Week 5

Binary Dependent Variables

POLI 6003 Maximum Likelihood
Rich Frank
University of New Orleans
September 20, 2012

Today we start to look at limited dependent variables.

Limited Dependent Variables

- Limited DVs are different from the continuous and unbounded DVs required for OLS to be BLUE.

- The main reason that OLS fails is that limited DVs are not normally distributed.

- If the variable is not normally distributed then the residuals are not normally distributed.

Limited Dependent Variables

- **Binary:**
  - 0 = no vote; 1 = vote
  - 0 = peace; 1 = war
  - 0 = not crying; 1 = crying
  - 0 = not in the military; 1 = in the military

Limited Dependent Variables

- **Ordered**: How much do you like the Saints?
  - I'm a Vikings fan, so I think Drew should drive off the twin span.
  - They're alright I guess, I don't watch games because they come on at the same time as Jeopardy
  - I go to Finn's to watch most games
  - I've got season tickets and my old 'Aints bag
  - Who are the Saints? I'm not religious.

Limited Dependent Variables

- Discrete Count:
  - Number of civil wars in Africa this year: 0,1,2,…∞
  - Number of attendees at my friend's bar mitzvah
  - Number of Kanye West's followers on Twitter

Limited Dependent Variables

- **Nominal**:
  - Do you eat Chocolate, Vanilla, Strawberry, or Chubby Hubby ice cream most frequently?
  - Do you take the train, bus, bike, or feet to class?

Limited Dependent Variables

- **Time to Failure:**
  - How long does a civil war last?
  - How long does a patient survive with cancer?
  - How long does a dictator last in office?

Binary Variables

- Binary DVs are the simplest type of model for limited dependent variables.
- Generally, we think of binary DVs as representing some underlying and unobserved continuous variable.
- If we could somehow observe and measure this underlying variable, we could just use OLS and call it a day.
- Unfortunately, we cannot.

Binary Variables

- A state might be at a risk for civil war without actually being at war, and we have a difficult time actually measuring this risk directly.

- Or a person might be very likely to vote, but gets arrested for an outstanding warrant on the way to the polls.

- We fall back on measuring what we can—whether civil war does or does not happen or whether someone votes or not.

- One way to approach modeling with such data is the *latent variable approach*.

- We think that there is an underlying propensity that produces the observed outcomes: 0, 1.

$$y^* = \mathbf{X}\beta + u$$

Where

$$y_i = \begin{cases} 1, & if \;\; y_i^* > \kappa \\ 0, & if \;\; y_i^* \leq \kappa \end{cases}$$

So why not use OLS?

- The only main difference is that the dependent variable is not continuous.

- Assume the normal regression model:

$$Y = \mathbf{X}\beta + u$$

  - Where we are interested in the conditional expectation of Y, $E(y_i | x_i)$, and where we want to interpret that expectation as a conditional probability, $p(y = 1 | x_i)$.

Several problems arise with binary DVs

1. The residuals are not normally distributed.

When $y_i = 1$, $u_i = 1 - x_i\hat{\beta}$

$y_i = 0$, $u_i = 0 - x_i\hat{\beta}$

Clearly the errors (like y_i) can take on only two values. Therefore, they are distributed according to the binomial rather than the normal distribution.

2. **The variance of the error term is heteroskedastic.**

- The DV and the error have the same probability distribution.

Y	Probability	u
y = 0	1 - p	$- x_i$
y = 1	p	$1 - x_i$

-And the variance of the disturbance term is:

$$var(u_i) = E(Y_i \mid X_i)[1 - E(Y_i \mid X_i)]$$
$$= p_i (1 - p_i)$$

- So the disturbance term variance depends on the conditional expectation of Y, which is conditional on X.

- This means that the variance of *u* depends on the independent variables.

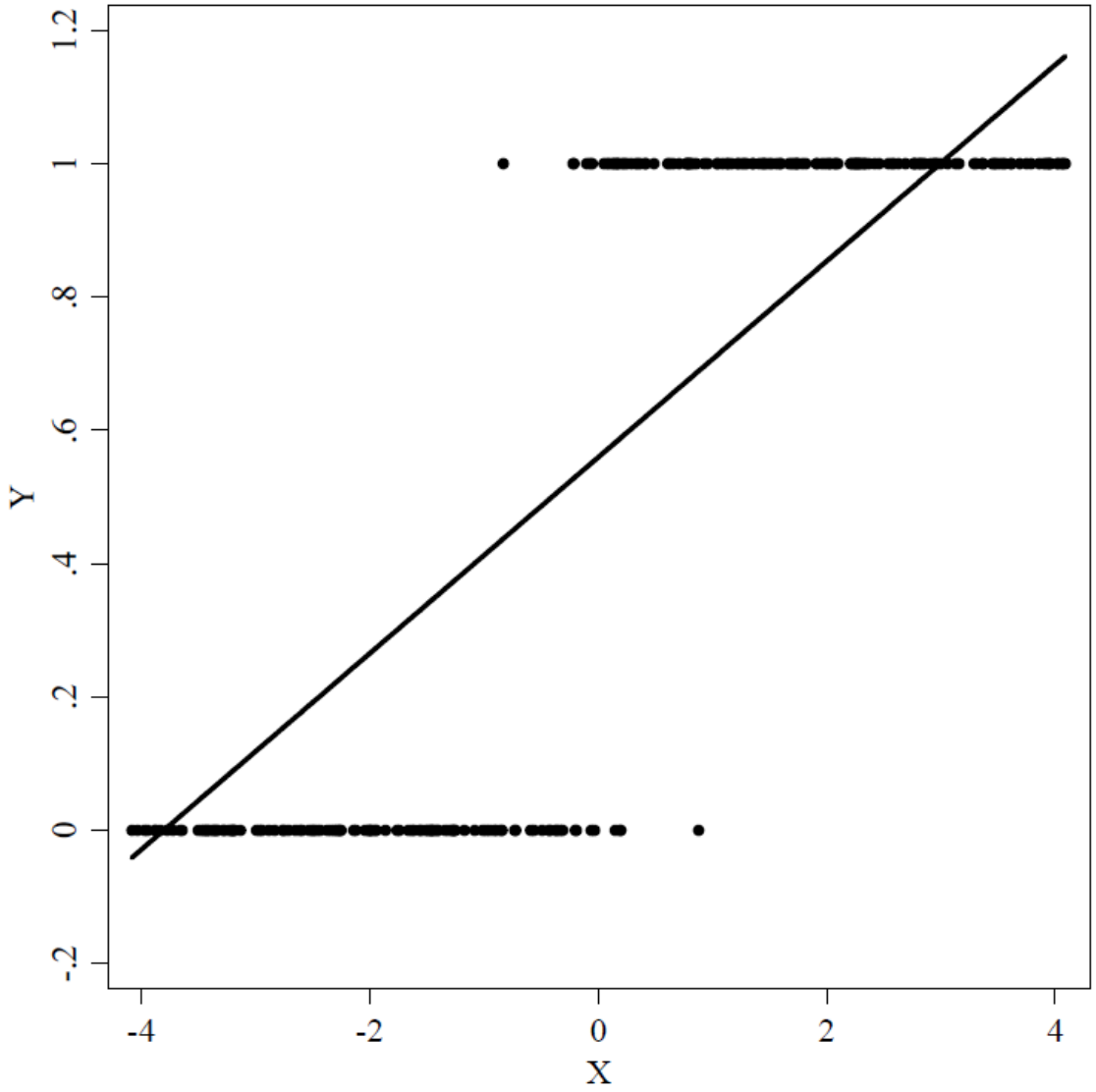
- Thus the variance of the error term is not homoskedastic.

Last problem…

3. The predictions of Y, the conditional expectation $E(y_i \mid x_i)$, is not bounded by 0 and 1.

-Thus the following expectation is not fulfilled.

$$0 \leq E(y_i \mid \boldsymbol{x}_i)$$


1.2
1
.8
.6
.4
.2
0
-.2
Y
-4
-2
0
2
4
X

Assumptions of Logit and Probit

1. The threshold separating Y = 0 and Y = 1 (the transition point) is Y* > 0.

2. The conditional expectation of the residuals is

$$E(u|X) = 0$$

3. The variances are fixed at $\frac{\pi^2}{3}$ (for logit—this is the variance of the logistic distribution) or 1 (for the standard normal (for probit).

