

ICPSR Categorical Data Analysis: Lab Guide for Stata

Shawna Rohrman and J. Scott Long – June 2009

1. The Lab Guide is divided into sections corresponding to class lectures. Each section includes both a review, which everyone should complete and an exercise, which is intended to get you started working more creatively with the commands.
2. We have provided a number of data sets, which you are welcome to use for the exercises. These include: `icpsr_nes3.dta`, `icpsr_science3.dta`, `icpsr_hsb3.dta`, and `icpsr_addhealth3.dta`. Codebooks are at the end of the Lab Guide. The codebooks include suggestions for variables you can use in the various exercises. This will save you time, giving you more time to try the statistical methods and require less time cleaning data.
3. For each substantive topic from lectures, there are computer assignments that are divided into two parts. The **review** quickly reviews the commands you will need using a specified dataset with variables we have selected. *Please* always start your assignment by completing the review. Next, there is a more open ended exercise that challenges you to use the methods in more creative ways. You can use your own dataset for the exercise, although you can also use the datasets mentioned in point 2 above.
4. As you work through the exercises (not the reviews), feel free to skip questions and explore commands in ways we have not suggested.
5. We provide a few examples of interpretation in the review sections. These are in shaded boxes.
6. If you want feedback on interpretation, write a paragraph or two and give this to us along with the relevant output from your log-file.
7. Although the command window can be used for exploring new commands, **exercises should always 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.

Contents

Section 1: Linear Regression – REVIEW	3
Section 1: Linear Regression – EXERCISE	7
Section 2: Models for Binary Outcomes – REVIEW	8
Section 2: Models for Binary Outcomes – EXERCISE	16
Section 3: Testing and Assessing Fit – REVIEW.....	17
Section 3: Testing and Assessing Fit – EXERCISE	22
Section 4: Models for Ordinal Outcomes – REVIEW.....	23
Section 4: Models for Ordinal Outcomes – EXERCISE	30
Section 5: Models for Nominal Outcomes – REVIEW.....	31
Section 5: Models for Nominal Outcomes – EXERCISE.....	41
Section 6: Models for Count Outcomes – REVIEW	42
Section 6: Models for Count Outcomes – EXERCISE.....	51
Appendix 1: Codebooks.....	59

Section 1: Linear Regression – REVIEW

The commands from this section are in `icda01a-linear-rev.do`.

___1.1R) Set-up your do-file.

```
capture log close
log using icda01a-linear-rev, replace text

// program:      icda01a-linear-rev.do
// task:         Review 1 - Linear Regression
// project:      CDA
// author:       your name \ today's date

// #1
// program setup
```

```
version 10
clear all
set linesize 80
matrix drop _all
```

___1.2R) Load the Data.

```
use icpsr_scireview3, clear
```

___1.3R) Examine the Data and Select Variables. First, describe the dataset to see a list of variables and some information about the dataset:

```
describe
```

Produce this output:

```
. describe
```

```
Contains data from icpsr_scireview3.dta
  obs:                264                Biochemist data for review -
                                         Some data artificially
                                         constructed
  vars:                 34                18 May 2009 10:14
  size:                16,632 (99.9% of memory free)  (_dta has notes)
```

```
-----
variable name      storage  display  value  variable label
                  type     format   label
-----
id                 float    %9.0g   ID number
cit1               int      %9.0g   Citations in PhD yrs -1 to 1
::output deleted::
jobprst           float    %9.0g   prstlb   Rankings of University Job.
* indicated variables have notes
-----
```

```
Sorted by:  jobprst
```

Use `keep` to select the dependent variable `totpub` and the three independent variables, `faculty`, `enrol`, and `phd`, which we use in the regression models later.

```
keep totpub faculty enrol phd
```

1.4R) Drop cases with missing data and verify. Use `misschk` to review the missing data. Then, use the `dropmiss` command to drop all the observations with missing data on your variables. Be sure to use the `obs` option, or the command will delete all variables with missing data (probably dropping all of your variables). The `force` option is necessary if your data have changed and have not been saved; since we dropped all variables except `totpub`, `faculty`, `enroll`, and `phd`, we need to use the `force` option.

```
misschk
dropmiss, force obs
```

Finally, you'll describe the data to verify that you've kept only the variables you want and deleted the missing:

```
describe
tab1 faculty enrol phd, m
```

Produce this output:

```
. misschk
```

Variables examined for missing values

#	Variable	# Missing	% Missing
1	enrol	0	0.0
2	phd	0	0.0
3	faculty	0	0.0
4	totpub	0	0.0

Missing for which variables?	Freq.	Percent	Cum.
_____	264	100.00	100.00
Total	264	100.00	

Missing for how many variables?	Freq.	Percent	Cum.
0	264	100.00	100.00
Total	264	100.00	

```
. dropmiss, force obs
(0 observations deleted)
```

```
. describe
```

```
Contains data from icpsr_scireview3.dta
obs:                264                Biochemist data for review -
                                         Some data artificially
                                         constructed
vars:                 4                18 May 2009 10:14
size:                3,960 (99.9% of memory free)  (_dta has notes)
```

variable name	storage type	display format	value label	variable label
---------------	--------------	----------------	-------------	----------------

```

enrol      byte    %9.0g          Years from BA to PhD.
phd        float   %9.0g          Prestige of Ph.D. department.
faculty    byte    %9.0g          fac1b1    1=Faculty in University
totpub     byte    %9.0g          Total Pubs in 9 Yrs post-Ph.D.

```

Sorted by:

Note: dataset has changed since last saved

```
. tab1 faculty enrol phd, miss
```

-> tabulation of faculty

1=Faculty in University	Freq.	Percent	Cum.
0_NotFac	123	46.59	46.59
1_Faculty	141	53.41	100.00
Total	264	100.00	

:: output deleted ::

___1.5R) Regression. Specifying a model is simple, with the dependent variable listed *first*, followed by independent variables. This command:

```
regress totpub faculty enrol phd
```

Produces this output:

```
. regress totpub faculty enrol phd
```

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

totpub	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
faculty	5.227261	1.297375	4.03	0.000	2.672561 7.78196
enrol	-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

___1.6R) Obtain Standardized Coefficients. `listcoef` lists the estimated coefficients for a variety of regression models. The `help` option includes details on the meaning of each coefficient. This command:

```
listcoef, help
```

Produces this output:

```
. listcoef, help
```

```
regress (N=264): Unstandardized and Standardized Estimates
```

```
Observed SD: 11.003911
```

SD of Error: 10.437764

totpub	b	t	P> t	bStdX	bStdY	bStdXY	SDofX
faculty	5.22726	4.029	0.000	2.6125	0.4750	0.2374	0.4998
enrol	-1.17488	-2.631	0.009	-1.6955	-0.1068	-0.1541	1.4432
phd	1.50690	2.339	0.020	1.5147	0.1369	0.1377	1.0052

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

___1.7R) Interpretation. Interpret one unstandardized and one standardized coefficient.

For a unit increase in the prestige of the doctoral department, the number of total publication is expected to increase by 1.5, holding other variables constant.

For a standard deviation increase in the length of time between enrollment and graduation (about 1.5 years), the number of total publication is expected to decrease by 1.7, holding other variables constant.

On average, scientists who take faculty positions have about a half a standard deviation more publications than scientists who do not take faculty positions.

___1.8R) Close Log File and Exit Do File.

```
log close  
exit
```

Section 1: Linear Regression – EXERCISE

The file `icda01b-linear-ex.do` contains an outline of this Exercise.

- ___ 1.1) Set-up your do-file.
- ___ 1.2) Load the data.
- ___ 1.3) Examine the data and select variables. Choose one continuous dependent variable and at least three independent variables (make sure one is binary and one is continuous) to use in a regression analysis.
- ___ 1.4) Drop cases with missing data and verify.
- ___ 1.5) Run an OLS regression.
- ___ 1.6) Obtain x-standardized, y-standardized, and fully standardized coefficients.
- ___ 1.7) Interpretation (you should write the answers as if they were part of a research paper)
 - ___ a. Interpret at least one unstandardized coefficient.
 - ___ b. Interpret at least one x-standardized, one y-standardized and one fully standardized coefficient.
- ___ 1.8) Close log and exit do-file

Section 2: Models for Binary Outcomes – REVIEW

For details about binary models and related Stata commands, see Chapter 4 of L&F (2005). The file `icda02a-binary-rev.do` contains these Stata commands.

___2.1R) Set-up your do-file.

```
capture log close
log using icda02a-binary-rev, replace text

// program:    icda02a-binary-rev.do
// task:      Review 2 - Binary Regression
// project:   CDA
// author:    your name \ today's date

// #1
// program setup

version 10
clear all
set linesize 80
matrix drop _all
```

___2.2R) Load the Data.

```
use icpsr_scireview3, clear
```

___2.3R) Examine data, select variables, drop missing, and verify.

```
describe
keep faculty fellow phd mcit3 mnas
misschk
dropmiss, force obs
```

```
:: output deleted ::
```

```
summarize
tab1 faculty fellow phd mcit mnas
```

Produces the following output:

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
fellow	264	.4128788	.4932865	0	1
mcit3	264	20.71591	25.44536	0	129
mnas	264	.0833333	.2769103	0	1
phd	264	3.181894	1.00518	1	4.66
faculty	264	.5340909	.4997839	0	1

```
. tab1 faculty fellow phd mcit mnas
```

```
-> tabulation of faculty
```

1=Faculty in University	Freq.	Percent	Cum.
-------------------------------	-------	---------	------

0_NotFac	123	46.59	46.59
1_Faculty	141	53.41	100.00

Total	264	100.00	

:: output deleted ::

___2.4R) **Binary Logit Model.** As with regression, the dependent variable is listed first. A probit model is run by simply changing **logit** to **probit**. This command:

```
logit faculty fellow phd mcit3 mnas
```

Produces this output:

```
. logit faculty fellow phd mcit3 mnas
```

```
Iteration 0:  log likelihood = -182.37674
Iteration 1:  log likelihood =  -164.019
Iteration 2:  log likelihood = -163.55936
Iteration 3:  log likelihood = -163.55534
Iteration 4:  log likelihood = -163.55534
```

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

faculty	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
fellow	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	.3639082	.5571229	0.65	0.514	-.7280327 1.455849
_cons	-.5806031	.4498847	-1.29	0.197	-1.462361 .3011547

___2.5R) **Store the estimation results.** It is sometimes handy to store estimation results to call up later in table format. You can do this using the command **eststo**. Here we store the estimates with the name **estlogit**.

```
eststo estlogit
```

___2.6R) **Predicted Probabilities.** We can compute and plot predicted probabilities for our observed data. Here we pick the name **prlogit** for the new variable that contains predicted values. Note that a predicted score is calculated for each case in the sample. These commands:

```
predict prlogit
label var prlogit "Logit: Predicted Probability"
sum prlogit
dotplot prlogit
graph export icda02a-binary-rev-fig1.emf , replace
```

Produce this output:

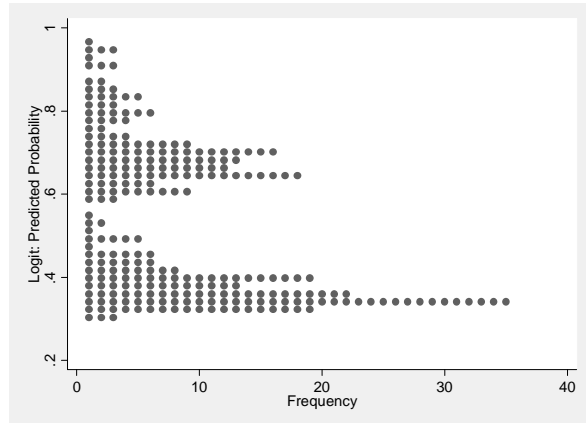
```
. predict prlogit
(option p assumed; Pr(faculty))

. label var prlogit "Logit: Predicted Probability"
```

```
. sum prlogit
```

Variable	Obs	Mean	Std. Dev.	Min	Max
prlogit	264	.5340909	.1828654	.3035647	.9665072

```
. dotplot prlogit
```



2.7R) Predict Specific Probabilities. `prvalue` computes the predicted value of our dependent variable given a set of values for our independent variables. Use the `x(variables = values)` option to set the values at which the variables will be examined. Use `rest(mean)` to set the other independent variables at their means. Note that whereas the command `predict` creates a new variable that contains the predicted score for each case in the sample, `prvalue` computes a predicted probability for a case with certain characteristics and hence does not create a new variable. This command:

```
prvalue , x(fellow=1 mnas=1) rest(mean)
```

Produces this output:

```
. prvalue , x(fellow=1 mnas=1) rest(mean)
```

```
logit: Predictions for faculty
```

```
Confidence intervals by delta method
```

		95% Conf. Interval	
Pr(y=1_Facult x):	0.7786	[0.5931,	0.9642]
Pr(y=0_NotFac x):	0.2214	[0.0358,	0.4069]

	fellow	phd	mcit3	mnas
x=	1	3.1818939	20.715909	1

The predicted probability of obtaining a faculty position for a scientist with both a fellowship and a mentor who was a member of the national association of scientists (NAS) and who is otherwise average is 0.78.

