Lecture Notes 12 – Advanced Topics
Econ 20150, Principles of Statistics
Kevin R Foster, CCNY
Spring 2012

Endogenous Independent Variables are Invalid

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

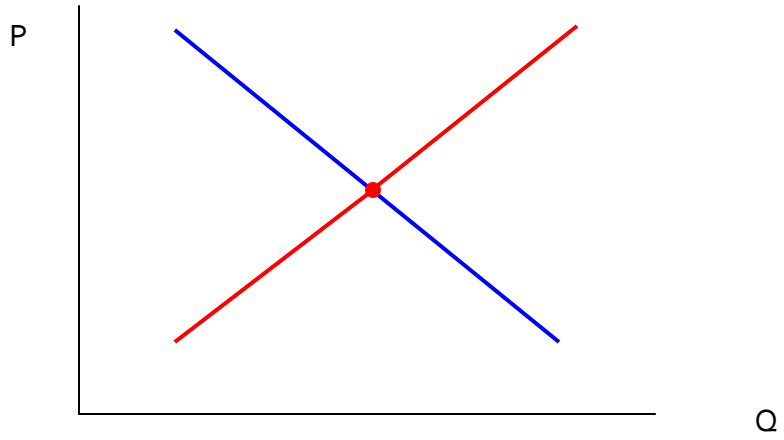
- **NEVER** regress Price on a Quantity or vice versa!
- Endogenous vs. Exogenous variables
    - Exogenous variables are generated from "exo" outside of the model; endogenous are generated from "endo" within the model. Of course this neat binary distinction rarely is matched by the world; some variables 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.

  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?

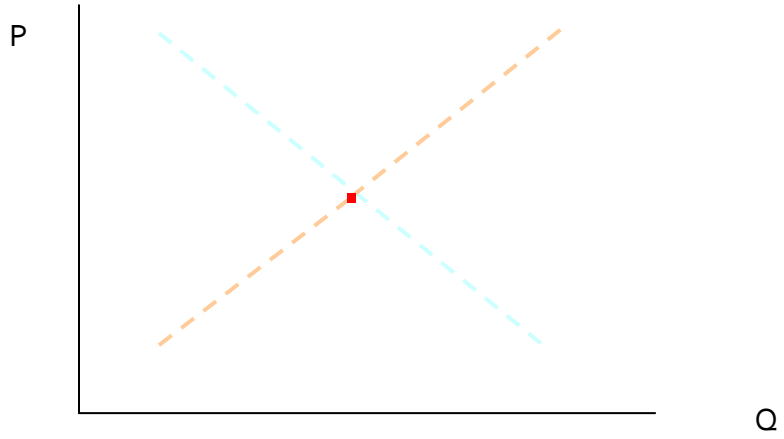
  In actuality, we don't observe either supply curves or demand curves. We only observe the values of price and quantity where the two intersect.

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

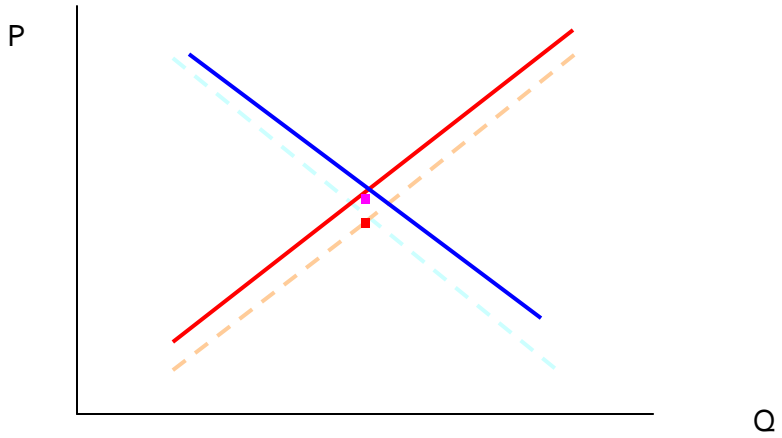
  For example, theory says we see this:



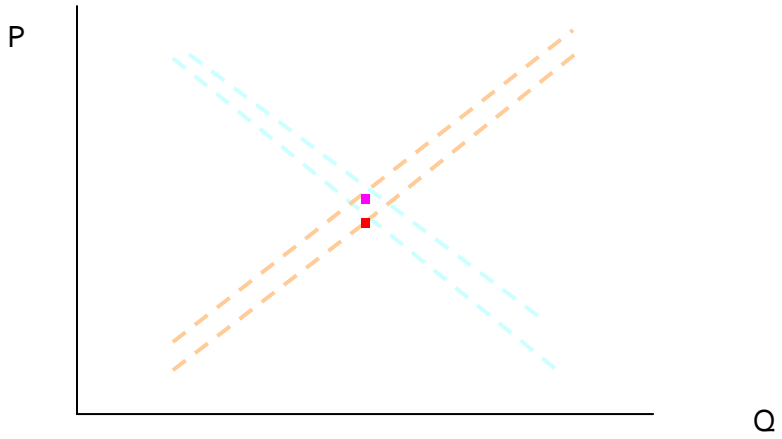
But in the world, we assume the dotted-lines and only actually observe the one intersection, the dot:



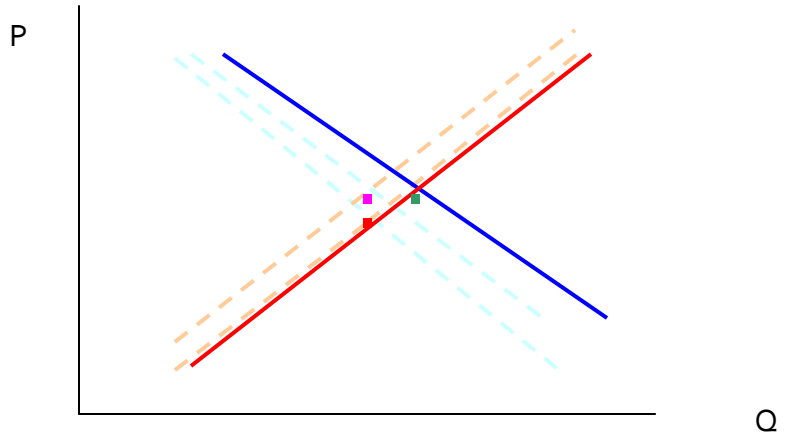
In the next time period, supply and demand shift randomly by a bit, so theory tells us that we now have:



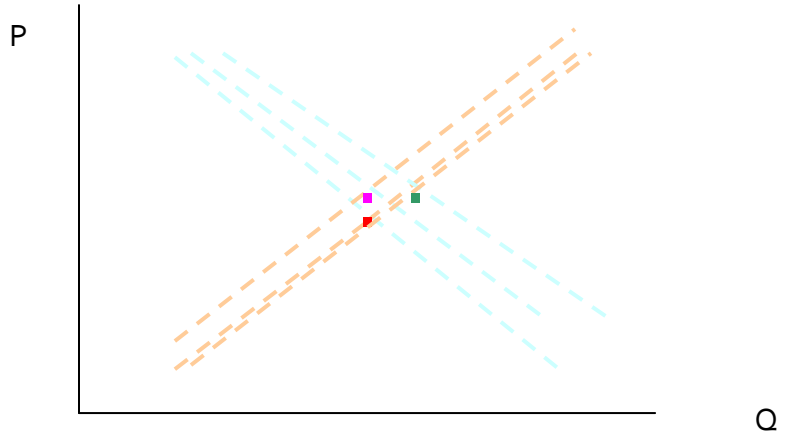
But again we actually now see just two points,



In the third period,



but again, in actuality just three points:



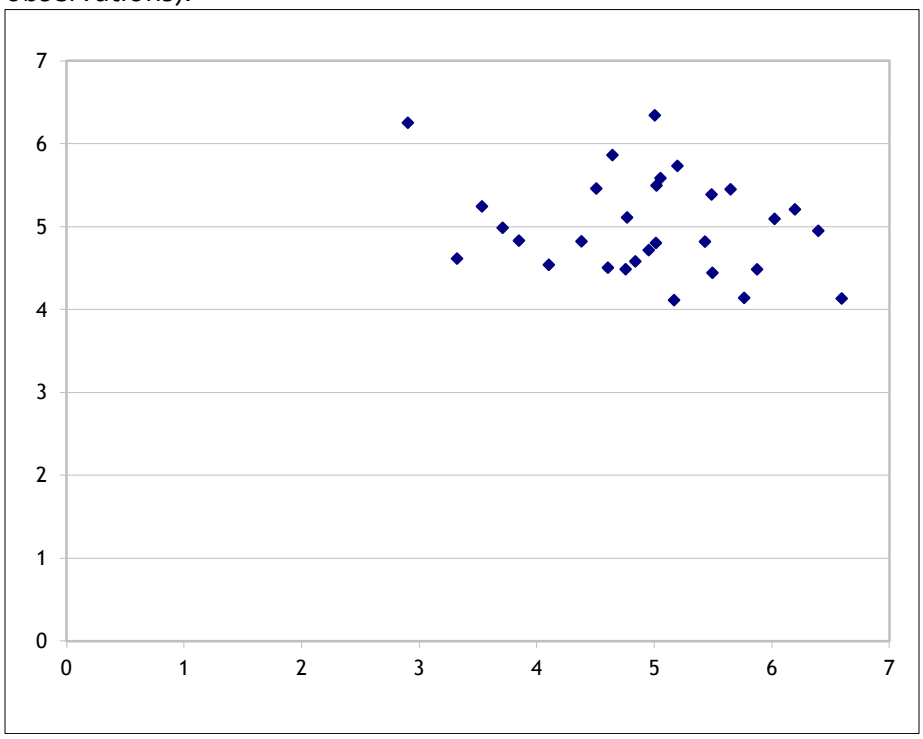
So if we tried to draw a regression line on these three points, we might convince ourselves that we had a supply curve. But do we, really? You should be able to re-draw the lines to show that we could have a down-sloping line, or a line with just about any other slope.

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.

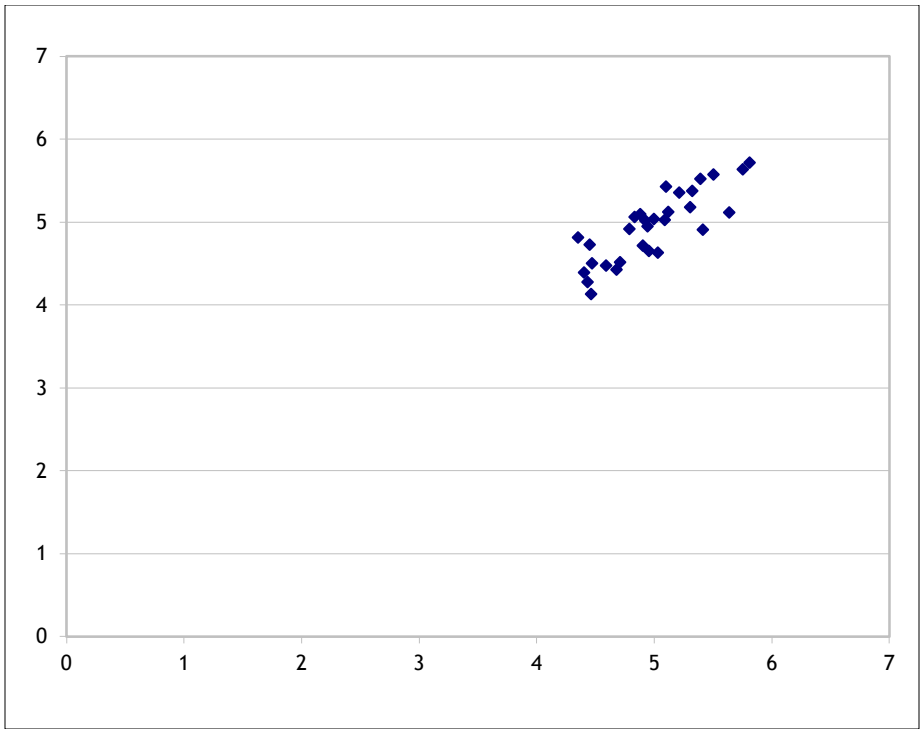
Recall that we can write a demand curve as $P_d = A - BQ_d$ and a supply curve as $P_s = C + DQ_s$, where generally A, B, C, and D are all positive real numbers. In equilibrium $P_d=P_s$ and $Q_d=Q_s$. For simplicity assume that A=10, C=0, and B=D=1. Without any randomness this would be a boring equation; 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 $P_s = (C+\varepsilon_c) + (D+\varepsilon_D)Q_s$).

