CDA ICPSR Lab Guide | 2017

Scott Long and Bianca Manago

Table of Contents

1 Introduction

The lab guide presents tools and exercises for categorical data analysis that correspond to the lectures. Use the guide as a starting point, then use the lecture notes for planning your more detailed analyses using sometimes more sophisticated and efficient methods. If you are unfamiliar with Stata or would just like a quick review, please refer to *Getting Started Using Stata*.

1. The guide is divided into parts corresponding to lectures. Each part includes a **review** which everyone should complete and an **exercise** you might want to work on after class. Do-files for the reviews have "review" in the name (e.g., cda16lab-brm-review.do). Do-files for exercises have "exercise" in the name (e.g., cda16lab-brm-exercise.do).
2. In the guide, Stata commands and output are in this `font`. In output, commands are preceded by "`.`" and "`>`". Do not include "`.`" or "`>`" in the do-file you write. Interpretation of results are shown in boxes. You should write your own interpretations. If you want feedback on your interpretation, write a paragraph and show it to the instructor along with your log-file.
3. The datasets cda-nes4.dta, cda-science4.dta, cda-hsb4.dta, cda-addhealth4.dta, and cda-hrs4.dta are available for the exercises. Codebooks are at the end of this guide, although cda-hrs4.dta does not have a codebook since it has so few variables. For quick analyses where you do not have to worry about cleaning the data, I suggest using cda-science4.dta.
4. Exercises should be completed using do-files. If you are not sure how to use a do-file see the *Getting Started with Stata Guide* for help.

How much should you do?

1. The workshop only lasts four days and you will be given a **lot** of new material. You should not expect to do all of the exercises. I suggest you start with the review do-files. For example, look at cda16lab-orm-review.do and run it. Change some of the commands. Add others. Run it again.
2. When you get home, apply the commands in the review files to your own data or explore doing other things with the sample datasets. At home you can look more closely at the do-files that produce the examples in the notes.

2 Using Stata

2.1 The working directory and sample files

Stata opens in a default working directory. You can find out what this is by looking in the lower left hand corner of the program or by typing **`cd`**.

Suppose that your working directory is on the C: drive. I suggest that you:

1. Create a directory C:\ICPSR\ using your file manager.
2. Copy the sample do-files and data for the class to that directory.
3. Each time you open Stata, change your working directory: cd c:\icpsr

If you are an experienced Stata user, feel free to do this some other way.

2.2 Using do-files

I want you to use do-file in lab to ensure that you are able to replicate your work. Here is an example of a do-file:

```
01     capture log close
02     log using jslong-a03lrm01, replace text
03     version 14.1
04     clear all
05     macro drop _all
06     set linesize 80
07     set scheme s1manual

08     //  Assignment 03: LRM
09     //  CDAicpsr 2017
10     local pgm jslong-a03lrm01
11     local who jslong Scott Long
12     local dte 2017-05-04
13      local graphtype emf // for mac use png or pdf

14     //  #1 load data

15     use dataset, clear

16     //  #2 task 2

17     //  #3 task 3

18     log close
19     exit
```

Lines 1-2 set up a log file where output from your do file will be saved. Line 1 ensures there are no log files that are already open. Line 2 opens the new log file. The name of the log files should match the name of the do-file. Line 18 closes the log file so additional results are not saved to the file. If you do not add a return after line 18, line 18 won't run. Line 19 makes sure that line 18 runs and tells Stata to ignore any later lines (you can put ideas, notes, etc. after line 19.

Lines 3-7 make sure that your results are not dependent on something you left in memory and accordingly makes it possible for your do-file to run later (i.e., it is needed for reproducibility). The scheme will make your graphs look the same on different computer.

Lines 8 to 12 document what you are doing. Line 8 is what the do-file is for, while line 9 is something you should include for all do-files in this class. Lines 10-12 are the what, who and when information which is useful later when looking at the output. Always update this information in a new do-file.

Line 13 indicates the type of graph you want to create. PNG files do not print well; EMF files do, but do not work with OS X.

Lines 8 and 9 are are comments. They do not run any Stata commands, but simply let you add notes. You can comment out single lines of text with an asterisk (`*`) or a double slash (`//`), or create blocks of comments starting with a `/*` and ending with `*/`.

The commands for analysis begin on line 14. To write legible do files, organize the content to make it easier to locate later. Grouping like commands (e.g., creating demographic variables, estimating nested models for one outcome) keeps the file orderly.

3 Continuous Outcomes

The commands from this section are in **`cda16lab-lrm-review.do`**.

#1 Load the Data

```
use cda-scireview4, clear
```

#2 and #3 Examine the Data and Select Variables

Begin by using the command **`codebook, compact`** to list variables, their labels, and summary statistics.

```
. codebook, compact

Variable   Obs Unique      Mean    Min    Max  Label
-------------------------------------------------------------------------------
id         264    264  58556.74  57001  62420  ID Number.
cit1       264     48  11.33333      0    130  Citations: PhD yr -1 to 1.
cit3       264     54  14.68561      0    196  Citations: PhD yr 1 to 3.
cit6       264     59  17.58712      0    143  Citations: PhD yr 4 to 6.
cit9       264     67  19.92803      0    214  Citations: PhD yr 7 to  9.
enroll     264      9  5.530303      3     14  Years from BA to PhD.
fel        264     96  3.191098      1   4.69  Fellow or PhD prestige.
<snip>
workuniv   264      2  .7045455      0      1  University work? (1=yes)
-------------------------------------------------------------------------------
```

Next, use **`keep`** to select the dependent variable **`pubtot`** and the three independent variables, **`workfac`**, **`enrol`**, and **`phd`**, which we use in the regression models later.

```
. keep pubtot workfac enrol phd
```

Section #3 in the do-file explores missing data.

#4 Regression

Specifying a model is simple, with the dependent variable listed *first* followed by independent variables. Prefacing an independent variable with **`i.`** indicates that it is a factor variable (i.e., a binary or categorical variable). By default, the category with the lowest value (in this case workfac=0) is the reference category. Prefacing a variable with **`c.`** indicates that a variable is continuous. If no prefix is specified, Stata assumes the variable is continuous (unless it is included in an interaction).

```
. regress pubtot i.workfac c.enrol c.phd

      Source |       SS           df       MS      Number of obs   =       264
-------------+----------------------------------   F(  3,    260)  =     10.77
       Model |  3519.43579         3  1173.14526   Prob > F        =    0.0000
    Residual |  28326.1968       260  108.946911   R-squared       =    0.1105
-------------+----------------------------------   Adj R-squared   =    0.1003
       Total |  31845.6326       263  121.086055   Root MSE        =    10.438

------------------------------------------------------------------------------
      pubtot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     workfac |
      1_Yes  |   5.227261   1.297375     4.03   0.000     2.672561     7.78196
      enroll |  -1.174879   .4465778    -2.63   0.009    -2.054249   -.2955094
         phd |   1.506904   .6442493     2.34   0.020     .2382931    2.775514
       _cons |   9.982767    3.33341     2.99   0.003     3.418849    16.54668
------------------------------------------------------------------------------
```


#5 Standardized Coefficients

`listcoef` displays the estimated coefficients along with standardized coefficients. The **`help`** option provides details on the meaning of each coefficient.

```
. listcoef, help

regress (N=264): Unstandardized and standardized estimates

  Observed SD: 11.0039
  SD of error: 10.4378

-------------------------------------------------------------------------------
             |        b        t    P>|t|    bStdX    bStdY   bStdXY     SDofX
-------------+-----------------------------------------------------------------
     workfac |
      1_Yes  |   5.2273    4.029    0.000    2.613    0.475    0.237     0.500
      enroll |  -1.1749   -2.631    0.009   -1.696   -0.107   -0.154     1.443
         phd |   1.5069    2.339    0.020    1.515    0.137    0.138     1.005
    constant |   9.9828    2.995    0.003        .        .        .         .
-------------------------------------------------------------------------------
       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
```

> For a unit increase in the prestige of the doctoral department, the total number of publications is expected to increase by 1.5, holding other variables constant (p<0.05, two-tailed test).
>
> For a standard deviation increase in the length of time between enrollment and graduation, about 1.5 years, the number of publication is expected to decrease by 1.7, holding other variables constant (p<0.01, two-tailed test).
>
> On average, scientists who take faculty positions have about a half a standard deviation more publications than scientists who do not take faculty positions (p<0.001, two-tailed test).

#6 Close log and exit program

```
log close
exit
```

We won't show this step in the rest of the guide. But, you always want to include this in your do-file.

4 Binary Outcomes

The commands for this section are in **`cda16lab-brm-review.do`**.

#1 Load the data

```
use cda-scireview4, clear
```

#2 Examine data, select variables, and verify

```
keep workfac fellow phd mcit3 mnas
```



```
tab1 fellow mnas workfac, miss // option miss includes missing values in table
codebook, compact
```

#3 Binary logit model

The dependent variable is listed first. A probit model is run by changing **`logit`** to **`probit`**.

```
. logit workfac i.fellow c.phd c.mcit3 i.mnas, nolog

Logistic regression                             Number of obs     =        264
                                                LR chi2(4)        =      37.64
                                                Prob > chi2       =     0.0000
Log likelihood = -163.55534                     Pseudo R2         =     0.1032

------------------------------------------------------------------------------
     workfac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      fellow |
      1_Yes  |   1.250155   .2767966     4.52   0.000     .7076434    1.792666
         phd |  -.0637186   .1471307    -0.43   0.665    -.3520894    .2246522
       mcit3 |   .0206156   .0071255     2.89   0.004     .0066498    .0345814
             |
        mnas |
      1_Yes  |   .3639082   .5571229     0.65   0.514    -.7280327    1.455849
       _cons |  -.5806031   .4498847    -1.29   0.197    -1.462361    .3011547
------------------------------------------------------------------------------
```

#4 Store the estimation results

It is sometimes necessary to store estimation results to restore later (e.g., when posting with **`margins`**). You do this using **`estimates store`**. Here we store the estimates with the name **`estlogit`**.

```
estimates store estlogit
```

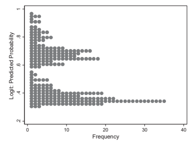
#5 Predicted probabilities for each observation

We can compute and plot predicted probabilities for each observations. We pick the name **`prlogit`** for the new variable that contains predictions.

```
. predict prlogit
(option pr assumed; Pr(workfac))

. label var prlogit "Logit: Predicted Probability"

. sum prlogit

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     prlogit |       264    .5340909    .1828654   .3035647   .9665072
```

The **`dotplot`** command is used create a histogram of the predicted probabilities:

```
. dotplot prlogit
. graph export `pgm'-phatdist.`graphtype', replace
```



#6 Predict specific probabilities

mtable computes a predictions and saves them in a table. Here we focus on the probability of our dependent variable for given values of the independent variables. The **at()** option sets the values where predictions are made. The **atmeans** option sets the other independent variables at their means.

- **predict** creates a new variable that contains predictions for each case in the sample.
- **mtable** computes predictions at specified values of the regressors and does not create a new variable.

We predict the probability of working as a faculty member for someone who has a postdoctoral fellowship and whose mentor was a member of the National Academy of sciences with other regressors held at their means:

```
. mtable, at(fellow=1 mnas=1) stat(ci) atmeans

Expression: Pr(workfac), predict()

   Pr(y)        ll        ul
-----------------------------
   0.779     0.593     0.964

Specified values of covariates

          |   fellow       phd     mcit3      mnas
----------+----------------------------------------
  Current |        1      3.18      20.7         1
```

> The predicted probability of obtaining a faculty position is 0.78 (95% CI: 0.59, 0.96) for an average scientist who began his career with a postdoctoral fellow after studying with a mentor who is in the National Academies of Sciences.

#7 Table of probabilities

mtable can make a table of predicted probabilities for combinations of values of independent variables.

```
. mtable, at(fellow=(0 1) mnas=(0 1)) stat(ci) atmeans

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



```
          |   fellow      mnas     Pr(y)        ll        ul
----------+--------------------------------------------------
        1 |        0         0     0.412     0.330     0.494
        2 |        0         1     0.502     0.232     0.771
        3 |        1         0     0.710     0.619     0.801
        4 |        1         1     0.779     0.593     0.964

Specified values of covariates

          |      phd     mcit3
----------+-------------------
  Current |     3.18      20.7
```

The same predictions can be obtained using **margins** which produces more output. The SPost **m*** commands are "wrappers" that make it easier to work with **margins**.

```
. margins, at(fellow=(0 1) mnas=(0 1)) atmeans

Adjusted predictions                            Number of obs   =        264
Model VCE    : OIM

Expression   : Pr(workfac), predict()

1._at        : fellow          =           0
               phd             =    3.181894 (mean)
               mcit3           =    20.71591 (mean)
               mnas            =           0

2._at        : fellow          =           0
               phd             =    3.181894 (mean)
               mcit3           =    20.71591 (mean)
               mnas            =           1

3._at        : fellow          =           1
               phd             =    3.181894 (mean)
               mcit3           =    20.71591 (mean)
               mnas            =           0

4._at        : fellow          =           1
               phd             =    3.181894 (mean)
               mcit3           =    20.71591 (mean)
               mnas            =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .4118608   .0417942     9.85   0.000     .3299457    .4937759
          2  |   .5019075   .1374789     3.65   0.000     .2324539    .7713612
          3  |   .7096895    .046453    15.28   0.000     .6186433    .8007358
          4  |   .7786445   .0946714     8.22   0.000     .5930921     .964197
------------------------------------------------------------------------------
```

#8 Discrete change at means with mlincom

mtable with the **post** option can be used to compute discrete changes. First, **mtable** computes the probabilities at the start and end values of the discrete change. With the **post** the predictions are left in memory for **mlincom** to use.

```
. mtable, at(fellow=(0 1)) atmeans post

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



```
          |   fellow     Pr(y)
----------+-------------------
        1 |        0     0.419
        2 |        1     0.716

Specified values of covariates

          |                          1.
          |      phd     mcit3      mnas
----------+-----------------------------
  Current |     3.18      20.7     .0833
```

mlincom computes the change in probability, that is, the discrete change. The numbers after **mlincom** refer to the numbered rows from **mtable** (e.g., row 2 minus row 1):

```
. mlincom 2-1, stats(all)

             |   lincom        se    zvalue    pvalue        ll        ul
-------------+------------------------------------------------------------
           1 |    0.297     0.061     4.888     0.000     0.178     0.416
```

> A scientist who receives a post-doctoral fellowship has a .30 higher probability of being on the faculty at a university than a scientist who does not receive a fellowship, holding other variables at their means (p<0.001, two-tailed test).

