Lecture Notes 14
Econ 29000, Principles of Statistics
Kevin R Foster, CCNY
Fall 2011

Need to have X causing Y not vice-versa or both!

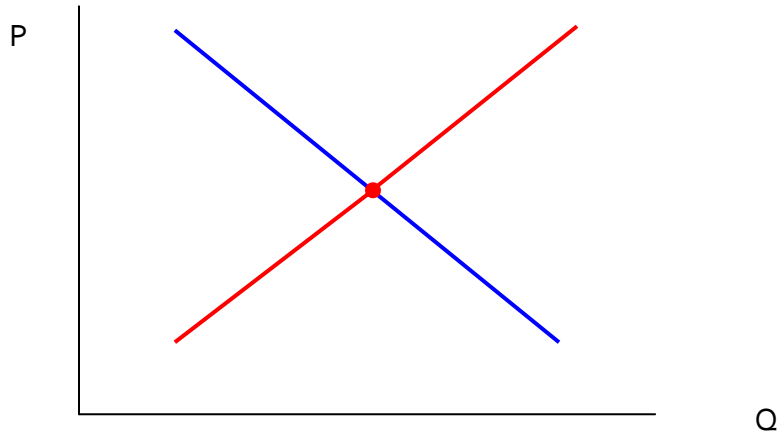
- Endogenous vs. Exogenous variables
    - Exogenous variables are generated from "exo" outside of the model; endogenous are generated from "endo" within the model. Of course this neat binary distinction rarely is matched by the world; some variables are more endogenous than others
- Data can only demonstrate correlations – we need theory to get to causation. "Correlation does not imply causation." Roosters don't make the sun rise.
- In any regression, the variables on the right-hand side should be exogenous while the left-hand side should be endogenous, so X causes Y, $X \to Y$. But we should always ask if it might be plausible for Y to cause X, $Y \to X$, or for both X and Y to be caused by some external factor. If we have a circular chain of causation (so $X \to Y$ and $Y \to X$) then the OLS estimates are meaningless for describing causation.
- **NEVER** regress Price on a Quantity or vice versa!

  Why never regress Price on Quantity? Wouldn't this give us a demand curve? Or would it give us a supply curve? Why would we expect to see one and not the other?

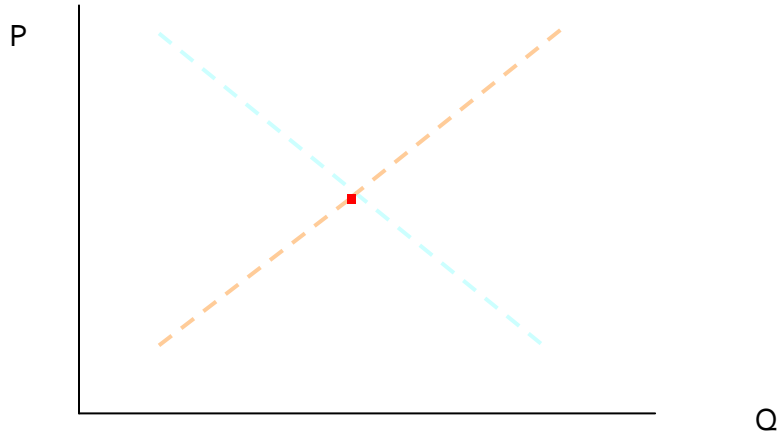
  In actuality, we don't observe either supply curves or demand curves. We only observe the values of price and quantity where the two intersect.

  If both vary randomly then we will not observe a supply or a demand curve. We will just observe a cloud of random points.

