Lecture 1: Time Series

Chris Conlon

NYU Stern

February 4, 2020

Packages for Today

Let's load some packages so that I can make some better looking plots:

What is Time Series?

Big Picture

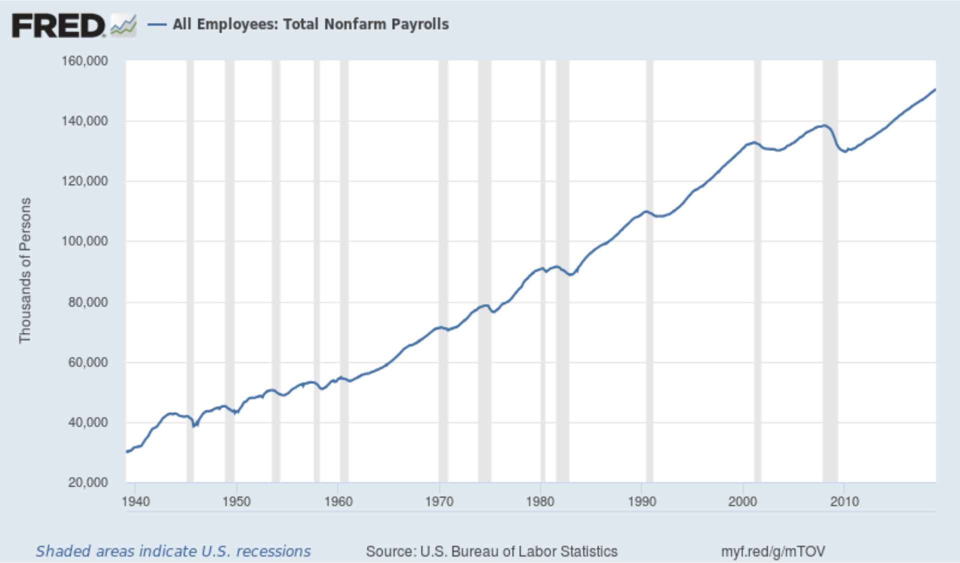
So far you have mostly studied cross sectional econometrics (subscript i):

- Individual observations are independent of one another (mostly).
- Large number of individuals allows us to do inference (LLN, CLT).

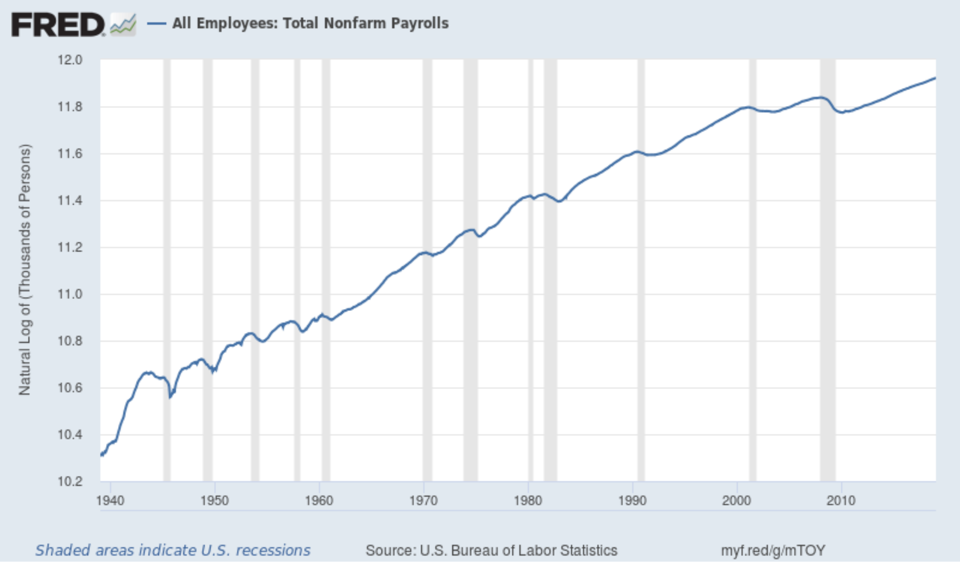
But suppose we observe a single object for may periods (subscript t):

- Now we worry that y_t is autocorrelated with y_{t-1}
- This means that independence is not going to hold.
- This presents a number of challenges addressed in time series econometrics.