#9 Discrete change at means using dydx()

Restoring estimates: After using **mtable** or **margins** with the **post** option, the logit estimates are no longer in memory since they have been replaced by the estimates from **margins**. To put the logit results back in memory (which is necessary for computing more predictions), we use **estimate restore**.

```
. estimates restore estlogit
(results estlogit are active now)
```

Using dydx(): Now we can compute additional predictions using these estimates. The results from the example using **mlincom** can be duplicated using the **dydx()** option with **mtable**. For variables with an **i.** prefix, **dydx()** computes a change from 0 to 1. For variables with a **c.** prefix or no prefix, **dydx()** computes the marginal change. Be careful since it is easy to compute incorrect results if you did not correctly specify the prefix for the independent variables in your regression model. Here we compute the discrete change for the variable fellow, which match the results above.

```
. mtable, dydx(fellow) atmeans stat(ci p)

Expression: Pr(workfac), predict()

 d Pr(y)        ll        ul         p
---------------------------------------
   0.297     0.178     0.416     0.000

Specified values of covariates

          |       1.                          1.
          |   fellow       phd     mcit3      mnas
----------+----------------------------------------
  Current |     .413      3.18      20.7     .0833
```


#10 <u>Average</u> discrete change with mchange

mchange computes the discrete change for some or all independent variables. Independent variables can be held at specific values using **at()** or at the means with **atmeans**. By default, however, the <u>average</u> discrete change is computed along with the p-value for a test that the marginal effect is 0.

```
. mchange

logit: Changes in Pr(y) | Number of obs = 264

Expression: Pr(workfac), predict(pr)

               |    Change    p-value
---------------+---------------------
fellow         |
 1 Yes vs 0 No |     0.285      0.000
phd            |
            +1 |    -0.014      0.665
           +SD |    -0.014      0.665
      Marginal |    -0.014      0.665
mcit3          |
            +1 |     0.004      0.002
           +SD |     0.111      0.002
      Marginal |     0.004      0.002
mnas           |
 1 Yes vs 0 No |     0.078      0.509

Average predictions

             |      0_No      1_Yes
-------------+---------------------
  Pr(y|base) |     0.466      0.534
```

The discrete change for fellow is different than before since **mchange** is computing the Average Marginal Effect (AME), whereas the first two discrete changes computed the Marginal Effect at the Mean (MEM). In the following interpretations, note the subtle yet crucial difference in wording for a discrete change computed using AME versus the wording of the earlier discrete change using MEM.

> On average, having a post-doctoral fellowship increases the probability of being faculty at a university by .29 (p<0.001, two-tailed test).
>
> On average, a standard deviation increase in the mentor's citations, about 25 citations, is expected to increase the probability of being a faculty member by 0.11 (p<0.01, two-tailed test).

#11 Plotting predicted probabilities

You might want to compute predicted probabilities across the range of a continuous variable for each of two groups and then plot these. **mgen** generates new variables containing predicted values and confidence intervals. These variables begin with the stem specified with **stub()**. The **predlabel()** option allows you to name what is being predicted.

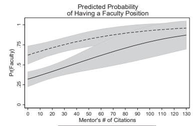
```
. mgen, at(fellow=1 mcit3=(0(5)130)) atmeans stub(fel1) predlabel(Fellow)

Predictions from: margins, at(fellow=1 mcit3=(0(5)130)) atmeans predict(pr)

Variable   Obs Unique      Mean       Min       Max  Label
--------------------------------------------------------------------------------
fel1pr1     27     27  .8361422   .621785  .9599656  Fellow
```



```
fel1ll1     27     27   .748555  .5078947  .8969149  95% lower limit
fel1ul1     27     27  .9237294  .7356753  1.023016  95% upper limit
fel1mcit3   27     27        65         0       130  Mentor's 3 yr citation.
--------------------------------------------------------------------------------

Specified values of covariates

                              1.
   fellow        phd        mnas
-------------------------------
        1   3.181894    .0833333

<snip>
```

After creating the variables with **mgen**, the following commands create the graph.

```
. graph twoway ///
>     (rarea fel1ul fel1ll fel1mcit3, col(gs10)) ///
>     (rarea fel0ul fel0ll fel1mcit3, col(gs10)) ///
>     (connected fel1pr fel1mcit, lpat(dash)  msym(i)) ///
>     (connected fel0pr fel1mcit, lpat(solid) msym(i)), ///
>     legend(on order(3 4)) ylab(0(.25)1) ytitle("Pr(Faculty)") ///
>     xlab(0(10)130) xtitle("Mentor's # of Citations") ///
>     title("Predicted Probability of Having a Faculty Position")
. graph export `pgm'-probplot.`graphtype', replace
```



> For an average scientist, receiving a fellowship increases the probability of being employed as a faculty member. When their mentors have not been cited, fellows have an advantage over non-fellows of nearly .30 and that advantage decreases gradually to about .10 for those with highly cited mentor.

You **cannot** use overlapping confidence intervals to determine if the differences in probabilities for fellows and non-fellows are significant. For this, you need to compute discrete changes.

#12 Computing Odds Ratios

The factor change in the odds and the standardized factor change are obtained with **listcoef**. **listcoef** can run after a probit model where it will compute standardized beta coefficients instead.

```
. listcoef, help

logit (N=264): Factor change in odds
```



```
 Odds of: 1_Yes vs 0_No

--------------------------------------------------------------------------
             |         b         z     P>|z|       e^b   e^bStdX     SDofX
-------------+------------------------------------------------------------
      fellow |
      1_Yes  |    1.2502     4.517     0.000     3.491     1.853     0.493
         phd |   -0.0637    -0.433     0.665     0.938     0.938     1.005
       mcit3 |    0.0206     2.893     0.004     1.021     1.690    25.445
             |
        mnas |
      1_Yes  |    0.3639     0.653     0.514     1.439     1.106     0.277
    constant |   -0.5806    -1.291     0.197         .         .         .
--------------------------------------------------------------------------
       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
   SDofX = standard deviation of X
```

> Obtaining a post-doctoral fellowship increases the odds of obtaining a faculty position by a factor of 3.5, holding other variables constant (p<0.001, two-tailed test).
>
> A standard deviation increase in mentor's citations, about 25, increases the odds of a faculty position by a factor of 1.7 (p<0.01, two-tailed test).

#13 Comparing Coefficients from Logit and Probit

Here we run a probit model using the same variables and store the results. We use **estimates table** to list the logit and probit estimates side-by-side. The logit estimates are around 1.7 times as large as the probit estimates. Why is this?

```
. probit workfac i.fellow c.phd c.mcit3 i.mnas, nolog
<SNIP>
. estimates store estprobit

. estimates table estlogit estprobit, b(%7.2f) t(%7.2f) stats(N) modelwidth(10)

----------------------------------------
    Variable |  estlogit    estprobit
-------------+--------------------------
      fellow |
      1_Yes  |      1.25         0.76
             |      4.52         4.56
             |
         phd |     -0.06        -0.04
             |     -0.43        -0.44
       mcit3 |      0.02         0.01
             |      2.89         2.97
             |
        mnas |
      1_Yes  |      0.36         0.23
             |      0.65         0.71
             |
       _cons |     -0.58        -0.35
             |     -1.29        -1.26
-------------+--------------------------
           N |       264          264
```

```
----------------------------------------
                            legend: b/t
```

5 Binary Outcomes: Advanced Post-estimation

The file **`cda16-lab-brm-advanced-review.do`** contains these Stata commands.

#1 Load the Data

```
use cda-scireview4, clear
```

#2 Examine data, select variables, and verify

```
keep workfac fellow phd mcit3 mnas
tab1 fellow mnas workfac, miss
codebook, compact
```

#3 Binary logit model

The same model is estimated as before.

```
logit workfac i.fellow c.phd c.mcit3 i.mnas, nolog
```

#4 Store the estimation results

It is sometimes necessary to store estimation results to restore later (e.g., when posting with **`margins`**). You do this using **`estimates store`**. Here we store the estimates with the name **`estlogit`**.

```
estimates store estlogit
```

#5 Use over() to compute local means

Using the **`over(`*`varlist`*`)`** options with **`atmeans`** computes local means for each combination of categorical variables listed in *varlist*. Only categorical variables can be included in the **`over()`** option. If **`atmeans`** is not specified, probabilities are computed as observed separately for the combination of categorical variables in *varlist*.

```
. mtable, over(fellow) atmeans

Expression: Pr(workfac), predict()

          |                                     1.
          |   fellow       phd     mcit3      mnas     Pr(y)
----------+--------------------------------------------------
        1 |        0      3.05      18.3     .0516     0.406
        2 |        1      3.37      24.2      .128     0.731

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

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

#6 Using if statements to obtain the same result

You can compute predictions using local means or observed values by using the **`if`** statement. This procedure is flexible and you can specify multiple variables in the **`if`** statement. The results below correspond to those above.


```
. quietly mtable if fellow == 0 , atmeans atvars(1.fellow phd mcit3 1.mnas) brief
. mtable if fellow == 1 , atmeans atvars(1.fellow phd mcit3 1.mnas) brief below

Expression: Pr(workfac), predict()

          |       1.                              1.
          |   fellow       phd     mcit3      mnas     Pr(y)
----------+--------------------------------------------------
        1 |        0      3.05      18.3     .0516     0.406
        1 |        1      3.37      24.2      .128     0.731
```

#7 Testing differences between ideal types

We provide the basic logic here of testing differences between ideal types by using **`mtable`** and multiple **`at()`** specifications. The same logic can be used in testing differences between local and global means, or just about anything! Refer to the lecture do-files for an exact and robust method of computing differences between groups using macros and multiple **`at()`** statements.

Including both multiple **`at()`** statements in a single **`mtable`** command and the **`post`** option computes the predicted probabilities of the different groups.

```
. mtable, at(fellow=0 mnas=0 phd=2 mcit=0) ///
>         at(fellow=1 mnas=1 phd=4 mcit=75) post

Expression: Pr(workfac), predict()

          |   fellow       phd     mcit3      mnas     Pr(y)
----------+--------------------------------------------------
        1 |        0         2         0         0     0.330
        2 |        1         4        75         1     0.911

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

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

Now **`mlincom`** is used to test the difference between these two ideal types.

```
. mlincom 2-1

             |   lincom    pvalue        ll        ul
-------------+----------------------------------------
           1 |    0.581     0.000     0.427     0.735
```

The probability of being a faculty member is 0.60 higher for scientists from elite backgrounds with successful mentors than those from adequate backgrounds (p<0.001).

#8 Computing marginal effects in subgroups

The **`if`** qualifier can be used with **`mchange`** as well to compute the marginal effects of variables. Below we compute the average marginal effect of **`mcit3`** for those receiving post-doctorate fellowships and whose mentors were members of the NAS.

```
. estimates restore estlogit
(results estlogit are active now)

. mchange mcit3 if fellow==1 & mnas==1 , amount(range) trim(10)

logit: Changes in Pr(y) | Number of obs = 14
```



```
Expression: Pr(workfac), predict(pr)

             |    Change   p-value
-------------+----------------------
mcit3        |
  10% to 90% |     0.240     0.024

Average predictions

             |      0_No     1_Yes
-------------+----------------------
  Pr(y|base) |     0.185     0.815

1: Sample selection: if fellow==1 & mnas==1  & e(sample)==1
```

We computed the change in **`mcit3`** over the trimmed range from the 10th to 90th percentiles. To make our interpretation understandable, we use **`centile`** with an **`if`** qualifier to obtain the number of citations at these locations.

```
. centile mcit3 if fellow==1 & mnas==1, centile(10 90)

                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
       mcit3 |        14          10          2.5               2    25.22268*
             |                    90           92        43.71093         127*

* Lower (upper) confidence limit held at minimum (maximum) of sample
```

On average, for scientists with postdoctoral fellowships and mentors who were members of the NAS, increasing the number of their mentor's citations from 2 to 92 increases the probability of having a faculty job by 0.24 (p<0.05).

#9 Second differences

Second differences can be computed by combing the **`dydx()`** and **`over()`** options. This computes the discrete change of the variable specified with **`dydx()`** restricting the sample to cases selected by the **`over()`** option. The **`post`** option saves results which allows **`mlincom`** to compute the second difference. First we estimate a logit model with an interaction between the two variables used for the second difference.

```
. logit workfac i.fellow##i.mnas c.phd c.mcit3, nolog
<snip>
```

Next, **`mtable`** computes the discrete change of **`fellow`** across each category of **`mnas`**.

```
. mtable, dydx(fellow) over(mnas) stat(ci) post

Expression: Pr(workfac), predict()

          |  d Pr(y)        ll        ul
----------+-----------------------------
     0 No |    0.296     0.175     0.417
    1 Yes |    0.158    -0.233     0.549

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

          |  No at()
----------+---------
```



```
  Current |       .n
```

We use **`mlincom`** to test if the discrete change of **`fellow`** is significantly different between categories of **`mnas`**.

```
. mlincom 1-2

             |   lincom    pvalue        ll        ul
-------------+----------------------------------------
           1 |    0.138     0.509    -0.272     0.548
```

Although the effect of having a postdoctoral fellowship is estimated to be 0.14 higher for scientists whose mentor was in the National Academy of Science, this difference is not statistically significant (p>0.10).

#10 Graphing discrete changes

In the last section we noted that overlapping confidence intervals do not necessarily indicate a lack of statistical significance. We now show how to graph discrete changes between categorical variables. First, we reproduce the graph from section 3.11.

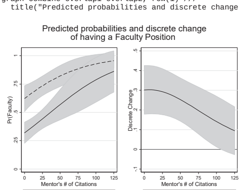
```
estimates restore estlogit
mgen, at(fellow=1 mcit3=(0(5)130)) atmeans stub(fel1) predlabel(Fellow)
   <snip>
mgen, at(fellow=0 mcit3=(0(5)130)) atmeans stub(fel0) predlabel(Not a Fellow)
   <snip>
graph twoway ///
    (rarea fel1ul fel1ll fel1mcit3, col(gs12)) ///
    (rarea fel0ul fel0ll fel1mcit3, col(gs12)) ///
    (connected fel1pr fel1mcit, lpat(dash)  msym(i)) ///
    (connected fel0pr fel1mcit, lpat(solid) msym(i))
    , ///
    legend(on order(3 4)) ylab(0(.25)1, grid gmin gmax) ytitle("Pr(Faculty)") ///
    xlab(0(10)130) xtitle("Mentor's # of Citations") name(overlap1, replace)
```

Next, we use **`mgen`** to compute the discrete change of **`fellow`** over the range of **`mcit3`**.

```
. mgen, dydx(fellow) at(mcit3=(0(5)130)) atmeans stub(fellowDC) ///
>   predlabel(Discrete change in fellow by mcit)

Predictions from: margins, dydx(fellow) at(mcit3=(0(5)130)) atmeans predict(pr)

Variable      Obs Unique      Mean       Min       Max  Label
--------------------------------------------------------------------------------
fellowDCd_~1   27     27  .2115119   .0870482  .3027023  Discrete change in f...
fellowDCll1    27     27  .0918198  -.0313136  .1798619  95% lower limit
fellowDCul1    27     27   .331204     .20541  .4260783  95% upper limit
fellowDCmc~3   27     27        65          0       130  Mentor's 3 yr citation.
--------------------------------------------------------------------------------

Specified values of covariates

       1.                     1.
   fellow        phd       mnas
-------------------------------
 .4128788   3.181894   .0833333
```

Finally we plot the discrete change and combine this graph with the one from section *3.11*. Note that the discrete change between levels of **`fellow`** is statistically significant at levels of **`mcit3`** where confidence intervals overlap.


```
. graph twoway ///
>   (rarea fellowDCul1 fellowDCll1 fellowDCmcit3, col(gs12)) ///
>   (connected fellowDCd_pr1 fellowDCmcit3, lpat(solid) msym(i)) ///
>   , ///
>   legend(on order(2)) ///
>   ylab(-0.1(.1).5, grid gmin gmax) ytitle("Discrete Change") ///
>   xlab(0(10)130) xtitle("Mentor's # of Citations") yline(0) name(overlap2, replace)

. graph combine overlap1 overlap2, row(1) ///
>   title("Predicted probabilities and discrete change of having a Faculty Position")
```



For an otherwise average scientist, having a postdoctoral fellowship increases the probability of being a faculty at a university. However, when the scientist's mentor has more than 100 citations, this difference is no longer statistically significant.

