12. Time series regression

Agoston Reguly

Data Analysis 2: Regression analysis

2020

Motivation

Heating and cooling are important uses of electricity. How does weather conditions affect electricity consumption? We are going to use monthly data on temperature and residential electricity consumption in Arizona. How to formulate a model which captures these factors?

Time series specialities

- We have *time series data* if we observe one unit across many time periods.
- There is a special notation:

$$y_t, \ t = 1, 2, \ldots, T$$

- Time series data presents additional opportunities as well as additional challenges to compare variables.
- Key issues:
  - Data wrangling: frequency and aggregation
  - Special nature of time series: stationarity or non-stationarity (random walk, trends, seasonality), SEs and serial correlation
  - Interpretation.

Data preparation

- Frequency of time series = time elapsed between two observations of a variable
  - Usual frequencies: yearly, quarterly, monthly, weekly, daily, hourly, minutes, seconds, ect.
- Practical problems with frequency:
  - There may be regular/irregular gaps between them: e.g. weekends for stock-exchange
    - If by nature: ignore them (there is no trading at weekends)
    - If it considered important: control for them (important news arrive at the weekends for trading)
  - Two variables have different frequencies
    - You have to harmonize them. (next slide)
- Extreme values (spikes) in your variable
  - Never drop them (destroys the time-series structure)
  - Use a dummy/step-indicator variable for/from that specific day.

Aggregation

- Regressions: to condition y_t on values of x_t the two variables need to be on the same frequency. When the frequency of y and x is different we need to adjust one of them.
  - Most often - aggregating the variable at higher frequency (e.g., weekly to monthly).
  - In really rare cases you interpolate your variable, which has lower frequency. This is similar to imputing. In general it is not advised to do it.
- Aggregation:
  - *Flow variables*: sum up the values within the interval. e.g. daily sales $\rightarrow$ weakly sales is the sum of daily sales.
  - *Stock variables*: take the end-period value. e.g. daily stock prices uses the closing price on a given day; or yearly GDP takes the last quarter's GDP value.
  - Other kinds of variables: usually take the average value (or other central tendency measure) e.g. daily prices to weakly price take the average.
    - You should think and check your variable: which value would be the most representative for that period? (e.g. variable with extreme value you may consider to use the median)

What is special in time series

- Many aspects of regression analysis remains, but time series regressions is special for several reasons.
  - Time series regression uncover patterns rather than evidence of causality.
  - Coefficient interpretation is based on conditional comparison
- Ordering matters – key difference to cross section
  - Trend - variables for later time periods will tend to be higher or lower than in the beginning.
  - Seasonality - cyclical component, such 4 seasons, months, days (e.g. shopping behaviour in December is expected to be different)
- Time series values are generally not independent from each other!

Stationarity

- *Stationarity*: a property of the time series itself. Key new concept.
- *Stationary time series* have the same conditional distribution (and same conditional expected value), at all times.

$$F(y_t|\cdot) = F(y_{t+\tau}|\cdot) \longrightarrow E\left[y_t|\cdot\right] = E\left[y_{t+\tau}|\cdot\right] = c, \quad \forall t \text{ and for any } \tau \in \mathbb{N}$$

- Stationarity means stability of the time series sampling distribution (usually we are interested in the expectation).
- Non-stationary time series are those that are not stable for some reason. Any of the equality does not hold.

Serial correlation - order and sign

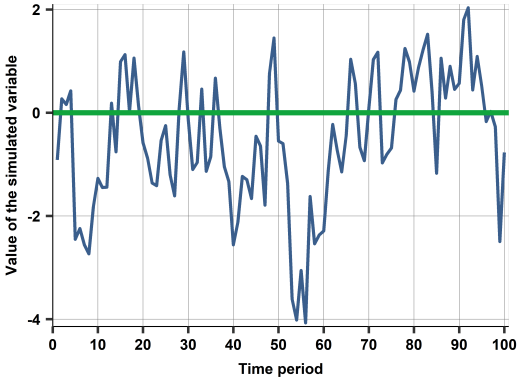
- Serial correlation means correlation of a variable with its previous values.
- The 1st order serial correlation coefficient is defined as