Employment

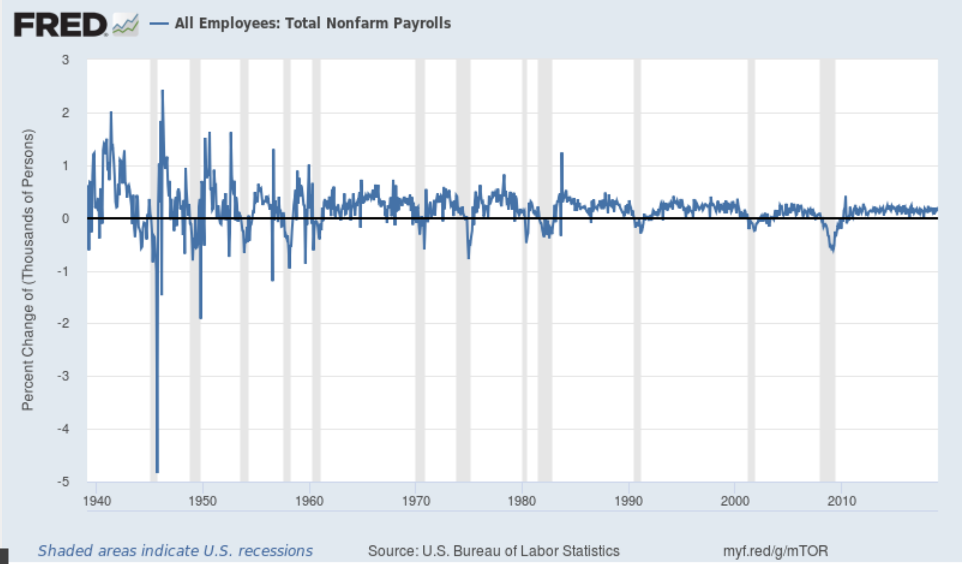

FRED
— All Employees: Total Nonfarm Payrolls
Thousands of Persons
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
1940
1950
1960
1970
1980
1990
2000
2010
Shaded areas indicate U.S. recessions
Source: U.S. Bureau of Labor Statistics
myf.red/g/mTOV

Employment



FRED
All Employees: Total Nonfarm Payrolls
Natural Log of (Thousands of Persons)
12.0
11.8
11.6
11.4
11.2
11.0
10.8
10.6
10.4
10.2
1940
1950
1960
1970
1980
1990
2000
2010
Shaded areas indicate U.S. recessions
Source: U.S. Bureau of Labor Statistics
myf.red/g/mTOY

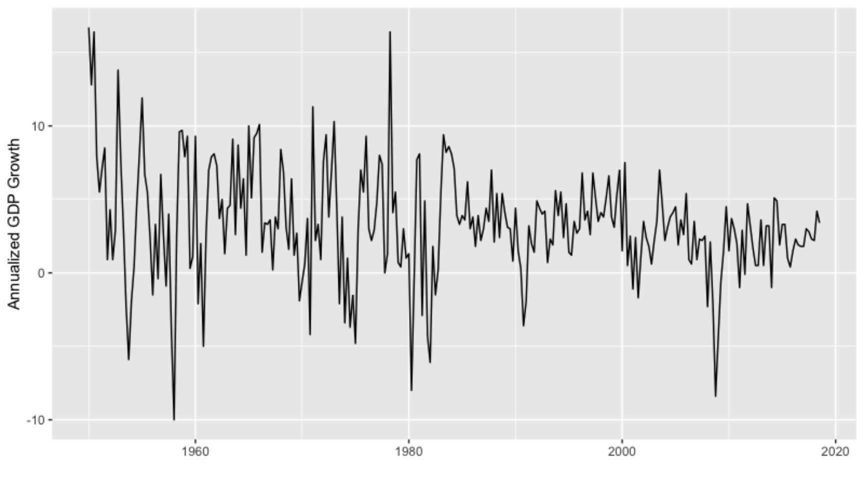
Employment



FRED
All Employees: Total Nonfarm Payrolls
Percent Change of (Thousands of Persons)
3
2
1
0
-1
-2
-3
-4
-5
1940
1950
1960
1970
1980
1990
2000
2010
Shaded areas indicate U.S. recessions
Source: U.S. Bureau of Labor Statistics
myf.red/g/mTOR

Do it ourselves

Do it ourselves



Theory of Time Series

We observe a sample $\{y_1, y_2, \ldots, y_{t-1}, y_t, y_{t+1}\}$.

- We call y_{y-1} the first lag of y_t.
- We call $\Delta y_t = y_t - y_{t-1}$ the first difference
- We might also want $\Delta \ln y_t = \ln y_t - \ln y_{t-1}$
- We can approximate percentage change as $100 \cdot \Delta \ln y_t$

Measure the correlation of a series with its own lagged values

- First autocovariance of y_t is $Cov(y_t, y_{t-1}) = \gamma(1)$.
- The jth autocovariance of y_t is $Cov(y_t, y_{t-j}) = \gamma(j)$.

