
API-202B Empirical Methods II

Session #7: Binary dependent variables

miguel_santos@hks.harvard.edu
@miguelsantos12

Today's class: Binary dependent variables



Politics

Sports

Science & Health

Economics

Culture

2016 ELECTION | 9:28 AM | FEB 17, 2016

Are You More Likely To Vote For A Woman Or A Man?

Republicans and Democrats answer that question very differently.

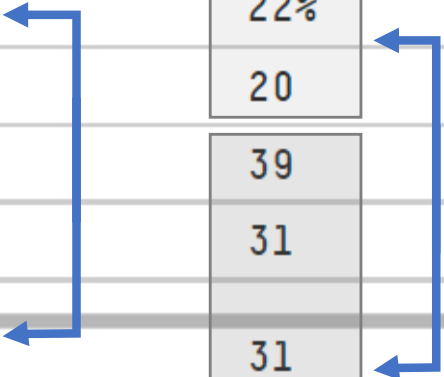
By CARL BIALIK

What's different about the outcome here than ones we've studied so far?

Are you more likely to vote for a woman or for a man?

Candidate gender preferences by presidential candidate support

SUPPORTERS OF ...	MORE LIKELY TO VOTE FOR ...		LESS LIKELY TO VOTE FOR ...	
	A WOMAN	A MAN	A WOMAN	A MAN
Hillary Clinton	40%	22%	6%	13%
Bernie Sanders	15	20	11	6
Donald Trump	13	39	28	7
Ted Cruz	12	31	20	4
Marco Rubio	9	31	23	3
Ben Carson	15	39	21	7
Jeb Bush	19	38	21	7



Poll of 2,030 registered voters, Feb. 10-11, 2016

What's different about the outcome here than ones we've studied so far?

Today's class: Binary dependent variables

- Introduction
- Percentage point vs. percent changes
- Linear probability model
- An example: Marriage in MA
- Probit model
- LPM vs. Probit models
- Takeaways
- Bonus:
 - STATA's "margins" command
 - Logit models

Introduction: Binary dependent variables

- The relationship between Y and X may not be a straight line, so this section of the course explores non-linear regressions
- A common example is when the dependent variable (Y) is binary:
 - Did a program impacted the decision of individuals attend college?
 - Are the banks' decisions on mortgages biased toward whites?
 - Were women more likely to vote for a women in the last election?
- We will study two regression models for binary dependent variables
 - I don't care, will use a linear probability model such as OLS
 - Probit model: using maximum likelihood estimation

A recap from API201 regarding percent and percentage points

- According to the Federal Reserve, the record of inflation in the United States over the previous 100 years (1917-2017) was 18.1%, registered the year after the end of World War II (1946). The year after (1947) – where the Marshall Plan was launched – ended with an inflation of only 8.8%
- Which of these statements regarding the events between 1946-1947 is true:
 1. The inflation rate decreased 51.4 percent
 2. The inflation rate decreased 9.3 percent
 3. The inflation rate decreased 51.4 percentage points
 4. The inflation rate decreased 9.3 percentage points
 5. 1) and 4) are true
 6. 2) and 3) are true
 7. None of them is true

Using a linear probability model (LPM)

- The **linear probability model (LPM)** is equivalent to the regressions we have seen so far, only that Y is a now dummy (not a continuous) variable:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

