Week 12: Linear Probability Models, Logistic and Probit

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- Modeling 1/0 outcomes
- The "wrong" but super useful model: Linear Probability Model
- Deriving logistic regression
- Probit regression as an alternative

Binary outcomes

- Binary outcomes are everywhere: whether a person died or not, broke a hip, has hypertension or diabetes, etc
- We typically want to understand what is the **probability** of the binary outcome given explanatory variables
- It's exactly the same type of models we have seen during the semester, the difference is that we have been modeling the conditional expectation given covariates: $E[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- Now, we want to model the probability given covariates: $P(Y = 1|X) = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$
- Note the function f() in there

Linear Probability Models

- We could actually use our vanilla linear model to do so
- If Y is an indicator or dummy variable, then $E[Y|X]$ is the proportion of 1s given X, which we interpret as the probability of Y given X
- The parameters are changes/effects/differences in the probability of Y by a unit change in X or for a small change in X
- If an indicator variable, then change from 0 to 1
- For example, if we model $died_i = \beta_0 + \beta_1 age_i + \epsilon_i$, we could interpret β_1 as the change in the probability of death for an additional year of age

Linear Probability Models

- The problem is that **we know that this model is not entirely correct**. Recall that in the linear model we assume $Y \sim N(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \sigma^2)$ or equivalently, $\epsilon_i \sim N(0, \sigma^2)$
- That is, Y distributes normal **conditional on Xs** or the error distributes normal with mean 0
- Obviously, a 1/0 variable can't distribute normal, and ϵ_i can't be normally distributed either
- We also know that we needed the normality assumption for **inference**, not to get best betas
- **The big picture**: Using the linear model for a 1/0 outcomes is mostly wrong in the sense that the SEs are not right
- Yet, the effects of covariates on the probability of the outcome are more often than not fine
- So LPM is the wrong but **super useful** model because changes can be interpreted in the **probability scale**

Example

- Data from the National Health and Nutrition Examination Survey (NHANES)
- We model the probability of hypertension given age

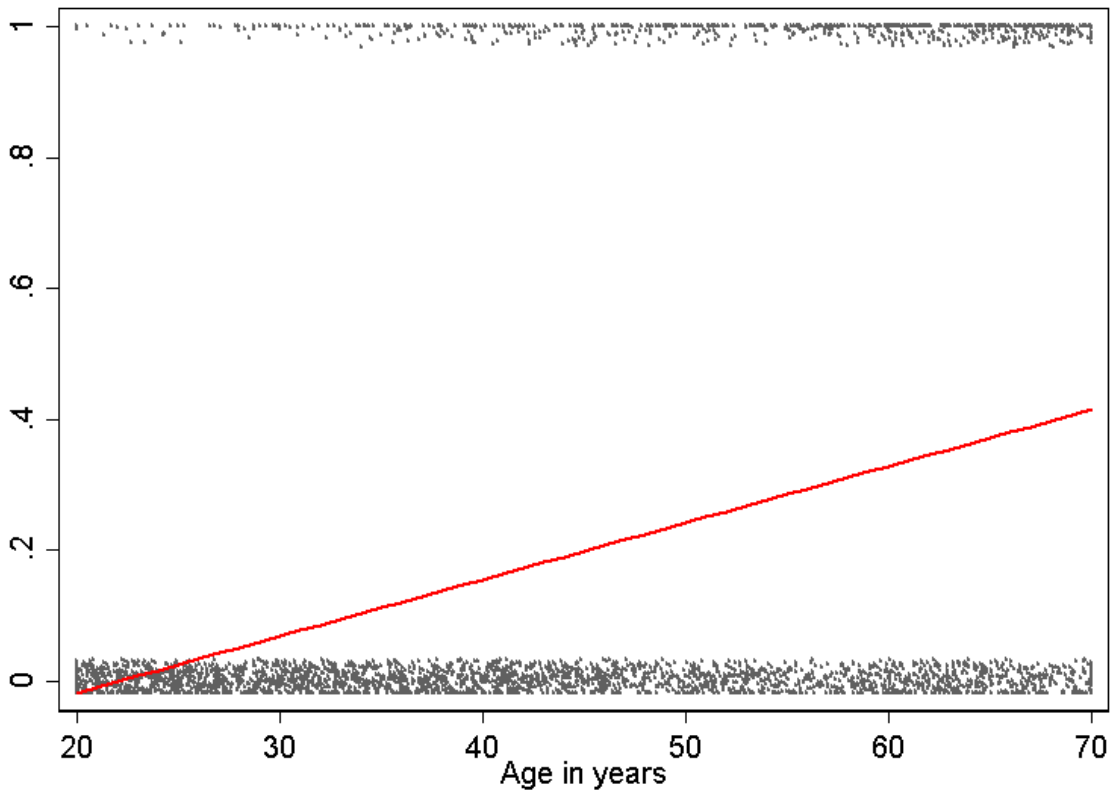
```
  reg htn age

      Source |       SS           df       MS      Number of obs   =     3,529
-------------+----------------------------------   F(1, 3527)      =    432.78
       Model |  59.3248737         1  59.3248737   Prob > F        =    0.0000
    Residual |  483.471953     3,527  .137077389   R-squared       =    0.1093
-------------+----------------------------------   Adj R-squared   =    0.1090
       Total |  542.796826     3,528  .153853976   Root MSE        =    .37024

------------------------------------------------------------------------------
         htn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0086562   .0004161    20.80   0.000     .0078404     .009472
       _cons |  -.1914229   .0193583    -9.89   0.000    -.2293775   -.1534682
------------------------------------------------------------------------------
```