Questions

1. How do we represent $Var(y_t)$?
2. Can we show that $\gamma(k) = \gamma(-k)$? (even function)
3. Can we show that $\gamma(0) \geq |\gamma(k)|$ for any k?
4. Does this imply that $|\gamma(k)| \geq |\gamma(k-1)|$?

We can also compute the autocorrelaton coefficient j:

$$Corr(y_t, y_{t-j}) = \frac{Cov(y_t, y_{t-j})}{Var(y_t)} = \frac{\gamma(j)}{\gamma(0)} = \rho(j)$$

With sample analogue

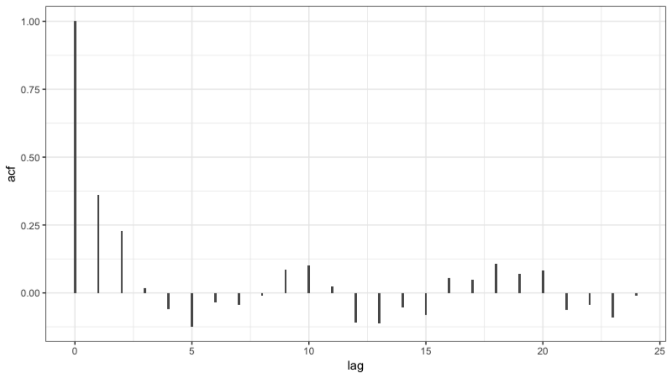
$$\widehat{Corr(y_t, y}_{t-j}) = \frac{\widehat{\gamma}(j)}{\widehat{\gamma}(0)} = \widehat{\rho}(j)$$

Which we can estimate via:

$$\widehat{\rho}(j) = \frac{1}{T}\sum_{t=j+1}^{T}(y_t - \bar{y})(y_{t-j} - \bar{y})$$

- Most software uses $\frac{1}{T}$ instead of d.o.f corrected $\frac{1}{T-j}$
- Some software uses mean of $\{y_{j+1}, y_T\}$ and $\{y_1, y_{T-j}\}$ instead of grand mean
- Can Correct autocorrelation between (y_t, y_{t-h}) removing dependence on $y_1, \ldots, y_{t-h+1}$ [PCF]

ACF PLOTS



Conceptually stationarity is one of the most important issues with time series:

- Basic idea: the future needs to look like the past (at least probabilistically)
- I cheated on previous slides and assumed stationarity. Why?
- Simplified: $Cov(y_t, y_{t-k})$ is allowed to depend on k but not on t.
    - Relationship between y_t and its lags is constant across time
- Formally we need the joint distribution $f(y_{s+1}, y_{s+2}, \ldots, y_{s+T})$ to be invariant to s.
- Weaker form: Covariance Stationary

We probably want something like an LLN or CLT:

- Independence is violated between (y_t, y_{t-k})
- Idea: consider a large value H and assume stationarity:
  - The block $(y_t, y_{t-1}, y_{t-2}, \ldots, y_{t-k})$ and $(y_{t+H}, y_{t-1+H}, y_{t-2+H}, \ldots, y_{t-k+H})$ are as if they are independent for some large enough choice of H.
  - How is H determined? The mixing rate of the time series?
  - In practice? Looking at the ACF function/plot

Hand Waving Technical Stuff

- Soemtimes people will talking about mixing properties or the mixing rate
- This tells us how far apart in time two observations are before we can treat them as if they are "independent".
- Another property is ergodicity

$$\sum_{k=0}^{\infty} |\gamma(k)| = \gamma(0)\tau < \infty$$

- τ is the correlation time
- We could look at the variance of $\overline{X}_t$ to derive this but
- It is as if we have $\frac{n}{1+2\tau}$ effective independent observations

$AR(1)$ Regression

Consider the first-order autoregression for a forecast:

$$y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$$

- No causal interpretation of (β_0, β_1).
- $\beta_1 = 0$ means that y_{t-1} is not informative about y_t.
- We can run this regression using OLS

$AR(1)$ Regression

Consider the first-order autoregression for a forecast:

$$y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$$

- No causal interpretation of (β_0, β_1).
- $\beta_1 = 0$ means that y_{t-1} is not informative about y_t.
- We can run this regression using OLS

$AR(1)$ Example

Wold Decomposition

Start with the AR(1) where ε_t is I.I.D with some variance σ^2:

$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \\
y_{t-1} &= \beta_0 + \beta_1 y_{t-2} + \varepsilon_{t-1} \\
y_{t-2} &= \beta_0 + \beta_1 y_{t-3} + \varepsilon_{t-2}
\end{aligned}
$$

Can we re-write the sequence as function of ϵ_t's only?

