Introduction to Choice Modeling

Data Science, General Assemb.ly

Nir Kaldero

University of California, Berkeley

Wednesday, 7 May , 2014

- Introduction to (applied) Choice Modeling
  - Learning how to leverage data & use predictive models
  - Takeaway: understand behavioral patterns & decision making process
- Discrete Choice Models
  - LPM – Linear Probability Model
  - Non-Linear Probability Models:
    - Logit (Log-Normal dist.)
    - Probit (normal dist.)
    - Nested-Logit
    - Random Coefficient (RD)
    - BLP
    - ….
- Practical Example
  - Motivation in Real-World Interface

How can we explain changes and differences between the choices we make – everyday?

- Choices?:
  - Whether I decide to work (be employed), or not?
  - Whether I decide to purchase 2% milk vs. non-fat milk?
  - Whether a firm decides to adopt a new technology?
  - Whether I decide to get married?
  - Whether Apple should invest in a new feature (or improve a current one)?

All of these are important everyday choices we want to understand

What can we do ?

- We can try to *understand how decisions are made* (what drives our decision to choose, behave, or act in a certain way..)
- We can try to *understand how different features/attributes affect* our decisions or our behavior

We will be able to make recommendations, create strategy, and polices

Example: Buy iPhone vs. Android?

How different attributes (e.g.: screen, design,.) or features (e.g.: Siri, Touch-Screen) affect our decision to buy an iPhone or other (Android)

Seems to be important for manufactures, marketers, and developers

MOTIVATION

In order to answer these questions we need to understand agents' behavior (e.g.: consumers, firms, policies)

→ we need to define and estimate Choice Models –Discrete (binary) or Continues

- We will focus on Discrete Choice Models

- Discrete Choice Models - A binary Choice:
  - All of these questions deal with binary choices – 0 or 1 (notation: $outcome \equiv y = (0,1)$)
    - Examples:
      - Be employed, or not? → emp(0,1)
      - Decided to purchase 2% milk or non-fat milk? → milk2%(0,1)
      - Firm decided to adopt a new technology? → platform(0,1)
      - Get married? → married(0,1)

FROM THEORY TO EMPIRICS

A Fresh Reminder:

- We are living in a new era! – Big Data
- There are things we know (x_k) and there are some that we don't know (u)
- OLS regression : $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3,...,+\beta_k x_k+u^{error}$

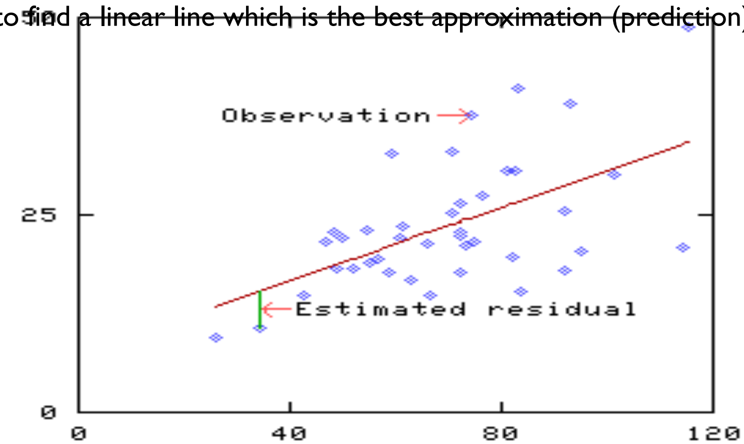
outcome

attributes

Un-known information for the econometrician

What are we trying to do? –Best Approximation

We want to find a linear line which is the best approximation (prediction) given all the data we have



- The method? - We minimize the 'error-term'/'residual' (distance between the points) : $MIN(u^2)=MIN((y-x\beta)^2)$
- OLS is a linear regression – the effect of the estimated parameters (β_k) on the outcome (y) is linear (i.e., constant)
- How do we interpret the results? – one unit change in x_k (increase/decrease) will change y by β_k ('linearity')

