# ICPSR 2018

# Modeling Categorical Outcomes

## *Advanced methods of interpretation*

Scott Long - β1 Draft - 2018-04-08



mco18-part1-beta1-2018-04-08.docx

# Table of contents

















# β1 Introduction

## Readings

*Long & Freese: Chapters 1 and 2*

- Check there for references to other sources

## Examples

1. Do-files and data for lecture examples are available
   - In Stata, run **`search mcosetup`**
   - **`mdo`***`year`***`-`***`topic`***`.do`**
2. Lectures do not show all of the code
3. Use these command files as templates for your analysis

## Cross-sectional models for categorical outcomes

1. Binary outcomes: binary logit and probit
2. Nominal outcomes: multinomial logit
3. Ordinal outcomes: ordinal logit and probit

### Focus on advanced methods of interpretation

1. *Telling a story* in the presence of nonlinearity
2. Regression coefficients are necessary but not sufficient
   - o Avoid signs and stars approach
3. Interpretation using predictions transform the estimated parameters
   - o Predictions conditional on values of regressors
   - o Marginal effects of regressors



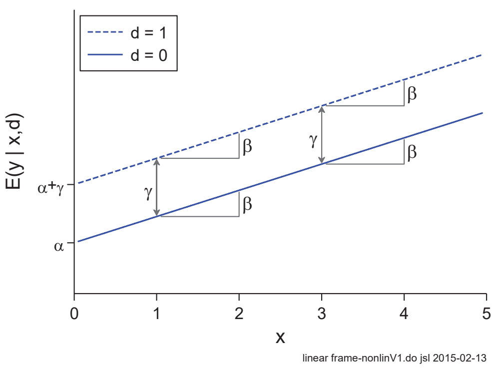
## Nonlinear models

1. In linear models the effect of $x_k$ on y does not depend on where it is evaluated
   - o Unless nonlinearities are introduced with interactions or transformations
2. In nonlinear models the effect of $x_k$ depends on:
   - o The value of $x_k$
   - o The values of other x's
3. Most models for categorical outcomes are implicitly nonlinear
4. In linear models, most of the work is done when the model is fit
5. In nonlinear models, the work begins
   - o Nonlinearity make things harder and more realistic



### Linear model

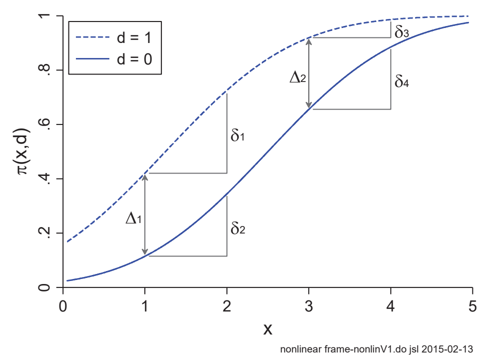
$$y = \alpha + \beta x + \gamma d$$






### Inherently nonlinear models

$$y = \frac{\exp(\alpha + \beta x + \gamma d)}{1 + \exp(\alpha + \beta x + \gamma d)}$$






### RHS (right-hand-side) variables are linear combinations

1. Notation
   - o $\mathbf{x}_i\boldsymbol{\beta} = \alpha + \beta x_i$
   - o $\mathbf{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Income}_i$
   - o $\mathbf{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1 \text{x}_{i1} + \beta_2 \text{x}_{i2} + \cdots + \beta_K \text{x}_{iK}$
2. Linear combinations can include
   - o Product terms (e.g., $x_3=x_1*x_2$)
   - o Transformed regressors (e.g., $x_1 = \sqrt{w_1}$ or $x_2 = w_1^2$)
3. With CDA, these enhancements lead to unexpected subtleties



## Software

1. How you interpret models depends on your software
   - o If post-estimation analysis is hard, you are unlikely to do it
2. Stata has great tools for post-estimation analysis
   - o **`margins`** and related commands
   - o **`suest`** and **`gsem`** for simultaneously fitting models
3. Other packages
   - o R
   - o SAS
   - o SPSS

## Roadmap

1. Linear regression model (LRM)
2. Binary regression models (BRM)
3. Estimation, testing, and fit
4. Testing marginal effects (ME)
5. Nonlinearities on the RHS (right-hand-side)
6. Comparing groups
7. Comparing effects across models
8. Nominal regression models (NRM)
9. Ordinal regression models (ORM)
10. Generalized marginal effects (GME)



*Tool: Locals in Stata*

1. Macros are abbreviations representing characters or numbers.
2. Syntax:

```
local local-name "string"
local local-name = expression
```

3. For example,

```
local rhs "var1 var2 var3 var4"
    local ncases = 198
```

4. To display a local:

```
. local OPTmark "msym(square circle) mcol(red blue) jitter(5)"
. di " `OPTmark'"
msym(square circle) mcol(red blue) jitter(5)
```

5. The opening quote ` and closing quote ' are different.

*Why is it called local?*

1. Local macros exist only when a do-file is running.
   - When that program ends, the macro disappears
2. This makes do-files robust since everything is defined in the do-file.



*Example: a provenance tag*

1. My do-files include a local to document provenance:

```
local pgm mypgm1
local dte 2018-04-02
local who Scott Long
local tag `pgm'.do `who' `dte'
```

2. I can display the tag:

```
. di "`tag'"
mypgm1.do Scott Long 2018-04-02
```

*Tool: Global macros*

1. Global macros are created as:

```
global vars "x1 x2 x3"
```

2. Content is retrieved using **$***globalname*

```
display "$vars"
```

3. Globals can make do-files fragile since they stay in memory until you delete them or leave Stata.



# β1 Linear regression

## Readings and examples

*Long & Freese: Chapters 3 and 4*

*mdo18-lrm-*.do*

## Objectives

1. Establish notation and terminology
2. Reinforce the ideas of linearity and nonlinearity
3. Explain identification
4. Introduce maximum likelihood estimation
5. Introduce **margins** based commands for post-estimation



## Notation

### Outcome = linear combination + error

1. $y_i = \alpha + \beta x_i + \varepsilon_i$
1. $Occupation = \beta_0 + \beta_1 Education + \beta_2 ParentEd + \beta_3 ParentOcc + \varepsilon$
2. $y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$

$$= \begin{bmatrix} 1 & x_{i1} & \ldots & x_{iK} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \varepsilon_i$$

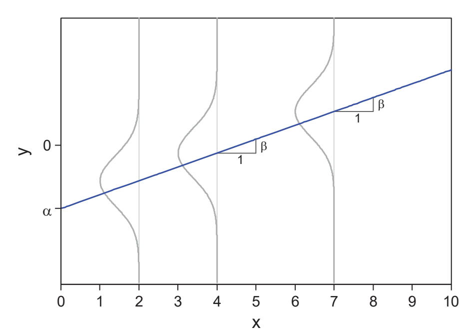
### ε is unexplained variation

1. Randomness
2. Unobserved heterogeneity.



## Assumptions

1. Linearity.
2. Not perfect collinearity.
3. E(ε|x)=0.
4. Homoscedasticity.
5. Uncorrelated errors.
6. Normality.



## Linearity

1. $y$ is linearly related to the $x$'s through the β's

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

   o A unit change in $x_1$ has a constant effect on y

## Collinearity

1. Multiple regression is used since the $x_k$'s are collinear
2. The $x_k$'s cannot be perfectly collinear

## Homoscedasticity

1. All observations have the same variance for ε.

$$Var(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2 \quad \text{for all } i$$

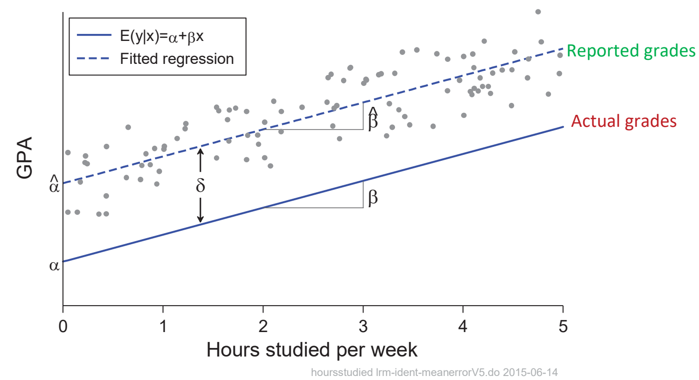
## Errors are uncorrelated

2. When would this assumption be violated? What are the consequences?
3. Imagine duplicating all observations and re-estimating. What changes?



## Conditional mean error and identification

1. We assume the average error is 0: $E(\varepsilon_i \mid \mathbf{x}_i) = 0$. How do you know?






## General principles of identification

1. Unidentified parameters cannot be estimated with more data.
2. Parameters are identified by:
   o Adding assumptions.
   o Using new kinds of data.
3. Identification is not all or nothing
   o Some parameters can be identified while others are not.
4. Combinations of unidentified parameters can be identified, while the individual parameters are not.
   o α+δ is identified, but α or δ are not individually identified.



# Interpretation with marginal effects

1. Marginal effects measure
   a. The change in the outcome
   b. for a change in one regressor
   c. holding other regressors constant.
2. Two types of marginal effects
   o Discrete change in E(y) as $x_k$ changes a fixed amount.
   o Marginal change in E(y) for an infinitely small change in a regressors.



## DC: Discrete change in E(y|x)

1. Start at $E(y \mid \mathbf{x}, x_3)$: expected value before change in $x_3$
1. Endg at $E(y \mid \mathbf{x}, x_3 + 1)$: expected value after change in $x_3$.
2. The discrete change for a change of 1 in $x_3$:

$$\frac{\Delta E(y \mid \mathbf{x}, x_3)}{\Delta x_3} = \text{End} - \text{Start}$$

$$= E(y \mid \mathbf{x}, x_3 + 1) - E(y \mid \mathbf{x}, x_3)$$

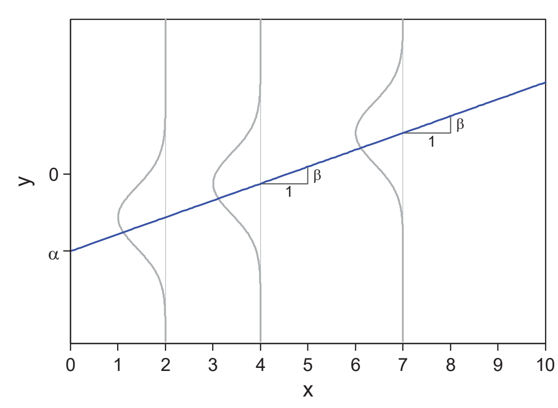
$$= \left[\beta_0 + \beta_1 x + \beta_2 x_2 + \underline{\beta_3 (x_3 + 1)}\right] - \left[\beta_0 + \beta_1 x + \beta_2 x_2 + \underline{\beta_3 x_3}\right]$$

$$= \beta_3$$

3. The amount of change does not depend on
   o The specific value of $x_3$
   o The specific values of the other x's that are held constant
4. Graphically,...



## Discrete change



## MC: Marginal change in E(y|x)

1. The instantaneous rate of change in E(y|x) as $x_k$ changes, holding other x's constant

$$\frac{\partial E(y\mid \mathbf{x})}{\partial x_k} = \frac{\partial\, \mathbf{x\beta}}{\partial x_k} = \beta_k$$

2. MC is the slope at a specific location
3. In the LRM, the MC does <u>not</u> depend on
   - o The value of $x_k$
   - o The values at which other $x$'s are held constant

## Marginal and discrete change in LRM

In linear models that do not have nonlinearities

$$\frac{\partial E(y\mid \mathbf{x})}{\partial x_k} = \frac{\Delta E(y\mid \mathbf{x})}{\Delta x_k} = \beta_k$$

.



## Simple interpretation due to linearity

### Continuous variables

For a unit increase in $x_k$ the expected change in y is $\beta_k$, holding other variables constant.

*For each additional year of education, income is expected to increase by $1,247, holding other variables constant.*

### Dummy variables coded as 0 and 1:

Having characteristic $x_k$ (as opposed to not having the characteristic) results in an expected change of $\beta_k$ in y, holding other variables constant.

*Being a female decreases the expected salary by $843, holding other variables constant.*



## Can you hold other variables constant?

1. Marginals assume one variable changes with other variables not changing
2. With linked variables this is mathematically impossible
   - o x and $x^2$ must change together
3. More generally
   - o Does it make <u>substantive</u> sense to change one regressor holding others constant?
   - o Can you increase education holding everything else constant?

## What does it mean when we say a variable is changing?

1. What does this <u>counterfactual</u> mean?
   - o Increase education by 4 years while holding income and occupation constant.
2. Does it make sense to imagine changing gender?



## Example: wages in Canada

Fox (2008) *Applied Regression Analysis and Generalized Linear Models* 2nd, p267. Survey of Labour & Income Dynamics, Ontario, Canada, 1994.

Model 1: $wages = \beta_0 + \beta_1 male + \beta_2 edyears + \beta_3 age + \varepsilon$

## Descriptive statistics - #0

```
. use slid-ontario01, clear
(Canada's 1994 Survey of Labor and Income Dynamics \ 2011-04-04)

. codebook, compact

Variable       Mean       Min       Max  Label
-----------------------------------------------------------------------
wages      15.54459       2.3     49.92  Hourly wages
male        .4978734        0         1  Is male?
age         36.95822       16        65  age in years
edyears     13.21191        0        20  years of education
-----------------------------------------------------------------------
N=3,997
```



## Fit M1 - #11

```
      Source |       SS           df       MS      Number of obs   =     3,997
-------------+----------------------------------   F(3, 3993)      =    590.67
       Model |  75828.1741         3  25276.058   Prob > F        =    0.0000
    Residual |  170869.757     3,993  42.7923258   R-squared       =    0.3074
-------------+----------------------------------   Adj R-squared   =    0.3069
       Total |  246697.931     3,996  61.736219   Root MSE        =    6.5416

------------------------------------------------------------------------------
       wages |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |    3.47367   .2070092    16.78   0.000     3.067817    3.879524
         age |   .2612932    .008664    30.16   0.000      .244307    .2782794
     edyears |   .9296491   .0342567    27.14   0.000     .8624868    .9968115
       _cons |  -8.124231   .5989773   -13.56   0.000    -9.298561   -6.949902
------------------------------------------------------------------------------
```

### Linear in wages

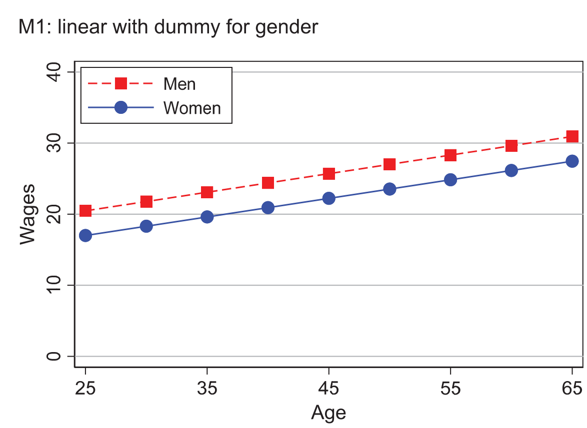
*For each additional year of age, <u>wages</u> are expected to increase by $0.26, holding other variables constant.*

*Being male increases wages by $3.47 at all ages and years of education.*

Graphically, on the next page...



## Plotting age and predicted wages - #13



# Standardized coefficients

1. Standardized coefficients remove the scale of variables.
2. In binary & ordinal models, standardization is required due to identification.

## Tool: Standardizing to 1

1. Standard deviation of $x_k$ : sd( $x_k$ ) = σ
2. Standard deviation of $\alpha x_k$ : sd( $\alpha x_k$ ) = α σ
3. Then: sd( $1/\sigma\, x_k$ ) = (1/σ) sd( $x_k$ ) = σ/σ = 1



## Standardizing coefficients by rescaling variables - #12

```
. egen Swages = std(wages)
. egen Sage = std(age)
. egen Sedyears = std(edyears)
. sum Swages wages Sage age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      Swages |     3,997    2.05e-09           1  -1.685654   4.374998
       wages |     3,997    15.54459     7.85724        2.3      49.92
        Sage |     3,997    8.64e-10           1  -1.745936   2.336036
         age |     3,997    36.95822      12.004         16         65

. * unstandardized variables
. regress wages male age edyears
::
. * y & x standardized
regress Swages male Sage Sedyears
::
. * x standardized
. regress wages male Sage Sedyears
::
. * y standardized
. regress Swages male age edyears
::
```

This is what **listcoef** does



## Standardized coefficients with listcoef - #12

```
. listcoef, help

regress (N=3997): Unstandardized and standardized estimates

  Observed SD:  7.8572
  SD of error:  6.5416

-------------------------------------------------------------------------------
             |         b         t     P>|t|     bStdX     bStdY    bStdXY      SDofX
-------------+-----------------------------------------------------------------
        male |    3.4737    16.780     0.000     1.737     0.442     0.221      0.500
         age |    0.2613    30.159     0.000     3.137     0.033     0.399     12.004
     edyears |    0.9296    27.138     0.000     2.823     0.118     0.359      3.037
    constant |   -8.1242   -13.564     0.000         .         .         .          .
-------------------------------------------------------------------------------
       b = raw coefficient
       t = t-score for test of b=0
   P>|t| = p-value for t-test
   bStdX = x-standardized coefficient
   bStdY = y-standardized coefficient
  bStdXY = fully standardized coefficient
   SDofX = standard deviation of X
```



Fully standardized coefficient

*For every standard deviation increase in age, wages are expected to increase by .399 standard deviations, holding other variables constant.*

```
             |         b         t     P>|t|     bStdX     bStdY    bStdXY      SDofX
-------------+---------------------------------------------------------------------
         age |    0.2613    30.159     0.000     3.137     0.033     0.399     12.004
```

x-standardized coefficient

*For every standard deviation increase in age, wages are expected to increase by $3.14, holding other variables constant.*

```
             |         b         t     P>|t|     bStdX     bStdY    bStdXY      SDofX
-------------+---------------------------------------------------------------------
         age |    0.2613    30.159     0.000     3.137     0.033     0.399     12.004
```

y-standardized coefficient

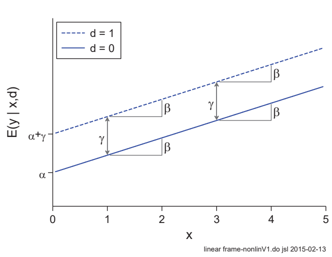
*Being a man increases the expected wages by .442 standard deviations, holding other variables constant.*

```
             |         b         t     P>|t|     bStdX     bStdY    bStdXY      SDofX
-------------+---------------------------------------------------------------------
        male |    3.4737    16.780     0.000     1.737     0.442     0.221      0.500
```

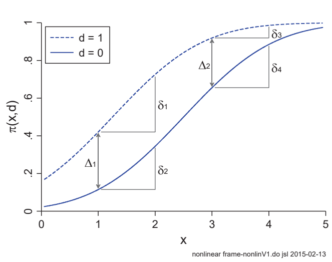


# Linear and nonlinear models

## A: Linear model




## B: Nonlinear model






## Nonlinear compared to linear models

### Marginal effect of $x_k$ in linear models

1. The size of the effect does not depend on the value of $x_k$
2. The size of the effect does not depend on the values of other $x$'s
3. Marginal change and discrete change are equal

$$\frac{\partial E(\cdot)}{\partial x_k} = \frac{\Delta E(\cdot)}{\Delta x_k}$$

### Marginal effect of $x_k$ in nonlinear models

1. The size of the effect does depend on the value of $x_k$
2. The size of the effect does depend on the values of the other $x$'s
3. Marginal and discrete change are usually unequal

$$\frac{\partial E(\cdot)}{\partial x_k} \neq \frac{\Delta E(\cdot)}{\Delta x_k}$$

## Nonlinear linear regression models

1. In a linear model, the $x$'s enter in the linear form $\mathbf{x\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$
2. The effects of regressors can be nonlinear by including transformations.

Quadratic:
$$y = \beta_0 + \beta_1 w_1 + \beta_2 w_1^2 + \varepsilon$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Loglinear:
$$y = \ln z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

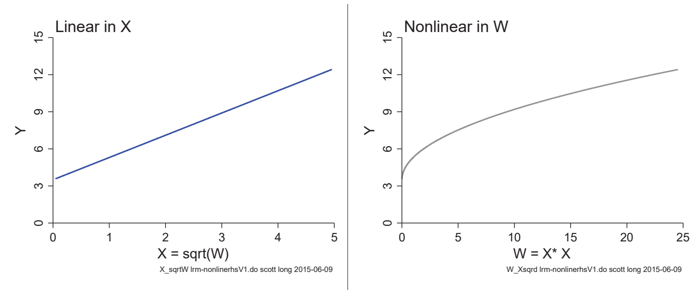
Square root:
$$y = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{w_2} + \varepsilon$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Graphically...



## Linear in sqrt(W); nonlinear in W

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \sqrt{w_2} + \varepsilon$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$






## Example: wages in Canada – continued

Model 1: $wages = \beta_0 + \beta_1 male + \beta_2 edyears + \beta_3 age + \varepsilon$

Model 2: $wages = \beta_0 + \beta_1 male + \beta_2 edyears + \beta_3 age + \beta_4 age^2 + \varepsilon$

Model 3: $wages = \beta_0^W + \beta_2^W edyears + \beta_3^W age + \beta_4^W age^2 + \varepsilon$

$wages = \beta_0^M + \beta_2^M edyears + \beta_3^M age + \beta_4^M age^2 + \varepsilon$

### Descriptive statistics - #0

```
Variable        Mean        Min        Max  Label
----------------------------------------------------------------
wages       15.54459        2.3      49.92  Hourly wages
male        .4978734          0          1  Is male?
age         36.95822         16         65  age in years
edyears     13.21191          0         20  years of education
----------------------------------------------------------------
N=3,997
```

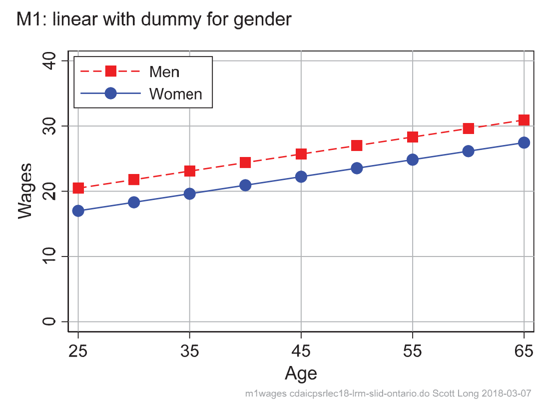
### M1: baseline regression - #11

$$wages = \beta_0 + \beta_1 male + \beta_2 edyears + \beta_3 age + \varepsilon$$

Plotting the effect of age, gender and wages...



## Plotting age and predicted wages - #13



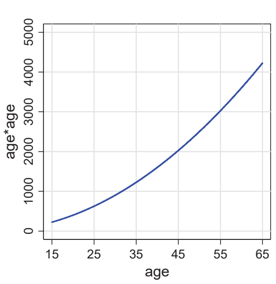



## M2: adding a squared term

1. In M1, the effect of age is: (a) always positive; or (b) always negative; or (c) always 0.
2. To allow the effect to be positive and negative, we add age-squared:

$$wages = \beta_0 + \beta_1 male + \beta_2 edyears + \beta_3 age + \beta_4 age^2 + \varepsilon$$

As age increases, age-squared increases faster




- The greater the age, the greater the impact of $\beta_{\text{age-squared}}$.
- If $\beta_{\text{age}}$ and $\beta_{\text{age-sq}}$ have different signs, the effect of age can change directions as the size of $age^2$ overwhelms the size of $age$.



## Specifying M2 with age-squared

1. I can create a squared variable with generate:

```
gen agesq = age*age
```

2. Factor syntax to implicitly create age-squared from **age**:

```
c.age#c.age
```

where **c.** indicates continuous; # indicates multiply

3. For example,

```
. sum agesq c.age##c.age
    Variable |       Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |   36.95822      12.004         16         65
       agesq |    1509.97     934.969        256       4225
 c.age#c.age |    1509.97     934.969        256       4225
```

4. Factor variables:
   - Created dynamically as needed
   - Disappear when not needed
   - Keep track of how variables are related
   - Extremely useful

## LRM that is quadratic in age - #22

```
. regress wages male c.age##c.age edyears
::
       wages |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   3.465888   .2017898    17.18   0.000     3.070267    3.861508
         age |   1.001166   .0517182    19.36   0.000     .8997691    1.102562
 c.age#c.age |  -.0096636   .0006664   -14.50   0.000    -.0109702    -.008357
     edyears |   .8312951   .0340748    24.40   0.000     .7644895    .8981007
       _cons |  -19.57354   .9820115   -19.93   0.000    -21.49883   -17.64825
------------------------------------------------------------------------------
```

The effect of being male

*Men are expected to earn $3.46 more than women with comparable characteristics.*

The effect of age

1. We can't interpret the coefficients for age and age-squared are:

   $\beta_{age}$ = 1.001166 and $\beta_{agesq}$ = -.0096636

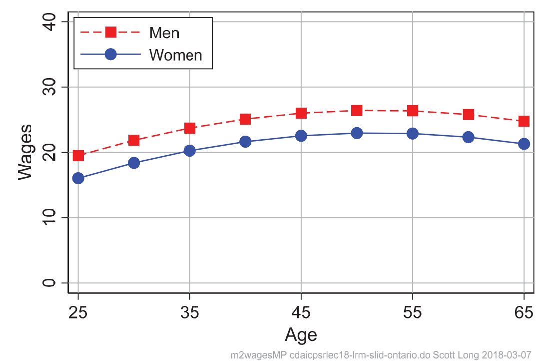
   since you can't increase age and hold age-squared constant; and vice versa.
2. Instead, we look at predictions or marginal effects of age



## Plotting age and wages - #23

The effect of age depends on your age.

M2: age-squared with dummy for gender






# Post-estimation predictions in Stata

## Native Stata commands

1. Predictions can be things like:
   - Expected values of the outcome
   - Marginal effects on the outcome
1. **`predict`** makes predictions at observed values of the regressors
2. **`margins`** makes predictions at observed or users specified values
   - Predictions can be averaged
3. **`marginsplot`** plots predictions



## SPost13 commands

1. These commands use **`margins`** for predictions
   - **`mtable`**: Create tables of predictions
   - **`mgen`**: Generate variables with predictions for plotting
   - **`mchange`**: Marginal effects
   - **`mlincom`**: Linear combinations of predictions.
2. These commands
   - Automatically construct multiple **`margins`** commands
   - Have compact output that combine results from multiple commands

## Stata or SPost?

1. Stata commands are more general and work with all models, but the output is more difficult.
2. SPost works for *most* cross-sectional models and is easier for many things.
3. To use **`marginsplot`**, you must use **`margins`**.



