Computational Social Science Masters

Introduction to Time Series

Graham Elliott

September 6, 2022

What Do We Mean By Time Series?

In many problems, rather than having observations over many units at a point in time (cross section) or a few points in time (longitudinal or panel data), we might have observations for a few units over many periods of time.

Examples
(a) Inventories - a company has a long history of sales and inventories of their own product
(b) Macroeconomic Data
(c) Stock market data
(d) Climate data

Why is Time Series a Thing?

When we have data ordered by time, the problems of analyzing the data and opportunities that the structure provides are different from cross sectional data.

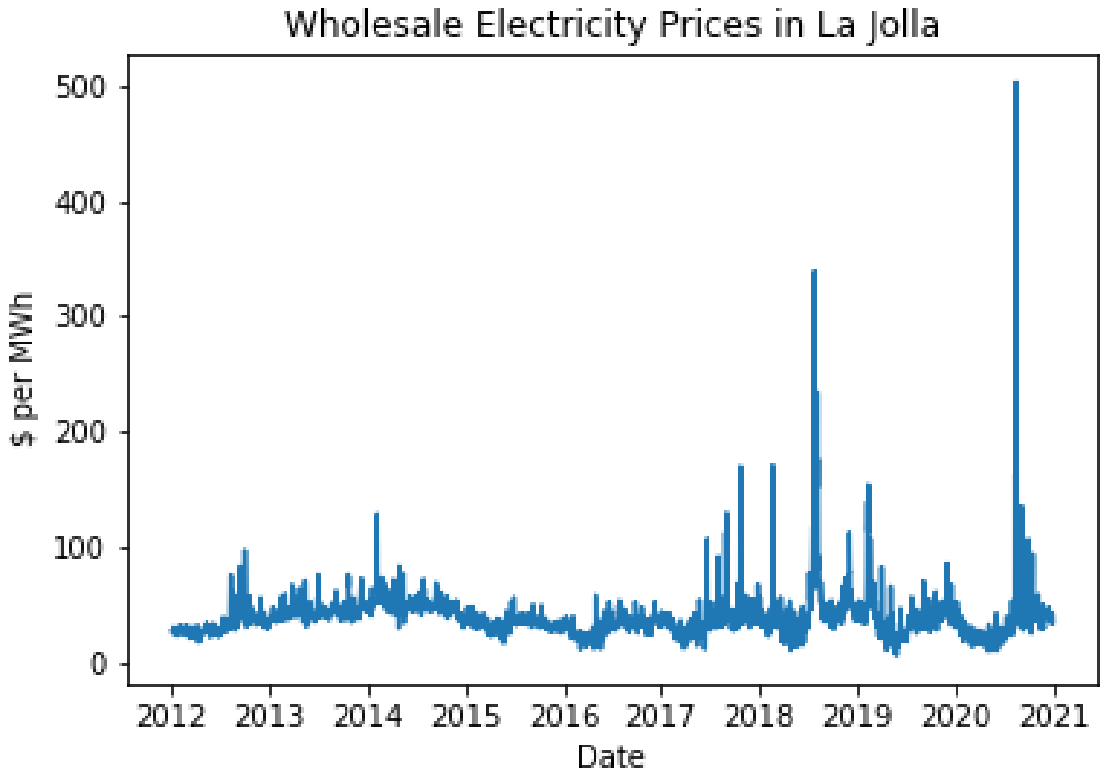
The main problem is that our common assumptions of independence between observations is generally untrue, need to adapt methods to account for the dependence in the data.

The opportunities arise from the additional structure of the data, that suggests modelling opportunities.

What Does Time Series Data Look Like?

We might write it as $\{y_t, x_t\}_{t=1}^{T}$

What Does Time Series Data Look Like?



What Can We Do with a Single Series

If we only observe a single series of data, can we do anything with it?

Possibilities

(a) Compute the average and/or the variance

(b) Model the dynamics

(c) Build forecasting models

Computing the mean

This you can already do, it is the inference part that is harder.

The sample average or sample mean is still

$$\bar{y}_T = \sum_{t=1}^{T} y_t.$$

But what happens to the estimate of the variance, which we need for t statistics or confidence intervals?

Computing the mean

You might recall formulas such as

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y).$$

Let T=2 and we have

$$\bar{y}_2 = \frac{1}{2} y_1 + \frac{1}{2} y_2.$$

Consider these as random variables, we have

$$Var(\bar{y}_2) = \frac{1}{4} Var(y_1) + \frac{1}{4} Var(y_2) + \frac{1}{2} Cov(y_1, y_2).$$

When there is temporal dependence we need to consider the covariance part.

Computing the mean

A standard approach is to use a 'HAC' estimator (also called robust)

$$VarEst = \sum_{j=-(T-1)}^{T-1} w(j)\hat{\gamma}(j) = \hat{\omega}^2$$

where here we have

$$\hat{\gamma}(j) = \frac{1}{T-1} \sum_{t=j+1}^{T} y_t y_{t-j}$$

Notice that if there is no dependence then this is close to our usual variance estimator. (assumes mean zero y's).