Let's make up some data and explore this...

```
clear
set obs 10000
gennorm x1 e, corr(.0)
gen y = 1 + x1 + e
gen yb= 0
replace yb=1 if y>=0
reg yb x1
predict yhat, xb
predict resid, r
probit yb x1
predict py1
```

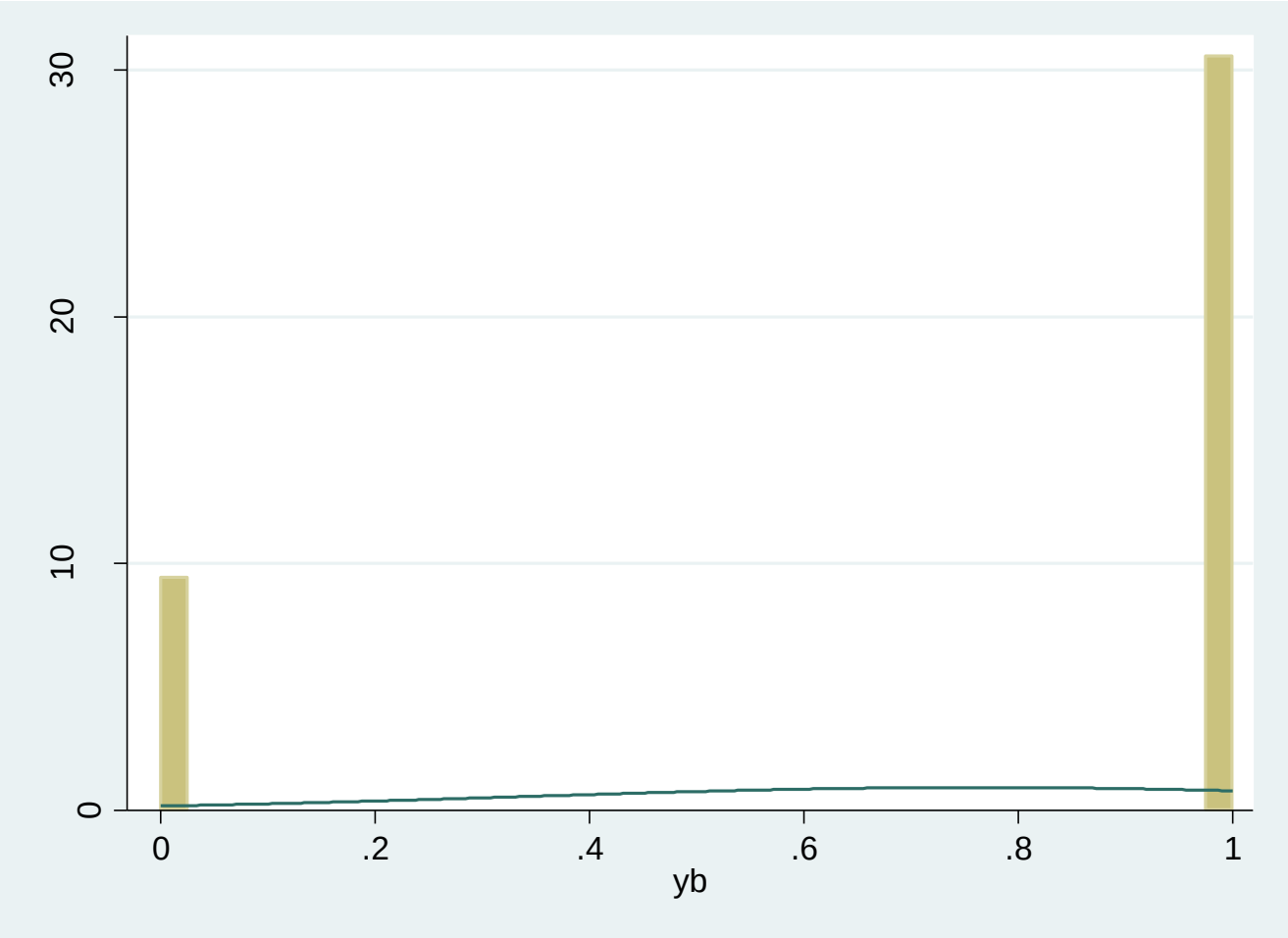
Let's take a look at these data

```
summary x1 e y yb yhat resid

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          x1 |     10000   -.0071565    1.005172  -3.579602   3.915892
           e |     10000     .004938    .9915043  -3.794222   3.700645
           y |     10000    .9977815    1.414712  -3.978064   6.202808
          yb |     10000       .7566     .429156          0          1
        yhat |     10000       .7566    .2253503  -.0443094   1.636111
-------------+--------------------------------------------------------
       resid |     10000   -6.67e-11    .3652288  -1.099107   .8482409
```

First, notice that the predicted y (yhat or $\hat{y}_i$) ranges from -.0443 to 1.636. This exceeds the bounds of the latent variable $P(Y = 1)$.

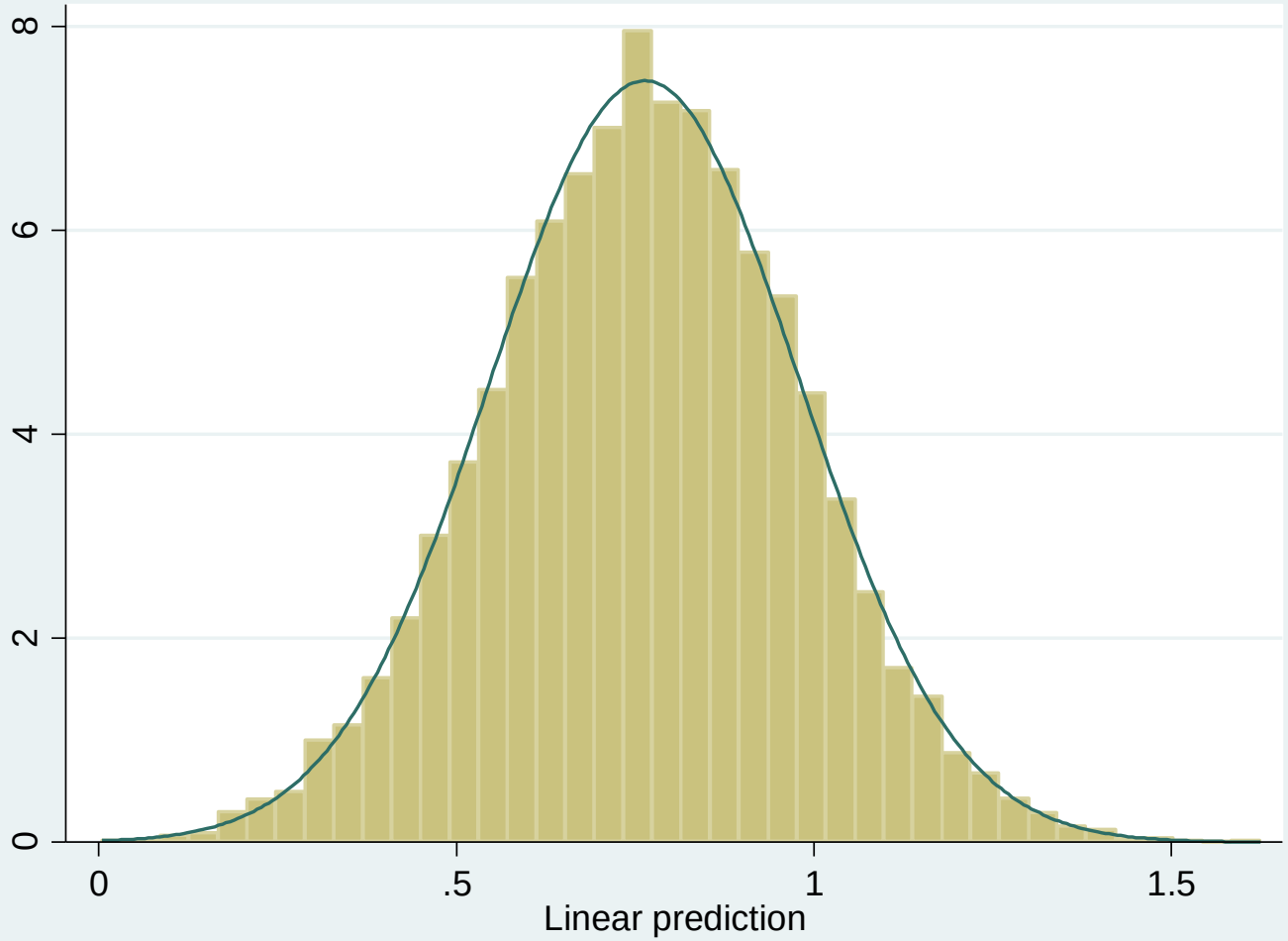
- Let’s look at some graphs.