## atspec: specifying values of regressors in margins and m*

atmeans: all regressors at their means.

```
margins, atmeans
```

at() for single values of regressors

```
margins, at(age=25 male=1 edyears=20) atmeans
```

Variables not specified are held at their mean.

at() with linked variables

```
margins, at(age=25) atmeans
```

If **`c.age#c.age`** is a regressor, predictions are made at 25*25 for age-squared.

at() for multiple values using a *numlist*

```
margins, at(age=(25(5)75) male=1 edyears=20) atmeans
```

Predictions are computed for **`age`** = 25, 30, 35, etc.

at() at multiple specified values

```
margins, at(age=25 male=1 edyears=20) ///
         at(age=60 male=0 edyears=12) atmeans
```



## M2 continued: Plotting predicted wages

### Predictions with mtable - #22

```
. mtable, atmeans at(age=(25(5)65) male=(0 1) edyears=20)

Expression: Linear prediction, predict()

          |     male       age        xb
----------+-----------------------------
        1 |        0        25    17.414
        2 |        0        30    19.292
        3 |        0        35    20.736
        4 |        0        40    21.749
        5 |        0        45    22.328
::
       13 |        1        40    25.034
       14 |        1        45    26.242
       15 |        1        50    26.898
       16 |        1        55    27.002
       17 |        1        60    26.554
       18 |        1        65    25.554

Specified values of covariates

          |  edyears
----------+---------
  Current |       20
```

## Make predictions with margins - #23

```
. margins, atmeans at(age=(25(5)65) male=(0 1) edyears=20)

Adjusted predictions                            Number of obs     =      3,997
Model VCE    : OLS

Expression   : Linear prediction, predict()

1._at        : male             =           0
               age              =          25
               edyears          =          20
2._at        : male             =           0
               age              =          30
               edyears          =          20
::
::
9._at        : male             =           0
               age              =          65
               edyears          =          20
10._at       : male             =           1
               age              =          25
               edyears          =          20
::
::
```



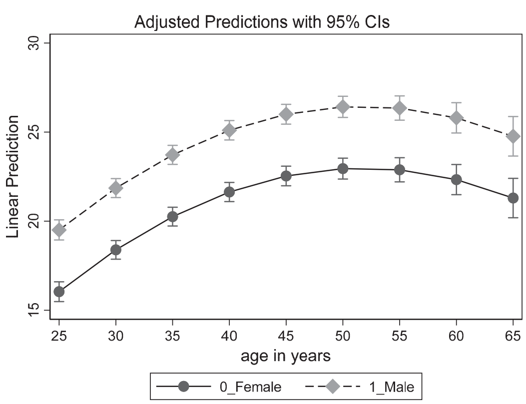
```
18._at       : male             =           1
               age              =          65
               edyears          =          20
```

| _at | Margin | Delta-method Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 1 | 16.04176 | .2836861 | 56.55 | 0.000 | 15.48557 | 16.59794 |
| 2 | 18.3901 | .2671511 | 68.84 | 0.000 | 17.86633 | 18.91386 |
| 3 | 20.25526 | .2685062 | 75.44 | 0.000 | 19.72883 | 20.78168 |
| 4 | 21.63724 | .2737742 | 79.03 | 0.000 | 21.10049 | 22.17399 |
| 5 | 22.53603 | .2808891 | 80.23 | 0.000 | 21.98534 | 23.08673 |
| 6 | 22.95165 | .2992583 | 76.70 | 0.000 | 22.36494 | 23.53837 |
| 7 | 22.88409 | .3452619 | 66.28 | 0.000 | 22.20719 | 23.561 |
| 8 | 22.33335 | .4321969 | 51.67 | 0.000 | 21.48601 | 23.1807 |
| 9 | 21.29943 | .5638506 | 37.77 | 0.000 | 20.19397 | 22.4049 |
| 10 | 19.50765 | .2883744 | 67.65 | 0.000 | 18.94227 | 20.07302 |
| 11 | 21.85598 | .2716595 | 80.45 | 0.000 | 21.32338 | 22.38859 |
| 12 | 23.72114 | .2725945 | 87.02 | 0.000 | 23.18671 | 24.25558 |
| 13 | 25.10312 | .2774585 | 90.48 | 0.000 | 24.55915 | 25.6471 |
| 14 | 26.00192 | .2842255 | 91.48 | 0.000 | 25.44468 | 26.55916 |
| 15 | 26.41754 | .3022107 | 87.41 | 0.000 | 25.82504 | 27.01004 |
| 16 | 26.34998 | .3477177 | 75.78 | 0.000 | 25.66826 | 27.0317 |
| 17 | 25.79924 | .4341173 | 59.43 | 0.000 | 24.94813 | 26.65035 |
| 18 | 24.76532 | .5653219 | 43.81 | 0.000 | 23.65697 | 25.87367 |



## Plotting with marginsplot: quick plots after margins - #24

```
. marginsplot
```



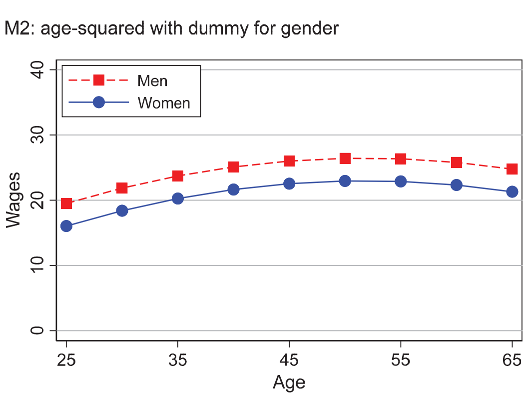



## Code: Adding options to marginsplot

```
marginsplot, noci /// #1
    ylab(0(10)40, labsize(*1.1) glwid(*.7) glcol(black*.3) grid gmin gmax) /// #2
    xlab(25(10)65, labsize(*1.1) glwid(*.7) glcol(black*.3) nogrid) /// #3
    legend(order(2 "Men" 1 "Women") ring(0) pos(11) rows(2)) /// #4
    plot1opts(lcol(blue*1.) lpat(solid) msym(O) msiz(*1.) mcol(blue*1.)) /// #5
    plot2opts(lcol(red*1.) lpat(dash) msym(S) msiz(*.9) mcol(red*1.)) /// #6
    plotopts(lwid(*1)) xtitle("Age") ytitle("Wages") /// #7
    title("M1: linear with dummy for gender" " ",ring(2) pos(11) size(*1)) /// #8
    caption("`graphname' `tag'", size(vsmall) pos(5) col(gs10)) /// #9
    scale(1.1) // #10
```



## M2: Plotting predicted wages






## M3: Interactions with gender

1. Let the coefficients differ by gender:

$$wages = \beta_0^W + \beta_2^W edyears + \beta_3^W age + \beta_4^W age^2 + \varepsilon$$

$$wages = \beta_0^M + \beta_2^M edyears + \beta_3^M age + \beta_4^M age^2 + \varepsilon$$

2. Fit separate models:

```
regress wages male c.age c.age#c.age edyears if female
regress wages male c.age c.age#c.age edyears if male
```

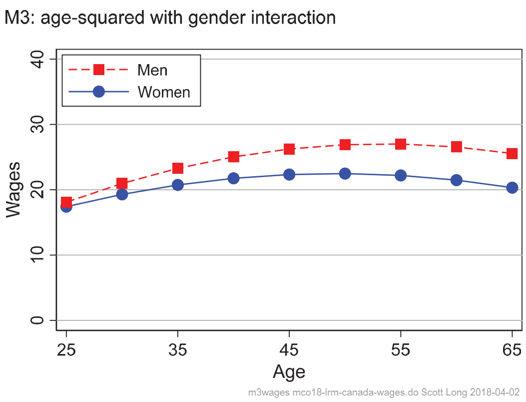
3. Or fit single model with interactions:

```
regress wages ibn.male ibn.male#(c.edyears c.age##c.age), nocon
```

   - **ibn** means no base value
   - For now, don't worry about the details

4. The predictions are shown in this graph...

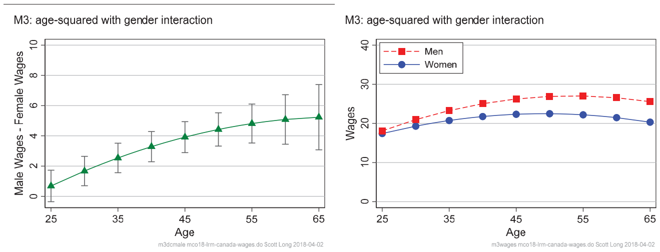
## Model 3 with Interactions






## Are wages of men greater than those of women?

Gender differences are significant when the CI crosses 0.






# Summary of nonlinear linear models

1. Nonlinearity has many forms.
2. With some forms, coefficients are easy to interpret (e.g., loglinear).
3. With other forms, coefficients have no direct interpretation.
4. Predictions can be used to interpret nonlinear models of any form.



# Estimation and testing

Details in *Estimating, Testing and Fit* lecture

## Estimation by OLS

1. OLS minimizes the sum of the squared residuals:

$$SSR = \sum_{i=1}^{N}\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right)^2 = \sum_{i=1}^{N}\left(\hat{\varepsilon}_i\right)^2$$

2. OLS has a simple "closed-form" formula

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

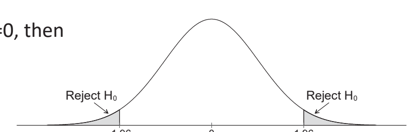


## Overview of hypothesis testing

| | Decision | |
|---|---|---|
| $H_0$: β=0 | **Accept $H_0$** | **Reject $H_0$** |
| **In fact β=0** | No error | Type I: Pr(reject true)=α<br>Area in the shaded tail.<br>Size of the test. |
| **In fact β≠0** | Type II: accept false<br>Power of test. | No error |

4. If the errors are normal and $\beta_k$=0, then

$$t_k = \frac{\hat{\beta}_k - 0}{\sqrt{Var\left(\hat{\beta}_k\right)}} \sim t_{N-K-1}$$






## Example of t-tests in regression - #11

```
. regress wages male age edyears

      Source |       SS           df       MS      Number of obs   =     3,997
-------------+----------------------------------   F(3, 3993)      =    590.67
       Model |  75828.1741         3  25276.058   Prob > F        =    0.0000
    Residual |  170869.757     3,993  42.7923258   R-squared       =    0.3074
-------------+----------------------------------   Adj R-squared   =    0.3069
       Total |  246697.931     3,996  61.736219   Root MSE        =    6.5416

------------------------------------------------------------------------------
       wages |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |    3.47367   .2070092    16.78   0.000     3.067817    3.879524
         age |   .2612932    .008664    30.16   0.000      .244307    .2782794
     edyears |   .9296491   .0342567    27.14   0.000     .8624868    .9968115
       _cons |  -8.124231   .5989773   -13.56   0.000    -9.298561   -6.949902
------------------------------------------------------------------------------
```

*Men have significantly higher wages than women (t=16.78, p<0.01 for a two-tailed test).*

*Each additional year of age increases expected wages by nearly a dollar, holding other variables constant. (p<.01 for a 2-tailed test).*

## Overview of continuous LHS

1. LRM is the foundation for CDA models
   - o Be careful about generalizing from LRM to other models!
2. Variables enter the model as **xβ**, called the index function.
   - o **xβ** allows flexible specifications through interactions and transformations.
   - o Complictions on the RHS make the LRM nonlinear
3. Nonlinearity makes interpretation more complicated
   - o Regression parameters no longer provide direct insights into effects.
   - o They are most useful for making predictions



# β1 Binary outcomes

## Readings and examples

*Long & Freese: Chapters 5 and 6*

- o See references in these chapter

*mdo18-brm-*.do*

## Objectives

1. Derive the binary regression model (BRM)
2. Explain interpretation using predictions.
   - o Interpreting predictions not parameters in nonlinear models
3. Applications of predictions and marginal effects



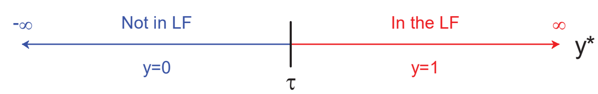
## Deriving the BRM

1. Binary logit and probit can be derived four ways.
   - o A nonlinear probability model
   - o A random utility model for chosing the optimal outcome
   - o Generalized linear model linking predictors and outcome
   - o Regresson on latent variable (LV) the generates observed outcomes
2. I focus on the LV approach
   - o It builds on LRM
   - o It highlights the scalar identication of parameters
   - o It generalizes easily to other models



## BRM as a latent variable model

1. The <u>unobserved propensity</u> $y^*$ generates the observed y:



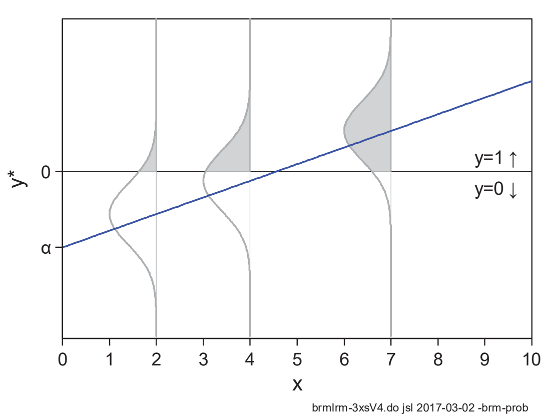

   where not all women in LF have the same propensity to work
2. A structural model regresses y* on the x's

   $y_i^* = \alpha + \beta x_i + \varepsilon_i$ or $y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$
3. The probability of observed y depends on $y^*$:

   $\Pr(y=1\,|\,\mathbf{x}) = \Pr(y^* > \tau\,|\,\mathbf{x})$
4. Graphically,....



## The structural model y* = α + βx + ε with Pr(y=1|x) shaded

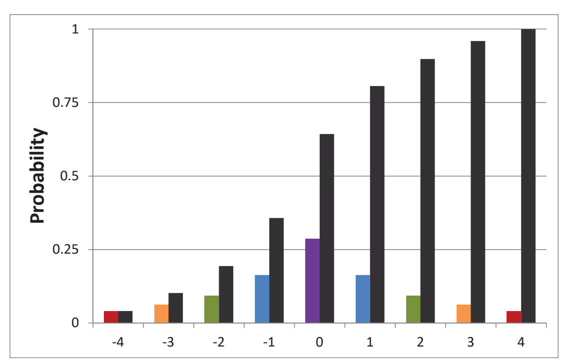





## Tool: PDF and CDF of probability distribution

1. y: -4 -3 -2 -1 0 1 2 3 4
2. PDF: Pr(y=-4), Pr(y=-3), Pr(y=-2), Pr(y=-1), Pr(y=0), Pr(y=1), Pr(y=2), Pr(y=3)
3. CDF: Pr(y≤-4), Pr(y≤-3), Pr(y≤-2), Pr(y≤-1), Pr(y≤0), Pr(y≤1), Pr(y≤2), Pr(y≤3)

Pr(*y*≤0) =
Pr(*y*=-4)
\+ Pr(*y*=-3)
\+ Pr(*y*=-2)
\+ Pr(*y*=-1)
\+ Pr(*y*=0)



## Errors in the latent variable model

The error is assumed to be normal or logistic.

### Normal errors

1. **Normal PDF**: standard deviation σ

$$\varphi(\varepsilon_P;\mu=0,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-\varepsilon_P^2}{2\sigma^2}\right)$$

2. **Standardized normal PDF**: standard deviation σ=1 simplifies distribution

$$\varphi^S(\varepsilon_P)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-\varepsilon_P^2}{2}\right)$$

3. **Standardized normal CDF**

$$\Phi^S\left(\varepsilon_P\right)=\int_{-\infty}^{\varepsilon}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)dt$$



## Logistic errors

1. **Standardized logistic PDF**: σ=1 makes distribution more complex

$$\lambda^S\left(\varepsilon_L\right)=\frac{\frac{\pi}{\sqrt{3}}\exp\left(\frac{\pi}{\sqrt{3}}\varepsilon_L\right)}{\left[1+\exp\left(\frac{\pi}{\sqrt{3}}\varepsilon_L\right)\right]^2}$$

2. **Standard logistic PDF**: σ=π/√3=1.81... is simpler.

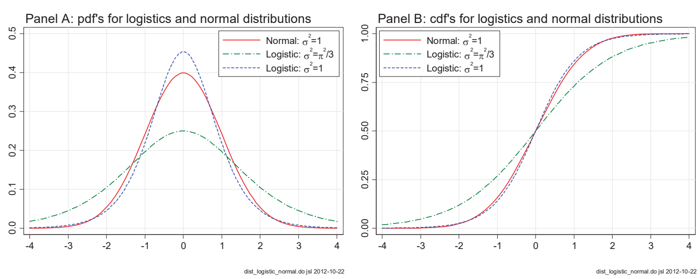
$$\lambda\left(\varepsilon_L\right)=\frac{\exp\left(\varepsilon_L\right)}{\left[1+\exp\left(\varepsilon_L\right)\right]^2}$$

3. **Standard logistic CDF**: σ=π/√3=1.81...

$$\Lambda(\varepsilon_L)=\frac{\exp\left(\varepsilon_L\right)}{1+\exp\left(\varepsilon_L\right)}$$

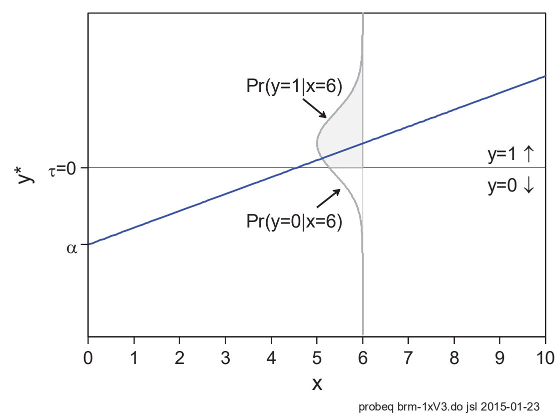


## PDF and CDF for normal and logit curves






## Computing Pr(y=1|x) from y*






## This is a CDF of the error distribution

See Long(1997) or Long and Freese (2014) for details.

1. For <u>probit</u> with standardized normal errors

$$\Pr\left(y=1\mid\mathbf{x}\right)=\Phi\left(\mathbf{x\beta}\right)=\int_{-\infty}^{\mathbf{x\beta}}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)dt$$

2. For <u>logit</u> with standard logistic errors

$$\Pr\left(y=1\mid\mathbf{x}\right)=\Lambda\left(\mathbf{x\beta}\right)=\frac{\exp\left(\mathbf{x\beta}\right)}{1+\exp\left(\mathbf{x\beta}\right)}$$

3. Using π() as shorthand for Pr(y=1|)

$$\pi\left(\mathbf{x\beta}\right)=\Pr\left(y=1\mid\mathbf{x}\right)=F\left(\mathbf{x\beta}\right)$$



## y* and Pr(y=1|x) for a single regressor

1. The structural equation is:

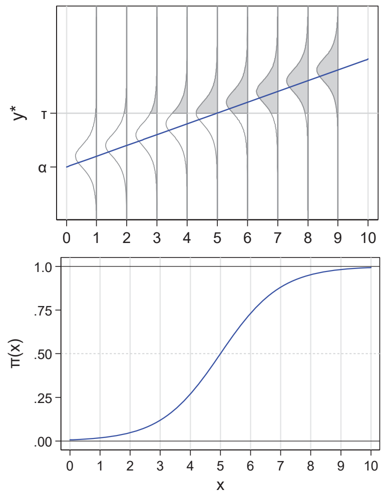
$$y^*=\alpha+\beta x+\varepsilon \quad \text{where } \varepsilon\sim N\left(0,1\right)$$

2. The probability equation is:

$$\Pr(y=1\mid x)=F(\alpha+\beta x)$$

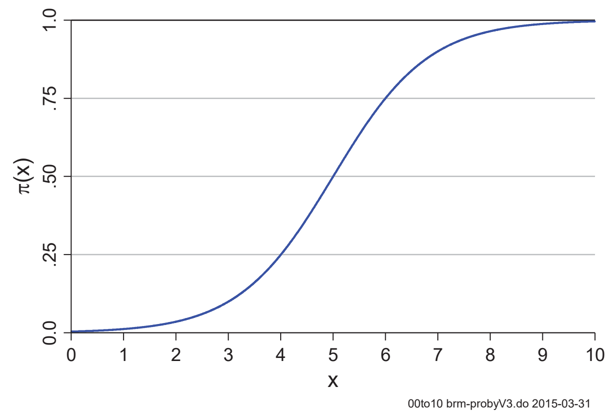
3. The link between $y^*$ and Pr($y$=1) leads to an S-shaped curve for Pr($y$=1|$x$)

Next page...

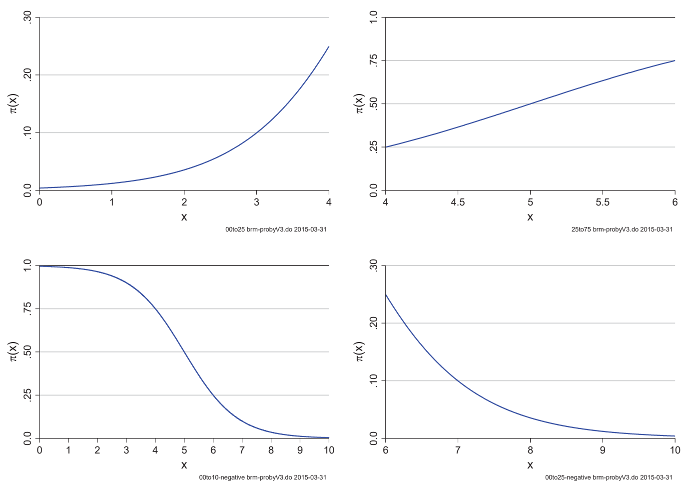





## Does the <u>empirical</u> relationship need to be S-shaped?



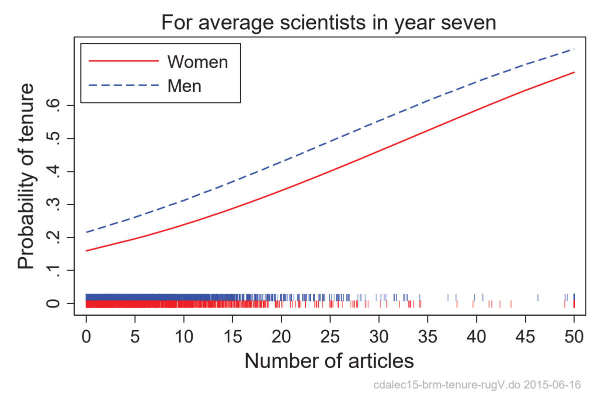





## On the support of the data

Where is your data? Where do you want to explore






## Scalar identification of β

1. The <u>true</u> structural model regresses y* on x:

   $y^* = \alpha + \beta x + \varepsilon$

2. Since y* and ε are unobserved, we cannot estimate their means or variances.
3. Suppose someone doubled the *unobserved* y*?

   $2y^* = 2\alpha + 2\beta x + 2\varepsilon$

4. Changing notation,

   $\underline{y}^* = \underline{\alpha} + \underline{\beta}x + \underline{\varepsilon}$

5. The true β and the imposter $\underline{\beta}$ are empirically indistinguishable
   - We can't interpret the estimated βs since we don't know the metric of y*
6. Stretching a graph illustrates this fundamental point:
   - See mco18-scalar identifcation demonstration 2018-04-03.docx



# Scalar identification in the BRM

1. Identification are critical for understanding the BRM
2. The regression coefficients are not identified; the probabilities are

## Arbitrary but necessary identifying assumptions

### Assumption 1: Mean of the errors (as with LRM)

$$E(\varepsilon \mid \mathbf{x}) = 0$$

### Assumption 2: Value of threshold

$$\tau = 0$$

### Assumption 3: Variance of the errors

$Var(\varepsilon \mid \mathbf{x}) = 1$ for probit

$Var(\varepsilon \mid \mathbf{x}) = \pi^2 / 3$ for logit

## Algebraic illustration of identification assumption 3

1. Consider the structural model for probit:

$$y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon \qquad \text{where } \mathrm{Var}(\varepsilon \mid \mathbf{x}) = 1$$

2. Multiply both sides by δ:

$$\delta y^* = \mathbf{x}(\delta\boldsymbol{\beta}) + \delta\varepsilon$$

3. We can't measure y* or ϵ and do not know **β**, so the change is unobservable.
4. For convenience, define:

$$y_L^* \equiv \delta y^* \qquad \boldsymbol{\beta}_L \equiv \delta\boldsymbol{\beta} \qquad \varepsilon_L \equiv \delta\varepsilon$$

5. Then:

$$y_L^* = \mathbf{x}\boldsymbol{\beta}_L + \varepsilon_L$$

6. And:

$$\mathrm{Var}(\varepsilon_L \mid \mathbf{x}) = \mathrm{Var}(\delta\varepsilon \mid \mathbf{x}) = \delta^2 \mathrm{Var}(\varepsilon \mid \mathbf{x}) = \delta^2$$

7. If $\delta \equiv \pi/\sqrt{3}$, then $\mathrm{Var}(\varepsilon_L \mid \mathbf{x}) = \pi^2/3$ as



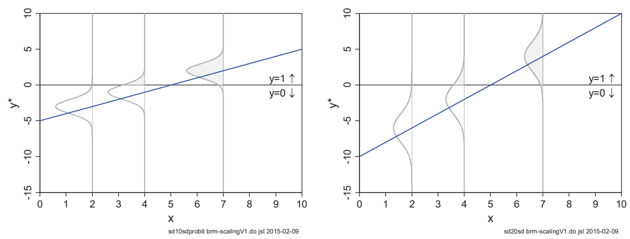
## Graphical illustration of identification assumption 3

The β's cannot be interpreted directly since their magnitude reflects:

a. The relationship between the *x*'s and *y**.

b. Arbitrary identifying assumptions.