Computing the mean

We can now compute t tests and confidence intervals as usual using $\hat{\omega}$ in place of our usual estimator of the square root of the variance of the data. t-statistics are now

$$t = \sqrt{T}\left(\frac{\bar{y}_T - \mu_0}{\hat{\omega}}\right)$$

Confidence intervals are now

$$\bar{y}_T \pm cv_{\alpha/2}\hat{\omega}/\sqrt{T}$$

Computing the mean

There are two important things to think about here

1. Does the mean actually mean anything here?

2. Are the covariances changing over time.

Computing the mean

Does the mean actually mean anything?
A lot of data, especially in macroeconomics but true more generally, exhibit trends in their time series. For example

(i) GDP per capita tends to grow over time

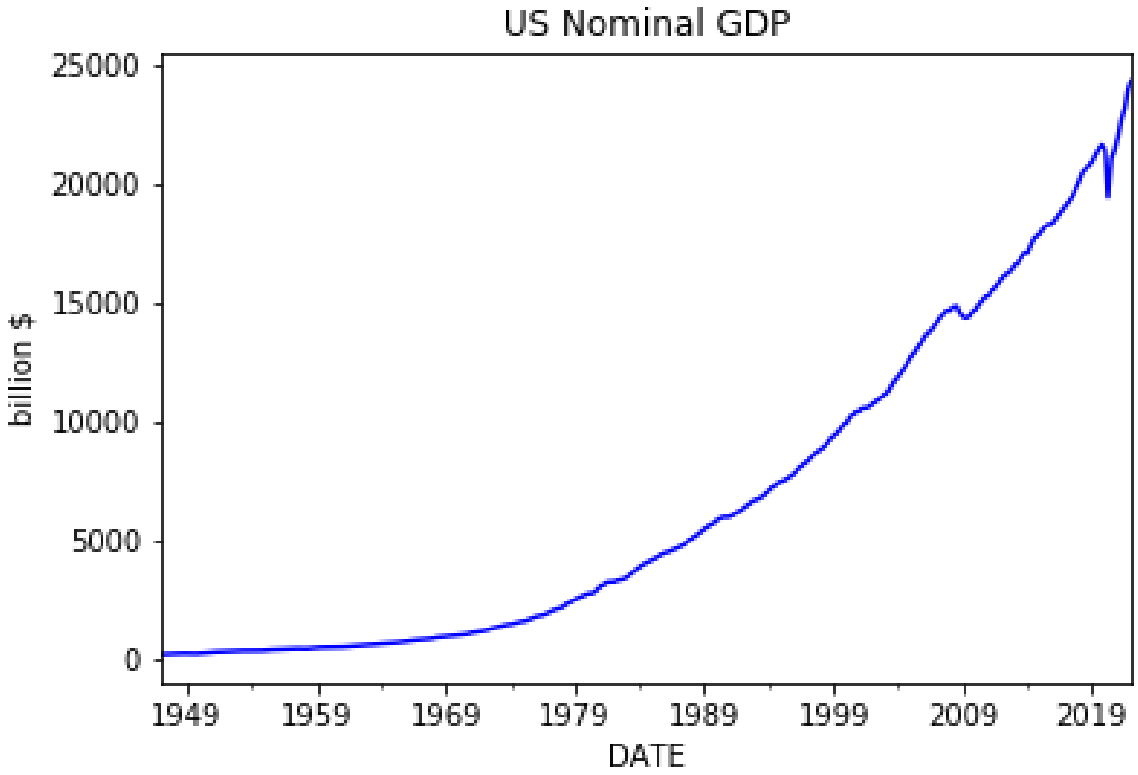
(ii) Stock market prices rise over time

(iii) Interest rates grew rapidly before the end of the 70's and have since generally followed a downward trend

(iv) Crime rates first rose post war through to the 90's, and have steadily fallen since then.

Computing the mean

Does the mean actually mean anything?



Computing the mean

Nonstationarity in the mean.

We could write a model for the data

$$y_t = \mu_t + u_t$$

Here the mean μ_t is explicitly changing, we are acknowledging that it is not a constant. So there is nothing constant to estimate.

All is not lost here - we can model the changes in the mean.

Computing the mean

A word of warning about nonstationarity in the mean.

Many equate nonstationarity with 'unit roots' in the data. They test for the null of a unit root, and A unit root model is one of MANY ways in

which the data can be nonstationary.

Computing the mean

What about our electricity data?

