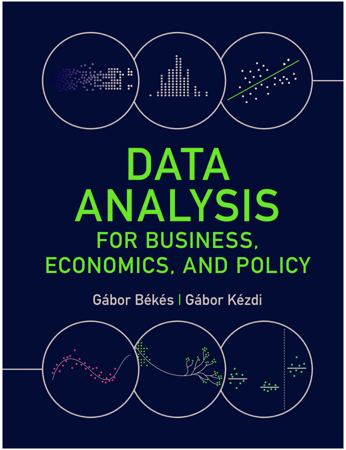
6 Time series regressions

Alice Kügler

Data Analysis 2 – **MS Business Analytics**: Regression Analysis

2023

Slides for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- **gabors-data-analysis.com**
  - Download all data and code: gabors-data-analysis.com/data-and-code/
- These slides are for Chapter 12

Motivation

You are considering investing in a company stock, and you want to know how risky that investment is. You have downloaded data on daily stock prices for many years. How should you define returns? How should you assess whether and to what extent returns on the company stock move together with market returns?

Heating and cooling are important uses of electricity. How do 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?

What is special in the analysis of time series data?

Time series specialties

- 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.

- Data wrangling novelties: frequency and aggregation
- Special nature of time series: serial correlation
- Coefficient interpretation

Data preparation

- Frequency of time series = time elapsed between two observations of a variable
- Practical problems with frequency:
  - There may be regular/irregular gaps between them: e.g. weekends for stock-exchange
  - Two variables have different frequencies
- Extreme values (spikes) in your variable

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.
- Aggregation:
  - *Flow variables*: sum up the values within the interval. e.g. daily sales $\rightarrow$ weekly 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
  - Other kinds of variables: usually take the average value

What is not special in time series

- Time series regressions is special for several reasons.
- Many aspects of regression analysis remain
  - Generalization, confidence intervals
  - Time series regression uncover patterns rather than evidence of causality
  - Practical data issues, missing observations, extreme values etc, remain
  - Coefficient interpretation is based on conditional comparison

What is special in time series

- Ordering matters – key difference to cross section.

- Trend - variables for later time periods will tend to be higher (lower).
- Seasonality - cyclical component, such 4 seasons, months, - every e.g. December value is expected to be different.
- Time series values are often not independent - correlated in time.

What is special in time series: trend

Define change (or fist difference): $\Delta y_t = y_t - y_{t-1}$

$$\text{Positive trend: } E[\Delta y_t] > 0 \tag{1}$$

$$\text{Negative trend: } E[\Delta y_t] < 0 \tag{2}$$

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

$$\text{Linear trend: } E[\Delta y_t] = constant \tag{3}$$

$$\text{Exponential trend: } E[\Delta ln(y_t)] = constant \tag{4}$$

What is special in time series: 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.

What is special in time series: stationarity

- **Stationary time series** have the same expected value and same distribution, at all times.
- Stationarity is a feature of the time series itself.
- Stationarity means stability (in expectations).

What is special in time series: non-stationarity

- **Non-stationary** time series are those that are not stable for some reason.
- Series that violate stationarity because the expected value is different at different times:
  - Has a trends
  - Has seasonality
  - Has some unstable patterns

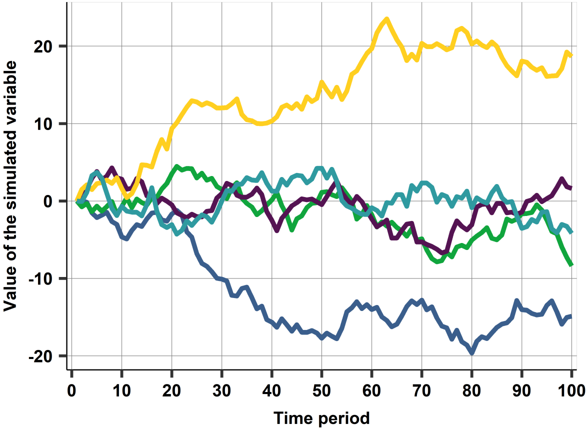
What is special in time series: non-stationarity

- Another example of non-stationary time series is the *random walk*.
- Random walk when y_t follows a random walk if its value in t is the same as in $(t-1)$ plus some totally random term: $y_t = y_{t-1} + e_t$.