2. Pr(*y*=1|**x**) is unaffected by the identifying assumption about Var(ε |**x**).




3. See mco18-scalar identification demonstration 2018-04-03.docx



## Comparing logit and probit with Mroz data - #2

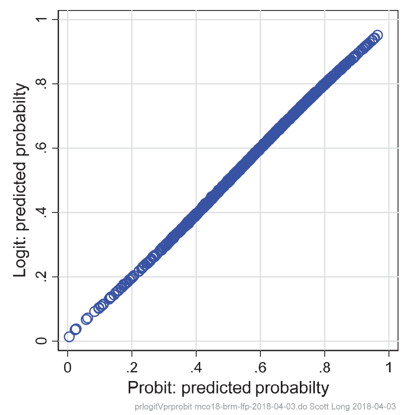
### Comparing regression coefficients and z-tests

```
logit  lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog
probit lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog
```

| | blm b | blm z | bpm b | bpm z | ratio b | ratio z |
|---|---|---|---|---|---|---|
| lfp | | | | | | |
| k5 | -1.392 | -7.182 | -0.840 | -7.480 | 1.657 | 0.960 |
| k618 | -0.066 | -0.916 | -0.041 | -0.975 | 1.593 | 0.939 |
| 1.wc | 0.798 | 3.367 | 0.482 | 3.481 | 1.655 | 0.967 |
| 1.hc | 0.136 | 0.659 | 0.074 | 0.596 | 1.841 | 1.106 |
| lwg | 0.610 | 3.677 | 0.371 | 3.894 | 1.644 | 0.944 |
| inc | -0.035 | -3.989 | -0.021 | -4.136 | 1.665 | 0.965 |



## Comparing predicted probabilities: r=.9998




```
. estimates restore blm
. predict prblm
. label var prblm ///
    "Logit: Pr(LFP|X)"
. estimates restore bpm
. predict prbpm
. label var prbpm ///
  "Probit: Pr(LFP|X)"
```



## Review of scalar identification in logit and probit

1. The magnitude of regression coefficients depends on the scale of the outcome
2. Since *y** is latent, we do not know its scale or variance
3. Therefore, the slopes are not identified
4. Estimated β's <u>cannot be directly interpreted</u> since they reflect
   - o The relationship between the *x*'s and *y**
   - o Arbitrary identifying assumption for Var(ε|**x**)
5. Scalar identification does not affect Pr(y=1|x)
   - o Probabilities can be interpreted without concern about identification
6. Scalar identification issue has profound implications for:
   - o Group comparisons
   - o Nested models
   - o Mediation effects



# Alternative derivations of the BRM

## Nonlinear probability model (see Theil)

1. Transform Pr(y=1|x) to the odds which range from 0 to ∞

$$\text{Odds}(1 \text{ versus } 0 \mid \mathbf{x}) = \Omega(\mathbf{x}) = \frac{\Pr(y=1 \mid \mathbf{x})}{\Pr(y=0 \mid \mathbf{x})}$$

2. Transform the odds to the logit or log odds which ranges from -∞ to ∞

$$\ln\left[\frac{\Pr(y=1 \mid \mathbf{x})}{\Pr(y=0 \mid \mathbf{x})}\right] = \mathbf{x}\boldsymbol{\beta}$$

3. Take the exponential of each side and solve for Pr(*y*=1|**x**)

$$\Pr(y=1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1+\exp(\mathbf{x}\boldsymbol{\beta})}$$

4. Or in terms of odds:

$$\Omega(\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots)$$

## Random Utility Model (RUM)

1. Two choices where

   Choice 0 provides utility $u_{0i}$

   Choice 1 provides utility $u_{1i}$

2. The utility received from a choice is modeled as

   $u_{0i} = \mathbf{x}_i\boldsymbol{\beta}_0 + \varepsilon_{0i}$

   $u_{1i} = \mathbf{x}_i\boldsymbol{\beta}_1 + \varepsilon_{1i}$

3. I chooses 0 if $u_{0i} > u_{1i}$ with $\Pr(u_{0i} > u_{1i} | \mathbf{x}) = \Pr(0 | \mathbf{x})$
4. If ε is normal, this is probit; if ε is extreme value type 2, logit



## Generalized linear model (GLM)

1. The observed y has a binomial distribution with mean

   $E(y) = \mu$

2. The linear predictor is

   $\eta = \mathbf{x}\boldsymbol{\beta}$

3. The link function:

   logit: $\ln[\mu/(1-\mu)] = \eta = \mathbf{x}\boldsymbol{\beta}$

   probit: $\Phi^{-1}(\mu) = \eta = \mathbf{x}\boldsymbol{\beta}$



## ML estimation

1. Since we can't estimate residuals, we can't use methods like OLS.
2. Maximum likelihood estimation choses the values of the parameters that makes the observed data more likely than any other values of the parameters
   - Pick paramters that make what you see most likely
3. Probability of what was observed for each observation

$$p_i = \begin{cases} \Pr(y_i = 1 | \mathbf{x}_i) & \text{if } y_i = 1 \text{ is observed} \\ 1 - \Pr(y_i = 1 | \mathbf{x}_i) & \text{if } y_i = 0 \text{ is observed} \end{cases}$$

4. If observations are independent, Pr(HH) = Pr(H)*Pr(H). Thus,

$$L(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} p_i$$

5. The estimates $\hat{\boldsymbol{\beta}}$ maximize $L(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X})$



### Comments on MLE

1. See lecture *Estimation, Testing, and Fit* for more information
1. ML estimates are asymptotically consistent, normal, and efficient
   - ML estimate are not necessarily bad in small samples, but small sample behavior is largely unknown
2. Numerical methods search for the maximum using the slope and change in slope of the likelihood equation
   - Numerical methods for ML estimation work very well "when your model is appropriate for your data" (Joreskog)
3. Cramer (1986:10) gives excellent advice

   Check the data, check their transfer into the computer, check the actual computations (preferably by repeating at least a sample by a rival program), and always remain suspicious of the results, regardless of the appeal.

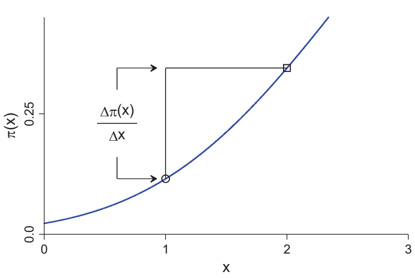


## Parameters, probability curves, and marginal effects

1. Consider the BRM:

$$\pi(x) = \Pr(y = 1 | x) = F(\alpha + \beta x)$$

2. Discrete change DC(x) is the change in Pr as x changes from 1 to 2:

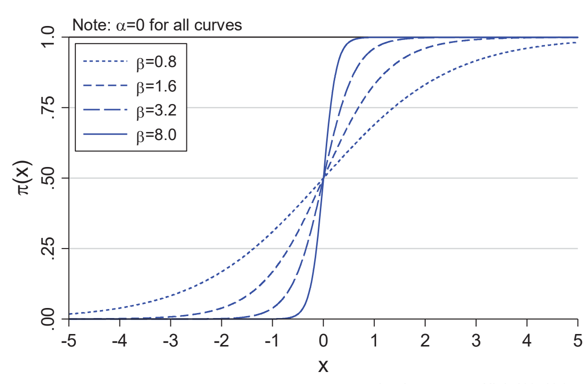



3. The size of DC9x) depends on α and β.

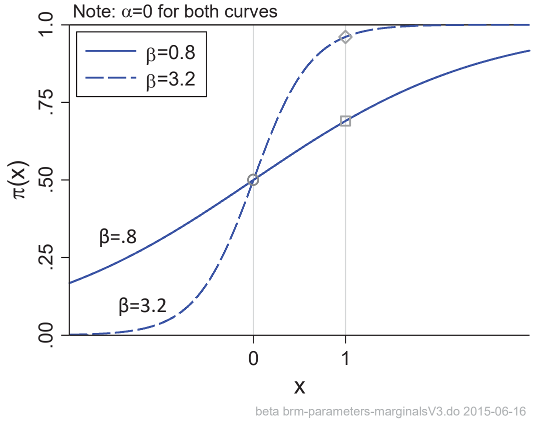


## Changing the slope β

The larger the slope, the smaller the Δx for a given ΔPr(y).



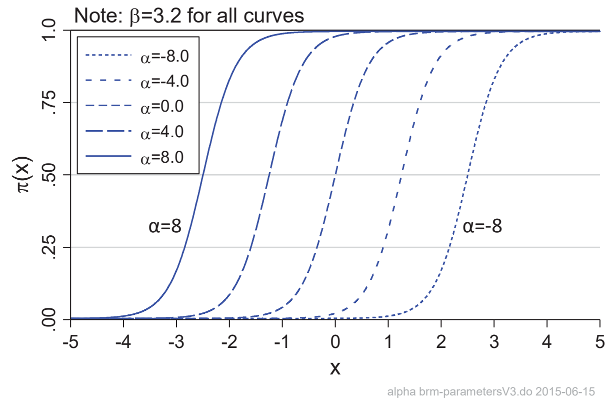
## Changing β changes the size of DC(x)





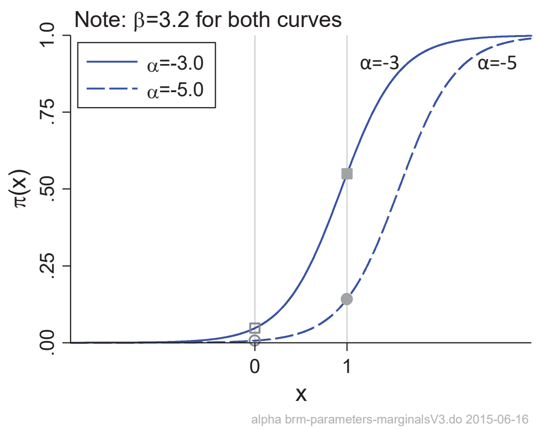

## Changing the intercept α

Smaller α shifts the curve right.






## Changing α changes the size of DC(x)






## How does the value of $x_2$ change $DC(x_1)$?

1. The model:

$$\Pr(y=1 \mid x_1, x_2) = F(-4 + .6x_1 + .5x_2)$$

2. If $x_2 = 0$,

$$\Pr(y=1 \mid x_1, x_2 = 0) = F(-4 + .6x_1 + [.5 \times 0])$$
$$= F(-4 + .6x_1)$$

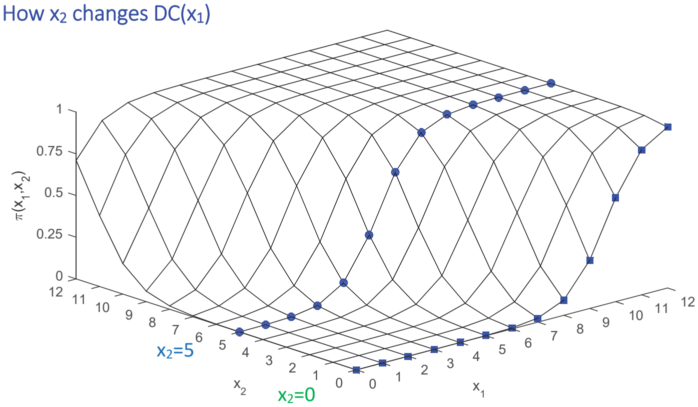
3. If $x_2 = 5$ (curve with circles on next page):

$$\Pr(y=1 \mid x_1, x_2 = 5) = F(-4 + .6x_1 + [.5 \times 5]) = F([-4 + 2.5] + .6x_1)$$
$$= F(-1.5 + .6x_1)$$

4. $DC(x_1)$ depends on values of other variable which shift the probability curve.
5. Graphically...



## How $x_2$ changes $DC(x_1)$






## Interpretation using predictions

1. Probabilities are the fundamental statistic for interpretation

$$\widehat{\pi}(\mathbf{x}) = \widehat{\Pr}(y=1 \mid \mathbf{x}) = F(\mathbf{x}\hat{\boldsymbol{\beta}})$$

2. Since model is nonlinear,

   *No single method of interpretation fully describes the relationship between a variable and the outcome.*

3. The critical decision is deciding at which values of **x** to examine the predictions.
   - This is substantive decision
4. Search for an elegant method that reflects substantive complexities.
   - Try many to find the right one

## Value of regressors for computing Pr(y=1|x)

1. In-sample predictions use observed values from the sample
2. Out of sample predictions use any values of the x's

### Key concepts

**On the support** are values where real data might be found

**Counterfactual experiments** imagine a variable changes holding others constant

**Average** could be a counterfactual

- o Who is .53 female?



## Ways to use predictions for interpretation

1. Predictions at observed values
2. Marginal effects
   - o Changes in predictions
3. Ideal types or profiles
   - o Predictions at values of substantive interest
4. Tables
   - o Predictions at multiple levels of several regressors
5. Graphs
   - o Predictions at many levels of regressors
6. Odds ratios
   - o A *ratios of ratios* of probabilities

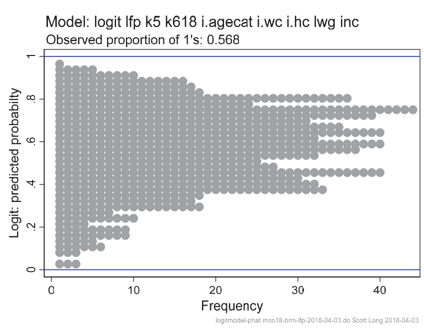


## In-sample predictions

1. In-sample predictions use observed $\mathbf{x}_i$'s

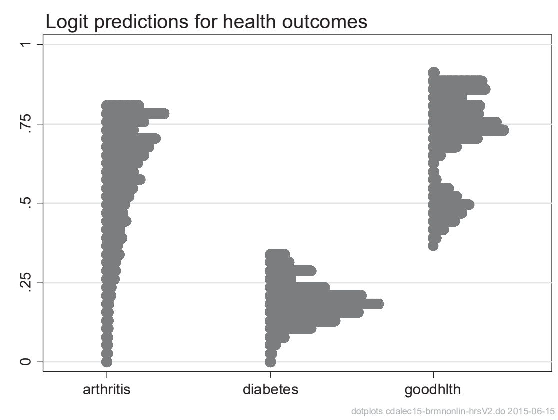
$$\hat{\pi}(\mathbf{x}_i) = \widehat{\Pr}(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i\hat{\boldsymbol{\beta}})$$

2. Examining these predictions for patterns and suspicious observations






## Predictions for health outcomes (details later)






### Code: in sample predictions and plots

*Make predictions*

```
estimates restore logitmodel
predict prlogit
label var prlogit "Logit: predicted probability"
```

*Compute mean prediction to add to graph*

```
qui sum prlogit // compute mean to include in graph
local mn = string(r(mean),"%5.3f") // store formatted string
```

*Dotplot/histogram*

```
dotplot prlogit, ///
    ylab(0(.2)1, nogrid) ylin(0 1, lcol(blue)) mcol(gs10) ///
    title(Model: logit lfp k5 … inc, pos(11)) ///
    subtitle("Observed proportion of 1's: `mn'", pos(11))
```



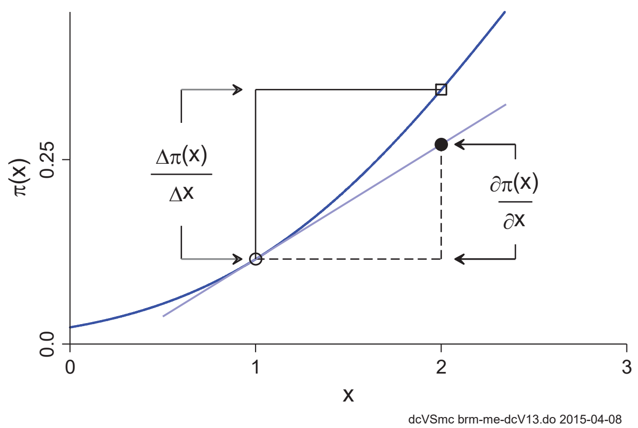
## Marginal effects: changes in probabilities

*The change in Pr(y|**x**) for a change of δ in $x_k$, holding other regressors at specific values.*

### Decisions when using MEs

1. How much change?
   - o An infinitely small change leads to the marginal change (MC).
   - o A finite change leads to a discrete change (DC).
2. Where is the change computed?
   - o The value of the ME depends on where it is evaluated
3. Since the value depends on where you compute the ME, how to you summarize the effect of a variable?

## Marginal change and discrete change






## Marginal change versus discrete change

I focus on DC but everything can be done with MC.

Marginal change

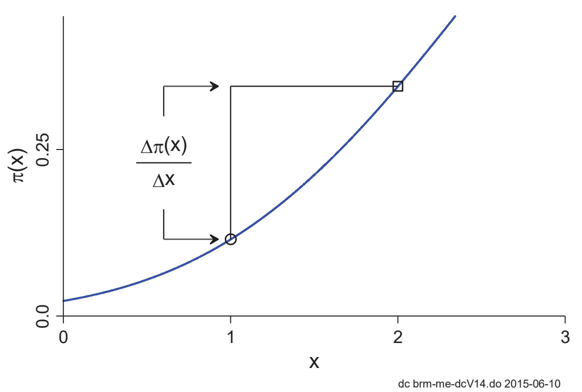
1. MC is the instantaneous rate of change
   o The speedometer reading
2. If probability curve is approximately linear, the MC tells you how much the probability would change for a unit change in $x_k$
   o If your speed is constant, the speedometer tells you how far you will go in an hour

Discrete change

1. DC is the change that occurs over a fixed distance.
2. I find the DC to be substantively clearer.
3. Unless your field uses MC, DC is more intuitive.



## Discrete change DC(x 1→2)




Here's how the DC is computed...



1. Compute probabilities at start and end values of $x_k$

   $\Pr(y=1 \mid \mathbf{x}^*, \text{Start } x_k)$: Starting probability given **x*** & start value $x_k$.

   $\Pr(y=1 \mid \mathbf{x}^*, \text{End } x_k)$: Ending probability after changing only $x_k$.

2. Discrete change

$$\frac{\Delta \Pr(y=1 \mid \mathbf{x})}{\Delta x_k} = \Pr(y=1 \mid \mathbf{x}^*, \text{End } x_k) - \Pr(y=1 \mid \mathbf{x}^*, \text{Start } x_k)$$

3. Interpretation

   *Changing $x_k$ from start to end changes the probability by DC($x_k$), holding other variables at the specific values.*

4. Example using means:

$$\frac{\Delta \Pr(y=1 \mid \mathbf{x})}{\Delta x_k} = \Pr(y=1 \mid wc=1, \bar{\mathbf{x}}) - \Pr(y=1 \mid wc=0, \bar{\mathbf{x}})$$

   *Attending college increases the probability of women being in the labor force by .19, holding other variables at their means.*

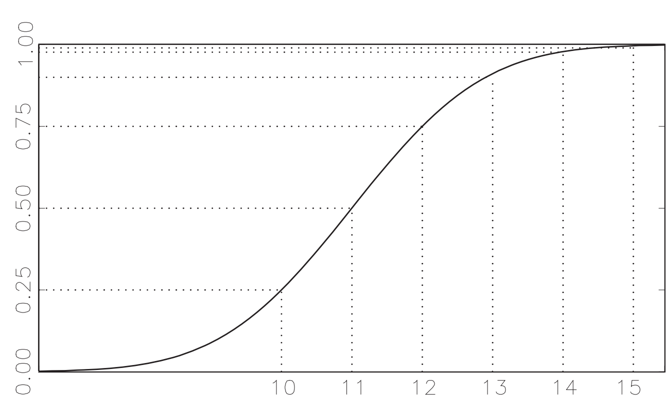


## What affects the size of the DC?

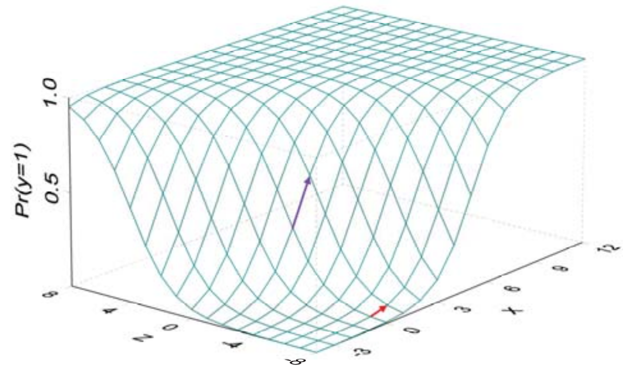
1. The regression coefficients as illustrated earlier
2. Start value of $x_k$
   o The curve changes more rapidly at some places
3. The amount of change in $x_k$
   o Bigger changes have bigger effects (assuming no polynomials)
4. Value of other regressors and their regression coefficients
   o Effectively, these change the intercept which changes the effect



## Effect of start value on DC(x+1)



## Effect of other variables on DC(x+1)





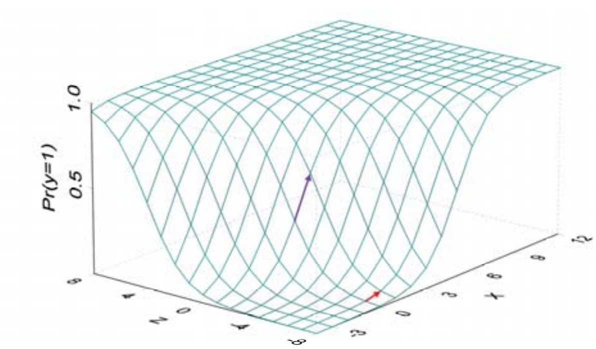

## Amount of change in $x_k$

1. 0 to 1 for binary variables: male compared to female
2. Fixed change
   - o Unit change: increase education by 1 year
   - o Standard deviation change: increase age by a standard deviation
   - o Minimum to maximum: lowest to highest income (or trimmed extremes)
   - o Four years of education or $10,000 of income
3. Changes in linked variables: increase age and age-squared
4. Changes in several variables: white males compared to black females



# Summarizing marginal effects

Since the ME depends on the levels of <u>all</u> variables in the model, how do you summarize the effect with a scalar value?.






## Common summary measures

<u>Marginal effects at representative values (MER)</u>

- o Look at values that are substantively interesting
- o Or at multiple sets of values (Madalla)

<u>Marginal effects at the mean (MEM)</u>

- o Use the mean as a representative values
- o Is anyone average? Is the mean a good summary?

<u>Average marginal effect (AME)</u>

- o Compute ME for each observation and then average

### Which is the best one?

The one that answers your substantive question!



## Discrete change at representative values (DCR)

Think of a specific set of values **x*** and compute DC($x_k$|**x***)

$$DCR: \frac{\Delta \Pr\left(y=1 \mid \mathbf{x}^*\right)}{\Delta x_k}$$

## Discrete change the mean (DCM)

Hold all variables held at theirs means

$$DCM: \frac{\Delta \Pr\left(y=1 \mid \bar{\mathbf{x}}\right)}{\Delta x_k}$$

## Average discrete change (ADC)

Compute the DC at each $\mathbf{x}_i$ and take the average.

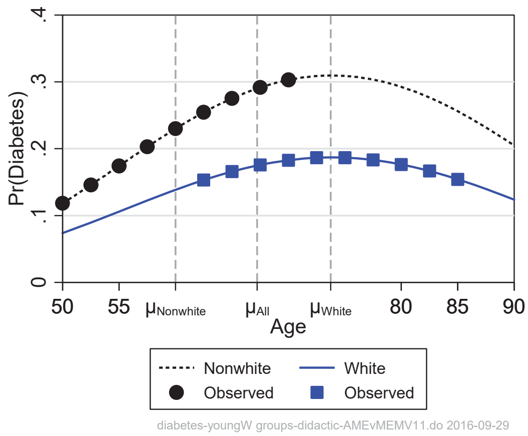
$$DC_i: \frac{\Delta \Pr\left(y=1 \mid \mathbf{x}_i\right)}{\Delta x_{ik}} \quad \text{the} \quad ADC = \frac{1}{N}\sum_{i=1}^{N} DC_i$$



## Which measure of change? ADC, DCM, DCR

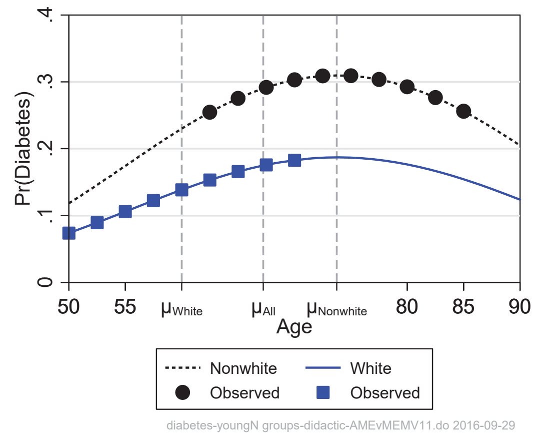
1. ADC and DCM can be similar, but are not asymptotically equivalent.
2. Traditionally, DCM prevailed since ADC requires N times more computation.
   - o Newer software computes both measures.
3. A critique of DCM is that the mean might not correspond to anyone.
   - a. The DC at the mean of binary x roughly averages the DC for the two groups.
   - b. DCR can use modal values of the binary variables, but this ignores everyone who is in a less well represented group.
   - c. DCR can be computed for both groups
4. Consider two examples illustrating what DCR and ADC can and cannot tell you

## Positive ADC for nonwhites; zero ADC for whites






## Zero ADC for nonwhites; positive ADC for whites






## Characteristics of the ADC

1. The ADC replaces one mean with another.
   - Computation **at the mean** is replaced by **the mean of**.
   - Means are only one characteristic of a distribution.
2. The ADC might not be close to the effect for anyone in the sample.
   - Suppose effects are small for men and large for women. The ADC does not indicate this difference.
   - If you are planning an intervention, are you interested in the average effect or the average for those you want to target (e.g., high risk youth)?
   - Later we look at the distribution of effects for all observations
3. The ADC reflects the regression surface and the distribution of values of x's in the sample



## Characteristics of the DCR

1. The representative values have to be substantively useful and meaningful.
2. It reflects the regression surface at a specific location that does *not* depend on the distribution of observations

## What do you want to know determines the best measure

1. The best measure is the one that addresses the goals of your research
2. What do you want to know?



## Testing marginal effects

1. The delta methods is most often used to computes standard errors.
2. You can test $H_0$: ME=0 or compute a confidence interval.
   - Is the effect of having another child significant?
3. More test complex hypotheses can be tested if the effects are computed *simultaneously*
   - Is effect of age the same for men and women?