$$
y_t = \underbrace{\beta_0 + \beta_1\beta_0 + \beta_1^2\beta_0}_{\widetilde{\beta}_0} + \beta_1\varepsilon_{t-1} + \beta_1^2\varepsilon_{t-2} + \varepsilon_t \ldots
$$

$$
y_t = \widetilde{\beta}_0 + \sum_{k=1}^{t} \beta^k \varepsilon_{t-k}
$$

Wold Decomposition

Our $AR(1)$ can be written as a $MA(\infty)$ moving average process:

$$y_t = \widetilde{\beta}_0 + \sum_{k=1}^{\infty} \beta^k \varepsilon_{t-k}$$

- We call this an $MA(\infty)$ process because it represents a β_1 weighted moving average of all past realizations of ε_t
- Wold's Theorem tells us we can write any stationary time series as the sum of a deterministc and stochastic component.

Wold Decomposition

Consider the Wold Representation of the $AR(1)$

$$y_t = \widetilde{\beta}_0 + \sum_{k=1}^{\infty} \beta_1^k \varepsilon_{t-k}$$

Assume that $\varepsilon \sim N[0, \sigma^2)$ and IID

$$E[y_t] = \widetilde{\beta}_0$$

$$V[y_t] = \sum_{k=1}^{\infty} \beta_0^k Var(\varepsilon_{t-k}) \to \frac{1}{1-\beta_1}\sigma^2$$

- Here stationarity requires $\beta_1 \in (0, 1)$.
- Note that as $\beta_1 \to 1$ implies that the series no longer converges
- This is what is known as a unit root

We could also construct an $AR(2)$

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$$

Or an $AR(p)$:

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \varepsilon_t$$

Or an $ARMA(p, q)$ which adds moving average terms:

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{p} \theta_k \varepsilon_{t-k}$$

- An important question is selecting the order of the lag p

What About Lag Selection

Think about the $AR(p)$ model, which order lag do we choose?

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \varepsilon_t$$

- More lags → Better Fit
- Potential for overfitting
- Bias vs. Variance tradeoff

Information Criteria

$$AIC(p) = \ln\left(\frac{SSR(p)}{T}\right) + (p+1)\frac{2}{T}$$

$$BIC(p) = \ln\left(\frac{SSR(p)}{T}\right) + (p+1)\frac{\ln T}{T}$$

The penalty is smaller for *AIC* than for *BIC*

- *AIC* estimates more lags (bigger p) than *BIC*
- *AIC* tends to overestimate p

There are other information criteria and ways to calculate.

$AR(p)$ Example: Auto-selecting

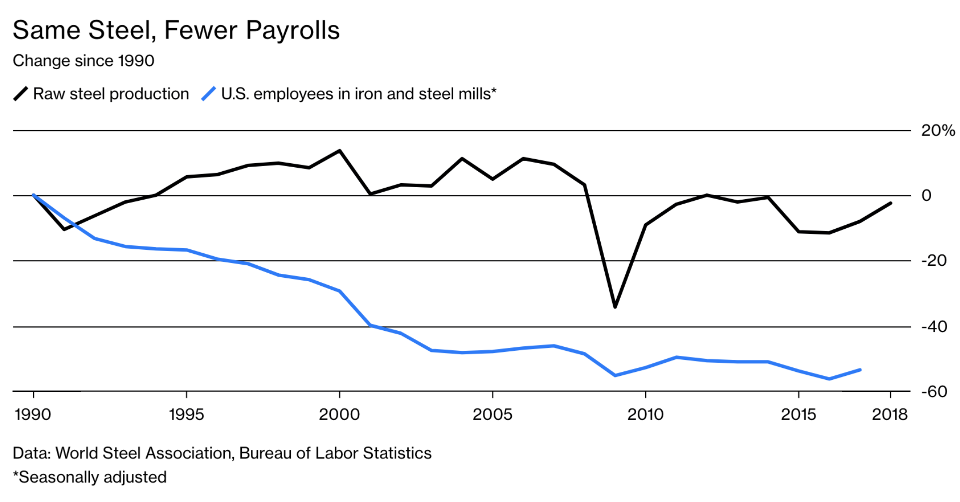
$ADL(p, r)$ models add the covariate X (and its lags). Usually contemporaneous X_t is excluded:

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{r} \theta_k X_{t-k} + \varepsilon_t$$

An important issue is Granger Causality

- This has nothing to do with actual causality
- Include $p > r$ lags of y_t. Does $(x_t, x_{t-1}, \ldots, x_{t-p})$ have any predictive value?
- Joint F-test of all coefficients on x_t lags