2.8R) Table of predicted probabilities. A table of predicted probabilities for different combinations of values of independent variables can be obtained with the command `prtab`. This command:

```
prtab fellow mnas
```

Produces this output:

```
. prtab fellow mnas
```

```
logit: Predicted probabilities of positive outcome for faculty
```

```
-----+-----
Postdoctoral fellow: |      Mentor NAS:
1=y,0=n.             |      1=yes,0=no.
                    |      0_OutNAS   1_InNAS
-----+-----
0_NoFellow           |      0.4119    0.5019
1_Fellow             |      0.7097    0.7786
-----+-----
```

```
          fellow      phd      mcit3      mnas
x=   .41287879  3.1818939  20.715909  .08333333
```

2.9R) Discrete Change. `prvalue` can also be used to compute the discrete change at specific values of the independent variables when the `save` and `dif` options are used. By default, discrete changes are calculated holding all other variables at their mean. Note that I use the `quietly` option to suppress the output of the first `prvalue` command since the information is repeated when the second `prvalue` command is executed. These commands:

```
quietly prvalue , x(fellow=1) rest(mean) save label(Fellow)
prvalue , x(fellow=0) rest(mean) dif label(NotFellow)
```

Produce this output:

```
. quietly prvalue , x(fellow=1) rest(mean) save label(Fellow)
. prvalue , x(fellow=0) rest(mean) dif label(NotFellow)
```

```
logit: Change in Predictions for faculty
```

```
Confidence intervals by delta method
```

	NotFellow	Fellow	Change	95% CI for Change
Pr(y=1_Facult x):	0.4192	0.7159	-0.2967	[-0.4156, -0.1777]
Pr(y=0_NotFac x):	0.5808	0.2841	0.2967	[0.1777, 0.4156]

	fellow	phd	mcit3	mnas
Current=	0	3.1818939	20.715909	.08333333
Saved=	1	3.1818939	20.715909	.08333333
Diff=	-1	0	0	0

A scientist who receives a post-doctoral fellowship has a .30 higher probability of being on faculty at a university than a scientist who does not receive a fellowship, holding other variables at their mean values. This difference is significant (95% CI: 0.18, 0.42).

2.10R) Discrete Change 2. `prchange` computes the discrete change for all independent variables but does not calculate a confidence interval for the discrete change. As with `prvalue`, values for specific independent variables can be set using the `x()` and `rest()` options. The `help` option provides a key. This command:

```
prchange , rest(mean) help
```

Produces this output:

```
. prchange , rest(mean) help
```

```
logit: Changes in Probabilities for faculty
```

	min->max	0->1	--1/2	--sd/2	MargEfct
fellow	0.2967	0.2967	0.3003	0.1516	0.3097
phd	-0.0576	-0.0154	-0.0158	-0.0159	-0.0158
mcit3	0.4775	0.0051	0.0051	0.1292	0.0051
mnas	0.0881	0.0881	0.0899	0.0250	0.0902

	0_NotFac	1_Facult
Pr(y x)	0.4526	0.5474

	fellow	phd	mcit3	mnas
x=	.412879	3.18189	20.7159	.083333
sd(x)=	.493287	1.00518	25.4454	.27691

Pr(y|x): probability of observing each y for specified x values
Avg|Chg|: average of absolute value of the change across categories
Min->Max: change in predicted probability as x changes from its minimum to its maximum
 0->1: change in predicted probability as x changes from 0 to 1
 --1/2: change in predicted probability as x changes from 1/2 unit below base value to 1/2 unit above
 --sd/2: change in predicted probability as x changes from 1/2 standard dev below base to 1/2 standard dev above
MargEfct: the partial derivative of the predicted probability/rate with respect to a given independent variable

A standard deviation increase in the number of mentor's citations (centered around the mean) increases the predicted probability of obtaining a faculty position by .13 for a scientist who is average on all other characteristics.

___2.11R) Plot predicted probabilities. It is often useful to compute predicted probabilities across the range of a continuous variable for two groups and then plot these. We do this using **prgen**. **prgen** generates a series of new variables containing predicted values and confidence intervals. The new variables begin with the stem you indicate in the option **gen()**. These commands:

```
prgen mcit3, x(fellow=1) rest(mean) from(0) to(130) gen(fel1) gap(5) ci  
prgen mcit3, x(fellow=0) rest(mean) from(0) to(130) gen(fel0) gap(5) ci  
label var fel1p1 "Fellow"  
label var fel0p1 "Not a Fellow"
```

```
graph twoway ///  
  (rarea fel1plub fel1pll fel1x, color(gs10)) ///  
  (rarea fel0plub fel0pll fel0x, color(gs10)) ///  
  (connected fel1p1 fel1x, lpattern(dash) msize(zero)) ///  
  (connected fel0p1 fel0x, lpattern(solid) msize(zero)), ///  
  legend(on order(3 4)) ///  
  ylabel(0(.25)1) ytitle("Pr(Fellow)") ///  
  xlabel(0(10)130) xtitle("Mentor's # of Citations") ///  
  title("Predicted Probability of Having a Faculty Postion")
```

```
graph export icda02a-binary-rev-fig2.emf , replace
```

Produce this output:

```
. prgen mcit3, x(fellow=1) rest(mean) from(0) to(130) gen(fel1) gap(5) ci
```

logit: Predicted values as mcit3 varies from 0 to 130.

```
      fellow      phd      mcit3      mnas
x=      1  3.1818939  20.715909  .08333333
```

```
. prgen mcit3, x(fellow=0) rest(mean) from(0) to(130) gen(fel0) gap(5) ci
```

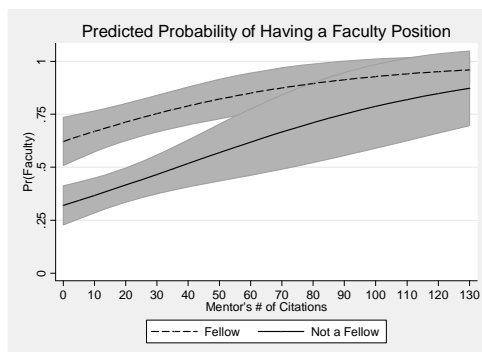
logit: Predicted values as mcit3 varies from 0 to 130.

```
      fellow      phd      mcit3      mnas
x=      0  3.1818939  20.715909  .08333333
```

```
. label var fel1p1 "Fellow"
```

```
. label var fel0p1 "Not a Fellow"
```

```
. graph twoway ///
>   (rarea fel1plub fel1p1lb fel1x, color(gs10)) ///
>   (rarea fel0plub fel0p1lb fel0x, color(gs10)) ///
>   (connected fel1p1 fel1x, lpattern(dash) msize(zero)) ///
>   (connected fel0p1 fel1x, lpattern(solid) msize(zero)), ///
>   legend(on order(3 4)) ///
>   ylabel(0(.25)1) ytitle("Pr(Fellow)") ///
>   xlabel(0(10)130) xtitle("Mentor's # of Citations") ///
>   title("Predicted Probability of Having a Faculty Postion")
```



For an average scientist, receiving a fellowship increases the probability of being employed as a faculty member when mentor's citations are below 50 or so. At higher levels of mentor's citations, there is no significant difference between scientists who received a fellowship and those who did not. In general, the probability of being employed as a faculty member increases as the number of mentor's citations increases and receiving a fellowship appears to be particularly useful when mentor's citations are low.

2.12R) Computing Odds Ratios. The factor change in the odds as well as the standardized factor change can be obtained with the command `listcoef`. Note that `listcoef` can also be run after estimating a

probit model but odds ratios cannot be computed for this model. Instead standardized beta coefficients are listed.

```
listcoef , help
```

Produces this output:

```
. listcoef , help
```

```
logit (N=264): Factor Change in Odds
```

```
Odds of: 1_Facult vs 0_NotFac
```

faculty	b	z	P> z	e^b	e^bStdX	SDofX
fellow	1.25015	4.517	0.000	3.4909	1.8528	0.4933
phd	-0.06372	-0.433	0.665	0.9383	0.9380	1.0052
mcit3	0.02062	2.893	0.004	1.0208	1.6897	25.4454
mnas	0.36391	0.653	0.514	1.4389	1.1060	0.2769

```

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 gaining a faculty position by a factor of 3.5, holding other variables constant. A standard deviation increase in mentor's citations (about 25) increases the odds of gaining a faculty position by a factor of 1.7.

2.13R) Compare the coefficients from Logit and Probit. Run a probit model using the same variables and store the results. Use `eststo` to store the estimation results and `esttab` to list these side-by-side with the estimation results from the logit model. Note that the logit estimates are around 1.7 times as large as the probit estimates. Why is this? These commands:

```
probit faculty fellow phd mcit3 mnas
eststo estprobit
esttab estlogit estprobit , mtitles(logit probit)
```

Produce these results:

```
. probit faculty fellow phd mcit3 mnas
```

```
Iteration 0: log likelihood = -182.37674
Iteration 1: log likelihood = -163.98754
Iteration 2: log likelihood = -163.73877
Iteration 3: log likelihood = -163.73838
```

```

Probit regression                               Number of obs   =           264
LR chi2(4)                                     =           37.28
Prob > chi2                                     =           0.0000
Pseudo R2                                       =           0.1022
Log likelihood = -163.73838

```

faculty	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
---------	-------	-----------	---	------	----------------------

fellow	.763915	.1675687	4.56	0.000	.4354863	1.092344
phd	-.0392676	.0897914	-0.44	0.662	-.2152556	.1367203
mcit3	.0118642	.003994	2.97	0.003	.0040362	.0196922
mnas	.2299521	.3252353	0.71	0.480	-.4074975	.8674016
_cons	-.3450294	.2743016	-1.26	0.208	-.8826506	.1925919

```
. eststo estprobit
. esttab estlogit estprobit , mtitles(logit probit)
```

	(1) logit	(2) probit
fellow	1.250*** (4.52)	0.764*** (4.56)
phd	-0.0637 (-0.43)	-0.0393 (-0.44)
mcit3	0.0206** (2.89)	0.0119** (2.97)
mnas	0.364 (0.65)	0.230 (0.71)
_cons	-0.581 (-1.29)	-0.345 (-1.26)
N	264	264

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

___2.14R) Close Log File and Exit Do File.

```
log close
exit
```

Section 2: Models for Binary Outcomes – EXERCISE

The file `icda02b-binary-ex.do` contains an outline of this exercise.

___ 2.1) Set-up your do-file

___ 2.2) Load your data.

___ 2.3) Examine the data and select your variables. Choose one binary dependent variable and at least three independent variables (make sure one is binary and one is continuous). Drop cases with missing data and verify.

___ 2.4) Estimate a binary logit model

___ 2.5) Store the results of the logistic regression.

___ 2.6) Predict probabilities for each observation. Make sure to label the new variable created by `predict`.

___ 2.7) Use `prvalue` to compute the predicted probability at some specific value of the independent variables. Interpret this.

___ 2.8) Create a table of predicted probabilities using `prtab`.

___ 2.9) Use `prvalue, save` and `prvalue, dif` to calculate a discrete change. Interpret this.

___ 2.10) Use `prchange` to calculate the discrete changes. Interpret a few of these.

___ 2.11) Use `prgen` to plot the predicted probabilities over the range of a continuous variable for the two levels of a binary variable (this is similar to what is done on page 59 of the notes). Interpret this.

___ 2.12) Obtain the factor change coefficients using `listcoef, help`. Interpret at least one of the unstandardized and one the standardized factor change coefficients.

___ 2.13) Compare the logit and probit coefficients by listing them side-by-side in a table. Why are the unstandardized coefficients different?

___ 2.14) Close log and exit do-file

Section 3: Testing and Assessing Fit – REVIEW

For a fuller discussion of testing and assessing fit, see Chapter 3 of L&F (2005). The file `icda03a-testing-rev.do` contains these Stata commands.

___3.1R) Set-up your do-file.