MODELS OF DISCRETE CHOICE

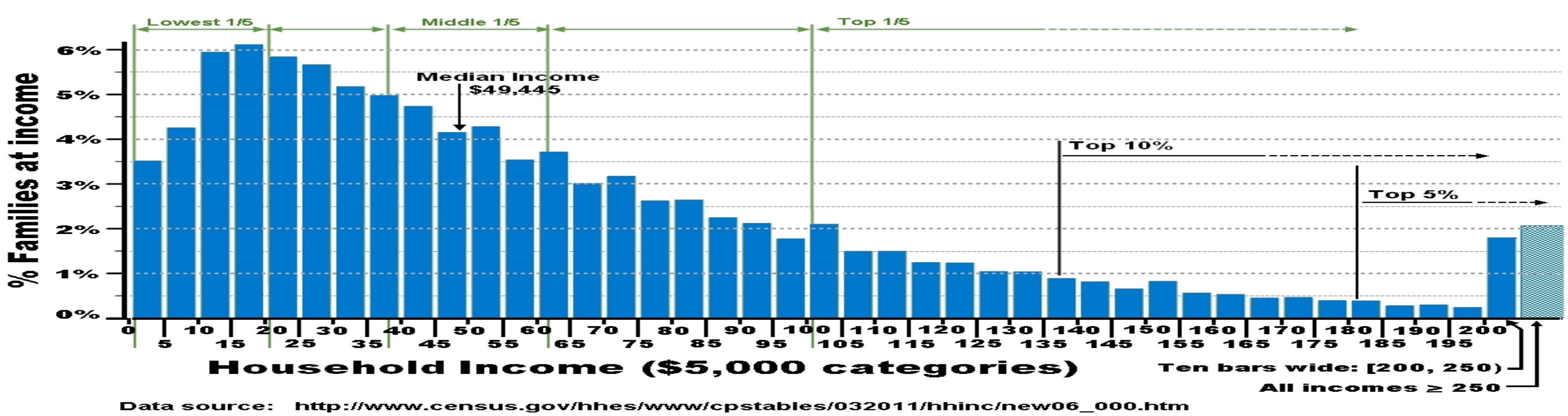
Three common models:

- LPM – Linear Probability Model
- Non-Linear Models (Advanced)
  - Probit (assuming Normal dist.)
  - Logit (assuming Log-Normal dist.)
- Each model has its own features (*assumptions*)
- Each model has its own pros and cons

The most important question in the industry (also in academia): *How to choose the 'right' model?*

A: depends on the *assumptions we make on the distribution of the error-term* (Log, Normal, etc.)

Example: It is known that income (proxy for employment) is Log-Normal distributed (Why?)



MODELS OF DISCRETE CHOICE – LPM

LPM – Linear Probability Model

In general the empirical model is:

$$y_{it} = \beta_{0t} + \beta_{1t} x_{1t} + \beta_{2t} x_{2t} + \beta_{3t} x_{3t} ,.., \beta_{kt} x_{3t} + u_{it} \; ; Where: y_{it} = 1 \; or \; 0$$

→ The LPM is a simple OLS regression with a binary dependent variable $y_{it} = emp(1,0)$

Why to choose this model:

- Pros: Easy to estimate and compute ☺
- It's generally accepted that the unknown information (unobserved to us) is normally distributed across our sample
  - Intuition: Choices are made in a *random way* (with a mean of 0 – on average)

Assumptions:

1. Exogenous – no correlation between x_k (the variables) and the error-term u_{it} → $corr(x_k, u) = 0$

   If $corr(x_k, u) \neq 0$ → the estimators (β_k) are biased!

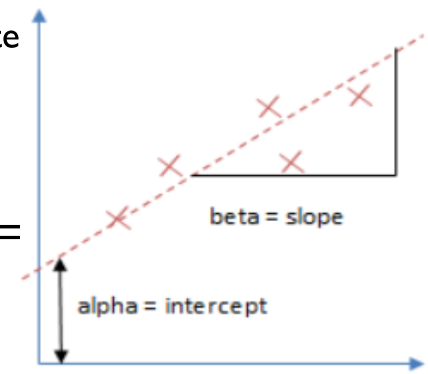
2. The error-term is normally distributed ($u \sim normal(\mu, \sigma^2)$)

MODELS OF DISCRETE CHOICE – LPM

In Practice - Since y_{it} is now a binary choice (1,0):

- The *outcome* (y) *gets a probability interpretation* (different from OLS)
- We *should define the 'Probability of Success'* - $Prob(y=1)$ based on our inte

How should we interpret the estimated coefficients (results - betas)?



