1. Time Series Data

Up until this point, we have studied cross-sectional or panel data. With those types of data, we could look for relationships between variables by running regressions with *many* different people, firms, schools, etc.

Time series data is different because it follows *one* unit over time. The variables involved are often *aggregate* measures, like GDP or unemployment. There are a couple of reasons to use aggregate data. One is when we actually want to learn more about an aggregate quantity (like GDP). The other is when we want to learn about individual-level behavior (e.g. demand for a product as a function of price) but only have access to more aggregated data.

What makes time series data different from cross-sectional data?

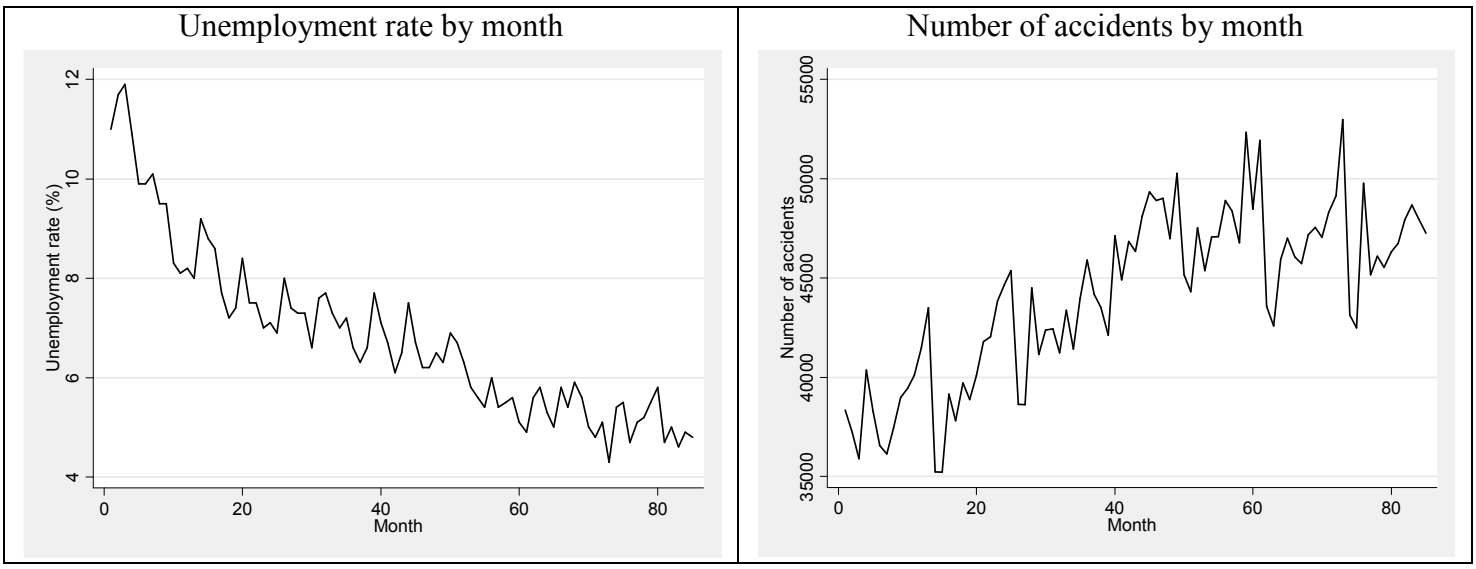
Cross-section	**Time series**
Requires a random sample. This is easy to define – just make sure to pick your observations randomly from the population.	The sample might not seem "random" – after all, there is only one realization of GDP each quarter, so how could we sample randomly among GDP realizations? For practical purposes, this isn't a problem for us. You can think of GDP (or whatever the variable) as the outcome of a random ("stochastic") process.
Observations are independent. Knowing the values of x and y for Observation 1 doesn't give us any information about the x and y of Observation 2.	Observations are not independent. Knowing (for example) GDP and unemployment in Year 1 gives me a pretty good guess of what they'll be in Year 2 because those variables change slowly over time.
Usually have a lot of observations—often enough to stop worrying about the degrees of freedom when we do t-tests.	Usually don't have many observations—just once a month, quarter, or year for some number of years. We must pay close attention to the degrees of freedom when performing a t-test (or look at the p-value from Stata).

2. "Time" as an Omitted Variable

The goal of a regression is often to find out if and by how much x affects y. The biggest obstacle to doing this is omitted variables bias. This is a huge problem with time series data because if x and y both have a time trend, we can confuse these trends for a relationship between x and y.

Example:
Do higher unemployment rates lower the number of car crashes? Fewer people are driving to work, so this could be true.