Datetime	Mean	Std
2012-12-31	33.027329	0.510849
2013-12-31	44.825278	0.385564
2014-12-31	50.354044	0.494663
2015-12-31	34.287818	0.304023
2016-12-31	31.068251	0.433761
2017-12-31	37.421917	0.807522
2018-12-31	48.832337	1.731254
2019-12-31	39.444670	0.932361
2020-12-31	36.737415	1.802023

Computing the mean

Are the covariances changing over time?

We often think that stock returns have time varying risk, this is essentially time variation in the variance.

If this does not change 'too much', our calculations are fine.

But sometimes it does change a lot, leading to poor inference. And in other situations we might actually want to understand this time variation.

Computing the mean

We could now write the model

$$y_t = \mu + h_t u_t$$

Here the variance of y_t is changing because of h_t

As with the mean, we could consider modelling h_t directly.

Popular choices of modelling approaches are (a) Stochastic Volatility models and (b) GARCH models and their variants.

Model the Dynamics

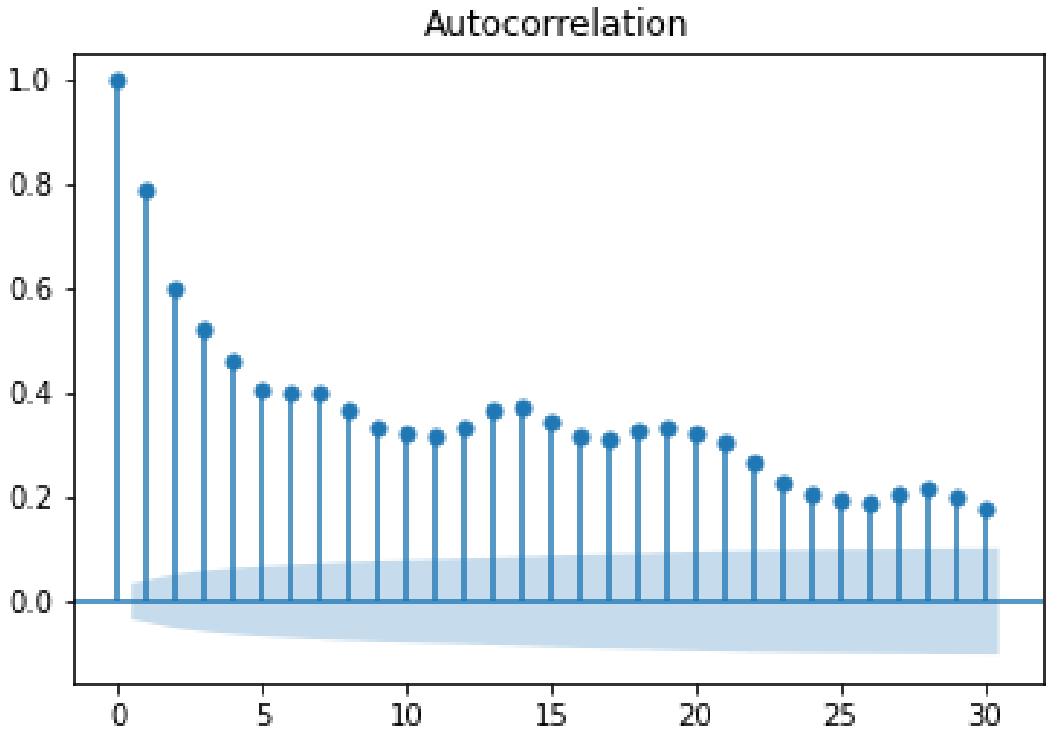
We might be interested in looking at the dynamic structure of y_t.

An old fashioned approach is to simply look at the autocovariances $\hat{\gamma}_t$

The Autocorrelation function is these autocovariances normalized by the variance.

Model the Dynamics

Consider the energy price data we saw before. The ACF is



Model the Dynamics

More commonly though we might model the dynamics in a more interpretable way.

An autoregressive model is one where past values of y_t are related to the current value in the following linear model.

$$y_t = \mu + \rho_1 y_{t-1} + ... + \rho_p y_{t-p} + \epsilon_t$$

We can run this as an OLS regression.

Model the Dynamics

The AR model with 8 lags is

```
from statsmodels.tsa.ar_model import AutoReg, ar_select_order

mod = AutoReg(bd.LMP, 8, old_names=False)
res = mod.fit()
print(res.summary())
```

```
                            AutoReg Model Results
==============================================================================
Dep. Variable:                    LMP   No. Observations:                 3288
Model:                     AutoReg(8)   Log Likelihood              -12763.524
Method:               Conditional MLE   S.D. of innovations             11.851
Date:                Tue, 21 Jun 2022   AIC                              4.951
Time:                        23:50:47   BIC                              4.969
Sample:                    01-09-2012   HQIC                             4.958
                         - 12-31-2020
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.8455      0.567     10.315      0.000       4.735       6.956
LMP.L1         0.8397      0.017     48.096      0.000       0.806       0.874
LMP.L2        -0.2116      0.023     -9.292      0.000      -0.256      -0.167
LMP.L3         0.1579      0.023      6.854      0.000       0.113       0.203
LMP.L4        -0.0126      0.023     -0.545      0.586      -0.058       0.033
LMP.L5        -0.0423      0.023     -1.824      0.068      -0.088       0.003
LMP.L6         0.0748      0.023      3.247      0.001       0.030       0.120
LMP.L7         0.0604      0.023      2.650      0.008       0.016       0.105
LMP.L8        -0.0138      0.017     -0.790      0.430      -0.048       0.020
```

Model the Dynamics

A more interpretable model is a moving average model.