β_k is the expected change in the probability of 'success' - $Prob(y_i =$

$$\beta_k = \partial Prob(y_i = 1|X) / \partial x_j \quad where \; x_j \in X$$

- The effect of β_k is linear on the outcome (y) and from here the name – LPM

Some bad news:

The expected (predicted) probability is not necessarily defined between 0-1 (does not make sense..)

LPM – REAL EXAMPLE

Question: How do having children affect married women's choice to work (be employed) ?

- Seems to be an important question in order to understand unemployment rate and to define optimal strategies/policies

Data – Israeli Labour Force Survey for the years 1985-2010 (a panel data – time series)

- Notation: Observation → i ; Year → t
- Variables: x_k

1. year – year of the survey
2. Sex – male (1), female (0)
3. Age
4. Marital status (1= married, 2= divorced, 3= widow, 4= single, 5= married live alone)
5. Schooling – years of education
6. Working_hours – number of hours at work (per week)
7. emp – 1 (yes) 0 (no) [if working_hours > 10 a week)]
8. ..
9. Controls (demographics), etc

- We need to choose from this huge data-set only: married women who have children

LPM – REAL EXAMPLE

We will use python in order to run an OLS simple regression with binary dependent variable - LPM model:

Source	SS	df	MS
Model	796.424037	7	113.774862
Residual	4880.18313	22760	.214419294
Total	5676.60717	22767	.249334878

Number of obs	=	22768
F(7, 22760)	=	530.62
Prob > F	=	0.0000
R-squared	=	0.1403
Adj R-squared	=	0.1400
Root MSE	=	.46305

emplo	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
schooling	.0317368	.0007409	42.84	0.000	.0302846	.0331891
age	.0617393	.0028811	21.43	0.000	.0560922	.0673864
age_sq	-.0007611	.000032	-23.81	0.000	-.0008238	-.0006985
children_0_4	-.0710904	.0048222	-14.74	0.000	-.0805422	-.0616386
children_5_9	-.0391121	.0043027	-9.09	0.000	-.0475458	-.0306785
children_10_14	-.0485788	.0044201	-10.99	0.000	-.0572426	-.039915
children_15_17	-.0499378	.0064346	-7.76	0.000	-.06255	-.0373256
_cons	-.9485538	.0620296	-15.29	0.000	-1.070136	-.8269716

- All the variables are statistically significant (p-value)
- All variables are consistent with our intuition (signs)
- How to interpret the results? (recall):
  - Each additional schooling year increases the *probability* of being employed by 3.2 biases point (0.317)
  - Having children between the ages of 0-4 decrease the *probability* of being employed by 7.1% (- 0.710)

This model – Discrete Choice – can help us *understand our behavior in real life circumstances*

LPM – EXAMPLE (AND SOME PROBLEMS..)

Problems with LPM model:

Variable	Obs	Mean	Std. Dev.	Min	Max
emp10	22768	.5260014	.4993344	0	1
emp_hat	22768	.5260014	.1870334	-.4886191	1.172533

- The predicted probability is not necessarily defined between 0-1

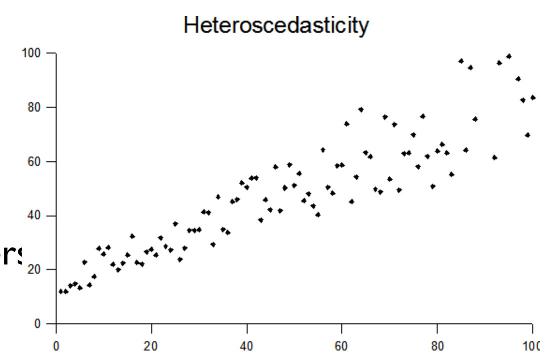
Why? For some observations that prediction of the model result is : $y_{it} \equiv emp_{it} < 0 \text{ or } emp_{it} > 1$