Distribution of y



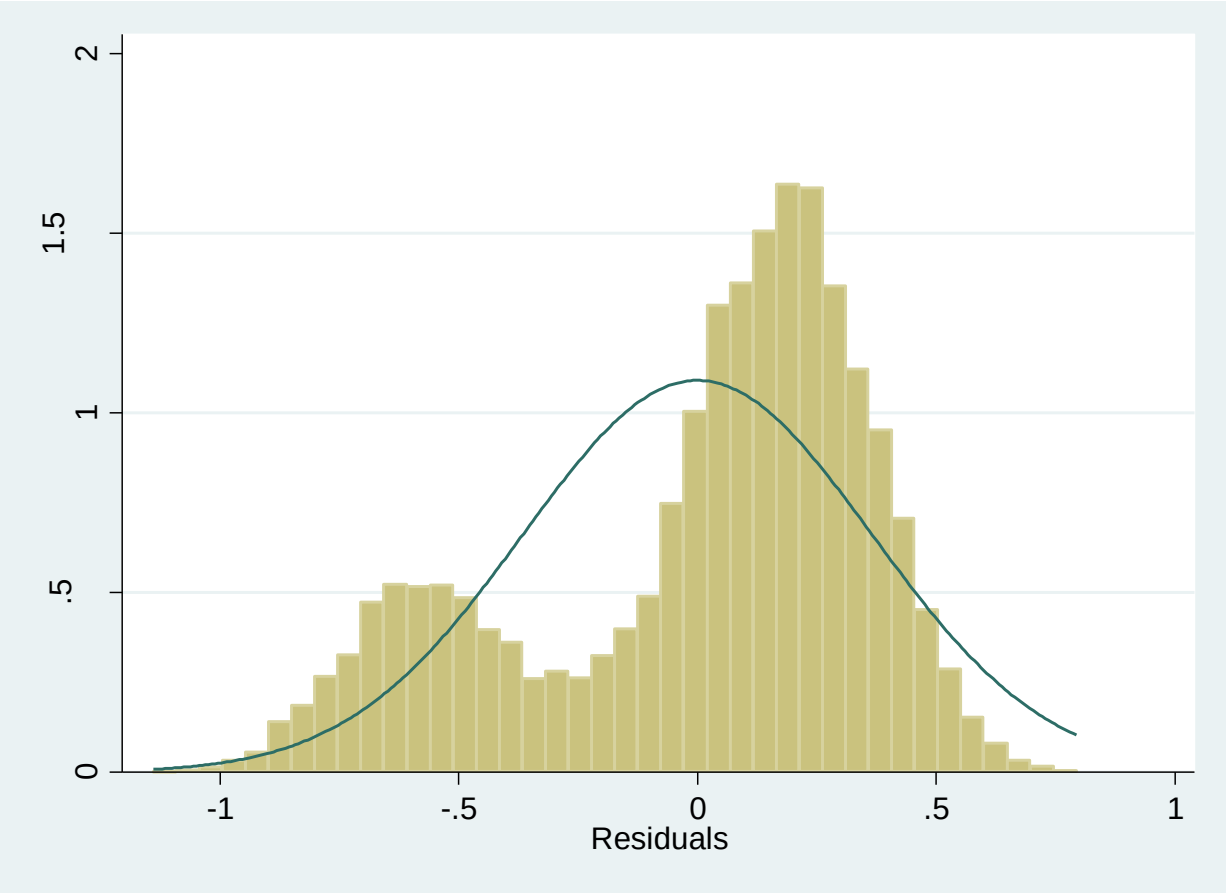
As you can see, Y is not normally distributed.

Distribution of $\widehat{y_i}$



A large number of predictions are above 1.

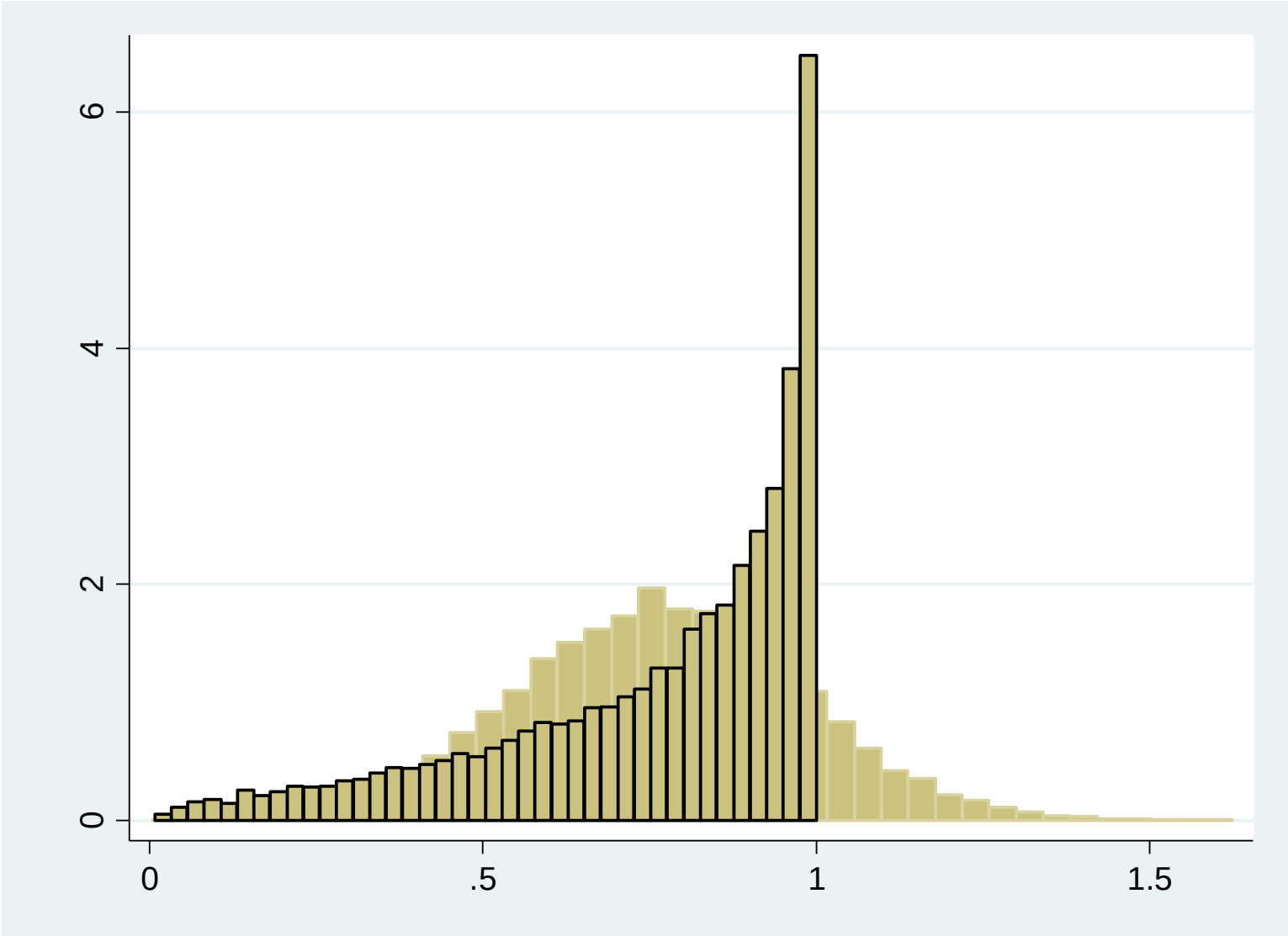
Distribution of Residuals



Residuals are also not normally distributed and appear more binomially distributed.

This is basically the first graph minus the second (why?).

Comparing OLS and Probit predictions



Stata code to create these tables

```
histogram yb, normal
histogram yhat, normal percent
histogram resid, normal
twoway (histogram yhat, density) ///
       (histogram py1, density blcolor(black) ///
       ), legend(off)
```

- So again $u_i = y_i - \hat{y}_i$

With a dichotomous variable:

$$u_i = \begin{cases} 1 - \hat{y}_i \ \ if \ \ y_i = 1 \\ 0 - \hat{y}_i \ \ if \ \ y_i = 0 \end{cases}$$

Let's use some real data now.

- For example, I downloaded data from Fearon and Laitin's (2003) famous APSR article on civil war and ran some models.

```
reg  onset lgdpenl1 colbrit colfra polity2l nwstate

      Source |       SS       df       MS              Number of obs =    6327
-------------+------------------------------           F(  5,  6321) =   14.63
       Model |  1.36243958     5  .272487916            Prob > F      =  0.0000
    Residual |  117.759735  6321  .018629922            R-squared     =  0.0114
-------------+------------------------------           Adj R-squared =  0.0107
       Total |  119.122175  6326  .018830568            Root MSE      =  .13649

------------------------------------------------------------------------------
       onset |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lgdpenl1 |  -.0108425     .00189    -5.74   0.000    -.0145476   -.0071375
     colbrit |   -.005164   .0039822    -1.30   0.195    -.0129705    .0026426
      colfra |  -.0120163   .0050562    -2.38   0.018    -.0219281   -.0021045
    polity2l |   .0001876   .0002629     0.71   0.476    -.0003279     .000703
     nwstate |   .0609367   .0109221     5.58   0.000     .0395257    .0823476
       _cons |   .1022667   .0150098     6.81   0.000     .0728424     .131691
------------------------------------------------------------------------------
```

```
. predict resid, r
(283 missing values generated)

. predict yhat, xb
(283 missing values generated)

. count if yhat<0
  398
```