6 Hypothesis Testing

The file **`cda16lab-test-review.do`** contains these Stata commands.

#1 Load the Data

```
use cda-scireview4, clear
```

#2 Examine data, select variables, and verify

```
keep workfac female fellow phd mcit3 mnas
tab1 workfac female fellow mnas, miss
codebook, compact
```

#3 Computing a z-test

z-scores are produced with the standard ML estimation commands. The z-scores are in the 4th column, labeled "z". Estimation results are stored with **`estimates store`** using the name **`base`**.

```
. logit workfac i.female i.fellow c.phd c.mcit3 i.mnas, nolog
```



```
Logistic regression                             Number of obs     =        264
                                                LR chi2(5)        =      41.72
                                                Prob > chi2       =     0.0000
Log likelihood = -161.51514                     Pseudo R2         =     0.1144

------------------------------------------------------------------------------
     workfac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
      1_Yes  |  -.5869003   .2911944    -2.02   0.044    -1.157631   -.0161698
             |
      fellow |
      1_Yes  |   1.118336   .2844612     3.93   0.000     .5608027     1.67587
         phd |    .002004   .1521298     0.01   0.989    -.2961648    .3001729
       mcit3 |   .0190813   .0072584     2.63   0.009     .0048551    .0333075
             |
        mnas |
      1_Yes  |   .3537104   .5652778     0.63   0.531    -.7542137    1.461635
       _cons |  -.5004836   .4539085    -1.10   0.270    -1.390128    .3891607
------------------------------------------------------------------------------

. estimates store base
```

#4 Single Coefficient Wald Test

The **`test`** command computes a Wald test that a single coefficient is equal to zero. Note that the name **`1.female`** exactly matches the output from the logit output. Entering "female" or "i.female" will result in an error. This can be confusing when working with factor variables. To find out that names you need to use, you can use the command **`logit, coeflegend`** .

```
. test 1.female

 ( 1)  [workfac]1.female = 0

           chi2(  1) =    4.06
         Prob > chi2 =    0.0439
```

The effect of being female on the probability of being a faculty member is significant at the .05 level (X^2=4.06, *df*=1, *p*=0.04).

#5 Multiple Coefficients Wald Test

We can also test if multiple coefficients are simultaneously equal to zero.

```
. test mcit3 1.mnas

 ( 1)  [workfac]mcit3 = 0
 ( 2)  [workfac]1.mnas = 0

           chi2(  2) =    7.78
         Prob > chi2 =    0.0204
```

The hypothesis that the effects of mentor's citations and mentor's membership in the NAS on the probability of being a faculty member are simultaneously equal to zero can be rejected at the .05 level (X^2=7.78, *df*=2, *p*=0.02).

#6 Equal Coefficients Wald Test

We can test whether the magnitude of the effect of being female equals the effect of having a fellowship. Since female and fellow have opposite signs, we multiple fellow by -1.

```
. test 1.female = -1*1.fellow

 ( 1)  [workfac]1.female + [workfac]1.fellow = 0

           chi2(  1) =    1.42
         Prob > chi2 =    0.2331
```

The effects of being a female and having a postdoctoral fellowship on the probability of being a faculty member are not significantly different (X^2=1.42, *df*=1, *p*=0.23).

#7 Single Coefficient LR Test

To test that the effect of female is zero, run the base model without **female** and compare it with the full model, stored earlier as **base**, using **lrtest** *estname1 estname2*.

```
. logit workfac i.fellow c.phd c.mcit3 i.mnas, nolog
<snip>
. estimates store dropfemale

. lrtest base dropfemale

Likelihood-ratio test                                 LR chi2(1)  =      4.08
(Assumption: dropfemale nested in base)               Prob > chi2 =    0.0434
```

The effect of being female on the probability of being a faculty member is significant at the .05 level (*LRX*^{2}=4.08, *df*=1, *p*=0.04).

#8 Multiple Coefficients LR Test

To test if the effects of **mcit3** and **mnas** are jointly zero, run the comparison model without these variables, store using **estimates store**, and then compare models using **lrtest**.

```
. logit workfac i.female i.fellow c.phd
<snip>
. estimates store dropmcit3mnas

. lrtest base dropmcit3mnas

Likelihood-ratio test                                 LR chi2(2)  =      9.19
(Assumption: dropmcit3mnas nested in base)            Prob > chi2 =    0.0101
```

The hypothesis that the effects of mentor's citations and the mentor's status in the NAS on the probability of being a faculty member are simultaneously equal to zero can be rejected at the .05 level (*LRX*^{2}=9.19, *df*=2, *p*<0.05).

#9 LR Test All Coefficients are Zero

To test that all of the regression coefficients are zero, we estimate the model with only an intercept, store the results, and compare the models using **lrtest**. This test statistic is identical to the one at the top of the estimation output for the full model shown in **4.3**.


```
. logit workfac
<snip>
. estimates store intercept

. lrtest base intercept

Likelihood-ratio test                                 LR chi2(5)  =     41.72
(Assumption: intercept nested in base)                Prob > chi2 =    0.0000
```

We can reject the hypothesis that all coefficients except the intercept are zero at the .01 level (*LRX*^{2}=41.72, *df*=5, *p*<0.01).

7 Measures of Fit

The file **cda16lab-fit-review.do** contains these Stata commands.

#1 Load the Data

```
use cda-scireview4, clear
```

#2 Examine data, select variables, and verify

```
keep workfac female fellow phd mcit3 mnas
tab1 workfac female fellow mnas, miss
codebook, compact
```

#3 Fit Statistics

fitstat computes measures of fit for your model. The **save** option saves the measures for subsequent comparisons. **dif** compares the measures for the current model with those of the saved model. Here we compare the base model to the model without **mcit3** and **mnas**.

```
. logit workfac i.female i.fellow c.phd c.mcit3 i.mnas
<snip>

. fitstat, save

                         |       logit
-------------------------+-------------
Log-likelihood           |
                   Model |    -161.515
          Intercept-only |    -182.377
-------------------------+-------------
Chi-square               |
       Deviance (df=258) |     323.030
              LR (df=5)  |      41.723
                 p-value |       0.000
-------------------------+-------------
R2                       |
                McFadden |       0.114
<snip>

. logit workfac i.female i.fellow c.phd
<snip>

. fitstat, dif

                         |     Current        Saved   Difference
-------------------------+--------------------------------------
```



```
Log-likelihood           |
                   Model |    -166.112     -161.515       -4.596
          Intercept-only |    -182.377     -182.377        0.000
-------------------------+--------------------------------------
Chi-square               |
        D (df=260/258/2) |     332.223      323.030        9.193
        LR (df=3/5/-2)   |      32.530       41.723       -9.193
                 p-value |       0.000        0.000        0.010
-------------------------+--------------------------------------
R2                       |
                McFadden |       0.089        0.114       -0.025
     McFadden (adjusted) |       0.067        0.081       -0.014
      McKelvey & Zavoina |       0.145        0.201       -0.055
            Cox-Snell/ML |       0.116        0.146       -0.030
  Cragg-Uhler/Nagelkerke |       0.155        0.195       -0.040
                   Efron |       0.120        0.151       -0.031
               Tjur's D  |       0.119        0.150       -0.030
                   Count |       0.659        0.678       -0.019
        Count (adjusted) |       0.268        0.309       -0.041
-------------------------+--------------------------------------
IC                       |
                     AIC |     340.223      335.030        5.193
        AIC divided by N |       1.289        1.269        0.020
        BIC (df=4/6/-2)  |     354.527      356.486       -1.959
-------------------------+--------------------------------------
Variance of              |
                       e |       3.290        3.290        0.000
                  y-star |       3.850        4.116       -0.266

Note: Likelihood-ratio test assumes current model nested in saved model.

Difference of    1.959 in BIC provides weak support for current model.
```

#4 Fit Statistics, Information measures only

fitstat with the **ic** option presents only information measures BIC and AIC. **ic** can be combined with the **save** and **dif** options.

```
. quietly logit workfac i.female i.fellow c.phd c.mcit3 c.mnas

. fitstat, ic

                         |       logit
-------------------------+-------------
AIC                      |
                     AIC |     335.030
          (divided by N) |       1.269
-------------------------+-------------
BIC                      |
              BIC (df=6) |     356.486
 BIC (based on deviance) |   -1115.565
   BIC' (based on LRX2)  |     -13.843
```

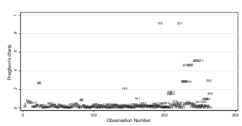
#5 Plotting Influential Cases Using dbeta

We compute influence using the command **predict, dbeta**. Then we sort our data in some meaningful way (here we choose to sort by **phd**). Next we generate the variable **index** whose values correspond to the rank order of **phd** (because of the way the data are sorted). Finally we plot the dbeta distance against the rank order of **phd**. You can also plot residuals as shown in the lecture notes.

```
. twoway scatter dbeta index, ysiz(1) xsiz(2) ///
>         xlab(0(100)300) ylab(0(.2)1., grid)
```



```
>         xscale(range(0, 300)) yscale(range(0, 1.)) ///
>         xtitle("Observation Number") msym(none) mlab(index) mlabposition(0)
. graph export `pgm'-cookplot.`graphtype'.emf, replace
```



8 Binary Outcomes: Complex Sampling and Nonlinearity

The file **cda16lab-brm-complications-review.do** contains these commands.

#1 Load the Data

```
use cda-hrs4, clear
```

#2 Examine data, select variables, and verify

```
keep arthritis age female ed11less ed12 ///
     ed1315 ed16plus secu kwgtr stratum
tab1 arthritis female ed11less ed12 ed1315 ed16plus, miss
codebook, compact
```

#3 Prepare Stata for svy commands

Always double check variables related to survey design to avoid careless mistakes, like using the wrong variables to define your sample.

```
. codebook secu kwgtr stratum, compact

Variable     Obs Unique      Mean  Min    Max  Label
-------------------------------------------------------------------------------
secu       18467      2  1.502356    1      2  sampling error computation unit
kwgtr      18467   4219  4144.727    0  16532  2006 weight: respondent level
stratum    18467     56  30.99767    1     56  stratum id
-------------------------------------------------------------------------------
```

Then declare that you are using a complex sampling design:

```
. svyset secu [pweight=kwgtr], ///
>     strata(stratum) vce(linearized) singleunit(missing)
```

#4 Examine Descriptive Statistics with and without Survey Variables

Next, look at descriptive statistics without survey adjustments and note how the survey adjustments affect variables. First we examine the mean and standard deviation without accounting for survey complexities.

```
. mean arthritis female age ed11less ed12 ed1315 ed16plus
<snip>
```



```
. estat sd

-------------------------------------
             |       Mean   Std. Dev.
-------------+-----------------------
   arthritis |   .5999673     .489918
      female |   .5898261    .4918785
         age |   68.54446      11.104
    ed11less |   .2433891    .4291397
        ed12 |   .3342239    .4717313
      ed1315 |    .209367    .4068679
    ed16plus |     .21302    .4094528
-------------------------------------
```

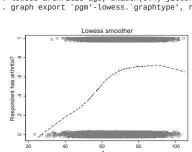
We compare these results to statistics accounting for survey complexities by adding **svy:** before **mean**.

```
. svy: mean arthritis female age ed11less ed12 ed1315 ed16plus
(running mean on estimation sample)
<snip>
. estat sd
-------------------------------------
             |       Mean   Std. Dev.
-------------+-----------------------
   arthritis |   .5699108     .495103
      female |   .5448029    .4980034
         age |   66.54182    10.37453
    ed11less |   .1958271    .3968478
        ed12 |   .3277077    .4693915
      ed1315 |   .2253607    .4178321
    ed16plus |   .2511045    .4336614
-------------------------------------
```

#5 Lowess plot

Now that we've set up our survey data, we can analyze nonlinearities in the right hand side of the model. A lowess plot shows a moving average of y as x changes. For key variables, a lowess plot can be a valuable first step in determining potential nonlinearities. Stata typically takes longer to produce lowess plots than other kinds of plots, so be patient. **lowess** does not support **svy:**, so these results are only exploratory.

```
. lowess arthritis age, bwidth(0.4) jitter(4) msym(oh)
. graph export `pgm'-lowess.`graphtype', replace
```



#6 Logit Models with Age, Age-squared, and Age-Cubed

Since the lowess plot suggests age has a nonlinear association with arthritis that cannot be captured by a logit model in which only age is included, we'll examine this more formally. We begin by estimating a model with only age, then add age-squared, and finally add age-cubed. After each regression, we compute a Wald test determining whether the age terms are simultaneously equal to zero. **logit** is preceded by **svy:** which means that the models are fit taking into account the complex survey design. A squared term is added by including the factor notation **c.age##c.age** as an independent. ## indicates that both age and age-squared are to be included in the model. To see independent variable names for Wald tests, include the command **logit, coeflegend** after running a logistic regression. First for the model with only age:

```
. * aM1: age
. svy: logit arthritis female ed11less ed1315 ed16plus age
(running logit on estimation sample)
<snip>
. estimates store aM1

. test age

Adjusted Wald test

 ( 1)  [arthritis]age = 0

       F(  1,    56) =  480.28
            Prob > F =    0.0000
```

Adding age-squared:

```
. * aM2: age + age^2
. svy: logit arthritis female ed11less ed1315 ed16plus c.age##c.age
(running logit on estimation sample)
<snip>
. estimates store aM2

. test age c.age#c.age

Adjusted Wald test

 ( 1)  [arthritis]age = 0
 ( 2)  [arthritis]c.age#c.age = 0

       F(  2,    55) =  272.38
            Prob > F =    0.0000
```