```
capture log close
log using icda03a-testing-rev, replace text

// program:      icda03a-testing-rev.do
// task:         Review 3 - Testing & Fit
// project:      CDA
// author:       your name \ today's date

// #1
// program setup

version 10
clear all
set linesize 80
matrix drop _all
```

___3.2R) Load the Data.

```
use icpsr_scireview3, clear
```

___3.3R) Examine data, select variables, drop missing, and verify.

```
describe
keep faculty female fellow phd mcit3 mnas
misschk
dropmiss, force obs
summarize
tab1 faculty female fellow phd mcit3 mnas
```

___3.4R) Computing a z-test. z-scores are produced with the standard estimation commands. In the output below the z-scores are in the 4th column. This command:

```
logit faculty female fellow phd mcit3 mnas, nolog
```

Produces this output:

```
. logit faculty female fellow phd mcit3 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
```

faculty	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
female	-.5869003	.2911944	-2.02	0.044	-1.157631	-.0161698
fellow	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	.3537104	.5652778	0.63	0.531	-.7542137	1.461634
_cons	-.5004836	.4539085	-1.10	0.270	-1.390128	.3891607

3.5R) Single Coefficient Wald Test. After estimation, the command `test` can compute a Wald test that a single coefficient is equal to zero. This command:

```
test female
```

Produces this output:

```
. test female

( 1)  female = 0

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

The effect of female is significant at the .05 level ($\chi^2=4.06$, $df=1$, $p=.04$).

3.6R) Multiple Coefficients Wald Test. We can also test if multiple coefficients are equal to zero. This command:

```
test mcit3 mnas
```

Produces this output:

```
. test mcit3 mnas

( 1)  mcit3 = 0
( 2)  mnas  = 0

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

The hypothesis that the effects of mentor's citations and mentor's status as an NAS member are simultaneously equal to zero can be rejected at the .05 level ($\chi^2=7.78$, $df=2$, $p=.02$).

3.7R) Equal Coefficients Wald Test. We can test that multiple coefficients are equal. This command:

```
test mcit3 = mnas
```

Produces this output:

```
. test mcit3 = mnas

( 1)  mcit3 - mnas = 0

           chi2( 1) =      0.35
       Prob > chi2 =      0.5545
```

The hypothesis that the effect of mentor's citations is equal to the effect of mentor's status as an NAS member cannot be rejected ($\chi^2=.35$, $df=1$, $p=.55$).

3.8R) LR Test - Store the Estimation Results. To conduct a likelihood ratio test you begin by storing the estimation results using the `eststo` command. We run the base model and then store the estimates with the name `base`.

```
logit faculty female fellow phd mcit3 mnas
eststo base
```

3.9R) Single Coefficient LR Test. To test that the effect of female is zero, run the base model without `female` and then compare with the full model using `lrtest estname1 estname2`. These commands:

```
logit faculty fellow phd mcit3 mnas
eststo nofemale
lrtest base nofemale
```

Produce this output:

```
. logit faculty fellow phd mcit3 mnas
:::output deleted:::
. eststo nofemale
. lrtest base nofemale
```

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

The effect of female is significant at the .05 level ($LRX^2=4.08$, $df=1$, $p=.04$).

3.10R) Multiple Coefficients LR Test. To test if the effects of `mcit3` and `mnas` are jointly zero, run the comparison model without these variables, store the estimation results, and then compare models using `lrtest`. These commands:

```
logit faculty female fellow phd
eststo nomcit3mnas
lrtest base nomcit3mnas
```

Produces this output:

```
. logit faculty female fellow phd
:::output deleted:::
. eststo nomcit3mnas

. lrtest base nomcit3mnas
```

```
Likelihood-ratio test                LR chi2(2) =      9.19
(Assumption: nomcit3mnas nested in base) Prob > chi2 =      0.0101
```

The hypothesis that the effects of mentor's citations and mentor's status as an NAS member are simultaneously equal to zero can be rejected at the .01 level ($LRX^2=9.19$, $df=2$, $p=.01$).

3.11R) LR Test All Coefficients are Zero. To test that all of the regressors have no effect, we estimate the model with only an intercept, store the estimation results again, and compare the models using `lrtest`. Note that this test statistic is identical to the one produced at the top of the estimation output for the full model (see page 16). These commands:

```
logit faculty
estimates store intercept
lrtest base intercept
```

Produce this output:

```
. logit faculty
:::output deleted:::
. eststo 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 ($LR\chi^2=41.72$, $df=5$, $p<.01$).

3.12R) Fit Statistics. `fitstat` computes measures of fit for your model. The `save` option saves the computed measures in a matrix for subsequent comparisons. `dif` compares the fit measures of the current model with those of the saved model. Here we compare the base model to the model without `mcit3` and `mnas`. We use the `quietly` option because we have already produced the estimation results above.

These commands:

```

quietly logit faculty female fellow phd mcit3 mnas
fitstat, save
quietly logit faculty female fellow phd
fitstat, dif

```

Produce this output:

```

. quietly logit faculty female fellow phd mcit3 mnas
. fitstat, save

```

Measures of Fit for logit of faculty

Log-Lik Intercept Only:	-182.377	Log-Lik Full Model:	-161.515
D(258):	323.030	LR(5):	41.723
		Prob > LR:	0.000
McFadden's R2:	0.114	McFadden's Adj R2:	0.081
ML (Cox-Snell) R2:	0.146	Cragg-Uhler(Nagelkerke) R2:	0.195
McKelvey & Zavoina's R2:	0.201	Efron's R2:	0.151
Variance of y*:	4.116	Variance of error:	3.290
Count R2:	0.678	Adj Count R2:	0.309
AIC:	1.269	AIC*n:	335.030
BIC:	-1115.565	BIC':	-13.843
BIC used by Stata:	356.486	AIC used by Stata:	335.030

(Indices saved in matrix fs_0)

```

. quietly logit faculty female fellow phd
. fitstat, dif

```

Measures of Fit for logit of faculty

	Current	Saved	Difference
Model:	logit	logit	
N:	264	264	0
Log-Lik Intercept Only	-182.377	-182.377	0.000
Log-Lik Full Model	-166.112	-161.515	-4.596
D	332.223(260)	323.030(258)	9.193(2)
LR	32.530(3)	41.723(5)	9.193(2)
Prob > LR	0.000	0.000	0.010
McFadden's R2	0.089	0.114	-0.025
McFadden's Adj R2	0.067	0.081	-0.014
ML (Cox-Snell) R2	0.116	0.146	-0.030
Cragg-Uhler(Nagelkerke) R2	0.155	0.195	-0.040
McKelvey & Zavoina's R2	0.145	0.201	-0.055

Efron's R2	0.120	0.151	-0.031
Variance of y*	3.850	4.116	-0.266
Variance of error	3.290	3.290	0.000
Count R2	0.659	0.678	-0.019
Adj Count R2	0.268	0.309	-0.041
AIC	1.289	1.269	0.020
AIC*n	340.223	335.030	5.193
BIC	-1117.524	-1115.565	-1.959
BIC'	-15.802	-13.843	-1.959
BIC used by Stata	354.527	356.486	-1.959
AIC used by Stata	340.223	335.030	5.193

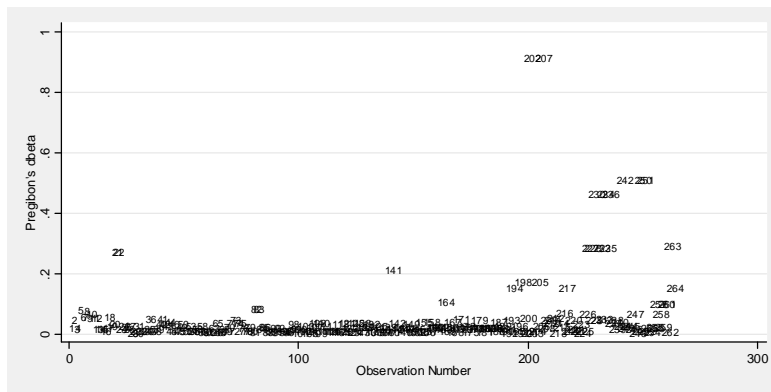
Difference of 1.959 in BIC' provides weak support for current model.

Note: p-value for difference in LR is only valid if models are nested.

3.13R) Plotting Influential Cases Using Cook's Distance. To plot outliers, we first compute Cook's distance using the command `predict , dbeta`. Then we sort our data in some meaningful way (here we choose to sort by `phd`). Next we generate a new variable called `index` whose values correspond to the rank order of `phd` (because of the way the data are sorted). Finally we plot the Cook's distance against the rank order of `phd`. These commands:

```
quietly logit faculty female fellow phd mcit3 mnas
predict cook, dbeta
sort phd
gen index = _n
tway scatter cook index, ysize(1) xsize(2) ///
xlabel(0(100)300) ylabel(0(.2)1., grid) ///
    xscale(range(0, 300)) yscale(range(0, 1.)) ///
    xtitle("Observation Number") ///
    msymbol(none) mlabel(index) mlabposition(0)
graph export icda03a-testing-rev-fig1.emf, replace
```

Produce this graph:



3.14) Close the Log File and exit Do File.

```
log close
exit
```

Section 3: Testing and Assessing Fit – EXERCISE

The file `icda03b-testing-ex.do` contains an outline of this Exercise. If you have done a lot of work already with testing, start with question 3.11 on examining outliers and influential observations.

___3.1) Set-up your do-file

___3.2) Load your data

___3.3) Examine the data and select your variables. Select one binary dependent variable and at least three independent variables. Again be sure to include at least one binary and one continuous independent variable. Drop cases with missing data and verify.

___3.4) Run a logit on the full model.

___3.5) Test the hypothesis that the effect of one of your independent variables is zero using the z-statistic. What is your conclusion? Use `eststo` to store estimation results.

___3.6) Use `test` to conduct a Wald test of the same hypothesis in 3.6. How is the specific value of the Wald test related to the z-test in 3.6?

___3.7) Now use the likelihood ratio test for the same hypothesis in 3.6.

___3.8) Test the hypothesis that the effects of two of your independent variables are simultaneously equal to zero using the Wald test. What is your conclusion?

___3.9) Now use the likelihood ratio test for the same hypothesis in 3.9.

___3.10) Use `fitstat` to compare two of your models. Which model do you prefer and why?

___3.11) Using your preferred model, use methods for detecting outliers and influential observations to evaluate weaknesses in your model. Based on what you find as extreme and/or influential cases, revise your model. Evaluate the revised model in terms of outliers and influential observations. Did things change?

___3.12) Close log & exit.

Section 4: Models for Ordinal Outcomes – REVIEW

For a fuller discussion of models for ordinal outcomes, see Chapter 5 of L&F (2005). The file `icda04a-ordinal-rev.do` contains these Stata commands.