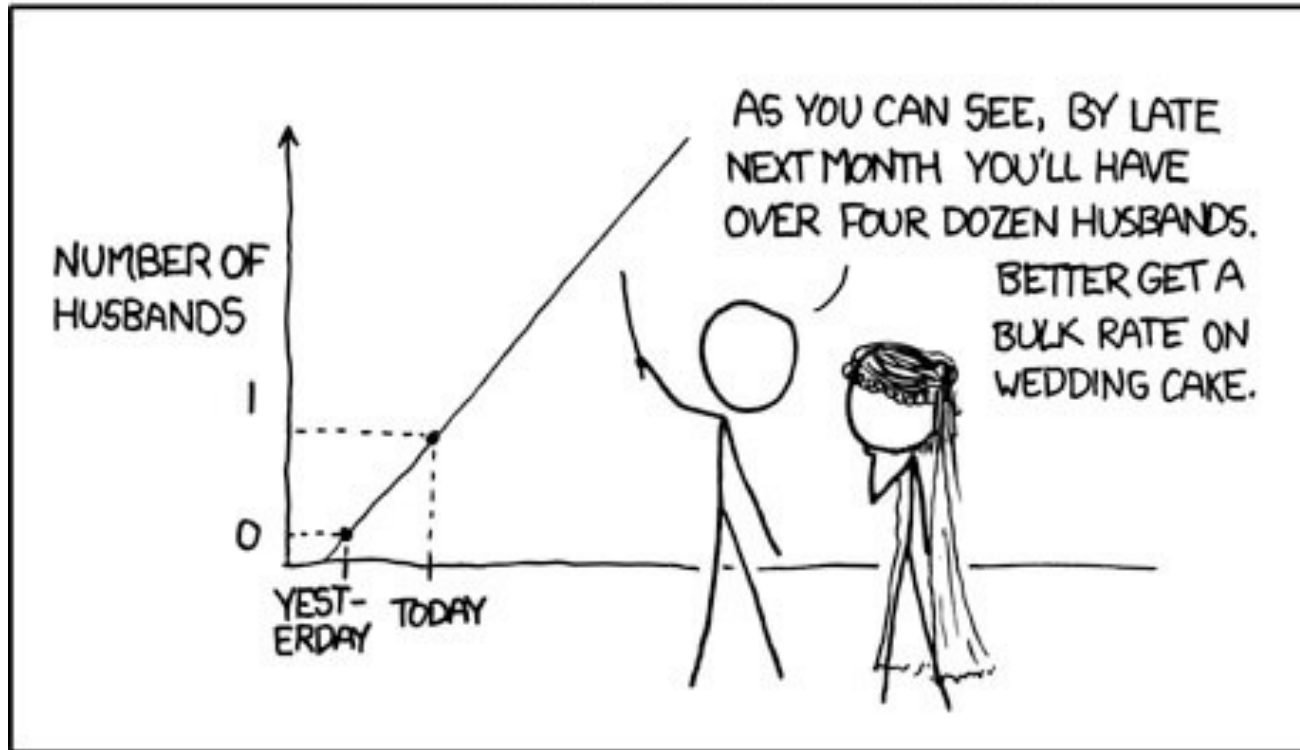
- Just as before, we use OLS to estimate the coefficients (i.e. we find the line of best fit by minimizing the sum of squared residuals)
- Having a binary dependent variable changes how we interpret coefficients:

A one unit increase in X is associated with a $100 * \beta_1$ percentage point change in the probability that $Y = 1$.

- Main advantage of the LPM: Coefficients are easy to interpret
- Main disadvantage of the LPM: can be below zero or above one