If w

Variable	Obs	Mean	Std. Dev.	Min	Max
emp_hat	497	-.0502759	.268012	-.4886191	1.172533

natic: (497/22,768)

- Another disadvantage of the LPM – Heteroscedasticity:
  - The variance across agents (observations) changes across our sample
  - Some observations ('agents') have different variabilities (std.) from others
  - Heteroscedasticity can invalidate statistical tests of significance
  - The estimators are not biased!



We can easily fix this in python

LPM – CONCLUSIONS

LPM –

- Easy to estimate (OLS regression)
- The predicted ('expected') probability is not necessarily between 0-1
- The effect of the parameters (β_{it}) on the expected/predicted probability is constant

  (each change in x_{it} will increase/decrease the probability in a constant fashion)

How can we overcome these crucial issues?

- There are more sophisticated models of discrete choice such as:
  - Probit (assuming standard normal distribution)
  - Logit (assuming standard log-normal distribution)

PROBIT/LOGIT MODEL

The general model (like LPM) tries to predict the 'probability of success':

$$\text{Prob}(y_{it}=1|\ x_j)=Prob(y_{it}=1\ |x1,\ x2,\ x3,\ x4,\ldots,\ x_k)$$

The general form of the model is: $Prob\, y_{it}=1 X=G(\beta_0+\beta_1 x_1+\beta_2 x2\ ,\ldots,+\beta_k x_k)$

$$s.t.:0<G(z)<1$$

- In order to ensure that the predicted values will be between 0-1 ($0<Prob(\cdot)<1$) we need to choose a function ($G(z)$) that satisfies this constrain
  - $G(z)$ - can also be a non-linear function (the effect of the β_{it} varies across observations)

There are two useful functions:

- The logistic function (Logit Model)
- The standard normal function (Probit Model)

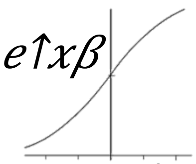
PROBIT/LOGIT MODEL

Functions properties

Logit

- In the logit model the function $(G(z))$:

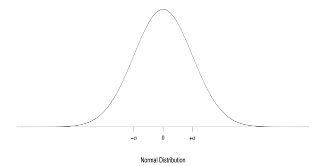
$$G(z)=\frac{e^{x\beta}}{1+e^{x\beta}}$$



This is the CDF of the standard logistic distribution function

Probit

- In the Probit model the function $(G(z))$:



$$G(z)=\int_{-\infty}^{z}\phi(v)dv=\theta(z)$$

And $\phi(z)=2\pi^{-1/2}\exp(-z^2/2)$

This is the CDF of the standard normal distribution function

Both functions are:

- Increasing
- ~Equal to 0 when Z goes to $-\infty$
- ~Equal to 1 when Z goes to ∞
- Symmetry around 0 : $1-G(z)=G(z)$

In general we can present logit/probit models as a sub-section of latent variable: $y^* =\beta_0 +x\beta+u$, $y=1\ if\ [y^* >0]$

ESTIMATION (IN PRACTICE)

These models are not linear (the functions) → we cannot estimate them using OLS methodology

How do we do it? – using Maximum Likelihood Estimation process

- The log-likelihood of the observations in the sample is:

$$\log L(\beta; y_1, x_1, y_2, x_2, y_3, x_3, \ldots, y_n, x_n) = \sum_{i=1}^{n} \{ y_i \log[G(x_i \beta)] + (1-y_i) \log[1-G(x_i \beta)] \}$$

- The function is non-linear and so there is no close form solution (analytic) for the estimators.