___4.1R) Set-up your do-file.

```
capture log close
log using icda04a-ordinal-rev, replace text

// program:      icda04a-ordinal-rev.do
// task:         Review 4 - Ordinal Regression
// project:      CDA
// author:       your name \ today's date

// #1
// program setup

version 10
clear all
set linesize 80
matrix drop _all
```

___4.2R) Load the Data.

```
use icpsr_scireview3, clear
```

___4.3R) Examine data, select variables, drop missing, and verify. Make sure to look at the distribution of the outcome variable, in this case `jobprst`.

```
describe
keep jobprst publ phd female
misschk
dropmiss, force obs
summarize
tab jobprst , m
```

```
:::output deleted:::
```

```
. tab jobprst , m
```

Rankings of University Job.	Freq.	Percent	Cum.
1_Adeq	29	10.98	10.98
2_Good	128	48.48	59.47
3_Strg	93	35.23	94.70
4_Dist	14	5.30	100.00
Total	264	100.00	

___4.4R) Ordered Logit. `ologit` and `oprobit` work in the same way. We only show `ologit`, but you might want to try what follows using `oprobit`. This command:

```
ologit jobprst publ phd female
```

Produces this output:

```
. ologit jobprst publ phd female
```

```
Iteration 0: log likelihood = -294.86055
Iteration 1: log likelihood = -255.97269
Iteration 2: log likelihood = -254.53567
Iteration 3: log likelihood = -254.51518
Iteration 4: log likelihood = -254.51518
```

```
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]	
publ	.1078786	.0481107	2.24	0.025	.0135833	.2021738
phd	1.130028	.1444046	7.83	0.000	.8470003	1.413056
female	-.6973579	.2617103	-2.66	0.008	-1.210301	-.1844152
/cut1	.9274554	.4268201			.0909033	1.764007
/cut2	4.003182	.4996639			3.023859	4.982505
/cut3	7.034637	.6296717			5.800503	8.26877

___4.5R) Predicted Probabilities in Sample. Use `predict` to compute predicted probabilities after `ologit` or `oprobit`. The `predict` command creates as many new variables as there are categories of the outcome variable so you will need to provide, in this case, 4 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. Here are the commands:

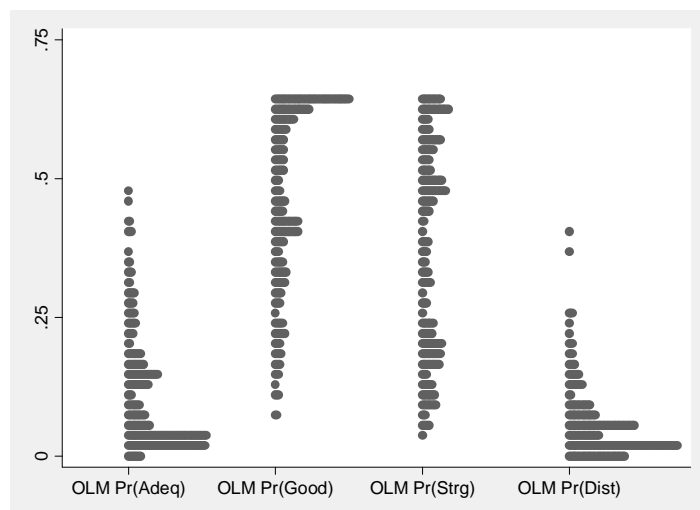
```
predict jpad jpgo jpst jpdi
label var jpad "OLM Pr(Adeq)"
label var jpgo "OLM Pr(Good)"
label var jpst "OLM Pr(Strg)"
label var jpdi "OLM Pr(Dist)"
```

___4.6R) Plot Predictions. An easy way to see the range of predictions is with the command `dotplot`.

These commands:

```
dotplot jpad jpgo jpst jpdi, ylabel(0(.25).75)
graph export icda04a-ordinal-rev-fig1.emf , replace
```

Produce and save this plot:



4.7R) Predict Specific Probabilities. `prvalue` computes the predicted value of the dependent variable given a set of values for the independent variables. Use the `x()` and `rest()` options to set the values at which the variables will be examined. These commands:

```
prvalue, x(female=1 phd=4) rest(mean)
```

Produce this output:

```
. prvalue, x(female=1 phd=4) rest(mean)
```

```
ologit: Predictions for jobprst
```

```
Confidence intervals by delta method
```

		95% Conf. Interval
Pr(y=1_Adeq x):	0.0413	[0.0170, 0.0655]
Pr(y=2_Good x):	0.4412	[0.3436, 0.5388]
Pr(y=3_Strg x):	0.4683	[0.3690, 0.5676]
Pr(y=4_Dist x):	0.0492	[0.0182, 0.0802]

```
x=      pub1      phd      female
      2.3219697      4      1
```

For a female from a distinguished university who is otherwise average, the probability of obtaining a distinguished job is .05.

4.8R) Graph Predicted Probabilities. Graphing predictions at specific, substantively informative values is usually more effective than the `dotplot` of sample predictions above. Here we use the command `prgen` to generate variables for graphing. We consider women from distinguished PhD programs (`phd=4`) and show how predicted probabilities are influenced by publications. `prgen` creates variables of both the predicted probabilities and the cumulative probabilities. Here we use `scatter` to plot the cumulative probabilities. These commands:

```
prgen pub1, x(female=1 phd=4) rest(mean) from(0) to(20) gen(pubpr)
label var pubprs1 "Pr(<=Adeq)"
label var pubprs2 "Pr(<=Good)"
label var pubprs3 "Pr(<=Strg)"
```

```

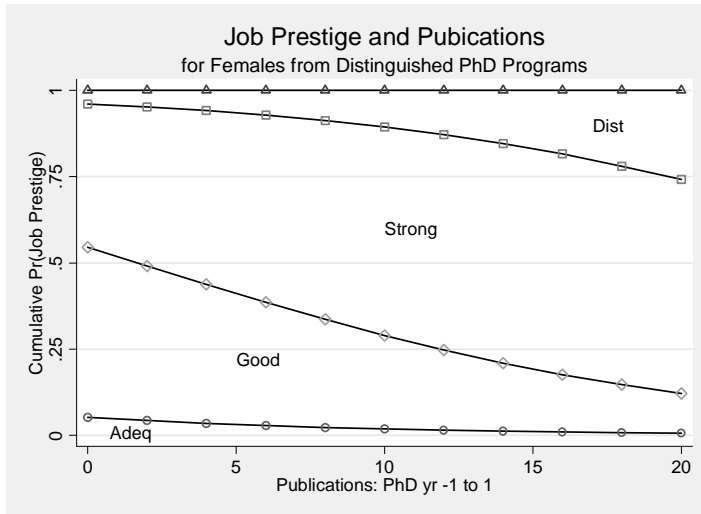
label var pubprs4 "Pr(<=Dist)"

graph twoway (connected pubprs1 pubprs2 pubprs3 pubprs4 pubprx, ///
  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 icda04a-ordinal-rev-fig2.emf , replace

```

Produce and save this plot:



The plot shows many things. First, the probability of obtaining a job in the lowest ranking category (adequate) is quite low for women scientists from high ranking universities regardless of the number of publications. Second, the probability of obtaining a prestigious job (strong or distinguished) increases as the number of publications increase. Third, the most dramatic change across the range of publications appears to be in the probability of obtaining a good job, which decreases as publications increase--this decrease is offset by an increase in the probability of obtaining a prestigious job.

4.9R) Compute the Marginal and Discrete Change. `prchange` computes marginal and discrete change at specific values of the independent variables. By default, the discrete and marginal change is calculated holding all other variables at their mean. Values for specific independent variables can be set using the `x()` and `rest()` options. These commands:

```
prchange, x(female=1 phd=4) rest(mean)
```

Produce this output:

```

. prchange, x(female=1 phd=4) rest(mean)

ologit: Changes in Probabilities for jobprst

pub1

```

Avg Chg	1_Adeq	2_Good	3_Strg	4_Dist
-----------	--------	--------	--------	--------

```

Min->Max      .2056791  -.04531779  -.36604039  .21180955  .19954866
  --+1/2      .01346503  -.00426875  -.0226613  .02188173  .00504833
  --+sd/2     .03470229  -.01103957  -.05836502  .05635044  .01305413
MargEfct      .01346829  -.00426717  -.0226694   .02189     .00504658

```

phd

```

          Avg|Chg|      1_Adeq      2_Good      3_Strg      4_Dist
Min->Max      .32923534  -.54076749  -.11770317  .56185113  .09661958
  --+1/2      .13745592  -.04651201  -.22839981  .22003269  .05487918
  --+sd/2     .13813134  -.04677188  -.22949082  .22107822  .05518444
MargEfct      .14108031  -.04469861  -.237462    .22929773  .05286288

```

female

```

          Avg|Chg|      1_Adeq      2_Good      3_Strg      4_Dist
0->1         .08272706  .02028089  .14517322  -.12050918  -.04494494

```

```

          1_Adeq      2_Good      3_Strg      4_Dist
Pr(y|x)    .04125749  .4412339   .4683077   .04920087

```

```

          publ      phd      female
x=      2.32197      4        1
sd(x)=  2.58074  1.00518  .476172

```

For a female scientist from a top ranked university, a change from the minimum to the maximum number of publications, a difference of 19 publications, increases the probability of receiving a *strong* job by .21.

4.10R) Confidence Intervals for Discrete Change. Although `prchange` computes discrete changes, it does not compute confidence intervals for these changes. To compute confidence intervals, you need to use `prvalue` with the `save` and `dif` options. These commands:

```

prvalue, x(female=1 phd=4 publ=0) rest(mean) label(lowpubs) save
prvalue, x(female=1 phd=4 publ=19) rest(mean) label(hipubs) dif

```

Produce this output:

```

. qui prvalue, x(female=1 phd=4 publ=0) rest(mean) label(lowpubs) save
. prvalue, x(female=1 phd=4 publ=19) rest(mean) label(hipubs) dif

```

ologit: Change in Predictions for jobprst

Confidence intervals by delta method

	Current	Saved	Change	95% CI for Change
Pr(y=1_Adeq x):	0.0071	0.0524	-0.0453	[-0.0770, -0.0136]
Pr(y=2_Good x):	0.1266	0.4926	-0.3660	[-0.5825, -0.1496]
Pr(y=3_Strg x):	0.6281	0.4163	0.2118	[0.0663, 0.3573]
Pr(y=4_Dist x):	0.2383	0.0387	0.1995	[-0.1043, 0.5034]

```

          publ      phd      female
Current=      19        4        1
  Saved=       0        4        1
  Diff=       19        0        0

```

For a female scientist from a top ranked university, a change from the minimum to the maximum number of publications, a difference of 19 publications, significantly increases the probability of receiving a *strong* job by .21 (95% CI: 0.07, 0.36).

4.11R) 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. This command:

```
listcoef , help
```

Produces this output:

```
. listcoef, help
```

```
ologit (N=264): Factor Change in Odds
```

```
Odds of: >m vs <=m
```

jobprst	b	z	P> z	e^b	e^bStdX	SDofX
pub1	0.10788	2.242	0.025	1.1139	1.3210	2.5807
phd	1.13003	7.825	0.000	3.0957	3.1139	1.0052
female	-0.69736	-2.665	0.008	0.4979	0.7174	0.4762

```

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. 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.

4.12R) Testing the Parallel Regression Assumption. `brant` performs a Brant test of the parallel regressions assumptions for the ordered logit model estimated by `ologit`. This command:

```
brant, detail
```

Produces this output:

```
. brant, detail
```

```
Estimated coefficients from j-1 binary regressions
```

	y>1	y>2	y>3
pub1	-.0017262	.10900112	.1431447
phd	.35049291	1.4136638	1.6052011
female	.47578945	-1.1219444	-2.1044684
_cons	.89122353	-4.9397923	-8.8968608

```
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

```
female |      11.26    0.004    2  
-----
```

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

___4.13R) Close the Log File and Exit Do File:

```
log close  
exit
```

Section 4: Models for Ordinal Outcomes – EXERCISE

The file `icda04b-ordinal-ex.do` contains an outline of this Exercise

- ___ 4.1) Set-up your do-file
- ___ 4.2) Load your data
- ___ 4.3) Examine the data and select your variables. Select one ordinal dependent variable. Select at least three independent variables (make sure one is binary and one is continuous). Drop cases with missing data and verify. Make sure to look at the distribution of your outcome variable.
- ___ 4.4) Estimate an ordered logit model.
- ___ 4.5) Predict probabilities for each observation. Make sure to label the new variables created by `predict`.
- ___ 4.6) Use `dotplot` to draw a dotplot of these predictions. What does this tell you?
- ___ 4.7) Use `prvalue` to compute the predicted probability at some specific value of the independent variables. Interpret this.
- ___ 4.8) Use `prgen` to plot the predicted probabilities over the range of a continuous variable. Interpret this.
- ___ 4.9) Use `prchange` to calculate the discrete changes. Interpret a couple of these.
- ___ 4.10) Use `prvalue,save` and `prvalue,dif` to calculate a discrete change with a confidence interval. Interpret this.
- ___ 4.11) Obtain the factor change coefficients using `listcoef, help`. Interpret at least one unstandardized and one standardized factor change coefficient.
- ___ 4.12) Use `brant` to test the parallel regression assumption. What is your conclusion?
- ___ 4.13) Close log and exit do-file

Section 5: Models for Nominal Outcomes – REVIEW

For details about models for multinomial outcomes and associated Stata commands, please read Chapter 6 of L&F (2005). File `icda05a-nominal-rev.do` contains these Stata commands.

___5.1R) Set-up your do-file.

