taking first and seasonal differences
- Examine event type after including linear trend and seasonal dummies

Transfer Functions for Regressors

Transfer Function for Regressors

X_{t-k}

$\omega(B)$



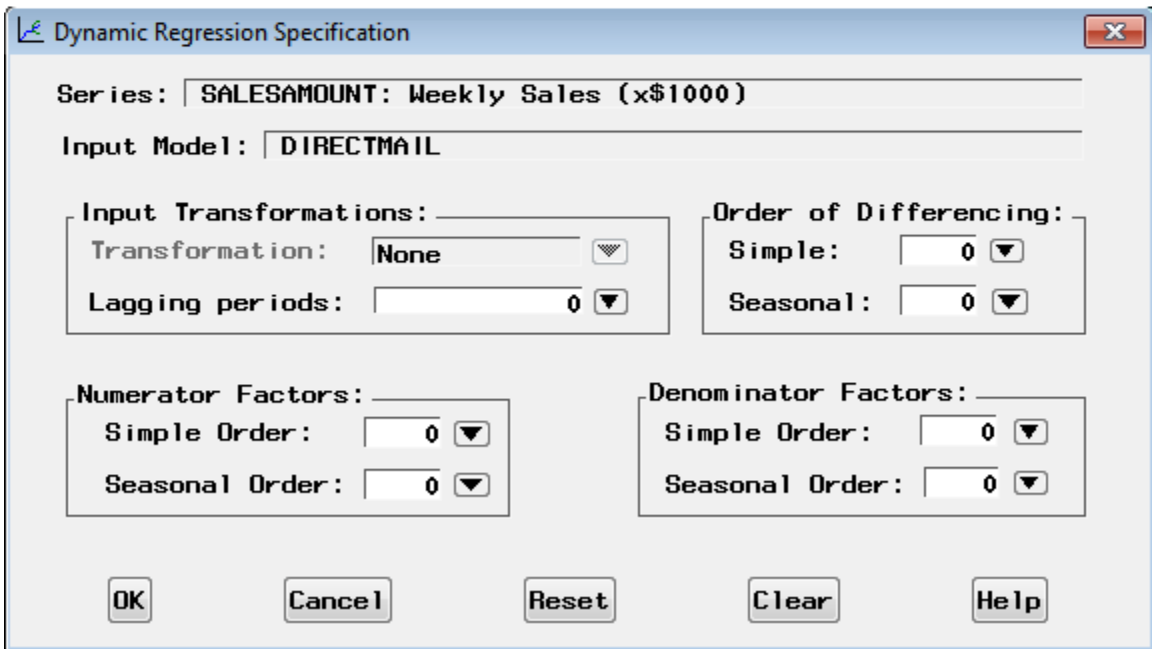
$\delta(B)$

- Differencing: if X has trend and seasonality
- Lagging periods: shift X to the past by k periods
- Seasonal Orders: replace B in $\omega(B)$ and $\delta(B)$ with B^S

Check this URL for more details on seasonal orders

Examples: Ordinary Regression

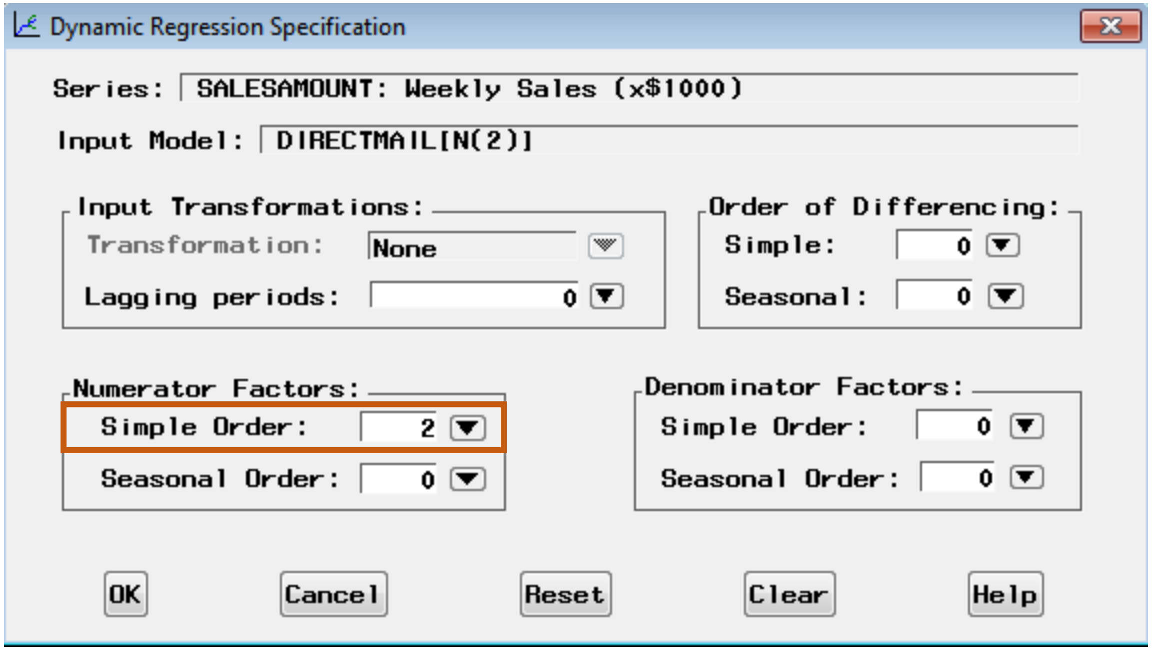
$Y_t = \omega_0 X_t + Z_t$, Z_t is an ARIMA error term



