Lecture Notes 8

Econ B2000, Statistics and Introduction to Econometrics

Kevin R Foster, the City College of New York, CUNY

Fall 2014

Instrumental Variables

- 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. Although Granger Causation from the logical inverse: not-correlate implies not-cause. If knowledge of variable X does not help predict Y, then X does not cause Y.

- 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. (So often need to watch dates – if the X variables are date (t-1) while Y is date t, then the causation is clearer than if all are dated t.) Example: oil prices and economic growth – high oil prices can choke off growth, but lower growth means less demand so lower oil prices.

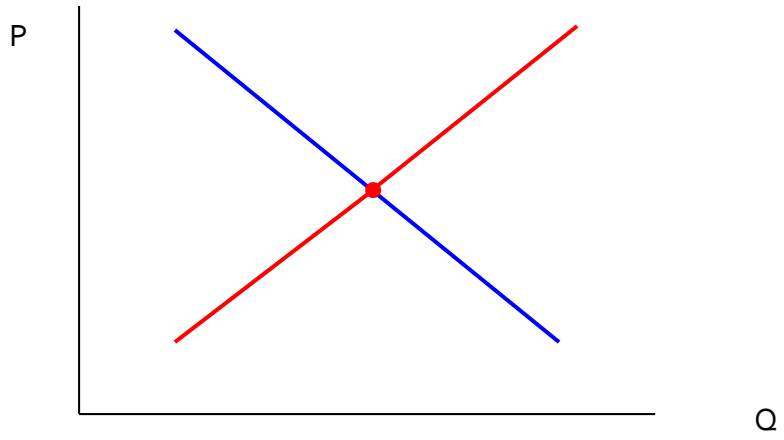
- **NEVER** regress Price on a Quantity or vice versa!

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?

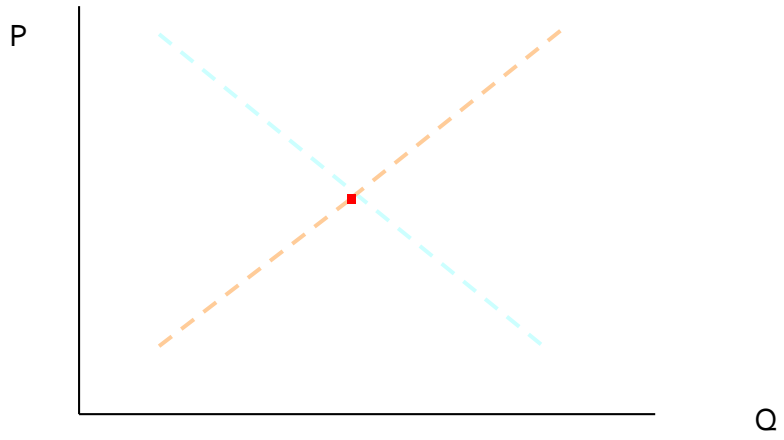
In actuality, we don't observe either supply curves or demand curves. We only observe the values of price and quantity where the two intersect.

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

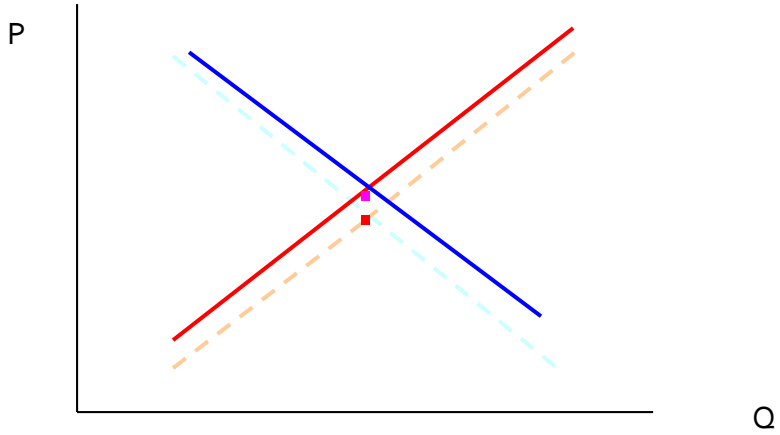
For example, theory says we see this:



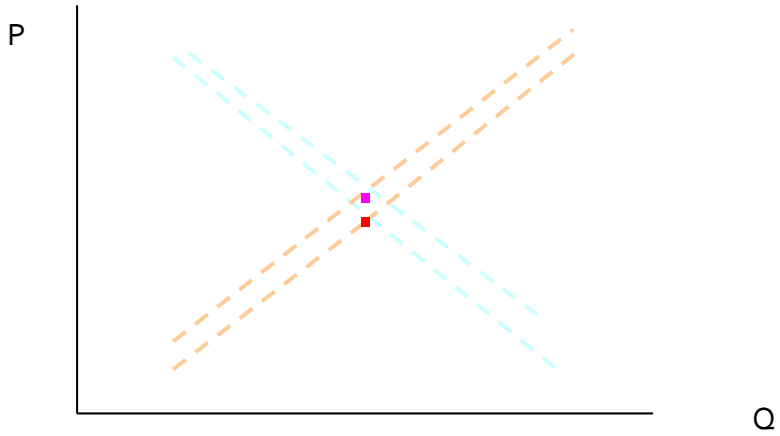
But in the world, we assume the dotted-lines and only actually observe the one intersection, the dot:



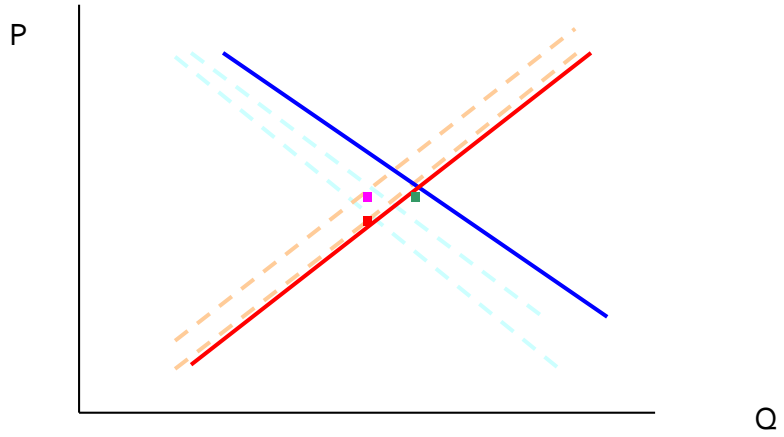
In the next time period, supply and demand shift randomly by a bit, so theory tells us that we now have:



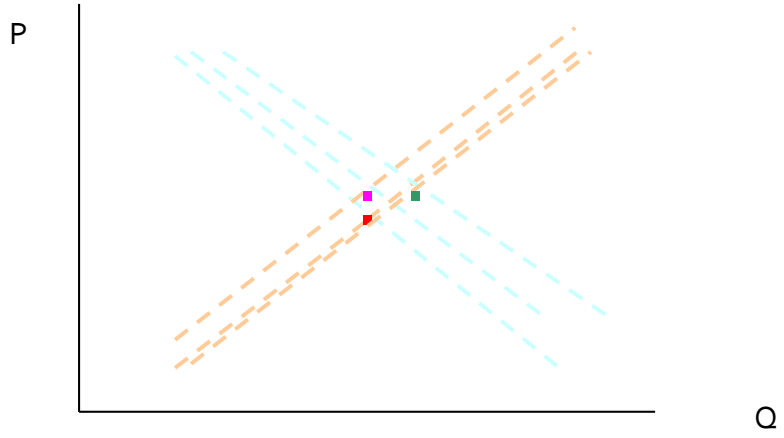
But again we actually now see just two points,



In the third period,



but again, in actuality just three points:



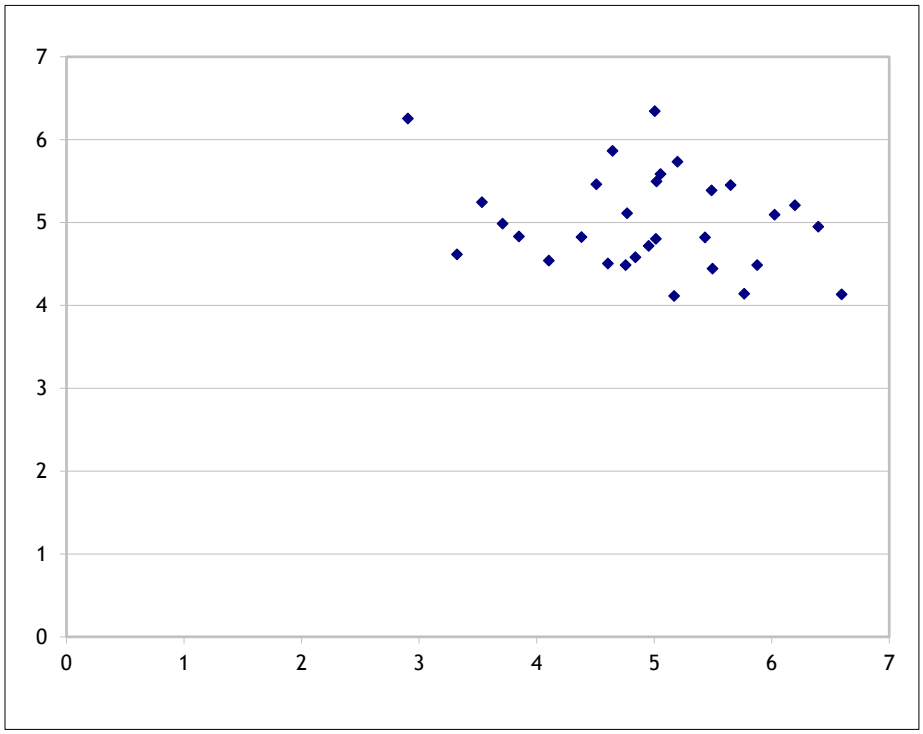
So if we tried to draw a regression line on these three points, we might convince ourselves that we had a supply curve. But do we, really? You should be able to re-draw the lines to show that we could have a down-sloping line, or a line with just about any other slope.

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.

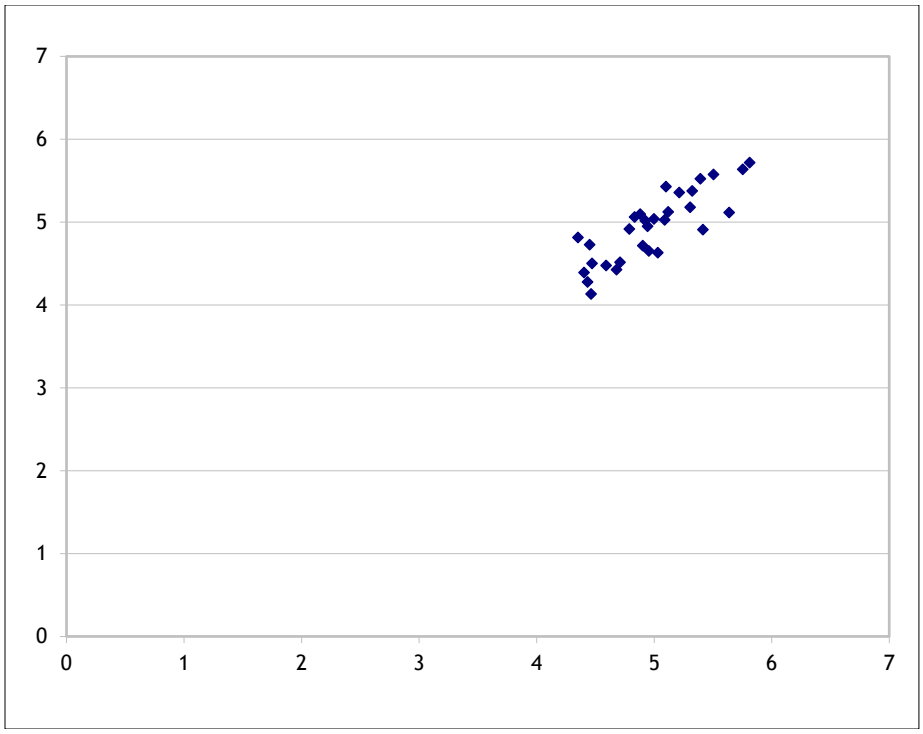
Recall that we can write a demand curve as $P_d = A - BQ_d$ and a supply curve as $P_s = C + DQ_s$, where generally A, B, C, and D are all positive real numbers. In equilibrium $P_d=P_s$ and $Q_d=Q_s$. For simplicity assume that A=10, C=0, and B=D=1. Without any randomness this would be a boring equation;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

Instrumental Variables Regression

- valid instrument, some Z_i for regression $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + u_i$
  - relevance: $corr(Z_i, X_i) \neq 0$ and
  - exogeneity: $corr(Z_i, u_i) = 0$
  - instrument explains X but NOT Y – can be excluded from list of variables explaining Y
- Two-Stage Least Squares (TSLS or 2SLS)
  - $X_i = \pi_0 + \pi_1 Z_i + v_i$, $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - get $X_i = \pi_0 + \pi_1 Z_i$ and regress Y_i on X_i