Adding age-cubed:

```
. * aM3: age + age^2 + age^3
. svy: logit arthritis female ed11less ed1315 ed16plus ///
> c.age c.age#c.age c.age#c.age#c.age
(running logit on estimation sample)
<snip>
. estimates store aM3

. test age c.age#c.age c.age#c.age#c.age

Adjusted Wald test

 ( 1)  [arthritis]age = 0
 ( 2)  [arthritis]c.age#c.age = 0
 ( 3)  [arthritis]c.age#c.age#c.age = 0
```

```
       F(  3,    54) =  175.52
            Prob > F =    0.0000
```

The **estimates table** command provides a concise way to view the three regression models.

```
. * tables of estimated coefficients
. estimates table aM1 aM2 aM3, title(Arthritis) ///
>     eform b(%9.3f) t(%9.2f) stats(N)

Arthritis

-----------------------------------------------
    Variable |    aM1          aM2          aM3
-------------+---------------------------------
      female |   1.779        1.813        1.815
             |   12.99        13.13        13.16
    ed11less |   1.206        1.217        1.217
             |    3.16         3.32         3.32
      ed1315 |   0.937        0.966        0.966
             |   -1.21        -0.62        -0.61
    ed16plus |   0.638        0.651        0.652
             |   -8.54        -8.09        -8.06
         age |   1.049        1.361        2.272
             |   21.92        12.11         3.32
             |
 c.age#c.age |                0.998        0.991
             |               -10.57        -2.67
             |
 c.age#c.age#|
       c.age |                             1.000
             |                              2.16
             |
       _cons |   0.046        0.000        0.000
             |  -19.54       -13.22        -4.07
-------------+---------------------------------
           N |   18375        18375        18375
-----------------------------------------------
                                    legend: b/t
```

#7 A closer look at the probabilities

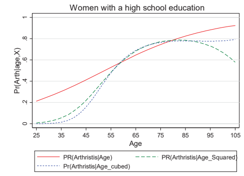
After determining that age, age-squared, and age-cubed are all significant, it is time to graph the predicted probabilities. We use **mgen** to create variables with predictions. Notice that as age changes, **mgen** uses **margins** to automatically increase age-squared and age-cubed.

```
. estimates restore aM1
(results aM1 are active now)

. mgen, at(age=(25(2.5)105) female=1 ed11less=0 ed1315=0 ed16plus=0) ///
>     stub(aM1) noci predlabel(PR(Arthristis|Age))
<snip>
. estimates restore aM2

. mgen, at(age=(25(2.5)105) female=1 ed11less=0 ed1315=0 ed16plus=0) ///
>     stub(aM2) noci predlabel(PR(Arthristis|Age_Squared))
<snip>
. estimates restore aM3

. mgen, at(age=(25(2.5)105) female=1 ed11less=0 ed1315=0 ed16plus=0) ///
>     stub(aM3) noci predlabel(Pr(Arthristis|Age_cubed))
<snip>
```


#8: Graph the probabilities

Now that we've created variables for the predicted probabilities with **mgen**, we can to make the graph below.

```
. graph twoway ///
>     (connected aM1pr aM1kage, msym(i) lcol(red)   lpat(solid)) ///
>     (connected aM2pr aM2kage, msym(i) lcol(green) lpat(dash)) ///
>     (connected aM3pr aM3kage, msym(i) lcol(blue)  lpat(shortdash)), ///
>     title("Women with a high school education") xtitle("Age") ///
>     ytitle("Pr(Arth|age,X)") xlabel(25(10)105) ylabel(0(.2)1, grid) ///
>     yline(0 1, lcol(gs13))
. graph export `pgm'-probplot.`graphtype', replace
```



9 Nominal Outcomes

cda16lab-nrm-review.do contains these Stata commands.

#1 Load the Data

```
use cda-scireview4, clear
```

#2 Examine data, select variables, and verify

Make sure to pay special attention to the distribution of the outcome variable **jobprst**.

```
keep jobprst pub1 phd female
tab1 pub1 female, miss
codebook, compact
```

#3 Multinomial Logit

mlogit estimates the multinomial logit model. The option **baseoutcome()** allows you to set the comparison category. **estimates store** stores estimation results for model comparison.

```
. mlogit jobprst c.pub1 c.phd i.female, baseoutcome(4) nolog

Multinomial logistic regression                 Number of obs   =        264
```



```
                                                LR chi2(9)      =     108.80
                                                Prob > chi2     =     0.0000
Log likelihood = -240.45919                     Pseudo R2       =     0.1845

------------------------------------------------------------------------------
     jobprst |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1_Adeq       |
        pub1 |  -.1577122   .1164937    -1.35   0.176    -.3860356    .0706112
         phd |  -2.227522   .5717459    -3.90   0.000    -3.348123   -1.106921
             |
      female |
      1_Yes  |   2.016045   1.168225     1.73   0.084    -.2736336    4.305724
       _cons |   8.952493   2.312129     3.87   0.000     4.420802    13.48418
-------------+----------------------------------------------------------------
2_Good       |
        pub1 |  -.2360238   .1027013    -2.30   0.022    -.4373146   -.0347329
         phd |  -2.473911   .5486436    -4.51   0.000    -3.549233   -1.398589
             |
      female |
      1_Yes  |   2.957967   1.104288     2.68   0.007     .7936018    5.122331
       _cons |    10.9781   2.257877     4.86   0.000     6.552745    15.40346
-------------+----------------------------------------------------------------
3_Strong     |
        pub1 |  -.1196281   .0831959    -1.44   0.150    -.2826891    .0434329
         phd |  -1.080595   .5279581    -2.05   0.041    -2.115374   -.0458166
             |
      female |
      1_Yes  |    1.76863   1.082655     1.63   0.102    -.3533356    3.890596
       _cons |   6.285116   2.216631     2.84   0.005     1.940598    10.62963
-------------+----------------------------------------------------------------
4_Dist       |  (base outcome)
------------------------------------------------------------------------------

. estimates store base
```

#4 Single Variable LR Test

In the MNLM, testing that a variable has no effect requires a test that *J*-1 coefficients are simultaneously equal to zero. For example, the effect of **i.female** involves three coefficients. We can use an LR test to test that all three are simultaneously equal to zero. First, we save the base model (which we did above); second, we estimate the model without **i.female** and store the estimation results; and third, we compare the two models using **lrtest** *estname1 estname2.*

```
. quietly mlogit jobprst c.pub1 c.phd, baseoutcome(4)
. estimates store dropfemale

. lrtest base dropfemale

Likelihood-ratio test                                 LR chi2(3)  =     19.17
(Assumption: dropfemale nested in base)               Prob > chi2 =    0.0003
```

The effect of gender on job prestige is significant at the .001 level (LRX^2=19.17, *df*=3, *p*=<.001).

Another way to do this is to use the command **mlogtest** after fitting the model. This saves your having to re-estimate the model minus the variable whose effect you want to test.

```
. estimates restore base
(results base are active now)

. mlogtest, lr
```



```
Likelihood-ratio tests for independent variables (N=264)

  Ho: All coefficients associated with given variable(s) are 0

             |       chi2   df   P>chi2
-------------+-------------------------
        pub1 |      5.600    3    0.133
         phd |     87.236    3    0.000
    1.female |     19.168    3    0.000
---------------------------------------
```

#5 Single Coefficient Wald Test

Wald tests can also be computed using the **test** command. For factor variables, you must enter the variable exactly as it is shown in the regression output, in this case **1.female**. To understand the labeling, **[1_Adeq]1.female** is the coefficient for category **1_Adeq** contrasted to the reference category **4_Dist**:

```
. test 1.female

 ( 1)  [1_Adeq]1.female = 0
 ( 2)  [2_Good]1.female = 0
 ( 3)  [3_Strong]1.female = 0
 ( 4)  [4_Dist]1o.female = 0
       Constraint 4 dropped

           chi2(  3) =   15.75
         Prob > chi2 =    0.0013
```

Again, you can automate this process using **mlogtest**.

```
. mlogtest, wald

Wald tests for independent variables (N=264)

  Ho: All coefficients associated with given variable(s) are 0

             |       chi2   df   P>chi2
-------------+-------------------------
        pub1 |      5.421    3    0.143
         phd |     56.559    3    0.000
    1.female |     15.748    3    0.001
---------------------------------------
```

The effect of gender on job prestige is significant at the .001 level (X^2=15.7, *df*=3, *p*=<.001).

#6 Combining Outcomes Test (low priority unless you need this test)

test can also compute a Wald test that two outcomes can be combined. Recall, that the coefficients for category **1_Adeq** were in comparison to the category **4_Dist**. Therefore, we are testing whether we can combine **1_Adeq** and **4_Dist**. Note that **[1_Adeq]** is necessary in specifying the test across categories and that **[1_Adeq]** does not equal **[1_adeq]** since syntax in Stata is case sensitive.

```
. test [1_Adeq]

 ( 1)  [1_Adeq]pub1 = 0
 ( 2)  [1_Adeq]phd = 0
 ( 3)  [1_Adeq]0b.female = 0
 ( 4)  [1_Adeq]1.female = 0
```



```
       Constraint 3 dropped

           chi2(  3) =   19.01
         Prob > chi2 =    0.0003
```

We can reject the hypothesis that *adequate* and *distinguished* are indistinguishable (X^2=19.0, *df*=3, *p*<.001).

This test could be done for combining other categories as well. For example we could test whether we can combine categories Adequate and Good by typing **test [1_Adeq=2_Good]**. But the easier way is to use **mlogtest**.

```
. mlogtest, combine

Wald tests for combining alternatives (N=264)

  Ho: All coefficients except intercepts associated with a given pair
      of alternatives are 0 (i.e., alternatives can be combined)

Alternatives tested|     chi2   df   P>chi2
-------------------+------------------------
  1_Adeq-  2_Good  |    5.189    3    0.158
  1_Adeq-3_Strong  |   19.884    3    0.000
  1_Adeq-  4_Dist  |   19.015    3    0.000
  2_Good-3_Strong  |   51.717    3    0.000
  2_Good-  4_Dist  |   31.132    3    0.000
3_Strong-  4_Dist  |    9.173    3    0.027
-------------------+------------------------
```

We cannot reject the hypothesis that categories *adequate* and *good* are indistinguishable (X^2=5.2, *df*=3, *p*=0.16).

#7 Testing for IIA (low priority unless you need this test)

mlogtest can be used to test the IIA (independence of irrelevant alternatives) assumption. While often recommended, this test is not very useful. Nonetheless, **mlogtest** computes both a Hausman and a Small-Hsiao test. Because the Small-Hsiao test requires randomly dividing the data into subsamples, the results will differ with successive calls of the command. To obtain test results that can be replicated, we set the seed used by the random-number generator. You can set the seed to whatever number you like. But when setting seeds in research that will be published, refer to the suggestions made in **help set seed**, as some seeds are more trustworthy than others.

```
. set seed 4415906
. mlogtest , iia

Hausman tests of IIA assumption (N=264)

  Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives

                |      chi2    df   P>chi2
----------------+-------------------------
         1_Adeq |     3.588     8    0.892
         2_Good |    17.887     8    0.022
       3_Strong |   -45.118     8        .
         4_Dist |    -0.222     8        .

  Note: A significant test is evidence against Ho.
```



```
  Note: If chi2<0, the estimated model does not meet asymptotic assumptions.

suest-based Hausman tests of IIA assumption (N=264)

  Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives

                |      chi2    df   P>chi2
----------------+-------------------------
         1_Adeq |     4.309     8    0.828
         2_Good |     9.915     8    0.271
       3_Strong |    21.271     8    0.006
         4_Dist |     4.377     8    0.822

  Note: A significant test is evidence against Ho.

Small-Hsiao tests of IIA assumption (N=264)

  Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives

                | lnL(full)  lnL(omit)      chi2    df   P>chi2
----------------+-----------------------------------------------
         1_Adeq |   -83.512    -72.740    21.543     8    0.006
         2_Good |   -70.925    -55.187    31.476     8    0.000
       3_Strong |   -76.846    -56.081    41.531     8    0.000
         4_Dist |  -112.991   -104.306    17.369     8    0.026

  Note: A significant test is evidence against Ho.
```

As is often the case with IIA tests, the evidence is mixed.

#8 Predicted Probabilities

mtable computes predicted probabilities for values of the independent variables. By default, **mtable** shows predicted probabilities for each outcome category. If you only want to list certain outcome categories, use the **outcome()** option.

```
. mtable, atmeans stat(ci)

Expression: Pr(jobprst), predict(outcome())

          |   1_Adeq    2_Good  3_Strong    4_Dist
----------+----------------------------------------
    Pr(y) |    0.128     0.513     0.344     0.014
       ll |    0.081     0.440     0.274    -0.004
       ul |    0.176     0.587     0.415     0.032

Specified values of covariates

          |                            1.
          |     pub1       phd    female
----------+-----------------------------
  Current |     2.32      3.18      .345
```

For an average scientist, the probability of being employed in a department rated as good is 0.51 (95% CI: 0.44, 0.59).

#9 Marginal and Discrete Change

We use **mchange** to calculate marginal and discrete change. By default, these are AME's. We only consider discrete change, specified by **amount(one sd)**.

```
. mchange, amount(one sd)

mlogit: Changes in Pr(y) | Number of obs = 264

Expression: Pr(jobprst), predict(outcome())

               |    1 Adeq     2 Good   3 Strong     4 Dist
---------------+--------------------------------------------
pub1           |
            +1 |     0.003     -0.021      0.011      0.007
       p-value |     0.732      0.079      0.293      0.084
           +SD |     0.007     -0.053      0.028      0.018
       p-value |     0.750      0.076      0.319      0.106
phd            |
            +1 |    -0.036     -0.201      0.144      0.093
       p-value |     0.002      0.000      0.000      0.022
           +SD |    -0.037     -0.202      0.145      0.093
       p-value |     0.002      0.000      0.000      0.022
female         |
 1 Yes vs 0 No |    -0.043      0.224     -0.116     -0.065
       p-value |     0.267      0.000      0.032      0.005

Average predictions

             |    1_Adeq     2_Good   3_Strong     4_Dist
-------------+--------------------------------------------
  Pr(y|base) |     0.110      0.485      0.352      0.053
```

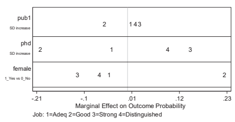
On average, being a female scientist is expected to decrease the probability of a job in a strong department by 0.12 (p<0.05, two-tailed test) and to decrease the probability of being in a distinguished department by 0.07 (p<0.01, two-tailed test).

On average, increasing PhD prestige by one level increases the probability on having a distinguished job by 0.09 (p<0.01, two-tailed test).