We have data from 1983-1989 on unemployment rate and number of car crashes for one state (I don't know which). Here are both variables, plotted vs. time:

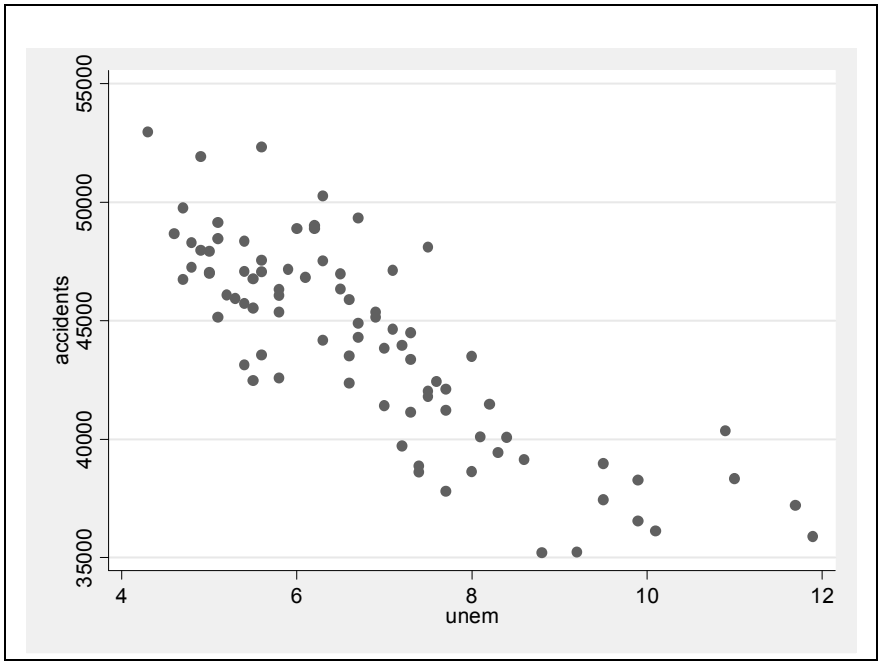


Over this time period in the 1980s, unemployment was falling. Number of accidents was rising.

We want to regress accidents on unemployment:

$$accidents_t = \beta_0 + \beta_1 unem_t + u_t$$

We know that a regression is just fitting a line through a set of points, so let's guess what the regression will find by sketching a graph:



What sign will $\hat{\beta}_1$ have? ____negative_____

Is $\hat{\beta}_1$ likely to be statistically significant? _____yes—the relationship appears very strong_____

Are you convinced that it was unemployment causing accidents to change? ____no_____

Looking at the top graphs, what's the most important omitted variable in our regression? ____time_____

3. Two Solutions for Time as an OV

Solution 1: Include a variable for time

Since time is the omitted variable (it "causes" accidents and is correlated with unemployment), we can just create a variable representing the month/quarter/year/etc. that we're in, and include that. This will control for a *linear trend* (a straight line) in the data. Letting $t = 1, 2, 3, ...$, we can write out the regression:

$$accidents_t = \beta_0 + \beta_1 unem_t + \beta_2 t + u_t$$

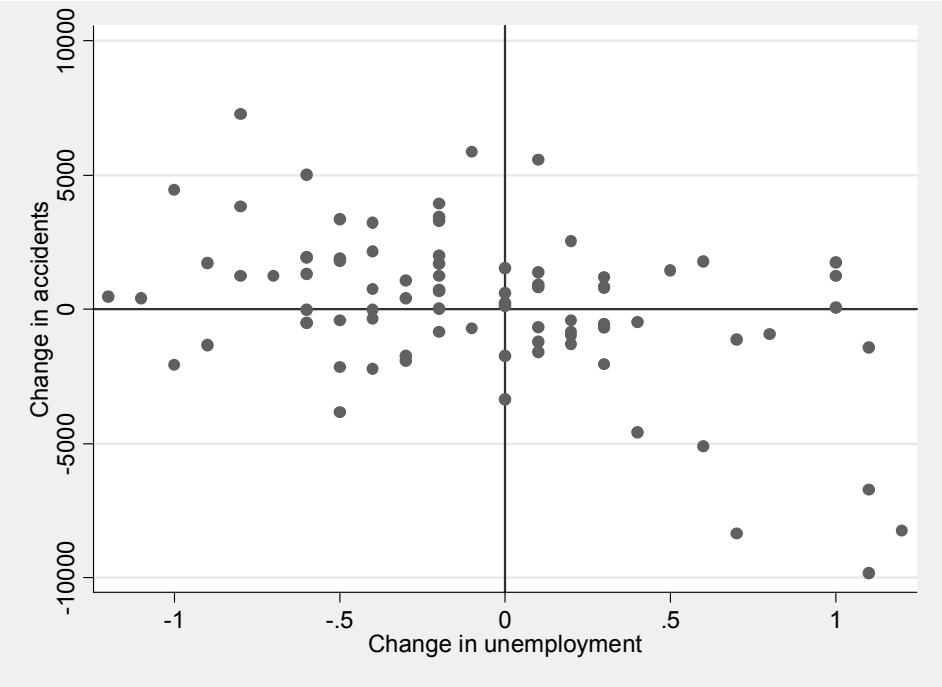
```
      Source |       SS       df       MS              Number of obs =      85
-------------+------------------------------           F(  2,    82) =   86.68
       Model |  1.0194e+09     2   509693963           Prob > F      =  0.0000
    Residual |   482178602    82  5880226.85           R-squared     =  0.6789
-------------+------------------------------           Adj R-squared =  0.6711
       Total |  1.5016e+09    84    17875792           Root MSE      =  2424.9

------------------------------------------------------------------------------
   accidents |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |  -1629.949   367.4759    -4.44   0.000    -2360.975   -898.9221
           t |   30.68196   25.41529     1.21   0.231    -19.87715    81.24107
       _cons |   53956.17   3541.389    15.24   0.000     46911.22    61001.12
------------------------------------------------------------------------------
```