Steel Production and Employment



$ADL(3,3)$ Example

Granger Test

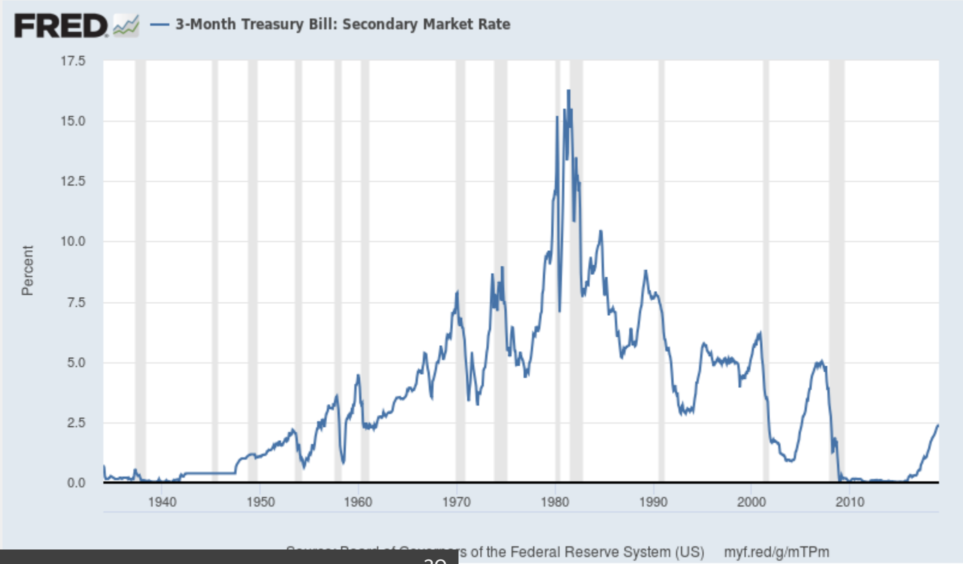
Significant! Hours predict output.

Granger Test: Other Direction

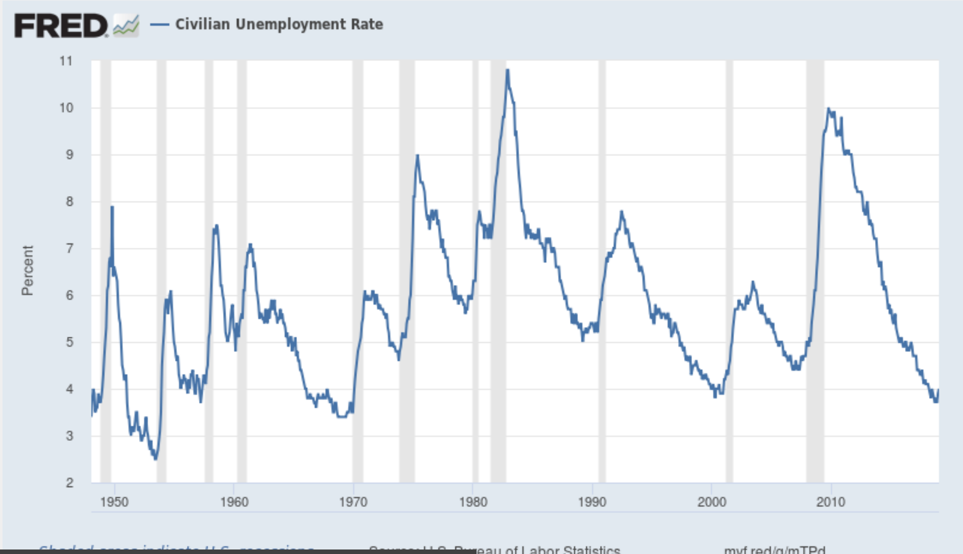
Not significant! Output does not predict hours.

Trends

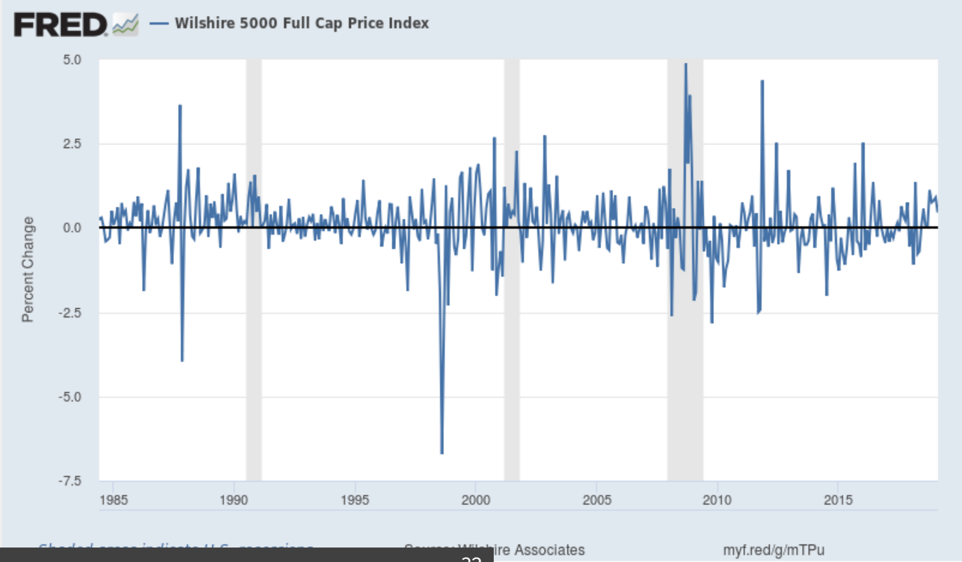
Which Series has a Trend?