- Time series variables that follow random walk change in completely random ways.
- 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.

What is special in time series: random walk

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



What is special in time series: random walk

- 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

- For stationary series, we need stability of patterns
- Avoid series with random walk when running regressions

What is special in time series: unit root

- We can test if a series is a random walk.
- Phillips-Perron test is based on this mode:

$$y_t = \alpha + \rho y_{t-1} + e_t \tag{5}$$

- This model represents a random walk if $\rho = 1$. This is called a unit root
  - Random walk test = testing if the series has a unit root.
- The Phillips-Perron test has hypothesis $H_0 : rho = 1$ against the alternative $H_A : rho < 1$.
- ie: we would like to *reject* the null of a unit, ie show a series is not a random walk
- Use 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
- More on unit root test: 12.U1
  - Many versions of unit root test.

What is special in 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:
  - Trend - expected value is different in later time periods than in earlier time periods
  - Seasonality - expected value is different in periodically recurring time periods
  - Random walk and similar series – variance keeps increasing over time
- 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
  - 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

Case study A: stocks

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 take ...
- Look at the data first

Case study: stock price and stock market index value

Microsoft, daily close price

S&P 500 index value, daily close

Time series comparisons - S&P 500 case study

- 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

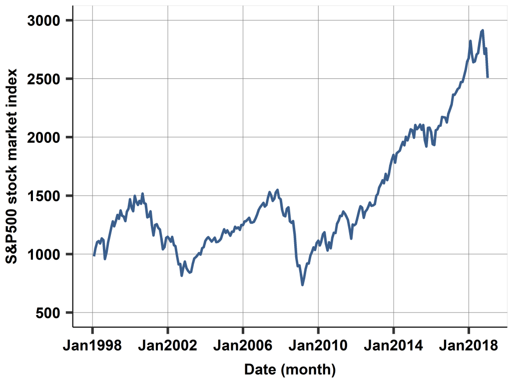
- Key decisions:
- Daily price = closing price
- Gaps will be overlooked
  - Friday-Monday gap ignored
  - Holidays (Christmas, 4 of July (when it would be a weekday)
- All values kept, extreme values are part of the process

Time series comparisons - S&P 500 case study

- In finance, portfolio managers often focus on monthly returns - this is the time horizon for which performance is measured and communicated to clients
- Hence, we choose monthly returns to analyze
- Take the last day of each month

Microsoft and S&P 500 stock prices - time series plot

Microsoft, daily close price

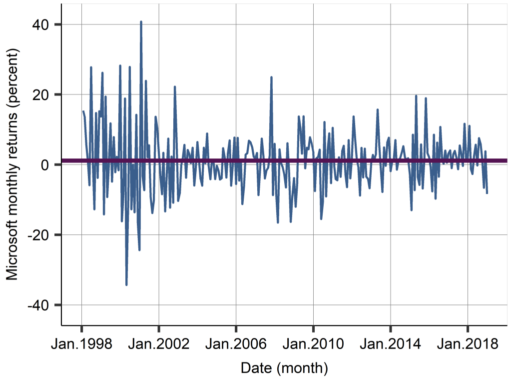


S&P 500 index value, daily close

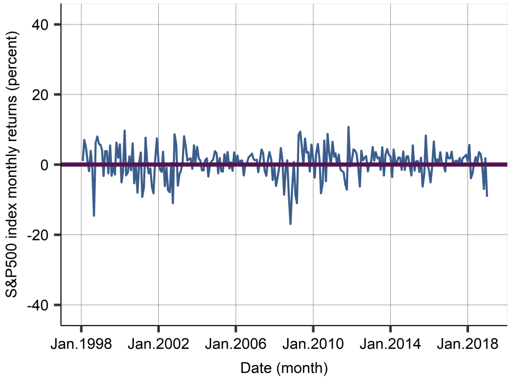
Microsoft and S&P 500 stock prices - decisions II