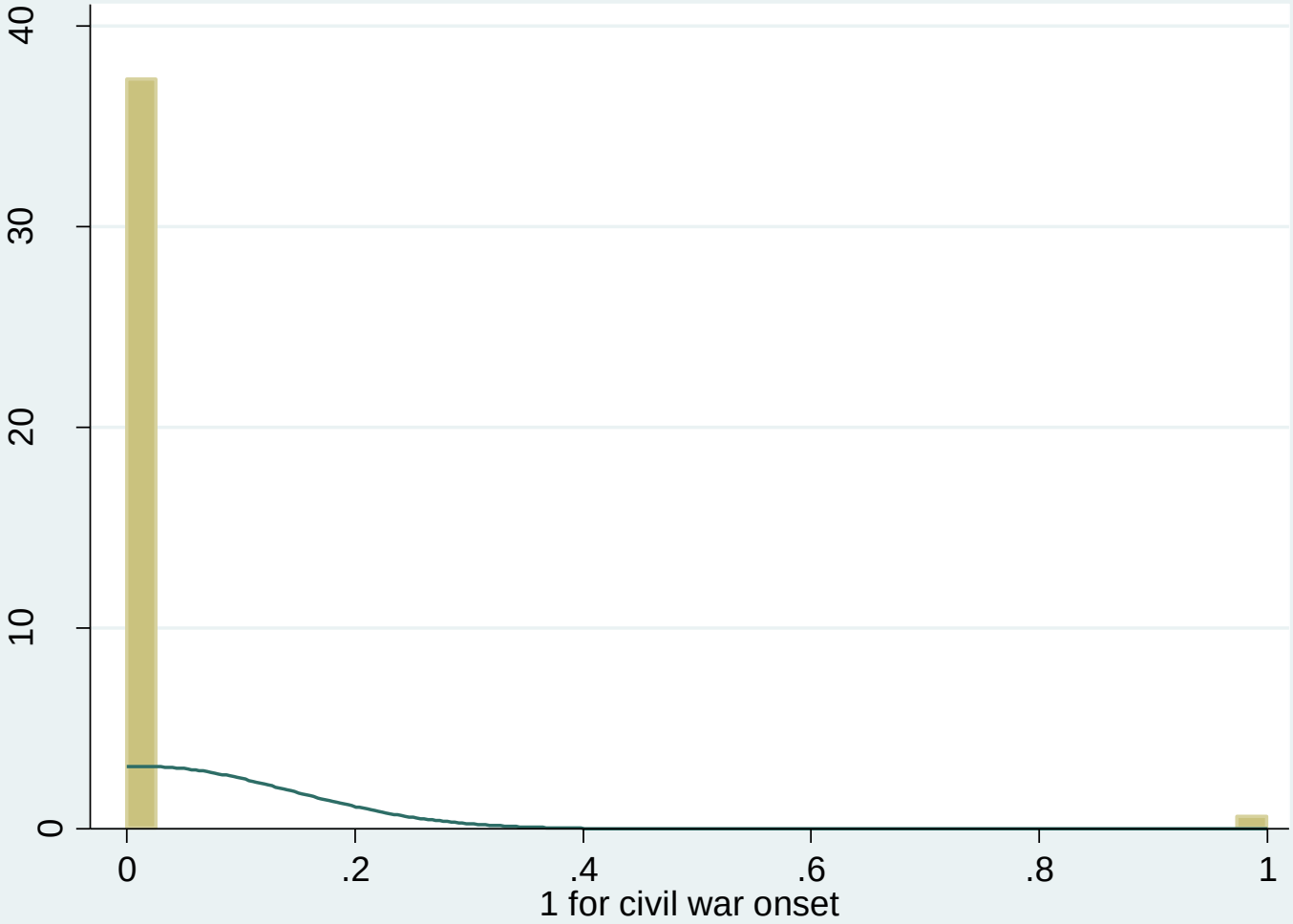
- You can see from the last command that over 6% of the predictions are out of bounds.

- Let's look at predictions about democracy.

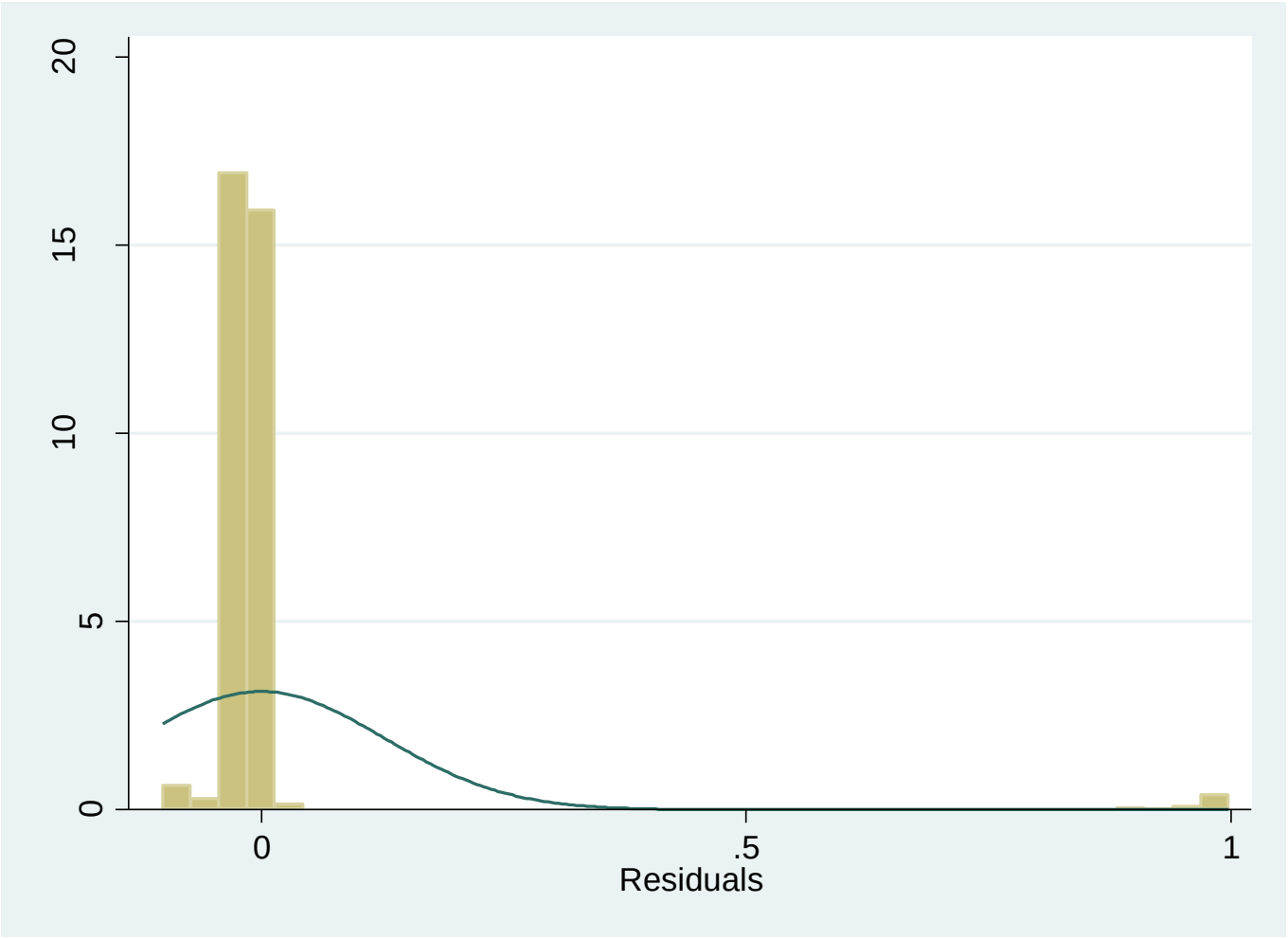
```
. sum polity2l if yhat<0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    polity2l |       398    1.630653    9.123379        -10         10


. tab  polity2l if yhat<0

     lagged |
   polity2, |
 except 1st |
 in country |
     series |      Freq.     Percent        Cum.
------------+-----------------------------------
        -10 |         52       13.07       13.07
         -9 |         63       15.83       28.89
         -8 |         39        9.80       38.69
         -7 |          3        0.75       39.45
         -6 |          1        0.25       39.70
         -4 |          3        0.75       40.45
         -2 |         18        4.52       44.97
          0 |          4        1.01       45.98
          8 |          7        1.76       47.74
          9 |         13        3.27       51.01
         10 |        195       48.99      100.00
------------+-----------------------------------
      Total |        398      100.00
```

- Polity is an index ranging from -10 to 10 coding states from the most autocratic to the most democratic.