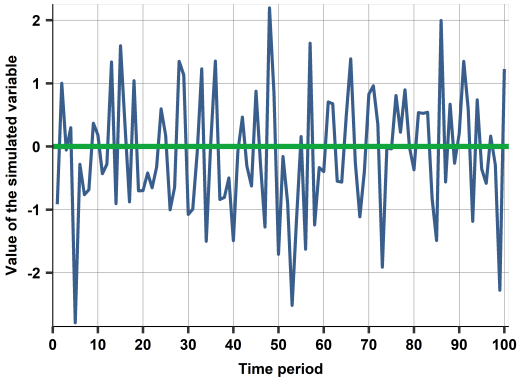
$$\rho_1 = Corr[y_t, y_{t-1}]$$

  - the 2nd order serial correlation coefficient is defined as $\rho_2 = Corr[y_t, y_{t-2}]$, ect.
- For a *positively serially correlated* variable, if its value was above average last time, it is more likely that it is above average this time, too.
- Special case: $\rho_1 = 0$ - no serial correlation, called "White Noise".
  - Meaning: like in cross-section, order does not matter.

Serial correlation - magnitude

Generated time series: $y_t = \rho y_{t-1} + \epsilon_t$ with $\rho_1 = 0.8$ (left) and $\rho_1 = 0$ (right). $\epsilon_t \sim N(0, 1)$





Non-stationarity - Random Walk

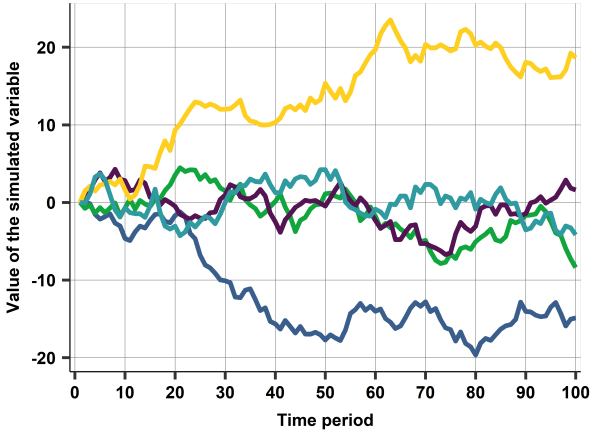
- Example of non-stationary time series is the *random walk*.
- Random walk when $\rho_1 = 1$ – also called a 'unit root'.
- Time series variables that follow random walk change in completely random ways.

$$y_t = y_{t-1} + \epsilon_t\,, \qquad \Delta y_t = y_t - y_{t-1} = \epsilon_t$$

- Whatever the previous change was the next one may be anything. Wherever it starts, a random walk variable may end up anywhere after a long time.
  - Random walks are impossible to predict
  - After a change, they don't revert back to some value or trend line but continue their journey from that point.
  - Spread rising from one interval to another

Random walk

- 5 simulated random walk path
- Each random walk series wanders around randomly.
- Further and further away as time passes



Unit root test

- Testing is complicated. (e.g. Dickey-Fuller sampling distribution)
- Phillips-Perron test is based on the following idea:

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t$$

- This model represents a random walk if $\rho = 1$ (with 'drift' if $\alpha \neq 0$)
- The Phillips-Perron test has hypothesis $H_0 : \rho = 1$ against the alternative $H_A : \rho < 1$.
- Statistical software calculate the p-value for this test.
- When the p-value is large (e.g., larger than 0.05), we don't reject the null, concluding that the time series variable follows a random walk (perhaps with drift).

Non-stationarity - (stochastic) Trend(s)

Change in variable (or the 'fist difference'): $\Delta y_t = y_t - y_{t-1}$

$$\text{Positive trend: } E[\Delta y_t] > 0$$

$$\text{Negative trend: } E[\Delta y_t] < 0$$

- A time series variable follows a *positive trend* if its change is positive on average. It follows a *negative trend* if its change is negative on average

$$\text{Linear trend: } E[\Delta x_t] = constant$$

$$\text{Exponential trend: } E[\Delta ln(x_t)] = constant$$

- Trend is *linear* if the change is the same on average and it is *exponential* if the change in the log of the variable is the same on average.