$$\frac{\omega(B)}{\delta(B)} = \omega_0$$

Examples: Lagged Regression

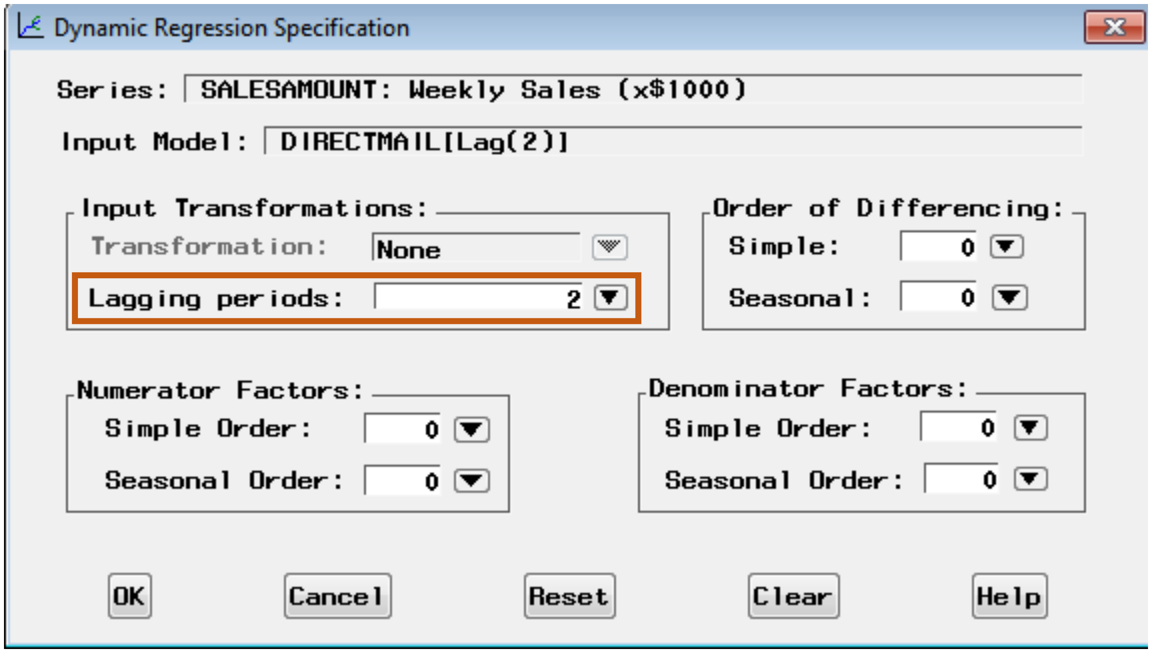
$$Y_t = \omega_0 X_t + \omega_1 X_{t-1} + \omega_2 X_{t-2} + Z_t$$