- An additional year of life increases the probability of hypertension by 0.8 percent. See, neat

Example

- Plot predicted values against observed values (note the jittering)

```
reg htn age
predict htnhat_lpm
scatter htn age, jitter(5) msize(tiny) || line htnhat_lpm age, sort ///
        legend(off) color(red) saving(hlpm.gph, replace)
graph export hlpm.png, replace
di _b[_cons] + _b[age]*20
```



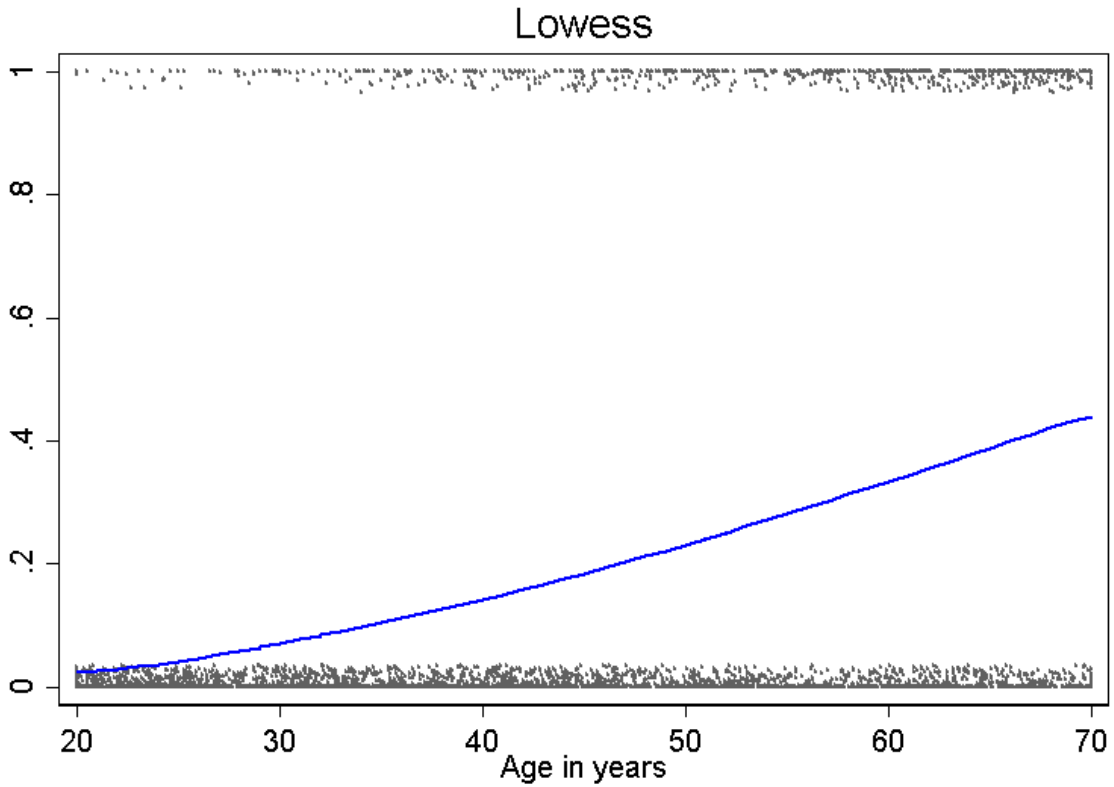
Linear Probability Models

- You can see the first problem with the LPMs
- The relationship between age (or any other variable) **cannot be linear**. Probabilities need to be constrained to be between 0 and 1
- In this example, the probability of hypertension for a 20 y/o is -.0182996
- Is this a big problem in this example? No, because **on average** the probability of hypertension is 0.19, which is not close to zero
- **Lowess** is neat for seeing 0/1 variables conditional on a variable