An MA model is one where past values of ϵ_t are related to the current value in the following linear model.

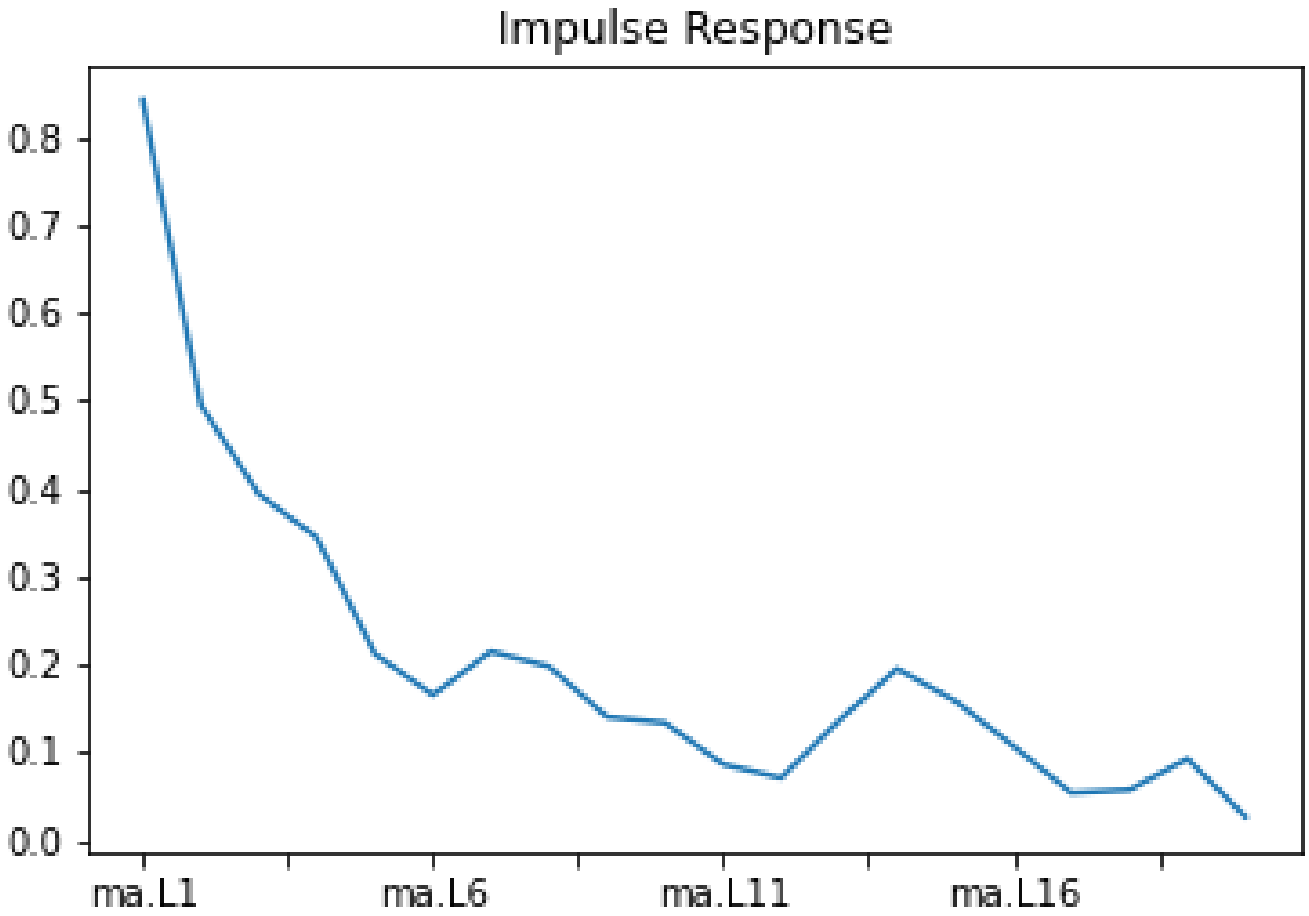
$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_p \epsilon_{t-p}$$

We cannot run this as an OLS regression.

What is nice about this is that the 'θ's' can be thought of as 'impulse responses' of the dynamic effect of a shock to y_t.