## Confidence intervals

1. Confidence intervals describe the distribution of estimators over repeated samples
   - The 95% CI indicates that we expect our estimate to fall within the CI 95 percent of the time in repeated sampling.
   - If the CI overlaps 0, you cannot reject that hypothesis that ME=0
2. You should not use overlapping CIs to conclude that effects are NOT significantly different
3. Details in *Testing Marginal Effects*



## Overview of mchange

```
. mchange, atmeans

logit: Changes in Pr(y) | Number of obs = 753

Expression: Pr(lfp), predict(pr)

                |    Change    p-value
----------------+---------------------
k5              |
             +1 |    -0.324      0.000
            +SD |    -0.180      0.000
:::

Predictions at base value

             | not in LF      in LF
-------------+---------------------
  Pr(y|base) |     0.422      0.578

Base values of regressors

           |                                  2.          3.          1.          1.
           |        k5       k618      agecat      agecat          wc          hc
-----------+-------------------------------------------------------------------------
        at |      .238       1.35        .385        .219        .282        .392
::
```

Code: options - help mchange for more information

Note that output in slides is sometimes edited

**`amount(one sd)`**: specify amount of change

**`atmeans`**: hold regressors at their means

**`stats(est pvalue ll ul)`**: show estimates, p-value, and CI

**`brief`**: reduce output

**`dec(#)`**: number of decimal digits



# Examples of marginal effects - #4

## MEM: marginal effects at the mean

```
. mchange, atmeans amount(one sd) stats(est p ll ul) dec(2)

logit: Changes in Pr(y) | Number of obs = 753

Expression: Pr(lfp), predict(pr)

               |   Change    p-value         LL          UL
---------------+--------------------------------------------
k5             |
            +1 |    -0.32       0.00      -0.40       -0.25
           +SD |    -0.18       0.00      -0.23       -0.13
k618           |
            +1 |    -0.02       0.34      -0.05        0.02
           +SD |    -0.02       0.34      -0.06        0.02
agecat         |
 40-49 vs 30-39 |   -0.15       0.00      -0.24       -0.05
   50+ vs 30-39 |   -0.31       0.00      -0.42       -0.19
   50+ vs 40-49 |   -0.16       0.00      -0.26       -0.06
wc             |
  college vs no |    0.19       0.00       0.09        0.28
hc             |
  college vs no |    0.03       0.51      -0.06        0.13
```



```
lwg            |
            +1 |     0.14       0.00       0.08        0.20
           +SD |     0.08       0.00       0.05        0.12
inc            |
            +1 |    -0.01       0.00      -0.01       -0.00
           +SD |    -0.10       0.00      -0.15       -0.05

Base values of regressors

           |                          2.          3.          1.          1.
           |       k5      k618    agecat      agecat          wc          hc
-----------+------------------------------------------------------------------
        at |      .24       1.4       .39         .22         .28         .39

           |      lwg       inc
-----------+----------------------
        at |      1.1        20

1: Estimates with margins option atmeans.
```



## A unit change: +1

$$\frac{\Delta \Pr\left(y=1 \mid \mathbf{x}^*\right)}{\Delta x_k} = \Pr\left(y=1 \mid \mathbf{x}^*, x_k^*+1\right) - \Pr\left(y=1 \mid \mathbf{x}^*, x_k^*\right)$$

```
               |   Change    p-value         LL          UL
---------------+--------------------------------------------
k5             |
            +1 |    -0.32       0.00      -0.40       -0.25
```

*For a woman who is average on all characteristics, an additional young child decreases the probability of being in the labor force by .32 (p<.01).*

Plugging in the specific values, the peculiarity of the mean is clear:

*For a woman who is average on all characteristics, increasing from .24 to 1.24 young child decreases the probability of being in the labor force by .32 (p<.01).*



## A standard deviation change: +SD

$$\frac{\Delta \Pr\left(y=1 \mid \mathbf{x}^*\right)}{\Delta x_k} = \Pr\left(y=1 \mid \mathbf{x}^*, x_k^*+s_k\right) - \Pr\left(y=1 \mid \mathbf{x}^*, x_k^*\right)$$

```
               |   Change    p-value         LL          UL
---------------+--------------------------------------------
            +1 |    -0.01       0.00      -0.01       -0.00
           +SD |    -0.10       0.00      -0.15       -0.05
```

*A standard deviation increases in family income, about $20,000, decreases the probability of being in the labor force by .10 (p<.01, two-tailed test), holding other regressors at their means.*



## A change from 0 to 1

Since wife's college was entered **`i.wc`** , the change is automatically from 0 to 1.

```
               |   Change    p-value         LL          UL
---------------+--------------------------------------------
wc             |
  college vs no |    0.19       0.00       0.09        0.28
hc             |
  college vs no |    0.03       0.51      -0.06        0.13
```

*If an average woman attends college, her probability of being in the labor force is .19 greater than that of a woman who does not attend college (p<.01). The effect of the husband attending college is small and not significant.*

## Change from the minimum to the maximum with trimming

1. This is a useful indication of the total possible effect of a variable:

$$\frac{\Delta \Pr(y=1|\mathbf{x}^*)}{\Delta x_k} = \Pr(y=1|\mathbf{x}^*, \max x_k) - \Pr(y=1|\mathbf{x}^*, \min x_k)$$

```
. mchange lwg inc, atmeans amount(range) dec(2) brief

             |    Change    p-value
-------------+----------------------
lwg          |
       Range |      0.67       0.00
inc          |
       Range |     -0.65       0.00
```

2. Option `trim()` removes extreme values:

```
. mchange lwg inc, atmeans amount(range) trim(5) dec(2) brief

             |    Change    p-value
-------------+----------------------
lwg          |
   5% to 95% |      0.27       0.00
inc          |
   5% to 95% |     -0.29       0.00
```



## AME: average marginal effects

1. Compute the DC for every observation at its observed values:

$$DC_i : \frac{\Delta \Pr(y=1|\mathbf{x}_i)}{\Delta x_{ik}}$$

2. Average the individual DCs:

$$ADC = \frac{1}{N}\sum_{i=1}^{N} DC_i$$

3. Consider the ADC(wc)

$$DC_i : \frac{\Delta \Pr(y=1|\mathbf{x}_i)}{\Delta wc(0 \rightarrow 1)} = \Pr(y=1|\mathbf{x}_i, wc=1) - \Pr(y=1|\mathbf{x}_i, wc=0)$$

$$ADC = \frac{1}{N}\sum_{i=1}^{N} DC_i$$



```
. mchange k5 wc, amount(one) dec(2) // <= no atmeans

logit: Changes in Pr(y) | Number of obs = 753

Expression: Pr(lfp), predict(pr)

               |    Change    p-value
---------------+----------------------
k5             |
            +1 |     -0.28       0.00
wc             |
 college vs no |      0.16       0.00

Average predictions

             | not in LF      in LF
-------------+----------------------
  Pr(y|base) |      0.43       0.57
```

No base values since we average over all cases.



## Comparing AME and MEM (excluding p-values)

1. AME for k5

   *On average having one more young child decreases the probability of being in the labor force by .28.*

2. MEM for k5

   *For someone who is average on all characteristics, having an additional young child is expected to decrease the probability of LFP by .32.*

3. AME for wc

   *On average women who attend college have a probability of being in the labor force that is .16 greater than those who do not attend college.*

   *The average impact of a women attending college is to increase her probability of LFP by.16.*

4. MEM for wc

   *If an average woman attends college, her probability of being in the labor force is .19 greater than that of an average woman who does not attend college.*



## MEM vs AME

1. MEM and AME answer different questions.
2. The AME is probably the best replacement for regression coefficients in the LRM.
   - o When comparing groups this is NOT necessarily the case
3. If MEM and AME differ, figure out what it tells you about the process.

```
               | AME       | MEM       |
               |    Change |    Change | AME-MEM
---------------+-----------+-----------+----------
k5         +SD |    -0.153 |    -0.180 |    0.027
k618       +SD |    -0.018 |    -0.021 |    0.003
---------------+-----------+-----------+----------
wc college vs  |     0.162 |     0.186 |   -0.024
---------------+-----------+-----------+----------
inc        +SD |    -0.086 |    -0.101 |    0.016
```



# Distribution of effects

*On average if a woman attends college her probability of labor force participation increase by .162.*

1. Averages do not indicate variation in the sample.
   - o The effect of college might be different for different people
2. This suggests looking at the distribution DC's for each observation:

$$DC_i : \frac{\Delta \Pr(y=1|\mathbf{x}_i)}{\Delta x_{ik}}$$

## Histogram of effects for wc #1

3. Using **`margins, generate()`** create variable **`DCwc1`** with DC(wc)

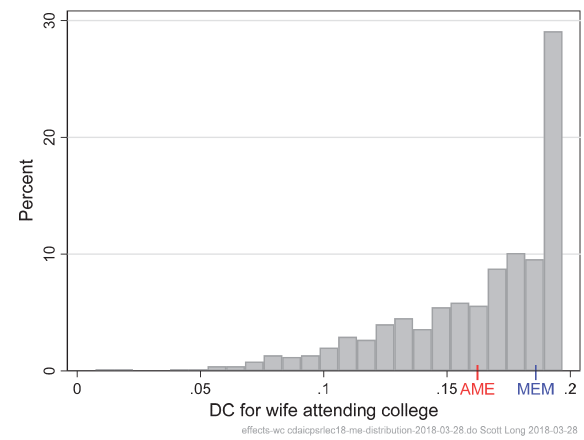
```
margins, dydx(wc) generate(DCwc)

Average marginal effects                        Number of obs     =        753
Model VCE    : OIM

Expression   : Pr(lfp), predict()
dy/dx w.r.t. : 1.wc

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          wc |
    college  |   .1624037    .0440211     3.69   0.000      .076124    .2486834
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

. codebook DCwc*, compact

Variable   Unique      Mean       Min       Max  Label
-------------------------------------------------------------------------
DCwc1           1         0         0         0  margins generate variabl...
DCwc2         753  .1624037  .0074083  .1968259  margins generate variabl...
-------------------------------------------------------------------------
```

4. The variable **dcwc2** had the effects for each case.

5. Plotting the results...



## Distribution of DC for wife attending college for wc






## Code for plotting the distribution of effects

```
margins, dydx(wc) // AME
local adc = el(r(b),1,2) // add ADC(wc) to local

margins, dydx(wc) atmeans // MEM
local dcm = el(r(b),1,2) // add DCM(wc) to local

histogram DCwc2, xlab(0(.05).20) ylab(0(10)30, grid) ///
    percent bin(25) color(gs10) fcolor(gs12) ///
    /// add labels for ADC and DCM
    text(-1.5 `adc' "ADC", color(red*.8)  placement(center)) ///
    text(-1.5 `dcm' "DCM", color(blue*.8) placement(center)) ///
    text( 0   `adc' "|"  , color(red*.8)  placement(center)) ///
    text( 0   `dcm' "|"  , color(blue*.8) placement(center))
```



## Effects of BMI on diabetes - #2

1. The example uses a model predicting diabetes from a later chapter.
2. BMI affects diabetes

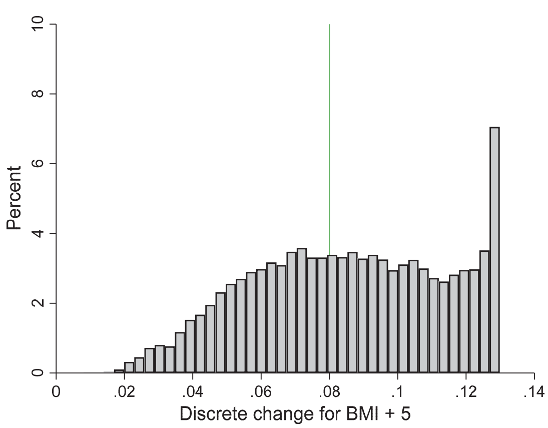
```
. sum bmi

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         bmi |     16,221    27.80409    5.796451   10.57755    82.6728
```

3. The ADC(bmi+5) is:

```
. mchange bmi, amount(sd) delta(5) decimal(8)

svy logit: Changes in Pr(y) | Number of obs = 16221

Expression: Pr(diabetes), predict(pr)

             |    Change     p-value
-------------+----------------------
bmi          |
      +delta | 0.08005615   0.00e+00

1: Delta equals 5.
```



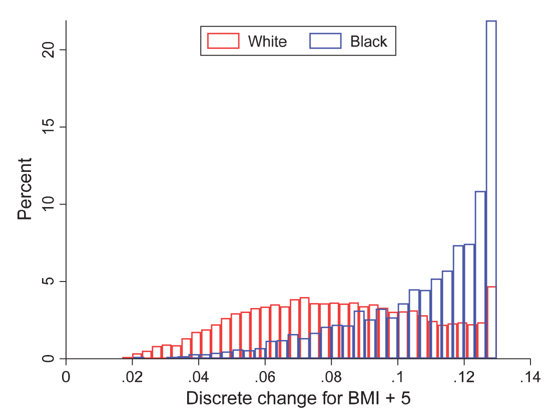
## Distribution of DC(bmi+sd)






## Distribution of DC(bmi+sd) by race

To show how affects vary by race



### Computing DC(bmi+sd)

1. The effects for each observation cannot be created with **`dydx()`** which computes MCs or DCs for **`i.`***variables*
2. I crate predicted probabilities at the observed BMI and observed + 5:

```
. margins, at(bmi=gen(bmi)) at( bmi=gen(bmi+5)) gen(PRbmi)

Predictive margins                              Number of obs     =     16,221
                                                Subpop. no. obs   =     15,677

Model VCE    : Linearized
Expression   : Pr(diabetes), predict()

1._at        : bmi             = bmi
2._at        : bmi             = bmi+5

             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .1793669   .0035909    49.95   0.000     .1721734    .1865604
          2  |    .259423     .00647    40.10   0.000     .2464621    .2723839

. codebook PRbmi*, compact

Variable    Unique      Mean       Min       Max  Label
-------------------------------------------------------------------------
PRbmi1       14798  .1984852   .013618   .980003  margins generate varia...
PRbmi2       14798  .2837495  .0227459  .9880413  margins generate varia...
-------------------------------------------------------------------------
```



3. Next, create the ADC for each observation:

```
. gen double DCbmi = PRbmi2 - PRbmi1
. lab var DCbmi "DC for increase of 5 in bmi"
```

4. To test if my computations are right, take the average which matches the results from **`mchange`**

```
. svy: mean DCbmi // verify this equals adc from mchange
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      56       Number of obs   =      16,248
Number of PSUs   =     112       Population size =  70,963,962
                                 Design df       =          56

             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       DCbmi |   .0800561   .0004647      .0791253     .080987
--------------------------------------------------------------

. gen double DCbmi = PRbmi2 - PRbmi1
. lab var DCbmi "DC for increase of 5 in bmi"
. svy: mean DCbmi // ADC to verify

             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       DCbmi |   .0800561   .0004647      .0791253     .080987
--------------------------------------------------------------
```



### *Code for plotting dual histograms*

```
twoway ///
  (hist DCbmi if race == 1, percent fcol(none) bcol(red*.8)) ///
  (hist DCbmi if race == 2, percent fcol(none) bcol(blue*.8)), ///
  xlab(0(.02).14) xtitle("Discrete change for BMI + 5") ///
  legend(symxsize(7) order(1 "White" 2 "Black") pos(12) ring(0)) ///
  scale(1.1) plotregion(margin(zero) lcol(white))
```



## Summary of marginal effects

1. A summary measure of the effect of a variable is often useful.
2. In LRM, the regression coefficients are used as long as nonlinearities (e.g., powers) are not included.
   - o The $\beta_x$ is DC(x) in this case
3. In BRM, regression coefficients are rarely the effect of interest.
   - o OR's are used, but are limited as discussed below.
4. Change in the probability is the best way to summarize effects.
   - o ADC and DCM are often close, but ADC is preferred as a single measure in most cases.
   - o Multiple DCR's might be the best approach.
5. But:

   *Summary measures are only summaries!*

6. Remember, *the model is nonlinear....*



## Predictions for ideal types or profiles - #6

1. What types of people are you interested in? Are there interesting clusters of characteristics that occur together?
2. Defining profiles makes you to think about where to look in the data
3. Comparing predictions across profiles helps you understand your data and the effects of variables
4. We will compute these types and later test if they have the same Pr(LFP)

```
                             |    Pr(y)        ll        ul
-----------------------------+-----------------------------
              Average person |    0.578     0.539     0.616
     Younger lower educ w kids |    0.159     0.068     0.251
        Young more educ w kids |    0.394     0.234     0.554
   Middle age higher educ w kids |    0.754     0.681     0.828
          Older w higher educ |    0.631     0.528     0.734
```



### An "average person"

1. **`mtable`** options
   - o **`atmeans`** to hold variables at their means.
   - o **`ci`** to include CI for predictions instead of p-value
   - o **`clear`** to start a new table
   - o **`rowname()`** to label the results
2. Make the predictions

```
. mtable, rowname(Average person) atmeans ci clear
Expression: Pr(lfp), predict()
                |    Pr(y)       ll       ul
----------------+---------------------------
 Average person |    0.578    0.539    0.616
Specified values of covariates
                                              2.        3.        1.        1.
           |       k5      k618      agecat    agecat        wc        hc
-----------+--------------------------------------------------------------
   Current |     .238      1.35        .385      .219      .282      .392
           |
           |      lwg       inc
-----------+--------------------
   Current |      1.1      20.1
```

## Confidence intervals

1. It usually is not interesting to test if a probability is 0.
2. Instead, confidence intervals are use to demonstrate the precision of the estimate.
3. For example,

   *The predicted probability of labor force participation for an average person is .58 with a 95% confidence interval from .54 to .62.*

   *The estimated probability of labor force participation is .58 (95%CI: .54, .62).*

   *Our results suggest that the predicted probability of labor force participation could be as small as .54 or as large as .62 with 95 percent confidence.*



## Young, lower class, less educated mom

1. We specify all values with **at()**:

```
* note: in 1975 $2.10 is min wage; .75 for lwg

. mtable, rowname(Younger lower educ w kids) ///
>     at(agecat=1 k5=2 k618=0 inc=10 lwg=.75 hc=0 wc=0) below ci twidth(28)

Expression: Pr(lfp), predict()

                            |    Pr(y)        ll        ul
----------------------------+------------------------------
             Average person |    0.578     0.539     0.616
   Younger lower educ w kids |    0.159     0.068     0.251

Specified values of covariates

          |                               2.         3.         1.         1.
          |       k5      k618      agecat     agecat         wc         hc
----------+---------------------------------------------------------------------
    Set 1 |     .238      1.35        .385       .219       .282       .392
  Current |        2         0           .          .          .          .

          |
          |      lwg       inc    agecat         wc         hc
----------+------------------------------------------------
    Set 1 |      1.1      20.1         .          .          .
  Current |      .75        10         1          0          0
```



## Young, more educated moms

1. Profile is defined as:

   **agecat==1 & k5==2 & k618==0 & wc==1 & hc==1**

2. Where should I hold **lwg** and **inc**?
   - Global means for the entire sample are too large.
   - Local means based on individuals who meet our profile are better.
3. Computing local means and saving them:

```
sum lwg if agecat==1 & k5==2 & k618==0 & wc==1 & hc==1
    local mnlwg = r(mean)
sum inc if agecat==1 & k5==2 & k618==0 & wc==1 & hc==1
    local mninc = r(mean)
```

4. Making the predictions

```
. mtable, at(agecat==1 k5==2 k618==0 wc==1 hc==1 inc=`mninc' lwg=`mnlwg') ///
>     rowname(Young more educ w kids) atmeans below ci twidth(28)
```



## Middle aged, educated dad with kids

```
sum inc if agecat==2 & k5==0 & wc==1 & hc==1
    local mninc = r(mean)
sum lwg if agecat==2 & k5==0 & wc==1 & hc==1
    local mnlwg = r(mean)
sum k618 if agecat==2 & k5==0 & wc==1 & hc==1
    local mnlk618 = r(mean)

mtable, at(agecat==2 k5==0 k618=`mnlk618' ///
           wc==1 hc==1 inc=`mninc' lwg=`mnlwg') ///
    rowname(Midage higher educ w kids) atmeans ci below twidth(28)
```

## More educated older couples

```
sum inc if agecat==3 & wc==1 & hc==1 & k618==0 & k5==0
    local mninc = r(mean)
sum lwg if agecat==3 & wc==1 & hc==1 & k618==0 & k5==0
    local mnlwg = r(mean)

mtable , at(agecat==3 k5==0 k618==0 wc==1 hc==1 inc=`mninc' lwg=`mnlwg') ///
    rowname(Older w higher educ) atmeans ci below  twidth(28)
```



## Summary of ideal types

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

                             |    Pr(y)        ll        ul
-----------------------------+------------------------------
              Average person |    0.578     0.539     0.616
    Younger lower educ w kids |    0.159     0.068     0.251
       Young more educ w kids |    0.394     0.234     0.554
Middle age higher educ w kids |    0.754     0.681     0.828
          Older w higher educ |    0.631     0.528     0.734

Specified values of covariates
::
```

1. Which variables seem most important?
2. In our commands for ideal types, we could add the option **statistics(ci)** to add confidence intervals to the table.
3. Later we consider testing if predictions are equal, such as:

   *Older women with higher education have significantly lower chances of being in the labor force than more educated middle aged with children.*



## Tables of predicted probabilities - #7

1. The ideal types suggest young children and wife's education are important
2. Predictions across categories of children and education summarize the effects

```
   Number      Did Not
   of Young    Attend     Attended
   Children    College    College     Difference
   -------------------------------------------
      0          .60        .77          .17
      1          .28        .46          .18
      2          .09        .17          .09    < due to rounding
      3          .02        .05          .03
   -------------------------------------------
```

3. Where do these numbers come from?

## Curves behind the table of probabilities

1. Let Θ be the linear combination of all variables except **k5** and **wc**.
2. The model is

$$\Pr\left(y=1\,|\,\mathbf{x}\right)=\Lambda\left(\beta_0+\beta_{k5}k5+\beta_{wc}wc+\Theta\right)$$
$$=\Lambda\left(\beta_0^*+\beta_{k5}k5+\beta_{wc}wc\right)$$

3. If **wc=0**

$$\Pr\left(y=1\,|\,\mathbf{x},wc=0\right)=\Lambda\left(\beta_0^*+\beta_{k5}k5\right)$$

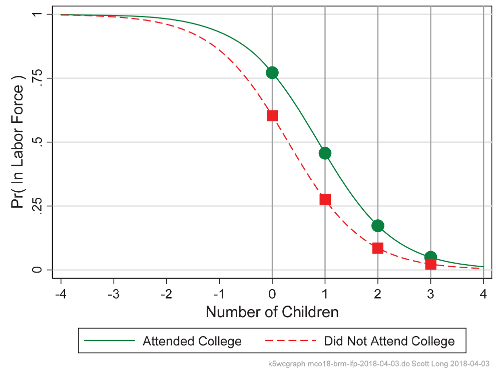
4. If **wc=1**

$$\Pr\left(y=1\,|\,\mathbf{x},wc=1\right)=\Lambda\left(\beta_0^*+\beta_{k5}k5+\beta_{wc}\right)$$
$$=\Lambda\left(\left[\beta_0^*+\beta_{wc}\right]+\beta_{k5}k5\right)$$
$$=\Lambda\left(\beta_0^{**}+\beta_{k5}k5\right)$$

5. These are parallel curves as shown on the next page.



| # Young Children | Not College | Attended College | Difference |
|---|---|---|---|
| 0 | .60 | .77 | .17 |
| 1 | .28 | .46 | .18 |
| 2 | .09 | .17 | .09 |
| 3 | .02 | .05 | .03 |






## Quick table for predictions by levels of two variables

```
. mtable, atmeans at(wc=(0 1) k5=(0 1 2 3))

Expression: Pr(lfp), predict()

          |       k5        wc     Pr(y)
----------+-----------------------------
        1 |        0         0     0.604
        2 |        0         1     0.772
        3 |        1         0     0.275
        4 |        1         1     0.457
        5 |        2         0     0.086
        6 |        2         1     0.173
        7 |        3         0     0.023
        8 |        3         1     0.049

Specified values of covariates

          |                   2.         3.         1.
          |     k618      agecat     agecat         hc        lwg        inc
----------+---------------------------------------------------------------
  Current |     1.35        .385       .219       .392        1.1       20.1
```



## Local and global means - #7.3

1. We held other variables at the global means
   - o Do college educated women without children have the same levels of income and wages and those without college and 3 young children?
2. Local means hold variables at levels local to other variables being examined held constant
   - o For example, the mean age for those with 3 young children
3. Predictions with local means are computed with **if** and **atmeans**
   - a. Create a selection variable that defines the group of interest.
   - b. Use **if** with **mtable** to select these cases.
   - c. The, **atmeans** compute means within the **if** group.



## Local means for tables using if

1. Select cases **if k5==0** and use **atmeans**

```
. mtable if k5==0, atmeans estname(k5_0) at(wc=(0 1) k5=0) atvars(1.wc)

       1.
       wc      k5 0
------------------
        0     0.583        <= prediction for k5==0 and wc==0
        1     0.757        <= prediction for k5==0 and wc==1

                             2.          3.           1.
       k5        k618    agecat      agecat           hc         lwg         inc
--------------------------------------------------------------------------------
    0.000       1.279     0.436       0.269        0.358       1.107      19.987
```



2. Adding predictions for **k5=1**

```
. mtable if k5==1, atmeans estname(k5_1) at(wc=(0 1) k5=1) atvars(_none) ///
>     right
```

   - o **right** places new results to the right of the current results
   - o **atvars(_none)** means don't add atvars to table

3. Adding predictions for **k5=2** and **k5=3**.

```
. mtable if k5==2, atmeans estname(k5_2) at(wc=(0 1) k5=2) atvars(_none) ///
>     right
. mtable if k5==3, atmeans estname(k5_3) at(wc=(0 1) k5=3) atvars(_none) ///
>     right

       1.
       wc      k5 0      k5 1      k5 2      k5 3
-------------------------------------------------
        0     0.583     0.337     0.154     0.017
        1     0.757     0.530     0.288     0.037
```

4. Next, compute the DC(wc|k5=j)

## DC(wc|k5=j) using local means

1. **dydx(***var***)** tells computes marginal effects for *var*.
   - If *var* is a **i.***var*, it computes DC; else MC

```
mtable if k5==0, atmeans dydx(wc) stat(est p) clear long ///
    roweqnm(DCwc) coleqnm(k5_0)
mtable if k5==1, atmeans dydx(wc) stat(est p) right long coleqnm(k5_1)
mtable if k5==2, atmeans dydx(wc) stat(est p) right long coleqnm(k5_2)
mtable if k5==3, atmeans dydx(wc) stat(est p) right long coleqnm(k5_3)
```

2. Results

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

          | k5_0      k5_1      k5_2      k5_3
          |  d Pr(y)   d Pr(y)   d Pr(y)   d Pr(y)
----------+----------------------------------------
DCwc      |
  d Pr(y) |    0.173     0.193     0.134     0.020
        p |    0.000     0.000     0.003     0.070

Specified values of covariates
::
```

3. The differences decrease with number of children and are not significant with three young children.



## Sensitivity review for global and local means

1. Did using local means change the conclusions?
   - Trends are similar.
   - Biggest differences are for one and two children.

```
             |      wc=0       wc=1     Change      pvalue
-------------+--------------------------------------------
global       |
        k5=0 |      0.60       0.77       0.17        0.00
        k5=1 |      0.27       0.46       0.18        0.00
        k5=2 |      0.09       0.17       0.09        0.01
        k5=3 |      0.02       0.05       0.03        0.09
-------------+--------------------------------------------
local        |
        k5=0 |      0.58       0.76       0.17        0.00
        k5=1 |      0.34       0.53       0.19        0.00
        k5=2 |      0.15       0.29       0.13        0.00
        k5=3 |      0.02       0.04       0.02        0.07
```

2. Substantively, I would draw the same conclusions
   - Which predictions would you use?



## Table of predictions

1. Tables can be very effective to show results for a few categorical variables
2. While graphs can be used for continuous variables, tables often work better
   - They are more compact
   - They are easier to see the specific result
3. The mtable command is a wrapper for margins to make predictions easier to read.
   - In the sample do-file, add **details** to the **mtable** commands to see the output from **margins**!
   - A few **mtable** tricks follow
   - See Long and Freese for detailed explanations



## * Local means for tables using over()

1. The **over(***overvars***)** option loops through the *overvars*
   - For each value of *overvars* it runs **mtable** or **margins** on observations that equal that value
2. The command

```
mtable, over(k5) at(wc=(0 1)) atmeans
```

Is equivalent to:

```
mtable if k5==0, at(wc=(0 1)) atmeans
mtable if k5==1, at(wc=(0 1)) atmeans
mtable if k5==2, at(wc=(0 1)) atmeans
mtable if k5==3, at(wc=(0 1)) atmeans
```

3. Using **over()** is quick but the output isn't pretty



```
. mtable, estname(k5_0) at(wc=(0 1)) atvars(1.wc k5) atmeans over(k5)

Expression: Pr(lfp), predict()

            |        1.
            |       wc        k5      k5_0
------------+-----------------------------
   0.k5#c.1 |        0         0     0.583
   1.k5#c.1 |        0         1     0.337
   2.k5#c.1 |        0         2     0.154
   3.k5#c.1 |        0         3     0.017
   0.k5#c.2 |        1         0     0.757
   1.k5#c.2 |        1         1     0.530
   2.k5#c.2 |        1         2     0.288
   3.k5#c.2 |        1         3     0.037

Specified values where .n indicates no values specified with at()

          |  No at()
----------+---------
  Current |       .n
```



## * Creating a nicer table

1. **mtable** stacks predictions from previous **mtable** results.
2. **clear** creates a new table dropping any prior results
3. **right** place estimates to the right.
4. **atvars(_none)** adds no new *atvars* to the table.
5. **dydx(wc)** requests a discrete change in **wc**.