  We are using numeric estimation in order to compute the values of each β_k

- The intuition behind the process:

1. Start with a random 'guess' about the magnitude of the coefficients (β_k)
2. Compute the log-likelihood function (from above)
3. With respect to the sign of the first derivative we choose another close 'guess' (higher or lower value) – and compute once again the log-likelihood
4. Continue (2-3) until you reach the point at which there is no change in the result of the log-likelihood expression formula (converge)

ESTIMATION - IN PRACTICE

- In order to compute the Logit Model in Python use (Lab):

$$Prob(y_i = emp(1)) = \beta_0 + \beta_1\ schooling + \beta_2\ age + \beta_3\ (age)^2 + \beta_4$$

```
Iteration 0:   log likelihood = -15750.775
Iteration 1:   log likelihood = -13979.924
Iteration 2:   log likelihood =  -13960.74
Iteration 3:   log likelihood = -13960.718
Iteration 4:   log likelihood = -13960.718

Logistic regression                               Number of obs   =      22768
                                                  LR chi2(7)      =    3580.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -13960.718                       Pseudo R2       =     0.1136
```

emp10	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
schooling	.1644975	.0041895	39.26	0.000	.1562863	.1727087
age	.2820811	.0136117	20.72	0.000	.2554026	.3087596
age_sq	-.0035001	.0001516	-23.09	0.000	-.0037973	-.003203
children_0_4	-.3870174	.0240956	-16.06	0.000	-.4342438	-.3397909
children_5_9	-.2057233	.0207747	-9.90	0.000	-.246441	-.1650056
children_10_14	-.2434284	.0211916	-11.49	0.000	-.2849632	-.2018935
children_15_17	-.2569445	.0308171	-8.34	0.000	-.3173451	-.196544
_cons	-6.805008	.2963469	-22.96	0.000	-7.385837	-6.224178

- Some questions:
  - Which coefficient is/are significant? Consistent with our intuition?
  - What is the [expected] probability that a women with 16 years of schooling, in the age of 31, and with 0-4 years old children – will go to work (be employed)?

ESTIMATION - IN PRACTICE [EXPECTED PROB]

- What is the [expected] probability that a women with 16 years of schooling, in the age of 31, and with 0-4 years old children – will go to work (be employed)?

- Let's do it together:
  - We want to compute
    - Schooling years :

```
Iteration 0:   log likelihood = -15750.775
Iteration 1:   log likelihood = -13979.924
Iteration 2:   log likelihood =  -13960.74
Iteration 3:   log likelihood = -13960.718
Iteration 4:   log likelihood = -13960.718

Logistic regression                               Number of obs   =      22768
                                                  LR chi2(7)      =    3580.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -13960.718                       Pseudo R2       =     0.1136
```

emp10	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
schooling	.1644975	.0041895	39.26	0.000	.1562863	.1727087
age	.2820811	.0136117	20.72	0.000	.2554026	.3087596
age_sq	-.0035001	.0001516	-23.09	0.000	-.0037973	-.003203
children_0_4	-.3870174	.0240956	-16.06	0.000	-.4342438	-.3397909
children_5_9	-.2057233	.0207747	-9.90	0.000	-.246441	-.1650056
children_10_14	-.2434284	.0211916	-11.49	0.000	-.2849632	-.2018935
children_15_17	-.2569445	.0308171	-8.34	0.000	-.3173451	-.196544
_cons	-6.805008	.2963469	-22.96	0.000	-7.385837	-6.224178

  - In order to compute y we need to calculate the logistic G(z) function: $G(z)=e^{x\beta}/1+e^{x\beta}$

$$y_i = G(z)=e^{x\beta}/1+e^{x\beta} = e^{(\beta_0+\beta_{schooling}16+\beta_{age}31+\beta_{agesqr}31^2+\beta_{child04}1)}/1+e^{(\beta_0+\beta_{schooling}16+\beta_{age}31+\beta_{agesqr}31^2+\beta_{child04}1)} = 0.8236$$

$$y_i = G(z)=e^{x\beta}/1+e^{x\beta} = e^{(\beta_0+\beta_{schooling}16+\beta_{age}45+\beta_{agesqr}45^2+\beta_{child04}1)}/1+e^{(\beta_0+\beta_{schooling}16+\beta_{age}45+\beta_{agesqr}45^2+\beta_{child04}1)} = 0.8538$$

$$\frac{\exp(-6.0805+16\times 0.1644+31\times 0.282+31^2\times(-0.0035)+1\times(-0.387))}{1+\exp(-6.0805+16\times 0.1644+31\times 0.282+31^2\times(-0.0035)+1\times(-0.387))}$$