- Monthly time series plot - easier to read
- Additional decision needed: the series are obviously non-stationary (Phillips-Perron test: very high p-value)
- Use returns - take difference :
- Returns: percent change of the closing prices: $100\% \frac{y_t - y_{t-1}}{y_t}$.
  - *monthly returns* - take the closing price for the last day of a month
    - 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)

Descriptive statistics on monthly returns

Monthly percentage returns on Microsoft stock and the S&P 500 index.

Variables	Min	Max	Mean	Std.dev.	N
Monthly returns on Microsoft (%)	-34.4	40.8	1.1	9.1	252
Monthly returns on the S&P500 (%)	-16.9	10.8	0.5	4.3	252

Source: stocks-sp500 dataset. December 31, 1997 to December 31, 2018, monthly frequency, N=252.

Regression with time series

Time series regressions

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

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

- Interpretations are 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.

Time series regression

- With time series data, we often estimate regressions in changes
- We use the Δ notation for changes

$$\Delta x_t = x_t - x_{t-1}$$

- The regression in changes is

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

  - α: y is expected to change by α when x does not change
  - β: y is expected to change by β more when x increases by one unit more

Time series regression

- We often have variables in relative or percentage changes,

$$pctchange(y_t)^E = \alpha + \beta pctchange(x_t)$$

- We can approximate relative differences by log differences, which are here log change: first taking logs of the variables and then taking the first difference
- $\Delta \ln(y_t) = \ln(y_t) - \ln(y_{t-1})$

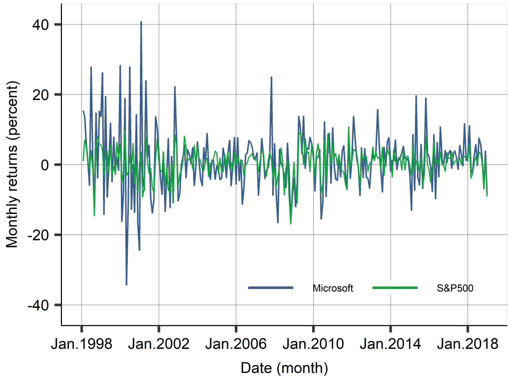
Returns on a company stock and market returns

- 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
- Focus on two years, so we can see the correlation better

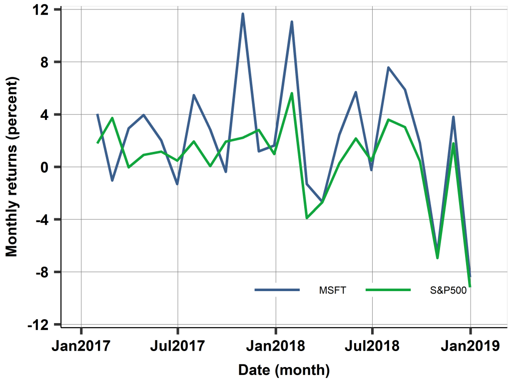
- We can estimate the regression formally
  - Monthly
  - Percent return

$$pctchange(MSFT_t)^E = \alpha + \beta pctchange(SP500_t) \tag{6}$$

Microsoft (in blue) and S&P 500 index (in green) returns - comparisons



The entire time series, 1998-2018



2017-18

Returns on a company stock and market returns

$$pctchange(MSFT_t) = \alpha + \beta pctchange(SP500_t) \tag{7}$$

- $\alpha = 0.54; \beta = 1.26$
- Intercept: returns on the Microsoft stock tend to be 0.54 percent when the S%P 500 index does not change.
- Slope: returns on the Microsoft stock tend to be 1.26% higher when the returns on the S&P 500 index are 1% higher. The 95% confidence interval is $[1.06, 1.46]$.
  - R-squared: 0.36
- First difference of log prices: estimate is 1.24
- Daily returns (percent): beta is 1.10