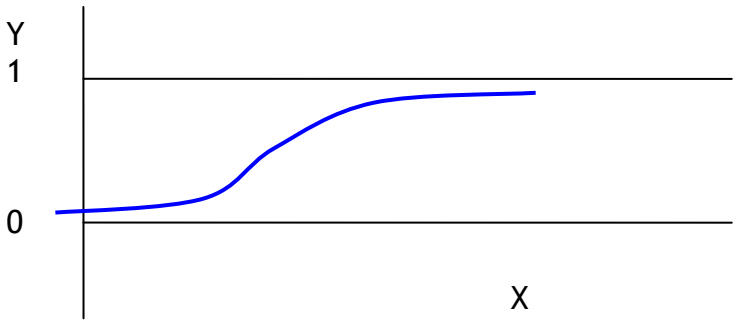
When you have a 0/1 variable as your dependent "y" variable…

…use SPSS "Binary Logistic"

Binary Dependent Variable Models

- Sometimes our dependent variable is continuous, like a measurement of a person's income; sometimes it is just a "yes" or "no" answer to a simple question. A "Yes/No" answer can be coded as just a 1 (for Yes) or a 0 (a zero for "no"). These zero/one variables are called dummy variables or **binary** variables. Sometimes the dependent variable can have a range of discrete values ("How many children do you have?" "Which train do you take to work?") – in this case we have a discrete variable. The binary and continuous variables can be seen as opposite ends of a spectrum.

- We want to explore models where our dependent variable takes on discrete values; we'll start with just binary variables. For example, we might want to ask what factors influence a person to go to college, to have health insurance, or to look for a job; to have a credit card or get a mortgage; what factors influence a firm to go bankrupt; etc.
- Linear Models such as OLS – NFG. These imply predicted values of Y that are greater than one or less than zero!
- Interpret our prediction of Y as being the probability that the Y variable will take a value of one. (Note: remember which value codes to one and which to zero – there is no necessary reason, for example, for us to code Y=1 if a person has health insurance; we could just as easily define Y=1 if a person is uninsured. The mathematics doesn't change but the interpretation does!)
- want to somehow "bend" the predicted Y-value so that the prediction of Y never goes above 1 or below zero, something like:



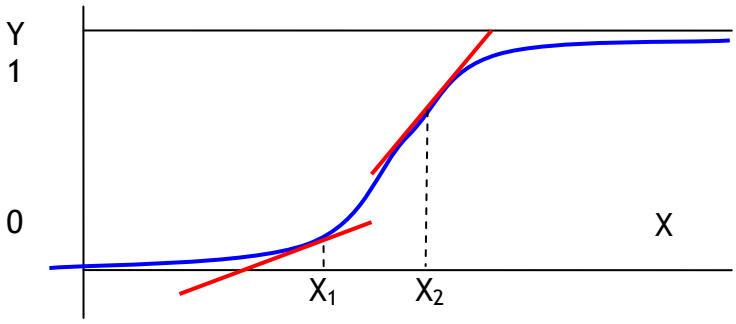
- Logit Model
  - $\Pr(Y=1|X) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$, where $F(z) = \frac{1}{1+e^{-z}}$
  - $\frac{\Delta \Pr}{\Delta X}$ is not constant
- Measures of Fit
  - no single measure is adequate; many have been proposed
  - What probability should be used as "hit"? If the model says there is a 90% chance of Y=1, and it truly is equal to one, then that is reasonable to count as a correct prediction. But many measures use 50% as the cutoff. Tradeoff of false positives versus false negatives – loss function might well be asymmetric
- How to do in SPSS:
  - for logit: `Analyze\Regression\Binary Logistic`…
    - SPSS will generate lots of output; you can safely ignore just about everything in "`Block 0`" and concentrate on "`Block 1`". The last table shows "`Variables in the equation`" with columns for B, S.E., Wald, df, Sig., and Exp(B). The column for B is the estimate of the coefficient and S.E. is its standard error, same as always. But we don't estimate a t-stat but instead a Wald stat (a more complicated formula, don't worry) which combines with df to get a Sig. (a p-value). As usual, if the Sig. (p-value) is less than 0.05 then the variable is significant at the 5% level and you can make confident deductions from it. For now don't

worry if you don't remember all of the details about the difference between t-tests and Wald tests from your stats classes. Just look at the calculated p-value to figure out which coefficients are significant. (Tests of multiple restrictions, which we did for the OLS model, are more complicated here so, again, don't worry about those now.)