Model the Dynamics

The impulse response from an MA model is



Forecasting

We might be interested in exploiting the dynamic structure of y_t to obtain forecasts

I will leave this to the forecasting module, but this is still done for many situations and has been a very useful tool in practice.

What Can We Do with More Than a Single Series

If we observe multiple time series of data, more opportunities arise

Possibilities
(a) Run regressions (build models) in the usual way.

(b) More complicated models of the dynamics

(c) Does one variable impact another?

What Can We Do with More Than a Single Series

First, what does the data look like?

Build Models in the usual way

We can consider our usual OLS regression

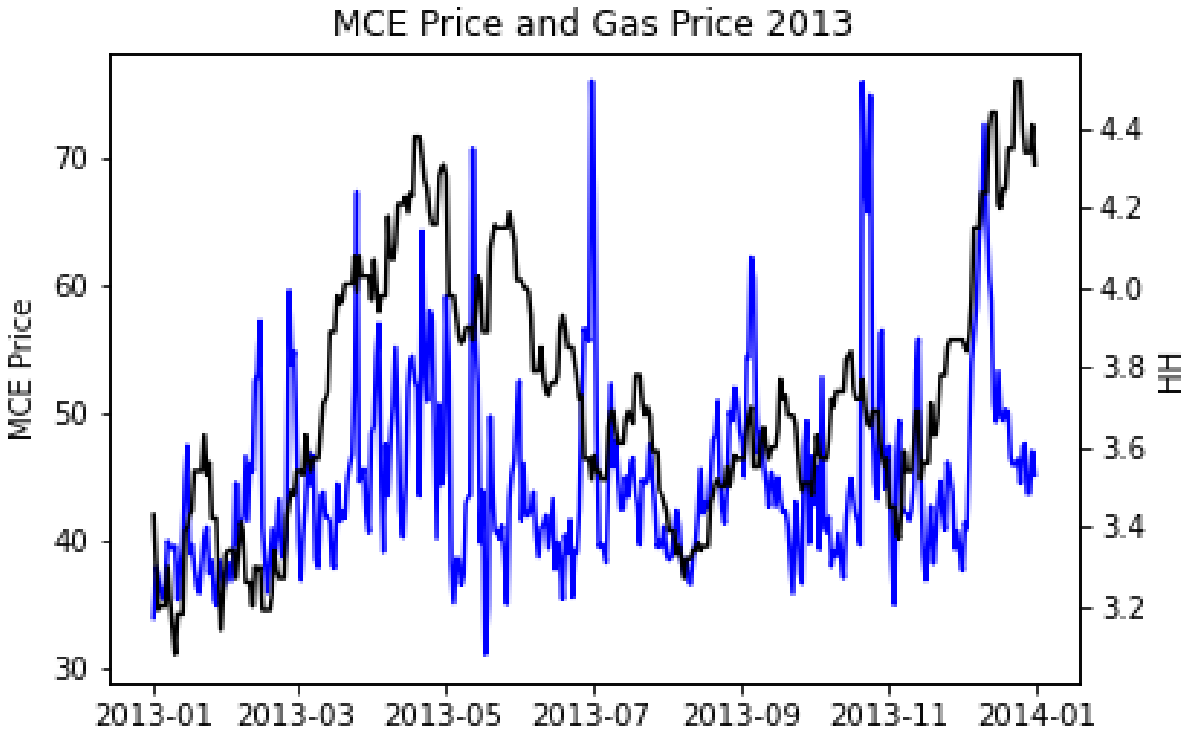
$$y_t = \mu_t + \beta' x_t + u_t$$

For stationary data, estimation is the same but again we need to fix the variance (use variance estimators robust to dependence).

Interpretation of the results is as usual.