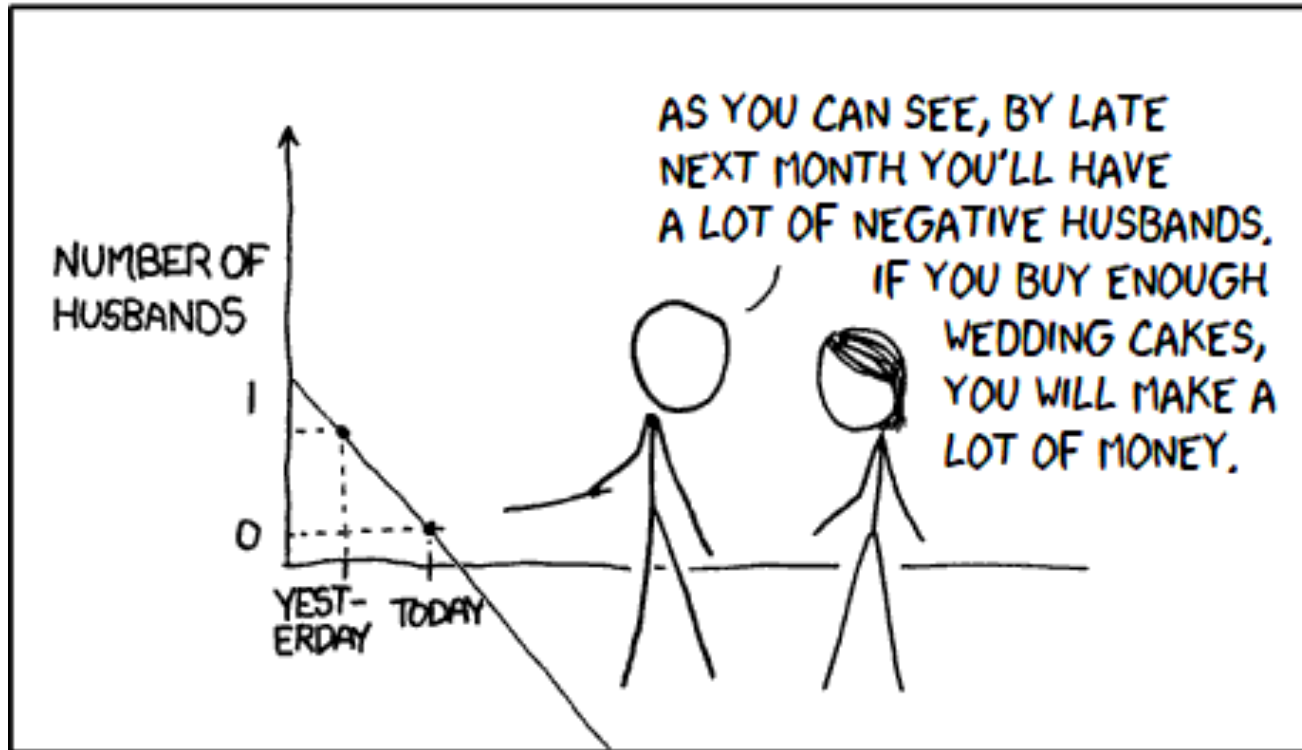
LPM: Predicted probabilities from this model can be above one...

MY HOBBY: EXTRAPOLATING



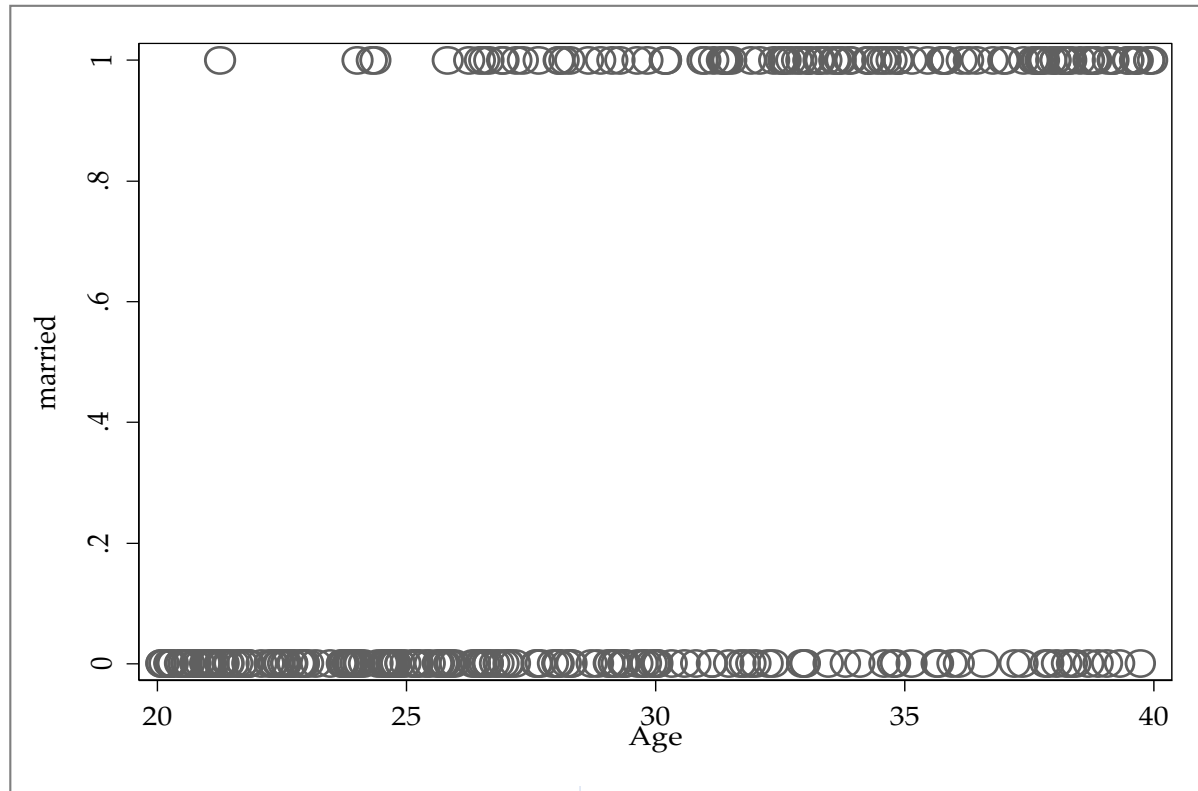
... or below zero (not very sensible from a math standpoint)