Which Series has a Trend?



Which Series has a Trend?



Two Kinds of Trends

- Deterministic Trends: $y_t = a \cdot t + \epsilon_t$ or $y_t = a \cdot t + b \cdot t^2 + \epsilon_t$
- Stochastic Trend: random and time varying trend (see how this works later)
- Random Walk: $Y_t = Y_{t-1} + \varepsilon_t$

$$Y_t = Y_{t-1} + \varepsilon_t, \quad E[\varepsilon_t] = 0, V[\varepsilon_t] = \sigma^2$$

- Best guess of tomorrow is today
- $E[y_{t+h}|y_t] = y_t$ for any t and h
- If Y_0 then $V(y_t) = t\sigma^2$

Generate Random Random Walks

Adding Drift

We an easily add a drift term β_0

$$Y_t = Y_{t-1} + \beta_0 + \varepsilon_t$$

- $E[y_{t+h}|y_t] = y_t + h \cdot \beta_0$ for any t and h
- If Y_0 then $V(y_t) = t\sigma^2$

Log stock prices are roughly RWD
(stock returns are random but positive on average)

Where are we heading?

Suppose we have a stochastic (random walk) trend:

- We no longer satisfy stationarity
- We can run OLS but we can't trust the results (not even a little bit)
  - Recall $AR(1)$ has non-convergent series!
  - Coefficients are biased towards zero
  - Not asymptotically normal
- We are going to want to transform things to return to stationary case
- Easy for RW trend because Δy_t is stationary!

$$y_t = y_{t-1} + \varepsilon_t$$
$$\Delta y_t = \varepsilon_t$$

A Simple Example: $AR(1)$

We can think about RWD as a special case of $AR(1)$ with $\beta_1 = 1$

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t \quad AR(1) \\
Y_t &= \beta_0 + Y_{t-1} + \varepsilon_t \quad \text{RWD} \\
\Delta Y_t &= \beta_0 + \varepsilon_t
\end{aligned}
$$

We call the β_1 case unit root because $1 - \beta_1 z = 0$ has root $z = \frac{1}{\beta_1}$ so that β_1 when $z = 1$.

Harder Example: $AR(2)$

This case is more complicated

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t \\ &= \beta_0 + (\beta_1 + \beta_2) Y_{t-1} - \beta_2 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t \\ &= \beta_0 + (\beta_1 + \beta_2) Y_{t-1} - \beta_2 (Y_{t-1} - Y_{t-2}) + \varepsilon_t \end{aligned}$$

Now difference Y_{t-1}:

$$\begin{aligned} Y_t - Y_{t-1} &= \beta_0 + \underbrace{(\beta_1 + \beta_2 - 1)}_{\delta} Y_{t-1} - \beta_2 \underbrace{(Y_{t-1} - Y_{t-2})}_{\Delta Y_t} + \varepsilon_t \\ \Delta Y_t &= \beta_0 + \delta Y_{t-1} - \beta_2 \Delta Y_t \varepsilon_t \end{aligned}$$

A Harder Example: $AR(2)$

What is a unit root now?

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} - \beta_2 \Delta Y_t \varepsilon_t$$

- $1 - \beta_1 z - \beta_2 z^2 = 0$ a unit root implies that $\beta_1 + \beta_2 = 1$
- If there is a unit root then $\delta = 0$
  - We can use this to construct a test for a unit root
- If $AR(2)$ has a unit root, then write as an $AR(1)$ in first differences

$$\Delta Y_t = \beta_0 - \beta_2 \Delta Y_t \varepsilon_t$$

The General Case $AR(p)$

What is a unit root now?