- $\hat{\beta}_1 = \frac{s_{ZY}}{s_{ZX}}$

- General Case:
  - $$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki} + \hat{\beta}_{k+1} W_{1i} + \hat{\beta}_{k+2} W_{2i} + \ldots + \hat{\beta}_{k+r} W_{ri} + u_i$$
  - X are endogenous regressors
  - W are exogenous regressors
  - $Z_{1i}, Z_{2i}, \ldots Z_{mi}$ are instruments
  - if m>k then "overidentified"; if m=k then just identified; if m<k then unidentified
  - still need:
    - $E\left(u_i \middle| W_{1i}, W_{2i}, \ldots W_{ri}\right) = 0$
    - X, W, Z are all i.i.d. with fourth moments
    - W not perfectly collinear
    - Instrument Relevance and Exogeneity
- Two-Stage Least Squares:
  - regress *X* on *Z* to get $\hat{X}$
  - then regress *Y* on *W* and $\hat{X}$
- Evaluating Instruments in the Real World
  - Weak instruments: check first-stage regression F-stat bigger than 10?
  - Examples:
    - cigarette tax to find effect of price
    - prison capacity in place of jail terms
    - random variation in births for class size
    - geography for heart attack treatment
    - number of immigrants 10 years ago for immigrant increase
    - Mariel boatlift, other policy shifts
    - deployment of police after 9/11 to estimate effects of police on crime
  - Bad examples of poor instruments:
    - weak instrument: month of birth on wage earnings
  - Many bad examples where instruments needed:
    - wage explained by schooling
    - health insurance explained by wage
    - wage explained by weight (discrimination against fat people?) vs wage explained by race/ethnicity (discrimination against minorities)
- Heckman 2-step for 2-part questions: first, "yes or no?"; next "how much?" Like 2SLS but first stage is a probit (we'll do that later)! Again need an exclusion restriction, some variable that explains the first step but not the second.
- Two-Stage Least Squares in SPSS:
  - run first-stage regression, save the predicted values
  - use predicted values in the second-stage prediction

Instrumental Variables Regression in R

There was a recent paper in the journal *Economic Inquiry*, by Cesur & Kelly (2013), "Who Pays the Bar Tab? Beer Consumption and Economic Growth in the United States," which concluded that beer consumption was bad for economic growth. I got data from the Brewer's Almanac, provided online by the Beer Institute (beerinstitute.org) and the Bureau of Economic Analysis (bea.gov). This is not quite the same data that the paper used (less complete) but it gives a flavor (bad pun) of the results.

You can download the R data from InYourClass. Then run this regression,

```
regression1 <- lm(growth_rates ~ beer_pc + gdp_L +
as.factor(st_fixedeff))
summary(regression1)
```

Where the growth rate of each state's GDP is a function of per-capita beer consumption, a lag of state GDP (reflecting the general idea that poorer states might grow faster), as well as state fixed effects (each state has its own intercept). This shows a positive and statistically significant coefficient on per-capita beer consumption. So beer is good for growth?!

As Homer Simpson put it, "To alcohol! The cause of – and solution to – all of life's problems." That circularity of causation makes the statistics more complicated.

Richer people have more money to buy everything including beer, so economic growth might cause beer consumption. One way out, suggested by the article authors, is to use an instrument for beer consumption – the tax on beer. This is a plausible instrument since it likely causes changes in beer consumption (higher price, lower consumption, y'know the demand curve) but it unlikely to be affected by economic growth. So estimate an instrumental variables equation,

```
iv_reg1 <- lm(beer_pc ~ beertax)
summary(iv_reg1)
```

And see that indeed there is a negative coefficient (hooray for demand curves!) although it is certainly a weak instrument (R^2 less than 1%). Use the predicted value of beer consumption per capita as an instrument in the regression in place of the endogenous variable,

```
pred_beer <- predict(iv_reg1)
iv_reg2 <- lm(growth_rates ~ pred_beer + gdp_L +
     as.factor(st_fixedeff))
summary(iv_reg2)
```

To note that now beer consumption seems to have negative effects on economic growth (only significant at 10% level; the article adds some other variables to get it significant). I put some other variables in the dataset that you might play with – see if you can find the opposite result! (R code from a simple summary at http://www.r-bloggers.com/a-simple-instrumental-variables-problem/)

Finally note that you can use the AER package and ivreg() procedure for better results, since these estimated standard errors won't be quite right – but that's just fine-tuning.

The basic idea of instrumental variables is that if we have some regression,

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

But X and Y are endogenous, then suppose we had some variable Z, which is uncorrelated with Y but still explains X, then we can make a supplementary regression,

$$X = \gamma_0 + \gamma_1 Z + u,$$

And get $\hat{X}$, the predicted values from that regression, then do the original regression as

$$Y = \beta_0 + \beta_1 \hat{X} + \varepsilon,.$$

Measuring Discrimination – Oaxaca Decompositions:

(much of this discussion is based on Chapter 10 of George Borjas' textbook on Labor Economics)

The regressions that we've been using measured the returns to education, age, and other factors upon the wage. If we classify people into different groups, distinguished by race, ethnicity, gender, age, or other categories, we can measure the difference in wages earned. There are many explanations but we want to determine how much is due to discrimination and how much due to different characteristics (chosen or given).

Consider a simple model where we examine the native/immigrant wage gap, and so measure $\overline{w}_N$, the average wages that natives get, and $\overline{w}_M$, the average wages that immigrants get. The simple measure, $\overline{w}_N - \overline{w}_M$, of the wage gap, would not be adequate if natives and migrants differ in other ways, as well.

Consider the effect of age. Theory implies that people choose to migrate early in life, so we might expect to see age differences between the groups. And of course age influences the wage. If natives and immigrants had different average wages solely because of having different average ages, we would conclude very different reasons for this than if the two groups had identical ages but different wages.

For example, in a toy-sized 1000-observation subset of CPS March 2005 data, there are 406 natives and 77 immigrants workers with non-zero wages. The natives averaged wage/salary of $37,521 while the immigrants had $32,507. The average age of the natives was 39.5; the average age of the immigrants was 42.1. We want to know how much of the difference in wage can be explained by the difference in age.

Consider a simple model that posits different simple regressions for natives and immigrants:

$$w_N = \beta_{0,N} + \beta_{1,N} Age + \varepsilon$$

$$w_M = \delta_{0,M} + \delta_{1,M} Age + \varepsilon \,.$$

We know that average wages for natives depend on average age of natives, $A\bar{g}e_N$:

$$\bar{w}_N = \beta_{0,N} + \beta_{1,N} A\bar{g}e_N$$

and for immigrants as well, wages depend on immigrants' average age, $A\bar{g}e_M$:

$$\bar{w}_M = \delta_{0,M} + \delta_{1,M} A\bar{g}e_M \,.$$

The difference in average wages is:

$$\bar{w}_N - \bar{w}_M = \left(\beta_{0,N} + \beta_{1,N} A\bar{g}e_N\right) - \left(\delta_{0,M} + \delta_{1,M} A\bar{g}e_M\right)$$

but we can add and subtract the cross term, , $\delta_{1,M} A\bar{g}e_N$ to get:

$$\bar{w}_N - \bar{w}_M = \left(\beta_{0,N} - \delta_{0,M}\right) + \left(\beta_{1,N} - \delta_{1,M}\right) A\bar{g}e_N + \delta_{1,M}\left(A\bar{g}e_N - A\bar{g}e_M\right) .$$

Each term can be interpreted in different ways. The first difference, $\left(\beta_{0,N} - \delta_{0,M}\right)$, is the difference in intercepts, the parallel shift of wages for all ages. The second, $\left(\beta_{1,N} - \delta_{1,M}\right) A\bar{g}e_N$, is the difference in how the skills are rewarded: if everyone in the data were to have the same age, immigrants and natives would still have different wages due to these first two factors. The third is $\delta_{1,M}\left(A\bar{g}e_N - A\bar{g}e_M\right)$, which gives the difference in wage attributable only to differences in average age (even if those were rewarded equally). The first two are generally regarded as due to discrimination while the last is not.