MY HOBBY: EXTRAPOLATING



Linear probability model (LPM): Marriage in MA

- Consider a random sub-sample of Massachusetts residents aged 20-40 from the 2009 American Community Survey
- Below is a graph of the relationship between an individual's age and whether that individual has ever been married:



Linear probability model (LPM)

- Here are some regression results from a multivariate linear probability model:

```
. reg married age female, robust
```

Linear regression

```
Number of obs = 300  
F( 2, 297) = 106.94  
Prob > F = 0.0000  
R-squared = 0.3422  
Root MSE = .39695
```

married	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0464059	.0033925	13.68	0.000	.0397294	.0530824
female	.0905176	.0462228	1.96	0.051	-.000448	.1814833
_cons	-1.021862	.0887222	-11.52	0.000	-1.196466	-.8472587

- What is the estimated sample regression function?

Linear probability model (LPM)

- How would you interpret the estimated coefficient on age?
- How would you interpret the estimated coefficient on female?
- What's the predicted probability of marriage for a:
 - 40 year old female?
 - 20 year old male?

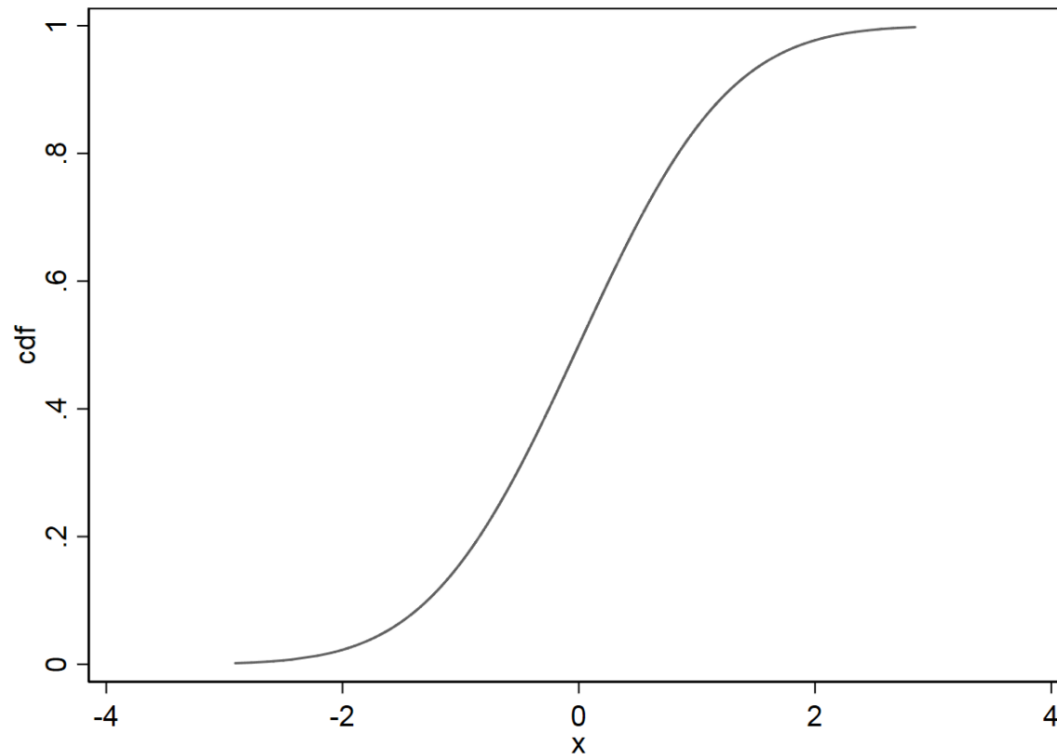
Linear probability model (LPM): The big picture