Lowess

- Lowess can show you how the relationship between the indicator variable and the explanatory variable looks like

```
scatter htn age, jitter(5) msize(vsmall) || line lhtn age, sort ///
        legend(off) color(red) saving(hlowess.gph, replace) ///
title("Lowess")
```



LPN, more issues

- The variance of a 1/0 (binary) depends on the values of X so there is always heteroskedasticity: $var(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})]$
- We know that we need this assumption for correct SEs and F tests. So we can correct SEs in LPMs using the robust option (Huber-White SEs; aka sandwich estimator)
- **Still, we do know that SEs are not totally correct because they do not distribute normal** either, even is we somehow correct for heteroskedasticity
- But at the very least, use the robust option by default

So why do we use LPMs?

- Not long ago, maybe 10 or 15 years ago (it's always 10 to 15 to 20 years ago), you couldn't use other alternatives with large datasets (logistic, probit)
- It would take too long to run the models or they wouldn't run; researchers would take a sample and run logit or probit as a sensitivity analysis
- The practice still lingers in HSR and health economics
- The main reason to keep using LPM as a **first step** in modeling, it's because the **coefficients are easy to interpret**
- In my experience, if the average of the outcome is not close to 0 or 1, not much difference between LPM or logit/probit (but SEs can change, although not by a lot)
- **But not a lot of good reasons to present LPM results in papers anymore**, except maybe in difference-in-difference models

One more time

- Other than interpretation of coefficients or a first pass to modeling, there are **NO GOOD REASONS TO USE THE LPM** model
- Some researches (ok, economists, mostly) truly love the LPN because the parameters are easy to interpret and often the effects are close enough
- Yet, in some cases, the effects could be off, too
- But it's the wrong model. Use a probit or logit, period

Logistic or logit model

- Logistic models can be derived in several ways, which makes learning confusing since you can read different versions
- In the MLE lecture we derived the model assuming that the outcome 1/0 distributes Bernoulli and that observations were iid . We will extend that example today
- An (almost) identical way is to assume that the **outcome comes from a Binomial distribution** since the Binomial is the sum of iid Bernoulli random variables
- A third way is to assume that there is a **latent and continuous** variable that distributes logistic (yes, there is also a logistic pdf), or probit, but we only get to observe a 1 or 0 when the latent variable crosses a threshold
- **You get to the same model** but the latent interpretation has a bunch of applications ins economics (for example, random utility models) and psychometrics (the latent variable is "ability" but you only observed if a person answers a question correctly, a 1/0)

Recall the Logit MLE from the MLE lecture

- Remember that if we have an iid Bernoulli rv we can write the **joint probability** of observing the data as:

  $L(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$
- As we saw before, taking the log makes the analytical and computational calculations easier:

  $lnL(p) = \sum_{i=1}^{n} y_i ln(p) + \sum_{i=1}^{n}(1-y_i)ln(1-p)$
- The log likelihood simplifies to:

  $lnL(p) = n\bar{y} ln(p) + (n - n\bar{y})ln(1-p)$
- If we take the derivative with respect to p and set it equal to zero we find the MLE estimators. The SEs can be calculated using the second derivatives (see the MLE lecture)