Non-stationarity - Seasonality

- There is seasonal variation, or simply *seasonality*, in a time series variable if its expected value changes periodically.
- Follows the seasons of the year, days of the week, hours of the day.
- Seasonality may be linear, when the seasonal differences are constant; it may be exponential, if relative differences (that may be approximated by log differences) are constant.
- Important real life phenomenon - many economic activities follow seasonal variation over the year, through the week or day.
  - Some variables are 'seasonally adjusted' which means this seasonal component is removed.
  - You may use year-in-year differences as your variable, but this is cumbersome to interpret.

Properties of time series - summary

- Stationary series are those where the expected value does not change, variance does not change over time: two observations at different points in time have the same mean and variance.
  - A series is 'stationary' if all time intervals are similar in this sense.
- We have seen three examples of non–stationarity:
  - Random walk and similar series – Variance keeps increasing over time, but there is no trend.
  - Trend - Expected value is different in later time periods than in earlier time periods
  - Seasonality - Expected value is different in periodically recurring time periods
- We care about this because regression with time series data variables that are not stationary are likely to give misleading results.

Practical implications

- Check if your variable is stationary
  - Visualize
  - Check serial correlation(s) (may use auto-correlation plot)
  - Do a unit-root test
- If there is a good reason to believe your variable trending (or RW)
  - Take differences Δy_t
  - Take percentage changes or log differences
  - In extremely rare cases you need to difference your variable twice, if your variable still has a unit root.
- If your variable has a seasonality
  - Use seasonality dummies in your regression
  - May consider to work with seasonal changes.

Microsoft and S&P 500 stock prices - data

- Daily price of Microsoft stock and value of S&P 500 stock market index
- The data covers 21 years starting with December 31 1997 and ending with December 31 2018.
- Many decisions to make...
- Look at data first

Microsoft and S&P 500 stock prices - ts plot

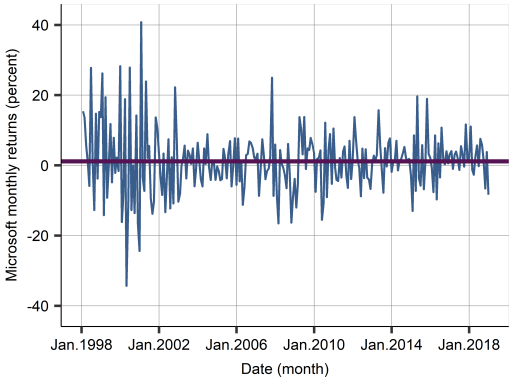
Microsoft, daily close price

S&P 500 index value, daily close

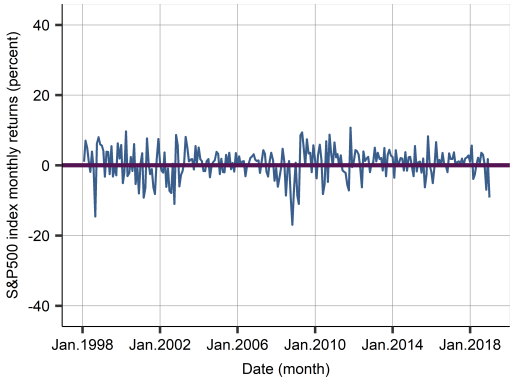
Microsoft and S&P 500 stock prices - decisions

- Frequency decisions:
  - Price: closing price
  - Gaps will be overlooked (neglect weekend)
  - Friday-Monday gap ignored (no special dummy)
  - Holidays: Christmas, 4 of July (when would be a weekday), ect. are ignored as well
- All values kept, extreme values part of process.
- Outcome: it is obviously non-stationary (Phillips-Perron test: very high p-value)
  - In finance, portfolio managers often focus on *monthly returns* - this is the time horizon for which performance are measured and communicated to clients. This means:
    - Monthly frequency: closing price for the last day of a month
    - Returns: percent change of the closing prices: $100\% \frac{y_t - y_{t-\mathbf{1}}}{y_t}$. (Alternative measure: first difference of log prices.)