- Because the LPM can yield predicted values below 0 or above 1, some people argue that such models should not be used
- We are, however, frequently interested not in \hat{Y} but in $\Delta\hat{Y}$, the change in the probability of Y associated with a change in X
- For this, the LPM performs well in many circumstances
 - Particularly when making predictions based on values of explanatory variables near their sample means
 - We generally get predictions below 0 or above 1 only far from the mean value of X
- In practice, linear probability models are often used as a first (and sometimes final) try for many kinds of estimation tasks

Probit model

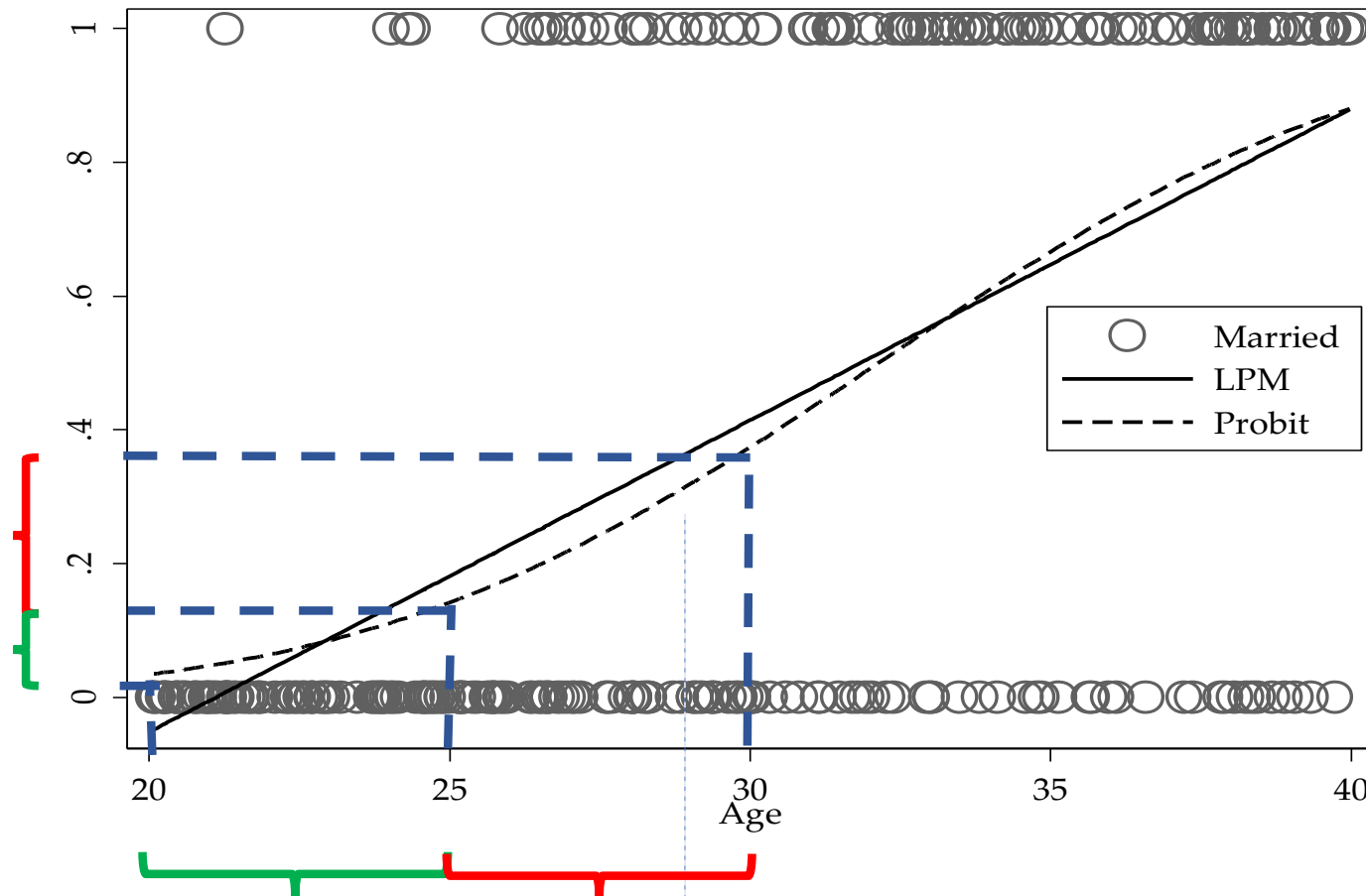
- **Probit** is a non-linear model that always predicts values between 0 and 1
- The probit model uses a standard normal cumulative distribution function (CDF) to model the probability of $Y=1$:

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_i)$$



What does it mean that the Probit model is non-linear?