The basic framework can be extended to other observable differences: in years of education, experience, or the host of other qualifications that affect people's wages and salaries.

From our discussions of regression models, we realize that the two equations above could be combined into a single framework. If we define an immigrant dummy variable as M_i, which is equal to one if individual *i* is an immigrant and zero if that person is native born, we can write a regression model as:

$$w_i = \beta_0 + \beta_1 Age_i + \beta_2 M_i + \beta_3 M_i Age_i + \varepsilon_i ,$$

where wages for natives depend on only β_0 and β_1, while the immigrant coefficients are $\delta_{0,M} = \beta_0 + \beta_2$ and $\delta_{1,M} = \beta_1 + \beta_3$. We construct $\bar{w}_N = \hat{\beta}_0 + \hat{\beta}_1 \overline{Age}_N$ and $\bar{w}_M = \hat{\beta}_0 + \hat{\beta}_2 + \left(\hat{\beta}_1 + \hat{\beta}_3\right) \overline{Age}_M$ so the Oaxaca decomposition is now:

$$\bar{w}_N - \bar{w}_M = -\beta_2 - \beta_3 A\bar{g}e_N + \left(\beta_1 + \beta_3\right)\left(A\bar{g}e_N - A\bar{g}e_M\right) .$$

We note that unobserved differences in quality of skills can be measured as instead being due to discrimination. In our example, suppose that natives get a greater salary as they age due to the skills which they amass, but immigrants who have language difficulties learn new skills more slowly. In this case, older natives would earn more, increasing the returns to aging. This would be reflected as lower coefficients on age for immigrants than natives, and so evidence of discrimination. If we had information on English-language ability (SAT, TOEFL or GRE scores, maybe?), then the regression would show that a lack of those skills led to lower wages – no longer would it be measured as evidence of discrimination.

But this elides the question of how people gain the "skills" measured in the first place. If a degree from a foreign university gets less reward than a degree from an American university, is this entirely due to discrimination? What fraction of the wage differential arises from skill differences? In the US, African-American and Hispanic children tend to go to lower-quality schools (as measured by test scores or teacher qualifications). The lower subsequent wages might not be due to labor market discrimination (if firms rationally pay less for lower skills) but still be due to societal discrimination.

Consider the sort of dataset that we've been working with. Regressing Age, an Immigrant dummy, and an Age-Immigrant interaction on Wage provides the following coefficient estimates (for the same sub-sample as before):

$$w_i = 7437 + 762.62 Age_i + 20,663.29 M_i - 658.06 Age_i M_i + \varepsilon_i$$

where the immigrant dummy is actually positive (neither the immigrant dummy nor the immigrant-age interaction term are statistically significant, but I ignore that for now). With the average ages from above (natives 39.5 years old; immigrants 42.1), we calculate the gap in average predicted wages (natives are predicted to make an average wage of $37,561; immigrants to make $32,502) is $5058.08. The two first terms in the Oaxaca decomposition, relating to unexplained factors such as "discrimination" $-\hat{\beta}_2 - \hat{\beta}_3 \overline{Age}_N$ account for $5329.95, while the difference in age accounts for just -$271.86 (a negative amount) – this means that the ages actually imply that natives and immigrants ought to be closer in wages so they are even farther apart. We might reasonably believe that much of this difference reflects omitted

factors (and could list out the important omitted factors); this is intended merely as an exercise.

Adding these additional variables is easy; I show the case for two variables but the model can be extended to as many variables as are of interest. Next consider a more complicated model, where now wages depend on Age and Education, so the two regressions for natives and immigrants are:

$$w_N = \beta_{0,N} + \beta_{1,N} Age + \beta_{2,N} Educ + \varepsilon$$

$$w_M = \delta_{0,M} + \delta_{1,M} Age + \delta_{2,M} Educ + \varepsilon .$$

We know that average wages for natives depend on average age and education of natives, $A\bar{g}e_N, Ed\bar{u}c_N$:

$$\bar{w}_N = \beta_{0,N} + \beta_{1,N} A\bar{g}e_N + \beta_{2,N} Ed\bar{u}c$$

and for immigrants as well, wages depend on immigrants' average age, $A\bar{g}e_M, Ed\bar{u}c_M$:

$$\bar{w}_M = \delta_{0,M} + \delta_{1,M} A\bar{g}e_M + \delta_{2,M} Ed\bar{u}c .$$

The difference in average wages is:

$$\bar{w}_N - \bar{w}_M = \left(\beta_{0,N} + \beta_{1,N} A\bar{g}e_N + \beta_{2,N} Ed\bar{u}c_N\right) - \left(\delta_{0,M} + \delta_{1,M} A\bar{g}e_M + \delta_{2,M} Ed\bar{u}c_M\right)$$

but we can add and subtract the cross terms, $\delta_{1,M} A\bar{g}e_N + \delta_{2,M} A\bar{g}e_N$ to get:

$$\bar{w}_N - \bar{w}_M = \left(\beta_{0,N} - \delta_{0,M}\right) + \left(\beta_{1,N} - \delta_{1,M}\right) A\bar{g}e_N + \delta_{1,M}\left(A\bar{g}e_N - A\bar{g}e_M\right) + \left(\beta_{2,N} - \delta_{2,M}\right) Ed\bar{u}c_N + \delta_{2,M}\left(Ed\bar{u}c_N - Ed\bar{u}c_M\right)$$

.

Again, the two terms showing the difference in average levels of external factors, $\left(A\bar{g}e_N - A\bar{g}e_M\right)$ and $\left(Ed\bar{u}c_N - Ed\bar{u}c_M\right)$, are "explained" by the model while the other terms showing the difference in the coefficients are "unexplained" and could be considered as evidence of discrimination.

Exercises:

1. Do the above analysis on the current CPS data.

2. If instead you used log wages, but still kept just age as the measured variable, is your answer substantially different than in the previous question? (Note that

the answers are in different units, so you have to think about how to convert the two answers.)

3. Consider other measures of skills, such as schooling and whatever other factors you consider important. How does this new regression change the Oaxaca decomposition?

4. What is the maximum fraction of wage difference that you can find (with different independent variables and regression specifications), related to discrimination? The minimum?