$$\frac{\omega(B)}{\delta(B)} = \omega_0 + \omega_1 B + \omega_2 B^2$$

Examples: Shifted Regression

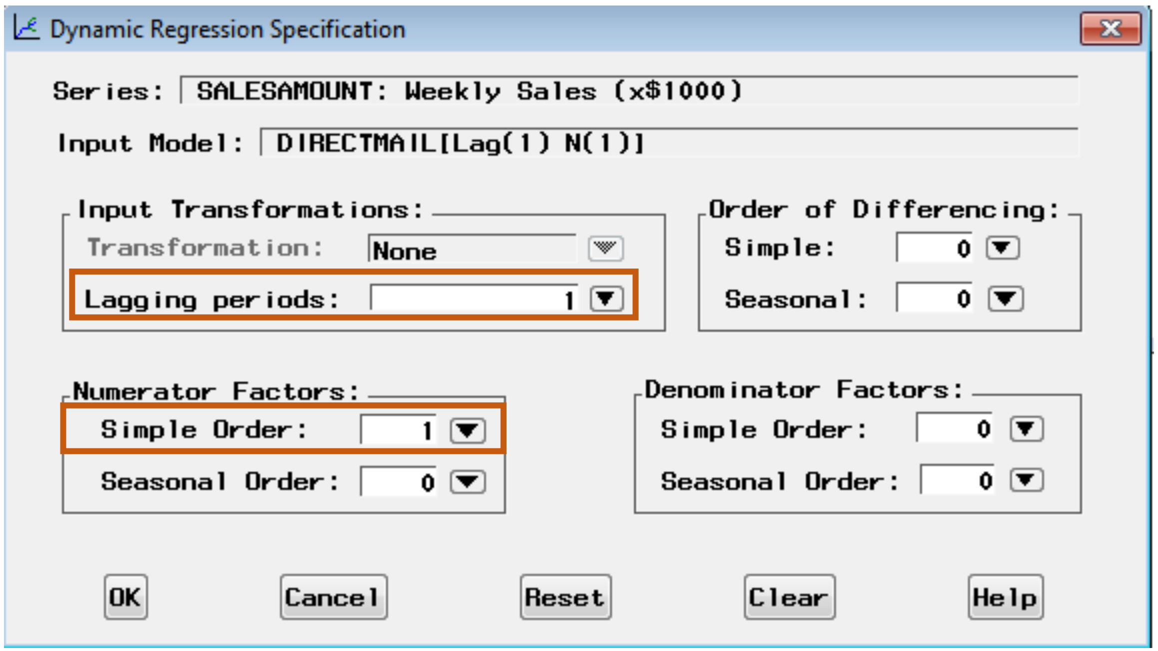
$$Y_t = \omega_2 X_{t-2} + Z_t$$



$$\frac{\omega(B)}{\delta(B)} = \omega_2 B^2$$

Examples: Shifted Regression with Lags

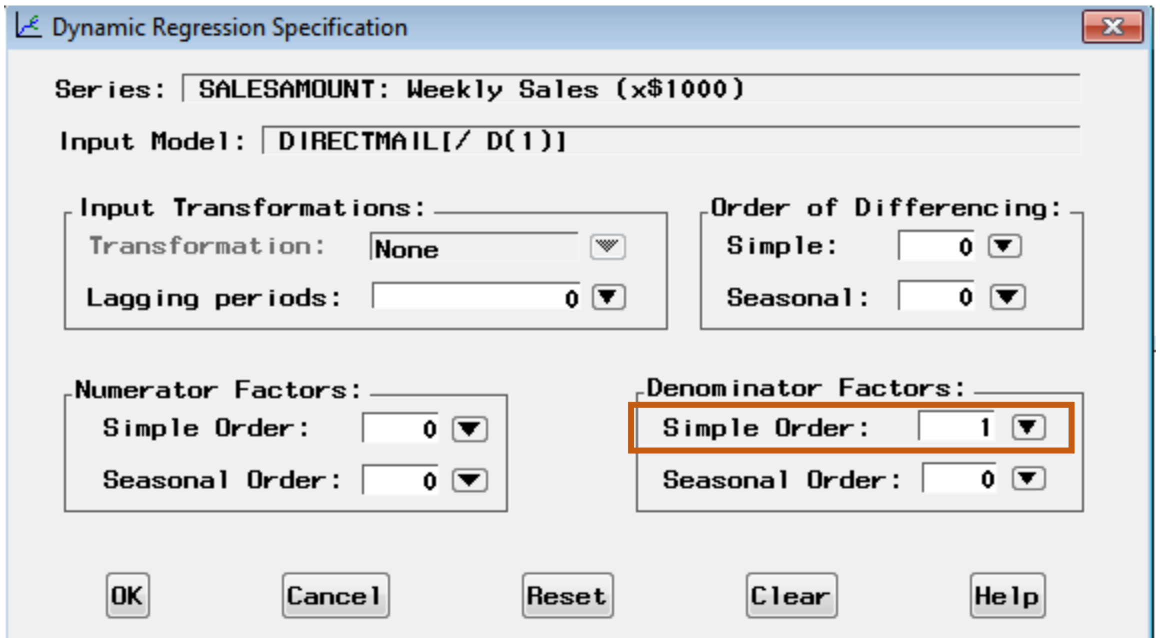
$$Y_t = \omega_0 X_{t-1} + \omega_1 X_{t-2} + Z_t$$