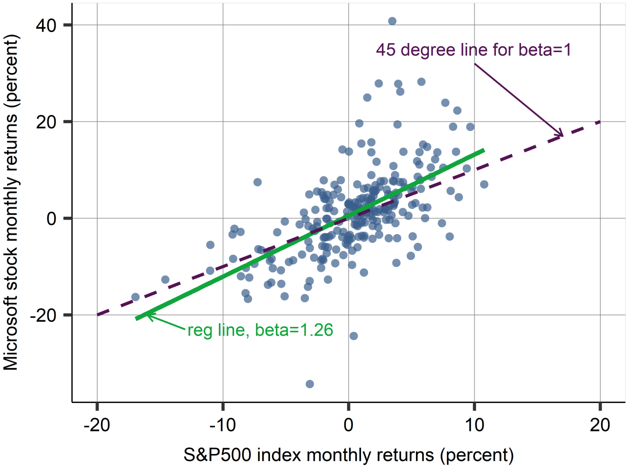
Returns on Microsoft and market returns: simple regression results

Dependent variable: monthly percentage returns on the Microsoft stock; explanatory variable: monthly percentage returns on the S&P500 index. December 31, 1997 to December 31, 2018, monthly frequency.

	(1)
Variables	**Microsoft returns**
S&P500 returns	1.26**
	(0.10)
Constant	0.54
	(0.45)
Observations	252
R-squared	0.36

Robust standard errors in parentheses
** p<0.01, * p<0.05

Returns on Microsoft and market returns: scatterplot and regression line



Issues to deal with: a laundry list

- Handling trend(s) and random walk (RW)
  - Transforming the series, such as taking first differences or percent change
- Handling seasonality and special events
  - Include dummies
- Reconsidering standard errors
- Dealing with serial correlation - taking time-to-build into account with lags

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.
    - In later time periods both tend to have higher values than in earlier time periods.

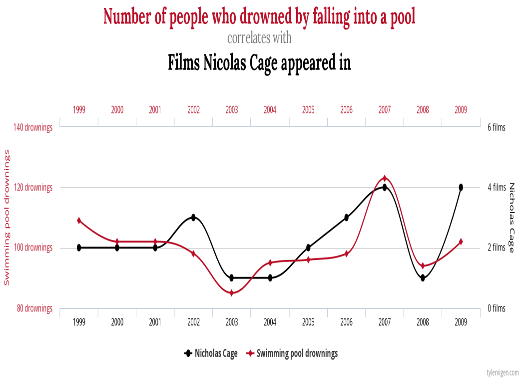
Trend & RW - spurious regression

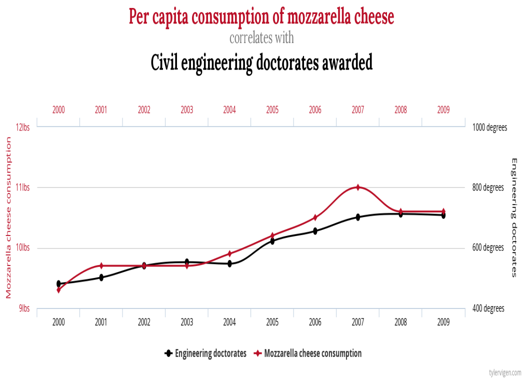
- Trends, seasonality, and random walks can present serious threats to uncovering meaningful patterns in time series data.
    - Example: time series regression in levels $y_t^E = \alpha + \beta x_t$.
    - If both y and x have a positive trend, the slope coefficient β will be positive whether the two variables are related or not.
        - 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 called **spurious correlation**.
    - Think of trend and seasonality as confounders (omitted variables)
        - Trend: global tendencies e.g. economic activity, fashion
        - Seasonality: e.g. weather, holidays, human habits (sleep)
- Another reason for spurious correlation is small sample size

Time series regressions: trends and seasonality

- Trend as confounder example
- A regression of the price of college education in the U.S. on the GDP of Germany over the past few decades
- Positive slope coefficient even though the two may not be related in any fundamental way.
- But U.S. GDP is correlated with both

Correlated time series. But ...