References:

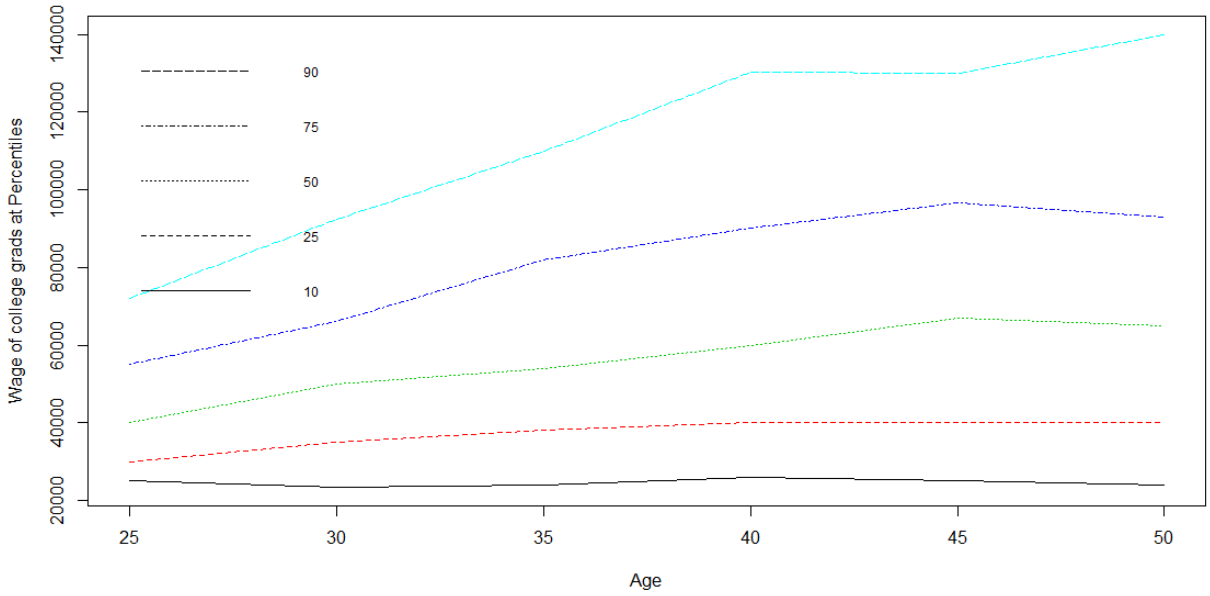
Borjas, George (2003). *Labor Economics*.

Oaxaca, Ronald (1973). "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14(3).

Quantile Regression

If you recall our discussion of heteroskedasticity in things like the Age-Wage relationship, there is a well-known tendency for younger workers to have more compressed earnings, which then fan out as people get older.

For example, if we use the 2013 CPS data, we can look at people aged 25-55 who are working full time for most of the year and, even if we focus on a single educational group, for example those with a 4-year degree, we can see the spread here:



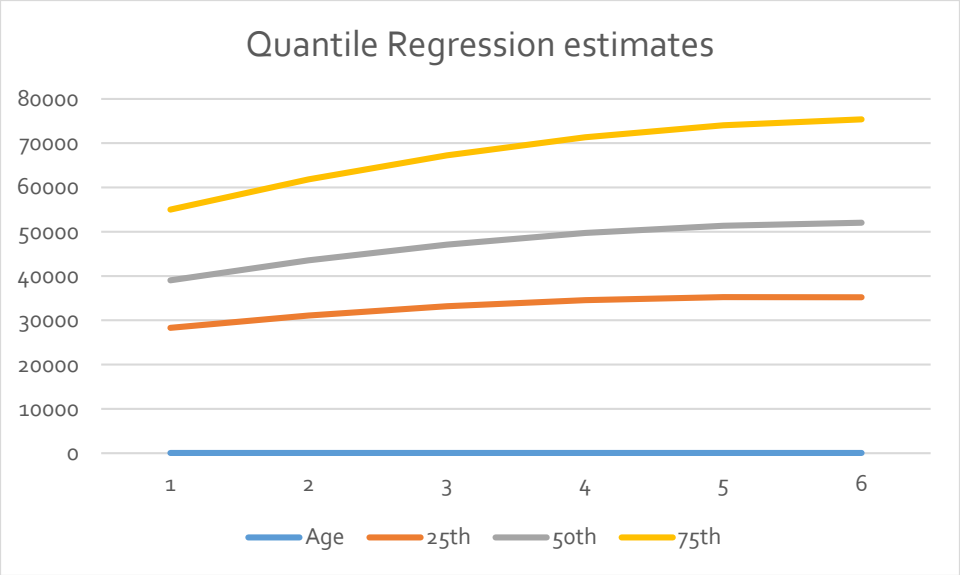
So the median worker saw a steady rise in wage: 30-yr-olds made $50,000 while 50-yr-olds made about $65,000; but those in the 25th percentile went from $35,000 at age 30 to $40,000 by 50; those in the 75th percentile went from $66,000 to $93,000.

One way to model these different results, for different percentiles, is with a quantile regression (mostly due to Roger Koenker), which uses a familiar regression framework to explain various percentiles.

In R this couldn't be easier: just use the "quantreg" package and call the `rq()` function instead of `lm()`. (Note that it's `rq` not `qr`; if you've done linear algebra you'll recall the QR matrix decomposition.)

```
p_tiles <- c(0.1, 0.25, 0.5, 0.75, 0.9)
quantreg1 <- rq(WSAL_VAL ~ Age + female + AfAm + Asian +
     Amindian + race_oth + Hispanic + educ_hs + educ_smcoll +
     educ_as + educ_bach + educ_adv + married + divwidsep +
     union_m + veteran + immigrant + immig2gen,  tau=p_tiles,
     data = dat_use)
summary(quantreg1)
plot(quantreg1)
```

Details are in the R file, cps3.R. This estimates age-wage profiles like this (for women with a 4-year degree):



Which shows the spread.

Non-Parametric Regression

Instead of assuming a functional form – that the age-wage profile is linear, or quadratic, or cubic, or whatever … just let the data determine the wiggles in the function.

Details in R program.

```
restrict2 <- as.logical(dat_use$educ_bach)
data3 <- subset(dat_use, restrict2)
NN <- length(data3$WSAL_VAL)
restrict3 <- as.logical(round(runif(NN,min=0,max=0.75)))
data4 <- subset(data3, restrict3)
library(np)
# note that this is rather computationally intensive!
model_nonparametric1 <- npreg(WSAL_VAL ~ Age, regtype = "ll", bwmethod
      = "cv.aic", gradients = TRUE, data = data4)
summary(model_nonparametric1)
npsigtest(model_nonparametric1)
plot(data4$Age,data4$WSAL_VAL, xlab = "age", ylab = "wage", cex=.1)
lines(data4$Age, fitted(model_parametric1), lty = 2, col = "red")
lines(data4$Age, fitted(model_nonparametric1), lty = 1, col = "blue")
```

A linear regression gives the expected value of Y given the values of X, under restriction that this expected value is a linear function. Quantile regression gives expected quantile of Y given X (again as a linear function). Nonparametric regression gives expected value of Y given X, subject to smoothness constraint (not linearity but still something).