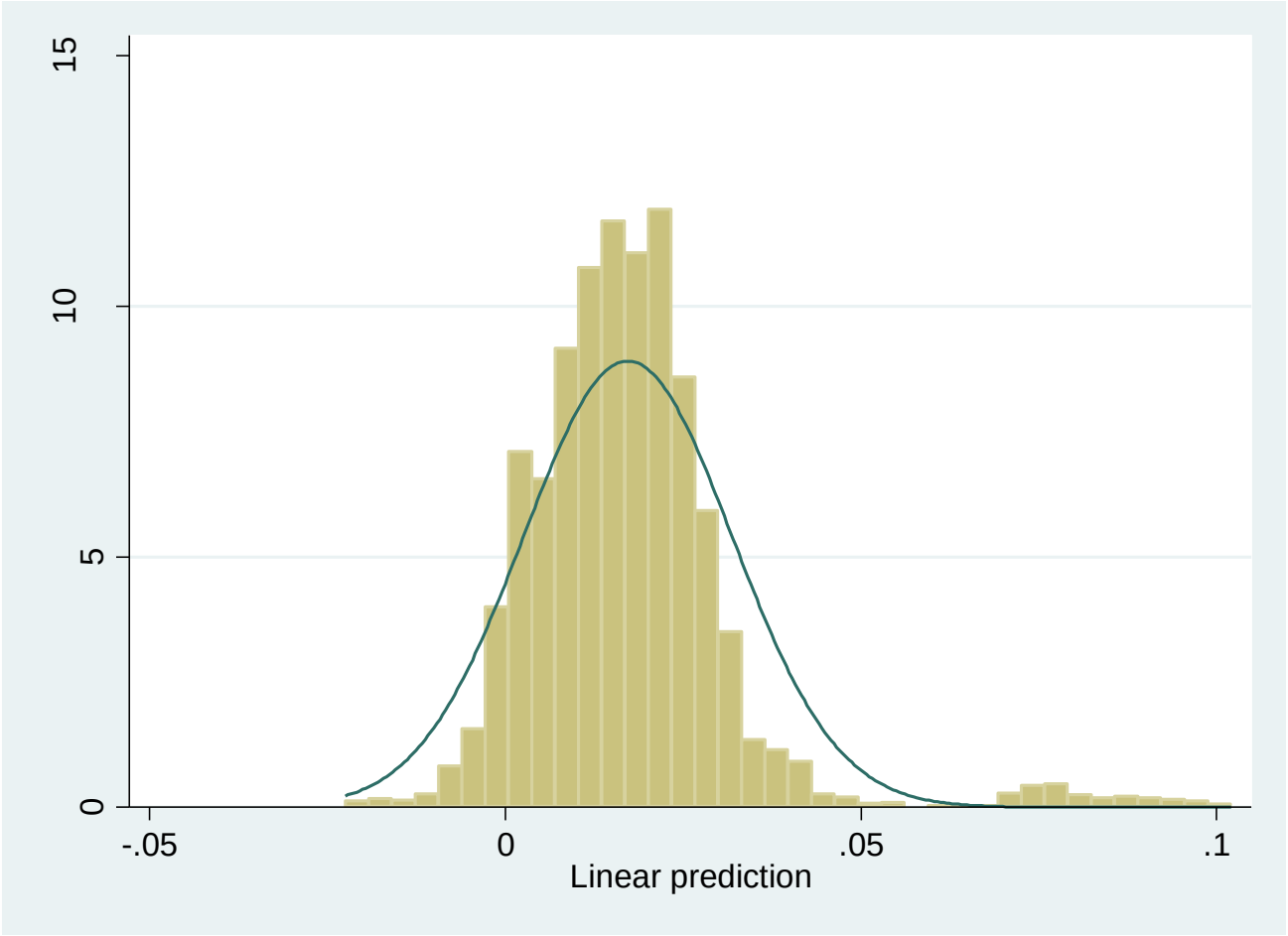
- You can see that most of the nonsense predictions are at both ends of the scale.



The DV is definitely not normally distributed.



Ditto the residuals.

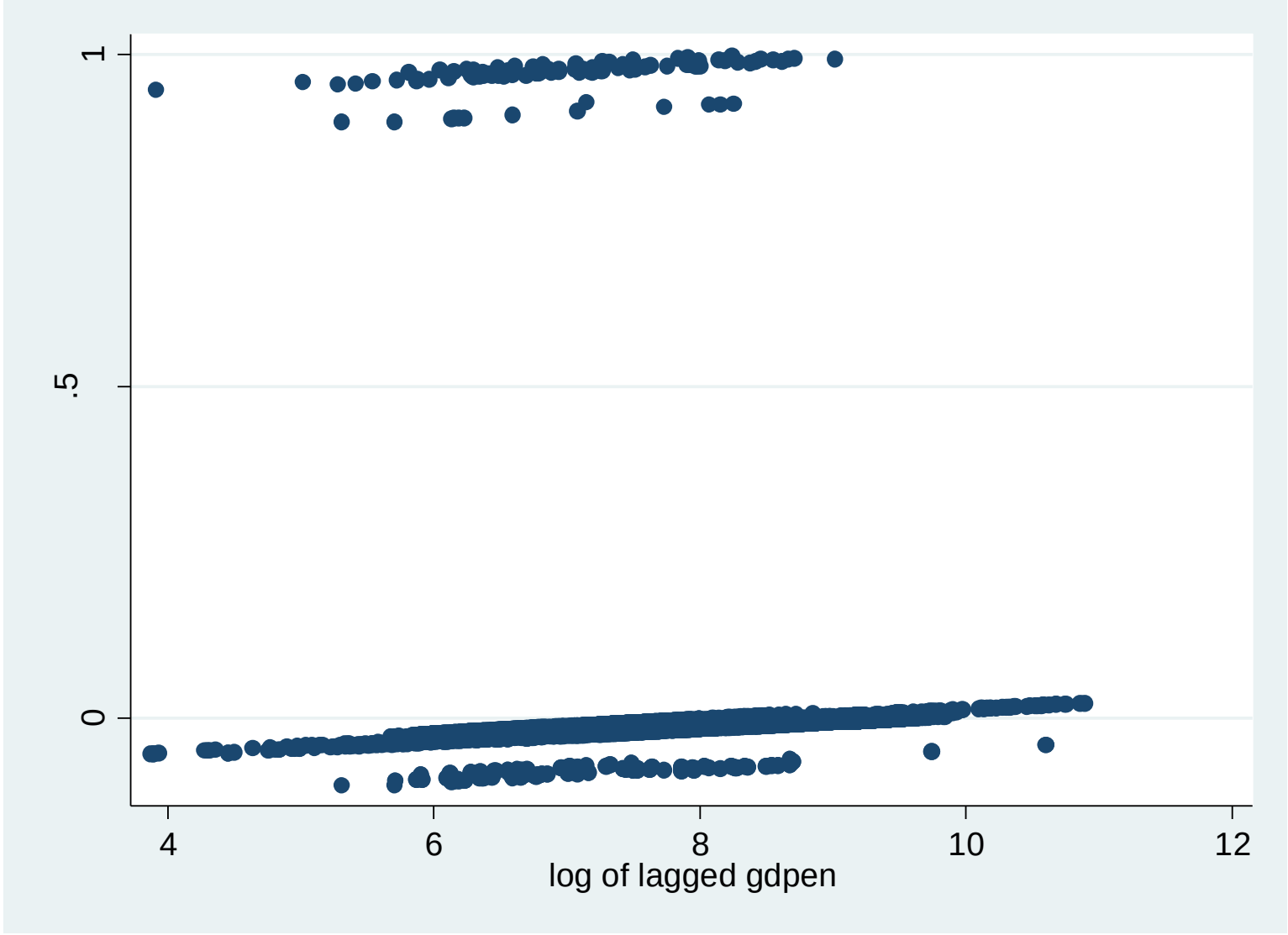


And the probability of civil war is predicted to be negative in some cases!

Residuals and Polity

Let's plot the disturbances against a variable or two of interest…

Residuals and Logged GDP



- So the linear model has problems when looking at latent variables observed dichotomously.

- However, there is *an additional problem*. The linear probability model constrains the relationship between P and X to be linear.

- This means that the rate of change towards P =1 (or 0) is constant for all values of X.

- The change between .99 to 1.0 is the same as between .50 and .51.

- However, we know that the probabilities of 0 and 1 are hardly common.

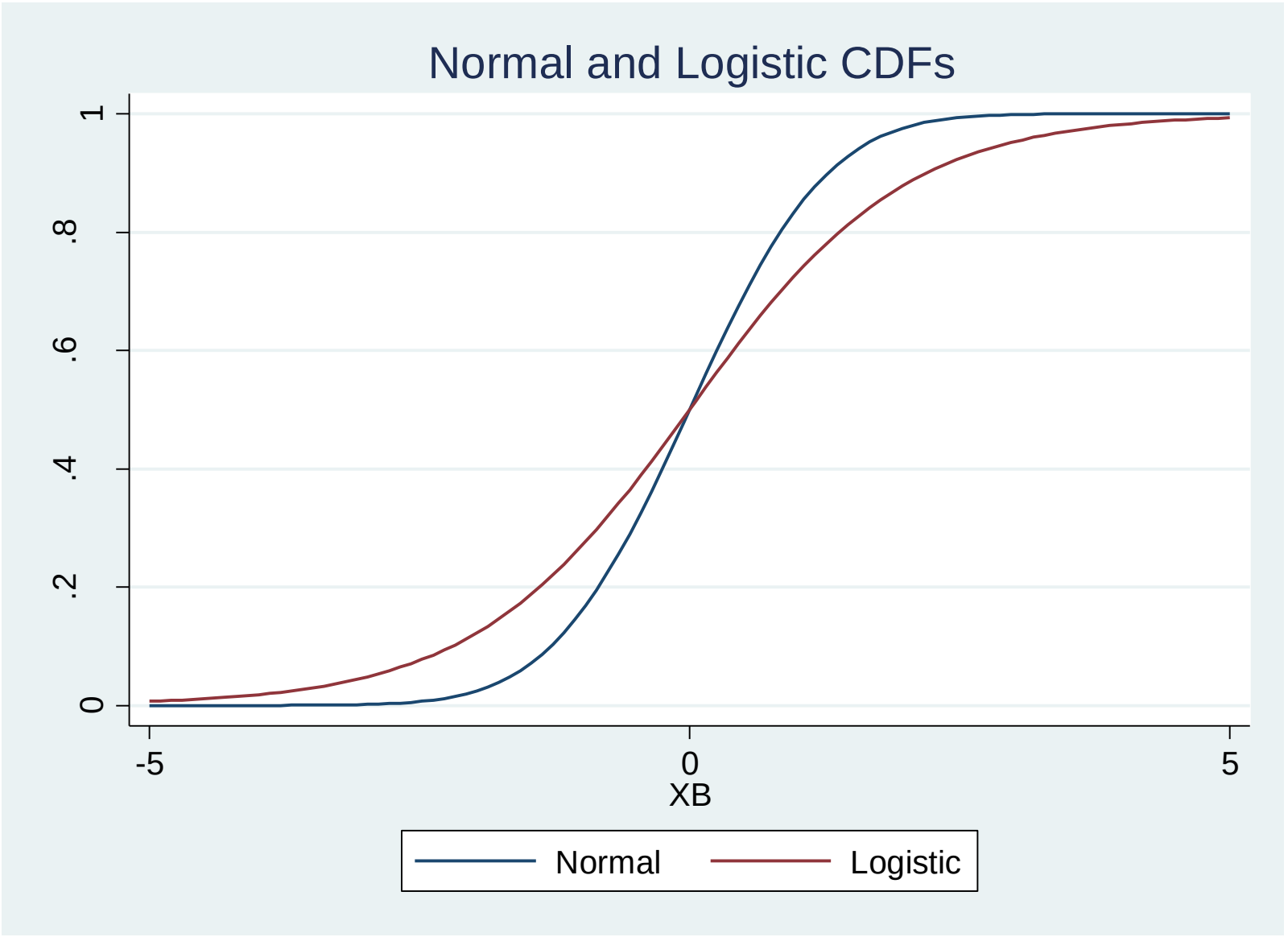
- Let’s say we are interested in whether a person owns a car or not is a function of income.