Microsoft and S&P 500 - index returns (pct)

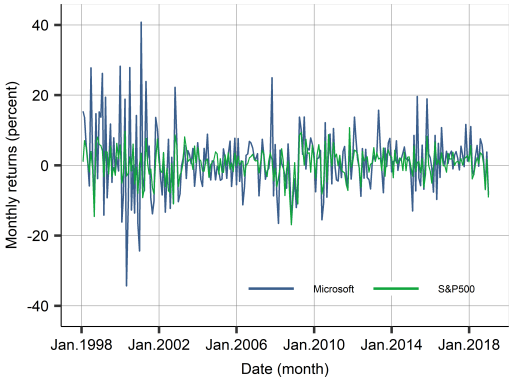


Microsoft, monthly return (pct)

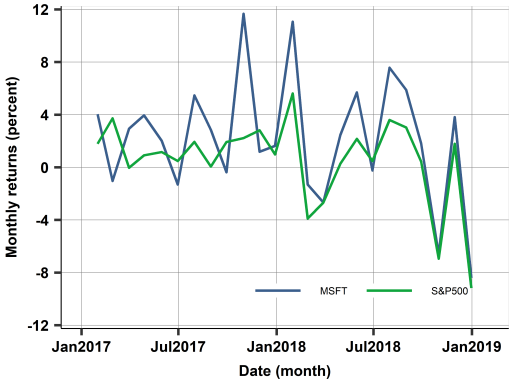


S&P 500 index value, monthly return (pct)

Microsoft and S&P 500 index returns - comparisons



The entire time series, 1998-2018



2017-18

Returns on a company stock and market returns

- Correlation in time series: the price of the Microsoft stock tends to increase when market prices increase, and it tends to decrease when market prices decrease.
- Market changes are smaller
- If focus on two years, we can see it better

- You can estimate the regression formally as well.
  - Read the book to get the argument about the co-movement of the returns.
  - Good for descriptive purposes, less for prediction.
  - Also note: it is still non-stationary because of changing variance...

Time series regressions

- Regression in time series data is defined and estimated the same way as in other data.

$$y_t^E = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + ...$$

- Interpretations similar to cross-section
  - β_0: We expect y to be β_0 when all explanatory variables are zero.
  - β_1: Comparing time periods with different x_1 but the same in terms of all other explanatory variables, we expect y to be higher by β_1 when x_1 is higher by one unit.

Issues to deal with

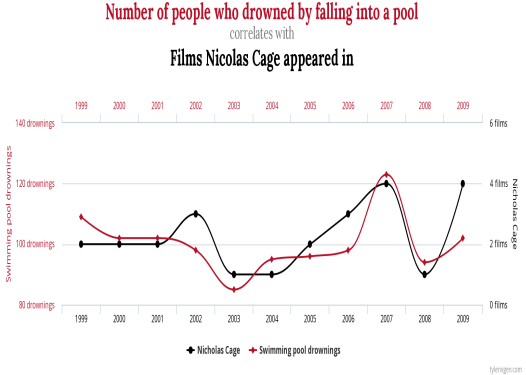
- Handling trend(s) and random walk (RW)
  - Transforming the series, such as taking first differences
- Handling seasonality and extreme values
  - Include dummies
- Standard errors for parameters
- Dealing with serial correlation
  - Considering lags and cummulative effects (in x_t)
  - May fix standard error problems if include lags of y_t

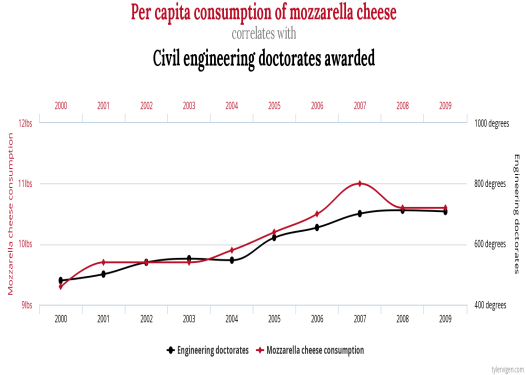
Trend & RW - Spurious regression

- Trends, seasonality, and random walks can present serious threats to uncovering meaningful patterns in time series data.
  - Example: time series regression in levels $y_t^E = \alpha + \beta x_t$.
  - If both y and x have a positive trend, the slope coefficient β will be positive whether the two variables are related or not.
  - That is simply because in later time periods both tend to have higher values than in earlier time periods.