Binary Dependent Variable Models

(Stock & Watson Chapter 9)

- 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"

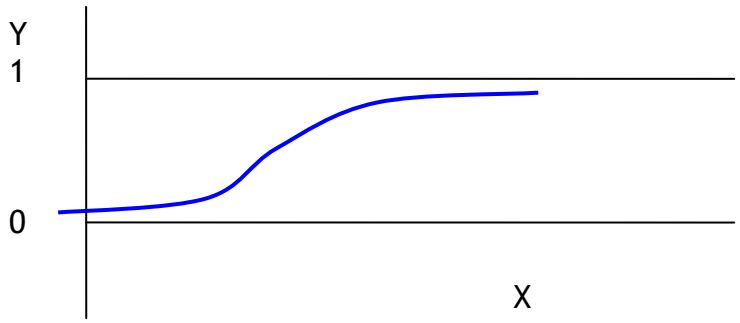
answer can be coded as just a 1 (for Yes) or a 0 (a zero for "no"). These zero/one variables are called dummy variables or **binary** variables. Sometimes the dependent variable can have a range of discrete values ("How many children do you have?" "Which train do you take to work?") – in this case we 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:



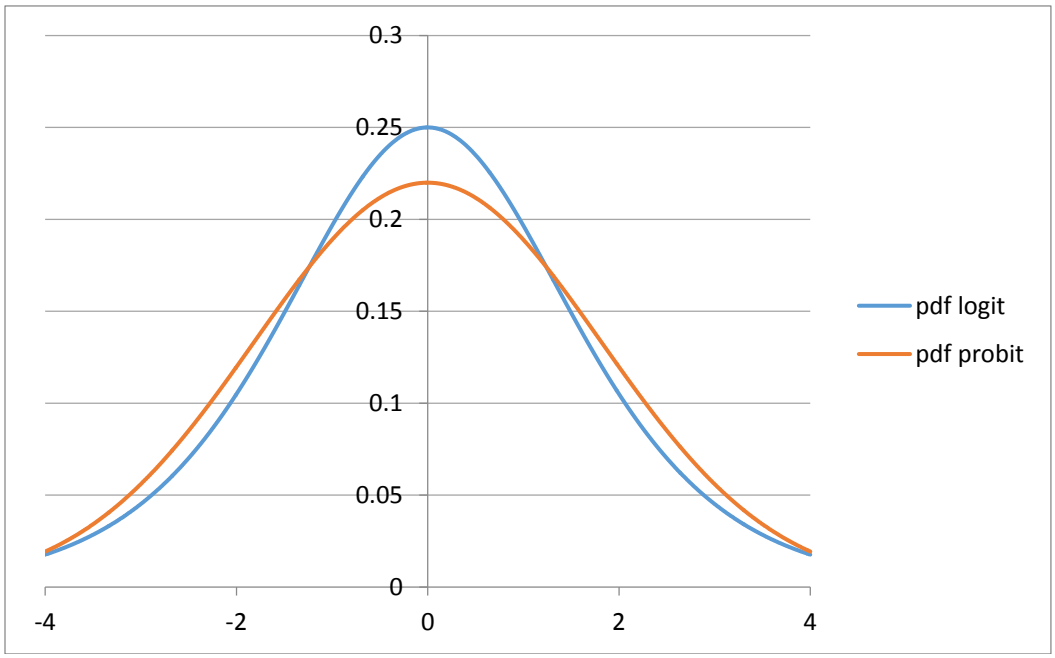
- Probit Model
  - $\Pr(Y=1|X)=\Phi(\beta_0+\beta_1X_1+\beta_2X_2)$ where $\Phi(\square)$ is the cdf of the standard normal
  - $\frac{\Delta \Pr}{\Delta X}$ is not constant

- Logit Model
  - $\Pr(Y=1|X)=F(\beta_0+\beta_1X_1+\beta_2X_2)$, where $F(z)=\frac{1}{1+e^{-z}}$

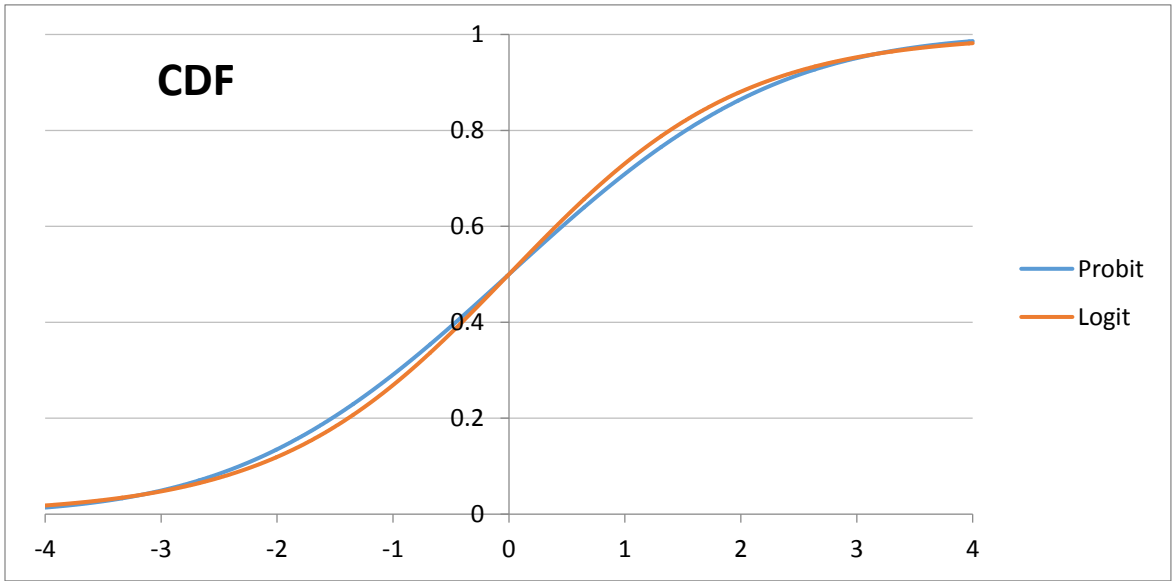
- $\frac{\Delta \Pr}{\Delta X}$ is not constant

- differences (Excel sheet: probit_logit_compare.xls)

Clearly the differences are rather small; it is rare that we might have a serious theoretical justification for one specification rather than the other.



(Note that the logit function given above has standard error of $\frac{\pi}{\sqrt{3}}$ so in the plots I scaled the probit by this factor).



- 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

Probit/Logit in R

For a logit estimation, just

```
regn_logit1 <- glm(Y ~ X1 + X2, family = binomial, data = data1)
```

for a probit estimation

```
regn_probit1 <- glm(Y ~ X1 + X2, family = binomial (link = 'probit'),
data = data1)
```

Example with CPS data

```
model_logit1 <- glm(health_ins ~ Age + I(Age^2) + female + AfAm +
      Asian + Amindian + race_oth + Hispanic + educ_hs + educ_smcoll +
      educ_as + educ_bach + educ_adv + married + divwidsep + union_m +
      veteran + immigrant + immig2gen, family = binomial, data =
      dat_use_hi)

summary(model_logit1)
regn_probit1 <- glm(health_ins ~ Age + female + AfAm + Asian +
      Amindian + race_oth + Hispanic + educ_hs + educ_smcoll + educ_as
      + educ_bach + educ_adv + married + divwidsep + union_m + veteran
+ immigrant + immig2gen, family = binomial (link = 'probit'), data =
      dat_use_hi)
summary(regn_probit1)
```

Then the estimation results from "`summary()`" should be familiar. The interpretation is also essentially unchanged: look at the individual t-statistics (formed by dividing coefficient estimates from standard errors) then get a p-value from that.