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdot + \beta_p Y_{t-p} + \varepsilon_t$$
$$\Delta Y_t = \beta_0 + \Delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_p \Delta Y_{t-p} + \varepsilon_t$$

With coefficients:

$$\begin{aligned}
\delta &= \beta_1 + \beta_2 + \cdots + \beta_p - 1 \\
\gamma_1 &= -(\beta_2 + \cdots + \beta_p) \\
\gamma_2 &= -(\beta_3 + \cdots + \beta_p) \\
\gamma_{p-1} &= -\beta_p
\end{aligned}$$

- Thus the $AR(p)$ becomes an $AR(p-1)$ in first differences.
- Again $\delta = 0$ tells us whether or not unit root is present

- Plot the Data: are there persistent long run movements?
- Run the Dickey-Fuller Test for unit roots

Dickey Fuller Test for $AR(1)$:

$$Y_t = \beta_0 + \beta_1 Y_{t_1} + \varepsilon_t$$
$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \mu \cdot t + \varepsilon_t$$

- $H_0 : \delta = 0$ vs $H_1 : \delta < 0$ (one sided test)
- The usual critical values for t-stats don't work (because at $\delta = 0$ things are non-normal).
- Software usually has adjusted critical values

Dickey Fuller Test

Which test do we want?

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \mu \cdot t + \varepsilon_t$$

- Can include the trend $\mu \cdot t$ or not
- Leads to different critical values
- Depends on whether y_t is stationary around a trend or not
- Need to choose number of lags first

Dickey Fuller Test: Example

Dickey Fuller Test: Example

Spurious Regression/Correlation

Imagine we have two series each with a trend

$$y_t = a_0 + a_1 t + \varepsilon_t$$
$$x_t = b_0 + b_1 t + \mu_t$$

- Both are related to t but neither has anything to do with each other.
- Regression of x_t on y_t can produce very high R^2

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$
$$y_t = \beta_0 + \beta_1 (b_0 + b_1 \cdot t + \mu_t) + \varepsilon_t$$
$$y_t = \underbrace{(\beta_0 + \beta_1 b_0)}_{\tilde{\beta}_0} + \underbrace{\beta_1 b_1}_{\tilde{\beta}_1} \cdot t + \underbrace{(\beta_1 \mu_t + \varepsilon_t)}_{\tilde{\varepsilon}_t}$$

Spurious Regression/Correlation

- This is a huge mistake and people make it all of the time
- `http://www.tylervigen.com/spurious-correlations`
- This problem is insidious: it seems obvious and then you do it

Applications of Time Series

Moving Average Models

We might want a trend but one that isn't a straight line.
Enter the simple q Moving average (SMA):

$$Y_t = \frac{Y_{t-1} + Y_{t-2} + \cdot + Y_{t-m}}{m}$$

- The average age of the data is around $\frac{m+1}{2}$ periods.
- We are always behind what is happening at time t
- As we include more lags, we use more data, but we get further behind today.
- Gets plotted a lot on stock market prices, etc.

Moving Average: S&P 500 w/ MA(60)

Simple Exponential Smoothing (SES)

We might want to weight older observations less and more recent observations more. Think about $L_t = E[Y_{t+1}|Y_t]$ our forecast of Y_{t+1}:

$$L_t = \alpha Y_t + (1-\alpha)L_{t-1}$$
$$E[Y_{t+1}|Y_t]] = \alpha Y_t + (1-\alpha)\hat{Y}_t$$

Notice that $\varepsilon_t \equiv Y_t - E[Y_t|Y_{t_1}]$ so that

$$E[Y_{t+1}|Y_t]] = \alpha E[Y_t|Y_{t-1}] + \alpha\varepsilon_t$$

Rewriting as a moving average

$$E[Y_{y+1}|Y_t] = \alpha[Y_t + (1-\alpha)Y_{t-1} + (1-\alpha)^2 Y_{t-2} + (1-\alpha)^3 Y_{t-3} + \ldots]$$

- Update the old forecast in direction of forecast error
- $\alpha = 0$ constant, $\alpha = 1$ RW

Decomposing Trends and Seasonality

Given some time series data how should we start?