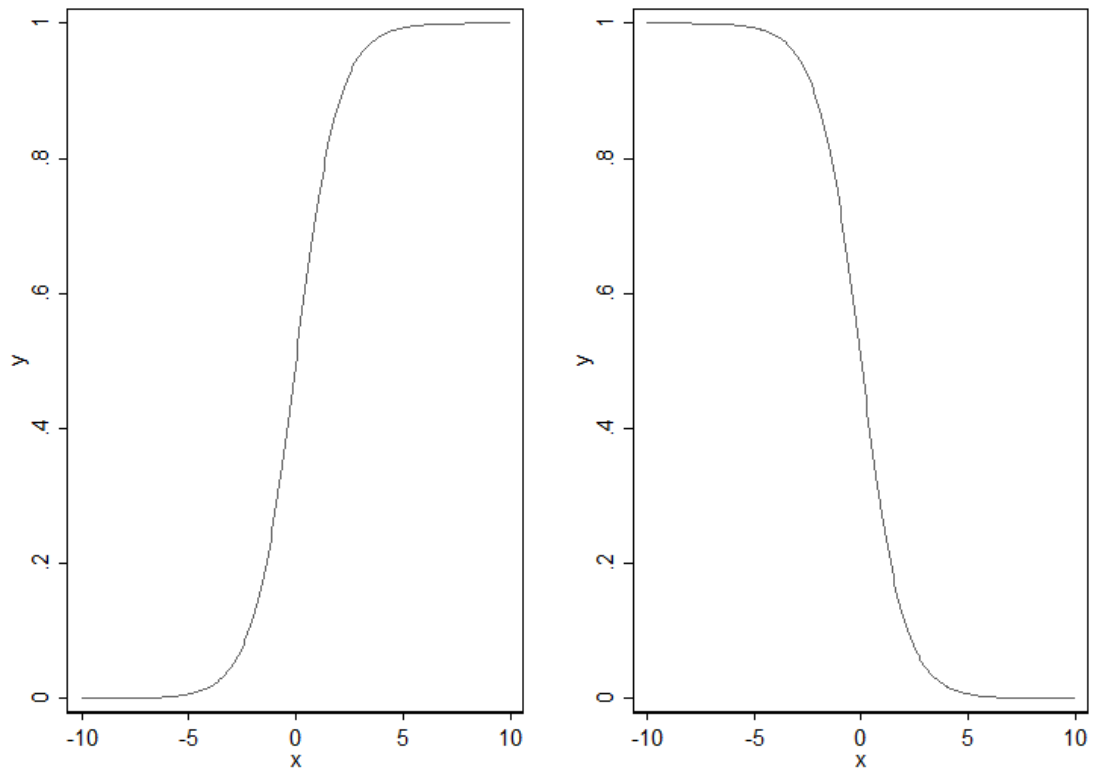
Adding covariates

- We can of course make p a function of covariates but the function can't be something like $p = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ since the p must be bounded between 0 and 1
- So we use a transformation. One of those transformations is the **logistic response function**:
  $\pi = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$, where x is any real number
- $\pi(x)$ is then restricted to be between 0 and 1
- **Confusion alert**: Note that there is a logistic response function and also the logistic distribution. Here we are using the response function. In the latent derivation, the logistic distribution is used
- See more here:
  `https://en.wikipedia.org/wiki/Logistic_distribution`

Logistic response function

- If we constrain the response to be between 0 and 1, it can't be linear with respect to X

```
twoway function y=exp(x) / (1+ exp(x)), range(-10 10) saving(l1.gph, replace)
twoway function y=exp(-x) / (1+ exp(-x)), range(-10 10) saving(l2.gph, replace)
graph combine l1.gph l2.gph, xsize(20) ysize(10)
graph export lboth.png, replace
```



Logistic or logit model

- Notice a couple of things.The effect of x on π is not linear; the effect depends on the value of x
- But we can make the function linear using the so-called logit transformation

  $ln(\frac{\pi}{1-\pi}) = x$
- I made you go the other way in one homework. If you solve for π you get to the logistic response function
- More general, the model is:
- $ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, which transformed is

  $p = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$
- Can also be written as: $p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}$

Logistic MLE

- When extending the log-likelihood function to make p a function of covariates we use the logistic response function to constraint p
- And that's what creates a bunch of interpretation problems. The **estimated betas are changes in the log-odds scale**
- From now on, always, always write logistic models like this:
- $ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$
- That's what Stata (SAS or R) estimate. The betas are **changes in the log-odds scale**
- We will see next class that we could write effects as **odds ratios** but I'll tell you that odds ratios are confusing

Yet another way

- What about if we assume there is a latent (unobserved) variable y^* that is continuous. Think about it as a measure of illness
- If y^* crosses a threshold, then the person dies. We only observe if the person died but we can't observe the **latent variable**
- We can write this problem as

  $P(y = 1|X) = P(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + u > 0) = P(-u < \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) = F(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$
- F() is the cdf of -u. If we assume logistic distribution, we get logistic regression, if we assume cumulative normal, we get a probit model
- See Cameron and Trivedi Chapter 14, section 14.3.1
- Both models are similar because the logistic distribution and the normal and almost the same