- 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 a probit or logit coefficient estimates are finding the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ that minimize $\sum_{i=1}^{n}\left(Y_i - f\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}\right)\right)^2$, whether the $f(\square)$ function is a normal c.d.f. or a logit c.d.f.

- Maximum Likelihood (ML) is a more sophisticated way to find these coefficient estimates – better than just guessing randomly.

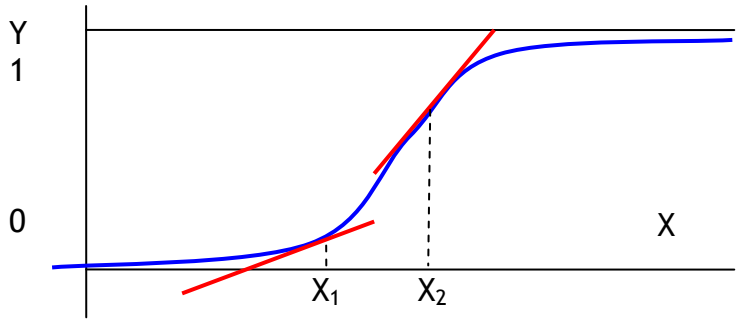
- For example the likelihood of any particular value from a normal distribution is the p.d.f., $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. If we have 2 independent observations, X_1, X_2 from a distribution that is known to be normally distributed with variance of 1 (to keep the math easy) then the joint likelihood is $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_1-\mu)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_2-\mu)^2}$. We want to find a value of μ that maximixes that function. This is an ugly function but we could note that any value of μ that maximizes the natural log of that function will also maximize the function itself (since $\ln(\square)$ is monotonic) so we take logs to get $\ln\left(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(X_1-\mu)^2 - \frac{1}{2}(X_2-\mu)^2$. Take the derivative with respect to μ and set it equal to zero to get $(X_1-\mu)+(X_2-\mu)=0$ so that $\mu = \frac{(X_1+X_2)}{2}$. You should be able to see that starting with n observations would get us $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ so the average is also the maximum-likelihood estimator. A maximum-likelihood estimator could be similarly derived in cases where we don't know the variance (interestingly, that ML estimator of the standard error divides by *n* not *(n – 1)* so it is biased but consistent).

- Maximizing the likelihood of the probit model is one or two steps more complicated but not different conceptually. Having a likelihood function with a first and second derivative makes finding a maximum much easier than the random hunt.

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 and probit estimations on the CPS 2013 dataset (R program shows this in detail) gives the following coefficient estimates (I am suppressing notation on significance since it is not important here):

	coefficient estimates	
	logit	probit
(Intercept)	-0.37783	-0.2473
Age	0.002625	0.002951
I(Age^2)	0.000133	0.000057
female	-0.13458	-0.07423
AfAm	-0.49067	-0.2879
Asian	0.295029	0.1695
Amindian	-0.68546	-0.4059
race_oth	-0.1998	-0.1172
Hispanic	-0.40528	-0.2429
educ_hs	0.84353	0.5237
educ_smcoll	1.215126	0.7426
educ_as	1.54497	0.9321
educ_bach	2.146008	1.254
educ_adv	2.536002	1.444
married	0.602157	0.3499
divwidsep	-0.16488	-0.09745
union_m	1.407863	0.7217
veteran	-0.18023	-0.1157
immigrant	-0.68214	-0.3973

immig2gen 0.071965 0.03768

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. African-Americans are less likely, along with Hispanics and Native Americans. Educational qualifications are positive and get larger.

But how large are these differences? For example, how much less likely to have health insurance are immigrants? It depends on the other variables. Intuitively, if a person is male, highly-educated, and married then he's probably insured (being an immigrant would him only slightly less so). So the change in probability associated with immigrant status would be low. At the opposite end, a woman without a high school diploma, who is single, is 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 30-yr-old non-immigrant African-American woman with an advanced degree, whose predicted probability of having health insurance is

$$= f\left(\begin{array}{c}\beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Female + \beta_4 Afam + \\ \beta_5 Asia + \beta_6 NativeAm + \beta_7 RaceOth + \beta_8 Hisp + \\ \beta_9 EdHS + \beta_{10} EdSmC + \beta_{11} EdAS + \beta_{12} Ed4 + \beta_{13} EdAdv \\ +\beta_{14} Marr + \beta_{15} DivWidSep + \beta_{16} Union + \beta_{17} Vet \\ +\beta_{18} Immig + \beta_{19} Imm2g + e\end{array}\right)$$

$$= f\left(\begin{array}{c}\beta_0 \cdot 1 + \beta_1 \cdot 30 + \beta_2 \cdot 30^2 + \beta_3 \cdot 1 + \beta_4 \cdot 1 + 0 + 0 \\ +0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 \\ +\beta_{16} \cdot 0 + \beta_{17} \cdot 0 + \beta_{18} \cdot 0 + \beta_{19} \cdot 0 + \beta_{20} \cdot 0 + \beta_{13} \cdot 1 \\ +0 \ldots\end{array}\right)$$

Summing the relevant coefficients (the intercept, female, and an advanced degree) gives a logit probability of

$$= f(-.378 + .079 + .120 - .135 - .491 + 2.536)$$

$$= \frac{1}{1+e^{-(-.378+.079+.120-.135-.491+2.536)}}$$

Which is 85.0%. For an otherwise-identical immigrant woman (also with an advanced degree) the probability is 0.74, so the change in probability is about 11 percentage points.

Comparing the probit estimates, we would just change the functional form and use the normal cdf instead of the logit function, so again from:

$$= f\left(\begin{array}{c}\beta_0 \cdot 1 + \beta_1 \cdot 30 + \beta_2 \cdot 30^2 + \beta_3 \cdot 1 + \beta_4 \cdot 1 + 0 + 0 \\ +0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 \\ +\beta_{16} \cdot 0 + \beta_{17} \cdot 0 + \beta_{18} \cdot 0 + \beta_{19} \cdot 0 + \beta_{20} \cdot 0 + \beta_{13} \cdot 1 \\ +0 \ldots\end{array}\right)$$

$$= f(-.247 + .089 + .051 - .074 - .288 + 1.444)$$

$$= pnorm(-.247 + .089 + .051 - .074 - .288 + 1.444) \text{ (in R)}$$

and find a probability for a non-immigrant woman as 0.835 and the immigrant woman to be 0.718, with a difference of 11.7 percentage points. These estimates from the logit and probit are very close.