#10 Plot Discrete Change

One difficulty with nominal outcomes is the many coefficients that need to be considered. To help you sort out the information, discrete change coefficients can be plotted using **mchangeplot**. We recommend adding a **note** to the plot that includes the values and value labels. **mchangeplot** must be run after **mchange**. We use **aspect(.4)** to change the vertical spacing of the graph.

```
. mchangeplot pub1 phd 1.female aspect(.4) ///
>         note(Job: 1=Adeq 2=Good 3=Strong 4=Distinguished)
. graph export `pgm'-mchangeplot.`graphtype', replace
```



The average marginal effects of a standard deviation change in PhD prestige and of being female are larger than the effects of a standard deviation change in publications. On average, a standard deviation increase in PhD prestige increases the probability of being in a strong (3) department and decreases the probability of being in a good (2) department by about .20. Being female increases the probability of being in a good (2) department by .22 and decreases the probability of being in a strong (3) one by .12.

We could check the output from **mchange** to determine if the effects are significant, or we could indicate this in the graph by using the **significance()** option.

#11 Odds Ratios

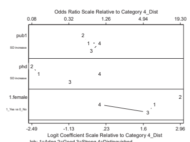
listcoef computes the factor change coefficients for each of the comparisons. The output is arranged by the independent variables.

```
. listcoef, help

mlogit (N=264): Factor change in the odds of jobprst

Variable: pub1 (sd=2.581)
-------------------------------------------------------------------------------
                             |         b        z    P>|z|       e^b   e^bStdX
-----------------------------+-------------------------------------------------
1_Adeq      vs 2_Good        |    0.0783    0.879    0.379     1.081     1.224
1_Adeq      vs 3_Strong      |   -0.0381   -0.412    0.680     0.963     0.906
1_Adeq      vs 4_Dist        |   -0.1577   -1.354    0.176     0.854     0.666
2_Good      vs 1_Adeq        |   -0.0783   -0.879    0.379     0.925     0.817
2_Good      vs 3_Strong      |   -0.1164   -1.623    0.105     0.890     0.741
2_Good      vs 4_Dist        |   -0.2360   -2.298    0.022     0.790     0.544
3_Strong    vs 1_Adeq        |    0.0381    0.412    0.680     1.039     1.103
3_Strong    vs 2_Good        |    0.1164    1.623    0.105     1.123     1.350
3_Strong    vs 4_Dist        |   -0.1196   -1.438    0.150     0.887     0.734
4_Dist      vs 1_Adeq        |    0.1577    1.354    0.176     1.171     1.502
4_Dist      vs 2_Good        |    0.2360    2.298    0.022     1.266     1.839
4_Dist      vs 3_Strong      |    0.1196    1.438    0.150     1.127     1.362
-------------------------------------------------------------------------------
<snip>
```

#12: Plot Odds Ratios

The odds ratios can be plotted in much the same way as the discrete changes by using the **mlogitplot** command. In the plot, a solid line indicates that the coefficient cannot differentiate between the two outcomes that are connected (i.e., the odds ratio is not significant). The significance level of the line is set with **linep()**.

```
. mlogitplot pub1 phd 1.female ///
>         note(Job: 1=Adeq 2=Good 3=Strong 4=Distinguished) linep(.1)
. graph export `pgm'-mlogitplot.`graphtype', replace
```

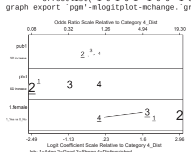
Here is a summary of the general pattern of effects:

The effects of publications are smallest, while the overall magnitude of effects of doctoral origin and being female being roughly equal. While doctoral prestige does not significantly affect the odds of working in adequate compared to a good department, it significantly increases the odds of strong and distinguished positions. Overall, being female increases the odds of less prestigious jobs.

#13: Adding Discrete Change to OR Plot

Information about the discrete change can be incorporated in the odds-ratio plot by using **mlogitplot, mchange**. Whereas the factor change in the odds is constant across the levels of all variables, the discrete change gets larger or smaller at different values of the independent variables. In the plot below, the discrete change is indicated by the size of the numbers with the area of the number proportional to the size of the discrete change. A number is underlined to indicate a negative discrete change. The **offsetlist** and **msizefactor** options "tweak" the graph to make it look better. Try experimenting with them. Try **help mlogitplot** for details.

```
. mlogitplot pub1 phd 1.female, ///
>         note(Job: 1=Adeq 2=Good 3=Strong 4=Distinguished) linep(.1) mchange ///
>         offsetlist(-1 0 1 0 1 -1 0 0 -1 0 1 -1) msizefactor(1.4)
. graph export `pgm'-mlogitplot-mchange.`graphtype', replace
```



10 Ordinal Outcomes

The file **cda16lab-orm-review.do** contains these Stata commands.

#1 Load the Data

```
use cda-scireview4, clear
```

#2 Examine data, select variables, and verify

Be sure to look at the distribution of the outcome variable, in this case **jobprst**.

```
keep jobprst pub1 phd female
tab1 jobprst female, miss
codebook, compact
```

#3 Ordered Logit

ologit and **oprobit** work in the same way. We only show **ologit**, but you could use **oprobit**.

```
. ologit jobprst c.pub1 c.phd i.female, nolog

Ordered logistic regression                     Number of obs   =        264
                                                LR chi2(3)      =      80.69
                                                Prob > chi2     =     0.0000
Log likelihood = -254.51518                     Pseudo R2       =     0.1368

------------------------------------------------------------------------------
     jobprst |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pub1 |   .1078786   .0481107     2.24   0.025     .0135833    .2021738
         phd |   1.130028   .1444046     7.83   0.000     .8470003    1.413056
             |
      female |
       1_Yes |  -.6973579   .2617103    -2.66   0.008    -1.210301   -.1844152
-------------+----------------------------------------------------------------
       /cut1 |   .9274554   .4268201                      .0909033    1.764007
       /cut2 |   4.003182   .4996639                      3.023859    4.982506
       /cut3 |   7.034637   .6296717                      5.800503     8.26877
------------------------------------------------------------------------------

. estimates store ologit
```

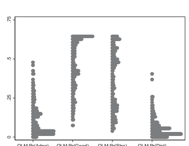
#4 Predicted Probabilities in Sample

predict computes predicted probabilities after **ologit** or **oprobit**. It creates as many new variables as there are categories of the outcome variable so you will need to provide variable names that correspond to the four outcome categories. The first variable contains the probability associated with the lowest outcome; the second the probability associated with the second outcome; and so on. Remember to label the newly created variables.

```
predict jpad jpgo jpst jpdi
label var jpad "OLM Pr(Adeq)"
label var jpgo "OLM Pr(Good)"
label var jpst "OLM Pr(Strg)"
label var jpdi "OLM Pr(Dist)"
```

An easy way to see the range of predictions is with the command **dotplot**.

```
. dotplot jpad jpgo jpst jpdi, ylabel(0(.25).75)
. graph export `pgm'-dotplot.`graphtype', replace
```

#5 Predict Specific Probabilities

mtable computes the predicted value for a set of values for the independent variables. Use the **at()** and **atmeans** options to set the values at which the variables will be examined.

```
. mtable, at(female=1 phd=4) atmeans stat(ci)

Expression: Pr(jobprst), predict(outcome())

          |   1_Adeq    2_Good  3_Strong    4_Dist
----------+----------------------------------------
    Pr(y) |    0.041     0.441     0.468     0.049
       ll |    0.017     0.344     0.369     0.018
       ul |    0.065     0.539     0.568     0.080

Specified values of covariates

          |     pub1       phd    female
----------+-----------------------------
  Current |     2.32         4         1
```

A female scientist with a doctorate from a distinguished university who is otherwise average has a probability of .05 of obtaining a distinguished job (95% CI: 0.02, 0.08).

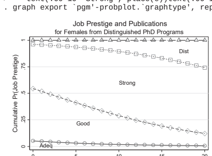
#6 Graph Predicted Probabilities

Graphing predictions as a continuous variable changes is a useful way to examine the effect of the variable. **mgen** creates variables for graphing. We consider women from distinguished PhD programs **(phd=4)** and show how predicted probabilities are influenced by publications. **mgen** creates variables of both the predicted probabilities and the cumulative probabilities. We plot the cumulative probabilities below. The lecture notes have example of plotting the probabilities for each category.

```
. mgen, at(female=1 phd=4 pub1=(0(1)20)) atmeans stub(pub)

Predictions from: margins, at(female=1 phd=4 pub1=(0(1)20)) atmeans predict(outc
> ome())

Variable   Obs Unique      Mean       Min       Max  Label
--------------------------------------------------------------------------------
pubpr1      21     21  .0223568  .0063504  .0523864  pr(y=1_Adeq) from margins
publl1      21     21  .0036257 -.0053376  .0215249  95% lower limit
pubul1      21     21  .0410879  .0180384   .083248  95% upper limit
```



```
pubpub1     21     21        10         0        20  Publications: PhD yr -...
pubCpr1     21     21  .0223568  .0063504  .0523864  pr(y<=1_Adeq)
pubpr2      21     21  .2839126  .1152733  .4925985  pr(y=2_Good) from margins
:: output omitted ::
--------------------------------------------------------------------------------

Specified values of covariates

      phd     female
---------------------
        4          1

. graph twoway (connected pubCpr1 pubCpr2 pubCpr3 pubCpr4 pubpub1, ///
>    title("Job Prestige and Publications") ///
>    subtitle("for Females from Distinguished PhD Programs") ///
>    ytitle("Cumulative Pr(Job Prestige)") xtitle("Publications: PhD yr -1 to 1") ///
>    xlabel(0(5)20) ylabel(0(.25)1, grid) msymbol(Oh Dh Sh Th) name(tmp2, replace) ///
>    text(.01 .75 "Adeq", place(e)) text(.22 5 "Good", place(e)) ///
>    text(.60 10 "Strong", place(e))text(.90 17 "Dist", place(e))), legend(off)
. graph export `pgm'-probplot.`graphtype', replace
```



The plot shows many things. For women with PhDs from distinguished programs, the probability of obtaining a job in the least prestigious programs, referred to as adequate, is low regardless of the number of publications. Second, the probability of obtaining a job in a good program decreases rapidly as the number of publications increases, with a corresponding increases in the probability of jobs in strong or distinguished programs. With twenty publications, over 80% of these women are predicted to be in these types of positions. Third, the increase in strong and distinguished jobs is offset by a corresponding decreases in good jobs.

#8 Discrete Change

mchange computes marginal and discrete change at specific values of the independent variables. Values for specific independent variables can be set using the **at()**. The below results are computed using AME.

```
. mchange, amount(one sd)

ologit: Changes in Pr(y) | Number of obs = 264

Expression: Pr(jobprst), predict(outcome())
```

```
               |    1 Adeq     2 Good   3 Strong     4 Dist
---------------+--------------------------------------------
pub1           |
            +1 |    -0.009     -0.012      0.015      0.005
       p-value |     0.027      0.028      0.024      0.045
           +SD |    -0.021     -0.031      0.037      0.015
       p-value |     0.020      0.034      0.021      0.056
phd            |
            +1 |    -0.064     -0.146      0.127      0.083
       p-value |     0.000      0.000      0.000      0.000
           +SD |    -0.065     -0.146      0.127      0.084
       p-value |     0.000      0.000      0.000      0.000
female         |
 1 Yes vs 0 No |     0.062      0.066     -0.097     -0.031
       p-value |     0.014      0.006      0.008      0.015

Average predictions

             |    1_Adeq     2_Good   3_Strong     4_Dist
-------------+--------------------------------------------
  Pr(y|base) |     0.104      0.470      0.371      0.055
```

On average, being a female scientist increases the probability of adequate and good job placements by .06 (p<0.05 and p<0.01 respectively, two-tailed test), and decreases the probability of strong jobs by .10 (p<0.01, two-tailed test) and distinguished jobs by .03 (p<0.05, two-tailed test).

If we wanted to compute predictions for women from distinguished departments who are average on other characteristics (i.e. MEM):

```
. mchange, at(female=1 phd=4) atmeans amount(one sd)

ologit: Changes in Pr(y) | Number of obs = 264

Expression: Pr(jobprst), predict(outcome())

               |    1 Adeq     2 Good   3 Strong     4 Dist
---------------+--------------------------------------------
pub1           |
            +1 |    -0.004     -0.023      0.022      0.005
       p-value |     0.037      0.028      0.026      0.066
           +SD |    -0.010     -0.059      0.054      0.015
       p-value |     0.028      0.027      0.021      0.080
phd            |
            +1 |    -0.028     -0.223      0.162      0.089
       p-value |     0.000      0.000      0.000      0.002
           +SD |    -0.028     -0.224      0.162      0.090
       p-value |     0.000      0.000      0.000      0.002
female         |
 1 Yes vs 0 No |     0.020      0.145     -0.121     -0.045
       p-value |     0.029      0.007      0.012      0.018

Predictions at base value

             |    1_Adeq     2_Good   3_Strong     4_Dist
-------------+--------------------------------------------
  Pr(y|base) |     0.041      0.441      0.468      0.049

Base values of regressors
```



```
             |      pub1        phd     female
-------------+---------------------------------
          at |      2.32          4          1

1: Estimates with margins option atmeans.
```

#9 Odds Ratios

The factor change in the odds can be computed for the ordinal logit model. Again we do this with the command **listcoef**. The **help** option presents a “key” to interpreting the headings of the output.

```
. listcoef, help

ologit (N=264): Factor change in odds

  Odds of: >m vs <=m

-------------------------------------------------------------------------
             |         b         z     P>|z|       e^b   e^bStdX     SDofX
-------------+-----------------------------------------------------------
        pub1 |    0.1079     2.242     0.025     1.114     1.321     2.581
         phd |    1.1300     7.825     0.000     3.096     3.114     1.005
             |
      female |
       1_Yes |   -0.6974    -2.665     0.008     0.498     0.717     0.476
   constant1 |    0.9275     2.173     0.030         .         .         .
   constant2 |    4.0032     8.012     0.000         .         .         .
   constant3 |    7.0346    11.172     0.000         .         .         .
   constant4 |         .         .         .         .         .         .
-------------------------------------------------------------------------
       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
   SDofX = standard deviation of X
```

The odds of receiving a higher ranked job are .50 times smaller for women than men, holding other variables constant (p<0.01, two-tailed test).

For a standard deviation increase in publications, about 2.6, the odds of receiving a higher ranked job increase by a factor of 1.3, holding other variables constant (p<0.05, two-tailed test).

#10 Testing the Parallel Regression Assumption

brant performs a Brant test of the parallel regressions assumptions for the ordered logit model.

```
. brant, detail

<snip>

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |     38.88    0.000     6
-------------+--------------------------
        pub1 |      2.76    0.252     2
         phd |     22.68    0.000     2
    1.female |     11.26    0.004     2
----------------------------------------
```



```
A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

There is strong evidence that the parallel regression assumption is violated (p<.001).

11 Count Outcomes

The file **cda16lab-crm-review.do** contains these Stata commands.

#1 Load the Data

```
use cda-scireview4, clear
```

#2 Examine data, select variables, and verify

Make sure to look at the distribution of the outcome variable, in this case, **pub6**.