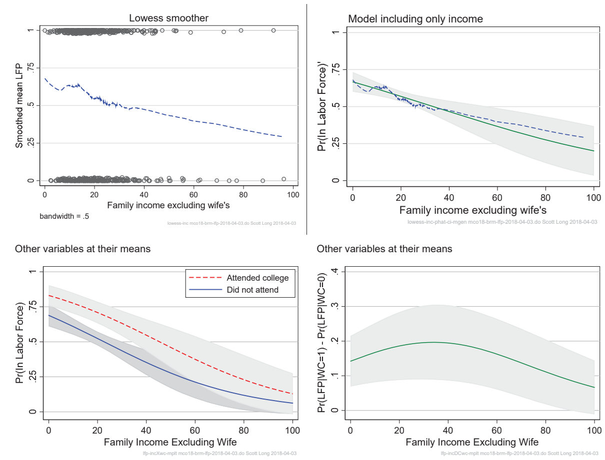
```
. qui mtable, atmeans at(wc=(0) k5=(0 1 2 3)) atvars(k5) ///
>     clear estname(NoCol)
. qui mtable, atmeans at(wc=(1) k5=(0 1 2 3)) atvars(_none) ///
>     right estname(College)
. mtable, atmeans dydx(wc) at(k5=(0 1 2 3)) atvars(_none) ///
>     right estname(Diff) stats(est p)

          |       k5     NoCol   College      Diff         p
----------+--------------------------------------------------
        1 |        0     0.604     0.772     0.168     0.000
        2 |        1     0.275     0.457     0.182     0.001
        3 |        2     0.086     0.173     0.087     0.013
        4 |        3     0.023     0.049     0.027     0.085
```

## Plotting predictions

1. For continuous variables, graphs can be effective
2. Non-parametric plots such as lowess let's you assess your functional form
3. Plots of predictions from your model can quickly summarize relationships
   - Multiple predictions can be included in one graph
4. Sometimes the graph shows you that you don't need the graph
5. Examples of plots



## Examples of graphs we will create






## Overview of plotting predictions

1. To get graphs to look the way you want is not fun
   - **marginsplot** is a great way to get quick plots
     - You can customize it like any **graph** command
     - It is difficult to combine results from multiple predictions
   - **mgen** creates variables with predictions to plot with **graph**
2. Creating graphs is irritating!
   - Use templates rather than starting from scratch
   - Use Stata's menu system to find options



## Tools for making graphs

1. Graphs have thousands of irritating options to make them look just right
2. Getting your graphs right is important
3. You also want them to be uniform

Locals for graph options

1. Create locals with options:
   `local ylab "0(.25)1., grid gmin gmax"`
2. Then **`` `ylab' ``** means **`0(.25)1., grid gmin gmax`**
3. All graph commands can use **`` ylabel(`ylab') ``**

Graph formats so graph print properly

1. Use EMF, EPS or PDF formats so your graphs scale

Graph captions so you know where it came from

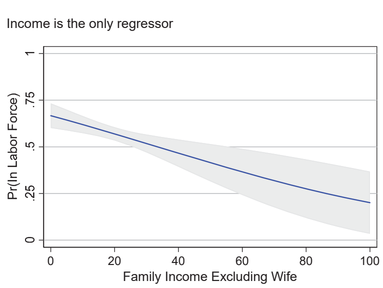
```
local graphname lfp-incXwc-mplt
marginsplot, … ///
    caption("`graphname' `tag'", size(*.5) pos(5) col(gs10)) scale(1.1)
```



## Lowess plots - #9

1. Is the relationship between income and LFP substantively reasonable?

Income is the only regressor




2. A lowess plot is non-parametric and does not constrain the shape of the relationship between a regressor and the outcome
3. A lowess is a first step in evaluating how a regressor is related to the outcome.

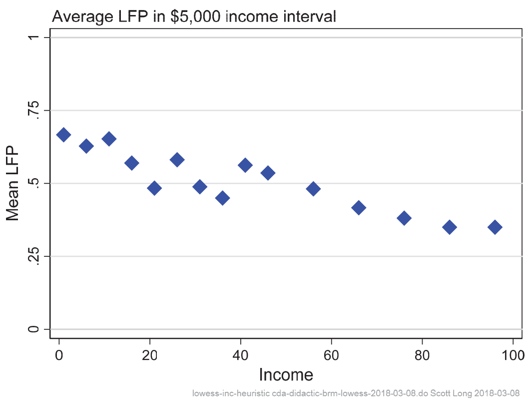


## Intuition behind a lowess plot

1. Compute mean LFP within income intervals of 5:

```
. sum lfp if inc>=0 & inc<5

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         lfp |        12    .6666667     .492366          0          1

::

. sum lfp if inc>=35 & inc<40

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         lfp |        18    .3888889    .5016313          0          1

::
```

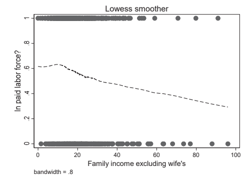
2. Plotting the means by income






## The lowess command

1. A lowess plot is a sophisticated way to do this that uses “sliding” intervals.
2. Simple running **`lowess lfp inc`** is often enough



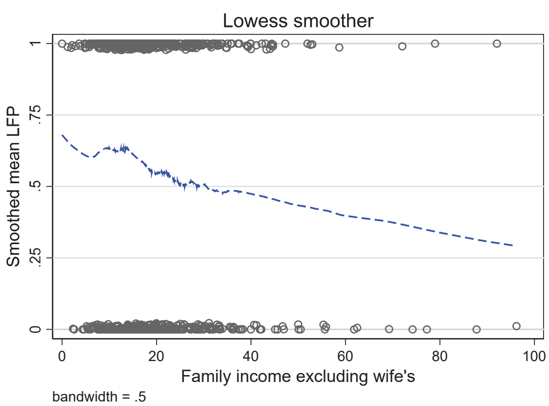

3. Options perfect the graph

```
sort inc
lowess lfp inc, jitter(3) generate(lowesslfp) bwidth(.5) ///
    msym(oh) lineopt(lcol(blue) lwid(*1.3)) ///
    xlab(0(20)100) ytitle(Smoothed mean LFP) ///
    ylab(0(.25)1., grid gmin gmax) yline(0 1, lcol(gs13)) ///
```

4. **`gen(lowesslfp)`** saves the predictions to a variable.

Graph on next page...



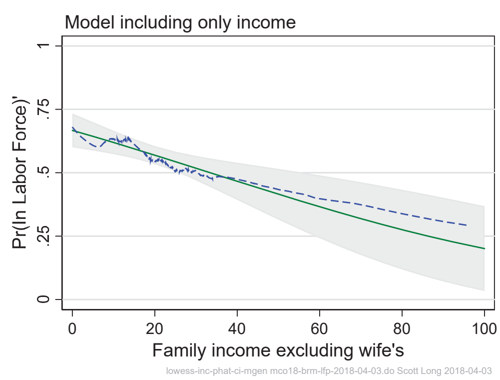



5. How would this compare to the predictions from logit?



## Predictions from logit

1. To assess the logit model, compare lowess to logit predictions
2. I am satisfied that my logit is a reasonable in how income is related to LFP






## Combining logit predictions with lowess

1. Fit the logit model

    **`logit lfp inc`**

    Or we could fit

    **`logit lfp k5 k618 i.agecat i.wc i.hc lwg inc`**

2. **`mgen`** computes predictions and saves predictions as variables:
3. Predict outcome as income increases from 0 to 100 by 5:

```
. mgen, at(inc=(0(5)100)) atmeans stub(PLT) predlabel(Logit prediction)

Predictions from: margins, at(inc=(0(5)100)) atmeans predict(pr)

Variable   Obs Unique      Mean       Min       Max  Label
---------------------------------------------------------------------------
PLTpr1      21     21  .4223433  .2008354  .6669906  Logit prediction
PLTll1      21     21   .320794  .0336831  .6007513  95% lower limit
PLTul1      21     21  .5238926  .3679877  .7332299  95% upper limit
PLTinc      21     21        50         0       100  Family income excluding
---------------------------------------------------------------------------

. label var PLTpr "Logit prediction"
```



4. Variables beginning with PLT are created by mgen:

```
. format %9.3g lfp inc PLTpr PLTll PLTul PLTinc
. list lfp inc PLTpr PLTll PLTul PLTinc in 1/25, clean nolabel

        Observed
        Variables             mgen variables

       lfp     inc   PLTpr1   PLTll1   PLTul1   PLTinc
  1.     1   -.029     .667     .601     .733        0
  2.     1     1.2     .644     .588     .699        5
  3.     0     1.5     .619     .573     .666       10
  4.     1    2.13     .595     .556     .633       15
  5.     1     2.2     .569     .534     .605       20
::
 15.     1       5     .319     .176     .462       70
 16.     1    5.12     .297     .146     .448       75
 17.     1    5.12     .276     .119     .433       80
 18.     1    5.32     .255    .0938     .417       85
 19.     0    5.33     .236    .0714     .401       90
 20.     1    5.49     .218    .0514     .385       95
 21.     0    5.55     .201    .0337     .368      100
 22.     0       6        .        .        .        .
 23.     0       6        .        .        .        .
 24.     1    6.02        .        .        .        .
 25.     1    6.25        .        .        .        .
```

5. Combine the variables created by **`mgen`** and **`lowess`**

```
local linPRopt  "msym(i) lcol(green) lpat(solid)"
local linLOWopt "msym(i) lcol(blue)  lpat(dash)"

graph twoway ///
    (rarea PLTul PLTll PLTinc, color(black*.1)) /// shaded CI
    (connected PLTpr PLTinc, `linPRopt') /// line for prob
    (connected lowesslfp inc, `linLOWopt'), ///
    subtitle("Model including only income", position(11)) ///
    ytitle("Pr(In Labor Force)'") ylab(0(.25)1., grid gmin gmax) ///
    xtitle("Family income excluding wife's") legend(off)
```



## Plot income in full model using marginsplot

1. Consider the full model

```
logit lfp k5 k618 i.agecat i.wc i.hc lwg inc
```

2. Compute predictions holding other variables at their means:

```
. margins, at(inc=(0(5)100)) atmeans

Adjusted predictions                            Number of obs     =        753

Expression   : Pr(lfp), predict()

1._at        : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               1.agecat        =    .3957503 (mean)
               2.agecat        =    .3851262 (mean)
               3.agecat        =    .2191235 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
               lwg             =    1.097115 (mean)
               inc             =           0
::
```

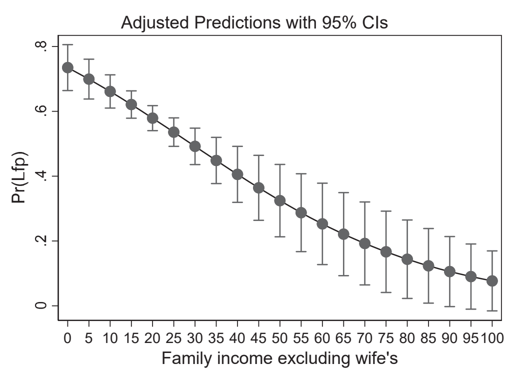


```
21._at       : k5              =    .2377158 (mean)
               k618            =    1.353254 (mean)
               1.agecat        =    .3957503 (mean)
               2.agecat        =    .3851262 (mean)
               3.agecat        =    .2191235 (mean)
               0.wc            =    .7184595 (mean)
               1.wc            =    .2815405 (mean)
               0.hc            =    .6082337 (mean)
               1.hc            =    .3917663 (mean)
               lwg             =    1.097115 (mean)
               inc             =         100

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .7349035   .0361031    20.36   0.000     .6641427    .8056643
::
         21  |   .0768617   .0472071     1.63   0.103    -.0156624    .1693858
------------------------------------------------------------------------------
```



3. All that is needed to plot predictions are the commands:

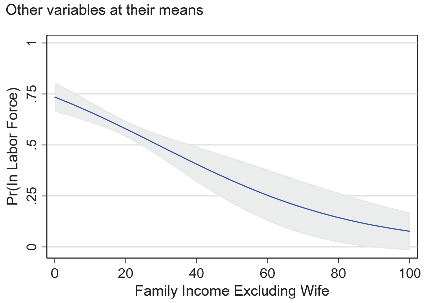
```
margins, at(inc=(0(5)100)) atmeans
marginsplot
```





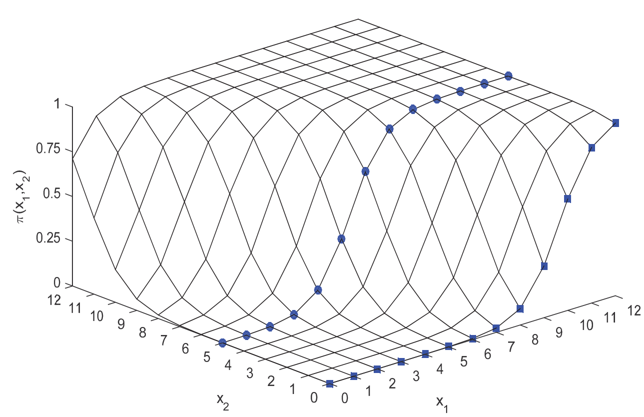

## Customizing marginsplot

```
local labYopt "labsiz(*1.1) glwid(*.7) glcol(black*.3) grid gmin gmax"
local labXopt "labsiz(*1.1) glwid(*.7) glcol(black*.3) nogrid"
local titleopt "ring(2) pos(11) size(*1)"
local lin1opt  "lcol(blue*1.) lpat(solid) msym(i)  msiz(*1.) mcol(blue*1.)"

marginsplot, recastci(rarea) ciopts(color(black*.1)) ///
    ylab(0(.25)1, `labYopt') xlab(0(20)100, `labXopt') ///
    plot1opts(`lin1opt') plotopts(lwidth(*1)) ///
    xtitle("Family Income Excluding Wife") ytitle("Pr(In Labor Force)") ///
    title("Other variables at their means" " ", `titleopt')
```

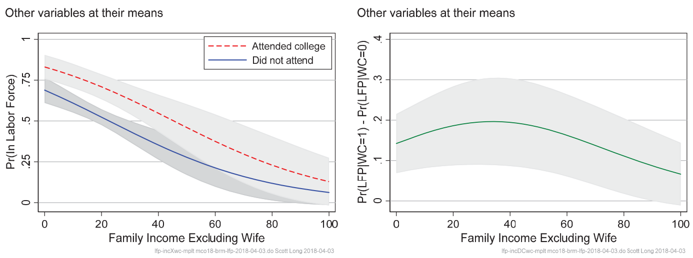





## Plotting predictions for multiple variables



## Predictions for income by wife's college - #10.3




*The probability of a woman being in the labor force decreases as family grows. For incomes, women who attend college are significantly more likely to be in the labor force, although the difference decreases at higher incomes.*

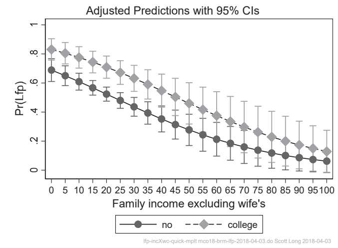


## Plotting predictions at two levels of wc

1. Let x* be the fixed values for all variable except **age** and **wc**.
2. Compute $\Pr\left(y=1 \mid \mathbf{x}^{*}, WC=0, INC\right)$ and $\Pr\left(y=1 \mid \mathbf{x}^{*}, WC=1, INC\right)$

```
margins, at(inc=(0(5)100) wc=(0 1)) atmeans
```

3. **marginsplot** is smart enough to know you want two curves. And quickly gives you enough information to know if you want to use the graph:




4. I can add the **noci** option to suppress the CIs.



## *Perfecting the marginsplot*

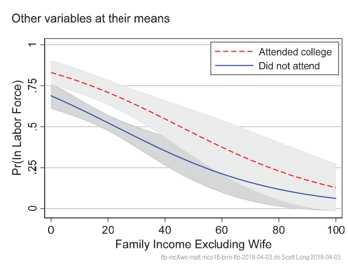
1. Or we make a presentation quality graph

```
marginsplot, recastci(rarea) ///
    ci1opts(color(black*.2)) ci2opts(color(black*.1)) ///
    plot1opts(`lin1opt') plot2opts(`lin2opt') ///
    plotopts(lwidth(*1.2))
    ylab(0(.25)1, `labYopt') xlab(0(20)100, `labXopt') ///
    xtitle("Family Income Excluding Wife") ///
    ytitle("Pr(In Labor Force)") ///
    title("Other variables at their means" " ", `titleopt') ///
    legend(order(4 "Attended college" 3 "Did not attend") ///
        ring(0) pos(1) rows(2)) ///
```



## DC(wc|inc): are the curves significantly different

1. Do women who go to college have higher rates of LFP for all levels of income?




2. The figure shows two curves with their CIs.
   - If the CI's do not overlap, predictions are significantly different.
   - If the CI's overlap, significance is unknown
3. We need to test if the predictions are significantly different



## Testing differences in predictions

1. We want to test

   $H_0$: DC(wc|inc) = 0

2. We compute

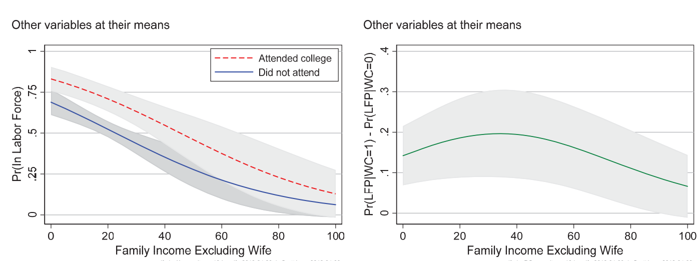
   [ Lower bound DC(wc|inc), Upper bound DC(wc|inc) ]

3. Since **wc** is entered into the model as **i.wc**, **margins, dydx(wc)** computes DC(wc).

```
margins, dydx(wc) at(inc=(0(5)100)) atmeans

local graphname lfp-incDCwc-mplt
marginsplot, recastci(rarea) ciopts(color(black*.1)) ///
    ylab(0(.1).4, `labYopt') xlab(0(20)100, `labXopt') ///
    plot1opts(`linDCopt') plotopts(lwid(*1)) ///
    xtitle("Family Income Excluding Wife") ///
    ytitle("Pr(LFP|WC=1) - Pr(LFP|WC=0)") ///
    title("Other variables at their means" " ", `titleopt') ///
    caption("`graphname' `tag'", size(*.5) pos(5) col(gs10)) scale(1.1)
. mgen, dydx(wc) at(inc=(`inc_rng')) atmeans stub(PLTdc) ///
>     predlabel(DC of wc by income)
```

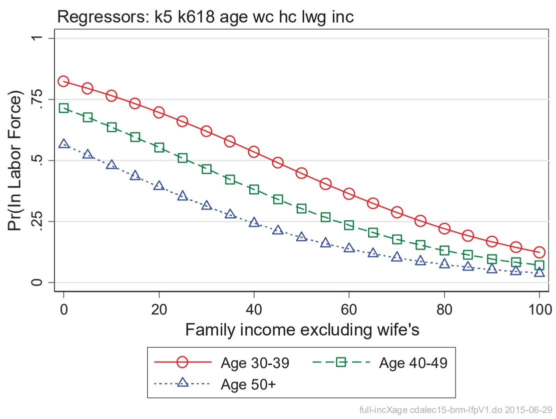


## Comparing overlapping CI's to tests of DC




Clearly, overlapping confidence intervals can be misleading

## The effect of income on LFP by age category






## Graphs for discovery versus presentation

1. <u>You need a graph to decide if you need a graph.</u>
   - If a graph is simple, you probably don't need it in a paper, but you need the graph to know you don't need it.
2. You need tools to create graphs quickly and must organize them efficiently or you won't do it.
   - Use templates to speed up the process of making graphs
   - Use a file viewer to quickly examine graphs



# Interpretation using odds ratios - #12

1. Odds ratios are a <u>common</u> and <u>unsatisfactory</u> method of interpretation.
2. Do you really want a ratio of ratios?

## Buying apples or pears

1. Are pears at \$.40 enough cheaper to buy instead of \$.45 apples?

| | | |
|---|---|---|
| *Cost index* for apples: | .818 | = (\$.45) / (\$1-\$.45) |
| *Cost index* for pears: | .667 | = (\$.40) / (\$1-\$.40) |
| *Cost index ratio*: | 1.23 | = (\$.45/(\$1-\$.45)) / (\$.4/(\$1-\$.4)) |
| *Cost difference*: | \$0.05 | = \$.45 - \$.40 |
| *Cost ratio*: | 1.120 | = \$.45 / \$.40 |

2. Which would you use to decide if you want apples?



## What is an odds ratio?

### Probability and odds at x and x+1

Probability: $\Pr(y=1\mid x)$ $\qquad$ $\Pr(y=1\mid x+1)$

Odds: $\Omega(x)=\dfrac{\Pr(y=1\mid x)}{\Pr(y=0\mid x)}$ $\qquad$ $\Omega(x+1)=\dfrac{\Pr(y=1\mid x+1)}{\Pr(y=0\mid x+1)}$

### The OR is a ratio of ratios of probabilities

Odds ratios: $OR(x\rightarrow x+1)=\dfrac{\Omega(x+1)}{\Omega(x)}$

*For a unit increase in x, the odds increase by a factor of OR(x) holding other variables constant.*



## Logit is linear in the log of the odds

1. A <u>logit</u> is the name for the log of the odds
2. The logit model is linear in the logit

$$\ln\left[\frac{\Pr(y=1\mid\mathbf{x})}{1-\Pr(y=1\mid\mathbf{x})}\right]=\ln\Omega(\mathbf{x})=\mathbf{x\beta}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$$

*For a unit change in $x_k$, the logit is expected to change by $\beta_k$, holding other variables constant.*

3. Linearity is fine, but what does a change of $\beta_k$ logits mean?

*Each additional young child decreases the logit of being in the labor force by 1.39, holding other variables constant.*

4. To understand the change in logit, we transform it to odds



## Change logit to odds and compute odds ratio (ORs)

1. Take the exponential of the logit with a focus on $x_3$:

$$\begin{aligned}\Omega(\mathbf{x})&=\exp\left[\ln\Omega(\mathbf{x})\right]=\exp(\mathbf{x\beta})\\&=e^{\beta_0+\beta_1x+\beta_2x_2+\beta_3x_3}\\&=e^{\beta_0}e^{\beta_1x_1}e^{\beta_2x_2}e^{\beta_3x_3}=\Omega(\mathbf{x},x_3)\end{aligned}$$

2. Let $x_3$ change by 1

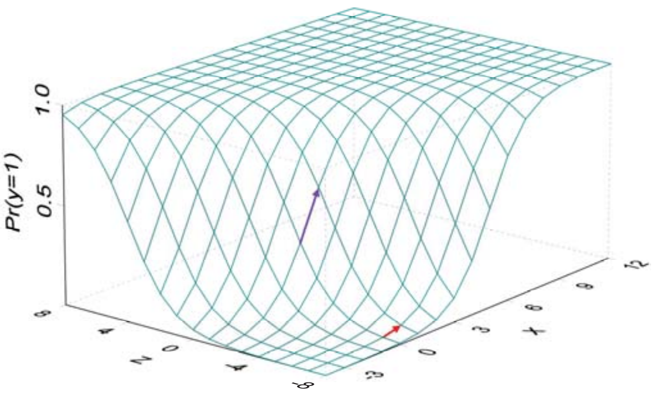
$$\begin{aligned}\Omega(\mathbf{x},x_3+1)&=e^{\beta_0}e^{\beta_1x_1}e^{\beta_2x_2}e^{\beta_3(x_3+1)}\\&=e^{\beta_0}e^{\beta_1x_1}e^{\beta_2x_2}e^{\beta_3x_3}e^{\beta_3}\end{aligned}$$

3. The odds ratio

$$\frac{\text{Ending }\Omega}{\text{Starting }\Omega}=\frac{\Omega(\mathbf{x},x_3+1)}{\Omega(\mathbf{x},x_3)}=\frac{e^{\beta_0}e^{\beta_1x_1}e^{\beta_2x_2}e^{\beta_3x_3}e^{\beta_3}}{e^{\beta_0}e^{\beta_1x_1}e^{\beta_2x_2}e^{\beta_3x_3}}=e^{\beta_3}$$

4. *The OR does not depend on the level of other variables*

## A change of 1 in x has the same OR everywhere






## Logit estimates

```
. logit lfp k5 k618 i.agecat i.wc i.hc lwg inc

Logistic regression                               Number of obs   =        753
                                                  LR chi2(8)      =     124.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -452.72367                       Pseudo R2       =     0.1207

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          k5 |  -1.391567   .1919279    -7.25   0.000    -1.767739   -1.015395
        k618 |  -.0656678    .068314    -0.96   0.336    -.1995607    .0682251
             |
      agecat |
          2  |  -.6267601    .208723    -3.00   0.003     -1.03585   -.2176705
          3  |  -1.279078   .2597827    -4.92   0.000    -1.788242   -.7699128
             |
        1.wc |   .7977136   .2291814     3.48   0.001     .3485263    1.246901
        1.hc |   .1358895   .2054464     0.66   0.508     -.266778    .5385569
         lwg |   .6099096   .1507975     4.04   0.000      .314352    .9054672
         inc |  -.0350542   .0082718    -4.24   0.000    -.0512666   -.0188418
       _cons |   1.013999   .2860488     3.54   0.000     .4533539    1.574645
------------------------------------------------------------------------------
```



## ORs with listcoef: interpretation on next page

```
. listcoef, constant help

logit (N=753): Factor Change in Odds

  Odds of: 1InLF vs 0NotInLF

----------------------------------------------------------------------
         lfp |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
          k5 |  -1.39157   -7.250   0.000   0.2487   0.4823     0.5240
        k618 |  -0.06567   -0.961   0.336   0.9364   0.9170     1.3199
    2.agecat |  -0.62676   -3.003   0.003   0.5343   0.7370     0.4869
    3.agecat |  -1.27908   -4.924   0.000   0.2783   0.5889     0.4139
        1.wc |   0.79771    3.481   0.001   2.2205   1.4319     0.4500
        1.hc |   0.13589    0.661   0.508   1.1456   1.0686     0.4885
         lwg |   0.60991    4.045   0.000   1.8403   1.4310     0.5876
         inc |  -0.03505   -4.238   0.000   0.9656   0.6651    11.6348
       _cons |   1.01400    3.545   0.000
----------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
```



## Odds ratio: factor change in the odds

1. For a unit change in $x_k$ the odds are expected to change by a factor of $\exp(\beta_k)$, holding other variables constant.

   For $\exp(\beta_k)>1$, the odds are $\exp(\beta_k)$ times larger.

   *By attending college her odds of LFP are 2.22 times larger, holding other variables constant.*

   For $\exp(\beta_k)<1$, the odds are $\exp(\beta_k)$ times smaller.

   *For each additional young child, the odds of LFP are .25 times smaller, ...*

2. For a standard deviation change in $x_k$, the odds are expected to change by a factor of $\exp(s_k\beta_k)$, holding other variables constant.

   *For a standard deviation increase in the log of wages the odds of LFP are 1.43 times larger, ...*



## **TODO DROP:** Percentage change in the odds

1. If the odds change by a factor of 2, they are 100% larger.
2. If the odds change by a factor of .5, they are 50% smaller.
3. In general, **%change = 100*(OR-1)**.