$$\frac{\omega(B)}{\delta(B)} = (\omega_0 + \omega_1 B)B$$

Examples: Infinite Past Regression

$$Y_t = \frac{\omega_0}{1 - \delta_1 B} X_t + Z_t$$



$$\frac{\omega(B)}{\delta(B)} = \frac{\omega_0}{1 - \delta_1 B}$$

Demo

Use advertising spending to predict sales

Chapter 5 p70-81

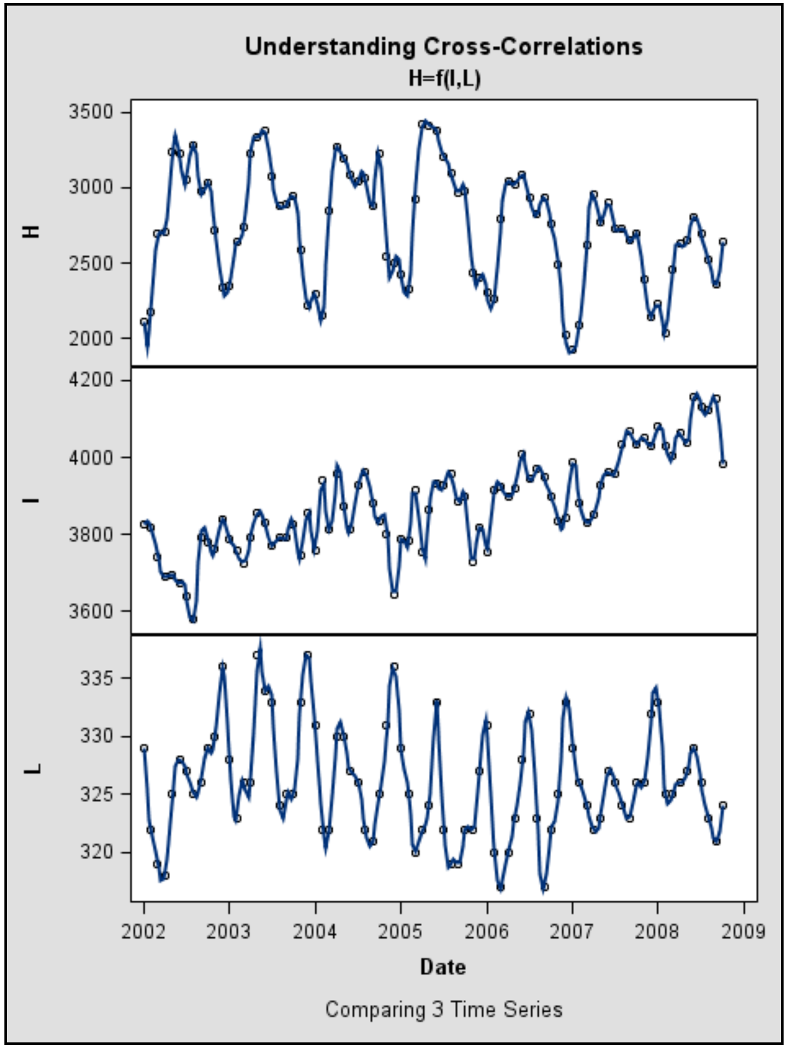
Cross-Correlation Function

Cross-Correlation Function (CCF)

- CCF(k) is the cross-correlation of target Y with input X at lag k.
  - A significant value at lag k implies that Y_t and X_{t-k} are correlated.
  - Spikes and decay patterns in the cross-correlation function can help determine the form of the transfer function.