```
capture log close
log using icda05a-nominal-rev, replace text

// program:      icda05a-nominal-rev.do
// task:         Review 5 - Nominal Regression
// project:      CDA
// author:       your name \ today's date

// #1
// program setup

version 10
clear all
set linesize 80
matrix drop _all
```

___5.2R) Load the Data.

```
use icpsr_scireview3, clear
```

___5.3R) Examine data, select variables, drop missing, and verify. Make sure to look at the distribution of the outcome variable, in this case `jobprst`.

```
describe
keep jobprst publ phd female
misschk
dropmiss, force obs
summarize
tab jobprst , m
```

```
:::output deleted:::
```

```
. tab jobprst , m
```

Rankings of University Job.	Freq.	Percent	Cum.
1_Adeq	29	10.98	10.98
2_Good	128	48.48	59.47
3_Strg	93	35.23	94.70
4_Dist	14	5.30	100.00
Total	264	100.00	

___5.4R) Multinomial Logit. `mlogit` estimates the multinomial logit model. The option `baseoutcome()` allows you to set the comparison category. `eststo` is used to store estimation results for model comparison. These commands:

```
mlogit jobprst publ phd female, baseoutcome(4) nolog
```

eststo base

Produce this output:

```
. mlogit jobprst publ phd 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						
publ	-.1577122	.1164934	-1.35	0.176	-.3860351	.0706106
phd	-2.227524	.5717345	-3.90	0.000	-3.348103	-1.106945
female	2.016046	1.168209	1.73	0.084	-.2736008	4.305693
_cons	8.952502	2.312074	3.87	0.000	4.420921	13.48408

2_Good						
publ	-.2360238	.102701	-2.30	0.022	-.437314	-.0347336
phd	-2.473913	.5486317	-4.51	0.000	-3.549211	-1.398615
female	2.957967	1.104271	2.68	0.007	.7936366	5.122298
_cons	10.97811	2.257819	4.86	0.000	6.552867	15.40336

3_Strg						
publ	-.1196281	.0831956	-1.44	0.150	-.2826885	.0434323
phd	-1.080598	.5279463	-2.05	0.041	-2.115353	-.0458419
female	1.768631	1.082638	1.63	0.102	-.3533006	3.890562
_cons	6.285125	2.216575	2.84	0.005	1.940718	10.62953

(jobprst==4_Dist is the base outcome)

```
. eststo base
```

5.5R) 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 **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 **female** and store the estimation results; and third, we compare the two models using `lrtest estname1 estname2`. These commands:

```
quietly mlogit jobprst publ phd, baseoutcome(4)
eststo nofemale
lrtest base nofemale
```

Produce this output:

```
. quietly mlogit jobprst publ phd, baseoutcome(4)
. eststo nofemale
. lrtest base nofemale
```

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

The effect of gender on job prestige is significant at the .01 level ($LR\chi^2=19.17, df=3, p<.001$).

Another way to do this is to use the command `mlogtest` after running the base model. This saves you the step of having to re-estimate the model minus the variable whose effect you want to test. These commands:

```
quietly mlogit jobprst pub1 phd female, baseoutcome(4) nolog
mlogtest, lr
```

Produce this output:

```
. quietly mlogit jobprst pub1 phd female, baseoutcome(4) nolog
. 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
female	19.168	3	0.000

5.6R) Single Coefficient Wald Test. Wald tests can also be computed using the `test` command. This command:

```
test female
```

Produces this output:

```
. test female
```

```
( 1)  [1_Adeq]female = 0
( 2)  [2_Good]female = 0
( 3)  [3_Strg]female = 0
```

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

Again you can automate this process using `mlogtest`. This command:

```
mlogtest, wald
```

Produces this output:

```
. 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	56.560	3	0.000
female	15.748	3	0.001

5.7R) Combining Outcomes 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. This command:

```
test [1_Adeq]
```

Produces this output:

```
. test [1_Adeq]

( 1)  [1_Adeq]pub1 = 0
( 2)  [1_Adeq]phd = 0
( 3)  [1_Adeq]female = 0

           chi2( 3) =    19.02
       Prob > chi2 =    0.0003
```

We can reject the hypothesis that *adequate* and *distinguished* are indistinguishable ($X^2(3)=19.02$ $p<.001$) and therefore conclude that these two categories cannot be combined.

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`. This command:

```
mlogtest, combine
```

Produces this output:

```
. 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_Strg	19.884	3	0.000
1_Adeq- 4_Dist	19.015	3	0.000
2_Good- 3_Strg	51.717	3	0.000
2_Good- 4_Dist	31.133	3	0.000
3_Strg- 4_Dist	9.174	3	0.027

We cannot reject the hypothesis that categories *adequate* and *good* are indistinguishable ($X^2(3)=5.19$ $p=0.158$) and conclude that these categories can be combined..

___5.8R) Testing for IIA. The `mlogtest` command can also be used to test the IIA (independence of irrelevant alternatives) assumption in multinomial logit models. 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. These commands:

```
set seed 4415906
mlogtest, iia
```

Produces this output:

```
. set seed 4415906
. mlogtest, iia
```

**** Hausman tests of IIA assumption (N=264)

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

Omitted	chi2	df	P>chi2	evidence
1_Adeq	3.590	8	0.892	for Ho
2_Good	17.722	8	0.023	against Ho
3_Strg	-45.122	8	---	---

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

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

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

Omitted	chi2	df	P>chi2	evidence
1_Adeq	4.309	8	0.828	for Ho
2_Good	9.917	8	0.271	for Ho
3_Strg	21.270	8	0.006	against Ho

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

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

Omitted	lnL(full)	lnL(omit)	chi2	df	P>chi2	evidence
1_Adeq	-86.808	-68.608	36.399	8	0.000	against Ho
2_Good	-76.978	-53.064	47.827	8	0.000	against Ho
3_Strong	-83.145	-58.028	50.234	8	0.000	against Ho

___5.9R) Predicted Probabilities. **prvalue** can be used as before to compute predicted probabilities for a given set of values of the independent variables. This command:

```
prvalue, rest(mean)
```

Produces this output:

```
. prvalue, rest(mean)
```

```
mlogit: Predictions for jobprst
```

```
Confidence intervals by delta method
```

		95% Conf. Interval	
Pr(y=1_Adeq x):	0.1285	[0.0811,	0.1759]
Pr(y=2_Good x):	0.5130	[0.4395,	0.5866]
Pr(y=3_Strg x):	0.3441	[0.2735,	0.4147]

```
Pr(y=4_Dist|x):      0.0143   [-0.0037,   0.0324]
```

```
      pub1      phd      female  
x= 2.3219697  3.1818939  .34469697
```

___5.10R) Marginal and Discrete Change. We can use `prchange` to calculate marginal and discrete change. This command:

```
prchange, rest(mean)
```

Produces this output:

```
. prchange, rest(mean)
```

```
mlogit: Changes in Probabilities for jobprst
```

```
publ
```

	Avg Chg	1_Adeq	2_Good	3_Strg	4_Dist
Min->Max	.22813503	-.0050713	-.45119875	.28733128	.16893877
+1/2	.01372091	.00318618	-.02744183	.0216375	.00261813
+sd/2	.03536417	.00819671	-.07072833	.0557389	.00679274
MargEfct	.01372404	.00318796	-.02744809	.02164446	.00261567

```
phd
```

	Avg Chg	1_Adeq	2_Good	3_Strg	4_Dist
Min->Max	.3980204	-.07734205	-.71869874	.65568398	.14035681
+1/2	.15514759	-.03767031	-.27262485	.28017618	.030119
+sd/2	.15590683	-.03785592	-.27395776	.28151292	.03030074
MargEfct	.15950273	-.03857619	-.28042927	.29138592	.02761955

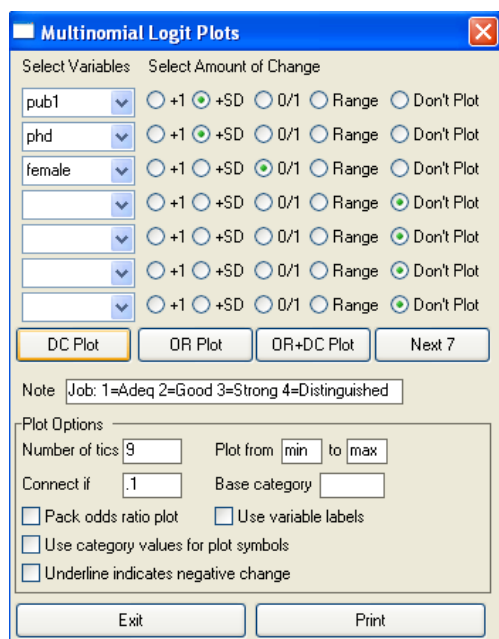
```
female
```

	Avg Chg	1_Adeq	2_Good	3_Strg	4_Dist
0->1	.14022193	-.04894972	.28044385	-.20246901	-.02902516

	1_Adeq	2_Good	3_Strg	4_Dist
Pr(y x)	.12849362	.51303595	.34413981	.01433065

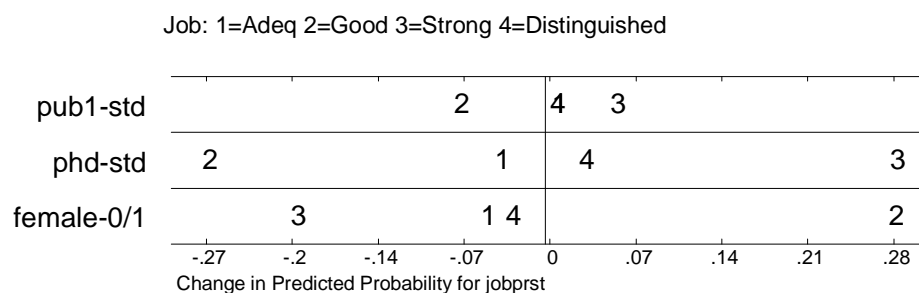
```
      pub1      phd      female  
x= 2.32197  3.18189  .344697  
sd(x)= 2.58074  1.00518  .476172
```

___5.11R) Plot Discrete Change. One difficulty with nominal outcomes is the many coefficients that need to be considered. To help you sort out all the information, discrete change coefficients can be plotted using `mlogview`. You can create these plots either by typing `mlogview` in the command window and then clicking the appropriate radio buttons in the pop-up window or by using the command `mlogplot` in your do-file. We recommend creating the plot using the pop-up window and then, when the desired plot is achieved, copying and pasting the syntax into your do-file. We also recommend adding a `note` to the plot that includes the values and value labels. Typing `mlogview` produces this pop-up:



Selecting the radio buttons above and clicking on the “DC Plot” button produces this output:

```
. mlogplot pub1 phd female, std(ss0) p(.1)
> note(Job: 1=Adeq 2=Good 3=Strong 4=Distinguished) dc ntics(9)
```



The effects of a standard deviation change in PhD prestige and of being female are larger than the effect of a standard deviation change in publications. A standard deviation increase in PhD prestige increases the probability of being in a *strong* (3) job and decreases the probability of being in *good* (2) job by about .28, for average scientists. Being female increases the probability of being in a *good* (2) job by about .28 and decreases the probability of being in a *strong* (3) job by about .20, for average scientists. None of the variables have much of an impact on the probability of obtaining the lowest (1) or highest (4) ranking jobs.

5.12R) Odds Ratios. `listcoef` computes the factor change coefficients for each of the comparisons. The output is arranged by the independent variables. This command:

```
listcoef, help
```

Produces this output:

```
. listcoef, help
```

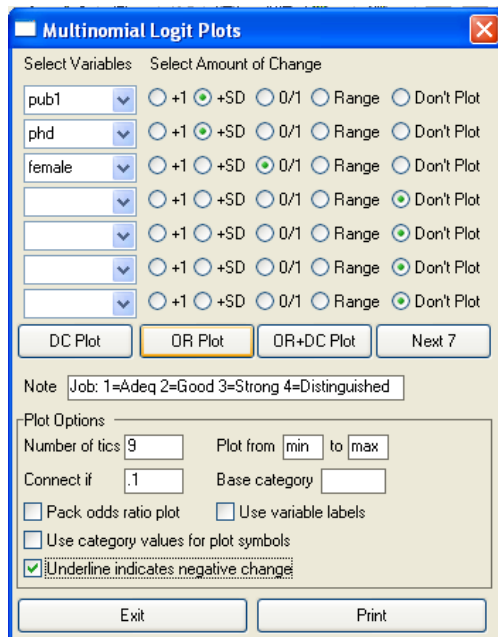
```
mlogit (N=264): Factor Change in the Odds of jobprst
```

Variable: pub1 (sd=2.5807363)

Odds comparing Alternative 1 to Alternative 2		b	z	P> z	e^b	e^bstdX
1_Adeq	-2_Good	0.07831	0.879	0.379	1.0815	1.2240
1_Adeq	-3_Strg	-0.03808	-0.412	0.680	0.9626	0.9064
1_Adeq	-4_Dist	-0.15771	-1.354	0.176	0.8541	0.6656
2_Good	-1_Adeq	-0.07831	-0.879	0.379	0.9247	0.8170
2_Good	-3_Strg	-0.11640	-1.623	0.105	0.8901	0.7405
2_Good	-4_Dist	-0.23602	-2.298	0.022	0.7898	0.5438
3_Strg	-1_Adeq	0.03808	0.412	0.680	1.0388	1.1033
3_Strg	-2_Good	0.11640	1.623	0.105	1.1234	1.3504
3_Strg	-4_Dist	-0.11963	-1.438	0.150	0.8873	0.7344
4_Dist	-1_Adeq	0.15771	1.354	0.176	1.1708	1.5023
4_Dist	-2_Good	0.23602	2.298	0.022	1.2662	1.8388
4_Dist	-3_Strg	0.11963	1.438	0.150	1.1271	1.3617

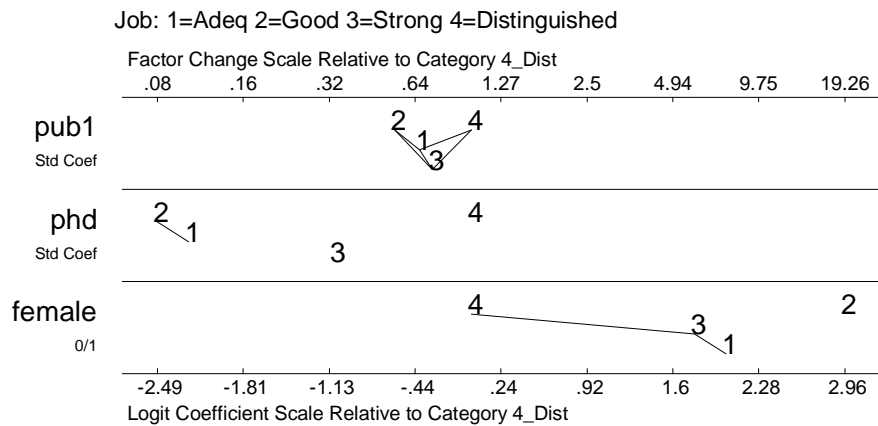
:::output deleted:::

5.13R) Plot Odds Ratios. The odds ratios can be plotted in much the same way as the discrete changes. In the plot, a solid line indicates that the coefficient cannot differentiate between the two outcomes that are connected. Typing `mlogview` in the command window produces this pop-up:



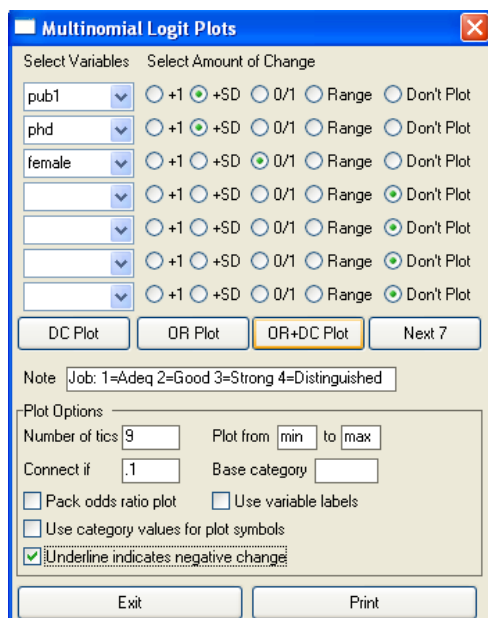
Highlighting these radio buttons and then clicking “OR Plot” produces this output:

```
. mlogplot pub1 phd female, std(ss0) p(.1) ///
> note(Job: 1=Adeq 2=Good 3=Strong 4=Distinguished) or sign ntics(9)
```



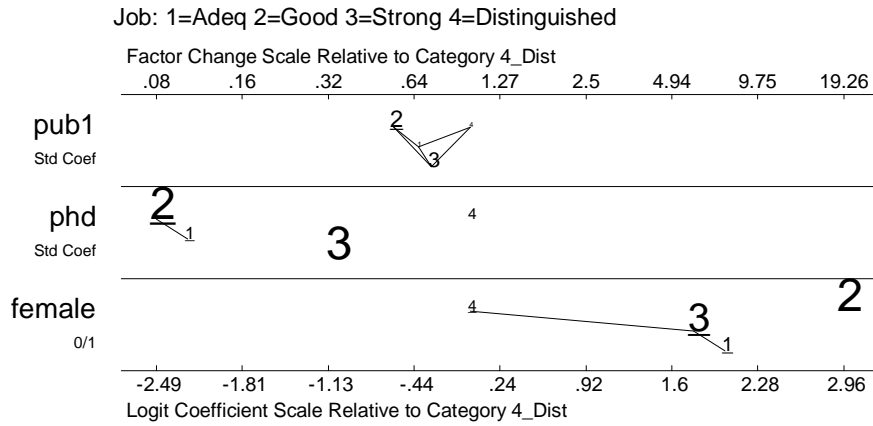
The effect of publications is the smallest, although a standard deviation increase in publications increases the odds of obtaining a *distinguished* (4) job compared to a *good* (2) job. A standard deviation increase in PhD prestige increases the odds of obtaining a *distinguished* (4) job relative to any other type of job, but fails to significantly distinguish between obtaining a *good* (2) job and an *adequate* (1) job. Being female increases the odds of obtaining a *good* (2) job relative to any other type of job, but does not significantly distinguish between obtaining a *distinguished* (4) and *strong* (3) job or a *strong* (3) and *adequate* (1) job.

5.14) Adding Discrete Change to OR Plot. Information about the discrete change can be incorporated in the odds-ratio plot. Remember that 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. Typing `mlogview` in the command window produces this pop-up:



Highlighting the radio buttons above and clicking the “OR+DC Plot” button produces this output:

```
. mlogplot pub1 phd female, std(ss0) p(.1)
> note(Job: 1=Adeq 2=Good 3=Strong 4=Distinguished) or dc sign ntics(9)
```



5.15R) Close Log File and Exit Do File.

```
log close
exit
```

Section 5: Models for Nominal Outcomes – EXERCISE

The file `icda05b-nominal-ex.do` contains an outline of this exercise.

___ 5.1) Set-up your do-file

___ 5.2) Load your data

___ 5.3) Examine the data and select your variables. Select one nominal dependent variable and at least three independent variables (make sure one is binary and one is continuous). Drop cases with missing data and verify. Make sure to look at the distribution of your outcome variable.

___ 5.4) Estimate a multinomial logit model.

___ 5.5) Use `mlogtest` to compute the LR test for each independent variable. Write your conclusion for at least one variable.

___ 5.6) Use `mlogtest` to compute the LR test that categories of the dependent variable can be combined. What do you find?

___ 5.7) Compute discrete changes and marginal effects using `prchange` with the independent variables held at some values.

___ 5.8) Use `mlogview` to plot the discrete changes. Write up an interpretation as if it were part of a publishable research paper. Note that you can use the output from 5.8 to determine the specific values.

___ 5.9) Use `listcoef`, `help` to compute the factor change coefficients.

___ 5.10) Use `mlogview` to plot the odds ratios. Write up an interpretation as if it were part of a publishable research paper. Note that you can use the output from 5.10 to determine the specific values.

___ 5.11) Now add discrete change to the odds ratio plot using `mlogview`. Do you see how discrete change and odds ratios give you different pieces of information?

___ 5.12) Close log and exit do-file

Section 6: Models for Count Outcomes – REVIEW

For more details about models for count outcomes, please read Chapter 7 of L&F (2005). The file `icda06a-count-rev.do` contains these Stata commands.

___6.1R) Set-up your do-file.

```
capture log close
log using icda06a-count-rev, replace text

// program:      icda06a-count-rev.do
// task:         Review 6 - Count Models
// project:      CDA
// author:       your name \ today's date

// #1
// program setup

version 10
clear all
set linesize 80
matrix drop _all
```

___6.2R) Load the Data.

```
use icpsr_scireview3, clear
```

___6.3R) Examine data, select variables, drop missing, and verify. Make sure to look at the distribution of the outcome variable, in this case, `pub6`.

```
describe
keep pub6 female phd enrol
misschk
dropmiss, force obs
summarize
tab1 pub6 female phd enrol
dotplot pub6
graph export icda06a-count-rev-fig1.emf , replace
```

___6.4R) Estimate the Poisson Regression Model. This command:

```
poisson pub6 female phd enrol, nolog
```

Produces this output:

```
. poisson pub6 female phd enrol, nolog
```

```
Poisson regression              Number of obs   =          264
                               LR chi2(3)           =          78.72
                               Prob > chi2          =          0.0000
Log likelihood = -839.78052      Pseudo R2       =          0.0448
```

	pub6	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	female	-.2408113	.069001	-3.49	0.000	-.3760508 -.1055719
	phd	.1882524	.0321844	5.85	0.000	.1251721 .2513328
	enrol	-.1325456	.0240756	-5.51	0.000	-.179733 -.0853582
	_cons	1.532594	.1699269	9.02	0.000	1.199543 1.865644

___6.5R) **Factor Changes.** `listcoef` computes the factor change coefficients. Note that factor change coefficients can also be computed after estimating the NBRM (coming up), but since the output is similar, we show only the output for PRM. This command:

```
listcoef, help
```

Produces this output:

```
. listcoef, help
```

```
poisson (N=264): Factor Change in Expected Count
```

```
Observed SD: 4.3102677
```

pub6	b	z	P> z	e^b	e^bStdX	SDofX
female	-0.24081	-3.490	0.000	0.7860	0.8917	0.4762
phd	0.18825	5.849	0.000	1.2071	1.2083	1.0052
enrol	-0.13255	-5.505	0.000	0.8759	0.8259	1.4432

```

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 .79, holding all other variables constant. A standard deviation increase in the number of years from enrollment to completion of PhD, about 1.4 years, decreases the expected number of publications by 17%, holding other variables constant.

___6.6R) **Marginal and Discrete Changes.** It is also possible to use the `prchange` command to compute the discrete change in the expected count/rate. This command:

```
prchange , rest(mean)
```

Produces this output:

```
. prchange , rest(mean)
```

```
poisson: Changes in Rate for pub6
```

	min->max	0->1	++1/2	++sd/2	MargEfct
female	-0.8666	-0.8666	-0.8996	-0.4276	-0.8975
phd	2.4511	0.4241	0.7026	0.7063	0.7016
enrol	-3.9991	-0.9629	-0.4943	-0.7140	-0.4940

```
exp(xb): 3.7269
```

```

      female      phd      enrol
x=   .344697   3.18189   5.5303
sd(x)= .476172   1.00518   1.44317

```

Being a female scientist decreases expected productivity by .86 publications, holding all other variables at their mean. A standard deviation increase (centered around the mean) in the number of years from enrollment to completion of PhD, about 1.4 years, decreases the expected rate of productivity .71, holding other variables at their mean.

___6.7R) Predicted Rate and Probabilities. We can use the `prvalue` command to generate confidence intervals for the discrete changes. Here we compare men and women who are average on all other characteristics. These commands:

```
quietly prvalue, x(female=0) rest(mean) save
prvalue, x(female=1) rest(mean) dif
```

Produce this output:

```
. quietly prvalue, x(female=0) rest(mean) save
. prvalue, x(female=1) rest(mean) dif
```

```
poisson: Change in Predictions for pub6
```

```
Confidence intervals by delta method
```

	Current	Saved	Change	95% CI for Change
Rate:	3.1828	4.0494	-.86662	[-1.3347, -0.3985]
Pr(y=0 x):	0.0415	0.0174	0.0240	[0.0081, 0.0399]
Pr(y=1 x):	0.1320	0.0706	0.0614	[0.0250, 0.0978]
Pr(y=2 x):	0.2100	0.1429	0.0671	[0.0316, 0.1026]
Pr(y=3 x):	0.2228	0.1929	0.0299	[0.0142, 0.0457]
Pr(y=4 x):	0.1773	0.1953	-0.0180	[-0.0346, -0.0013]
Pr(y=5 x):	0.1129	0.1582	-0.0453	[-0.0711, -0.0195]
Pr(y=6 x):	0.0599	0.1068	-0.0469	[-0.0714, -0.0224]
Pr(y=7 x):	0.0272	0.0618	-0.0345	[-0.0524, -0.0167]
Pr(y=8 x):	0.0108	0.0313	-0.0204	[-0.0313, -0.0096]
Pr(y=9 x):	0.0038	0.0141	-0.0102	[-0.0160, -0.0045]

	female	phd	enrol
Current=	1	3.1818939	5.530303
Saved=	0	3.1818939	5.530303
Diff=	1	0	0

For scientists who are average on all other characteristics, women are expected to have almost .90 fewer publications than men, with estimated bounds for the 95% confidence interval at -1.33 and -.40. Women are more likely than men to have between 0 and 3 publications and men are more likely than women to have more than 3 publications.