INTERPRETING THE RESULTS (LOGIT)

- Since $y_i = f(0,1)$ we cannot interpret the estimators/coefficients β_i as we did in the simple OLS model.
  - Recall (OLS): one unite change (+/- 1) in x_i increases/decreases the outcome y_i by β_i
- We need to go back to the functions and compute the predicted value for y_i

  (because the function $G(z)$ is not a linear function (OLS)
  - However, the sign of the estimators (β_i) can be interpret immediately – always in the same direction
- There are 4 types of variables, and therefore there are 4 cases for interpreting the coefficients:
  1. Case 1: x_i is a *continues* variable (think about angles ($0° - 360°$), age, incme….)
     1. You need to compute the direct effect (or 'odds ratio')

$$\beta_j = \partial Prob(y=1)/\partial x_j = \partial G(\beta_0 + x\beta)/\partial x_j = g(\beta_0 + x\beta)\cdot\beta_j,$$

$$where\ g(z) in\ logit\ equals\ to: g(z) = e^{\beta_0 + x\beta}/(1+e^{x\beta})^2$$

Pay attention that in this model the effect of a singular estimator (β_j) depends on all other estimators in the regression $(\beta_0 + x\beta)$, $where\ x\beta = \beta_1 x_1 + \beta_2 x_2$ ….

INTERPRETING THE RESULTS (LOGIT)

1. Case II: x_i is a *dummy (binary)* variable (insurance (1,0))

   1. You need to compute the difference between $|x_i(=1) - x_i(0)|$

$$\beta_j = \partial Prob(y=1)/\partial x_1 = G(\beta_0 + \beta_1 1 + \beta_j x_j) - G(\beta_0 + 0 + \beta_j x_j)$$

$$where\ G(z) in\ logit\ equals\ to: G(z) = e^{x\beta} / 1 + e^{x\beta} \quad [CDF]$$

There are many more cases of course.. Python can do the job for

Let's go back to our example –

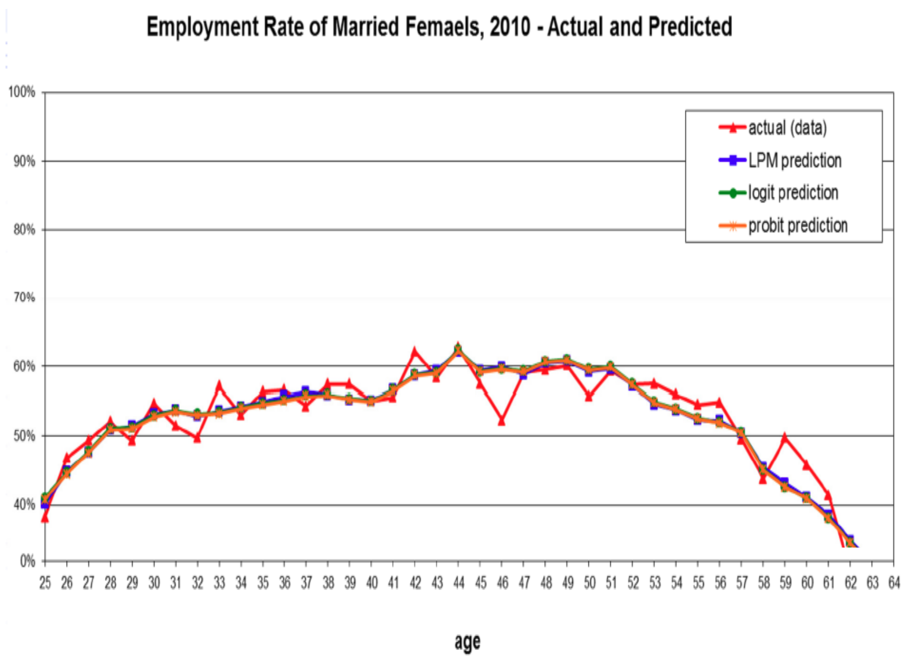
	LPM	logit	probit
schooling	0.032	0.164	0.098
age	0.062	0.282	0.173
age_sq	-0.001	-0.004	-0.002
children_0_4	-0.071	-0.387	-0.234
children_5_9	-0.039	-0.206	-0.124
children_10_14	-0.049	-0.243	-0.147
children_15_17	-0.050	-0.257	-0.156
_cons	-0.949	-6.805	-4.133