Compare logistic distribution and normal

- Stata doesn't have a logistic distribution but you can simulate any probability distribution using the uniform distribution and plugging in into the inverse of the pdf you want

```
clear
set seed 123456
set obs 5000
gen u = uniform()

* Simulate logistic distribution
gen l = -ln((1 - u)/u)
sum l

* Simulated normal with same parameters
gen n = rnormal(r(mean), r(sd))

* Plot
kdensity l, bw(0.3) gen(xl dl)
kdensity n, bw(0.3) gen(xn dn)
line dl xl, sort color(red) || line dn xn, sort ///
   title("Logistic (red) vs normal distribution") ytitle("Density") ///
   xtitle("x") legend(off)
```

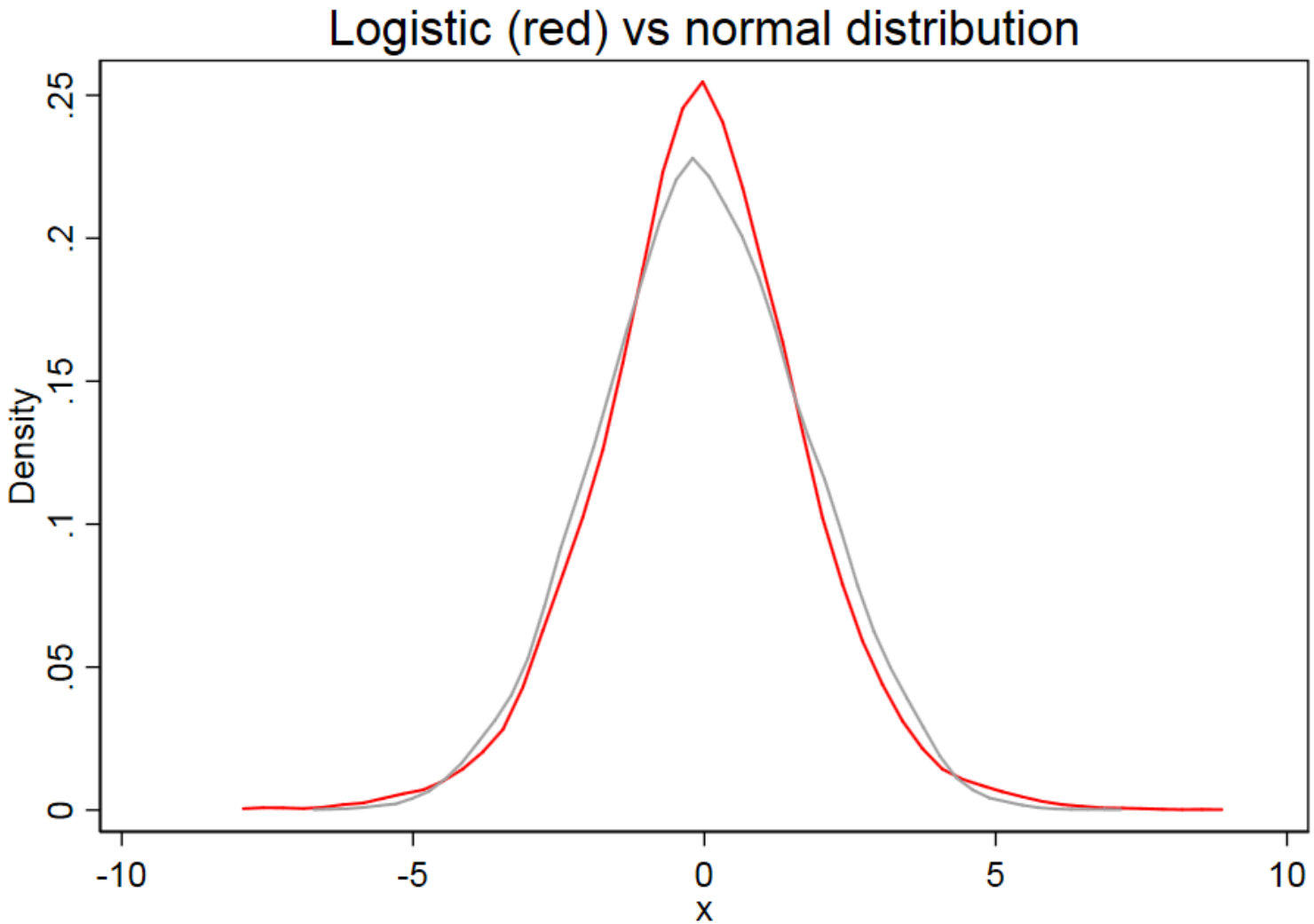
Standard logistic vs standard normal

- Does it make much of a difference if we use one vs the other?

```
clear
set seed 123456
set obs 5000
gen u = uniform()

* Simulate logistic distribution
gen l = -ln((1 - u)/u)
sum l

* Simulated normal with same parameters
gen n = rnormal(r(mean), r(sd))

* Plot
kdensity l, bw(0.3) gen(xl dl)
kdensity n, bw(0.3) gen(xn dn)
line dl xl, sort color(red) || line dn xn, sort ///
   title("Logistic (red) vs normal distribution") ytitle("Density") ///
   xtitle("x") legend(off)
graph export logvsnorm.png, replace
```

Logistic vs normal

- Assuming either one as the latent distributions makes little difference



Big picture

- Not a big difference in the **probability scale** between probit and logit
- If you are an economist you run probit models; for the rest of the world, there is the logistic model
- **IMPORTANT**: There is a **big difference** in terms of **interpreting a regression output** because the coefficients are estimated in **different scales**
- In the logistic model the effect of a covariate can be made linear in terms of the odds-ratio; you can't do the same in probit models
- We will see examples of both