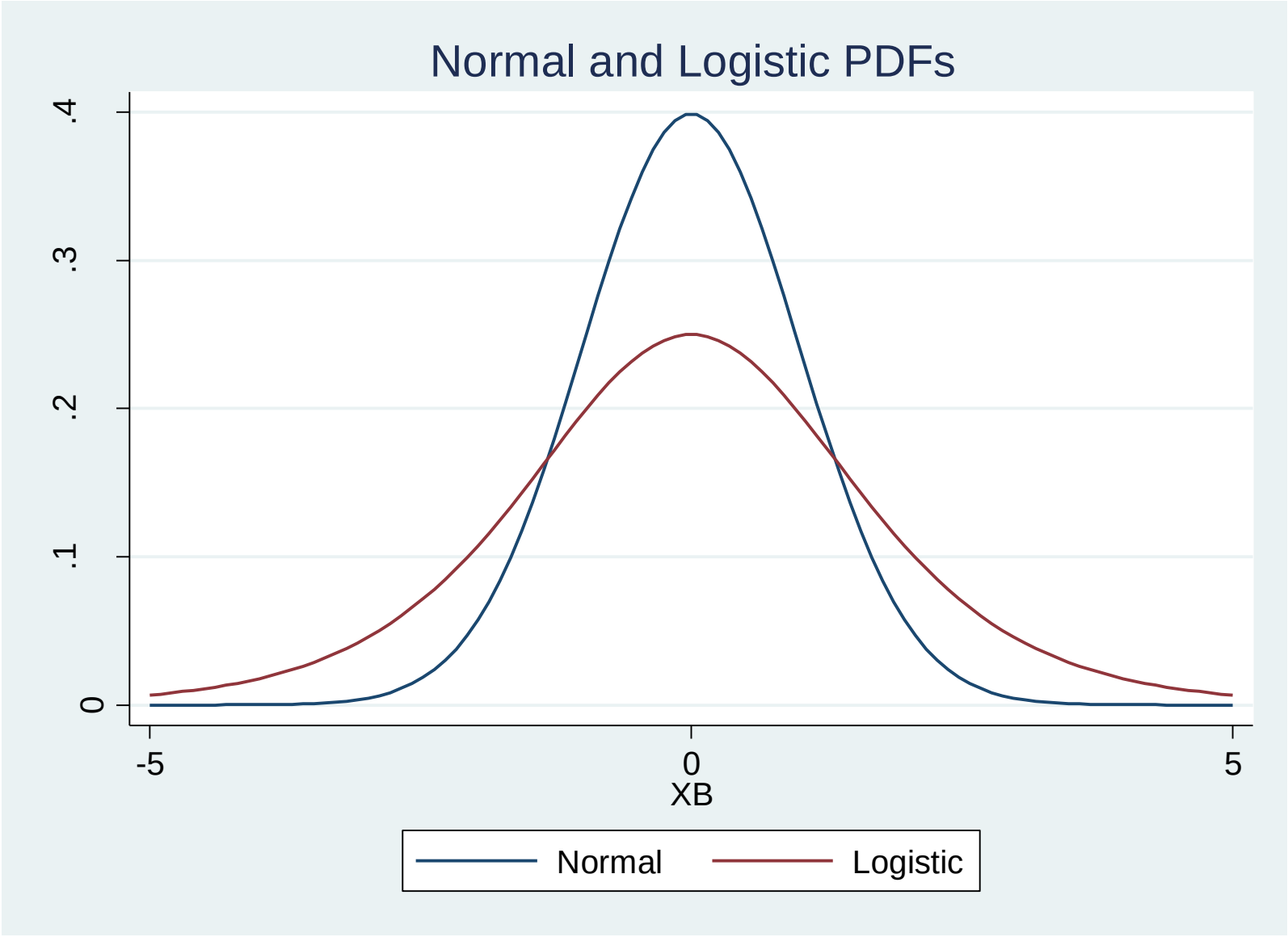
- The difference between 0 and $1,000 income is probably not going to have the same effect as shifting from $25,000 to $26,000 or $150,000 to $151,000.

- What we want is a non-linear distribution to better capture the non-linear relationship between P and X.

- This is why we turn to other distributions like the cumulative distribution of a random variable, which looks better because it follows an S-shaped curve.

- Using the CDF to represent a binary variable generally leads us to the two most common CDFs: the logistic and the normal.


Normal and Logistic CDFs
0
.2
.4
.6
.8
1
-5
0
5
XB
Normal
Logistic


Normal and Logistic PDFs
0
.1
.2
.3
.4
-5
0
5
XB
Normal
Logistic

- These models are based on the Bernoulli distribution.

$$Y = \begin{cases} 1, & \pi \\ 0, & 1 - \pi \end{cases}$$

And we want to allow the parameter π to vary across cases *i* instead of keeping it fixed to one value.

$$Y = \begin{cases} 1, & \pi_i \\ 0, & 1 - \pi_i \end{cases}$$

- We then want to estimate π_i as a function of some independent variables, so

$$\pi_i = f(X\beta)$$

- This function must also meet the bounds

($0 \leq \pi_i \leq 1$) that the LPM did not, and allow the X's to vary (as widely as possible) so that $-\infty < x < +\infty$.

- This allows the $X\beta$'s to have the same range.

- However, the function also has to be able to give us values that are bounded by 0 and 1.

$$0 \leq f(X\beta) \leq 1$$

- This leads us to the logistic and normal probability functions.

Logit

- The logit model is based on the inverse of the gamma distribution we saw last week where the probability Y = 1 is:

$$P_i = E(\mathrm{Y} = 1 \mid \mathrm{X_i}) = \frac{1}{1 + e^{-X\beta}}$$

And the probability Y = 0 is:

$$1 - P_i = E(\mathrm{Y} = 0 \mid \mathrm{X_i}) = \frac{1}{1 + e^{X\beta}}$$

- This means that the odds ratio can be written as:

$$\frac{P_i}{1-P_i} = \frac{1+e^{X\beta}}{1+e^{-X\beta}} = e^{X\beta}$$

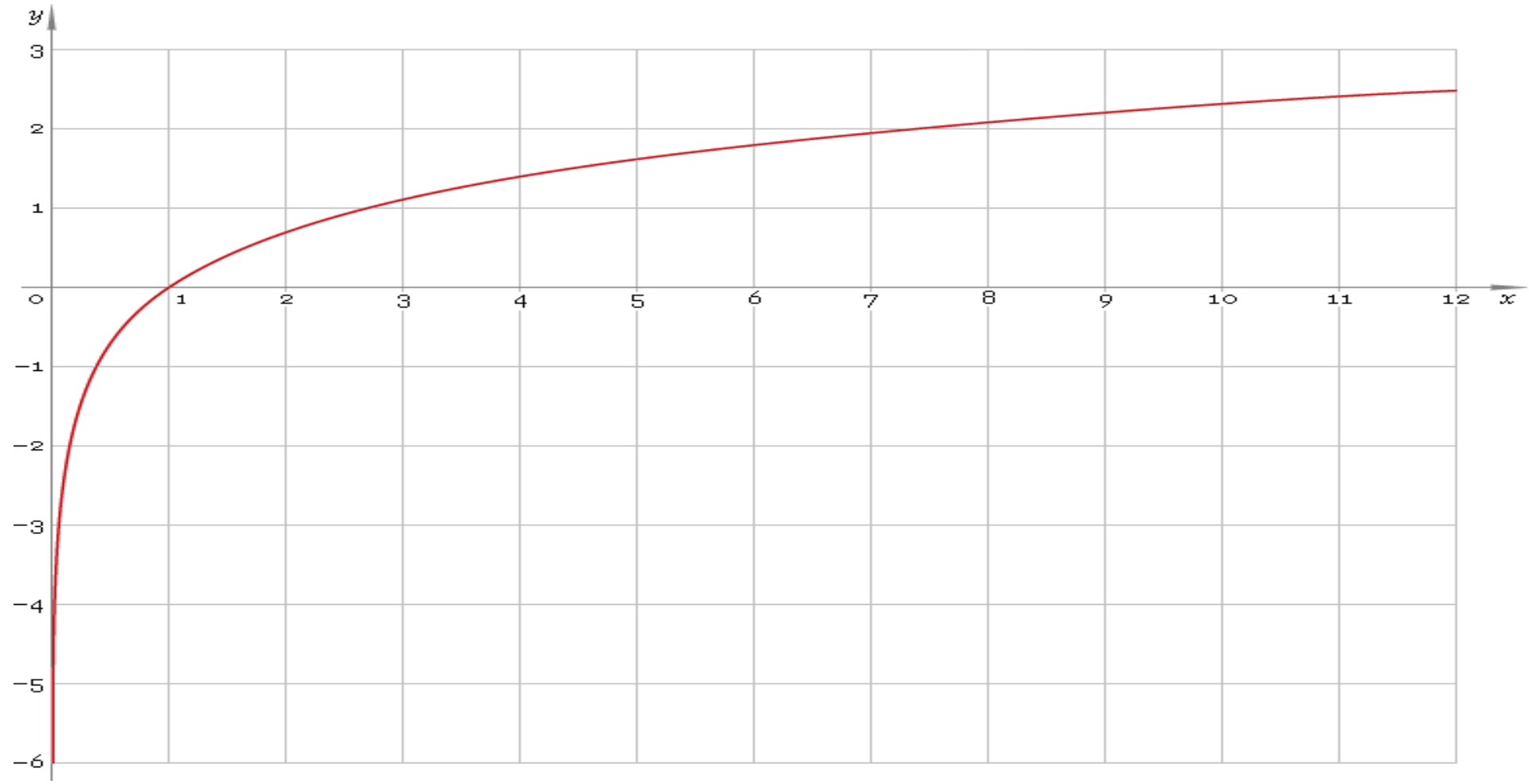
And again in order to estimate the model we take the natural log of both sides:

$$\text{Ln}\,\frac{P_i}{1-P_i} = \ln\,(e^{X\beta}) =$$

$$L_i = X\beta$$

Where L_i is the log of the odds ratio (known as the logit).

Quick sidebar on the natural log



- Remember that $X\beta$ stands for the scalar $\beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots X_k\beta_k$.

- This means that the model is linear both in X and in the parameters.

- But P and X are now not linearly related. Rather it is L, the log-odds ratio is linearly related to X.

- To more easily interpret it you can transform the effects into probabilities.

$$P_i = \frac{1}{1 + e^{-X\beta}}$$

- We do this by holding the other variables constant at their means and modes and vary the variable of interest.

- Now we are getting somewhere.

- We can now calculate the maximum likelihood using the following function:

$$\text{Ln } L = \sum_{i=1}^{n} y_i \ln\left(\frac{1}{1 + e^{(X\beta)}}\right) + (1 - y_i) \ln\left(1 - \frac{1}{1 + e^{(X\beta)}}\right)$$

- This means that the estimated β_k is the change in the log-odds ratio given a unit change in X_k. This is straightforward to understand but not straightforward to interpret.

Probit

- The probit model is a cousin of the logit model and is based on the normal distribution rather than the logistic.

- It is arguably more complicated because interpretation required reference to the probabilities under the normal curve instead of just using a calculator.

- Again remember the Bernoulli distribution.

$$Y = \begin{cases} 1, & \pi \\ 0, & 1 - \pi \end{cases}$$

Where:

$$\pi_i = f(X\beta)$$

and for probit = $\Phi(X\beta)$

Where $P(Y = 1) = \Phi(X\beta)$

and $P(Y = 0) = 1 - \Phi(X\beta)$

- The maximum likelihood function is therefore:

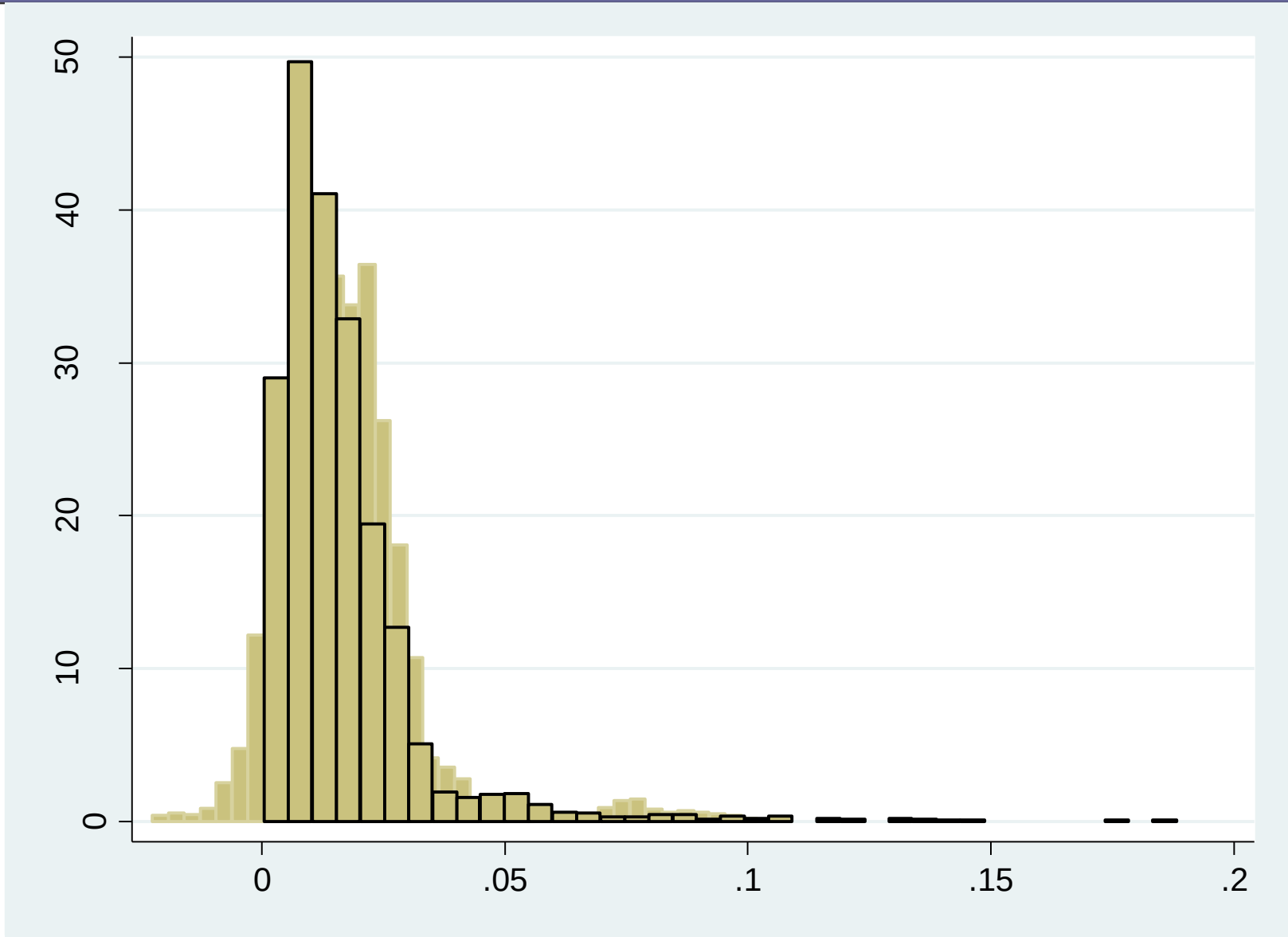
$$\text{Ln } L = \sum_{i=1}^{n} y_i \ln\Phi(X\beta) + (1- y_i)\ln\Phi(X\beta)$$

Let's try probit with the Fearon and Laitin data

```
. probit onset lgdpenl1 colbrit colfra polity2l nwstate

Iteration 0:   log likelihood = -534.46236
Iteration 1:   log likelihood = -505.50461
Iteration 2:   log likelihood = -502.37603
Iteration 3:   log likelihood = -502.37368
Iteration 4:   log likelihood = -502.37368

Probit regression                                 Number of obs   =       6326
                                                  LR chi2(5)      =      64.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -502.37368                       Pseudo R2       =     0.0600

------------------------------------------------------------------------------
       onset |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lgdpenl1 |  -.2817146   .0466158    -6.04   0.000      -.37308   -.1903493
     colbrit |  -.1110464   .0959122    -1.16   0.247     -.299031    .0769381
      colfra |  -.2006893   .1150581    -1.74   0.081    -.4261991    .0248205
    polity2l |   .0073926   .0065902     1.12   0.262    -.0055241    .0203092
     nwstate |    .681184   .1536992     4.43   0.000     .3799391    .9824288
       _cons |   -.018186   .3519194    -0.05   0.959    -.7079353    .6715634
------------------------------------------------------------------------------
```

- Substantively, the findings are similar to OLS, which is important to note. This speaks to the robustness of least squares.