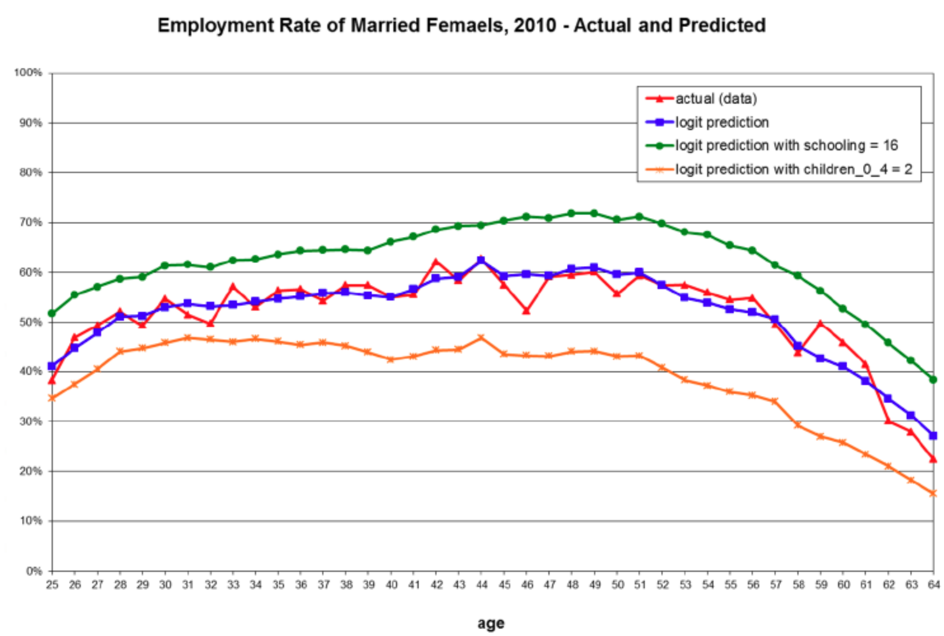
- The sign of the coefficients are all the same (direction)
- We cannot intuitively interpret the magnitude of the coefficients in the logit/probit models

ESTIMATION (& FITNESS)

We can see that we overcome the biggest issue with the LPM model ($0 > Prob(\cdot) < 1$):

Variable	Obs	Mean	Std. Dev.	Min	Max
emp10	22768	.5260014	.4993344	0	1
emp_lpm	22768	.5260014	.1870334	-.4886191	1.172533
emp_logit	22768	.5260014	.1901581	.0052596	.9680831
emp_probit	22768	.5247376	.1885883	.0008062	.9791964





EXAMPLE & CONCLUSIONS

Discrete Choice Models are very useful in order to understand (and predict) consumers' behavior

Allowing us to create Optimal Strategies

Suppose you are the Head of Marketing at Target

- You want to understand why consumers are choosing 2% milk vs. fat-milk?
  (effects on revenues, promotions, demand, prices, etc.)

- You Have the Data! (e.g.: purchases, product's attributes, expenses, # time-bought..)

- You can use Discrete Choice Models ($y = milk2\%(1,0) | x$) in order to better understand your consumers → predict their behavior

- It has a large effect on defining Optimal Strategies (Operational, Marketing, etc.)
  - Tailored Promotions (and discounts – shifting demand)
  - Psychological Manipulations (buy 1 pay 3$, buy 2 pay 6$ - buying in bundle)
  - "Healthier Campaigns" – converting consumers to buying healthier products
  - Large effect on revenues and operations

LAB