- The calculation of CCF can be tricky

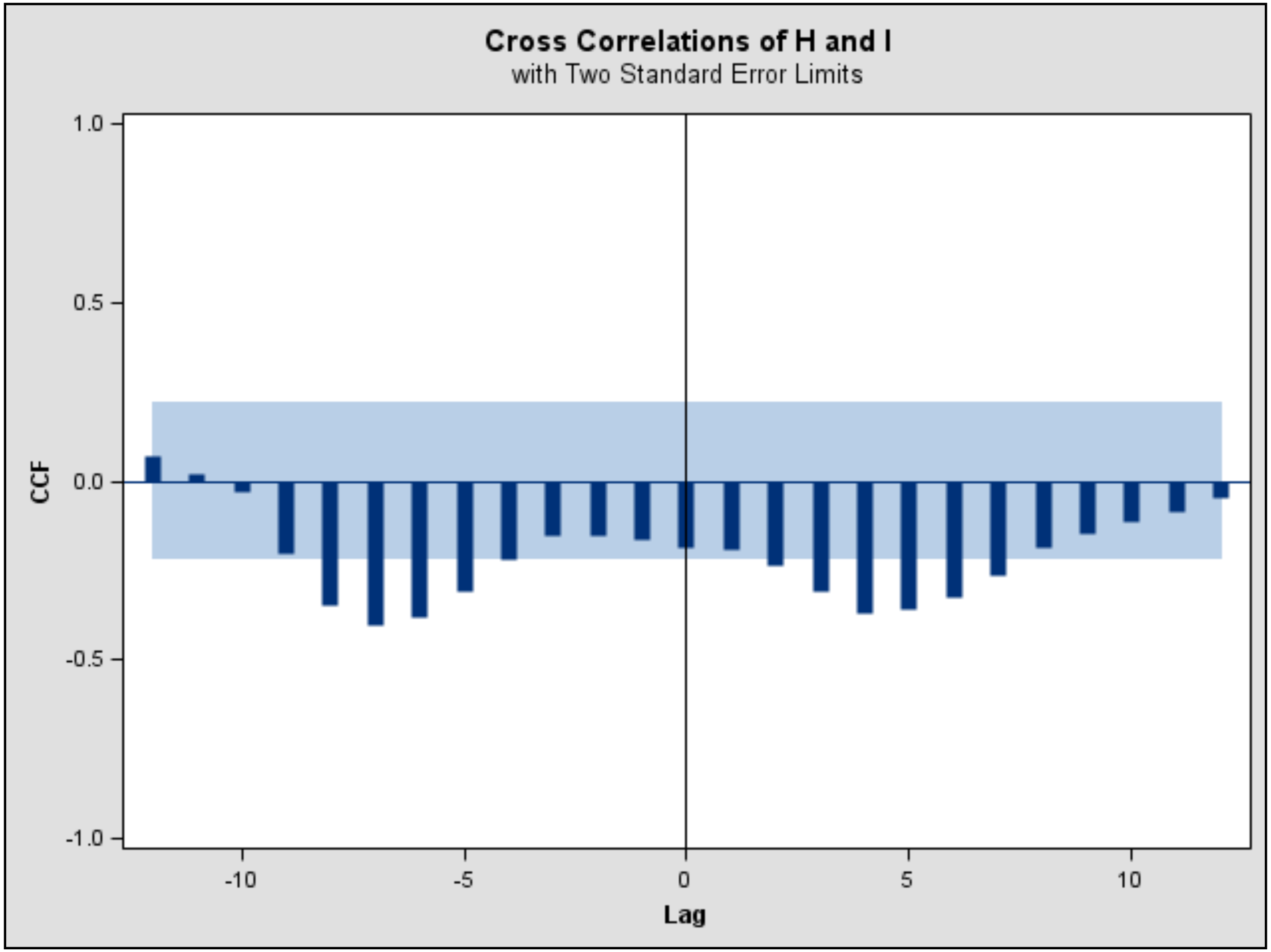
Example



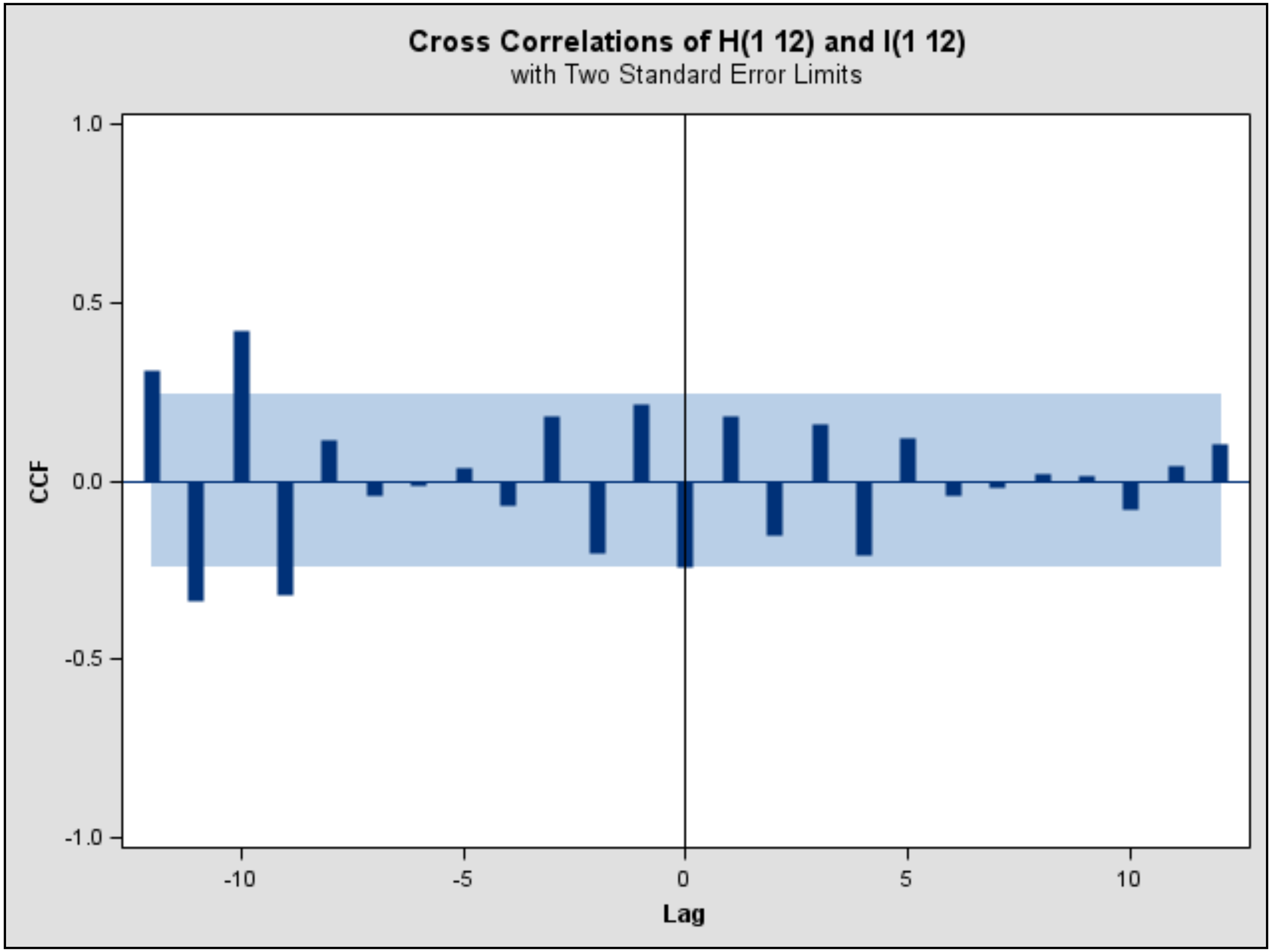
Three series (shifted and scaled):

- Housing Starts (H) for the U.S.
- Motor Vehicle Injuries (I) occurring in a large U.S. metropolitan area
- Lowest Tide Gauge Mark (L) for a San Francisco monitoring station

CCF before Removing Trend and Seasonality



CCF after Removing Trend and Seasonality



Takeaway

- When to consider CCF
  - Not sure if an input variable is a good predictor
  - Not sure about the appropriate transfer function for a regressor
- Be careful about spurious CCF
  - Two time series with trend (or seasonality) will usually appear to be correlated.
  - Trend and seasonal components should be removed before calculating the CCF.
- A more direct and probably better approach: adding the regressor into the model and see if the prediction performance improves

Further Readings

- Forecasting Chapter 5
- https://onlinecourses.science.psu.edu/stat510/node/72