```
keep pub6 female phd enrol
tab1 pub6 female, miss
codebook, compact
```

#3 Estimate the Negative Binomial Regression Model

```
. nbreg pub6 i.female c.phd c.enrol, nolog

Negative binomial regression                    Number of obs     =        264
                                                LR chi2(3)        =      20.59
Dispersion     = mean                           Prob > chi2       =     0.0001
Log likelihood = -642.723                       Pseudo R2         =     0.0158

------------------------------------------------------------------------------
        pub6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
      1_Yes  |  -.2822292   .1382637    -2.04   0.041     -.553221   -.0112373
         phd |   .1995909   .0651859     3.06   0.002     .0718288     .327353
      enroll |   -.150895   .0480431    -3.14   0.002    -.2450578   -.0567322
       _cons |   1.607418   .3379749     4.76   0.000     .9449989    2.269836
-------------+----------------------------------------------------------------
    /lnalpha |   -.203673   .1255831                     -.4498113    .0424654
-------------+----------------------------------------------------------------
       alpha |   .8157291   .1024418                      .6377485     1.04338
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =  394.12 Prob>=chibar2 = 0.000
```

Because there is significant evidence of overdispersion (G^2=394.12, $p<.001$), the negative binomial regression model is preferred to the Poisson regression model.

#4 Factor Changes

listcoef computes the factor change coefficients.

```
. listcoef, help

nbreg (N=264): Factor change in expected count
```



```
  Observed SD:  4.3103

-------------------------------------------------------------------------
             |         b         z     P>|z|       e^b   e^bStdX     SDofX
-------------+-----------------------------------------------------------
      female |
      1_Yes  |   -0.2822    -2.041     0.041     0.754     0.874     0.476
         phd |    0.1996     3.062     0.002     1.221     1.222     1.005
      enroll |   -0.1509    -3.141     0.002     0.860     0.804     1.443
    constant |    1.6074     4.756     0.000         .         .         .
-------------+-----------------------------------------------------------
alpha        |
     lnalpha |   -0.2037         .         .         .         .         .
       alpha |    0.8157         .         .         .         .         .
-------------------------------------------------------------------------
  LR test of alpha=0: 394.12   Prob>=LRX2 = 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 expected count for unit increase in X
 e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
   SDofX = standard deviation of X
```

Being a female scientist decreases the expected number of publications by a factor of .75, holding other variables constant (p<0.05, two-tailed test).

A standard deviation increase in the number of years from enrollment to completion of the PhD, about 1.4 years, decreases the expected number of publications by 15%, holding other variables constant (p<0.01, two-tailed test).

#5 Discrete Change

mchange computes the discrete change in the expected count/rate. The changes below are AME's. To compute them using MEM, simply add the option **atmeans**.

```
. mchange

nbreg: Changes in mu | Number of obs = 264

Expression: Predicted number of pub6, predict()

               |    Change    p-value
---------------+----------------------
female         |
 1 Yes vs 0 No |    -1.048      0.036
phd            |
            +1 |     0.861      0.008
           +SD |     0.865      0.008
      Marginal |     0.778      0.004
enroll         |
            +1 |    -0.546      0.001
           +SD |    -0.762      0.001
      Marginal |    -0.588      0.003

Average prediction

    3.896
```


On average, being a female scientist is expected to decrease productivity by 1.0 publication (p<0.05, two-tailed test).

The average effect of an additional year in graduate school decreases productivity by 0.55 publications (p<0.01, two-tailed test).

#6 Expected Count

Use **mtable** to compute the expected count of publications for average men and average women. **mtable** is run 3 times, with the option **below** stacking the current **mtable** results below the previous **mtable** results. Note that **rowname()** is used to label each of the rows.

```
. quietly mtable, at(female=0) stat(ci) atmeans rowname(Men)
. quietly mtable, at(female=1) stat(ci) atmeans rowname(Women) below
. mtable, dydx(female) stat(ci) atmeans rowname(Change) below

Expression: Predicted number of pub6, predict()

          |        mu        ll        ul
----------+------------------------------
      Men |     4.088     3.456     4.719
    Women |     3.083     2.399     3.766
   Change |    -1.005    -1.939    -0.072

Specified values of covariates

          |                                 1.
          |    female       phd    enroll    female
----------+----------------------------------------
    Set 1 |         0      3.18      5.53         .
    Set 2 |         1      3.18      5.53         .
  Current |         .      3.18      5.53      .345
```

For scientists who are average on other characteristics, women are expected to have about 1.0 fewer publications than men (95% CI: -1.94, -0.07).

#7 Predicted Rate and Probabilities

mtable can also calculate the predicted probabilities for specific levels of the outcome variable, as well as the discrete change in the probabilities. This is done using the **pr()** option. The option **roweq** is used to name the different sections of the table rows.

```
. quietly mtable, at(female=0) atmeans roweq(Men_) pr(0(1)5)
. quietly mtable, at(female=1) atmeans roweq(Women_) pr(0(1)5) below
. mtable, dydx(female) stat(est pvalue) atmeans roweq(Change) pr(0(1)5) below

Expression: Marginal effect of Pr(pub), predict(pr(5))

          |         0         1         2         3         4         5
----------+------------------------------------------------------------
Men       |
        1 |     0.166     0.156     0.134     0.111     0.090     0.072
Women     |
        1 |     0.214     0.188     0.150     0.115     0.087     0.065
Change    |
  d Pr(y) |     0.049     0.032     0.016     0.004    -0.003    -0.007
        p |     0.049     0.043     0.036     0.053     0.261     0.093
```



```
Specified values of covariates

          |                                 1.
          |    female       phd    enroll    female
----------+----------------------------------------
    Set 1 |         0      3.18      5.53         .
    Set 2 |         1      3.18      5.53         .
  Current |         .      3.18      5.53      .345
```

For scientists who are average on all other characteristics, women have a higher probability than men of having no publications (p<0.05, two-tailed test), while men have a higher probability of having five publications (p<0.1, two-tailed test).

#8 ZIP Model

The **zip** command with the **inf(*indvars*)** option estimates a Zero-Inflated Poisson Regression Model. You can “inflate” the same set of variables that are used in the PRM portion of the model or an entirely different set of variables. Here we “inflate” using the variable **phd**.

```
. zip pub6 i.female c.phd c.enrol, inf(c.phd) nolog

Zero-inflated Poisson regression                Number of obs     =        264
                                                Nonzero obs       =        212
                                                Zero obs          =         52

Inflation model = logit                         LR chi2(3)        =      48.74
Log likelihood  = -758.0032                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        pub6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pub6         |
      female |
      1_Yes  |  -.1210631   .0710846    -1.70   0.089    -.2603864    .0182602
         phd |   .1400257   .0334849     4.18   0.000     .0743964     .205655
      enroll |  -.1306837   .0250179    -5.22   0.000    -.1797178   -.0816496
       _cons |   1.838966   .1749225    #51   0.000     1.496124    2.181808
-------------+----------------------------------------------------------------
inflate      |
         phd |  -.2383082   .1657934    -1.44   0.151    -.5632572    .0866408
       _cons |  -.7539084   .5332584    -1.41   0.157    -1.799076     .291259
------------------------------------------------------------------------------
```

#9 ZINB Model

We can use the same types of commands for the ZINB. The results are stored using **estimates store**.

```
. zinb pub6 i.female c.phd c.enrol, inf(c.phd) nolog

Zero-inflated negative binomial regression      Number of obs     =        264
                                                Nonzero obs       =        212
                                                Zero obs          =         52

Inflation model = logit                         LR chi2(3)        =      18.91
Log likelihood  = -642.2026                     Prob > chi2       =     0.0003

------------------------------------------------------------------------------
        pub6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
pub6         |
      female |
      1_Yes  |  -.2708994   .1371918    -1.97   0.048    -.5397905   -.0020084
         phd |   .1745669   .0695427     2.51   0.012     .0382657    .3108682
      enroll |  -.1527173    .047032    -3.25   0.001    -.2448984   -.0605362
       _cons |   1.739814   .3498874     4.97   0.000     1.054047     2.42558
-------------+----------------------------------------------------------------
inflate      |
         phd |  -.5440498   .8665119    -0.63   0.530    -2.242382    1.154282
       _cons |  -1.456929   2.082817    -0.70   0.484    -5.539175    2.625316
-------------+----------------------------------------------------------------
    /lnalpha |  -.3514184   .2107589    -1.67   0.095    -.7644982    .0616614
-------------+----------------------------------------------------------------
       alpha |   .7036893   .1483088                      .4655675    1.063602
------------------------------------------------------------------------------

. estimates store estzinb
```

#10 Factor Change

Factor change coefficients can be computed after estimating the ZIP or ZINB models using **`listcoef`**. Since the output is similar, we show only the output for ZINB. The top half of the output, labeled Count Equation, contains coefficients for the factor change in the expected count for those in the Not Always Zero group. The bottom half, labeled Binary Equation, contains coefficients for the factor change in the odds of being in the Always Zero group compared with the Not Always Zero group.

```
. listcoef, help

zinb (N=264): Factor change in expected count

  Observed SD:  4.3103

Count equation: Factor change in expected count for those not always 0

-------------------------------------------------------------------------
             |         b         z    P>|z|       e^b   e^bStdX     SDofX
-------------+-----------------------------------------------------------
      female |
      1_Yes  |   -0.2709    -1.975    0.048     0.763     0.879     0.476
         phd |    0.1746     2.510    0.012     1.191     1.192     1.005
      enroll |   -0.1527    -3.247    0.001     0.858     0.802     1.443
    constant |    1.7398     4.972    0.000         .         .         .
-------------+-----------------------------------------------------------
alpha        |
     lnalpha |   -0.3514         .        .         .         .         .
       alpha |    0.7037         .        .         .         .         .
-------------------------------------------------------------------------
       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 expected count for unit increase in X
 e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
   SDofX = standard deviation of X

Binary equation: factor change in odds of always 0
-------------------------------------------------------------------------
             |         b         z    P>|z|       e^b   e^bStdX     SDofX
-------------+-----------------------------------------------------------
         phd |   -0.5440    -0.628    0.530     0.580     0.579     1.005
    constant |   -1.4569    -0.699    0.484         .         .         .
-------------------------------------------------------------------------
       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
   SDofX = standard deviation of X
```

Among those who have the opportunity to publish, a standard deviation increase PhD prestige increases the expected rate of publication by a factor of 1.2, holding other variables constant (p<0.05, two-tailed test).

A standard deviation increase in PhD prestige decreases the odds of not having the opportunity to publish by a factor of 0.58, although this is not significant (z=-0.63, p=0.53).

#11 Predicted Probabilities and Expected Counts

The ZINB model has 3 types of post-estimation results we are interested in: the expected count, the probability of always being zero, and the predicted probability of various levels of the outcome. By default **`mtable`** computes the expected count. To compute the probability of being always zero, include the **`predict(pr)`** option. To compute the predicted probability of various levels of the outcome variable, include the **`pr()`** option.

```
. quietly mtable, at(phd=(1 4)) atmeans long stat(ci)
. quietly mtable, at(phd=(1 4)) atmeans long stat(ci) noatvar right ///
>     estname(Always0) predict(pr)
. mtable, at(phd=(1 4)) atmeans long stat(ci) noatvar colstub(pr) right pr(0 1 9)

Expression: Pr(pub), predict(pr(9))

          |       phd        mu   Always0       pr0       pr1       pr9
----------+------------------------------------------------------------
       mu |         1     2.339     0.119     0.316     0.182     0.012
       ll |         1     1.470    -0.166     0.111     0.102     0.004
       ul |         1     3.208     0.404     0.521     0.262     0.021
       mu |         4     4.367     0.026     0.155     0.139     0.032
       ll |         4     3.692    -0.068     0.099     0.102     0.026
       ul |         4     5.042     0.120     0.210     0.176     0.038

Specified values of covariates

          |        1.
          |   female    enroll
----------+-------------------
    Set 1 |     .345      5.53
    Set 2 |     .345      5.53
  Current |     .345      5.53
```

An average scientist from a distinguished university is expected to have 4.4 publications (95% CI: 3.69, 5.04), while an average scientist from an adequate university is expected to have 2.3 publications (95% CI: 1.47, 3.21).

For an average scientist from an adequate university, the probability of having no publications because the scientist does not have the opportunity to publish is 0.12 (95% CI: -0.17, 0.40). Thus most of the 0's for average scientists are for those who are "potential publishers".

For an average scientist from a low prestige university, the probability of having no publications, either because the scientist does not have the opportunity to publish or because the scientist is a potential publisher who by chance did not publish, is 0.32 (95% CI: 0.11, 0.52).

For an average scientist from a high prestige university, the probability of having 9 publications is 0.03 (95% CI: 0.026, 0.038).

#12 Discrete Change for Predicted Probabilities and Expected Counts

To compute the discrete change of the different types of predicted values above, we can use **`margins, post`** followed by **`mlincom`**. The results are stacked into an easy to read table with **`mlincom`** by specifying the **`add`** option. Note that estimation results need to be restored before each **`margins, post`** by using **`estimates restore`**.

```
. quietly margins, at(phd=(1 4)) atmeans post
. quietly mlincom 2-1, rowname(Expected_y) stat(all) estname(Change)
. estimates restore estzinb

. quietly margins, at(phd=(1 4)) atmeans predict(pr) post
. quietly mlincom 2-1, rowname(Always_0) stat(all) estname(Change) add
. estimates restore estzinb

. quietly margins, at(phd=(1 4)) atmeans predict(pr(0)) post
. quietly mlincom 2-1, rowname(Pr_y=0) stat(all) estname(Change) add
. estimates restore estzinb

. quietly margins, at(phd=(1 4)) atmeans predict(pr(1)) post
. quietly mlincom 2-1, rowname(Pr_y=1) stat(all) estname(Change) add
. estimates restore estzinb

. quietly margins, at(phd=(1 4)) atmeans predict(pr(9)) post
. mlincom 2-1, rowname(Pr_y=9) stat(all) estname(Change) add
. estimates restore estzinb

             |   Change        se    zvalue    pvalue        ll        ul
-------------+------------------------------------------------------------
  Expected_y |    2.028     0.619     3.278     0.001     0.816     3.241
    Always_0 |   -0.093     0.163    -0.573     0.566    -0.412     0.226
      Pr_y=0 |   -0.162     0.120    -1.343     0.179    -0.398     0.074
      Pr_y=1 |   -0.043     0.046    -0.941     0.347    -0.134     0.047
      Pr_y=9 |    0.020     0.006     3.354     0.001     0.008     0.031
```

For an average scientist, attending a distinguished university compared to an adequate university is expected to increase productivity by slightly over two publications (p<0.01, two-tailed test).

For an average scientist, attending a distinguished university compared to an adequate university does not affect the probability of having no publications as a result of not having the opportunity to publish (z=-0.573, p=0.566).

For an average scientist, attending a high prestige university compared to a low prestige university increases the probability of having 9 publications (95% CI: 0.008, 0.031).

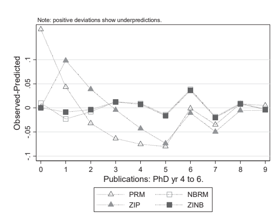
#13 Compare models

`countfit` compares the fit of PRM, NBRM, ZIP, and ZINB, optionally generating a table of estimates, a table of differences between observed and average estimated probabilities, a graph of these differences, and various tests and measures of fit.