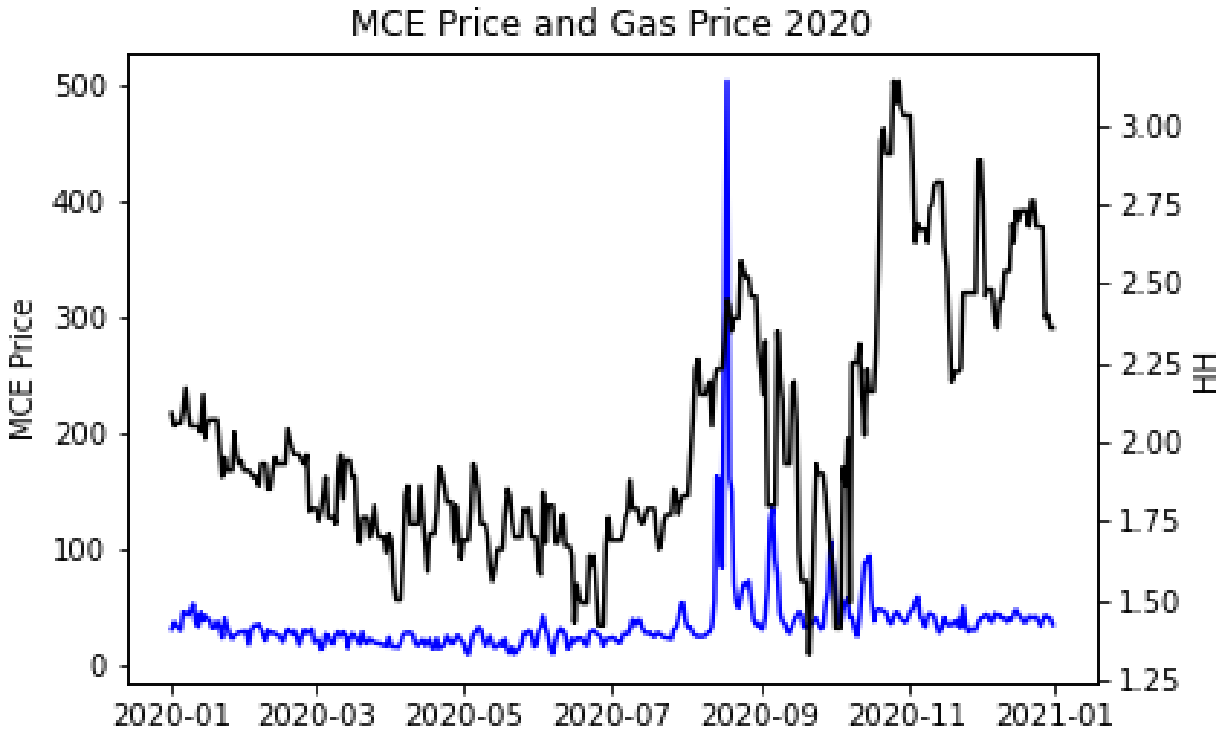
Build Models in the usual way

The relationship between electricity and gas prices



Build Models in the usual way

The relationship between electricity and gas prices



Build Models in the usual way

The relationship between electricity and gas prices

```
# run a regression
bdv = bd[['LMP','HenryHub']]
bdv=bdv['2013']
reg = smf.ols('LMP ~ HenryHub',data = bdv).fit(cov_type='HAC',cov_kwds={'maxlags
print(reg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    LMP   R-squared:                       0.103
Model:                            OLS   Adj. R-squared:                  0.101
Method:                 Least Squares   F-statistic:                     13.42
Date:                Sun, 26 Jun 2022   Prob (F-statistic):           0.000286
Time:                        18:04:10   Log-Likelihood:                -1226.4
No. Observations:                 365   AIC:                             2457.
Df Residuals:                     363   BIC:                             2465.
Df Model:                           1
Covariance Type:                  HAC
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     17.5202      7.345      2.385      0.017       3.125      31.915
HenryHub       7.3249      2.000      3.663      0.000       3.406      11.244
==============================================================================
Omnibus:                      132.258   Durbin-Watson:                   0.710
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              408.657
Skew:                           1.673   Prob(JB):                     1.82e-89
Kurtosis:                       6.960   Cond. No.                         46.5
==============================================================================
```

Nonstationarity Again

We still have to be careful about nonstationarity.

Causal Relationships?

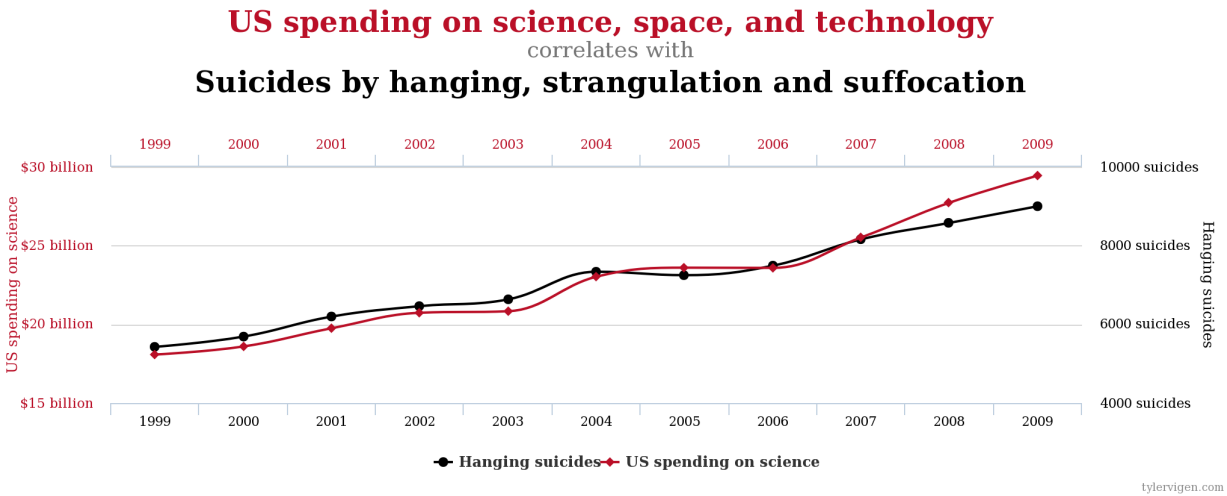


Figure: The difficulties of study!!!