- Associations between variables only because of the effect of trends are said to be spurious correlation.
  - trend and seasonality are confounders (omitted variables)
    - trend: global tendencies e.g. population growth, economic activity, fashion, technology, ect.
  - seasonality: e.g. weather, holidays, leisure time, sleeping and eating habits, open/close time of shops, ect.

Trend & RW - Seemingly correlated time series. But....





These and similar graphs from http://tylervigen.com/spurious-correlations

Trend & RW - solution: first differences

A linear regression in differences for both y and x is the following

$$\Delta y_t^E = \alpha + \beta \Delta x_t$$

- Coefficients same interpretation as before, but use *"when"* and *"change"*
- Because variables denote changes...
  - α is the average change in y when x doesn't change.
  - The slope coefficient on Δx_t shows how much *more* y is expected to change when x changes by one more unit.
  - "more" – needed as we expect y to change anyway by α, when x doesn't change.
    - The slope shows how y is expected to change when x changes, in addition to α.
  - If you have more explanatory variables: same interpretation but need to add that all other variables are remains the same/unchanged.

Trend & RW - side note

- It is not necessary that you take difference for both y and x variables.
  - Substantive: which transformation makes more sense for interpretation.
  - Statistical: avoid trends/unit roots in your variables. (stronger reason!)
- For most applications, time series regression involving using differences or log differences.
- Take differences unless you have a good reason not to:
  - One such case is when your variable is already a difference, GDP growth = difference of levels of GDP in percentage
- Deterministic trend: if you think there is a simple deterministic trend behind both variables you may consider to use levels and a simple time-trend variable: $y_t = \alpha + \beta x_t + \delta t$. However, this is rarely the case.
- Co-integration: if there is a common-stochastic trend behind both (or all) your variable. Special treatment, we not cover here.

Seasonality in time series regressions

- Capturing seasonality also crucial.
- Higher the frequency – the more important.
  - People behave differently on different hours and days
  - Weather varies over months
  - Holidays, quarters, ect.

- Have seasonal dummies if seasonality is stable.
  - Often good enough.
  - Reminder: skip one of them to avoide collinearity.

$$y_t = \alpha + \beta x_t + \delta_{Jan} + \delta_{Feb} + \cdots + \delta_{Nov}$$

  - Pattern may vary over time. If it does, solutions must capture exact pattern – difficult, not covering here.

Standard errors in time series regressions

- Serial correlation ($Corr[y_t, y_{t-1}] \neq 0$), makes the usual standard error estimates wrong.
  - When the dependent variable is serially correlated - classical heteroskedasticity robust SE is wrong - sometimes very wrong leading to false generalization.
  - More precisely it is serial correlation in residuals, but think about is as serial correlation in y_t is okay
- Use new SE - the Newey-West SE
  - procedure incorporates the structure of serial correlation of the regression residuals
  - Fine if heteroskedasticity as well
  - Need to specify lags. If enough data, frequency and seasonality should help, Months - 12 lags should be good.
- An alternative solution is to have lagged dependent variable in the regression

$$y_t = \alpha + \beta x_t + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} \cdots$$

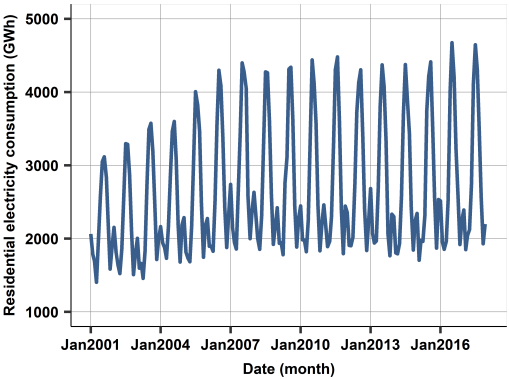
Electricity consumption and temperature

- Monthly weather and electricity data for Phoenix, Arizona
- January 2001 and ends in Dec 2017. Overall 204 month
- The weather data includes “cooling degree days" and “heating degree days" per month.
- Cooling degree days and heating degree days are daily temperatures transformed in a simple way and then added up within a month.