- Let us see how both the LPM and probit models fit the marriage data



- The probit's slope predictions are similar to those of the LPM near X's sample mean, but can differ far from that sample mean.

Probit model

- The coefficient's sign and statistical significance can be read directly from the Stata output
- Unlike in the LPM, the **magnitude of probit coefficients** have no direct interpretation because the slope changes with X
- The only way to compute change in Y associated with a change in X is:
 - Compute the predicted probability for the initial value of X
 - Compute the predicted probability for the changed value of X
 - Take the difference of those two predicted probabilities.
- Computing the predicted probability is done by plugging the value of X into the probit equation, and looking for the result into the normal cumulative distribution function

Comparing LPM and Probit models: Using LPM on marriage data

- We will ask how both models to predict changes in the probability of marriage between the ages of 20 and 22, and between 30 and 32
- Here is the output from the linear probability model (LPM):

```
. reg married age, robust
```

married	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0466116	.0033851	13.77	0.000	.0399498	.0532734
_cons	-.984162	.0914573	-10.76	0.000	-1.164146	-.804178

```
. di _b[_cons]+20*_b[age]  
-.05192965
```

```
. di _b[_cons]+22*_b[age]  
.04129359
```

```
. di _b[_cons]+30*_b[age]  
.41418655
```

```
. di _b[_cons]+32*_b[age]  
.50740978
```

- What can you say about the relation between age and marriage?
- What's the difference in predicted marriage rates between
 - 20 and 22 year olds?
 - 30 and 32 year olds?

Comparing LPM and Probit models: Using LPM on marriage data

- Here is the output from the probit model:

married	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.1502393	.0158322	9.49	0.000	.1192087	.1812699
_cons	-4.828632	.4944724	-9.77	0.000	-5.79778	-3.859484

```
. di normprob(_b[_cons]+20*_b[age])  
.03408769
```

```
. di normprob(_b[_cons]+22*_b[age])  
.06383343
```

```
. di normprob(_b[_cons]+30*_b[age])  
.37393369
```

```
. di normprob(_b[_cons]+32*_b[age])  
.49163319
```

- What can you say about the relation between age and marriage?
- What's the difference in predicted marriage rates between
 - 20 and 22 year olds?
 - 30 and 32 year olds?

Where does Stata get those probabilities?

$$\Pr(Y = 1|X) = \Phi(-4.8286 + 0.1502 * X_i)$$

$$\Pr(Y = 1|X = 20) = \Phi(-4.8286 + 0.1502 * 20)$$

$$\Pr(Y = 1|X = 20) = \Phi(-1.82) = 0.0344 = 3.44\%$$

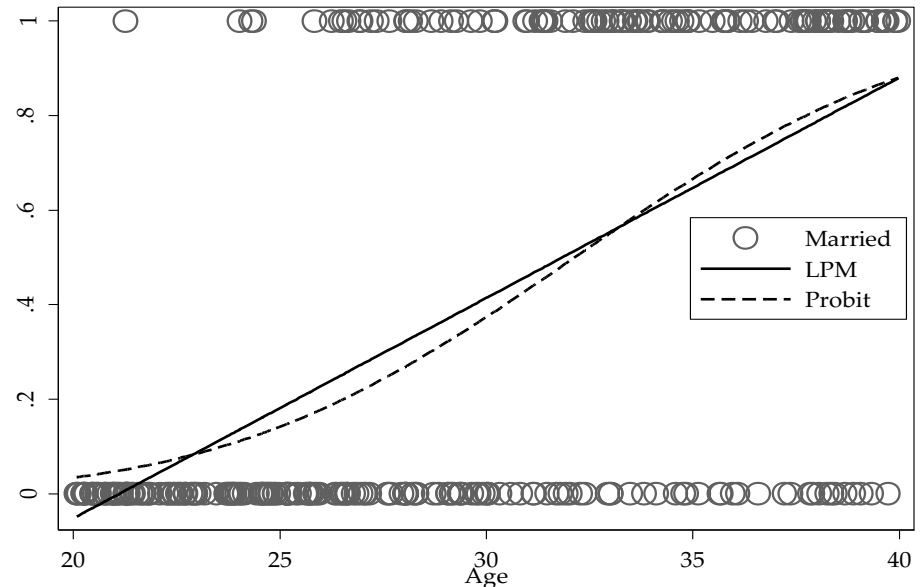
Table 1: Table of the Standard Normal Cumulative Distribution Function $\Phi(z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776