Digression: How to simulate a logistic model

```
set seed 12345678
set obs 10000
gen x1 = rchi2(1)+2

* Make the probability a function of x1
gen pr1 = exp(-1+0.1*x1) /(1+exp(-1+0.1*x1))

* Generate a one-trial binomial (Bernoulli)
gen y = rbinomial(1, pr1)
list y in 1/10
     +---+
     | y |
     |---|
  1. | 0 |
  2. | 1 |
  3. | 1 |
  4. | 0 |
  5. | 1 |
     |---|
  6. | 0 |
  7. | 0 |
  8. | 0 |
  9. | 1 |
 10. | 0 |
     +---+
```

Simulation

- Note that the parameters match the ones I used in the gen step in previous slide

```
 logit y x1

Iteration 0:   log likelihood = -6351.6567
Iteration 1:   log likelihood = -6328.5462
Iteration 2:   log likelihood = -6328.4883
Iteration 3:   log likelihood = -6328.4883

Logistic regression                             Number of obs     =     10,000
                                                LR chi2(1)        =      46.34
                                                Prob > chi2       =     0.0000
Log likelihood = -6328.4883                     Pseudo R2         =     0.0036

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0998315   .0145838     6.85   0.000     .0712478    .1284153
       _cons |  -1.004065   .0493122   -20.36   0.000    -1.100715   -.9074147
------------------------------------------------------------------------------
```

Big picture

- In the next weeks, we will (again) focus on parameter interpretation
- You already have most of the tools to do modeling but we need to adapt them to these types of models
- The tricky part about logistic or probit models or other types of models is to move from the world in which relationships are linear to the world in which they are not
- There is no R^2 but there are other ways to check the predictive ability and fit of models. Mostly, we do it using the estimated log-likelihood