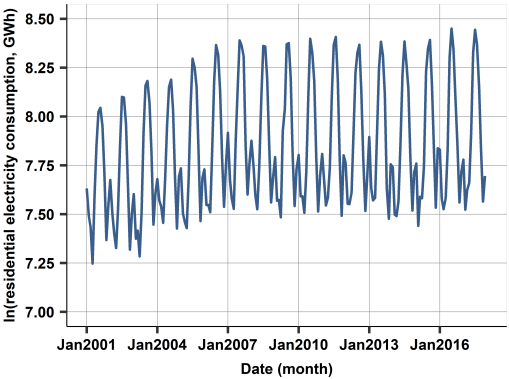
Electricity consumption and temperature - details

- The cooling degree days measure takes the average temperature within each day, subtracts a reference temperature (65F, or 18C), and adds up these daily values.
- If the average temperature in a day is, say, 75F (24C), the cooling degree is 10F (6C). This would be the value for one day.
- Then we would calculate the corresponding values for each of the days in the month and add them up.
  - Days when the average temperature is below 65F have zero values.
- For heating degree days it's the opposite: zero for days with 65F or warmer, and the difference between the daily average temperature and 65F when lower.
  - For example, with 45F (7C), the heating degree is 20F (11C).

(Log) Electricity consumption

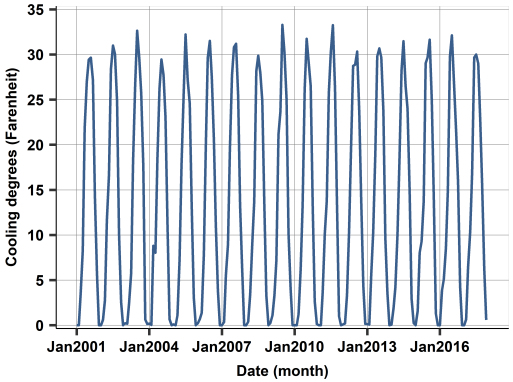


Electricity consumption

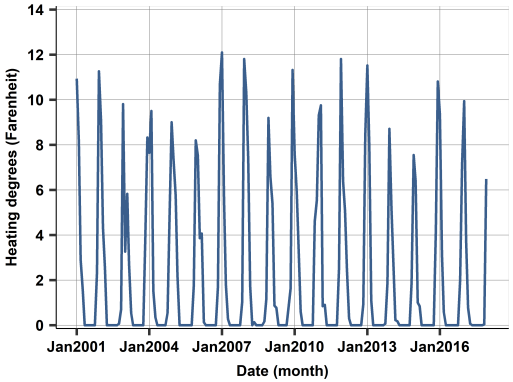


Log of electricity consumption

Average cooling/heating degrees



Average cooling degrees



Average heating degrees

Wrangling and modelling decisions

- There is an exponential trend in electricity –> use log difference
- For easier interpretation, take first difference (FD) of cooling days and heating days as well.
  - Natural question: How much does electricity consumption change when temperature changes?
- In this example, taking first difference does not make a huge difference, would not be a (big) mistake to keep in levels
  - Another option could be to take 12-month difference
- Add monthly dummies, January (December to January) is reference
- Use Newey-West standard errors in parentheses; ** $p < 0.01$, * $p < 0.05$

Model estimates

VARIABLES	(1) ΔlnQ	(2) ΔlnQ
ΔCD	0.031**	0.017**
	(0.001)	(0.002)
ΔHD	0.037**	0.014**
	(0.003)	(0.003)
month = 2, February		-0.274**
month = 3, March		-0.122**
....		
month = 7, July		0.058**
month = 8, August		-0.085**
month = 9, September		-0.176**
....		
month = 12, December		0.067**
Constant	0.001	0.092**
	(0.002)	(0.013)

Model results

- Simple (wrong) model:
  - In months when cooling degrees increase by one degree and heating degrees do not change, electricity consumption increases by 3.1 percent, on average.
  - When heating degrees increase by one degree and cooling degrees do not change, electricity consumption increases by 3.7 percent, on average.
- BUT, monthly dummies matter (confounder), reduce slope coefficient estimates.
  - The reference month is January: constant (when cooling and heating degrees stay the same), electricity consumption increases by about 9% from December to January.
  - The other season coefficients compare to this change:
    - February: the January to February change is 28 percentage points lower than in the reference month, December to January.
    - That was +9%, so electricity consumption decreases by about 19% on average to February from January when cooling and heating degrees stay the same.