   100% = 100*(2-1) Double odds, is 100% increase

   -50% = 100*(.5-1) Halve odds, is 50% decrease

4. For example

   *By attending college her odds of LFP are 124 percent larger, holding other variables constant.*

   *For an additional young child, the odds of LFP are 77 percent smaller, ...*

   *For a standard deviation increase in the log of wages the odds of LFP are 43 percent larger, ...*

5. To compute these: `listcoef, percent`



## Interpreting odds ratios (ORs)

1. OR is a multiplicative coefficient.
   - Positive effects are greater than one
   - Negative effects are between zero and one
2. Magnitudes of positive and negative ORs are compared by taking the inverse of the negative effect (or vice versa).
   - A positive OR=2 has the same magnitude as a "negative" OR=1/2.
   - An OR=1/10 is larger than OR=2.
3. The effect on the odds of the event not occurring is the inverse of the OR of the event occurring.

   *Being ten years older makes the odds of not being in the labor force 1.9 (=1/.52) times greater, holding other variables constant.*

## Additional examples of ORs

```
. listcoef, constant help

logit (N=753): Factor Change in Odds

  Odds of: 1InLF vs 0NotInLF

----------------------------------------------------------------------
         lfp |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
          k5 |  -1.39157   -7.250   0.000   0.2487   0.4823     0.5240
        k618 |  -0.06567   -0.961   0.336   0.9364   0.9170     1.3199
    2.agecat |  -0.62676   -3.003   0.003   0.5343   0.7370     0.4869
    3.agecat |  -1.27908   -4.924   0.000   0.2783   0.5889     0.4139
        1.wc |   0.79771    3.481   0.001   2.2205   1.4319     0.4500
        1.hc |   0.13589    0.661   0.508   1.1456   1.0686     0.4885
         lwg |   0.60991    4.045   0.000   1.8403   1.4310     0.5876
         inc |  -0.03505   -4.238   0.000   0.9656   0.6651    11.6348
       _cons |   1.01400    3.545   0.000
----------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
. listcoef, constant percent help
```

Interpretations on next page...



**k5:**

*For each additional young child, the odds of employment are decreased by a factor of .25, holding other variables constant.*

```
     |     b         z     P>|z|     e^b    e^bStdX      SDofX
  k5 |  -1.39157  -7.250   0.000   0.2487   0.4823     0.5240
```

**lwg:**

*For a standard deviation increase in wages, the odds of employment are 1.43 times greater, holding other variables constant.*

```
     |     b          z     P>|z|     e^b    e^bStdX      SDofX
 lwg |   0.60991   4.045   0.000   1.8403   1.4310     0.5876
```



## Odds do not translate linearly into probabilities

1. "For a unit increases in X the odds of Y are increase by a factor of OR, holding other variables constant."
   - Where the increase in X begins does not matter
   - The levels of other variables does not matter
2. This seems to make interpretation as simple as βs in linear regression
3. Except the meaning of a given factor change depends on p.
4. Think about doubling the odds of being a victim of a crime
   a. If the odds are 1/100,000,000, they become 2/100,000,000
   b. If the odds are 1/10, they become 2/10
   c. Do these mean the same things in terms of the probability of being a victim?



## OR compared to Pr(y) for groups

1. Two logit models are estimated

```
logit tenure pub phdyr if female==1
logit tenure pub phdyr if female==0
```

where $\exp\left(\hat{\beta}_{pub}^{Women}\right) = \exp\left(\hat{\beta}_{pub}^{Men}\right) = 2$.

2. Suppose these are the probabilities and odds for men and women:
   $p_M = .500$ → $\Omega_M = .500/(1-.500) = 1.000$
   $p_W = .050$ → $\Omega_W = .050/(1-.050) = 0.053$
3. How does doubling the odds change the probability?
   $2^*\Omega_M = 2.000$ → $p_M = 2.000/(2.000+1) = .667$
   $2^*\Omega_W = 0.105$ → $p_W = 0.105/(0.105+1) = .095$
4. Then,
   Δ pM/ Δ pub = .167 = (.667 - .500)
   Δ pW/ Δ pub = .045 = (.095 - .050)
5. Are the effects equal for men and women?



## Advanced for the curious: The OR as a marginal effect

### Computing ORs with predictions and margins

Estimate the model

```
. logit lfp k5 k618 i.agecat i.wc i.hc lwg inc, or nolog

Logistic regression                              Number of obs    =        753

         lfp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
<snip>
```

Compute probabilities and odds

```
. predict double Pinc
. label var  Pinc "Pr(inc)"
. gen double Oinc = Pinc / (1-Pinc)
. label var  Oinc "Odds(inc)"
```

Increase income by 1 and compute probabilities and odds

```
. replace inc = inc + 1 // dangerous to change your data!
. predict double Pincplus
. label var  Pincplus "Pr(x=inc+1)"
. gen double Oincplus = Pincplus / (1 - Pincplus)
. label var  Oincplus "Odds(x=inc+1)"
```



Compute the odds ratio for a unit increase in income

```
. gen double ORinc = Oincplus / Oinc
. label var  ORinc "Odds(x=inc+1) / Odds(x=inc)"
```

The average equals the odds ratio

```
. sum ORinc // average odds ratio

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       ORinc |       753    .9655531    7.06e-09   .9655531   .9655532
```

The logit results

```
         lfp | Odds Ratio   Std. Err.      z
-------------+-----------------------------------
   logit inc |   .9655531   .0079868    -4.24
```

Using margins to compute odds at inc and inc+1

```
. mtable, at(inc=generate(inc)) at(inc=generate(inc+1)) ///
>     expression(predict(pr)/(1-predict(pr))) post

Expression: , predict(pr)/(1-predict(pr))

          |   Margin
----------+---------
        1 |    2.011
        2 |    1.941
```

Estimate the odds ratio

```
. nlcom  (_b[2._at]/_b[1._at]) // estimate OR

       _nl_1:  _b[2._at]/_b[1._at]

             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .9655531   .0079868   120.89   0.000     .9498992     .981207
------------------------------------------------------------------------------
```

Testing if the OR=1 (NOT 0!)

```
. testnl (_b[2._at]/_b[1._at]) = 1 // test OR = 1

  (1)  (_b[2._at]/_b[1._at]) = 1

               chi2(1) =       18.60
           Prob > chi2 =        0.0000
. di sqrt(18.60)
4.3127717
```

The logit results

```
         lfp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  logit  inc |   .9655531   .0079868    -4.24   0.000     .9500254     .9813346
```



Compute the OR for probit

```
. use binlfp4, clear
. probit lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog

Probit regression                               Number of obs   =        753

         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
<snip>
         inc |  -.0210541   .0048205    -4.37   0.000     -.030502   -.0116061

. //  #2d compute marginal prediction

. mtable , ///
>     at(inc=generate(inc)) ///
>     at(inc=generate(inc+1)) ///
>     expression(predict(pr)/(1-predict(pr))) post

Expression: , predict(pr)/(1-predict(pr))

           |   Margin
 ----------+---------
         1 |    2.164
         2 |    2.085
```



```
. nlcom  (_b[2._at]/_b[1._at]) // estimate OR

       _nl_1:  _b[2._at]/_b[1._at]

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   .9634678   .0085774   112.33   0.000     .9466565    .9802791
------------------------------------------------------------------------------

. testnl (_b[2._at]/_b[1._at]) = 1 // test OR = 1

  (1)  (_b[2._at]/_b[1._at]) = 1

               chi2(1) =       18.14
           Prob > chi2 =        0.0000
```

The logit results

```
         lfp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  logit  inc |   .9655531   .0079868    -4.24   0.000     .9500254     .9813346
```



# Overview of models for binary outcomes

## Why so much time on BRM

1. BRM is foundation for many models for ordinal, nominal, and count variables.
2. A deep understanding of BRM makes other models easier to understand.

## Key points

1. Interpretation requires understanding nonlinearity and substance
2. No single method of interpretation is always best
   - Try alternative methods to find which one works best.
3. There are subtle ways in which models for categorical outcomes differs from those for linear regression
   - Be careful about taking what you know about LRM and appying it to BRM.
   - Be careful about interpreting LRM if there are nonlinearities on the RHS



# β1 Estimation, testing, and fit

## Readings and examples

*Long & Freese: 3.1, 3.2, 3.3*

*mdo18-test-fit-*.do; mdo18-svy-*.do*

## Outline

1. Estimation of regression coefficients with SRS
2. Estimation of regression coefficients with complex samples
3. Compound tests of regression coefficients
4. Assessing fit with IC measures
5. $R^2$-type measures of fit



# Estimation with simple random sampling

## Linear regression with OLS

1. OLS minimizes the sum of the squared residuals:

$$SSR = \sum_{i=1}^{N}\left(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}\right)^2 = \sum_{i=1}^{N}\left(\hat{\varepsilon}_i\right)^2$$

2. OLS has a simple "closed-form" formula:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

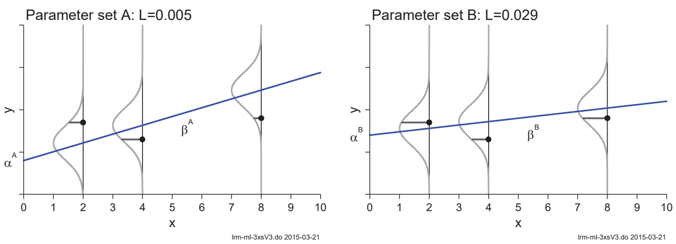
3. The covariance matrix for the estimates

$$\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = Var\left(\hat{\boldsymbol{\beta}}\text{ for X and Z}\right) = \begin{pmatrix} Var\left(\hat{\beta}_X\right) & Cov\left(\hat{\beta}_X,\hat{\beta}_Z\right) \\ Cov\left(\hat{\beta}_Z,\hat{\beta}_X\right) & Var\left(\hat{\beta}_Z\right) \end{pmatrix}$$

TODO: Drop section in LRM on estimation?

## Maximum likelihood estimation in LRM

1. MLE maximize the likelihood of what you observe.




2. For LRM, MLE gives essentially the same results as OLS



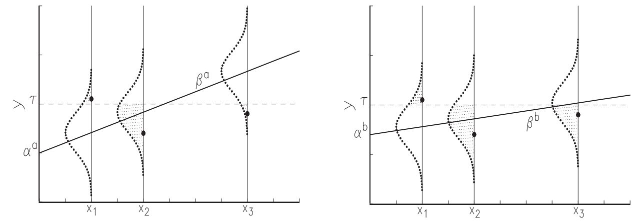
## MLE for binary logit and probit

1. We observe y=1 or y=0. $p_i$ is the probability of observing what was observed

$$p_i = \begin{cases} \Pr(y_i = 1 \mid \mathbf{x}_i) & \text{if } y_i = 1 \text{ is observed} \\ 1 - \Pr(y_i = 1 \mid \mathbf{x}_i) & \text{if } y_i = 0 \text{ is observed} \end{cases}$$

2. If observations are independent the likelihood is $L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} p_i$

3. Which is better?





## Properties of ML estimators

1. Under general conditions, the ML estimates are <u>asymptotically</u>
   - o Consistent: mean of the sampling distribution approaches the true value.
   - o Efficient: data are used as well as possible.
   - o Normal: sampling distribution becomes normal.

### When is the N large enough to justify MLE?

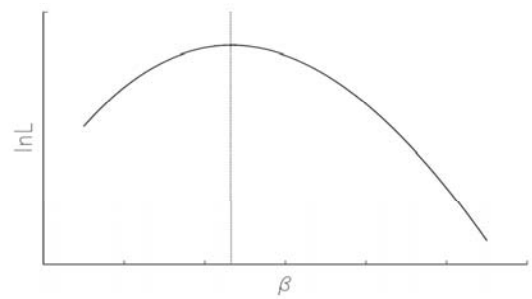
1. It is risky to use MLE for N<100. N>500 is generally safe
2. N's should be larger in some cases
   - o If there are a *lot of parameters*, more observations are needed
   - o Data are <u>ill-conditioned</u> or little variation in the dependent variable
3. Some models seem to require more observations (e.g., ordinal regression)
4. Small depends on the size of smallest outcome. "Rare events" methods deal with this.

### Exact estimation

Run **`help exlogistic`** for details.



## Maximizing the likelihood and numerical methods




1. Algebraic maximization of *ln L*(**β**|**X**,**y**) is not possible
2. Numerical methods search for the maximum using the slope and change in slope of the likelihood equation (i.e., first and second derivatives)
3. Here is the intuition of what happens and what can go wrong



### Numerical methods and climbing a hill

1. Numerical methods are like finding the top of a hill when blindfolded
   - o What direction do you go?
   - o How big of a step will you take? Always the same?
   - o What would it take to make sure you were at the top?
   - o What would you want to know before playing this game?
   - o Will you end up at the same place as another person? Why? Why not?
2. Estimates of coefficients are usually very close in different software, with perhaps small differences in standard errors



## What if problems occur with ML?

1. Types of problems
   - o lack of convergence
   - o convergence to the wrong answer
   - o extremely large standard errors
   - o Instability with minor model changes
2. What to do if you encounter problems
   - o Verify the model specification
   - o Verify the variables and the sample
   - o Rescale varibles with extremely large/small variances
3. If a very large proportion of cases are in one of the categories of the outcome, convergence may be difficult. Firth regression or extreme value logit.

## Perfect Prediction - #1

1. Perfect prediction occurs when the value of a predictor perfectly predicts the outcome

```
   Mentor is | Pubs greater than 10?
       male? |     LoPub      HiPub |     Total
-------------+----------------------+----------
Female mentor |         4          0 |         4
             |    100.00       0.00 |    100.00
-------------+----------------------+----------
  Male mentor |       293          6 |       299
             |     97.99       2.01 |    100.00
-------------+----------------------+----------
       Total |       297          6 |       303
             |     98.02       1.98 |    100.00
```

2. The 0 leads to the following problem
   - The odds of LoPub if female mentor are 4/0 which is undefined.
   - The odds of HiPub if female mentor are 0/4=0.
3. Logit drops the four cases with female mentors since their $p_i$ in the likelihood function is 1.
4. Logit on next page...



```
. logit hipub i.mmale phd, nolog

note: 0.mmale != 0 predicts failure perfectly
      0.mmale dropped and 4 obs not used
```

This means: female mentors are low publishers with probability 1.

```
note: 1.mmale omitted because of collinearity

Logistic regression                               Number of obs   =        299
                                                  LR chi2(1)      =       0.23
                                                  Prob > chi2     =     0.6320
Log likelihood = -29.276794                       Pseudo R2       =     0.0039

------------------------------------------------------------------------------
       hipub |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mmale |
Female men.. |          0  (empty)
 Male mentor |          0  (omitted)
         phd |  -.1927085   .4023944    -0.48   0.632    -.9813871    .5959701
       _cons |  -3.293021   1.272882    -2.59   0.010    -5.787824   -.7982179
------------------------------------------------------------------------------
```



## Overall

1. Numerical methods for ML estimation work very well "when your model is appropriate for your data" (Joreskog)
2. Cramer (1986:10) gives excellent advice

   *Check the data, check their transfer into the computer, check the actual computations (preferably by repeating at least a sample by a rival program), and always remain suspicious of the results, regardless of the appeal.*

3. Perhaps, especially if the results are appealing!



# Estimation with complex samples

## Readings and examples

Heeringa, S., West, B.T., & Berglund, P.A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman Hall/CRC. (HWB)

StataCorp Stata Survey Data Reference Manual. StataCorp LP: College Station, TX.

*Long & Freese, 100-103*

## Overview

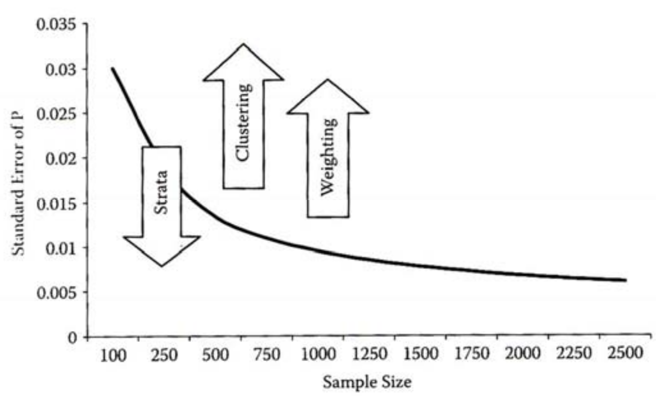
1. Standard software assumes <u>a simple random sample (SRS)</u>
   - Each person in the population has the same probability of selection
   - A person being selected does not affect the probability of another person being selected.
2. SRS is conceptually and mathematically simple, but impractical.



3. Most major datasets use a complex sampling designs.
   - <u>Clustering</u>: clusters are sampled; all cases in cluster are included.
   - <u>Stratification</u>: strata are chosen, not sampled; sampling occurs within strata.
   - <u>Sampling weights</u>: different cases represent different proportions of the population.
4. Complex sampling can:
   - Reduce costs
   - Increases or decrease sampling variability
   - Increase the representation of subpopulations
5. If you do not adjust for complex sampling
   - Variances of estimates are usually underestimated
   - Estimates might be biased
6. Estimation with complex sampling is simple
7. Post-estimation commands work with complex estimation



## Complex sampling designs



## Clustering

1. Clusters or primary sampling units (PSUs) divide the population into distinct and exhaustive groups
   - o Clusters are naturally occurring groups such as blocks in a neighborhood, classes within a school
2. People in a cluster tend to be more similar than people in the population
   - o The makes the sample behave as if it were "smaller"
   - o Since cases are not independent, statistical efficiency is lost

## Stratification

1. Individuals are in disjoint and exhaustive strata based on known characteristics
   - o Racial groups; gender; rural/urban; large/small hospitals; country
2. Size within strata is fixed, not random
3. Different sampling fractions can be used for subpopulations
4. When individual strata are more homogeneous than the population, there is an increase in efficiency. It can "make your sample larger"



## Sampling weights

1. Weights are probabilities of selection.-
2. The probability of inclusion differing across individuals
3. Weights are the share of the population represented be a single observation

## The effective N

1. Each sampling complication changes the "effective N" in the sample (HWB 34)

| Design | Estimator | $\bar{y}$ | $se(\bar{y})$ | Effective n |
|---|---|---|---|---|
| *SRS* | $\bar{y}_{SRS}$ | 40.77 | 2.41 | 32.0 |
| *Clustered* | $\bar{y}_{CL}$ | 40.77 | 3.66 | 13.9 |
| *Stratified* | $\bar{y}_{ST}$ | 40.77 | 2.04 | 44.4 |
| *Stratified, clustered* | $\bar{y}_{CL,ST}$ | 40.77 | 2.76 | 24.4 |

2. The actual n is the same with each design; the effective n varies by design
3. The SE's reflect the change in the "effective n" caused by the design



## Using Stata for survey data

1. There are many subtle points involving the survey commands. Here I provide only an overview. For details see *Stata Survey Data* manual.
2. Always check with the data provider on how to adjust for complex sampling
3. Using **svy** commands involves two steps
   - a. **svyset** to describe the design
   - b. **svy:** for commands such as **svy: logit**



## Example: Health and Retirement Study

1. My example examines

```
arthritis  1=arthritis 0=no arthritis
```

2. Regressors

```
female     Is female?
age        Age at 2006 interview
ed11less   Ed years <= 11?
ed12       Ed years = 12?
ed1315     Ed years 13-15?
ed16plus   Ed years 16 or more?
```

3. The variables the describe the complex sample are:

```
secu     sampling error computation unit
kwgtr    2006 weight: respondent level
stratum  stratum id
```

4. In practice it can be hard to be sure which variables to use.



## Declaring the survey design

1. The design is specified

```
. svyset secu            /// clusters
>     [pweight=kwgtr],    /// weights
>     strata(stratum)     /// stratum
>     vce(linearized) singleunit(missing) // method of compute SE's

      pweight: kwgtr
          VCE: linearized
  Single unit: missing
     Strata 1: stratum
         SU 1: secu
        FPC 1: <zero>
```

2. The output means:

**vce(linearized)** : linearization for estimating standard errors.

**singleunit(missing)** : stratum with single sampling unit is missing.



## Effects of svy adjustment on descriptive statistics

1. Non-survey estimates:

   **sum** *var*

2. Survey adjusted estimates:

   **svy : mean** *var*

   **estat sd**

3. Comparison:

| | srsMean | svyMean | Ratio | srsSD | svySD | Ratio |
|---|---|---|---|---|---|---|
| arthritis | 0.60 | 0.57 | 1.05 | 0.49 | 0.50 | 0.99 |
| age | 68.50 | 66.50 | 1.03 | 11.13 | 10.38 | 1.07 |
| female | 0.59 | 0.54 | 1.08 | 0.49 | 0.50 | 0.99 |
| ed11less | 0.24 | 0.20 | 1.24 | 0.43 | 0.40 | 1.08 |
| ed12 | 0.33 | 0.33 | 1.02 | 0.47 | 0.47 | 1.00 |
| ed1315 | 0.21 | 0.23 | 0.93 | 0.41 | 0.42 | 0.97 |
| ed16plus | 0.21 | 0.25 | 0.85 | 0.41 | 0.43 | 0.94 |

## Effects of survey adjustments on regressions

```
// no survey adjustment

logit arthritis age i.female i.ed4cat
estimates store nosvy
predict nosvyphat
label var nosvyphat "nosvy phat"

// weights and cluster but not stratum

logit arthritis age i.female i.ed4cat ///
    [pweight=kwgtr], cluster(secu)
estimates store wtclstr
predict wtclstrphat
label var wtclstrphat "wtclstr phat"

// weights, clusters, and stratification

svyset secu [pweight=kwgtr], ///
    strata(stratum) vce(linearized) singleunit(missing)
svy: logit arthritis age i.female i.ed4cat
estimates store svy
predict svyphat
label var svyphat "svy phat"
```



```
. // #9 tables of estimated coefficients

--------------------------------------------------
    Variable |     srs       wtclstr       svy
-------------+------------------------------------
         age |     1.046       1.049       1.049
             |     29.57      910.60       21.92  < = t-value
      female |     1.759       1.779       1.779
             |     17.68       12.10       12.99
    ed11less |     1.162       1.206       1.206
             |      3.50        2.57        3.16
      ed1315 |     0.961       0.937       0.937
             |     -0.92       -0.94       -1.21
    ed16plus |     0.703       0.638       0.638
             |     -8.20      -11.47       -8.54
       _cons |     0.054       0.046       0.046
             |    -26.60     -226.92      -19.54
-------------+------------------------------------
           N |     18341       16862       18375
--------------------------------------------------
                                       legend: b/t
. pwcorr nosvyphat wtclstrphat svyphat

             | nosvyphat wtclstrphat  svyphat
-------------+---------------------------
 wtclstrphat |   0.9984    1.0000
     svyphat |   0.9984    1.0000    1.0000
```



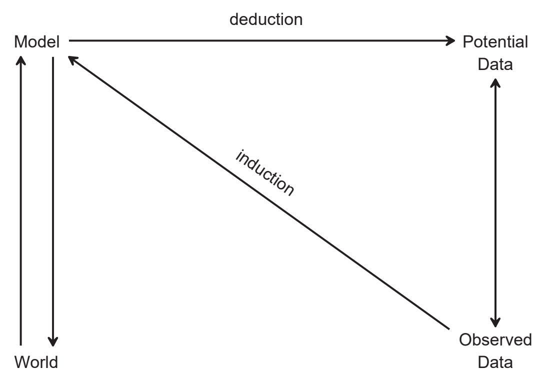
# Hypothesis testing of regression coefficients

1. Hypothesis testing is critical for the effective use of regression models
2. A quick review of the theory of hypothesis testing
3. Wald and LR tests for regression coefficients with a focus on testing multiple coefficients
   - We are more interested in tests of marginal effects, but this lecture explains critical features of testing
4. There are many ways to invalidate standard testing. See this great review:

   Young and Holsteen. 2015. Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis. *Sociological Methods and Research*.



## Barnett's model of inference




test-barnettV1.do jsl 2015-03-12



## The importance of off diagonal element

1. Let $y = \beta_0 + \beta_x x + \beta_z z + \varepsilon$
2. The covariance matrix the X and Z coefficients:

$$\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = Var\left(\hat{\boldsymbol{\beta}} \text{ for X and Z}\right) = \begin{pmatrix} Var\left(\hat{\beta}_X\right) & Cov\left(\hat{\beta}_X, \hat{\beta}_Z\right) \\ Cov\left(\hat{\beta}_Z, \hat{\beta}_X\right) & Var\left(\hat{\beta}_Z\right) \end{pmatrix}$$

3. The diagonal provides the standard errors for tests of single coefficients.
4. Off-diagonal elements reflect how the regression plane "rocks"
   - These are essential for tests of multiple coefficients.



## What affects the variance of an estimate?

1. Let:

   $y = \beta_0 + \beta_x x + \beta_z z + \varepsilon$

2. If $\rho_{XZ}$ is the correlation between X and Z, then:

$$Var\left(\hat{\beta}_X\right) = \frac{\sigma_\varepsilon^2}{N\sigma_X^2\left(1-\rho_{XZ}^2\right)}$$

### Each component affects the variance

Increasing $N$ decreases $Var\left(\hat{\beta}_X\right)$

Increasing $\sigma_X^2$ decreases $Var\left(\hat{\beta}_X\right)$

Increasing $\rho_{XZ}^2$ increases $Var\left(\hat{\beta}_X\right)$

Increasing $\sigma_\varepsilon^2$ increases $Var\left(\hat{\beta}_X\right)$

## Testing individual regression coefficients

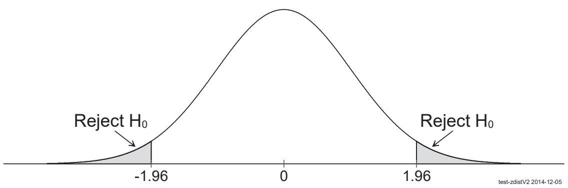
1. Standard output provides tests of regression coefficients
2. If $H_0$: $\beta_k = \beta_k^*$ is true, the ML estimator is

$$\hat{\beta}_k \overset{a}{\sim} \text{Normal}\left(\beta_k^*, \text{Var}\left(\hat{\beta}_k\right)\right)$$

3. The test statistics for $H_0$: $\beta_k = 0$ is

$$z = \left(\hat{\beta}_k - 0\right) / \hat{\sigma}_{\hat{\beta}_k}$$

4. If $H_0$ is true, then $z$ is distributed normally:






5. Two types of errors are possible when testing

| | **Decision** | |
|---|---|---|
| $H_0$: β=0 | **Accept $H_0$** | **Reject $H_0$** |
| **In fact β=0** | No error | **Type I: Pr(reject true)=α**<br>Size of test (the shaded tail). |
| **In fact β≠0** | **Type II: accept false**<br>Power of test. | No error |



## z-test of β's for logit - #11

```
. logit lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog

Logistic regression                             Number of obs   =        753
                                                LR chi2(8)      =     124.30
                                                Prob > chi2     =     0.0000
Log likelihood = -452.72367                     Pseudo R2       =     0.1207

------------------------------------------------------------------------------
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          k5 |  -1.391567   .1919279    -7.25   0.000    -1.767739   -1.015395
        k618 |  -.0656678    .068314    -0.96   0.336    -.1995607    .0682251
::
        1.wc |   .7977136   .2291814     3.48   0.001     .3485263    1.246901
        1.hc |   .1358895   .2054464     0.66   0.508     -.266778    .5385569
         lwg |   .6099096   .1507975     4.04   0.000      .314352    .9054672
         inc |  -.0350542   .0082718    -4.24   0.000    -.0512666   -.0188418
       _cons |   1.013999   .2860488     3.54   0.000     .4533539    1.574645
------------------------------------------------------------------------------
```

*Having young children has a significant effect on the probability of working (z=-7.25, p<0.01 for a two-tailed test).*

*The effect of having older children is not significant (z=-.96, p=.34).*



## Hypothesis for multiple coefficients

1. Our model:

   `logit lfp k5 k618 i.agecat i.wc i.hc lwg inc`

2. Tests involving multiple coefficients
   - Kids have no effect on LFP  $H_0$: $\beta_{k5} = \beta_{k618} = 0$
   - Education has effect on LFP  $H_0$: $\beta_{wc} = \beta_{hc} = 0$
3. Consider algebraic statements and probabilistic statements.