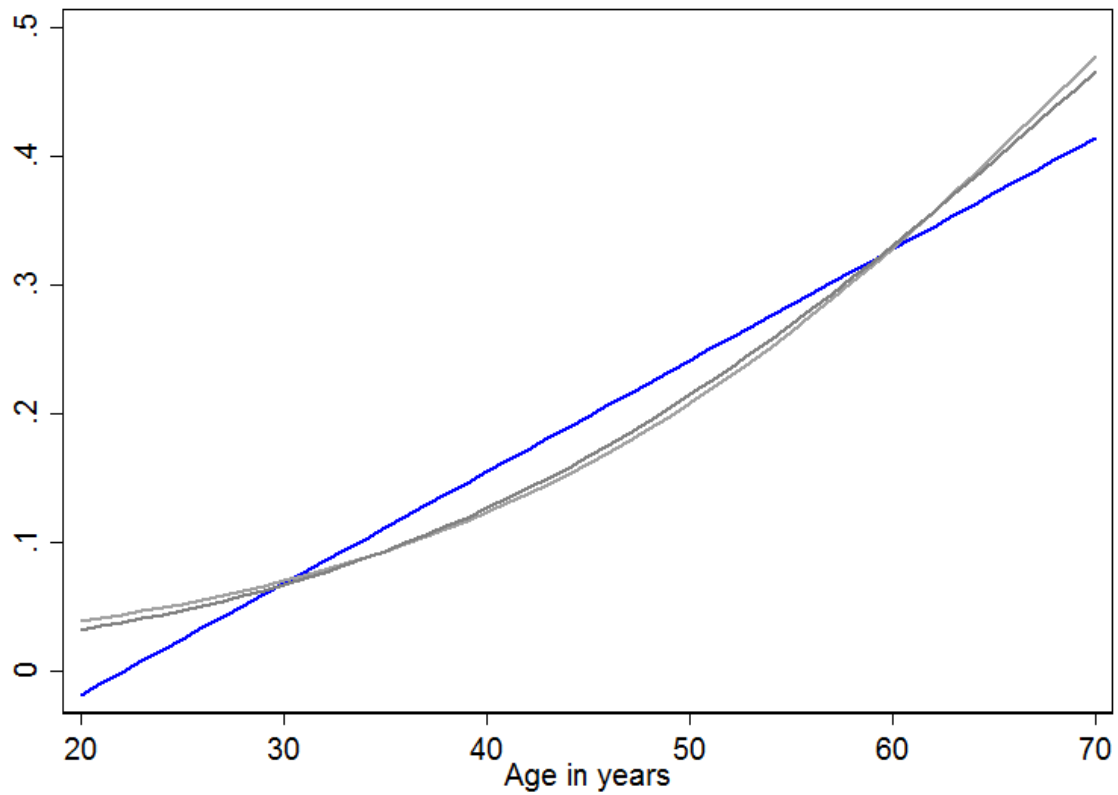
Comparing LPM, logit, and probit models

- For now, let's compare models

```
* LPM
qui reg htn age
est sto lpm
predict hatlpm

* Logit
qui logit htn age
est sto logit
predict hatlog

* Probit
qui probit htn age
est sto prob
predict hatprob

line hatlpm age, sort color(blue) || line hatlog age, sort || ///
     line hatprob age, sort legend(off) saving(probs.gph, replace)
graph export prob.png, replace
```

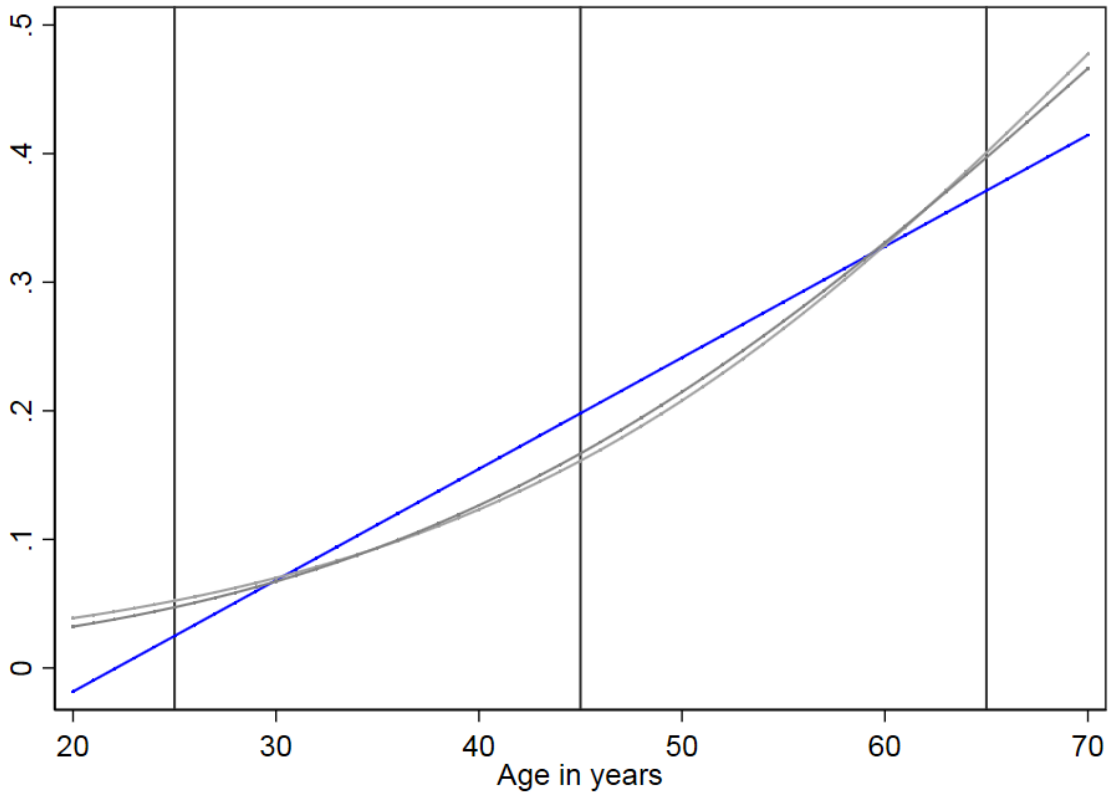
Predicted probabilities

- Note that probabilities are not linear with probit and logit even though we wrote models in the same way
- Note that there is practically no difference between the logit and probit models in this example



But wait, look closely... What about effects?