- Details of estimation
  - recall that OLS just gives a convenient formula for finding the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ that minimize the sum $\sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}\right)\right)^2$. If we didn't know the formulas we could just have a computer pick values until it found the ones that made that squared term the smallest.
  - similarly the logit coefficient estimates are finding the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ that minimize $\sum_{i=1}^{n}\left(Y_i - f\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}\right)\right)^2$, where the $f(\bullet)$ function is a logistic function.
- **Properly Interpreting Coefficient Estimates**:

Since the slope, $\frac{\Delta Y}{\Delta X} = \frac{\Delta \Pr}{\Delta X}$, the change in probability per change in X-variable, is always changing, the simple coefficients of the linear model cannot be interpreted as the slope, as we did in the OLS model. (Just like when we added a squared term, the interpretation of the slope got more complicated.)

Return to the picture to make this much clearer:



The slope at X_1 is rather low; the slope at X_2 is much steeper.

The effect of the coefficients now interacts with all of the other variables in the model: for example the effect of a person's gender on their probability of having health insurance will depend on other factors like their age and educational level. Women are generally less likely to have their own insurance than men, but how much less? Among young people with very low education, neither men nor women are very likely to be insured; among older

people with very high education both are very likely insured. The biggest difference is toward the middle.

For example, very simple logit estimations on the CPS 2008 dataset gives the following coefficient estimates (I am suppressing notation on significance since it is not important here):

	Logit
female	-0.428
afam	0.220
asian	0.252
Amindian	0.012
Hispanic	-0.028
ed_hs	0.987
ed_smcol	1.180
ed_coll	1.652
ed_adv	1.927
marrd	0.492
divwidsp	0.875
union	1.336
veteran	0.088
immig	-0.277
imm2gen	-0.067
Intercept	-1.303

The probability of having health insurance varies for different socioeconomic groups. We can interpret the signs in a straightforward way: the negative coefficients on the "female" variable indicate that women are less likely to have health insurance. Surprisingly, African-Americans are more likely, along with Asians and Native Americans (although the last is not significant). Hispanics are less likely although this is also not significant.

But how large are these differences? For example, how much less likely to have health care are immigrants? It depends on the other variables. Intuitively, if a person is male, highly-educated, married, and unionized then he's probably insured (being an immigrant would them only slightly less so). So the change in probability associated with immigrant status would be low. At the opposite end, a woman without even a high school diploma, who is single, might already be unlikely to be insured. Immigrant status hardly changes this. Only in the middle will there be a big effect.

We can calculate it straightforwardly, though.

Consider, say, a non-immigrant woman with an advanced degree, whose predicted probability of having health insurance is =

$$f\left(\begin{array}{l}\beta_0+\beta_1 Female+\beta_2 Afam+\beta_3 Asian+\beta_4 Amindian+\beta_5 Hispanic\\+\beta_6 Ed\_hs+\beta_7 Ed\_Scoll+\beta_8 EdColl+\beta_9 Ed\_adv+\beta_{10} Marrd\\+\beta_{11} DivWidS+\beta_{12} Union+\beta_{13} Veteran+\beta_{14} Immig+\beta_{15} Imm2gen+e\end{array}\right)$$

$$= f\left(\begin{array}{l}\beta_0+\beta_1 1+\beta_2 0+\beta_3 0+\beta_4 0+\beta_5 0\\+\beta_6 0+\beta_7 0+\beta_8 0+\beta_9 1+\beta_{10} 0\\+\beta_{11} 0+\beta_{12} 0+\beta_{13} 0+\beta_{14} 0+\beta_{15} 0+e\end{array}\right)$$

Summing the 3 relevant coefficients (the intercept, female, and an advanced degree) gives a logit probability of $f\left(-1.303-0.428+1.927\right)=\frac{1}{1+e^{-(-1.303-0.428+1.927)}}=0.5487$. For an otherwise-identical immigrant woman (also with an advanced degree) the probability is 0.4796, so the change in probability is about 7%.

Compare the change in probabilities for a married male with an advanced degree who is a union member, who is either an immigrant or not. Now the probability of having insurance is, by the logit, 0.9206 for the non-immigrant and 0.8979 for the immigrant, a change of just 2.3%. From the probit the estimated probabilities are 0.9298 for the non-immigrant and 0.9045 for the immigrant, a change of 2.5%. This is because a married male with an advanced degree who is a union member is already highly likely to have health insurance, so the difference of being an immigrant or not makes only a small change compared with the previous example of a female with a high education (but unmarried and not in a union).

If you have a dependent "y" variable that has a few discrete values like 1, 2, 3, 4…

…use SPSS Multinomial Logit.