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

Time series regressions: trends and seasonality

- In a regression, we shall deal with trends
- Replacing variables in the regression with their first differences
  - Variables in differences – no trends – likely to to be stationary
  - Could be log difference for exponential trends

- In a regression, we shall deal with seasonality
- Including binary season variables in regressions
  - Look at pattern, figure out if quarters, months, weeks, days of week, etc.
  - Or work with year-on-year changes instead of first differences

Trend & RW - solution: first differences

We use the Δ notation to denote a first difference:

$$\Delta y_t = y_t - y_{t-1} \tag{8}$$

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.

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, ect
- Have seasonal dummies if seasonality is stable

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

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

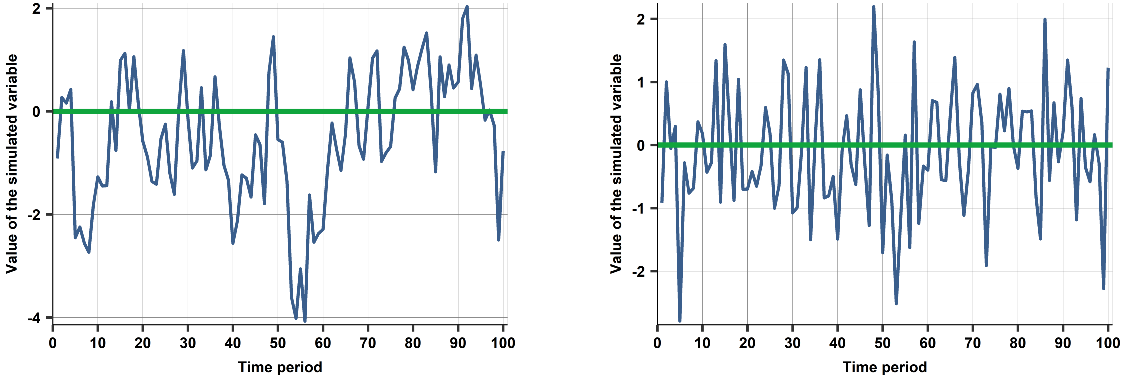
What is special in time series: serial correlation

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

$$\rho_1 = Corr[x_t, x_{t-1}] \tag{9}$$

  - the 2nd order serial correlation coefficient is defined as $\rho_2 = Corr[x_t, x_{t-2}]$;
- 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.
- $\rho_1 = 0$ - no serial correlation. "White Noise"
  - Like cross-section, order does not matter.

Two simulated series: rho=0.8 (left), rho=0 (right)



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 thinking about is as serial correlation in y_t is okay.

Standard errors in time series regressions

- In most time series, there will be some serial correlation
- Two solutions:
- 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.
- Have lagged dependent variable in the regression (one lag is usually enough)

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

  - Coefficients will also change
- Either one is fine.

Case study B: electricity consumption and temperature

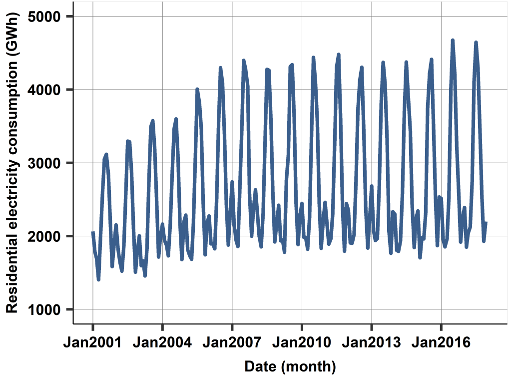
Electricity consumption and temperature

- Monthly weather and electricity data for Phoenix, Arizona
- January 2001 and ends in Dec 2017. Overall 204 months
- 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.