## Algebraic relationships among parameters in hypothesis

1. Consider X and Z from this regression:

   $y = \beta_0 + \beta_X X + \beta_Z Z + \ldots + \epsilon$

2. Hypotheses are *algebraic* statements.

   $H_A$: $\beta_X = 0$  <= income has no effect

   $H_B$: $\beta_Z = 0$  <= wealth has no effect

   $H_C$: $\beta_X = \beta_Z$  <= income & wealth have equal effects

   $H_D$: $\beta_X = \beta_Z = 0$  <= income & wealth have no effects

3. If $H_A$ and $H_B$ are *true*, then $H_C$ and $H_D$ *must be true.*
   - If $\beta_X = 0$ and $\beta_Z = 0$ then mathematically $\beta_X = \beta_Z = 0$



## Statistical conclusions from hypothesis tests

1. Consider two tests of hypotheses:

   $H_A$: $\beta_X = 0$  => test results says $H_A$ might be true or might not

   $H_B$: $\beta_Z = 0$  => test results says $H_B$ might be true or might not

2. Do results from these tests provide insights regarding

   $H_C$: $\beta_X = \beta_Z$

   $H_D$: $\beta_X = \beta_Z = 0$

3. Accepting $H_A$ and $H_B$ does not imply you will accept either $H_C$ or $H_D$!
   - Who stole my wallet?
4. Consider the formula from the LRM and the effect of collinearity:

$$y = \beta_0 + \beta_x x + \beta_z z + \varepsilon$$

$$Var\left(\hat{\beta}_X\right) = \frac{\sigma_\varepsilon^2}{N\sigma_X^2\left(1-\rho_{XZ}^2\right)}$$

## Wald tests of joint hypotheses

1. ML theory shows that:

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \text{Normal}\left(\boldsymbol{\beta}, \mathbf{Var}(\hat{\boldsymbol{\beta}})\right)$$

2. With three coefficients:

$$\text{Var}\begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_X \\ \hat{\beta}_Z\end{pmatrix} = \begin{pmatrix}\sigma^2_{\hat{\beta}_0} & \sigma_{\hat{\beta}_0,\hat{\beta}_X} & \sigma_{\hat{\beta}_0,\hat{\beta}_Z} \\ \sigma_{\hat{\beta}_X,\hat{\beta}_0} & \sigma^2_{\hat{\beta}_X} & \sigma_{\hat{\beta}_X,\hat{\beta}_Z} \\ \sigma_{\hat{\beta}_Z,\hat{\beta}_0} & \sigma_{\hat{\beta}_Z,\hat{\beta}_X} & \sigma^2_{\hat{\beta}_Z}\end{pmatrix}$$

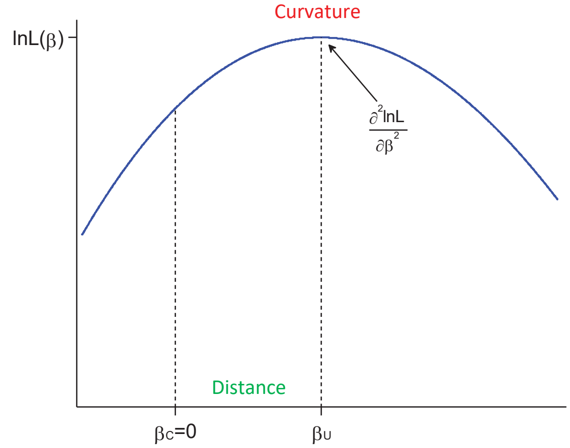
3. $\sigma_{\hat{\beta}_X,\hat{\beta}_Z}$ indicates how the regression plane rocks as the sample changes.

4. The Wald test measures:
   - How far estimates are from hypothesized values.
   - How flat the likelihood functions is.

Graphically...

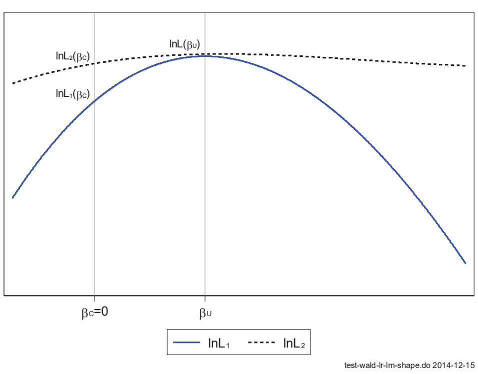


## Wald test and the log likelihood function






## Curvature of ln L curve and the Wald test




1. The flatter the curve, the less "significant" the distance from estimate to constraint

2. How would increasing the sample size affect the curvature?

3. What if the model is "nearly" unidentified?



## Wald test of linear constraints

1. Consider linear constraints $\mathbf{Q}\boldsymbol{\beta} = \mathbf{0}$.
   - $\boldsymbol{\beta}$ is vector of parameters
   - $\mathbf{Q}$ is matrix that combine the β's
2. Examples:
   - $\mathbf{Q}\boldsymbol{\beta} = \beta_1 - \beta_2 = 0$
   - $\mathbf{Q}\boldsymbol{\beta} = \beta_1 = 0$
   - $\mathbf{Q}\boldsymbol{\beta} = \beta_1 = \beta_2 = 0$
3. The Wald statistic equals:

$$W = \left[\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{0}\right]' \left[\mathbf{Q}Var\left(\hat{\boldsymbol{\beta}}\right)\mathbf{Q}'\right]^{-1} \left[\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{0}\right] \sim \chi^2$$

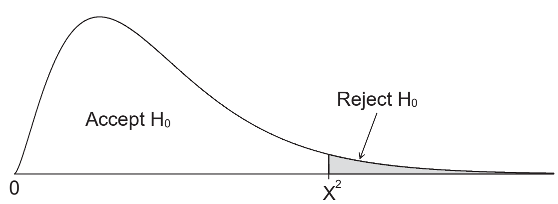
[Distance] [Curvature] [Distance]

4. See Long 1997 for details.



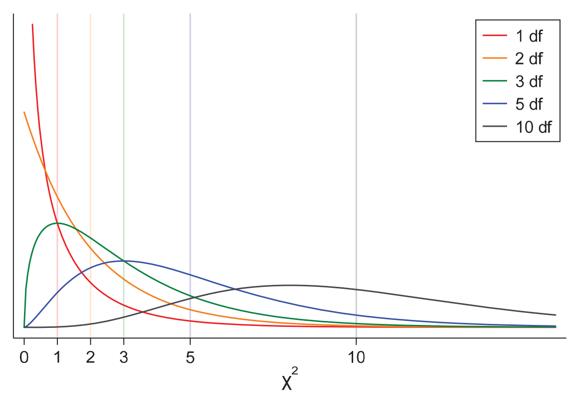
## Sampling distribution of the Wald test

If $H_0$ is true, as N increases the sampling distributions of *W* converges to the chi-square distribution with degrees of freedom equal to the number of constraints being tested.






## Chi-square distribution and degrees of freedom



## Example: Wald tests of regression coefficients - #3

The model is:

```
logit lfp k5 k618 i.agecat i.wc i.hc lwg inc
estimates store logitmodel
```

$H_0$: $\beta_{k5} = 0$

```
. test k5

 ( 1)  [lfp]k5 = 0

           chi2(  1) =   52.57
         Prob > chi2 =    0.0000
```

*The effect of having young children on entering the labor force is significant at the .01 level ($X^2(1)=52.6$).*

Note

Chi-square 52.57 equals the z-value squared -7.25*-7.25.



## How do you know the names of coefficients to use in test?

```
. logit, coeflegend

-------------------------------------------
         lfp |      Coef.  Legend
-------------+-----------------------------
          k5 |  -1.391567  _b[k5]
        k618 |  -.0656678  _b[k618]
             |
      agecat |
      40-49  |  -.6267601  _b[2.agecat]
        50+  |  -1.279078  _b[3.agecat]
             |
          wc |
    college  |   .7977136  _b[1.wc]
             |
          hc |
    college  |   .1358895  _b[1.hc]
         lwg |   .6099096  _b[lwg]
         inc |  -.0350542  _b[inc]
       _cons |   1.013999  _b[_cons]
-------------------------------------------
```



## #14 $H_0$: $\beta_{wc} = \beta_{hc} = 0$

```
. test 1.wc 1.hc // joint test

 ( 1)  [lfp]1.wc = 0
 ( 2)  [lfp]1.hc = 0

           chi2(  2) =   17.83
         Prob > chi2 =    0.000
```

*We can reject the hypothesis that the effects of the husband's and the wife's education are simultaneously zero (X2(2)=17.83, p<.01).*

## #15 $H_0$: $\beta_{wc} = \beta_{hc}$

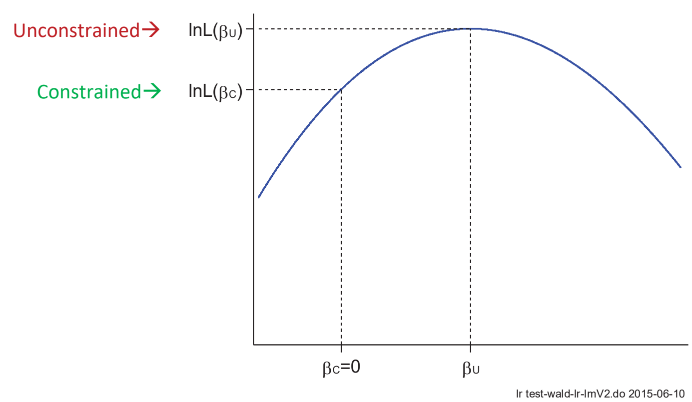
```
. test 1.wc = 1.hc

 ( 1)  [lfp]1.wc - [lfp]1.hc = 0

           chi2(  1) =    3.24
         Prob > chi2 =    0.0719
```

*The hypothesis that the effects of husband's and wife's education are equal is rejected marginally at the .05 level (X2(1) =3.24, p=.07).*



## LR test of *nested* models

The LR test is an alternative to the Wald test.






## Nested models

1. A constrained model = unconstrained model + constraints.
2. Constraints can be things like
   - o A coefficient is 0
   - o Two coefficients are equal
3. Let $M_C$ be the constrained model.
4. Let $M_U$ be the unconstrained model.
5. $M_C$ is ***nested*** in $M_U$.
6. Consider these models:

   M1: $\Pr(y=1|\mathbf{x}) = \Lambda(\beta_0 + \beta_1x_1 + \beta_2x_2)$

   M2: $\Pr(y=1|\mathbf{x}) = \Lambda(\beta_0 + \beta_1x_1 + \beta_3x_3)$

   M3: $\Pr(y=1|\mathbf{x}) = \Lambda(\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_4x_4)$

   M4: $\Pr(y=1|\mathbf{x}) = \Lambda(\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4)$
7. Which are nested?



## Example: LR tests of regression coefficients - #4

H0: $\beta_{wc} = \beta_{hc} = 0$

Full model

```
. logit lfp k5 k618 i.agecat i.wc i.hc lwg inc

Iteration 0:   log likelihood =  -514.8732
::
Iteration 4:   log likelihood = -452.72367

Logistic regression                               Number of obs   =        753
                                                  LR chi2(8)      =     124.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -452.72367                       Pseudo R2       =     0.1207
::

. estimates store full
```

Restricted model

```
. logit lfp k5 k618 i.agecat           lwg inc, nolog
::
. estimates store dropwchc
```

LR test of Ho: wc = hc = 0

```
. lrtest full dropwchc

Likelihood-ratio test                                 LR chi2(2)  =     18.68
(Assumption: dropwchc nested in full)                 Prob > chi2 =    0.0001
```

*The hypothesis that the effects of the husband's and the wife's education are simultaneously equal to zero can be rejected at the .01 level (LRX2(2)=18.7).*



## Summary on testing

1. Under general conditions, the tests are asymptotically equivalent
   - Statisticians generally prefer LR
   - In practice, convenience determines which is used
2. LR and Wald tests can be used with other models using MLE
3. Wald tests can be used when LR cannot
   - With survey estimation, LR tests are not possible
4. Testing multiple coefficients is often critical for your work
5. Avoid these pitfalls:
   - Testing things you aren't interested in (regression coefficients?)
   - Not testing things you are interested in (marginal effects?)
6. Never "add" the results of two or more tests!



## Information criteria to assess fit

1. More complex models fit better at the cost of more parameters.
2. Likely you prefer a model that fits better without “too many” parameters
3. Two information criteria are commonly used to compare fix and complexity

   AIC: Akaike's information criterion

   BIC: Bayesian information criterion
4. These criteria formalize the tradeoff between fit and complexity
5. IC are computed as

   IC = - Fit + Complexity

   = - 2ln*L* + Function of N and # of parameters
   - Fit is negative; more negative is a better fit
   - Complexity is positive so more positive is worse fit
6. A model with a smaller IC is preferred.



## Computing IC measures

1. Define

   N = number of observations

   k = number of parameters

   lnL = log likelihood
2. Then

   IC = fit + complexity

   AIC = -2ln*L* + 2*k // smaller complexity penalty

   BIC = -2ln*L* + ln(N)*k // larger complexity penalty
3. BIC prefers more parsimonious models than AIC



## Comparing models

1. Estimate multiple models
2. Select the model with the smallest IC
3. Consider models M1 and M2

   a. ΔBIC = BIC1 - BIC2

   b. If ΔBIC > 0 choose M2 (BIC1 > BIC2)

   c. If ΔBIC < 0 choose M1 (BIC1 < BIC2)
4. While BIC is <u>not</u> a statistical test, Raftery suggests degrees of evidence

| *Absolute ΔBIC* | *Strength of Evidence* |
|---|---|
| 0 - 2 | Weak |
| 2 - 6 | Positive |
| 6 - 10 | Strong |
| >10 | Very strong |



## Software variations in IC measures

1. BIC in Stata

$$\text{BIC} = \left[-2\ln(likelihood)\right] + \left[\ln N * k\right]$$

   where *k* is the number of parameters
2. BIC'

$$\text{BIC}' = \left[-G^2(M)\right] + \left[df_k^{'} \ln N\right]$$

   $G^2$=LR chi-squared and $df_k^{'}$ = # of regressors (not parameters)
3. BIC deviance or BIC in Raftery's notation

$$\text{BIC}^D = [D] - [df \ln N]$$

   where D is the deviance with *df* = N - (# of parameters).
4. Critically,

$$\text{BIC}_1 - \text{BIC}_2 = \text{BIC}_1^{'} - \text{BIC}_2^{'} = \text{BIC}_1^D - \text{BIC}_2^D$$

## Example: Comparing models with IC

### Adding inc-squared and dropping k618 & hc

```
. use binlfp4, clear
. logit lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog
::
. estimates store m1

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
          m1 |    753   -514.8732   -452.7237      9     923.4473    965.0639
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

. qui fitstat, ic save

. logit lfp k5      i.agecat i.wc      lwg c.inc##c.inc, nolog
::
. estimates store m2

. estat ic
::
```



```
. estimates table m1 m2, stats(N bic) b(%9.3f) t(%6.2f)

    Variable |    m1          m2
-------------+------------------------
          k5 |  -1.392      -1.385
             |   -7.25       -7.27
        k618 |  -0.066
             |   -0.96
    agecat  2 |  -0.627      -0.585
             |   -3.00       -2.87
           3 |  -1.279      -1.186
             |   -4.92       -5.08
          wc |   0.798       0.904
             |    3.48        4.36
          hc |   0.136
             |    0.66
         lwg |   0.610       0.631
             |    4.04        4.19
         inc |  -0.035      -0.065
             |   -4.24       -3.47
 c.inc#c.inc |               0.000
             |                1.88
-------------+------------------------
           N |     753         753
         bic | 965.064     956.484
--------------------------------------
                          legend: b/t
```



## `fitstat` for IC measures

1. SPost **`fitstat`** command compares BIC and AIC statistics

```
. logit lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog
. quietly fitstat, ic save

. logit lfp k5      i.agecat i.wc      lwg c.inc##c.inc, nolog
. fitstat, ic diff

                         |     Current        Saved   Difference
-------------------------+----------------------------------------
AIC                      |
                     AIC |     919.491      923.447       -3.956
           (divided by N) |       1.221        1.226       -0.005
-------------------------+----------------------------------------
BIC                      |
          BIC (df=8/9/-1) |     956.484      965.064       -8.580
  BIC (based on deviance) |   -4031.438    -4022.857       -8.580
     BIC' (based on LRX2) |     -79.887      -71.307       -8.580

Difference of    8.580 in BIC provides strong support for current model.
```

2. There is strong support for the model that adds income-squared and drops **k618** and **hc**



## Pseudo $R^2$'s

1. It would be great to have a single number to summarize model fit.
2. Such a measure would aid in comparing competing models.
   - Within a substantive area, measures of fit might provide a rough index of whether a model is adequate.
   - If prior models of LFP routinely have values of .4 for a given measure, you expect analyses with a different sample or with revised measures of the variables to have a similar value for that measure.
3. Long (1997) warns

   I am unaware of convincing evidence that selecting a model that maximizes the value of a given measure of fit results in a model that is optimal in any sense other than the model having a larger value of that measure.
4. Still, these measures are commonly used in the literature and you should use the measure that is commonly used in your field. But, do not over-interpret it!



## Summary

1. IC measures can be valuable for selecting models that are not nested
   - Do not over use these measures
   - Think about your models
2. Scalar measures of fit are sometimes required by referees, but are often of little value.



# β1 Testing marginal effects

## Readings and examples

*Long & Freese: Chapters ???*

- See references in these chapter

*Mize, Doan and Mize – forthcoming working paper*

*mco18-test-meffects-*.do a*

## From regression coefficients to marginal effects

1. Our interest is in regression coefficients to estimates predictions and estimate marginal effects.
2. Predictions

$$\text{Logit: } \widehat{\Pr}(y=1\mid \mathbf{x}) = \Lambda\left(\mathbf{x}\hat{\boldsymbol\beta}\right) = \frac{\exp\left(\mathbf{x}\hat{\boldsymbol\beta}\right)}{1+\exp\left(\mathbf{x}\hat{\boldsymbol\beta}\right)}$$

3. Discrete change

$$\frac{\Delta\widehat{\Pr}(y=1\mid \mathbf{x})}{\Delta(x_k : start \rightarrow end)} = \widehat{\Pr}(y=1\mid x_k = end, \mathbf{x}) - \widehat{\Pr}(y=1\mid x = x_k = start, \mathbf{x})$$

4. Marginal change

$$\frac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k} = f\left(\mathbf{x}\hat{\boldsymbol\beta}\right)\widehat{\beta}_k$$

5. Standard errors computed with delta method , bootstrapping, or simulation.



## Testing regression coefficients and marginal effects

1. A marginal effect depends on all parameters and the **x** where estimated:

$$\frac{\partial \Pr(y=1\mid \mathbf{x})}{\partial x_k} = f\left(\mathbf{x}\boldsymbol\beta\right)\beta_k$$

2. The size of the effect depends on:
   - All of the $\beta_j$'s, not just $\beta_k$
   - The values of the x's where the effect is evaluated
3. Does $\beta_k$=0 imply $\partial \Pr(y=1\mid\mathbf{x})/\partial x_k$ =0?
   - If you know $\beta_k$=0, then $\partial \Pr(y=1\mid\mathbf{x})/\partial x_k$ =0
   - If you accept $H_0$: $\beta_k$=0, $\partial \Pr(y=1\mid\mathbf{x})/\partial x_k$ might be 0 or might not



## Tests of $\beta_k$ and MC(wc) can give different results #1

1. Fit the logit and test $\beta_{wc}$

```
. logit lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog
::
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
::
          wc |
     college  |   .7977136   .2291814     3.48   0.001     .3485263    1.246901
::
------------------------------------------------------------------------------
```

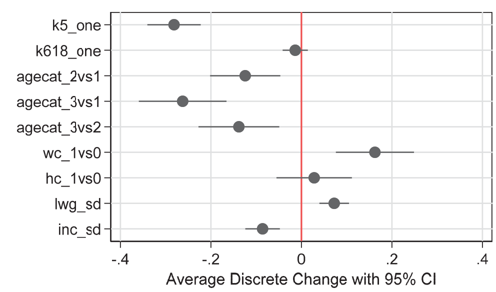
2. Compute DC(wc) for different numbers of young children

```
               |   Change    p-value
---------------+-------------------
DCR(wc | k5=0) |    0.168     0.000
DCR(wc | k5=1) |    0.182     0.001
DCR(wc | k5=2) |    0.087     0.013
DCR(wc | k5=3) |    0.027     0.085
```

3. The significance of DCR(wc) depends on the number of young children.
   - Does this make more substantive sense than saying that young children has a significant effect on LFP?



## Comparing marginal effects from the same equation

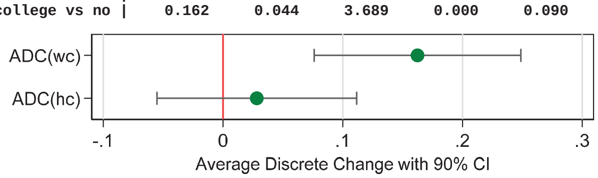



1. We can determine the size and significance of DCs
2. We can compare the size of two DCs
3. How do we test if two effects have the same size?
   - We must estimate multiple effects simultaneously



## Testing DC(hc) = DC(wc) - #2

```
               |   Change   Std Err   z-value   p-value        LL        UL
---------------+------------------------------------------------------------
hc             |
 college vs no |    0.028     0.043     0.663     0.508    -0.042     0.098
wc             |
 college vs no |    0.162     0.044     3.689     0.000     0.090     0.235
```




1. Can I conclude?

   *A woman attending college has a significantly larger effect on LFP than that of her husband attending college.*

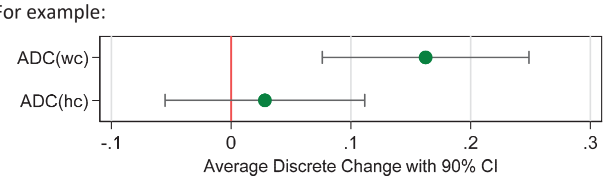


### Overlapping Confidence Intervals

1. The 90% confidence interval [ Lower level, Upper level ] can be interpreted as:

   *With repeated samples we would expect our prediction to be within the CI 90% of the time.*

2. For example:

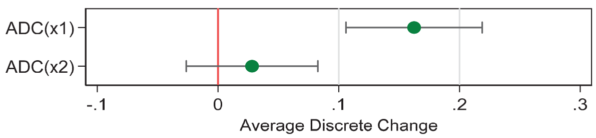



3. We conclude

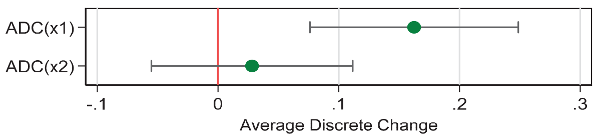
   *Our results suggest that the effect of a woman attending college could be as small as .090 or as large as .235 with 90 percent confidence. The effect of the husband's college is expected to fall between -.042 and .098.*

4. Can we conclude that DC(wc)=DC(hc)?

CIs do not over: The effects are significantly different.




CIs overlap: We cannot tell if the effects are significantly different




Conclusion

Make the formal test!