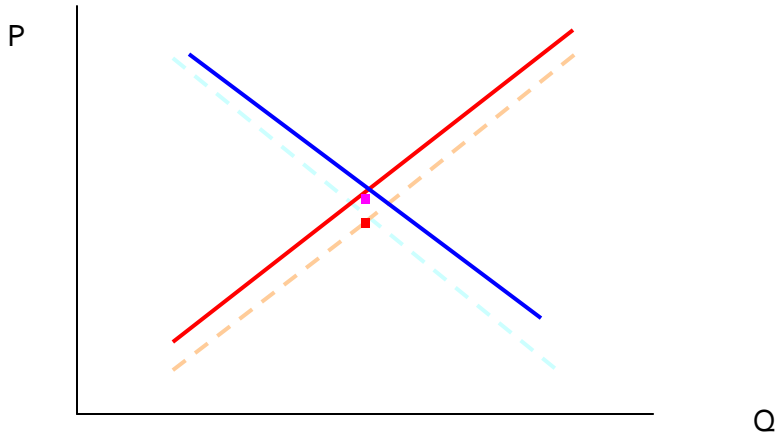
  For example, theory says we see this:



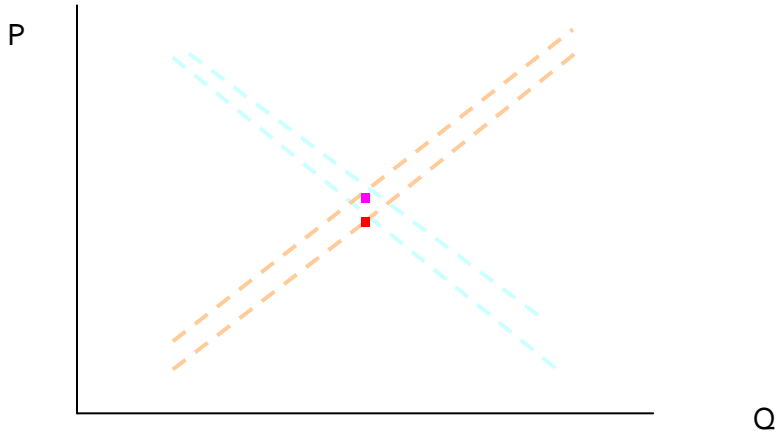
But in the world, we assume the dotted-lines and only actually observe the one intersection, the dot:



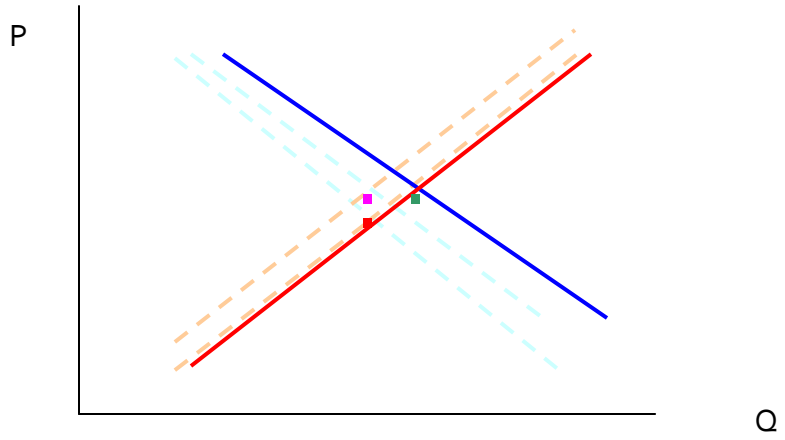
In the next time period, supply and demand shift randomly by a bit, so theory tells us that we now have:



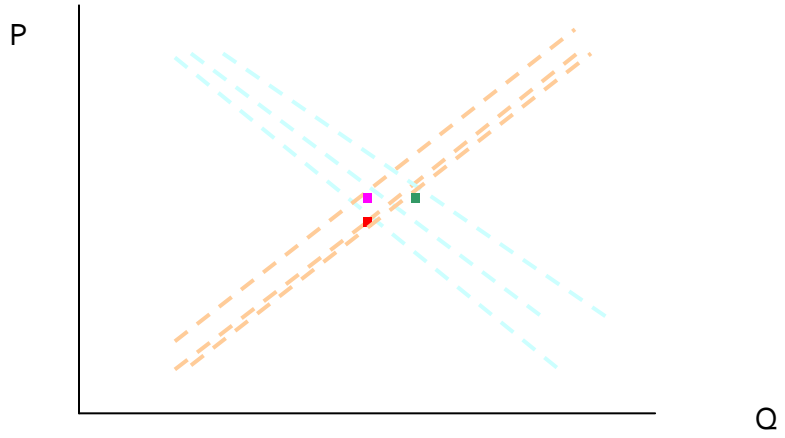
But again we actually now see just two points,



In the third period,



but again, in actuality just three points:



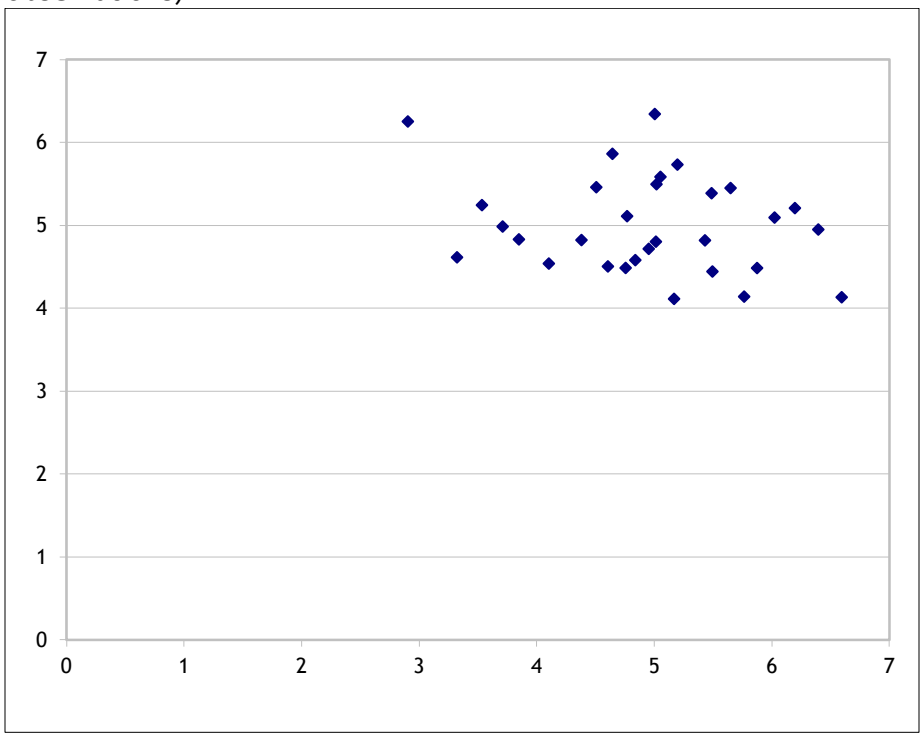
So if we tried to draw a regression line on these three points, we might convince ourselves that we had a supply curve. But do we, really? You should be able to re-draw the lines to show that we could have a down-sloping line, or a line with just about any other slope.

So even if we could wait until the end of time and see an infinite number of such points, we'd still never know anything about supply and demand. This is the problem of endogeneity. The regression is not identified – we could get more and more information but still never learn anything.

We could show this in an Excel sheet, too, which will allow a few more repetitions.