```
. countfit pub6 i.female c.phd c.enrol, inf(c.phd) ///
>     graphexport(`pgm'-countfit.`graphtype', replace)

--------------------------------------------------------------------------------
                    Variable |    PRM         NBRM          ZIP          ZINB
-----------------------------+--------------------------------------------------
pub6                         |
                      female |
                      1_Yes  |     0.786        0.754        0.895        0.836
                             |     -3.49        -2.04        -1.57        -1.19
 Prestige of Ph.D. department. |   1.207        1.221        1.151        1.231
                             |      5.85         3.06         4.19         3.19
          Years from BA to P.. |   0.876        0.860        0.879        0.871
                             |     -5.51        -3.14        -5.14        -2.82
                    Constant |     4.630        4.990        6.213        4.532
                             |      9.02         4.76        10.44         4.45
-----------------------------+--------------------------------------------------
lnalpha                      |
                    Constant |                  0.816                     0.735
                             |                  -1.62                     -2.14
-----------------------------+--------------------------------------------------
inflate                      |
                             |
                      female |
                      1_Yes  |                               2.006     2.60e+06
                             |                                2.04         0.02
 Prestige of Ph.D. department. |                             0.759        1.430
                             |                               -1.66         0.49
          Years from BA to P.. |                             1.028        1.370
                             |                                0.23         0.68
                    Constant |                               0.351        0.000
                             |                               -1.24        -0.02
-----------------------------+--------------------------------------------------
Statistics                   |
                       alpha |                  0.816
                           N |       264          264          264          264
                          ll |  -839.781     -642.723     -755.914     -641.263
                         bic |  1701.865     1313.326     1556.436     1332.709
                         aic |  1687.561     1295.446     1527.828     1300.526
--------------------------------------------------------------------------------
                                                                     legend: b/t
Comparison of Mean Observed and Predicted Count

          Maximum       At       Mean
Model     Difference    Value    |Diff|
---------------------------------------------
PRM         0.163         0      0.051
NBRM        0.038         6      0.015
ZIP         0.100         1      0.033
ZINB        0.037         6      0.012

PRM: Predicted and actual probabilities

Count   Actual    Predicted    |Diff|   Pearson
------------------------------------------------
0        0.197       0.034      0.163   205.490
1        0.144       0.100      0.044     4.992
```



```
2        0.129       0.161      0.032     1.688
3        0.121       0.185      0.064     5.777
4        0.095       0.170      0.075     8.815
5        0.053       0.133      0.080    12.712
6        0.091       0.092      0.001     0.003
7        0.023       0.057      0.035     5.546
8        0.042       0.033      0.009     0.589
9        0.023       0.018      0.005     0.371
------------------------------------------------
Sum      0.917       0.983      0.507   245.982
<snip>

Tests and Fit Statistics

PRM             BIC=  1701.865  AIC=  1687.561  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs NBRM       BIC=  1313.326  dif=   388.539  NBRM    PRM   Very strong
                AIC=  1295.446  dif=   392.115  NBRM    PRM
                LRX2=  394.115  prob=    0.000  NBRM    PRM   p=0.000
-------------------------------------------------------------------------
  vs ZIP        BIC=  1556.436  dif=   145.429  ZIP     PRM   Very strong
                AIC=  1527.828  dif=   159.733  ZIP     PRM
                Vuong=   4.358  prob=    0.000  ZIP     PRM   p=0.000
-------------------------------------------------------------------------
  vs ZINB       BIC=  1332.709  dif=   369.155  ZINB    PRM   Very strong
                AIC=  1300.526  dif=   387.035  ZINB    PRM
-------------------------------------------------------------------------
NBRM            BIC=  1313.326  AIC=  1295.446  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs ZIP        BIC=  1556.436  dif=  -243.110  NBRM    ZIP   Very strong
                AIC=  1527.828  dif=  -232.382  NBRM    ZIP
-------------------------------------------------------------------------
  vs ZINB       BIC=  1332.709  dif=   -19.384  NBRM    ZINB  Very strong
                AIC=  1300.526  dif=    -5.080  NBRM    ZINB
                Vuong=   0.834  prob=    0.202  ZINB    NBRM  p=0.202
-------------------------------------------------------------------------
ZIP             BIC=  1556.436  AIC=  1527.828  Prefer  Over  Evidence
-------------------------------------------------------------------------
  vs ZINB       BIC=  1332.709  dif=   223.726  ZINB    ZIP   Very strong
                AIC=  1300.526  dif=   227.302  ZINB    ZIP
                LRX2=  229.302  prob=    0.000  ZINB    ZIP   p=0.000
-------------------------------------------------------------------------
```

Datasets for CDA Exercises

There are the datasets that we provide for exercises.

cda-science4 (cda-scireview4) contains information on the careers of 308 Ph.D. biochemists. (Note that cda-scireview4 has dropped missing cases and therefore contains information on 264 scientists.) This data set is based on data collected by Scott Long with funding from the National Science Foundation. Please note that some variables have been modified.

cda-hsb4 contains 1647 observations on 68 variables from the 1983 High School and Beyond Study.

cda-nes4 contain 2487 observations on 45 variables from the 1992 National Election Study.

cda-addhealth4 contains 2146 observations on 126 variables. It is an extract from the 1994-95 wave of the Add Health public use dataset, and contains information on the hobbies and activities of students aged 12-21, including delinquent behavior and drug/alcohol use. The dataset also includes information about the relationships between the respondents and their parents

The codebooks and data are like those you will encounter in the real world. They attempt to be accurate, but they probably are not. That means that it is up to you to make sure that the descriptions correspond to the distribution of the data in the file. As always in such things, *caveat emptor.*

cda-science4.dta (cda-scireview4): Codebook for Science Data

id	ID Number of scientist
cit1	Number of Citations: PhD year -1 to 1.
cit3	Number of Citations: PhD year 1 to 3.
cit6	Number of Citations: PhD year 4 to 6.
cit9	Number of Citations: PhD year 7 to 9.
enroll	Number of years it took to get a Ph.D. after receipt of B.A.
fel	Prestige of Ph.D. if scientist is not a fellow; prestige of fellowship department if a fellow. Ranges from 0.75 to 5.00. See phd for details on scores.
felclass	Fellow or Ph.D. prestige class: 1: adequate; 2: good; 3: strong; 4: distinguished
fellow	Postdoctoral fellow? (1=yes; 0=no)
female	Female? (1=yes; 0=no)
jobimp	Prestige of first job if first job is as a university faculty member. Ranges from 0.75 to 5.00. See phd for details on prestige scores. Imputed.
jobprst	Prestige of job: 1: adequate; 2: good; 3: strong; 4: distinguished.
mcit3	Mentor's # of citations for 3 year period ending the year of the student's Ph.D.
mcitt	Mentor's total # of citations in 1961.
mmale	Was mentor a male? (1=yes; 0=no)
mnas	Was mentor in National Academy of Science? (1=yes; no)
mpub3	Mentor's 3 year publications.
nopub1	No pubs PhD year -1 to 1? (1=yes; 0=no)
nopub3	No pubs PhD year 1 to 3? (1=yes; 0=no)

nopub6	No pubs PhD year 4 to 6? (1=yes; 0=no)
nopub9	No pubs PhD year 7 to 9? (1=yes; 0=no)
phd	Prestige of PhD department. Ranges from 0.75 to 5.00. All prestige variables can be broken into categories as follows: 0.75 1.99 is adequate; 2.00 2.99 is good; 3.00 3.99 is strong; and 4.00 5.00 is distinguished.
phdclass	Prestige class of Ph.D. department. 1: adequate; 2: good; 3: strong; 4: distinguished
pub1	Number of Publications: PhD year -1 to 1.
pub3	Number of Publications: PhD year 1 to 3.
pub6	Number of Publications: PhD year 4 to 6.
pub9	Number of Publications: PhD year 7 to 9.
pubtot	Total Pubs in 9 years post-Ph.D.
work	Type of first job. 1: Faculty in university; 2: Academic research; 3: College teacher; 4: Industrial research; 5: Administration
workadmn	Work in Administration? (1=yes; 0=no)
workfac	Work as Faculty in University? (1=yes; 0=no)
worktch	Work in Teaching? (1=yes; 0=no)
workuniv	Work in University? (1=yes; 0=no)

cda-hsb4.dta: Codebook for 1983 High School and Beyond Study

id ID number of respondent

sex 1: male; 2: female

male, female

0: no; 1: yes

region Region of country respondent lives in

1: New England 2: Mid Atlantic 3: South Atlantic 4: East South Central
5: West South Central 6: East North Central 7: West North Central 8: Mountain
9: Pacific

hsprog: High School program.

1: general 2: academic 3: agricultural 4: business 5: distributive educ.
6: health 7: home economics 8: technical 9: trade/industrial

algebra2, **geometry**, **trig**, **calc**, **physics**, **chem**: Did you take ...?

0: no; 1: yes

hsgrades: What are your grades in HS?

.5: Mostly below D's 1: Mostly D's 1.5: Mostly C's & D's 2: Mostly C's
2.5: Mostly B's & C's 3: Mostly B's 3.5: Mostly A's & B's 4: Mostly A's

mathabs: Are your math grades mostly A's and B's?

englabs: Are your English grades mostly A's and B's?

busiabs: Are your business grades mostly A's and B's?

0: no; 1: yes

remengl: Have you taken remedial English? **remmath**: Have you taken remedial math?

advengl: Have you taken advanced English? **advmath**: Have you taken advanced math?

0: no; 1: yes

hmwktime: How much time do you spend on homework each week?

1: None is assigned 2: Don't do any 3: Less than 1 hour 4: 1 to 3 hours
5: 3 to 5 hours 6: 5 to 10 hours 7: 10 or more hours

workage: Age you first worked.

11: age 11 or less 12 to 19: ages 12 to 19 respectively21: never worked

hrswork: Hours worked last week. **hrslstyr**: Hours worked per week last year

1: none 2: 1 to 4 3: 5 to 14 4: 15 to 21
5: 22 to 29 6: 30 to 34 7: 35 or more

varsport: Did you participate in varsity sports?

pepclub: In pep club, cheerleading, or other activity?

1: no; 2: participant; 3: leader/officer

livealon: Did you live alone while attending HS? **livedad**: With your father while attending HS?
livemale: With other male guardian? **livemom**: With mother?
livfemal: With other female guardian? **livsibs**: With any brothers or sisters?
livgrand: With your grandparent(s)?

0: no 1: yes

momwkel: Did your mother work while you were in elementary school?

momwkhs: Did your mother work while you were in HS?
momwkpre: Did your mother work before you were in elementary school?

1: no paid work 2: part time work 3: full time work 4: DK 5: NA

dadocc: Father's occupation. **momocc**: Mother's occupation.

1: not living with father 2: clerical 3: craftsman 4: farmer
5: homemaker 6: laborer 7: manager/admin 8: military
9: operative 10: professional 11: advanced professional 12: proprietor
13: protective service 14: sales 15: school teacher 16: service
17: technical 18: never worked 19: DK

daded: Father's education level. **momed**: Mother's education level.

1: not living with father 2: less than HS degree 3: HS or equivalent degree
4: vocational less than 2 years 5: vocational 2 or more years 6: college less than 2 years
7: college 2 or more years 8: college graduate 9: masters degree
10: PhD/MD advanced degree 11: DK

dadhsgrd: Dad graduate high school? **momhsgrd:** Mom graduate high school?
dadcoll: Dad graduate college? **momcoll:** Mom graduate college?

0: no 1:yes

mommonit: Mother monitors your school work? **dadmonit**: Father monitors your school work?

1: yes 2: no 3: NA

talkpar: How often do you talk to your parents?

1: rarely or never 2: less than once a week
3: once or twice a week 4: almost every day

dadplans, momplans: How much did your father/father influence your HS plans?

1: not at all 2: somewhat 3: a great deal

edattain: What educational level do you expect to attain?

momatain: What educational level does your mother expect you to attain?

lowed: What is the lowest educational level you would be satisfied with?

1: Less than HS 2: HS graduate 3: vocational < 2 years 4: vocational 2+ years
5: college < 2 years 6: college 2+years 7: college graduate 8: masters degree
9: PhD/MD degree 10: DK

compserv: Which would you chose if forced into compulsory service?

1: military 2: public service 3: undecided 4: avoid both

earnings: How much have you made this year?

0: None .5: <$1K 2: $1K-$3K 4: $3K-$5K 6: $5K-$7K
8: $7K-$9K 10: $9K-$11K 12: $11K-$13K 14: $13K-$15K 15: $15K+

expenses: How many expenses do you have?

0 .5 1.5 2.5 3.5
4.5 6 8.5 10

netearn: Net earnings this year **sumearn**: Net earnings from last year.

0 100 450 900 1600 2000

agewed: Age you expect to be married. **agekid**: have your first child. **agejob**: have first full time job.

agehome: move out on your own. **ageeduc**: finish your education.

See values when tabulating these variables

age: 15 to 20 is actual years; 21 = 21 years and older.

race: Respondent's race

1: Black 2: White 3: American Indian 4: Asian/Pacific Islander 5: Other

white: White? **black:** Black? **amerind:** American Indian?
asian: Asian? **othrace:** Other race?

0: no 1: yes

origin: Respondent's national origin/country of origin

1: Mexican 2: Cuban 3: Puerto Rican 4: Latin American
5: Afro-American 6: West Indian 7: Alaskan 8: American Indian
9: Chinese 10: Filipino 11: Indian: other 12: Japanese
13: Korean 14: Vietnamese 15: Pacific Islander 16: Asian: other
17: English/Welsh 18: French 19: German 20: Greek
21: Irish 22: Italian 23: Polish 24: Portuguese
25: Russian 26: Scottish 27: Europe-other 28: Fr. Canadian
29: Canadian 30: USA. 31: Other

religion:

1: Baptist 2: Methodist 3: Lutheran 4: Presbyterian
5: Episcopalian 6: Other Protestant 7: Catholic 8: Other Christian
9: Jewish 10: Other 11: None

relProt: Protestant? **relCath:** Catholic? **relJew:** Jewish?
relOth: Other religion? **relNone:** No religion?

0: no 1: yes

religper: Do you consider yourself a religious person?

1: not at all 2: somewhat 3: very much

politics: Political ideology

1: conservative 2: moderate 3: liberal 4: radical 5: none 6: DK

fincome: Family income

3.5 9.5 14 18 22.5 31.5 38

college: Type of college you plan to attend

1: four year college 2: two year college

pubpriv: Do you plan to attend a public or private college?

1: public college 2: private college

instate: Do you plan to attend a college in your state?

0: No 1: Yes

ses: Socioeconomic status

1: low 2: medium 3: high

cda-nes4.dta: Codebook for 1992 National Election Study

caseid: ID number of respondent

prebush, **preclint**, **preperot**: Feelings about each candidate prior to the 1992 presidential election.

postbush, **postclin**, **postpero**: Feelings about each candidate after the 1992 presidential election.

Feeling thermometers range from 0 to 100 with higher score being more favorable. 50 is neutral.

partyid: Political party identification

1: Strong Democrat 2: Weak Democrat 3: Indep-leaning Democrat
4: Independent 5: Indep-leaning Republican 6: Weak Republican
7: Strong Republican 8: Other

abortion: View on abortion

1: Never permitted by law 2: If rape, incest, life threatening 3: If need is established
4: Abortion as personal choice 5: Law should not be involved 6: Other

election: Who do you think you will vote for?

1: Bush 2: Clinton 3: Perot 7: Other

religion: Religious affiliation

1: Protestant 2: Catholic 3: Jewish 4: Other

relProt: Protestant? **relCath:** Catholic? **relJew:** Jewish? **relOth:** Other religion?

0: no 1: yes

age: 17-90 is actual years; 91 = 91 years and older.

marital: Marital status

1: Married, living with spouse 2: Never married 3: Divorced
4: Separated 5: Widowed 6: Unmarried partners

married: Married?

0: no 1: yes

educatio: Education level.

1: 8th grade or less 2: Some High School 3: High school 4: More than 12 years
5: Jr. college degree 6: BA level degrees 7: Advanced degree

collgrad: College graduate? **hsgrad:** High School graduate?

0: no 1:yes

occup: Occupational code.

1: Executive, administrative and managerial 8: Service except protective & household
2: Professional specialty occupations 9: Farming, forestry, and fishing occup.
3: Technicians and related support occup. 10: Precision production, craft and repair

4: Sales occupations 11: Machine operators, assemblers, inspectors
5: Administrative support, including clerical 12: Transport &material moving occup.
6: Private household 13: Handlers, equipment cleaners, laborers
7: Protective service 14: Member of the armed forces

fincome: Family income.

Tabulate variable to see values.

sex: Respondent's sex

1: Male 2: Female

male: Male?
female: Female?

0: no 1:yes

race: Respondent's race

1: White 2: Black 3: American Indian/Alaskan 4: Asian/Pacific Islander

white: White? **black:** Black? **amerind:** American Indian? **asian:** Asian?

0: no 1: yes

didvote: Did you vote this November? **regvote**: Were you registered to vote?

0: No 1: Yes

regvote: Were you registered to vote?

0: No 1: Yes 6: Not required

presvote: Presidential vote. **prefvote**: Did not vote, but preferred

1: Bush 2: Clinton 3: Perot 7: Other

canparty: Which party(ies) did the candidate you contributed to belong to?

whichpar: To which party did you give money?

1: Republican 2: Both 3: Democratic 7: Other

campaign*:	Did you talk to people about voting for or against a party or candidate?
contact:	Were you contacted by any person intent on showing you who to vote for?
support*:	Did you wear or display a campaign button, sticker, or sign?
attend*:	Did you attend any political meetings, rallies etc. in support of a candidate?
enlist:	Did anyone enlist you to attend a political rally, meeting, speech, or dinner?
partywrk*:	Did you do any work for one of the parties or candidates?
askwork:	Did anyone ask you to do any work for one of the parties or candidates?
taxretur*:	Did you make a political contribution on your income tax return this year?
fundcam:*	Did you give any money to an individual candidate running for public office?
fundpart*:	Did you give any money to a political party during this election year?
fundgrp*:	Did you give money to any other group that supported or opposed candidates?
contvote:	This year, did anyone talk to you about registering or getting out to vote?
mailfund:	Did you receive any mail requests asking you to contribute to a party/candidate?
contmail:	Did you contribute any money because of the mail you received?
phonfund:	Did you receive any phone requests asking you to contribute to a party/candidate?
contphon:	Did you contribute any money because of the phone calls you received?
persfund:	Did you receive any personal requests asking you to contribute to a party/candidate?
contpers:	Did you contribute any money because of the personal contacts you received?

0: no 1: yes

* These variables used to create **polacts** using code in Stata Guide.

alotmail: How many mail requests for contributions to a candidate/party did you receive?

alotphon: How many phone requests for contributions to a candidate/party did you receive?

persalot: How many personal requests for contributions to a candidate/party did you receive?

1: not very many   5: quite a few

cda-addhealth4: Codebook for 1994-95 Add Health Public Data extract

Note: missing values for all variables have these meainings

.d:Don't know   .n: Not applicable   .r: Refused   .s: Skip

caseid: Respondent's case ID number

gswgt1: Grand sample weight

cluster2: Sample cluster, stratum 2

The syntax for setting the survey weights is:

```
svyset, clear
svyset [pweight=gswgt1], strata(cluster2)
```

age: Respondent's age (calculation includes months; ranges from 11.4167 to 20.1667).

sex: Respondent's sex   1: Male   2: Female

male: Male?   **female:** Female?   0: no   1:yes

hispanic: Hispanic origin?   **white:** Non-Hispanic white?

black: Non-Hispanic Black or African American?   **asian:** Non-Hispanic Asian or Pacific Islander?

othrace: Another race?

0: No   1: Yes

bornUS: Respondent born in the United States?

0: No   1: Yes

hobbies: During the past week, how many times did you do hobbies, such as collecting baseball cards, playing a musical instrument, reading, or doing arts and crafts?

videos: During the past week, how many times did you watch television or videos, or play video games?

skating: During the past week, how many times did you go roller-blading, roller-skating, skate-boarding, or bicycling?

sport: During the past week, how many times did you play an active sport, such as baseball, softball, basketball, soccer, swimming, or football?

exercise: During the past week, how many times did you do exercise, such as jogging, walking, karate, jumping rope, gymnastics or dancing?

friends: During the past week, how many times did you just hang out with friends?

0: None   1: 1-2 times   2: 3-4 times   3: 5+ times

hrstv: How many hours a week do you watch television?

hrsvideo: How many hours a week do you watch videos?

hrscomp: How many hours a week do you play video or computer games?

hrsradio: How many hours a week do you listen to the radio?

Continuous variables starting at 0,

brthctrl: If you wanted to use birth control, how sure are you that you could stop yourself and use birth control once you were highly aroused or turned on?

1: Very unsure   2: Moderately unsure   3: Neither sure or unsure
4: Moderately sure   5: Very sure   6: Never want to use birth control

intlgnce: Compared with other people your age, how intelligent are you?

1: Moderately below average   2: Slightly below average   3: About average
4: Slightly above average   5: Moderately above average   6: Extremely above average

<u>How often in the past week did you experience the following?</u>

bothered: You were bothered by things that usually don't bother you.

appetite: You didn't feel like eating, your appetite was poor.

blues: You felt that you could not shake off the blues, even with help from your family and your friends.

mindfoc: You had trouble keeping your mind on what you were doing.

depressed: You felt depressed.   **tired:** You felt that you were too tired to do things.

failure: You thought your life had been a failure.   **fearful:** You felt fearful.

talkless: You talked less than usual.   **lonely:** You felt lonely.

unfrndly: People were unfriendly to you.   **sad:** You felt sad.

dislike: You felt that people disliked you.   **getstart:** It was hard to get started doing things.

living: You felt life was not worth living.

0: Never   1: Some   2: A lot   3: Mostly

goodas: You felt that you were just as good as other people.

hopeful: You felt hopeful about the future.   **happy:** You were happy.

enjlife: You enjoyed life.

0: Mostly   1: A lot   2: Some   3: Never

depress: Depression scale, above 19 items added together.

momeduc: How far in school did your mom go?

dadeduc: How far in school did your dad go?

1: eighth grade or less   2: more than 8th grade, but not HS grad
3: business/trade/vocational instead of HS   4: high school graduate
5: completed a GED   6: business/trade/vocational after HS
7: went to college, but did not graduate   8: graduated from a college/univ
9: prof. training beyond a 4yr college/univ.   10: Never went to school.
11: Went, but R doesn't know what level.   12: R doesn't know if went to school.

momcoll: Mom graduated from college?   **dadcoll:** Dad graduated from college?

momhsgrd: Mom graduated from high school?   **dadhsgrd:** Dad graduated from high school?

0: No   1: Yes

mombrnUS: Was your mom born in the United States?

dadbrnUS: Was your dad born in the United States?

0: No   1: Yes

<u>Which of the things listed on this card have you done with your mother in the past 4 weeks?</u>

momshop: gone shopping

momsport: played a sport

momrel: gone to a religious service or church-related event

momlife: talked about someone you're dating, or a party you went to

mommovie: gone to a movie, play, museum, concert, or sports event

momprob: had a talk about a personal problem you were having

mombehav: had a serious argument about your behavior

momgrades: talked about your school work or grades

momproj: worked on a project for school

momoth: talked about other things you're doing in school

momnone: didn't do any of these things with your mom

0: No   1: Yes

actsmom: Number of above activities respondent did with mom, except talk about personal problems, argue about behavior, and talk about grades (range 0-7)

<u>Which of these things have you done with your father in the past 4 weeks?</u>

dadshop: gone shopping

dadsport: played a sport

dadrel: gone to a religious service or church-related event

dadlife: talked about someone you're dating, or a party you went to

dadmovie: gone to a movie, play, museum, concert, or sports event

dadprob: had a talk about a personal problem you were having

dadbehav: had a serious argument about your behavior

dadgrades: talked about your school work or grades

dadproj: worked on a project for school

dadoth: talked about other things you're doing in school

dadnone: didn't do any of these things with your dad

0: No   1: Yes

actsdad: Number of above activities respondent did with dad, except talk about personal problems, argue about behavior, and talk about grades (range 0-7)

momrshp: Overall, you are satisfied with your relationship with your mother.

dadrshp: Overall, you are satisfied with your relationship with your father.

0: No   1: Yes

momcare: How much do you think your mom cares about you?

dadcare: How much do you think your dad cares about you?

adultcare: How much do you feel that adults care about you?

tchrcare: How much do you feel that your teachers care about you?

prntscare: How much do you feel that your parents care about you?

frndscare: How much do you feel that your friends care about you?

1: Not at all   2: Very little   3: Somewhat   4: Quite a bit   5: Very much   6:DNApply

<u>How much do you agree with the following statements?</u>

goodqual: You have a lot of good qualities.

proud: You have a lot to be proud of.

likeself: You like yourself just the way you are.

doright: You feel like you are doing everything just about right.

accepted: You feel socially accepted.

loved: You feel loved and wanted.

1: Strongly disagree   2: Disagree   3: Neither   4: Agree   5: Strongly agree

esteem: Self-esteem scale, six above items added together

abpledge: Have you taken a public or written pledge to remain a virgin until marriage? (0: No; 1: Yes)

havesex: Have you ever had sexual intercourse? (0: No; 1: Yes)

smokereg: Have you ever smoked cigarettes regularly, that is, at least 1 cigarette every day for 30 days?

0: No   1: Yes

dayssmok: During the past 30 days, on how many days did you smoke cigarettes? (range 0-30)

numcigs: During the past 30 days, on days you smoked, how many cigarettes did you smoke daily? (range 0-60)

numdrinks: Think of all the times you have had a drink during the past 12 months. How many drinks did you usually have each time? (range: 0-90)

daysdrink: During the past 12 months, on how many days did you drink alcohol?

drink5: Over the past 12 months, on how many days did you drink five or more drinks in a row?

daysdrunk: Over the past 12 months, on how many days have you gotten drunk or "very, very high" on alcohol?

1: Never   2: 1 to 2 days   3: Once a month   4: A few times a month

5: Once a week   6: A few times a week   7: Daily

potlife: During your life, how many times have you used marijuana? (range 0-900)

potlstmo: During the past 30 days, how many times did you use marijuana? (range 0-800)

<u>In the past 12 months, how often did you ...</u>

graffiti: paint graffiti or signs on someone else's property or in a public place?

damage: deliberately damage property that didn't belong to you?

lieprnts: lie to your parents or guardians about where you had been or whom you were with?

shoplift: take something from a store without paying for it?

fight: get into a serious physical fight?

injureoth: hurt someone badly enough to need bandages or care from a doctor or nurse?

runaway: run away from home?

stealcar: drive a car without its owner's permission?

stealGT50: steal something worth more than $50?

burglar: go into a house or building to steal something?

weapon: use or threaten to use a weapon to get something from someone?

selldrugs: sell marijuana or other drugs?

stealLT50: steal something worth less than $50?

grpfight: take part in a fight where a group of your friends was against another group?

rowdy: act loud, rowdy, or unruly in a public place?

0: None   1: 1-2 times   2: 3-4 times   3: 5+ times

delinq: Number of the above items respondent did at least once in the last 12 months. (range 0-15)

leavehome: How much do you feel that you want to leave home?

famundrst: How much do you feel that people in your family understand you?

famfun: How much do you feel that you and your family have fun together?

famattn: How much do you feel that your family pays attention to you?

1: Not at all   2: Very little   3: Somewhat
4: Quite a bit   5: Very much   6: Does not apply

relig: What is your religion?

0: none   1: Adventist   2: African Methodist Episcopal, AME Zion, CME
3 Assemblies of God   4: Baptist   5: Christian Church (Disciples of Christ)
6: Christian Science   7: Congregational   8: Episcopal
9: Friends/Quaker   10: Holiness   11: Jehovah's Witness
12: Latter Day Saints (Mormon)   13: Lutheran   14: Methodist
15: National Baptist   16: Pentecostal   17: Presbyterian
18: United Church of Christ   19: other Protestant   20: Baha'i

21: Buddhist   22: Catholic   23: Eastern Orthodox
24: Hindu   25: Islam, Muslim   26: Jewish
27: Unitarian   28: other religion

relProt: Protestant?   **relCath**: Catholic?   **relJew:** Jewish?

relOth: Other religion?   **relNone:** No religion?

0: No   1: Yes

service: In the past 12 months, how often did you attend religious services?

1: Never   2: Less than once a month
3: Less than once a week   4: Once a week or more

pray: How often do you pray?

1: Never   2: Less than once a month   3: Once a month
4: Once a week   5: Once a day

wantcoll: On a scale of 1 to 5, where 1 is low and 5 is high, how much do you want to go to college?

likelycol: On a scale of 1 to 5, where 1 is low and 5 is high, how likely is it that you will go to college?

1: Low   2   3   4   5: High

AHvocab: Add Health Picture Vocabulary Test standardized score (range 16-137)

RAWvocab: Add Health Picture Vocabulary Test raw score (range 4-87)

*higher score indicates better performance

CDAicpsr2017 Lab Guide 20917-05-04