___6.8R) The Negative Binomial Regression Model. We can use the same types of commands for the NBRM. Note that this model typically takes longer to converge. The output is similar to that of the PRM, except for the results at the bottom where you will find overdispersion parameter (alpha) and a likelihood ratio test for overdispersion. This command:

```
nbreg pub6 female phd enrol
```

Produces this output:

```
. nbreg pub6 female phd enrol
```

Negative binomial regression

Number of obs = 264
LR chi2(3) = 20.59
Prob > chi2 = 0.0001
Pseudo R2 = 0.0158

Dispersion = mean
Log likelihood = -642.723

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
pub6						
female	-.2822292	.1382637	-2.04	0.041	-.553221	-.0112373
phd	.1995909	.0651859	3.06	0.002	.0718288	.327353
enrol	-.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.

6.9R) 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” the variable `phd`. This command:

```
zip pub6 female phd enrol, inf(phd)
```

Produces this output:

```
. zip pub6 female phd enrol, inf(phd)
```

Zero-inflated Poisson regression

Number of obs = 264
Nonzero obs = 212
Zero obs = 52

Inflation model = logit
Log likelihood = -758.0032

LR chi2(3) = 48.74
Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
pub6						
female	-.1210631	.0710846	-1.70	0.089	-.2603864	.0182602
phd	.1400257	.0334849	4.18	0.000	.0743964	.205655
enrol	-.1306837	.0250179	-5.22	0.000	-.1797178	-.0816496
_cons	1.838966	.1749225	10.51	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

6.10R) The ZINB Model. We can use the same types of commands for the ZINB. This command:

```
zinb pub6 female phd enrol, inf(phd)
```

Produces this output:

```
. zinb pub6 female phd enrol, inf(phd)
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
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	

pub6						
female	-.2708994	.1371918	-1.97	0.048	-.5397905	-.0020084
phd	.1745669	.0695427	2.51	0.012	.0382657	.3108682
enrol	-.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

6.11R) 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. This command:

```
listcoef, help
```

Produces this output:

```
. listcoef

zinb (N=264): Factor Change in Expected Count

Observed SD: 4.3102677

Count Equation: Factor Change in Expected Count for Those Not Always 0
```

pub6	b	z	P> z	e^b	e^bStdX	SDofX
female	-0.27090	-1.975	0.048	0.7627	0.8790	0.4762
phd	0.17457	2.510	0.012	1.1907	1.1918	1.0052
enrol	-0.15272	-3.247	0.001	0.8584	0.8022	1.4432

ln alpha	-0.35142					
alpha	0.70369	SE(alpha) = 0.14831				

Binary Equation: Factor Change in Odds of Always 0

Always0	b	z	P> z	e^b	e^bStdX	SDofX
phd	-0.54405	-0.628	0.530	0.5804	0.5788	1.0052

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. A standard deviation increase in PhD prestige decreases the odds of not having the opportunity to publish by a factor of .58, although this is not significant (z=-0.63, p=0.53).

6.12R) Discrete and Marginal Changes. For the ZIP and ZINB models, `prvalue` produces two probabilities of zero counts: the probability of being zero and the probability of being always zero. This command:

```
quietly prvalue , x(phd=1) rest(mean) save
prvalue , x(phd=0) rest(mean) dif
```

Produces this output:

```
. quietly prvalue, x(phd=1) rest(mean) save
. prvalue, x(phd=4) rest(mean) diff
```

zinb: Change in Predictions for pub6

	Current	Saved	Change
Expected y:	4.3668	2.3388	2.0281
Pr(Always0 z):	0.0258	0.1191	-0.0933
Pr(y=0 x,z):	0.1545	0.3162	-0.1617
Pr(y=1 x):	0.1389	0.1824	-0.0435
Pr(y=2 x):	0.1277	0.1438	-0.0161
Pr(y=3 x):	0.1106	0.1068	0.0037
Pr(y=4 x):	0.0928	0.0769	0.0159
Pr(y=5 x):	0.0764	0.0543	0.0221
Pr(y=6 x):	0.0621	0.0379	0.0242
Pr(y=7 x):	0.0500	0.0261	0.0238
Pr(y=8 x):	0.0399	0.0179	0.0220
Pr(y=9 x):	0.0317	0.0122	0.0195

x values for count equation

	female	phd	enrol
Current=	.34469697	4	5.530303
Saved=	.34469697	1	5.530303
Diff=	0	3	0

z values for binary equation

	phd
Current=	4
Saved=	1
Diff=	3

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 .32. The probability of having no publications because the scientist does not have the opportunity to publish is .12. Thus most of the 0s for average scientists are due to being in the group who are “potential publishers.”

We find that scientists from low prestige universities who are otherwise average are expected to have 2 fewer publications than scientists from high prestige universities who are otherwise average.

__6.13R) Compare model using countfit. `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. This command:

```
countfit pub6 female phd enrol, inf(female phd enrol) ///
graph(icda06a-count-rev-fig2.emf , replace)
```

Produces this output:

```
. countfit pub6 female phd enrol, inf(female phd enrol) ///
>      graph(icda06a-count-rev-fig2.emf , replace)