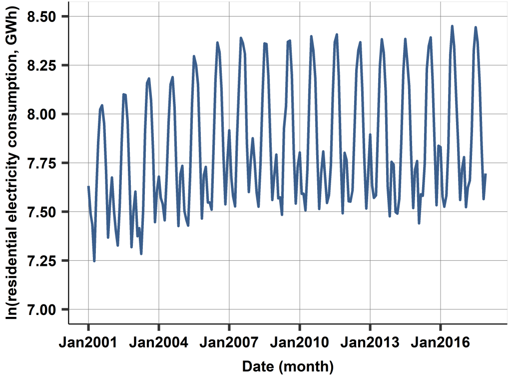
Electricity consumption and temperature - details

- 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 is 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

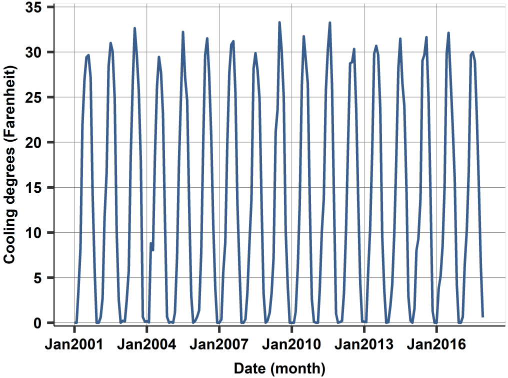


Electricity consumption

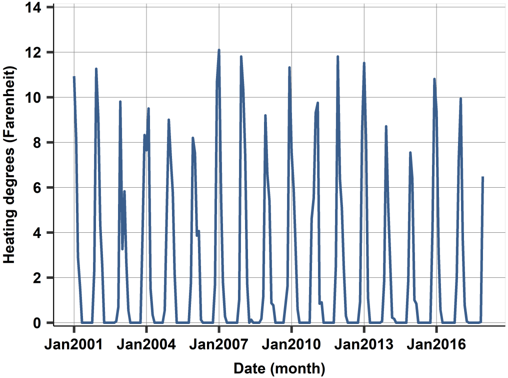


Log of electricity consumption

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, and it would not be a (big) mistake to keep in levels
  - Another option could be to take 12-months difference
- Add monthly dummies, January (December to January) as 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)
Observations	203	203

Model results

- Simple (1) 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.
- Model (2) with monthly dummies.
  - 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.

Electricity consumption and temperature – different SE estimates

	(1)	(2)	(3)
	Simple SE	Newey–West SE	Lagged dep.var
Variables	$\Delta \ln Q$	$\Delta \ln Q$	$\Delta \ln Q$
ΔCD	0.017**	0.017**	0.017**
	(0.002)	(0.002)	(0.002)
ΔHD	0.014**	0.014**	0.014**
	(0.002)	(0.003)	(0.002)
Lag of $\Delta \ln Q$			-0.002
			(0.062)
Month dummies	YES	YES	YES
Observations	203	203	202
R-squared	0.951		0.951

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

Electricity consumption and temperature – different SE estimates

- To correct for serial correlation, compare simple SE model with two correctly specified models
- with Newey-West SE
- with lagged dependent variable

- SE marginally different, and with lagged values, coefficients are also similar up to 3 digits
  - Similar, not the same
  - Note: sometimes substantial difference (if strong serial correlation)

Propagation effect: changes and lags

Changes take an impact in several periods later (time-to-build):

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

- 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.
- Often need a few, not many lags

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 a 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.

Findings

- How would you summarize our findings?

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 do not 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 is very complicated beyond this.
- But: These steps often good enough.

Main takeaways

- Regressions with time series data allow for additional opportunities, but they pose additional challenges, too
  - Regressions with time series data help uncover associations from changes and associations across time
  - Trend, seasonality, and random walk-like non-stationarity are additional challenges
  - Do not regress variables that have a trend or seasonality; without dealing with them they produce spurious results