Propagation effect: changes and lags

- Potential causal scenario where changes take an impact in several periods later:

$$\Delta y_t^E = \alpha + \beta_0 \Delta x_t + \beta_1 \Delta x_{t-1} + \beta_2 \Delta x_{t-2}$$

- Coefficients – how y is expected to change after a one-time change in x, i.e., when x changes in one time period *but not afterwards*.
- β_0 shows the contemporaneous association: what to expect in the same time period.
- β_1 shows the once-lagged association: what to expect in the next time period.

Interpretation of lags and cumulative effect

$$\Delta y_t^E = \alpha + \beta_0 \Delta x_t + \beta_1 \Delta x_{t-1} + \beta_2 \Delta x_{t-2}$$

- β_0 = how many units more y is expected to change within the same time period when x changes by one more unit (and it didn't change in the previous two time periods).
- β_1 = how much more y is expected to change *in the next time period* after x changed by one more unit – provided that it didn't change at other times.

- Cumulative effect:

$$\beta_{cumul} = \beta_0 + \beta_1 + \beta_2$$

Testing the cumulative effect

- To get a SE on the cumulative effect, do a trick and transformation, and estimate a different model

$$\Delta y_t^E = \alpha + \beta_{cumul}\Delta x_{t-2} + \delta_0\Delta(\Delta x_t) + \delta_1\Delta(\Delta x_{t-1})$$

- the β_{cumul} in this regression is exactly the same as $\beta_0 + \beta_1 + \beta_2$ in the previous regression.
  - Other two right-hand-side variables strange and we do not care, but needed.
- Typically estimate both. One with lags to see patterns. One with cumulative second to test the cumulative value.

Lag selection

- Lag selection is a practical question
- Think about theory, domain knowledge. This may drive your call.
- Try out a few lags. Few depends on the size of your dataset.
  - Few dozen observations - need to be picky (2-4)
  - 10-20 years of monthly data, can try all months
- Watch for seasonality. Often need lags to capture 12 months, 4 quarters, etc.
- Try a few versions. Choose based on coefficient significance.

Electricity consumption and temperature - use lags

- Go back to model
- Add 2 lags - for both cooling and heating days
- And keep monthly dummies

Model summary with lags

VARIABLES	(1) $\Delta \ln Q$	VARIABLES	(2) $\Delta \ln Q$
ΔCD	0.020**	ΔCD cumulative coeff	0.027**
	(0.002)		(0.005)
ΔCD 1st lag	0.006**		
	(0.002)		
ΔCD 2nd lag	0.001		
	(0.002)		
ΔHD	0.019**	ΔHD cumulative coeff	0.030**
	(0.003)		(0.007)
ΔHD 1st lag	0.011**		
	(0.003)		
ΔHD 2nd lag	0.000		
	(0.003)		
Observations	201		201
R-squared	0.957		0.957
Month binary variables	Yes		Yes

Standard errors in parentheses
** p<0.01, * p<0.05

Thinking about the results

- Cumulative effect is the same as the sum of the lags.
- Interestingly evidence of lagged effect: there is propagation effect.
  - Not straightforward to answer why we see this pattern.
    - People take time to react to weather change
    - Or captures some correlated other variable
- Overall: Temperature is strongly associated with residential electricity consumption in Arizona, even when seasonality is captured

Main lessons learnt

- Temperature explains a large part of electricity consumption, i.e. hotter than average summers and cooler than average winters lead to substantially higher electricity consumption.
  - Months matter on their own right as well.
- We had to deal with the strong seasonality in both electricity consumption and temperature.
  - We included month binary variables, and the estimated coefficients became smaller (about half the original for cooling degree days, and about one third the original value for heating degree days)
- If there is serial correlation in the dependent variable, we need to adjust standard error estimation.
  - Most general solution is to use Newey-West standard errors.

Summary of the process

- Decide on frequency; deal with gaps if necessary.
- Plot the series. Identify features and issues.
- Handle trends by transforming variables (Often: first difference).
- Specify regression that handles seasonality, usually by including season dummies.
- Include or don't include lags of the right-hand-side variable(s).
- Handle serial correlation.
- Interpret coefficients in a way that pays attention to potential trend and seasonality.

- Time series econometrics very complicated beyond this.
- But: These steps often good enough.