- In the linear model the effect of X (age in this case) is always the same regardless of age. It's the slope for small changes and when we use marginal effects
- See graph again. Around 45 the effect will be identical (similar slopes). But not around 25 or 65



Effect at different points

- Comparing logit and LPM for now (not probit)
- For the linear model, the effect is, well, linear, so always the same: 0.0086562

```
 qui reg htn age
margins, dydx(age) at(age=(25 45 65))
Conditional marginal effects                    Number of obs     =      3,529
Model VCE    : OLS
Expression   : Linear prediction, predict()
dy/dx w.r.t. : age
1._at        : age             =          25
2._at        : age             =          45
3._at        : age             =          65
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0086562   .0004161    20.80   0.000     .0078404     .009472
          2  |   .0086562   .0004161    20.80   0.000     .0078404     .009472
          3  |   .0086562   .0004161    20.80   0.000     .0078404     .009472
------------------------------------------------------------------------------
```

Effect at different points

- But for logit the effect will be different. Note that the difference is **quite large**

```
qui logit htn age
margins, dydx(age) at(age=(25 45 65)) vsquish
Conditional marginal effects                    Number of obs     =      3,529
Model VCE    : OIM
Expression   : Pr(htn), predict()
dy/dx w.r.t. : age
1._at        : age             =          25
2._at        : age             =          45
3._at        : age             =          65
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0030922   .0001522    20.32   0.000     .0027939    .0033904
          2  |   .0084346   .0004166    20.25   0.000      .007618    .0092511
          3  |   .0149821   .0009574    15.65   0.000     .0131057    .0168585
------------------------------------------------------------------------------
```

But, does this matter?

- Well, yes and no
- If you are using the model to understand how age affects the probability of hypertension, yes, of course it matters
- If you want to understand, the effect of age on $p(htn)$ for most people (on average) then it doesn't matter because the average age on the sample is 45 and at 45 the effect is the same
- Yet another reason to use the LPM as a first pass, not as your final tool
- **At the very least you must understand these issues**
- I'm with Will Manning on this issue. The LPM is the wrong model (but useful)

Coefficients in the estimation scale?

- The coefficients are **very different** because they are **measured in different scales**

```
. est table lpm logit prob, p

-------------------------------------------------
    Variable |    lpm         logit         prob
-------------+-----------------------------------
-            |
         age |  .00865616
             |     0.0000
       _cons | -.19142286
             |     0.0000
-------------+-----------------------------------
htn          |
         age |                .0623793     .0352797
             |                  0.0000       0.0000
       _cons |              -4.4563995   -2.5542014
             |                  0.0000       0.0000
-------------------------------------------------
                                    legend: b/p
```

Similar in probability scale

- We will learn how to make them comparable using the probability scale, which is what we really care about. The margins command is computing an **average effect across values of age**

```
. * LPM
qui reg htn age
 margins, dydx(age)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0086562   .0004161    20.80   0.000     .0078404     .009472
------------------------------------------------------------------------------

* Logit
qui logit htn age
margins, dydx(age)
------------+-----------------------------------------------------------------
         age |   .0084869   .0004066    20.87   0.000       .00769    .0092839
------------------------------------------------------------------------------

* Probit
qui probit htn age
margins, dydx(age)
-------------+----------------------------------------------------------------
         age |   .0084109   .0003917    21.47   0.000     .0076432    .0091787
------------------------------------------------------------------------------
```

Summary

- There are many ways to arrive to the logistic model
- The big picture is that we are trying to model a probability, which must be bounded between 0 and 1
- If it's bounded, then the effect of any covariate on the probability cannot be linear
- The tricky part is to learn how to interpret parameters. The short story is that we estimate (assuming one predictor) $log(\frac{p}{1-p}) = \beta_0 + \beta_1 x$ but we care about $p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$
- To make life easier, logistic models are often interpreted using odds ratios but odds ratios can be misleading
- In the probit model, we interpret parameters as shifts in the cumulative normal, even less intuitive