. countfit pub6 female phd enrol, inf(phd)
```

Variable		PRM	NBRM	ZIP	ZINB
pub6					
Female: 1=female,0..		0.786	0.754	0.886	0.763
		-3.49	-2.04	-1.70	-1.97
Prestige of Ph.D. department.		1.207	1.221	1.150	1.191
		5.85	3.06	4.18	2.51
Years from BA to P..		0.876	0.860	0.877	0.858
		-5.51	-3.14	-5.22	-3.25
Constant		4.630	4.990	6.290	5.696
		9.02	4.76	10.51	4.97
lnalpha					
Constant			0.816		0.704
			-1.62		-1.67
inflate					
Prestige of Ph.D. department.				0.788	0.580
				-1.44	-0.63
Constant				0.471	0.233
				-1.41	-0.70
Statistics					
alpha			0.816		
N		264.000	264.000	264.000	264.000
ll		-839.781	-642.723	-758.003	-642.203
bic		1701.865	1313.326	1549.462	1323.437
aic		1687.561	1295.446	1528.006	1298.405

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
::: output omitted :::				
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

NBRM: Predicted and actual probabilities

Count	Actual	Predicted	Diff	Pearson
0	0.197	0.187	0.010	0.142
1	0.144	0.167	0.023	0.834
2	0.129	0.136	0.008	0.115
3	0.121	0.109	0.013	0.382
::: output omitted :::				
8	0.042	0.032	0.009	0.728
9	0.023	0.025	0.003	0.069
Sum	0.917	0.903	0.145	12.875

ZIP: Predicted and actual probabilities

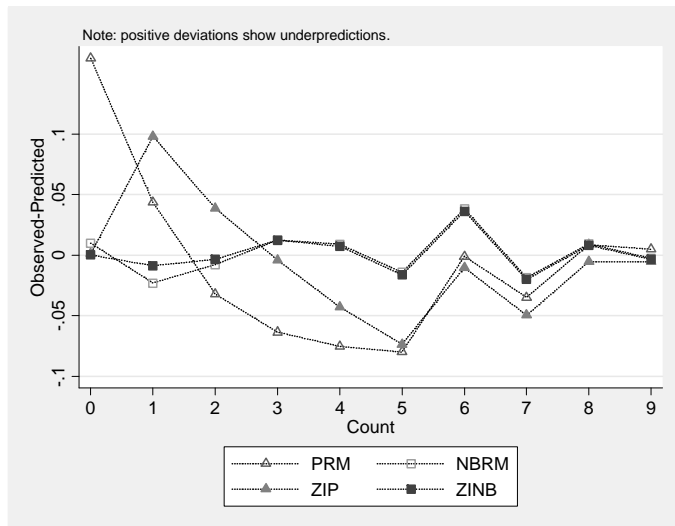
Count	Actual	Predicted	Diff	Pearson
0	0.197	0.197	0.000	0.000
1	0.144	0.044	0.100	59.264
2	0.129	0.088	0.041	4.940
3	0.121	0.124	0.003	0.016
::: output omitted :::				
8	0.042	0.048	0.006	0.193
9	0.023	0.028	0.006	0.306
Sum	0.917	0.969	0.332	89.027

ZINB: Predicted and actual probabilities

Count	Actual	Predicted	Diff	Pearson
0	0.197	0.194	0.003	0.010
1	0.144	0.156	0.012	0.247
2	0.129	0.133	0.004	0.040
3	0.121	0.109	0.012	0.373
::: output omitted :::				
8	0.042	0.033	0.008	0.567
9	0.023	0.026	0.003	0.109
Sum	0.917	0.904	0.123	11.496

Tests and Fit Statistics

PRM	BIC=	229.814	AIC=	6.392	Prefer	Over	Evidence
vs NBRM	BIC=	-158.725	dif=	388.539	NBRM	PRM	Very strong
	AIC=	4.907	dif=	1.485	NBRM	PRM	
	LRX2=	394.115	prob=	0.000	NBRM	PRM	p=0.000
vs ZIP	BIC=	84.385	dif=	145.429	ZIP	PRM	Very strong
	AIC=	5.787	dif=	0.605	ZIP	PRM	
	Vuong=	4.358	prob=	0.000	ZIP	PRM	p=0.000
vs ZINB	BIC=	-139.341	dif=	369.155	ZINB	PRM	Very strong
	AIC=	4.926	dif=	1.466	ZINB	PRM	
NBRM	BIC=	-158.725	AIC=	4.907	Prefer	Over	Evidence
vs ZIP	BIC=	84.385	dif=	-243.110	NBRM	ZIP	Very strong
	AIC=	5.787	dif=	-0.880	NBRM	ZIP	
vs ZINB	BIC=	-139.341	dif=	-19.384	NBRM	ZINB	Very strong
	AIC=	4.926	dif=	-0.019	NBRM	ZINB	
	Vuong=	0.834	prob=	0.202	ZINB	NBRM	p=0.202
ZIP	BIC=	84.385	AIC=	5.787	Prefer	Over	Evidence
vs ZINB	BIC=	-139.341	dif=	223.726	ZINB	ZIP	Very strong
	AIC=	4.926	dif=	0.861	ZINB	ZIP	
	LRX2=	229.302	prob=	0.000	ZINB	ZIP	p=0.000



___ 6.14R) Close the Log File.

log close
exit

Section 6: Models for Count Outcomes – EXERCISE

The file `icda06b-count-ex.do` contains an outline of this exercise.

___ 6.1) Set-up your do-file.

___ 6.2) Load your data.

___ 6.3) Examine the data and select your variables. Choose one count dependent variable and at least three independent variables (make sure one is binary and one is continuous). Drop cases with missing data and verify. Make sure to look at the distribution of your outcome variable.

___ 6.4) Estimate a NBRM.

___ 6.5) List the factor change coefficients using `listcoef`, `help`. Interpret a few of the unstandardized and standardized factor change coefficients.

___ 6.6) Test NBRM against PRM and write up the results as though it were part of a research paper. (Hint: use the LR test results at the end of the NBRM output)

___ 6.7) Use `prvalue,save` and `prvalue,dif` to calculate a discrete change. Interpret this.

___ 6.8) Estimate the same model using `zip` and `zinb`.

___ 6.9) List the factor change coefficients using `listcoef`, `help` for either the ZIP or ZINB model. Interpret a few of the unstandardized and standardized factor change coefficients.

___ 6.10) Use `prvalue,save` and `prvalue,dif` to calculate a discrete change. Interpret this.

___ 6.11) Use `countfit` to compare count models. Which model do you prefer and why?

___ 6.12) Close log and exit do-file.

Data Sets for CDA Workshop

Scott Long - June 2009

There are three data sets that are provided for the computer exercises.

- **icpsr_science3 (icpsr_scireview3)** contains information on the careers of 308 Ph.D. biochemists. (Note that icpsr_scireview3 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 these data are *not* the actual data that were collected.
- **icpsr_hsb3** contains 1647 observations on 68 variables from the 1983 High School and Beyond Study.
- **icpsr_nes3** contain 2487 observations on 45 variables from the 1992 National Election Study.
- **icpsr_addhealth3** 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*.

icpsr_science3.dta (icpsr_scireview3): Codebook for Science Data

id	ID number of scientist.
cit#	Number of citations over 3-year period ending in career year # (for #=1, 3, 6, 9)
enrol	Number of years it took to get a Ph.D. after receipt of B.A.
faculty	Faculty in a college or university? 0: No 1: Yes
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.
fellow	Postdoctoral fellow? 0: No 1: Yes
felclass	Fellow or Ph.D. prestige class. 1: adequate 2: good 3: strong 4: distinguished
female	Female? 0: No 1: Yes
job	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.
jobclass	Prestige class of 1st 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? 0: No 1: Yes
mnas	Was mentor in National Academy of Science? 0: No 1: Yes
mpub3	Mentor's # of articles in 3 year period ending year of the student's Ph.D.
nopub#	No articles in 3 year period ending year # after Ph.D. (for #=1, 3, 6, 9) 0: No 1: Yes
phd	Prestige of Ph.D 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
pub#	Number of publications over 3-year period ending # (for #=1, 3, 6, 9)
totpub	Total Pubs in 9 Yrs post-Ph.D.
work	Type of first job 1: Faculty in university 2: Academic research 3: College teacher 4: Industrial research 5: Administration
workadm	Work in administration? 0: No 1: Yes
worktch	Work in teaching 0: No 1: Yes
workuniv	Work in university? 0: No 1: Yes

Suggestions for variable sets by model:

Linear Regression Model:	Y:	totcit (created in the Stata Guide)
	C:	fel
	D:	mnas
	X:	enrol
Binary Regression Model:	Y:	nopub3
	C:	phd
	D:	female
	X:	enroll
Multinomial Logit Model:	Y:	work
	C:	pub1
	D:	female
	X:	phd
Count Model:	Y:	pub9
	C:	mcit3
	D:	workuniv
	X:	fellow

icpsr_hsb3.dta: 1983 High School and Beyond Study

id: ID number of respondent

sex 1: male 2: female

male 0: no 1: yes

female 0: no 1: yes

region 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?

1: Mostly below D's 2: Mostly D's 3: Mostly C's & D's 4: Mostly C's
5: Mostly B's & C's 6: Mostly B's 7: Mostly A's & B's 8: 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.

1: age 11 or less 2 to 9: ages 12 to 19 respectively 11: never worked

hrswork: Hours worked last week.

1: none 2: 1 to 4
5: 22 to 29 6: 30 to 34

hrs1styr: Hours worked per week last year

3: 5 to 14 4: 15 to 21
7: 35 or more

varsport: Did you participate in varsity sports?

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

1: non participant 2: participant 3: leader/officer

livealon: Did you live alone while attending HS?

livemale: With other male guardian?

livfemal: With other female guardian?

livgrand: With your grandparent(s)?

0: no 1: yes

livedad: With your father while attending HS?

livemom: With mother?

livesibs: With any brothers or sisters?

momwork: Did your mother work while you were in HS?

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

premomwk: 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.

1: not living with father
4: farmer
7: manager/admin
10: professional
13: protective service
16: service
19: DK

momocc: Mother's occupation.

2: clerical 3: craftsman
5: homemaker 6: laborer
8: military 9: operative
11: advanced professional 12: proprietor
14: sales 15: school teacher
17: technical 18: never worked

daded: Father's education level.

1: not living with father
4: vocational less than 2 years
7: college 2 or more years
10: PhD/MD advanced degree

momed: Mother's education level.

2: less than HS degree 3: HS or equivalent degree
5: vocational 2 or more years 6: college less than 2 years
8: college graduate 9: masters degree
11: DK

dadhsgrd: Dad graduate high school?

dadcoll: Dad graduate college?

0: no

momhsgrd: Mom graduate high school?

momcoll: Mom graduate college?

1: yes

mommonit: Mother monitors your school work?

1: yes 2: no

dadmonit: Father monitors your school work?

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: How much did your father influence your HS plans? **momplans:** your mother?

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?

- | | | | |
|----------------|--------------|---------------|----------------|
| 1: None | 2: <\$1K | 3: \$1K-\$3K | 4: \$3K-\$5K |
| 5: \$5K-\$7K | 6: \$7K-\$9K | 7: \$9K-\$11K | 8: \$11K-\$13K |
| 9: \$13K-\$15K | 10: \$15K+ | | |

expenses: How many expenses do you have?

- | | | | |
|------------------|----------------|-------------------|--------------|
| 1: None, at home | 2: None: other | 3: Less than \$1K | 4: \$1K-\$2K |
| 5: \$2K-\$3K | 6: \$3K-\$4K | 7: \$4K-\$5K | 8: \$5K-\$7K |
| 9: \$7K-\$10K | 10: \$10K+ | | |

netearn: Net earnings this year

- | |
|------------------|
| 1: none |
| 4: \$600-\$1,200 |

sumearn: Net earnings from last year.

- | | |
|--------------------|----------------|
| 2: less than \$200 | 3: \$300-\$600 |
| 5: \$1,200-\$2,000 | 6: \$2,000+ |

agewed: Age you expect to be married.

agekid: ...to have your first child.

agejob: ...to have your first full time job.

agehome: ...to move out on your own.

ageeduc: ...to finish your education.

- | | | | |
|--------------------|---------------|-------------|-------------|
| 1: Don't expect to | 2: already am | 3: under 18 | 4: 19 |
| 5: 20 | 6: 21 | 7: 22 | 8: 23 |
| 9: 24 | 10: 25 | 11: 26 | 12: 27 |
| 13: 28 | 14: 29 | 15: 30 | 16: over 30 |

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?

relOth: Other religion?

0: no

1: yes

relCath: Catholic?

relNone: No religion?

relJew: Jewish?

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 liberal

5: none

6: DK

fincome: Family income

1: Under \$7K

2: \$7-\$12K

3: \$12-\$16K

4: \$16-\$20K

5: \$20-\$25K

6: \$25-38K

7: >\$38K

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?

1: home state

2: another state

ses: Socioeconomic status

1: low

2: medium

3: high

Suggestions for variable sets by model:

Binary Regression Model:

Y: dadmonit

C: hsgrades

D: chem

X: ses

Ordinal Logit Model:

Y: talkpar

C: agehome

D: female

X: dadplans

Multinomial Logit Model:

Y: talkpar
C: agehome
D: female
X: dadplans

Y: varsport
C: edattain
D: male
X: momplans

icpsr_nes3.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.

(NOTE: Feeling thermometers range from 0 to 100. The higher the score, the more favorable the view of the candidate. 50 is a neutral score.)

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 minor party	

abortion: View on abortion

1: Abortion never permitted by law	2: Only if rape, incest, or life threatening
3: Only 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?

0: no

relCath: Catholic?

1: yes

relJew: Jewish?

relOth: Other religion?

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

marital: Marital status

1: Married and 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: 9th-11th grades	3: High school
4: More than 12 years	5: Jr. college degree	6: BA level degrees
7: Advanced degree		

collgrad: College graduate?

0: no

1: yes

hsgrad: High School graduate?

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: Transportation and material moving occup. |
| 6: Private household | 13: Handlers, equipment cleaners, laborers |
| 7: Protective service | 14: Member of the armed forces |

fincome: Family income.

- | | | | |
|----------------------|----------------------|-------------|------------|
| 1: <3K | 2: 3-5K | 3: 5-7K | 4: 7-9K |
| 5: 9-10K | 6: 10-11K | 7: 11-12K | 8: 12-13K |
| 9: 13-14K | 10: 14-15K | 11: 15-17K | 12: 17-20K |
| 13: 20-22K | 14: 22-25K | 15: 25-30K | 16: 30-35K |
| 17: 35-40K | 18: 40-45K | 19: 45-50K | 20: 50-60K |
| 21: 60-75K | 22: 75-90K | 23: 90-105K | 24: >105K |
| 66: Below 25K but NA | 77: Above 25K but NA | | |

income: Income, recoded to midpoints of fincome (66 and 77 = missing)

sex: Respondent's sex

- | | |
|---------|-----------|
| 1: Male | 2: Female |
|---------|-----------|

male: Male?

- 0: no

female: Female?

- 1: yes

race: Respondent's race

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

white: White?

- 0: no

black: Black?

amerind: American Indian?

- 1: yes

asian: Asian?

didvote: Did you vote this November?

- 0: No

regvote: Were you registered to vote?

- | | |
|--------|-----------------|
| 1: Yes | 6: Not required |
|--------|-----------------|

presvote: Presidential vote.

- 1: Bush

prefvote: Did not vote, but preferred

- | | | |
|------------|----------|----------|
| 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 are used to create **polacts**; the code for creating this variable is in the Stata Guide provided by the instructor.

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

Suggestions for variable sets by model:

Linear Regression Model:	Y: preclint C: age D: hsgrad X: fundcam
Binary Regression Model:	Y: campaign C: income D: male X: age
Ordinal Logit Model:	Y: partyid (recoded to smaller # categories) C: income D: relProt X: educatio
Multinomial Logit Model:	Y: partyid (recoded to smaller # categories) C: income D: relProt X: education
MNLM 2nd Outcome:	Y: abortion C: income D: relProt X: partyid
Count Model:	Y: polacts (created in the Stata Guide) C: prebush D: collgrad X: married

icpsr_addhealth3: Codebook for 1994-95 wave of Add Health Public Data extract

Note: missing values for all variables are as follows:

.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

Note: the syntax for setting the survey weights is:

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

age: Respondent's age (calculation includes months; ranges from 12.4167 to 21.1667).

sex: Respondent's sex

1: Male 2: Female

male: Male?

0: no

female: Female?

1: yes

hispanic: Are you of Hispanic origin?

hhwhite: Are you white?

nhblack: Are you Black or African American?

nhasian: Are you Asian or Pacific Islander?

raceoth: Are you of another race?

0: No 1: Yes

bornus: 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?

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: Somewhat unsure | 3: Neither sure or unsure |
| 4: Somewhat sure | 5: Very sure | 6: Never want to use birth control |

intlgnc: 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 |
| 4: Moderately above average | 6: Extremely above average |

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.

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

minfoc: 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.

hopeful: You felt hopeful about the future.*

failure: You thought your life had been a failure.

fearful: You felt fearful.

happy: You were happy.*

talkless: You talked less than usual.

lonely: You felt lonely.

unfrndly: People were unfriendly to you.

enjlfe: You enjoyed life.*

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 |
|----------|---------|----------|-----------|

(Variables marked with an asterisk (*) are coded as follows:

- | | | | |
|-----------|----------|---------|----------|
| 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

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

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

1: Not at all

2: Very little

3: Somewhat

4: Quite a bit

5: Very much

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

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: none

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)

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

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: none

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

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 agree nor disagree

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?

havensex: 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?

numcigs: During the past 30 days, on the days you smoked, how many cigarettes did you smoke each day?

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?

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

4: A few times a week

7: Daily

potlife: During your life, how many times have you used marijuana?

potlstmo: During the past 30 days, how many times did you use marijuana?

In the past 12 months, how often did you ...

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?

injuroth: 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)

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?

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

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

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: AME, 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 5: High

AHvocab: Add Health Picture Vocabulary Test standardized score

RAWvocab: Add Health Picture Vocabulary Test raw score

Suggestions for variable sets by model:

Linear Regression Model:
Y: AHvocab
C: hrstv
D: dadcoll
X: depress

Binary Regression Model:
Y: havesex
C: age
D: dadcoll
X: depress

Ordinal Logit Model:
Y: pray
C: actsmom
D: female
X: nhblack

Multinomial Logit Model:
Y: momrshp
C: age
D: dadcare
X: goodqual

Count Model:
Y: delinq
C: esteem
D: racewhite
X: adultcare

Last Revised 2009-05-19