This might be a good method when we can eyeball the variables over time and see a *constant, straight-line trend* in the data. Otherwise…

Solution 2: First-difference the data and run the regression on that data

When we ask if *x* causes *y*, it's the same as asking if increasing *x* causes an increase (or decrease) in *y*. In other words, are Δx and Δy related?

Defining $\Delta unem_t = unem_t - unem_{t-1}$ and $\Delta accidents_t = accidents_t - accidents_{t-1}$, we can graph them:



Turning this graph into a regression:

$$\Delta accidents_t = \beta_0 + \beta_1 \Delta unem_t + u_t$$

```
      Source |       SS       df       MS              Number of obs =      85
-------------+------------------------------           F(  1,    83) =   21.87
       Model |   145825267     1   145825267           Prob > F      =  0.0000
    Residual |   553314119    83  6666435.17           R-squared     =  0.2086
-------------+------------------------------           Adj R-squared =  0.1990
       Total |   699139387    84  8323087.94           Root MSE      =  2581.9

------------------------------------------------------------------------------
  daccidents |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dunem |  -2374.307   507.6535    -4.68   0.000     -3384.01   -1364.605
       _cons |  -61.30242   282.4886    -0.22   0.829    -623.1608     500.556
------------------------------------------------------------------------------
```

Luckily, we can interpret the first-differenced results *as if we were interpreting a "non-differenced" regression*. How do we interpret $\hat{\beta}_1$?

(Causal interpretation) Holding all else equal, a 1 percentage point increase in unemployment leads to 2,374 less car accidents per month.
("Predicted" interpretation) A 1 percentage point increase in unemployment decreases predicted accidents by 2,374.

4. Growth Rates

Suppose we want to know how much a variable grows per period, on average. This is really simple:

$$\log(y) = \beta_0 + \beta_1 t + u_t$$

where *t* is a number that goes up by 1 each time period ($t = 1, 2, 3, ...$). Then $\hat{\beta}_1$ is the average growth rate. We know this just by interpreting the coefficient as we always do:

"A 1-period increase in time raises predicted *y* by $\hat{\beta}_1 \times 100\%$." So if time is measured in years,

"Each year, *y* is predicted to increase by $\hat{\beta}_1 \times 100\%$."

5. Seasonality

Some variables are seasonal, i.e. they are always much higher in certain months than others. The seasonality itself can be interesting (are car accidents higher in December when people are traveling?). Or, it can be a nuisance that makes it hard to see the relationship we want to see. To control for seasonality, *add dummy variables for all months* (excluding one, of course). Example with linear trend (excluding January):

$$accidents_t = \beta_0 + \beta_1 unem_t + \beta_2 t + \delta_1 feb_t + \delta_2 mar_t + \cdots + \delta_{11} dec_t + u_t$$

```
      Source |       SS       df       MS              Number of obs =      85
-------------+------------------------------           F( 13,    71) =   21.62
       Model |  1.1987e+09    13  92207481.4           Prob > F      =  0.0000
    Residual |   302869270    71  4265764.36           R-squared     =  0.7983
-------------+------------------------------           Adj R-squared =  0.7614
       Total |  1.5016e+09    84    17875792           Root MSE      =  2065.4

------------------------------------------------------------------------------
   accidents |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |  -1056.939   354.1083    -2.98   0.004    -1763.011   -350.8672
           t |   61.78422   23.66835     2.61   0.011     14.59088    108.9776
         feb |  -755.2448   1104.827    -0.68   0.496    -2958.207    1447.717
         mar |   3375.545   1110.457     3.04   0.003     1161.357    5589.734
         apr |   804.0844   1128.654     0.71   0.479    -1446.388    3054.557
         may |   1295.495   1136.384     1.14   0.258      -970.39    3561.381
         jun |   1094.341   1117.825     0.98   0.331    -1134.538     3323.22
         jul |   2484.717   1105.367     2.25   0.028     280.6772    4688.757
         aug |   2790.583   1121.445     2.49   0.015      554.486     5026.68
         sep |   2029.618   1140.287     1.78   0.079    -244.0488    4303.285
         oct |   3626.985   1148.213     3.16   0.002     1337.513    5916.458
         nov |   3331.344   1143.326     2.91   0.005     1051.616    5611.072
         dec |   4360.414    1103.09     3.95   0.000     2160.915    6559.912
       _cons |   46640.28   3632.162    12.84   0.000     39397.95     53882.6
------------------------------------------------------------------------------
```