Causal Relationships?

Centewall, Journal of the American Medical Association (1992).

Paper examines the causal effect of Television on various crimes.

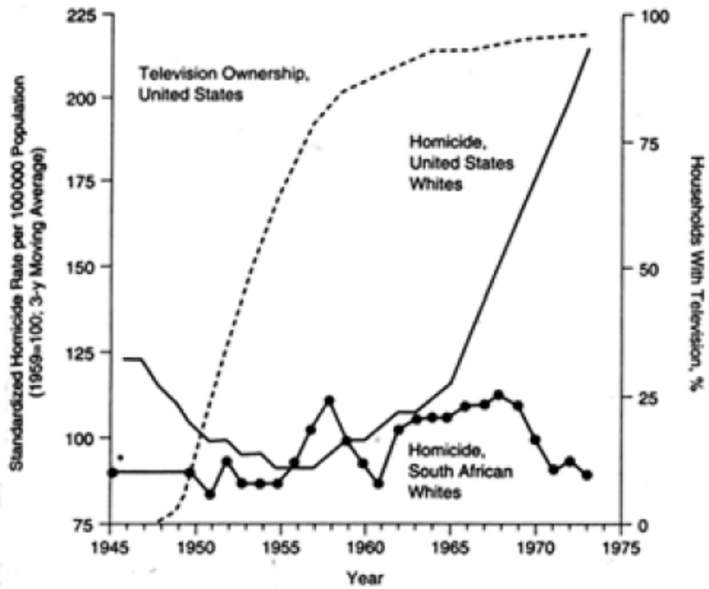


Fig 2. Television ownership and white homicide rates, United States and South Africa, 1945 through 1973. Asterisk denotes 6 year average. Note that television broadcasting was not permitted in South Africa prior to 1975 (from Centewall,8 and reprinted by permission of Academic Press).

"If, hypothetically, television technology had never been developed, there would today be 10,000 fewer homicides each year in the US, 70,000 fewer rapes …."

Figure: From JAMA

Causal Relationships?