Comparing LPM and Probit models

- Note the similarities and differences in the models' predictions:

Predicted probability	LPM	Probit
Age = 20	-5.193% (zero)	3.478%
Age = 22	4.129%	6.383%
Difference	9.322% (4.129%)	2.905%
Age = 30	41.419%	37.393%
Age = 32	50.741%	49.163%
Difference	9.322%	11.770%



- You might have guessed these patterns from graphs of predicted values

Takeaways

- When our dependent variable is binary we can run a LPM or a Probit model
- LPM is easy to interpret and performs well in many instances, in particular when making predictions based on values of X 's near their sample means
 - LPM can yield predicted values below 0 or above 1, usually when estimating the probability for X 's far from their mean value
- The probit is a non-linear model that always predicts values between 0-1; uses a standard normal CDF to model the probability of $Y=1$
- Unlike in the LPM, the magnitudes of probit coefficients have no direct interpretation because the slope changes with X ; coefficients are better interpreted as probabilities, and changes in probabilities
- Computing the predicted probability is done by plugging the value of X into the normal cumulative distribution function
- The only way to compute change in Y associated with a change in X is: 1)
- Compute the predicted probability for the initial value of X
- Compute the predicted probability for the changed value of X
- Take the difference of those two predicted probabilities

A note on coming PSETs

- There's a modified version of Problem Set 2
- Please check if you have this version, Armando uploaded it on Wednesday
- Hand it in in Hard-copy and – please – stick to the format!
- Remember to be environmentally friendly and print on both sides. If you don't know how to set it, HKS library staff can help you

Vocabulary

- Percent versus percentage points
- Binary variable
- Linear probability model (LPM)
- Probit model
- Standard normal cumulative distribution function (CDF)
- Bonus vocabulary:
 - Marginal effects in a probit model
 - Logit model
 - Maximum likelihood estimation

Bonus: STATA's “margins” command in a probit model

- A simple way to generate interpretable probit coefficients is to use Stata's “margins” command after running a probit model
- This command evaluates the slope of the probit curve at a specific value of X to generate a marginal effect (slope at that point X)
- Typically, people evaluate the slope at the mean value of X
- The mean age in this sample is 29.4 years old

```
. margins, dydx(age) atmeans
```

```
Conditional marginal effects  
Model VCE      : OIM
```

```
Number of obs   =      300
```

```
Expression      : Pr(married), predict()
```

```
dy/dx w.r.t.    : age
```

```
at              : age          = 29.40959 (mean)
```

	dy/dx	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0551015	.0056533	9.75	0.000	.0440212	.0661818

Bonus: STATA's “margins” command in a probit model

- What is the interpretation of the marginal effect estimated above?
- For probits, additional explanatory variables complicate the picture slightly
- If we had used age and gender as explanatory variables, we could not simply calculate the difference in marriage probability between 20 and 22 year olds because that difference varies by gender
- We would either have to calculate those probabilities separately for males and females or compute the marginal effect of age at the mean age and mean gender (!) of the sample

Bonus: The Logit model

- Sometimes – in situations where the dependent variables is binary – you will see a **logit model**, which uses a **logistic cumulative distribution** - function to model the probability:

$$Pr(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- Logits are used in medical literature, where they are called logistic regressions and coefficients are reported as odds-ratios
- Logits and probits generally yield similar results to each other because the two distributions have roughly similar shapes
- Both logit and probit are also called methods of maximum likelihood estimators