Econometrics
Exercise: Binary Dependent Variables

NOTE: You will need the package 'mfx' and 'mgcv' for this exercise. So please **before class** open up RStudio, go to the top panel and select Tools, then select Install Packages, then type in mfx and hit Install. Then redo and type mgcv and hit install.

Load the data 'smoke_data.csv'.

1. First estimate the probability...
a. a worker is a smoker
b. a worker is a smoker if they work in a place with a smoking ban
c. a worker is a smoker if they work in a place without a smoking ban
(Hint: You can simply use the 'mean(·)' function to find the mean of a variable. Remember if you want a variable you need to do 'data$var'. And also if you want a subset you can do something like: 'data$var1[data$var2 == 1]' - this singles out the var1 data for which var2 = 1.)

2. Use a linear probability model to determine if this difference is significant
a. Interpret the coefficient on 'smkban'

3. Re-estimate your LPM controlling for: female, age, age squared, hsdrop, hsgrad, colsome, colgrad, black, and hispanic.
a. How do the answers from 2 and 3 differ?
b. Why do you think this is the case? Why did it change, specifically?

Lets now look at these using a probit model.

First bring up the mgcv library: 'library(mgcv)'

Now to run a probit regression we will use the 'glm' call instead of the 'lm' or 'plm' call. Other than that the only difference is we need to tell it what the 'link' function is - the function connecting our linear βX to our Y - here it is a normal distribution for example. So:

my.output $<-$ glm(y$\sim$x1+x1, data = mydata, family = binomial(link = "probit"))

Then you can just as usual: summary(my.output)

4. Rerun your full regression from 3 using a probit model.
a. Is the effect of a smoking ban significant?
b. Interpret the coefficient on 'smkban'

Marginal Effects

If you thought some about 4.a you saw it was sort of a trick question. The coefficient, except for the sign, does not really have an interpretation. Thats because our model is:

$Y = \Phi(\beta_0 + \beta_1 smkban + ... + \beta_k X)$

So to really interpret what our coefficients mean we want marginal effects - how does changing X change the expectation of Y.

Now in the simple linear regression case β was the marginal effect, but not here.

Now there are two ways to do this, I really think one is better.
A. Calculate the marginal effect for each person, and take an average of this.
B. Calculate the marginal effect for an 'average person' - the marginal effect of x_1 when all other xs are set to the population average

Now I prefer A, unless you are asking an specific question like how does the smoking ban affect a black woman with a college degree.

To do this lets bring up the mfx library: 'library(mfx)'

And from this package we will use the 'probitmfx' call. The only special thing we need to do is tell it if we want A (atmean = F) or B (atmean = T) - note that B is the default so if you don't tell it which type it will do B.

myresults $<-$ probitmfx(y$\sim$x1+x2,data = mydata, atmean = F)

Now we can't just do 'summarize' here. If we want the estimated marginals and SEs we tell it:

myresults$mfxest

And it gives the estimated marginals, SEs, z-values, and p-values.

5. So now, interpret the coefficient on 'smkban'

6. Use your results from 4 to answer the following:
Someone is a male, white, non-Hispanic, 20 year old high school dropout.
Say he works in a place without a smoking ban, what is the probability he smokes?
Now say he does work in a place with a smoking ban, what is the probability he smokes?
Does this difference differ from the results of 5? Why?

7. Do you think there are any reasons to think we should not think of this as a causal effect?