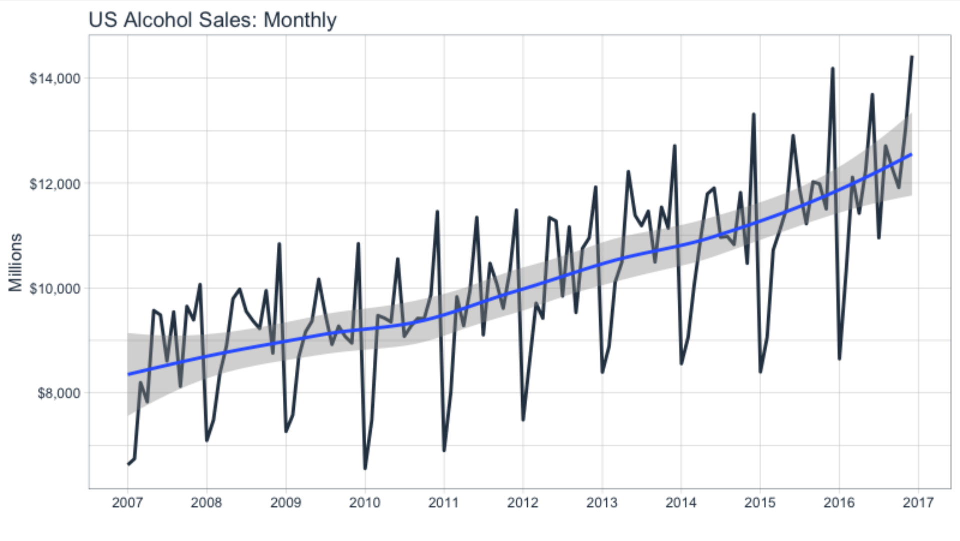
- Plot the series
- Try and decompose the series
  - Extract trends
  - Look for seasonality
  - Remainder should be random

Loading Alcohol Data :

https://fred.stlouisfed.org/series/S4248SM144NCEN

Plotting Alcohol Data

Alcohol Example



Rearranging Alcohol Data

Notice the strong seasonal pattern (December and June)

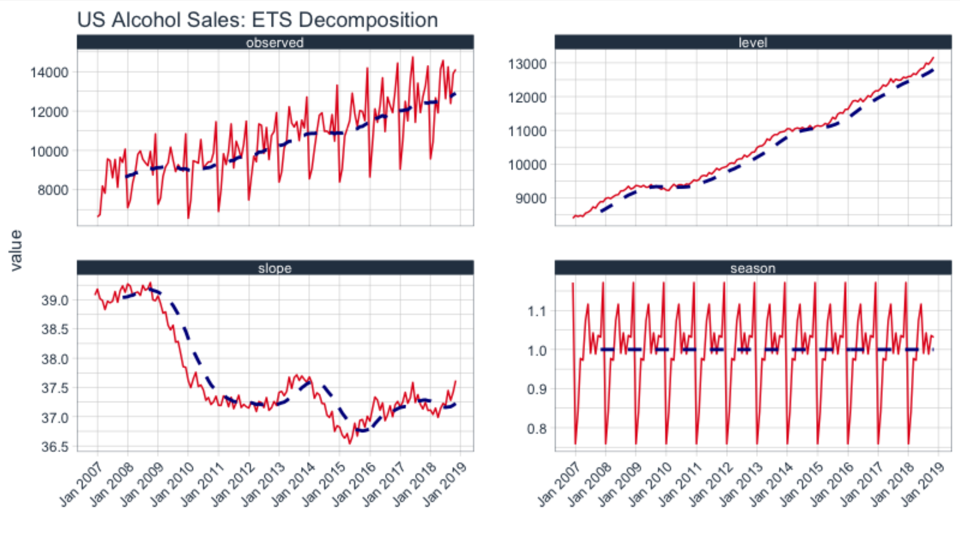
Rearranging Alcohol Data

Apply Error Trend Seasonal Decomposition (ETS) to data. These are not really interpretable on their own:

Rearranging Alcohol Data

Run the equivalent of `decompose` on the data:

ETS/decompose Example



Consider Auto-Regressive Integrated Moving Average $ARIMA(p, d, q)$

- Autoregressive p terms like $AR(p)$: lags of y_{t-p}
- Integrated d Differenced out unit roots
- Moving Average q include lags of forecast errors ϵ_{t-h}

Denote by (p, d, q)

- $(0, 0, 0) + c$ constant model
- $(0, 1, 0)$ RW
- $(0, 1, 0) + c$ RW w/ drift
- $(1, 0, 0)$ $y_t \sim y_{t-1}$
- $(1, 1, 0)$ $\Delta y_t \sim \Delta y_{t-1}$
- $(2, 1, 0)$ $\Delta y_t \sim \Delta y_{t-1} + \Delta y_{t-2}$
- $(0, 1, 1)$ SES model
- $(0, 1, 1) + c$ SES with constant trend

More Serious: X-13 ARIMA

Lots of government economic series are seasonally adjusted

- The Census uses X-13 software to seasonally adjust most series
- Also popular is Bank of Spain (SEATS) adjustment
- available in R package `seasonal`
- `https://github.com/christophsax/seasonal/wiki/Examples-of-X-13ARIMA-SEATS-in-R`

Next time: Panel Data

- Linear Model
- Serial Correlation
- Fixed Effects, Random Effects
- Dynamic Panel: Arellano Bond, etc.