## Formally testing if MEs are equal

1. To test:

$$H_0: \Delta_1 = \Delta_2$$

2. Compute the statistics:

$$z = \frac{\hat{\Delta}_1 - \hat{\Delta}_2}{\sqrt{Var\left(\hat{\Delta}_1 - \hat{\Delta}_2\right)}}$$

3. The variance of the difference is:

$$Var\left(\hat{\Delta}_1 - \hat{\Delta}_2\right) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\sigma}_{1,2}$$

4. To estimates $\hat{\sigma}_{1,2}$ we need to simultaneously estimate the effects
   - In special cases $\hat{\sigma}_{1,2}$ is known to be 0



## Joint estimation and testing of effects - #4

1. Fit the model
2. Jointly estimate ADC(wc) and ADC(hc)

```
. margins, dydx(wc hc) post
::
dy/dx w.r.t. : 1.wc 1.hc
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          wc |
     college |   .1624037   .0440211     3.69   0.000      .076124    .2486834
             |
          hc |
     college |   .0281828    .042534     0.66   0.508    -.0551824    .1115479
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```



3. Testing if DC(wc)=DC(hc)

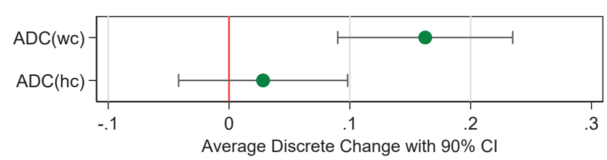
```
. test 1.wc = 1.hc
 ( 1)  1.wc - 1.hc = 0

           chi2(  1) =    3.33
         Prob > chi2 =    0.0680
```

4. We conclude:

   *The effects of the wife and the husband attending college on labor force participation are not significantly different (p>.05).*

   Or:

   *The effects of the wife and the husband attending college on labor force participation are significantly different (p<.10).*






Code: posting results from margins

1. Fit the model and store the estimates

```
logit lfp k5 k618 i.agecat i.wc i.hc lwg inc, nolog
estimates store logitmodel
```

2. Compute the effects and post the results

```
margins, dydx(wc hc) post
```

   - **post** replaced the logit estimates in memory with those from **margins**

```
. matlist e(b)
             |        0b.          1.         0b.          1.
             |        wc          wc          hc          hc
-------------+--------------------------------------------
          y1 |         0    .1624037           0    .0281828
. matlist e(V) // covariance for predictions
             |        0b.          1.         0b.          1.
             |        wc          wc          hc          hc
-------------+--------------------------------------------
       0b.wc |         0
        1.wc |         0    .0019379
       0b.hc |         0           0           0
        1.hc |         0   -.0008315           0    .0018091
```



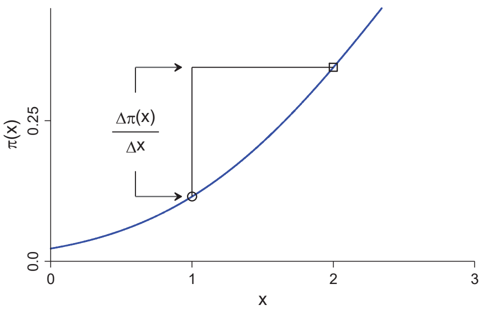
3. Test if the effects are equal

```
. test 1.wc = 1.hc
::
. lincom 1.wc - 1.hc

 ( 1)  1.wc - 1.hc = 0

             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1342209    .073553     1.82   0.068    -.0099403    .2783822

. mlincom   1 - 2, stats(all)

             |   lincom         se    zvalue    pvalue        ll        ul
-------------+----------------------------------------------------------
           1 |    0.134      0.074     1.825     0.068    -0.010     0.278
```

4. Restore the regression estimates

```
estimates restore logitmodel
```

## Comparing more complex effects




1. For DC($x_a$) compute Pr(y|$x_a$=start, **x**) and Pr(y|$x_a$=end, **x**)
2. For DC($x_b$) compute Pr(y|$x_b$=start, **x**) and Pr(y|$x_b$=end, **x**)
3. To test Ho: DC($x_a$) = DC($x_b$), estimate:

   [Pr(y|$x_a$=end, **x**)-Pr(y|$x_a$=start, **x**)] - [Pr(y|$x_b$=end, **x**)-Pr(y|$x_b$=start, **x**)]



Code: margins, at(*var*=gen(*expression*))

1. `at(`*var*`=gen(`*expression*`))`
   - predictions with *var* equal to the expression)
2. `at(x=gen(x+1))`
   - predictions at values one larger than the observed x
3. `at(x=gen(x))`
   - predictions at the observed values of x



## Testing if DC(inc+sd)=DC(lwg+sd) - #5

1. Compute standard deviations

```
. qui sum inc
. local sdinc = r(sd)
. qui sum lwg
. local sdlwg = r(sd)
```

2. Estimate four probabilities

```
. margins, at(inc=gen(inc)) at(inc=gen(inc+`sdinc')) ///
>          at(lwg=gen(lwg)) at(lwg=gen(lwg+`sdlwg')) post

Expression   : Pr(lfp), predict()

1._at        : inc             = inc
2._at        : inc             = inc+11.63479853339243
3._at        : lwg             = lwg
4._at        : lwg             = lwg+.5875564251146244

             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .5683931   .0166014    34.24   0.000      .535855    .6009312
          2  |   .4825886   .0257951    18.71   0.000     .4320312    .5331459
          3  |   .5683931   .0166014    34.24   0.000      .535855    .6009312
          4  |   .6408189   .0228361    28.06   0.000      .596061    .6855768
------------------------------------------------------------------------------
```



3. Compute DC(inc+sd)=DC(lwg+sd)

```
. qui mlincom 2-1, rowname(DCinc+sd) stats(all) clear
.     mlincom 4-3, rowname(DClwg+sd) stats(all) add

             |   lincom        se    zvalue    pvalue        ll        ul
-------------+------------------------------------------------------------
    DCinc+sd |   -0.086     0.019    -4.404     0.000    -0.124    -0.048
    DClwg+sd |    0.072     0.017     4.344     0.000     0.040     0.105
```

Confirm DCs are correct

```
. mchange inc lwg, stats(est se z p ll ul) amount(sd) width(8)

logit: Changes in Pr(y) | Number of obs = 753

Expression: Pr(lfp), predict(pr)

             |   Change   Std Err   z-value   p-value        LL        UL
-------------+------------------------------------------------------------
inc      +SD |   -0.086     0.019    -4.404     0.000    -0.124    -0.048
lwg      +SD |    0.072     0.017     4.344     0.000     0.040     0.105
```



Test that the ADCs are equal but *opposite*

```
. test (2._at-1._at)=(-1*(4._at-3._at))

 ( 1)  - 1bn._at + 2._at - 3._at + 4._at = 0

           chi2(  1) =    0.27
         Prob > chi2 =    0.6023
```

*The magnitude of the effects of income and wages are not significantly different (p=.60).*



Test equality of DCs by computing second difference

```
. mlincom (2-1)+(4-3), rowname(2nd difference) stats(all)

             |   lincom        se    zvalue    pvalue        ll        ul
-------------+------------------------------------------------------------
2nd differ~e |   -0.013     0.026    -0.521     0.602    -0.064     0.037.
```

Or:

```
. lincom (2._at-1._at)+(4._at-3._at)

 ( 1)  - 1bn._at + 2._at - 3._at + 4._at = 0

             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0133787    .025672    -0.52   0.602    -.0636949    .0369374
```

Tool: Summary of commands for comparing two DC's

```
sum x1
local Dx1 = r(sd) // change in x1; can be any value
sum x2
local Dx2 = r(sd) // change in x2; can be any value

margins, at(x1=gen(x1)) at(x1=gen(x1+`Dx1')) /// atmeans for DCM
         at(x2=gen(x2)) at(x2=gen(x2+`Dx2')) post

mlincom (2-1)-(4-3) // test of equality
mlincom (2-1)+(4-3) // test of equal magnitude
```



## Comparing ideal types and profiles - #6

1. In the lecture *Binary Regression Model* we computed predicted probabilities for these ideal types:

```
                             |    Pr(y)        ll        ul
-----------------------------+-------------------------------
              Average person |    0.578     0.539     0.616
     Younger lower educ w kids |  0.159     0.068     0.251
      Young more educ w kids |    0.394     0.234     0.554
Middle age higher educ w kids |   0.754     0.681     0.828
          Older w higher educ |   0.631     0.528     0.734
```

2. I want to say:

   *Among those with higher education, women who are middle aged with young children are no more likely to be in the labor force than older women whose children are no longer living at home.*

3. To justify this, I need to jointly estimate the probabilities.



## Estimate profiles simultaneously

```
. mtable, clear ci ///
>     at((mean) _all) ///
>     at(agecat=1 k5=2 k618=0    wc=0 hc=0 inc=10   lwg=0.75) ///
>     at(agecat=1 k5=2 k618==0   wc=1 hc=1 inc=16.6 lwg=1.62) ///
>     at(agecat=2 k5=0 k618=1.37 wc=1 hc=1 inc=27.7 lwg=1.41) ///
>     at(agecat=3 k5=0 k618==0   wc=1 hc=1 inc=27.9 lwg=1.38) ///
>     post

Expression: Pr(lfp), predict()

          |                         2.        3.        1.        1.
          |      k5      k618    agecat    agecat        wc        hc
----------+------------------------------------------------------------
        1 |    .238      1.35      .385      .219      .282      .392
        2 |       2         0         0         0         0         0
        3 |       2         0         0         0         1         1
        4 |       0      1.37         1         0         1         1
        5 |       0         0         0         1         1         1
          |
          |     lwg       inc     Pr(y)        ll        ul
----------+--------------------------------------------------
        1 |     1.1      20.1     0.578     0.539     0.616
        2 |     .75        10     0.159     0.068     0.251
        3 |    1.62      16.6     0.394     0.234     0.555
        4 |    1.41      27.7     0.754     0.681     0.827
        5 |    1.38      27.9     0.630     0.527     0.733
```



## Test if probabilities are equal

4. Estimate differences using the posted predictions:

```
. mlincom 4 - 5, rowname(MidEdDad-OldHiEd) clear twidth(20)

                     |   lincom    pvalue        ll        ul
---------------------+----------------------------------------
   MidEdDad-OldHiEd |    0.124     0.007     0.034     0.214
```

5. My initial impression was wrong and I conclude:

   *Young mothers with higher education have significantly higher chances of being in the labor force than older women with higher education who no longer have children at home (p<.01)*



## *Using the returned atspec from mtable

1. To avoid typing in values in the **`at()`** specification
   - Use the **`mtable`** return **`r(atspec)`** to save the atspec
   - Run **`mtable`** with multiple **`at()`**'s
2. See Long and Freese 2014, page 275+ for details



## Summary on testing marginal effects

1. Too often researchers use only the default tests from the estimation command
   - They test things they aren't interested in
   - They don't test things they are interested in
2. The methods above let you test many useful hypotheses
3. Remember: tests of regression coefficients and marginal effects do not always give the same result.
4. Overlapping CIs do not indicate a the estimates are equal
5. To test if MEs are equal, estimate them jointly
6. Later we extend this idea to tests across models

# β1 Nonlinearities on the RHS

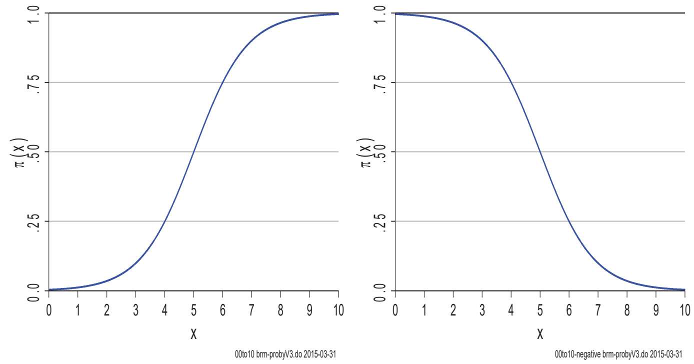
## Readings and examples

*Long & Freese: pages 301-302*

*mdo18-nonlin-*.do*

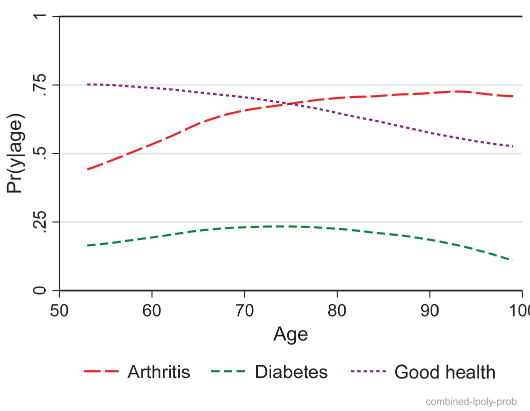


## Probabilities do not always get larger or smaller






## Real data might looks like this





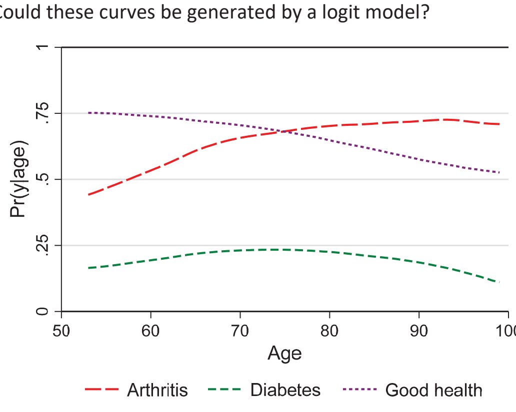

## Overview

1. Assume that $\mathbf{x}\boldsymbol{\beta}$ does not have power terms or interactions.
2. Then as $x_k$ increases, Pr(y|**x**) must *always* increases or *always* decreases.
   - The is required by the *parametric form* of the logit and probit model
3. Substantively, does this make sense?
   - Should the probability only increase or only decreases with changes in $x_k$?
   - Should the maximum probability be 1.0?
   - The minimum 0.0?
4. Nonparametric smoothers do not assume any form for the relationship between one x and the outcome
   - Lowess (`lowess`) and local polynomial smoothing (`lpoly`)
5. I often start analyses with a nonparametric fit of key regressors to the outcome
   - Here's why



## Nonparametric smoothing to assess nonlinearities

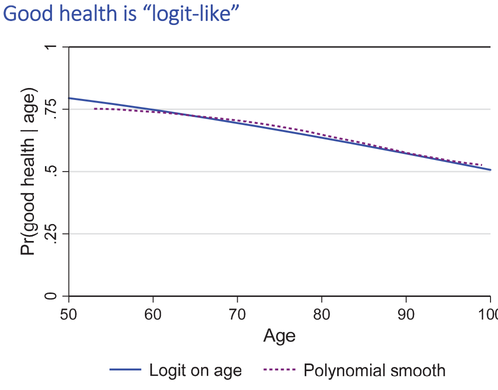
Could these curves be generated by a logit model?





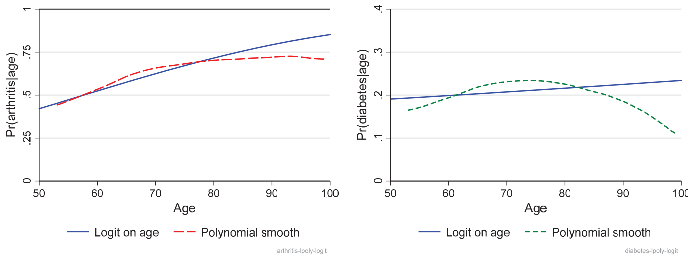

## Could these data be generated by a binary logit model? - #2

### Good health is "logit-like"



## Arthritis and diabetes are not “logit-like”

What is the substantive cost of assuming a logit-like functional form?






## Adding nonlinearities to a model

1. Consider model where x is age with other controls

$$\Pr\left(y=1\mid \mathbf{x}\right)=\Lambda\left(\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\cdots\right)$$

2. x, $x^2$ and $x^3$ are *linked* since you when x changes $x^2$ and $x^3$ must change

    If x=1, then $x^2$=1 and $x^3$=1

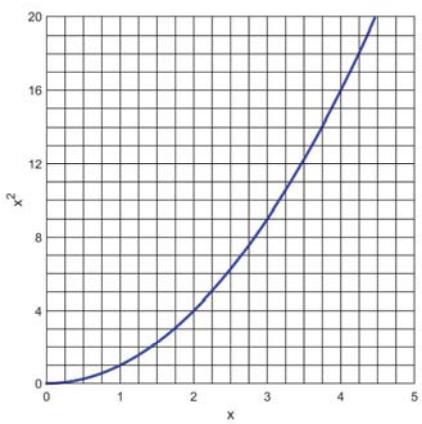
    If x=2, then $x^2$=4 and $x^3$=8

    If x=3, then $x^2$=9 and $x^3$=27

3. Polynomials on the RHS allow the probability curve to:
    - o Change directions as $x_k$ increases
        : a hill, a valley, or a snake
    - o Level off at values other than 1 or 0
4. This is how polynomials lead to nonlinearities…

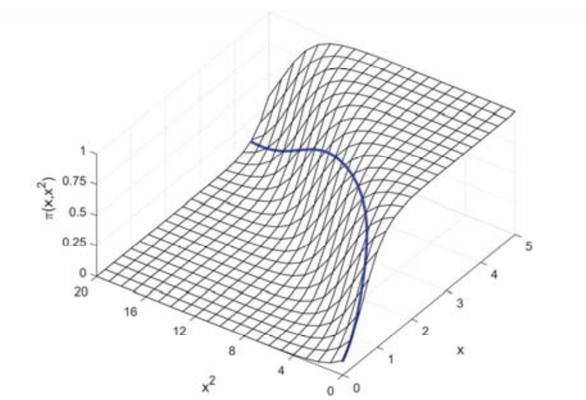


## Top view of logit with x and $x^2$



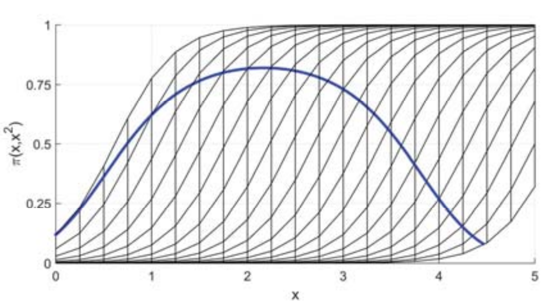



## Front view of logit with x and $x^2$






## Side view of logit with x and $x^2$






## Logit models for diabetes - #3

1. To address the nonparametric results, add age and age-squared to the model
2. To select the model
    - o AIC and BIC to compare fits
    - o Compare predictions and marginal effects

### Fit models and store estimates

```
svy: logit diabetes c.age i.female i.ed4cat, or
est store dMage1 // age

svy: logit diabetes c.age##c.age i.female i.ed4cat, or
est store dMage2 // age + age-squared

svy: logit diabetes c.age c.age#c.age c.age#c.age#c.age i.female i.ed4cat
est store dMage3 // age + age-squared + age-cubed

estimates table dMage1 dMage2 dMage3, title(diabetes) ///
    eform b(%9.5f) p(%9.3f)
```

## Logit estimates for diabetes models

The coefficients provide little insight into which model to choose

| Variable | dMage1 | dMage2 | dMage3 |
|---|---|---|---|
| female | | | |
| female | 0.80854 | 0.81816 | 0.81815 |
| | 0.000 | 0.000 | 0.000 |
| ed4cat | | | |
| 12 years | 0.66281 | 0.65679 | 0.65678 |
| | 0.000 | 0.000 | 0.000 |
| 13-15 years | 0.54123 | 0.55383 | 0.55378 |
| | 0.000 | 0.000 | 0.000 |
| 16+ years | 0.44993 | 0.45797 | 0.45794 |
| | 0.000 | 0.000 | 0.000 |
| age | 1.00656 | 1.29691 | 1.25235 |
| | 0.003 | 0.000 | 0.511 |
| c.age#c.age | | 0.99819 | 0.99869 |
| | | 0.000 | 0.784 |
| c.age#c.age#c.age | | | 1.00000 |
| | | | 0.915 |
| _cons | 0.25513 | 0.00004 | 0.00010 |
| | 0.000 | 0.000 | 0.254 |

legend: b/p



## IC measures from non-svy model fitting - #3.2

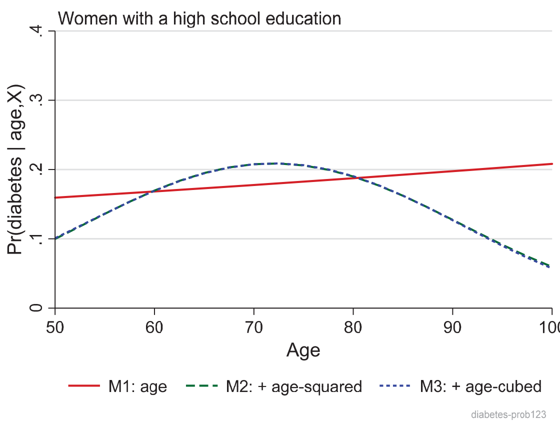
1. Since IC measures are not defined with survey estimation, models are estimate without adjusting for the complex sampling

| | nosvyM1 | nosvyM2 | nosvyM3 |
|---|---|---|---|
| bic | 17569.00 | 17458.86 | 17467.79 |
| aic | 17522.40 | 17404.50 | 17405.66 |

2. Results:
   - BIC gives M2 a 10 points advantage over M3
   - AIC gives M2 a 1 point advatntage over M3;
   - No support for M1
3. IC measures support M2



## How do the predictions compare? - #3.3






## Comparing DC(age+10) across models - #3.4

1. Does the effect of age differ across models?

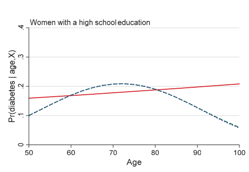
| | Change | p-value | Std Err |
|---|---|---|---|
| M1 ADC | 0.010** | 0.003 | 0.003 |
| M2 ADC | 0.004 | 0.206 | 0.003 |
| M3 ADC | 0.003 | 0.335 | 0.003 |
| M1 DCR@50 | 0.009** | 0.002 | 0.003 |
| M2 DCR@50 | 0.072*** | 0.000 | 0.007 |
| M3 DCR@50 | 0.071*** | 0.000 | 0.017 |
| M1 DCR@70 | 0.010** | 0.004 | 0.003 |
| M2 DCR@70 | -0.018*** | 0.000 | 0.005 |
| M3 DCR@70 | -0.018* | 0.030 | 0.008 |
| M1 DCR@90 | 0.011** | 0.006 | 0.004 |
| M2 DCR@90 | -0.070*** | 0.000 | 0.004 |
| M3 DCR@90 | -0.072*** | 0.000 | 0.014 |

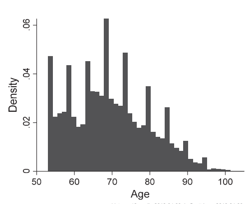
*≤.05; **≤.01; ***≤.001

2. Which model would you choose? Why is ADC misleading?

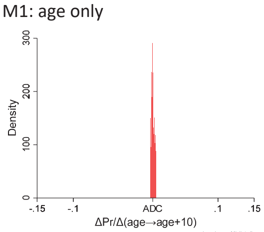


## Why ADC can be misleading

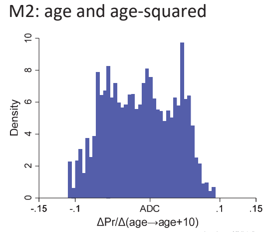



M1: age only




M2: age and age-squared






# Logit models for arthritis

## Logit estimates for arthritis models - #4.1

| Variable | aMage1 | aMage2 | aMage3 |
|---|---|---|---|
| female | 1.77543 | 1.80948 | 1.81087 |
| | 0.000 | 0.000 | 0.000 |
| ed4cat | | | |
| 12 years | 0.82788 | 0.82101 | 0.82109 |
| | 0.003 | 0.002 | 0.002 |
| 13-15 years | 0.77455 | 0.79218 | 0.79310 |
| | 0.000 | 0.001 | 0.001 |
| 16+ years | 0.52825 | 0.53507 | 0.53543 |
| | 0.000 | 0.000 | 0.000 |
| age | 1.04844 | 1.35998 | 2.28835 |
| | 0.000 | 0.000 | 0.002 |
| c.age#c.age | | 0.99813 | 0.99076 |
| | | 0.000 | 0.014 |
| c.age#c.age#c.age | | | 1.00003 |
| | | | 0.043 |
| _cons | 0.05711 | 0.00001 | 0.00000 |
| | 0.000 | 0.000 | 0.000 |

legend: b/p

## Choosing a model

### What does substantive research tell you?

Does Pr(arthritis | age)=1.0 make sense?

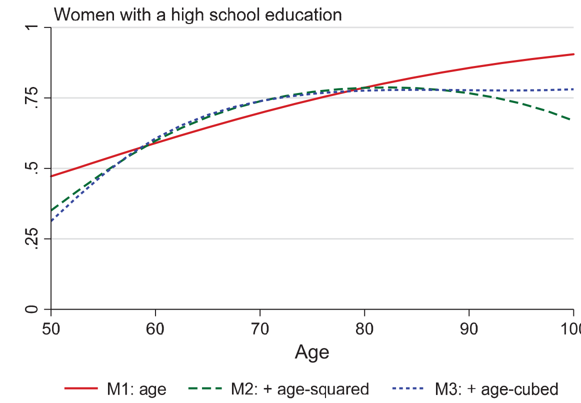
### IC measures from <u>non-svy</u> model fitting

```
             |  nosvyM1      nosvyM2      nosvyM3
-------------+------------------------------------
         bic |  22094.79     21909.34     21914.01
         aic |  22048.19     21854.98     21851.89
```

- BIC which prefers simpler models, points to M2
- AIC which allows more complexity, points to M3



### How do the predictions compare?

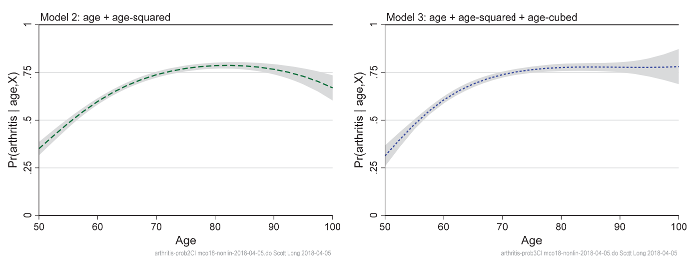



The main differences between M2 and M3 occur beyond 90.



### Confidence intervals for predictions

The biggest differences occur where there is the least precision




- The CIs beyond 90 overlap
- Tools in *Comparing Marginal Effects* let you test if they are different



## What is the effect of age on arthritis?

1. Does the model affect the effect of age?
2. Which model would you choose? Is it time to consult a rheumatologist?

```
           |   Change   p-value   Std Err
---------- +-----------------------------
M1 ADC     |   0.101***  0.000     0.004
M2 ADC     |   0.092***  0.000     0.004
M3 ADC     |   0.099***  0.000     0.005

M1 DCR@50  |   0.116***  0.000     0.005
M2 DCR@50  |   0.236***  0.000     0.012
M3 DCR@50  |   0.278***  0.000     0.025

M1 DCR@70  |   0.104***  0.000     0.004
M2 DCR@70  |   0.056***  0.000     0.005
M3 DCR@70  |   0.044***  0.000     0.009

M1 DCR@90  |   0.063***  0.000     0.001
M2 DCR@90  |  -0.111***  0.000     0.022
M3 DCR@90  |   0.004     0.932     0.047
---------- +---------------------------------
*≤.05; **≤.01; ***≤.001
```



## Code

### <u>Local polynomials</u>

```
lpoly diabetes age if age<100, gen(d_age d_poly) nograph n(200) bwidth(5)
label var d_poly "Diabetes"
```

### <u>IC measures</u>

```
qui {
    logit diabetes age i.female i.ed4cat, or
    est store nosvyM1
    logit diabetes c.age##c.age i.female i.ed4cat, or
    est store nosvyM2
    logit diabetes c.age c.age#c.age c.age#c.age#c.age i.female i.ed4cat
    est store nosvyM3
}
estimates table nosvyaMage1 nosvyaMage2 nosvyaMage3, ///
    stats(bic aic) keep(age c.age#c.age c.age#c.age#c.age) ///
    b(%9.5f) p(%9.3f) stfmt(%9.2f)
```

### <u>Predictions for probability plots</u>

```
est restore dMage1
mgen, at(age=(50(2.5)100) female=1 ed4cat=2) ///
    atmeans stub(dM1) replace
```

### <u>Plot command with CI</u>

```
est restore aMage3
```



```
mgen, at(age=(50(2.5)100) female=1 ed4cat=2) /// predictions for plot
    atmeans stub(aM3) replace
local graphname "arthritis-prob3CI"
graph twoway ///
    (rarea aM3ul aM3ll1 aM1age, color(gs13) lw(none)) /// shaded CI
    (connected aM3pr1 aM1age, $M3line ) , ///
    title("Model 3: age + age-squared + age-cubed", position(11)
size(*.8)) ///
    xtitle("Age") xlab(50(10)100) ///
    $ytitlea  $ylab  yline(0 1, lcol(black)) ///
    legend(off) $nogapnoline scale(1.1) ///
    caption("`graphname' `tag'", $captionopt)
graph export `pgm'-`graphname'.$graphfmt, replace
```

### <u>Effects of age</u>

```
estimates restore dMage1
mchange age, amount(delta) delta(10) stats(est se p)
mchange age, amount(delta) delta(10) stats(est se p) atmeans at(age=50)
mchange age, amount(delta) delta(10) stats(est se p) atmeans at(age=70)
mchange age, amount(delta) delta(10) stats(est se p) atmeans at(age=90)

estimates restore dMage2
::
```

<u>*How would you test if the effects differ across models?*</u>

Try to figure this out after *Comparing Marginal Effects*

## Summary of nonlinearities on the RHS

1. Always consider nonlinearities on the RHS
   - What are your substantive expectations?
   - Do not let the functional form of logit/probit dictate what you find
2. Nonlinearities on the RHS can create models where
   - Predictions do not plateau at 1
   - Predictions do not uniformly increase or decrease
   - Predictions are more linear or less linear the a “linear” logit
3. Starting with a nonparametric plot is often valuable
4. Compare the substantive implications of the model