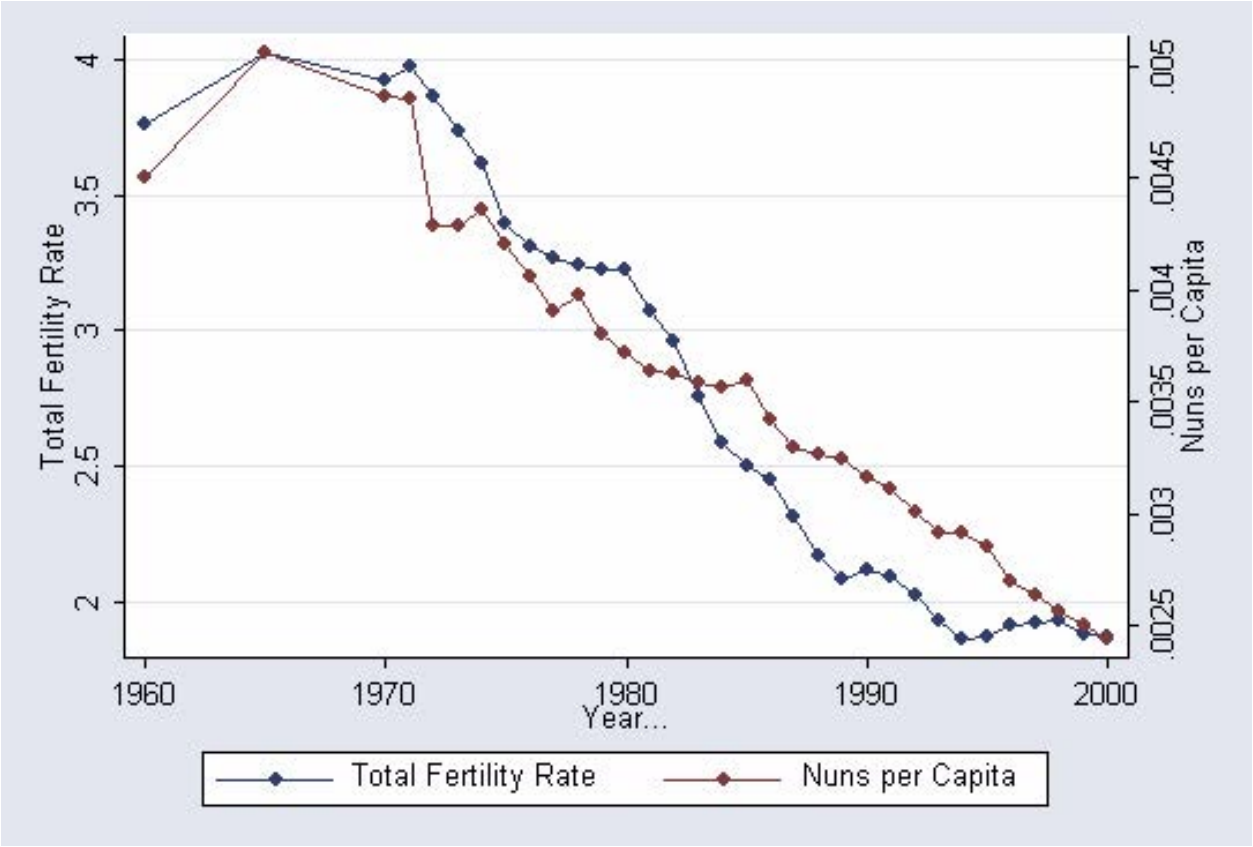


Figure: Ireland 1960-2000

Spurious regression

Scatter plot with data that grows over time

More Complicated Models of the Dynamics

The relationship between electricity and gas prices can include dynamics

The most common approach is a linear extension of the AR model called a Vector Autoregression (VAR).

This model allows us to see not only dynamics of one variable or another but also dynamics generated by one variable on another.

For example we could think about the effect of a shock to Gas prices on the path of electricity prices.

More Complicated Models of the Dynamics

The Vector Autoregression (VAR)

More Complicated Models of the Dynamics - 2020

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Thu, 09, Jun, 2022
Time:                     16:50:30
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                    2.17755
Nobs:                     363.000    HQIC:                   2.08706
Log likelihood:          -1384.11    FPE:                    7.59403
AIC:                      2.02735    Det(Omega_mle):         7.30941
--------------------------------------------------------------------
Results for equation LMP
===============================================================================
                coefficient       std. error           t-stat            prob
-------------------------------------------------------------------------------
const              -0.711146         6.726452          -0.106           0.916
L1.LMP              0.745284         0.052438          14.213           0.000
L1.HenryHub        -5.121350        11.331032          -0.452           0.651
L2.LMP             -0.180122         0.064954          -2.773           0.006
L2.HenryHub         2.747083        15.249493           0.180           0.857
L3.LMP              0.163476         0.052449           3.117           0.002
L3.HenryHub         7.668243        11.320699           0.677           0.498
===============================================================================

Results for equation HenryHub
===============================================================================
                coefficient       std. error           t-stat            prob
-------------------------------------------------------------------------------
const               0.082839         0.031458           2.633           0.008
L1.LMP              0.000141         0.000245           0.574           0.566
L1.HenryHub         0.900740         0.052993          16.997           0.000
L2.LMP             -0.000214         0.000304          -0.704           0.481
L2.HenryHub         0.086924         0.071319           1.219           0.223
L3.LMP              0.000389         0.000245           1.585           0.113
L3.HenryHub        -0.033936         0.052945          -0.641           0.522
===============================================================================

Correlation matrix of residuals
                LMP  HenryHub
LMP        1.000000  0.075959
HenryHub   0.075959  1.000000
```

More Complicated Models of the Dynamics -2014

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Wed, 13, Jul, 2022
Time:                     16:06:09
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                    1.21487
Nobs:                     362.000    HQIC:                   1.12420
Log likelihood:          -1205.96    FPE:                    2.89903
AIC:                      1.06437    Det(Omega_mle):         2.79008
--------------------------------------------------------------------
Results for equation LMP
===============================================================================
                coefficient       std. error           t-stat            prob
-------------------------------------------------------------------------------
const              8.256718         2.254025            3.663           0.000
L1.LMP             0.638983         0.052918           12.075           0.000
L1.HenryHub        8.972225         1.099618            8.159           0.000
L2.LMP            -0.000124         0.060811           -0.002           0.998
L2.HenryHub       -7.483958         1.667048           -4.489           0.000
L3.LMP            -0.061668         0.049851           -1.237           0.216
L3.HenryHub        1.510263         1.188987            1.270           0.204
===============================================================================

Results for equation HenryHub
===============================================================================
                coefficient       std. error           t-stat            prob
-------------------------------------------------------------------------------
const              0.223614         0.102011            2.192           0.028
L1.LMP             0.003925         0.002395            1.639           0.101
L1.HenryHub        1.153558         0.049766           23.180           0.000
L2.LMP            -0.003831         0.002752           -1.392           0.164
L2.HenryHub       -0.593651         0.075446           -7.869           0.000
L3.LMP             0.002306         0.002256            1.022           0.307
L3.HenryHub        0.360039         0.053810            6.691           0.000
===============================================================================

Correlation matrix of residuals
                LMP  HenryHub
LMP        1.000000  0.028084
HenryHub   0.028084  1.000000
```