- However, the predictions are different.


50
40
30
20
10
0
0
.05
.1
.15
.2

So how to chose between logit and probit?

- Unless we have theoretical expectations that inform us about the data in the tails of the distribution (and we really do not have detailed enough theories to tell us this) it really makes no difference which one you choose.

- We will talk more about interpreting logit and probit models next week.

How to interpret the intercept in ML models?

- A good question from last week.
- Long (1997:62-3) has the answer (*see?*).
- The intercept shifts the curve to the right or left, but the slope does not change.

- Also notice Panel B on page 63 for a way to think about how different betas shape the rate of change in the probability that y=1

Scobit

- Let's now turn to the Nagler (1994) article.

- He starts by discussing a crucial assumption of logit and probit—those most sensitive to change are clustered around .5.

- Another assumption is that the models are inherently interactive—the effects of one X on Y depends on the values of the other Xs.

- Basically, Nagler allows the response curve to vary according to the data by estimating an additional parameter, α.

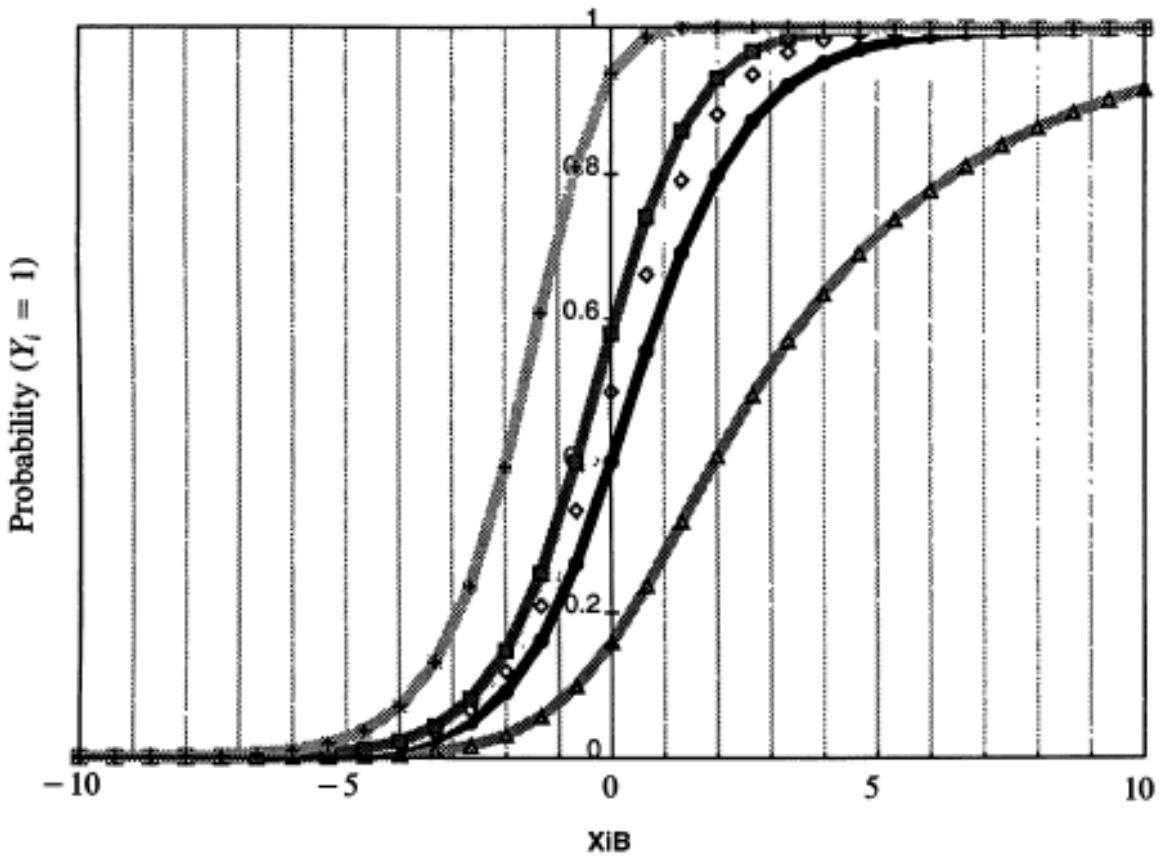
- Keeping the same notation as above:

$$P_i = \left(\frac{1}{(1 + e^{-X\beta})^{\alpha}} \right)$$

Where $\alpha > 0$

- Notice that the logistic distribution is nested within the Burr-10 distribution when $\alpha = 1$.

Figure 2.A. Cumulative Distribution for Scobit
[Y − Axis = 1 − $F(-X_i\beta)$]



Key:

▲α = .25

●α = .75

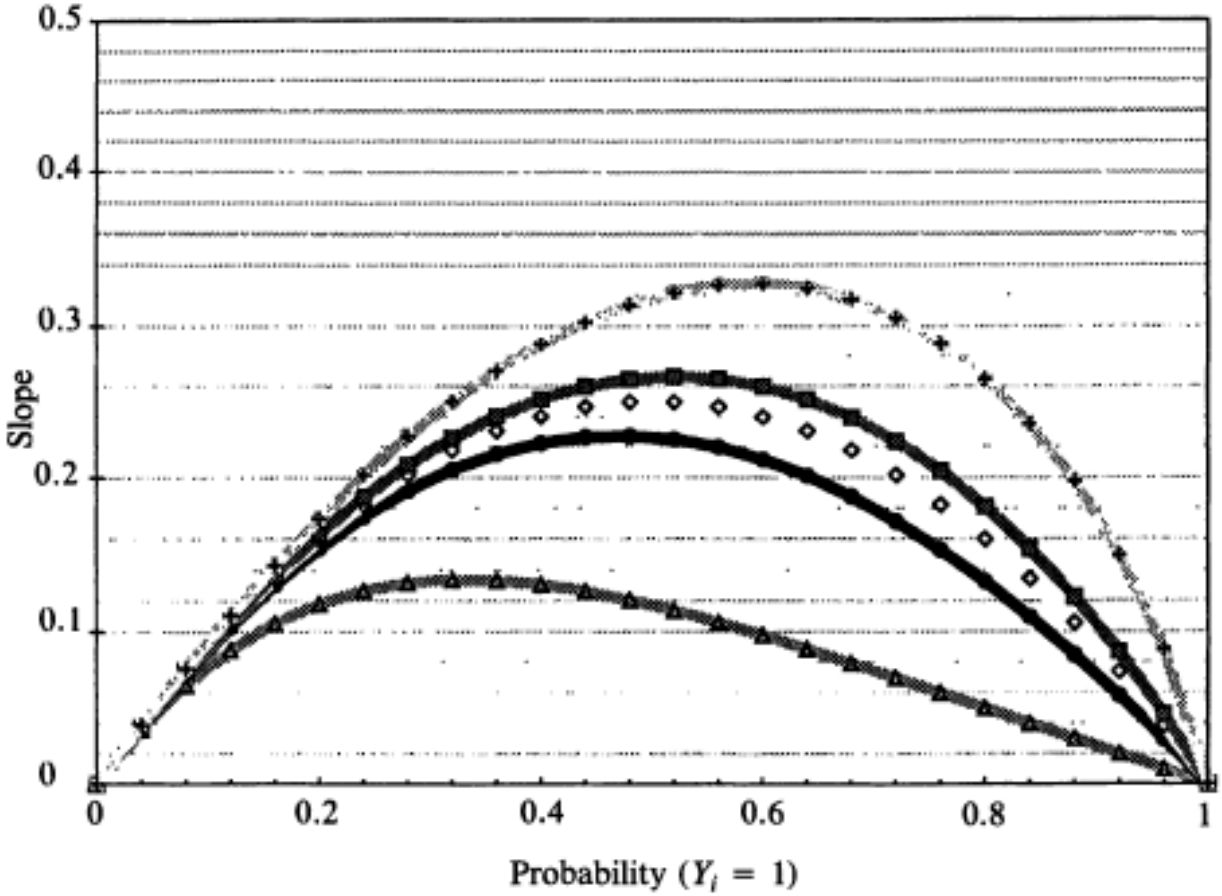
◇α = 1 (logit)

■α = 1.25

✱α = 4

Source: Nagler 1994: 237.

Figure 2.B. Slope for Scobit at Different Values of α versus Probability $Y = 1$



Key:

▲ α = .25

● α = .75

◇ α = 1 (logit)

■ α = 1.25

✱ α = 4

Source: Nagler 1994: 238.

- To conclude, binary dependent variables require shifting to maximum likelihood because what we know about the DV forces us to conclude that the assumptions underlying OLS are violated.

- You can chose among a number of different methods depending on the nature of your data, the theoretical knowledge you have about how they were generated, as well as (in the case of probit vs. logit) personal preference.

- Questions?

- Let’s play with the Fearon & Laitin (2003) data a bit more and see if we can make some substantive interpretations of the results that are more interesting and easier to understand than just estimated coefficients.