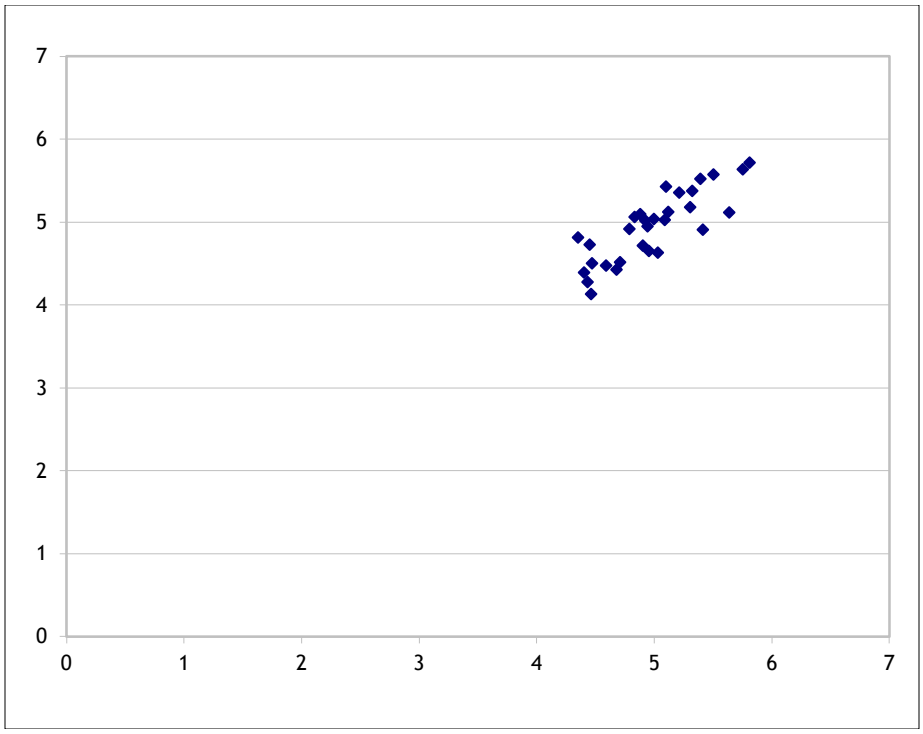
Recall that we can write a demand curve as $P_d = A - BQ_d$ and a supply curve as $P_s = C + DQ_s$, where generally A, B, C, and D are all positive real numbers. In equilibrium $P_d=P_s$ and $Q_d=Q_s$. For simplicity assume that A=10, C=0, and B=D=1. Without any randomness this would be a boring equation; solve to find 10 – Q = Q and Q*=5, P*=5. (You did this in Econ 101 when you were a baby!) If there were a bit of randomness then we could write $P_d = A - BQ_d + \varepsilon_d$ and $P_s = C + DQ_s + \varepsilon_s$. Now the equilibrium conditions tell that $10 - Q + \varepsilon_d = Q + \varepsilon_s$ and so $Q^* = \frac{10 + \varepsilon_d - \varepsilon_s}{2} = 5 + \frac{(\varepsilon_d - \varepsilon_s)}{2}$ and $P^* = 5 + \frac{\varepsilon_d - \varepsilon_s}{2} + \varepsilon_s = 5 + \frac{\varepsilon_d + \varepsilon_s}{2}$.

Plug this into Excel, assuming the two errors are normally distributed with mean zero and standard deviation of one, and find that the scatter looks like this (with 30 observations):



You can play with the spreadsheet, hit F9 to recalculate the errors, and see that there is no convergence, there is nothing to be learned.

On the other hand, what if the supply errors were smaller than the demand errors, for example by a factor of 4? Then we would see something like this:



So now with the supply curve pinned down (relatively) we can begin to see it emerge as the demand curve varies.

But note the rather odd feature: variation in demand is what identifies the supply curve; variation in supply would likewise identify the demand curve. If some of that demand error were observed then that would help identify the supply curve, or vice versa. So sometimes in agricultural data we would use weather variation (that affects the supply of a commodity) to help identify demand.

If you want to get a bit crazier, experiment if the slope terms have errors as well (so $P_d = (A+\varepsilon_a) - (B+\varepsilon_b)Q_d$ and $P_s = (C+\varepsilon_c) + (D+\varepsilon_D)Q_s$).