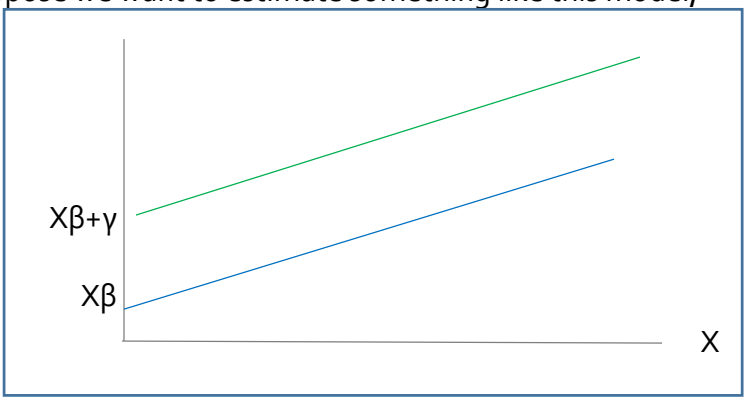
Compare the change in probabilities for a divorced 45-yr-old white male without any degree, who is either an immigrant or not. Now the probability of having insurance is, by the logit, 0.461 for the non-immigrant and 0.302 for the immigrant, a change of 15.9 percentage points. From the probit the estimated probabilities are 0.462 for the non-immigrant and 0.311 for the immigrant, a change of 15.1 percentage points. This is because such a person is already less likely to have health insurance, so the difference of being an immigrant or not makes a bigger difference compared with the previous example.

The details of this calculation are in an Excel spreadsheet that you can download.

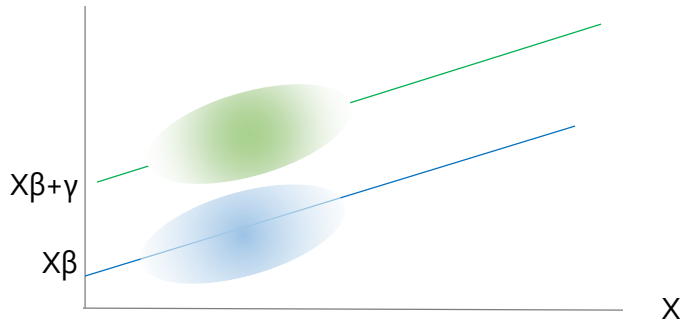
Propensity Score Models

Again Gelman & Hill give a nice explanation. Ordinarily we look at estimating dummy variable coefficients using the whole set of data, so we want to estimate the coefficient on D in the equation, $y = X\beta + \gamma D + \varepsilon$ (where $X\beta$ includes all of the rest of the model variables). If the X variables are very similar for those with D=0 and D=1, then we are likely to get a good estimate of the effect of D (the γ coefficient). But if the values of the of X variables are very different, between those with D=0 and those with D=1, then we need to be sure that the model is very accurate.

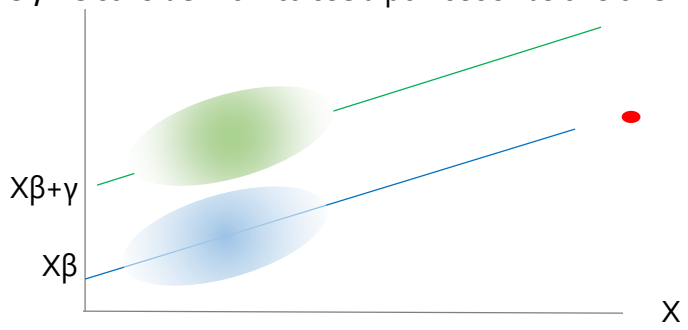
As a simple example, consider again the sort of model we'd discussed about dummy variables – suppose we want to estimate something like this model,



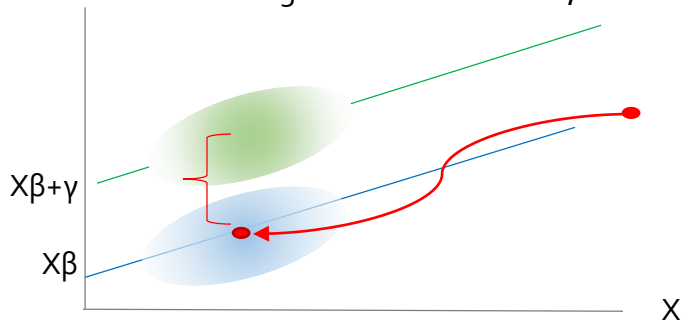
If the data for D=1 and D=0 are similar, then this can be well estimated:



If, however, we consider how to use a point such as this one:



Then what the model is essentially doing is using the estimate of β to shift that down to a comparable location then measuring the vertical distance, so:



But what if the estimate of β is a bit off? What if, instead of a simple linear function like Xβ, we have some nonlinear part? Or an interaction of X and D that is omitted? In that case the new point might be just contributing noise.

So a propensity score model would just compare D=1 values with those certain D=0 values that have X-values that are "close" – leaving out the X-values that are far away. If X is uni-dimensional then defining "close" is pretty easy (as in the graph above) but if X has multiple dimensions then this becomes more difficult – recall our discussion of k-nearest-neighbor for machine learning!

To do this in R, start with a logit model of the 'treatment' – which for this example is whether the person is female. Then use this estimated distance to match.

```
model1 <- glm(female ~ Age + I(Age^2) + AfAm + Asian + Amindian + race_oth
      + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv
      + married + divwidsep + union_m + veteran + immigrant + immig2gen,
      family= binomial, data = dat_use)
X_dist <- model1$fitted
Y_est <- dat_use$WSAL_VAL
```

```
tr_est <- dat_use$female

require('Matching')
# this is numerically intensive
model_match <- Match(Y=Y_est, Tr=tr_est, X=X_dist, M=1, version='fast')
summary(model_match)
```

This estimates the female wage disadvantage to be -18776, compared to a linear regression model where the dummy variable gets an estimate of -19296, so not much of a difference in this case, although other situations might find a bigger difference in estimates.

Alternately we could consider education, which is a bit more of a "treatment" and look at the effect of getting an advanced degree compared with getting a bachelor's degree.

```
use_varb2 <- as.logical(dat_use$educ_bach + dat_use$educ_adv)
dat_use2 <- subset(dat_use,use_varb2) # 19231 obs

model2 <- glm(educ_adv ~ Age + I(Age^2) + female + AfAm + Asian + Amindian +
      race_oth + Hispanic + married + divwidsep + union_m + veteran +
      immigrant + immig2gen, family= binomial, data = dat_use2)
X_dist <- model2$fitted
Y_est <- dat_use2$WSAL_VAL
tr_est <- dat_use2$educ_adv

require('Matching')
model_match2 <- Match(Y=Y_est, Tr=tr_est, X=X_dist, M=1, version='fast')
summary(model_match2)

modelcompare2 <- lm(WSAL_VAL ~ Age + I(Age^2) + female + AfAm + Asian +
      Amindian + race_oth + Hispanic + educ_adv + married + divwidsep +
      union_m + veteran + immigrant + immig2gen, data = dat_use2)
summary(modelcompare2)
```

Here's a good explanation, http://ftp.iza.org/dp1588.pdf